

***IA Neuro-Simbólica para Descoberta de
Marcadores Genéticos e Epigenéticos de Longevidade***

Projeto de Pesquisa

Prof. Alberto Ferreira De Souza

Dados do Proponente

Nome: Alberto Ferreira de Souza.

Vínculo institucional: Professor Emérito da Universidade Federal do Espírito Santo (UFES), com vínculo formal à UFES, mais especificamente ao Departamento de Informática e ao Programa de Pós-Graduação em Informática, conforme RESOLUÇÃO Nº 26/99 do Conselho Universitário da UFES (https://daocs.ufes.br/sites/daocs.ufes.br/files/field/anexo/resolucao_26_1999.pdf).

Mini-Currículo: Alberto Ferreira de Souza é Professor Emérito de Ciência da Computação da Universidade Federal do Espírito Santo (UFES). Ele é Bacharel em Engenharia (Cum Laude) em Engenharia Eletrônica e Mestre em Ciências em Engenharia de Sistemas e Computação pela Universidade Federal do Rio de Janeiro (UFRJ), e Ph.D. em Ciência da Computação pelo *University College London*, Reino Unido. Ele é autor / co-autor de uma patente Norte Americana e de mais de 150 publicações. Foi editor dos anais de seis conferências nacionais e internacionais (três patrocinadas pelo IEEE) e é Senior Member do IEEE. O Professor de Souza ocupou as seguintes posições administrativas na UFES e em outros órgãos: Coordenador do curso de Engenharia de Computação - UFES (1994-1996), membro da Diretoria do Conselho Regional de Engenharia, Arquitetura e Agronomia do Espírito Santo (CREA-ES) - CREA-ES (1995-1996), Chefe do Departamento de Informática - UFES (2000), Vice-Diretor do Centro Tecnológico da UFES (2001-2004), Coordenador do Laboratório de Computação de Alto Desempenho (LCAD) - LCAD (2002-2019), Diretor Superintendente do Instituto de Tecnologia da UFES (2003-2004), Presidente do Comitê Gestor da Rede Metropolitana de Alta Velocidade de Vitória (METROVIX) - METROVIX (2005-2008), Pró-Reitor de Planejamento e Desenvolvimento Institucional da UFES (2004-2008), Diretor de Pesquisa da UFES (2012-2013), Membro do Conselho de Administração da Fundação Espírito-santense e de Tecnologia (FEST) - FEST (2004-2014), e Diretor Presidente do Instituto de Inteligência Computacional Aplicada (I2CA) do Espírito Santo. Alberto Ferreira de Souza é Comendador da Ordem Rubem Braga. CV Lattes: <http://lattes.cnpq.br/7573837292080522>.

Área do Conhecimento Predominante e Áreas do Conhecimento Correlatas

Grande Área: Ciência da Computação.

Área predominante: Inteligência Artificial (IA), subárea Aprendizado de Máquina.

Áreas correlatas: Bioinformática e Biologia Computacional (genômica), Gerontologia (longevidade) e Saúde Pública.

O projeto é multidisciplinar, combinando IA, ontologias, genômica e biologia do envelhecimento.

Orçamento Detalhado

Tabela 1: Recursos Aprovados pelo CNPq

Item	Descrição	Valor Unit. U\$	Valor Unit. R\$	Quant	Valor Total R\$
	Bolsas				
1	Bolsa de Pós-Doutorado no Exterior (PDE) - de um ano	34.886,00	189.940,29	1	R\$ 189.940,27
2	Bolsa de Desenvolvimento Tecnológico e Inovação no Exterior Sênior (DES) - de um mês	6.396,00	34.823,66	3	R\$ 104.470,98
3	Bolsa de Desenvolvimento Tecnológico e Inovação no Exterior Junior (DEJ) - de um mês	4.796,00	26.112,30	2	R\$ 52.224,60
	Itens de Custeio				
4	Passagens Aéreas Internacionais para Visitas Técnico-Científicas às Universidades Parceiras do Projeto e para Participação em Eventos	1.275,02	6.941,99	3	R\$ 20.825,98
5	Diárias Internacionais para Visitas Técnico-Científicas às Universidades Parceiras do Projeto e para Participação em Eventos	370,00	2.014,50	12	R\$ 24.174,02
		R\$/US\$	5,4445993		
			Total Bolsas		R\$ 346.635,85
			Total Custeio		R\$ 45.000,00
			Total		R\$ 391.635,85

Dados Gerais do Projeto

Título: IA Neuro-Simbólica para Descoberta de Marcadores Genéticos e Epigenéticos de Longevidade

Title: Neuro-Symbolic AI for the Discovery of Genetic and Epigenetic Longevity Markers

Palavras-chave: Inteligência Artificial, Aprendizado Profundo, IA Neuro-Simbólica, Longevidade, Genômica, Bioinformática

Keywords: Artificial Intelligence, Deep Learning, Neuro-Symbolic AI, Longevity, Genomics, Bioinformatics

Resumo

O envelhecimento humano é um dos maiores desafios científicos e sociais contemporâneos, pois está diretamente associado ao declínio funcional e ao aumento da incidência de doenças crônicas. Nas últimas décadas, avanços em genômica e biologia do envelhecimento revelaram que a taxa de envelhecimento pode ser modulada por fatores genéticos e epigenéticos, abrindo novas perspectivas de intervenção clínica. Dentro desse cenário, esta proposta busca aplicar métodos de inteligência artificial neuro-simbólica para identificar marcadores genéticos e epigenéticos associados à longevidade saudável, aproveitando a capacidade preditiva de redes neurais profundas e a interpretabilidade de ontologias e grafos de conhecimento biológico.

O problema central a ser abordado é a dificuldade de descobrir biomarcadores robustos e biologicamente interpretáveis de longevidade, o que é crucial para avançar na compreensão dos mecanismos do envelhecimento e apoiar o desenvolvimento de futuras terapias. Nossa hipótese é que modelos neuro-simbólicos, ao combinar predição neural e raciocínio baseado em conhecimento, podem identificar assinaturas moleculares da longevidade de forma mais precisa e explicável do que abordagens puramente neurais. Essa hipótese é sustentada por avanços recentes, como o uso de *transformers* genômicos e a aplicação de ontologias na biologia computacional, que já demonstraram capacidade de revelar padrões sutis em dados genômicos complexos.

A estratégia metodológica proposta integra (i) predição com um módulo *deep learning* de sequências genômicas de grande escala com (ii) um módulo simbólico para validação e explicação das predições, criando um pipeline neuro-simbólico com potencial de produzir descobertas interpretáveis.

Pesquisadores vinculados à *City St George's, University of London*, e à *University of Twente* atuarão como colaboradores acadêmicos de caráter consultivo, por meio de discussões técnico-científicas, intercâmbio de conhecimento, palestras, seminários e reuniões científicas. Não haverá, contudo, compartilhamento ou acesso aos dados genômicos utilizados no projeto, execução de etapas do mesmo pelas instituições estrangeiras, desenvolvimento de módulos sob responsabilidade dessas instituições, decisão metodológica ou responsabilidade institucional estrangeira. Essa interlocução acadêmica contribuirá para fortalecer a inserção internacional dos pesquisadores brasileiros e ampliar o impacto científico da pesquisa no campo da expansão da longevidade saudável.

Abstract

Human aging is one of the greatest contemporary scientific and societal challenges, as it is directly associated with functional decline and an increased incidence of chronic diseases. In recent decades, advances in genomics and the biology of aging have revealed that the rate of aging can be modulated by genetic and

epigenetic factors, opening new perspectives for clinical intervention. Within this scenario, this proposal aims to apply neuro-symbolic artificial intelligence methods to identify genetic and epigenetic markers associated with healthy longevity, combining the predictive capacity of deep neural networks with the interpretability of ontologies and biological knowledge graphs.

The central problem to be addressed is the difficulty of discovering robust and biologically interpretable biomarkers of longevity, which is crucial to advancing the understanding of aging mechanisms and supporting the development of future therapies. Our hypothesis is that neuro-symbolic models, by combining neural prediction and knowledge-based reasoning, can identify molecular signatures of longevity more accurately and explainably than purely connectionist approaches. This hypothesis is supported by recent advances, such as the use of genomic transformers and the application of ontologies in computational biology, which have already demonstrated the ability to reveal subtle patterns in complex genomic data.

The proposed methodological strategy integrates (i) prediction with a deep learning module trained on large-scale genomic sequences and (ii) a symbolic module for validating and explaining predictions, creating a neuro-symbolic pipeline with the potential to yield interpretable discoveries.

Researchers affiliated with City St George's, University of London, and the University of Twente will act as academic collaborators in a consultative capacity, through technical-scientific discussions, knowledge exchange, lectures, seminars, and scientific meetings. However, there will be no sharing of or access to the genomic data used in the project, execution of project stages by the foreign institutions, development of modules under the responsibility of those institutions, methodological decision-making, or foreign institutional responsibility. This academic interaction will contribute to strengthening the international engagement of the Brazilian researchers and expanding the scientific impact of the research in the field of healthy longevity extension.

Objetivo Geral

O objetivo geral deste projeto é desenvolver e empregar uma plataforma de inteligência artificial (IA) neuro-simbólica para a análise de dados genômicos de humanos e de espécies modelo, visando identificar biomarcadores de longevidade — genes, variantes genéticas e padrões epigenéticos associados à vida longa — de forma acurada, robusta e interpretável. Esse objetivo se desdobra na criação de um pipeline computacional capaz de prever a predisposição à longevidade a partir de dados genômicos brutos e, ao mesmo tempo, explicar biologicamente tais predições ao vinculá-las a processos moleculares conhecidos do envelhecimento, como reparo de DNA, estabilidade genômica e senescência celular.

A motivação central para o desenvolvimento do projeto reside no fato de que, apesar do envelhecimento ser o principal fator de risco para doenças crônicas e degenerativas, estudos recentes indicam que sua taxa pode ser modulada geneticamente e que intervenções em mecanismos moleculares podem atrasar ou até reverter aspectos do processo de envelhecimento humano. A identificação de biomarcadores de longevidade, contudo, ainda enfrenta o desafio de extrair conhecimento confiável de conjuntos massivos e heterogêneos de dados, exigindo métodos que combinem a capacidade preditiva de redes neurais profundas com a interpretabilidade do raciocínio simbólico. Nossa hipótese é que modelos neuro-simbólicos são particularmente adequados para esse propósito, pois permitem não apenas detectar assinaturas sutis em grandes volumes de sequências genômicas, mas também contextualizar suas descobertas à luz do conhecimento biológico existente.

A estratégia metodológica envolve o uso de modelos neurais profundos, como *transformers* genômicos de longo contexto, treinados com dados de múltiplas espécies e validados por meio de um módulo simbólico baseado em ontologias e grafos de conhecimento biológico. Esse módulo atuará tanto na validação das predições quanto na sua explicação, propiciando interpretações mais confiáveis e potencialmente aplicáveis

em contextos clínicos. Espera-se, com isso, não apenas avançar a compreensão dos mecanismos moleculares associados à longevidade, mas também abrir caminho para futuras intervenções terapêuticas baseadas em alvos validados por IA.

Os Profs. Alberto Ferreira de Souza (coordenador deste projeto) e Thiago Oliveira dos Santos, membros da equipe da UFES, possuem relevante experiência em redes neurais em geral e modelos neurais profundos em particular, bem como sua aplicação na área de saúde, conforme atestado por sua produção científica nestas áreas. O Prof. José Geraldo Mill, titular do Departamento de Ciências Fisiológicas da UFES, pesquisador 1A do CNPq e coordenador do centro ELSA (Estudo Longitudinal de Saúde do Adulto) no Espírito Santo, atuará na avaliação médica e fisiológica dos marcadores identificados, fornecendo a correlação clínica e epidemiológica necessária à validação biomédica dos resultados. A Profa. Débora Dummer Meira, também membro da equipe da UFES, possui relevante experiência em biologia molecular, genômica e biotecnologia, conforme também atestado por sua produção científica. Pesquisadores vinculados à *City St George's, University of London*, incluindo o Prof. Artur Garcez, e à *University of Twente*, incluindo o Prof. Giancarlo Guizzardi, atuarão como colaboradores acadêmicos de caráter consultivo, por meio de discussões técnico-científicas, intercâmbio de conhecimento, palestras, seminários e reuniões científicas. Não haverá, contudo, compartilhamento ou acesso aos dados genômicos utilizados no projeto, execução de etapas do mesmo pelas instituições estrangeiras, desenvolvimento de módulos sob responsabilidade dessas instituições, decisão metodológica ou responsabilidade institucional estrangeira. Essa interlocução acadêmica contribuirá para fortalecer a inserção internacional dos pesquisadores brasileiros e ampliar o impacto científico, tecnológico e social do projeto.

General Objective

The overall objective of this project is to develop and employ a neuro-symbolic artificial intelligence (AI) platform for the analysis of genomic data from humans and model species, aiming to accurately, robustly, and interpretably identify biomarkers of longevity – genes, genetic variants, and epigenetic patterns associated with extended lifespan. This objective unfolds into the creation of a computational pipeline capable of predicting predisposition to longevity from raw genomic data while simultaneously providing biological explanations for these predictions by linking them to well-established molecular processes of aging, such as DNA repair, genomic stability and cellular senescence.

The central motivation for this project lies in the fact that, although aging is the main risk factor for chronic and degenerative diseases, recent studies indicate that its rate can be genetically modulated and that interventions at the molecular level can delay or even reverse aspects of the human aging process. Identifying longevity biomarkers, however, still faces the challenge of extracting reliable knowledge from massive and heterogeneous datasets, which demands methods that combine the predictive power of deep neural networks with the interpretability of symbolic reasoning. Our hypothesis is that neuro-symbolic models are particularly well suited to this purpose, as they allow not only the detection of subtle signatures across large volumes of genomic sequences but also the contextualization of such discoveries in light of existing biological knowledge.

The methodological strategy involves the use of deep neural models, such as long-context genomic transformers, trained with data from multiple species and validated through a symbolic module built on biological ontologies and knowledge graphs. This module will operate both as a validation layer and as an explanatory component, providing more reliable and potentially clinically applicable interpretations. In doing so, the project aims not only to advance the understanding of molecular mechanisms associated with longevity but also to pave the way for future therapeutic interventions based on AI-validated targets.

Profs. Alberto Ferreira de Souza (coordinator of this project) and Thiago Oliveira dos Santos, members of the UFES team, have significant experience in neural networks in general and deep learning models in particular,

as well as their applications in healthcare, as evidenced by their scientific production in these areas. Prof. José Geraldo Mill, Full Professor at the UFES Department of Physiological Sciences, CNPq Researcher Level 1A, and coordinator of the ELSA (Longitudinal Study of Adult Health) center in Espírito Santo, will contribute to the medical and physiological evaluation of the identified markers, providing the clinical and epidemiological correlation necessary for the biomedical validation of the results. Prof. Débora Dummer Meira, also a member of the UFES team, has extensive experience in molecular biology, genomics, and biotechnology, likewise demonstrated by her scientific production. Researchers affiliated with City St George's, University of London, including Prof. Artur Garcez, and with the University of Twente, including Prof. Giancarlo Guizzardi, will act as academic collaborators in a consultative capacity, through technical-scientific discussions, knowledge exchange, lectures, seminars, and scientific meetings. However, there will be no sharing of or access to the genomic data used in the project, execution of project stages by the foreign institutions, development of modules under the responsibility of those institutions, methodological decision-making, or foreign institutional responsibility. This academic interaction will contribute to strengthening the international engagement of the Brazilian researchers and expanding the scientific, technological, and societal impact of the project.

1 - Introdução

A ciência há muito busca compreender e minimizar os efeitos negativos do processo de envelhecimento humano, responsável pelo nosso declínio funcional e predisposição a doenças crônicas. Sabe-se que o envelhecimento é o principal fator de risco para patologias como câncer [1], Alzheimer [2], doenças cardiovasculares [3] e outras. Ele é, em essência, caracterizado pela acumulação de danos químicos em nossas moléculas e células, pela perda progressiva de nossa integridade fisiológica, e o aumento de nossa vulnerabilidade à morte [4], [5]. Entretanto, pesquisas realizadas nas últimas décadas evidenciam que a taxa de envelhecimento não é fixa e pode ser modulada por vias genéticas, ou seja, alterações no DNA, e bioquímicas, ou seja, alterações no epigenoma [4], [5].

DNA, Epigenoma e Longevidade

O DNA é a molécula fundamental que armazena as instruções genéticas em todos os organismos vivos. Organizado em uma estrutura molecular em forma de dupla hélice, o DNA é composto por uma cadeia de nucleotídeos, ou bases nitrogenadas, que podem ser de quatro tipos diferentes: adenina (A), timina (T), guanina (G) ou citosina (C). Estas bases se emparelham de forma específica – A com T e G com C –, o que estabelece uma complementaridade entre as duas fitas da hélice. Esta complementaridade oferece redundância e resistência a falhas, e é fundamental em processos tais como a replicação celular [6].

A sequência de bases do DNA funciona como um código, escrito em um alfabeto de quatro letras, que especifica tudo que é necessário para a síntese de proteínas, o controle do funcionamento das células e, em última análise, do organismo como um todo – o DNA humano possui cerca de 3 bilhões de pares de bases. Essa longa molécula não se encontra nua no interior de nossas células. Ela está associada a proteínas chamadas histonas, formando a cromatina, que permite compactar de forma conveniente os quase dois metros de comprimento do nosso DNA dentro do núcleo microscópico de nossas células. A cromatina organiza-se em unidades discretas chamadas cromossomos – os seres humanos possuem 46 em cada célula somática (23 pares, que oferecem mais um nível de redundância e resistência a falhas). O grau de compactação da cromatina, controlado pela organização em nucleossomos (ver Figura 1), influencia diretamente no acesso às instruções genéticas e na regulação da expressão gênica (o quanto frequentemente trechos do DNA são transcritos em RNA [6]).

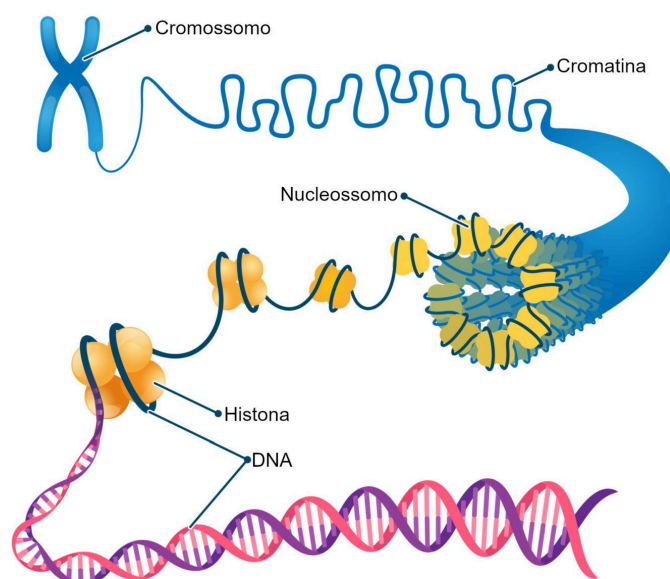


Figura 1: DNA, histona, nucleossomo, cromatina e cromossomo.

O DNA herdado dos pais – 23 cromossomos da mãe e 23 do pai – é fixo e o mesmo em todas as células de um indivíduo (exceto os óvulos e espermatozoides, que só possuem metade do número de cromossomos, que são uma mistura dos da mãe com os do pai). Além da sequência fixa do DNA, o funcionamento de cada célula depende de um conjunto de marcações químicas mais dinâmicas – presentes nele e nas histonas – conhecido como epigenoma [6]. Essas marcações, como a metilação do DNA e a acetilação das histonas [6], não alteram a sequência de bases do DNA, mas modulam a forma como a informação genética é lida dele. Por meio destas marcações, células com o mesmo DNA podem assumir identidades distintas, como neurônios, células da pele ou fibras musculares.

O epigenoma é suficientemente estável para manter a identidade celular ao longo do tempo, mas é também flexível para responder a sinais internos e externos, ajustando a expressão dos genes conforme necessário. Assim, as interações entre o código genético e as marcações epigenéticas possibilitam, tanto a diversidade funcional das células em um organismo multicelular, quanto a sua adaptabilidade às condições ambientais.

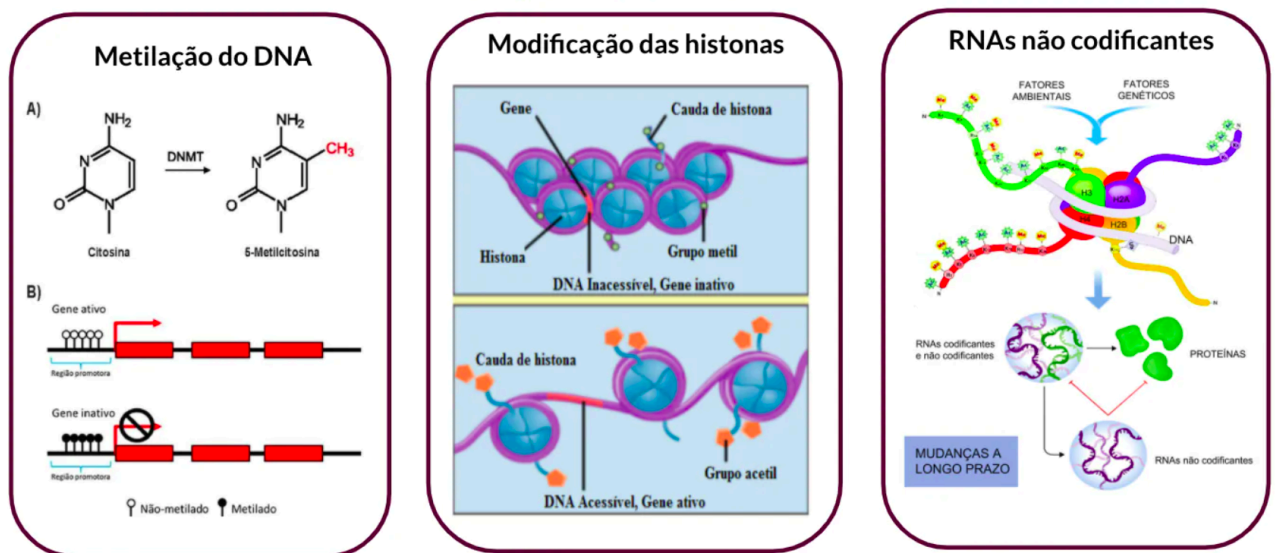


Figura 2: Informação epigenética: Metilação do DNA; Modificações de histonas; RNAs não codificantes (<https://brainlatam.com/blog/neuroepigenetica-e-sociedade-como-o-ambiente-pode-modificar-a-expressao-de-nossos-genes-modulando-nossa-neurofisiologia-e-comportamento-2977>).

A Figura 2 ilustra os três tipos fundamentais de informação epigenética. O primeiro é a metilação do DNA, geralmente ocorrendo na citosina (em mamíferos, cerca de 99% da metilação ocorre na citosina [7]), que atua como um sinal químico capaz de silenciar genes ao reduzir a acessibilidade dos mesmos à maquinaria bioquímica de transcrição. O segundo são as modificações das histonas, como acetilação, metilação ou fosforilação, que alteram o grau de compactação da cromatina e, consequentemente, facilitam ou dificultam a leitura do DNA. Por fim, os RNAs não codificantes, que exercem funções regulatórias ao se ligar a moléculas de RNA mensageiro ou interagir com a cromatina, modulando a expressão gênica de forma dinâmica [6].

O epigenoma pode ser compreendido como um sistema que controla a expressão do DNA (que partes do mesmo estão ativas ou inativas) por meio de escritas e leituras químicas sobre o DNA, sobre a cromatina e por meio de RNA regulatório advindo do DNA. Esse processo de escrita e leitura químicas no DNA e cromatina é mediado fundamentalmente por proteínas especializadas chamadas enzimas epigenéticas [6], que são produzidas a partir da própria informação genética do organismo. Algumas atuam como escritores (*writers*), adicionando marcas químicas específicas; outras como leitores (*readers*), reconhecendo essas marcas e transmitindo sinais à maquinaria celular; e outras como apagadores (*erasers*), removendo marcações quando necessário [6]. Essas enzimas são codificadas pelos genes e sintetizadas pelas células em resposta a programas de desenvolvimento do organismo ou a sinais ambientais – durante a formação de um organismo, elas estabelecem os padrões que guiam a diferenciação celular; no organismo adulto, continuam

regulando a atividade dos genes para manter o funcionamento celular e permitir ajustes dinâmicos às condições internas e externas.

Evidências científicas mostram que alterações no DNA e no epigenoma, que se acumulam ao longo da vida, estão associadas ao envelhecimento. Os “marcadores do envelhecimento” (*Hallmarks of Aging* [4], [5]), isto é, instabilidade genômica, encurtamento de telômeros, alterações epigenéticas, perda de proteostase, senescência celular, entre outros, fornecem um arcabouço conceitual que explica como a disfunção molecular acumulada ao longo do tempo pelo organismo leva à degeneração celular e tecidual [4], [5]. Contudo, intervenções em modelos animais já demonstraram que é possível atrasar ou reverter aspectos do envelhecimento ao atacar esses marcadores [8], [9]. Por exemplo, a expressão transitória de fatores de reprogramação celular (fatores de Yamanaka [10]) em camundongos induziu rejuvenescimento epigenético, revertendo sinais moleculares da idade e prolongando a vida de animais com envelhecimento acelerado [11]. Essa reprogramação parcial restaurou funções fisiológicas e reforçou a ideia de que o envelhecimento em mamíferos pode ser tratado como um “problema de software”: um epigenoma corrompido, ou em estado avançado de desenvolvimento, poderia ser reinicializado para um estado mais jovem [12]. Em linha com esta ideia, algumas terapias gênicas focadas em longevidade obtiveram sucesso notável – em um estudo feito por Bernardes de Jesus et al. [13], uma única aplicação de telomerase em camundongos aumentou sua vida em até 24%, sem incidência extra de câncer. Esses avanços pioneiros ilustram o potencial de intervenções deste tipo na extensão da longevidade saudável.

IA e a Análise de Dados Biológicos em Larga Escala

O avanço da Inteligência Artificial (IA) tem transformado nossa capacidade de analisar dados biológicos em larga escala [14]. Trabalhos seminais, como o AlphaFold2 [15] e o RoseTTAFold [16] para predição de estruturas de proteínas, ESM2 [17] e ESM3 [18] para geração de novas proteínas, e o AlphaFold3 [19] para a predição de estruturas de proteínas, DNA, RNA, íons e outras moléculas associadas à vida, utilizaram redes neurais profundas com mecanismos de atenção (*transformers*) [20] para modelar e estudar estruturas biomoleculares e interações entre elas de maneira eficaz.

No trabalho científico em que foi proposto o Enformer [21], por exemplo, foram empregadas arquiteturas *transformer* [20] para prever expressão gênica e estados da cromatina a partir de sequências de DNA. Mais recentemente, essa abordagem foi adotada em *single-cell genomics* (o estudo genômico em nível de célula individual, permitindo analisar a variabilidade entre células que compõem um mesmo tecido ou organismo) por modelos como o scGPT [22], o GeneFormer [23] e o scBERT [24]. Essas redes neurais profundas conseguem detectar padrões genéticos sutis em conjuntos de dados genômicos complexos, revelando biomarcadores e alvos terapêuticos que passariam despercebidos por métodos tradicionais [25].

Em 2023, a DeepMind (subsidiária do Google que criou a série AlphaFold) apresentou o AlphaMissense [26], um modelo capaz de prever a patogenicidade de variantes (mutações) missense – alterações pontuais no DNA que causam a substituição de um único aminoácido em uma proteína. Ele é uma adaptação do AlphaFold2, que aproveita o contexto estrutural de proteínas previsto pelo AlphaFold2 e dados de frequência populacional de variantes para avaliar os efeitos das mutações. O modelo prevê a probabilidade de uma variante missense ser clinicamente patogênica, classificando-a como provável benigna, provável patogênica ou incerta [26].

Neste ano de 2025 (27 de junho), a DeepMind apresentou o AlphaGenome, um modelo que recebe como entrada uma longa sequência de DNA – de até 1 milhão de bases – e prediz diversas propriedades moleculares que caracterizam sua atividade regulatória [27]. Além disso, AlphaGenome pode avaliar os efeitos de variantes genéticas, comparando suas predições associadas a sequências mutadas com predições associadas a sequências não mutadas. No contexto deste projeto, redes neurais profundas, como AlphaFold3, Enformer, AlphaMissense, AlphaGenome e outras que buscaremos criar, serão investigadas e empregadas na predição e/ou identificação de assinaturas genéticas e epigenéticas associadas à longevidade saudável [28], [29].

Redes Neurais Profundas

Redes neurais profundas são tipicamente empregadas para emular nossas capacidades cognitivas e a inteligência derivada destas capacidades. Nossas capacidades cognitivas fundamentais estão relacionadas aos nossos sentidos e aos diversos tipos de memória que possuímos. De destaque, podemos citar (i) a nossa memória de longo prazo de dados sensoriais e conceitos associados, (ii) nossa memória procedural, (iii) nossa memória episódica, e (iv) nossa memória de trabalho [30].

Os recentes avanços na área de aprendizado profundo (*Deep Learning* [31]) nos permitem hoje implementar sistemas com memória de longo prazo de dados sensoriais e conceitos associados que rivalizam com a memória humana. Uma rede neural profunda (*Deep Neural Network* – DNN [31]) que recebe como entrada uma imagem e retorna uma classe associada a cada objeto detectado na imagem (Figura 1, [32]) nada mais é que um tipo de memória de longo prazo de dados sensoriais e conceitos associados.

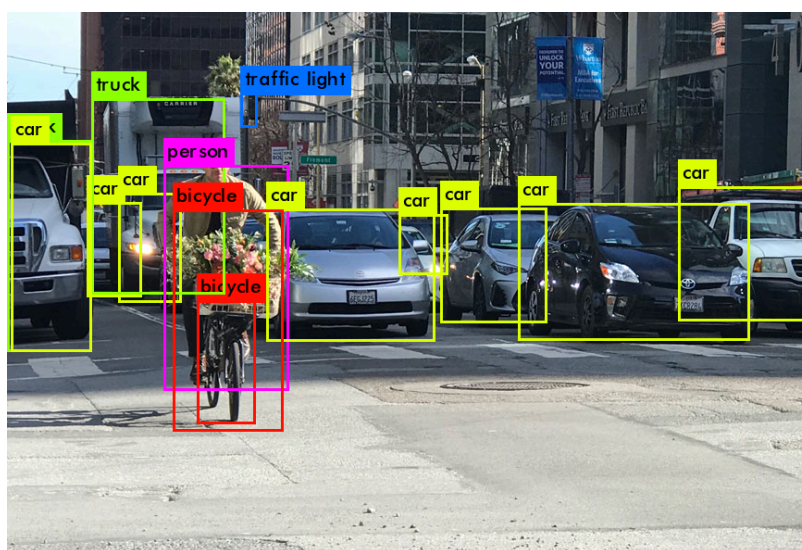


Figura 1: Saída de uma rede neural profunda para detecção de objetos em imagens.

Sistemas com poderosas memórias procedurais, que são também memórias de longo prazo, já foram implementados por meio de redes neurais, especialmente *Long Short Term Memory* (LSTM [33]) e *transformers* [20], acopladas eventualmente a outros tipos de redes neurais profundas. Hoje já existem, também, aplicações em que memórias procedurais artificiais rivalizam em desempenho com alguns tipos de memória procedural humana, como aquelas associadas à conversão de voz em texto [34] e à produção de texto [35], [36]. Estas duas aplicações de *transformers*, ou, mais especificamente, *Generative Pre-trained transformers* (GPTs [36]), demonstram o poder desta arquitetura de rede neural profunda. Sistemas neurais modernos com tal capacidade são hoje comumente designados sistemas de Inteligência Artificial Generativa.

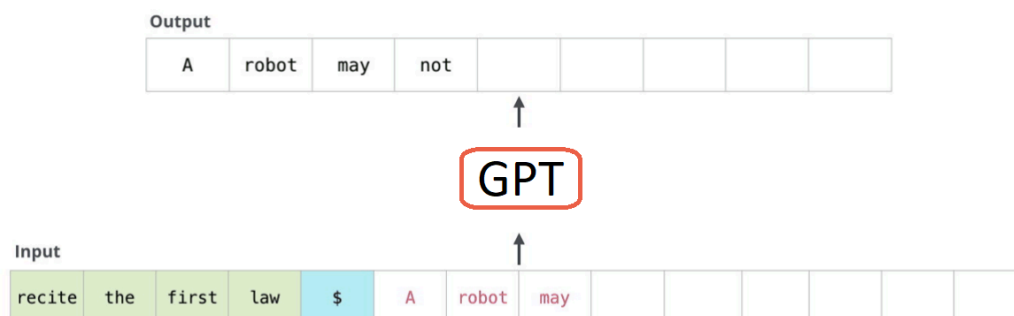


Figura 2: *Generative Pre-trained Transformer* (GPT).

Uma GPT (Figura 2) é um tipo de rede neural profunda que pode ser empregada na implementação de modelos de linguagem de larga escala (*Large Language Models* – LLMs [36], [37]). GPTs são capazes de gerar

textos semelhantes aos produzidos por humanos a partir de uma entrada baseada em texto, ou *prompt*, fornecida. Um usuário “alimenta” a GPT com o *prompt*, que é seguido de um marcador de fim (na Figura 2, o *prompt* é “*recite the first law*” e o marcador de fim, “\$”), e a GPT gera informações coerentes a partir deste *prompt* em formato de texto, um *token* de cada vez (um *token* representa uma palavra, parte de uma palavra, pontuação, símbolos, etc.). Cada *token* gerado é inserido na entrada e contribui para geração de um próximo *token* até que a GPT gere um *token* de fim texto (não mostrado na Figura 2). Para ser capaz de gerar textos (ou outros dados com sons e imagens) coerentes de saída, uma GPT precisa ser treinada com bilhões ou até trilhões de *tokens* advindos da Internet, livros, etc. A GPT-3 [36], precursora da ChatGPT [38] e da recente OpenAI GPT-5 (<https://openai.com/index/introducing-gpt-5/>), foi treinada com 300 bilhões de *tokens*. O treinamento consistiu na modificação dos valores das sinapses que interligam os neurônios da GPT-3, além de outros parâmetros da arquitetura, na direção de fazê-la predizer (a saída de uma GPT é estocástica [36]) o próximo *token* da sequência de *tokens* presentes no conjunto de dados de treinamento.

As sinapses constituem a maior parte do total de parâmetros que definem a dimensão (ou complexidade) de uma GPT. A GPT-3, por exemplo, possui cerca de 175 bilhões de parâmetros [36] que precisam ser ajustados no seu processo de treinamento. Os detalhes exatos do tempo de treinamento da GPT-3 não foram divulgados publicamente; no entanto, sabemos que o treinamento de LLMs como a GPT-3 requer hoje uma quantidade significativa de tempo e de recursos computacionais. O treinamento de um modelo de linguagem dessa magnitude pode levar vários meses, dependendo do hardware e da configuração utilizados. Além disso, otimizações e ajustes no modelo podem prolongar ainda mais o tempo de treinamento.

Uma vez concluído o treinamento, os parâmetros de uma GPT não são mais alterados e, na produção de textos (ou outros dados de saída) a partir de *prompts* de entrada, a GPT deve se valer apenas do conhecimento memorizado em suas sinapses e outros parâmetros da arquitetura durante o treinamento. Deste modo, GPTs são, na verdade, memórias procedurais estáticas de longo prazo – elas possuem memória sobre como devem proceder para gerar, sequencialmente, *tokens* coerentes de saída a partir de *tokens* de entrada.

Há ainda muito o que se investigar e desenvolver nas áreas de memórias de longo prazo de dados sensoriais e conceitos associados e de memórias procedurais, mas os grandes desafios científicos e tecnológicos estão hoje no desenvolvimento de sistemas computacionais que viabilizem a implementação de memórias episódicas e de trabalho. Nossa memória episódica, que é também uma memória de longo prazo, nos permite armazenar incrementalmente sequências de eventos conectados no tempo dos quais participamos (episódios); enquanto que, nossa memória de trabalho, de curto prazo, agrupa informações sensoriais com dados recuperados de todas as nossas memórias de longo prazo (de dados sensoriais e conceitos associados, procedural, e episódica) para produzir uma gama de comportamentos. A janela de contexto das GPTs (onde são depositados os *prompts* e onde a GPT escreve sua saída) emula nossa memória de trabalho e hoje já é super humana – janelas de contexto de bilhões de *tokens* podem armazenar, e dar acesso à GPT, ao equivalente a dezenas de livros simultaneamente.

Tecnologia que emule apropriadamente nossa memória episódica, se disponível para robôs e outros sistemas autonomizáveis, vai revolucionar a indústria e outros setores da atividade humana (o uso de GPTs para este fim já vem sendo investigado [39], [40], [41], [42]). Ou seja, é de fundamental importância o desenvolvimento de tecnologia para agregar memória de longo prazo adicional automaticamente (aprendizado de máquina contínuo [43], [44]). Existem várias doenças humanas, tais como Alzheimer e diversos tipos de amnésia causadas por trauma, que mostram como ficamos debilitados quando perdemos a capacidade de formar novas memórias de longo prazo. Assim, sistemas baseados em IA generativa serão mais poderosos e úteis se forem capazes de formar novas memórias de longo prazo continuamente.

Ontologias, Grafos de Conhecimento e IA Neuro-Simbólica

Acreditamos que a combinação de técnicas de IA generativa com ontologias [45], [46] e grafos de conhecimento [47] pode ser uma alternativa para a implementação de mecanismos para agregar memória de longo prazo adicional automaticamente a sistemas de IA. Além disso, acreditamos que a integração de

técnicas de IA generativa com conhecimento biológico estruturado, por exemplo, ontologias de dados genéticos [48], [49], [50], [51] e grafos de conhecimento médico [48], [52], tem o potencial de permitir não só prever biomarcadores de longevidade, mas também de manter a interpretabilidade desta predição, propiciando a elucidação de seus mecanismos subjacentes.

Ontologias são representações formais e explícitas de domínios de conhecimento e, no contexto destes domínios, elas permitem definir conceitos, entidades e relações de maneira não ambígua [45], [46]. Em bioinformática e genômica, por exemplo, ontologias podem ser usadas para padronizar a terminologia (descrições de genes, fenótipos, doenças, etc.) e fornecer uma semântica compartilhada para dados biológicos de diferentes fontes [51]. Seu uso propicia que sistemas de IA incorporem conhecimento prévio e “compreendam” melhor os dados, aumentando a interoperabilidade entre bases de dados distintas e a coerência das inferências geradas [50], [51]. Além disso, o uso de ontologias fundamentais (*foundational ontologies*), como a UFO (*Unified Foundational Ontology*) proposta por Guizzardi, Almeida e colegas, assegura que diferentes ontologias de domínio estejam alinhadas a princípios ontológicos comuns – facilitando a integração de informações e a interpretação consistente dos resultados [46].

Grafos de conhecimento, por sua vez, organizam informações como um conjunto de entidades (nós) e suas relações (arestas) dentro de um grafo semântico [47]. Nesses grafos, cada nó pode representar, por exemplo, uma entidade biológica (como um gene, uma proteína ou uma droga) e cada aresta pode representar um relacionamento significativo entre elas (por exemplo, “codifica”, “interage com” ou “regula”) [48], [52]. Diferentemente de bases de dados tradicionais, os grafos de conhecimento integram dados heterogêneos sob um vocabulário unificado (que pode ser definido por uma ontologia), permitindo consultas complexas e inferências sobre conexões indiretas na base de conhecimento. Essa abordagem já impulsionou avanços importantes – por exemplo, na descoberta de novos alvos terapêuticos e fármacos, integrando dados genômicos, clínicos e químicos em um único grafo consultável [48], [52]. Ao proporcionar um arcabouço para raciocínio simbólico automatizado, os grafos de conhecimento permitem que algoritmos de IA identifiquem padrões de forma mais interpretável, pois cada predição pode ser relacionada a caminhos e relações bem definidas no grafo subjacente.

A IA neuro-simbólica, por outro lado, refere-se a abordagens híbridas que unem métodos simbólicos de IA (baseados em conhecimento explícito, lógicas e regras) com métodos conexionistas (redes neurais e aprendizado estatístico) [53], [54], [55], [56]. O objetivo dessa integração é aproveitar o melhor dos dois mundos: de um lado, a capacidade de aprendizado de padrões complexos fornecida pelas redes neurais; de outro, a interpretabilidade e o raciocínio explícito proporcionados pelas representações simbólicas [55]. Conforme arguido por Garcez e Lamb [56], sistemas neuro-simbólicos bem projetados podem em muitos casos alcançar desempenho preditivo comparável ao de modelos puramente neurais e, ao mesmo tempo, fornecer explicações para suas predições.

A combinação de aprendizado profundo com conhecimento estruturado tem sido considerada por vários pesquisadores como uma possível “terceira onda” da IA, pois promete conciliar a aprendizagem neural em larga escala com o raciocínio baseado em conhecimento [56]. Em estudos sobre longevidade, por exemplo, um modelo neuro-simbólico poderia tirar proveito de um grafo de conhecimento das vias moleculares do envelhecimento (construído sob uma ontologia pré-definida), juntamente com dados genômicos de larga escala, para gerar predições sobre fatores pró-longevidade, que viriam acompanhadas de justificativas baseadas em mecanismos biológicos conhecidos e compreensíveis ao pesquisador.

Há boas razões para caracterizar, por exemplo, o AlphaFold3 [19] (AF3) como IA neuro-simbólica (o Prêmio Nobel de Química de 2024 foi concedido a Demis Hassabis e John Jumper, da Google DeepMind, pelo desenvolvimento da família de modelos AlphaFold, e a David Baker, pela sua contribuição em design computacional de proteínas). Além de seus componentes neurais de última geração – Pairformer e módulo de difusão generativa [19] –, o pipeline deste sistema incorpora e opera sobre informações estruturadas que funcionam como “conhecimento explícito”. Na entrada da arquitetura do AF3, as MSAs¹ e os templates estruturais em formato PDB² resumem homologia e correspondência espacial de dados de entrada de forma

¹ *Multiple Sequence Alignments* [57] – alinhamentos de sequências biológicas homólogas que revelam posições e padrões conservados em diferentes espécies.

² *Protein Data Bank* [58] – base de dados de estruturas 3D de macromoléculas determinadas experimentalmente.

simbólica. Na saída de AF3, as perdas (*loss*) por violações estereoquímicas impõem restrições físico-químicas ao resultado e, crucialmente, todo o ecossistema no qual AF3 se apoia e para o qual gera resultados é ontologicamente estruturado e grafo-centrado: o PDBx/mmCIF³ atua como um dicionário/ontologia de dados para estruturas biomoleculares; o SIFTS⁴ mapeia, no nível de resíduo (bases individuais), UniProt⁵ a PDB; a UniProt RDF⁶/OWL⁷ e a *Gene Ontology* (*Gene Ontology Consortium* [51]), juntamente com iniciativas como o PDBe-KB⁸ e o 3D-Beacons⁹, definem ontologia e grafo de conhecimento federado que conectam estruturas experimentais e preditas a anotações e evidências experimentais.

Nessa arquitetura, o aprendizado é continuamente guiado e validado por estruturas e restrições simbólicas, e as predições são interpretáveis no contexto de grafos de conhecimento biomoleculares, alinhando-se, assim, à definição de Garcez, Lamb e outros para integração neuro-simbólica entre aprendizado e representação/raciocínio (ver detalhamento da arquitetura de AF3 nos materiais suplementares de [19] e em <https://elanapearl.github.io/blog/2024/the-illustrated-alphafold/>). Vale destacar que a arquitetura de AlphaGenome [27] possui, entre suas saídas, uma saída ontológica.

Um dos objetivos deste projeto de pesquisa será incorporar mais aspectos neuro-simbólicos à arquitetura de AlphaGenome ou de sistemas equivalentes. Acreditamos que essa sinergia entre IA generativa, ontologias, grafos de conhecimento e biologia acelerará descobertas de alvos genéticos e/ou epigenéticos para ampliação da longevidade saudável e viabilizará o design in silico de intervenções promissoras. Uma vez identificadas intervenções promissoras, técnicas de edição genética de precisão – como sistemas CRISPR-Cas9 (*Clustered Regularly Interspaced Short Palindromic Repeats* - Cas9) [57], [58], [59], [60] e suas variantes (Cas12a [61], edição de bases [62] e *prime editing* [63]) – podem ser empregadas para testá-las e, assim, desenvolver terapias inovadoras para a ampliação da longevidade saudável [64].

Edição de DNA

O CRISPR-Cas9 originou-se na natureza no contexto do sistema imunológico de bactérias e archaea, nas quais fragmentos de DNA viral conhecidos pelo seu sistema imunológico (as sequências “CRISPR”) guiam enzimas Cas para cortar material genético invasor [65]. Em 2012, pesquisadores liderados por Jennifer Doudna e Emmanuelle Charpentier demonstraram que esse mecanismo poderia ser adaptado para editar genomas de células eucarióticas [59], inaugurando a era da edição gênica simplificada. Em reconhecimento a essa descoberta, Doudna e Charpentier receberam o Prêmio Nobel de Química de 2020.

Molecularmente, o complexo Cas9-RNA [57] busca uma sequência alvo no DNA de uma célula em que esteja e induz uma quebra de dupla fita [57], [59]. Os mecanismos de reparo celular, então, entram em ação, possibilitando inserções ou deleções imprecisas no DNA (via junção não-homóloga de pontas [59]) ou edições precisas se um molde de reparo for fornecido (reparo dirigido por homologia [57]). Diferentemente das técnicas anteriores de edição gênica, como as nucleases de dedos de zinco (ZFNs [58]) e as nucleases baseadas em efetores TAL (TALENs [58], [66]), o CRISPR-Cas9 é muito mais fácil de projetar e empregar, demandando apenas a síntese de um RNA guia correspondente ao alvo desejado, em vez de reengenharia proteica complexa. Essa simplicidade e eficiência, aliadas ao menor custo, tornaram o CRISPR amplamente acessível e aceleraram sua adoção em laboratórios do mundo todo.

³ PDB Exchange / macromolecular Crystallographic Information File (<https://mmcif.mrc.org/>, [58]) – esquema de dados usado pelo PDB, comumente referenciado junto com mmCIF como PDBx/mmCIF.

⁴ *Structure Integration with Function, Taxonomy and Sequences* [59].

⁵ *Universal Protein Resource* [60] – uma base abrangente de conhecimento sobre proteínas.

⁶ *Resource Description Framework* [61] – um modelo W3C para intercâmbio de dados na web usando triplas sujeito-predicado-objeto.

⁷ *Web Ontology Language* (<https://www.w3.org/OWL/>) – uma linguagem W3C para definir ontologias e permitir raciocínio lógico na web.

⁸ *Protein Data Bank in Europe – Knowledge Base* [62] – uma base de conhecimento comunitária integrando anotações em torno das entradas PDB.

⁹ *3D-Beacons Network* [63] – uma rede federada que expõe estruturas 3D experimentais e preditas de proteínas via uma API comum.

Do ponto de vista histórico, após a prova de conceito em 2012-2013 por Doudna e Charpentier, houve uma explosão de aplicações de CRISPR em múltiplos organismos. Em poucos anos, surgiram diversas variantes da plataforma original, por exemplo: nucleases Cas alternativas, como a Cas12a [61], foram descobertas; e técnicas derivadas, como a edição de bases [62], que permite trocar uma única letra do DNA sem corte de fita dupla, e o *prime editing*, que realiza as tarefas de “buscar e substituir” em sequências específicas [63], foram desenvolvidas para aumentar a precisão e o espectro de edições possíveis. Essas inovações mitigaram limitações do Cas9 clássico, reduzindo efeitos fora do alvo e permitindo correções pontuais de mutações causadoras de doenças.

A medicina tem sido transformada pelo CRISPR nos últimos anos. Em 2023 foi aprovado o primeiro tratamento somático utilizando edição gênica por CRISPR-Cas9, voltado para curar a anemia falciforme [67], [68]. Ensaios clínicos demonstraram resultados muito relevantes – 29 de 31 pacientes tratados permaneceram livres de crises vaso-oclusivas graves causadas pela doença por pelo menos um ano após o tratamento, evidenciando o potencial curativo da abordagem [67]. A aprovação do tratamento pela FDA (EUA) e pela agência regulatória do Reino Unido em finais de 2023 [68] sinalizou o início de uma nova era de terapias gênicas de precisão.

Além da anemia falciforme, uma série de terapias experimentais com CRISPR estão em desenvolvimento clínico. Por exemplo, editando-se geneticamente células do sistema imunológico, pesquisadores criaram linfócitos T modificados para atacar cânceres resistentes [69]. Em outra frente, a empresa Intellia Therapeutics demonstrou pela primeira vez a edição gênica in vivo (dentro do corpo) ao inativar um gene no fígado de pacientes com amiloidose hereditária: uma única infusão de nanopartículas com CRISPR resultou em reduções superiores a 90% na proteína patogênica (transtirretina) de forma duradoura, sem eventos de segurança graves [70]. Adicionalmente, novas variantes deste tipo de terapia, como as mencionadas edição de bases [70] e *prime editing* [68], começam a entrar em testes clínicos. Em maio de 2025, foi reportado o primeiro uso do *prime editing* em um paciente humano [68] – um adolescente com imunodeficiência crônica, no qual a técnica corrigiu mutações genéticas e restaurou a função de enzimas-chave em suas células imunológicas. Esses avanços ilustram o progresso rápido rumo a terapias genéticas de dose única para diversas doenças.

Edição do Epigenoma

Assim como a edição do DNA, a edição do epigenoma utiliza a plataforma CRISPR, mas com nucleases cataliticamente inativas (dCas9) acopladas a enzimas moduladoras de expressão gênica [71]. Esta técnica de edição do epigenoma permite ligar ou desligar genes sem alterar a sequência de DNA ou induzir quebras na dupla hélice [71]. Por exemplo, a fusão da dCas9 à DNA metiltransferase 3A (dCas9-DNMT3A) realiza a metilação de promotores específicos, silenciando genes-alvo; inversamente, a fusão de dCas9 à enzima TET1 (dCas9-TET1) remove marcas de metilação e pode reativar genes silenciados [71]. De forma análoga, domínios reguladores de transcrição também podem ser editados – dCas9-KRAB (inibidores) para interferência com CRISPRi e dCas9-VPR ou similares (ativadores) para ativação com CRISPRa – inibindo ou aumentando a expressão de genes endógenos sem necessidade de qualquer corte no DNA. Ferramentas CRISPR permitem também modificações de histonas de forma seletiva, recriando estados ativos ou reprimidos da cromatina [71].

Já existem vários exemplos relevantes de edição de epigenoma para o tratamento de doenças. Em neurônios, por exemplo, já se demonstrou o silenciamento preciso de genes neurotóxicos ligados à doença de Alzheimer: a entrega de um editor dCas9-DNMT3A reduziu significativamente os níveis da proteína tau patogênica em neurônios por meio da metilação dirigida do gene *APP* (*Amyloid Precursor Protein* gene) [72]. Estratégia semelhante foi empregada para modular genes metabólicos: pesquisadores usaram um editor epigenético para silenciar o gene *PCSK9* no fígado (via metilação da região promotora), reduzindo a expressão de receptores de LDL e abrindo caminho para diminuir o colesterol de forma controlada [73].

No contexto de oportunidades de ampliação da longevidade saudável, as tecnologias de edição epigenética abrem horizontes para intervenções inéditas. Uma aplicação evidente é o alongamento de telômeros em células somáticas: ao contrário da edição gênica tradicional, que inseriria novos genes de

telomerase, pode-se usar CRISPRa para reativar o gene endógeno *TERT* (gene que codifica a telomerase). De fato, um estudo recente mostrou que a ativação epigenética do promotor de *TERT* em linfócitos T humanos quiescentes estendeu seu tempo de replicação e retardou a entrada em senescência celular em pelo menos três meses [74]. Essa estratégia usou dCas9 acoplada a ativadores de transcrição, e enzimas como p300/HAT e TET1 para remodelar localmente a cromatina, induzindo a expressão de telomerase de forma controlada. O rejuvenescimento desses linfócitos ocorreu sem sinais de transformação maligna ou proliferação desregulada: as células mantiveram sua identidade e não exibiram instabilidade genômica, indicando que a intervenção epigenética foi segura e específica [74]. Em princípio, abordagens análogas poderiam ser aplicadas para eliminar células senescentes ou neutralizar seus efeitos deletérios nos tecidos envelhecidos (por exemplo, induzindo seletivamente a apoptose via ativação/silenciamento de genes chave) ou, ainda, reativar genes de regeneração que são silenciados com a idade. Como essas intervenções epigenéticas não alteram a sequência genética subjacente, elas oferecem um meio de restaurar padrões de expressão gênica juvenis por meio de alterações epigenéticas associadas ao envelhecimento.

Por fim, vale destacar que a expressão transitória de fatores de pluripotência (fatores de Yamanaka, [10]) pode “rejuvenescer” células envelhecidas, revertendo marcas de idade e restaurando funções perdidas; porém, acarretam riscos de tumorigênese e perda da identidade celular. Ferramentas CRISPR epigenéticas podem oferecer uma alternativa: pesquisas demonstraram que é possível induzir células-tronco pluripotentes a partir de fibroblastos humanos utilizando apenas CRISPRa para ativar os loci endógenos de OCT4, SOX2, KLF4, MYC e LIN28A [75]. Ademais, estudos de reprogramação parcial em modelos animais sugerem que reverter somente parte do epigenoma ao estado jovem já é suficiente para melhorar múltiplos aspectos do envelhecimento – incluindo restauração da capacidade regenerativa tecidual, prolongamento da saúde celular e até extensão da vida [11]. Em teoria, editores epigenéticos poderiam reproduzir esses efeitos ao amenizar o “*drift*” epigenético associado ao envelhecimento de forma precisa, restaurando padrões de expressão de genes benéficos sem necessidade de inserções ou deleções gênicas.

Desafios e Oportunidades

Não obstante os relevantes resultados recentes, desafios de entrega (*delivery*) do produto do editor genético ou epigenético aos tecidos alvo permanecem como um obstáculo central. Muitos dos sucessos iniciais envolvem células sanguíneas editadas *ex vivo*, o que requer transplante celular – um procedimento complexo. Entregar complexos CRISPR diretamente a órgãos como pulmões ou cérebro ainda esbarra em barreiras de biodistribuição, resposta imunológica ao vetor ou à proteína Cas9, e limitações de tamanho de vetores virais. Estratégias em investigação correntemente incluem o uso de vetores virais adeno-associados modificados [76], partículas lipídicas nanoparticuladas [77] e até novas proteínas Cas9 menores [78], para alcançar células somáticas específicas de forma eficaz e segura. A eficiência, especificidade e segurança dos sistemas de entrega precisam ser ampliadas para viabilizar todo o potencial terapêutico dessas técnicas de edição genética e epigenética.

Outra preocupação relevante é a ocorrência de mutações fora do alvo (*off-target*), ou seja, edições não intencionais em sequências semelhantes ao alvo [79]. Embora melhorias nas plataformas CRISPR e métodos de bioinformática tenham reduzido significativamente esse risco, é importante o acompanhamento de longo prazo dos pacientes tratados para monitorar efeitos inesperados.

Inteligência Artificial para Genômica

Desenvolver uma terapia de ampliação da longevidade saudável efetiva e personalizada é extremamente ambicioso. Exige enfrentar a complexidade biológica do envelhecimento – um processo multifatorial – e navegar pelas etapas de descoberta, validação e aplicação clínica, normalmente longas e custosas. No entanto, o impacto potencial é enorme: estender a vida saudável da população transformaria economias, reduzindo drasticamente o fardo de doenças relacionadas à idade, por exemplo.

Do ponto de vista de investimento, a área de longevidade tem atraído aportes bilionários globalmente. Iniciativas como Calico Labs (<https://www.calicolabs.com>), Altos Labs (<https://www.altoslabs.com>) e outras

startups e entidades sem fins lucrativos atuando na área são indicativos do interesse crescente – e já vem produzindo importantes resultados, como a recente criação de CRISPR sintético por meio de uma rede neural GPT ([80] <https://www.youtube.com/watch?v=EnbfoFUFm2s>).

Neste projeto de pesquisa, buscaremos desenvolver ciência e tecnologia que viabilizem a implementação de inteligências artificiais (IAs) para o apoio à investigação na área de genômica. Mais especificamente, buscaremos investigar como empregar IA neuro-simbólica para análise de dados genômicos de humanos e espécies modelo, a fim de identificar biomarcadores de longevidade (genes, variantes genéticas ou padrões epigenéticos associados à longevidade) de forma precisa e interpretável.

2 - Objetivo Geral

O objetivo geral deste projeto é desenvolver e empregar uma plataforma de inteligência artificial (IA) neuro-simbólica para a análise de dados genômicos de humanos e de espécies modelo, visando identificar biomarcadores de longevidade — genes, variantes genéticas e padrões epigenéticos associados à vida longa — de forma acurada, robusta e interpretável. Esse objetivo se desdobra na criação de modelos computacionais capazes de prever a predisposição à longevidade a partir de dados genômicos brutos e, ao mesmo tempo, explicar biologicamente tais predições ao vinculá-las a processos moleculares conhecidos do envelhecimento, como reparo de DNA, estabilidade genômica, senescência celular e autofagia celular [4], [5].

A motivação central reside no fato de que, apesar do envelhecimento ser o principal fator de risco para doenças crônicas e degenerativas [1], [2], [3], estudos recentes indicam que sua taxa pode ser modulada geneticamente e que intervenções em mecanismos moleculares podem atrasar ou até reverter aspectos do processo [4], [5]. A identificação de biomarcadores de longevidade, contudo, ainda enfrenta o desafio de extrair conhecimento confiável de conjuntos massivos e heterogêneos de dados, exigindo métodos que combinem a capacidade preditiva de redes neurais profundas com a interpretabilidade do raciocínio simbólico. Nossa hipótese é que modelos neuro-simbólicos [53], [54], [55], [56] são particularmente adequados para esse propósito, pois permitem não apenas detectar assinaturas sutis em grandes volumes de sequências genômicas, mas também contextualizar suas descobertas à luz do conhecimento biológico existente.

Nossa estratégia metodológica envolve o uso de arquiteturas de deep learning de última geração, como *transformers* genômicos de longo contexto [27], [21], treinados com dados de múltiplas espécies e validados por meio de um módulo simbólico baseado em ontologias [45], [46] e grafos de conhecimento biológico [47], [48], [49], [50], [51], [52]. Esse módulo atuará tanto na validação das predições quanto na sua explicação, propiciando interpretações mais confiáveis e potencialmente aplicáveis em contextos clínicos. Espera-se, com isso, não apenas avançar a compreensão dos mecanismos moleculares associados à longevidade, mas também abrir caminho para futuras intervenções terapêuticas baseadas em alvos validados por IA.

Os Profs. Alberto Ferreira de Souza (coordenador deste projeto) e Thiago Oliveira dos Santos, membros da equipe da UFES, possuem relevante experiência em redes neurais em geral e modelos neurais profundos em particular, bem como sua aplicação na área de saúde, conforme atestado por sua produção científica nestas áreas. O Prof. José Geraldo Mill, titular do Departamento de Ciências Fisiológicas da UFES, pesquisador 1A do CNPq e coordenador do centro ELSA (Estudo Longitudinal de Saúde do Adulto) no Espírito Santo, atuará na avaliação médica e fisiológica dos marcadores identificados, fornecendo a correlação clínica e epidemiológica necessária à validação biomédica dos resultados. A Profa. Débora Dummer Meira, também membro da equipe da UFES, possui relevante experiência em biologia molecular, genômica e biotecnologia, conforme também atestado por sua produção científica. Pesquisadores vinculados à *City St George's, University of London*, incluindo o Prof. Artur Garcez, e à *University of Twente*, incluindo o Prof. Giancarlo Guizzardi, atuarão como colaboradores acadêmicos de caráter consultivo, por meio de discussões técnico-científicas, intercâmbio de conhecimento, palestras, seminários e reuniões científicas. Não haverá, contudo, compartilhamento ou acesso aos dados genômicos utilizados no projeto, execução de etapas do mesmo pelas instituições estrangeiras, desenvolvimento de módulos sob responsabilidade dessas instituições, decisão metodológica ou responsabilidade institucional estrangeira. Essa interlocução acadêmica

contribuirá para fortalecer a inserção internacional dos pesquisadores brasileiros e ampliar o impacto científico, tecnológico e social do projeto.

3 - Objetivos Específicos

Para atingir o objetivo geral proposto, definimos os seguintes objetivos específicos:

1. **Investigar modelos (redes neurais profundas) genômicos** capazes de analisar sequências longas de DNA e identificar *features* de interesse. Em particular, serão investigados *transformers* genômicos de longo contexto (Enformer [21] e AlphaGenome [27], por exemplo) para capturar dependências de longa distância nas sequências, visando detectar variantes regulatórias, motivos em promotores (*sequence motifs*) ou padrões epigenéticos ligados à manutenção celular, reparo de DNA, estabilidade genômica, senescência celular, regeneração tecidual, etc. Faz parte deste objetivo específico desenvolver um Módulo Neural para nosso *pipeline de análise genômica baseado em IA neuro-simbólica*.
2. **Integrar conhecimento biológico estruturado** ao nosso *pipeline de análise genômica baseado em IA neuro-simbólica*. Isso inclui investigar/desenvolver ontologias do domínio do envelhecimento e grafos de conhecimento sobre o tema, e usá-los como parte da entrada, e/ou para treinar/validar e/ou explicar as previsões do Módulo Neural. Concretamente, será desenvolvido um Módulo Simbólico capaz de fazer inferências sobre as bases curadas (*Human Ageing Genomic Resources* - HAGR [81], [82], LongevityMap [83], GenAge [84], etc.) associadas a cada *feature* detectada pelo Módulo Neural, filtrando sinais espúrios e enriquecendo os resultados com informações de contexto (funções da *feature*, vias relacionadas, disfunções associadas à mesma relatadas na literatura, etc.). Usaremos as saídas deste Módulo Simbólico para, também, fazer *fine-tuning* dos modelos genômicos do Módulo Neural em busca de melhorar o desempenho do sistema como um todo. Empregaremos uma abordagem inspirada em *Reinforcement Learning from Human Feedback* (RLHF [85]) e alternativas mais recentes de *reinforced learning* [86] para este *fine-tuning*, mas usando feedback do Módulo Simbólico.
3. **Identificar genes, variantes e assinaturas genéticas e epigenéticas candidatas à longevidade** por meio da análise de dados reais com o nosso *pipeline de análise genômica baseado em IA neuro-simbólica*. Usaremos o pipeline, a ser desenvolvido dentro do escopo dos Objetivos Específicos 1 e 2 para a análise dos dois conjuntos de dados genômicos humanos previstos neste projeto: o Welllderly Study, como grupo experimental, e o 1000 Genomes Project – high coverage, como grupo controle (ver a Seção 5 - Metodologia). Espera-se identificar um conjunto de marcadores genéticos e/ou epigenéticos de longevidade, priorizados por relevância estatística e biológica.
4. **Interpretar e Validar os marcadores identificados** à luz do conhecimento existente. Cada marcador genético ou epigenético identificado será verificado quanto a: (a) presença em bancos de longevidade (HAGR [81], [82], LongevityMap [83], GenAge [84], etc.), (b) evidências em literatura (estudos prévios que o associam à longevidade ou doenças relacionadas ao envelhecimento), e (c) participação em vias reconhecidas do envelhecimento (por exemplo, gene envolvido em reparo de DNA, resposta ao estresse oxidativo, autofagia celular, etc.). Essa etapa possibilitará consolidar os achados e gerar hipóteses biológicas plausíveis.
5. **Explorar correlações evolutivas** de longevidade através de genômica comparativa. Como objetivo exploratório, os marcadores identificados em humanos serão comparados com outros previamente descobertos em espécies longevas (como as baleias-da-Groenlândia ou as tartarugas-das-galápagos, por exemplo) para verificar se há convergência evolutiva. Estudos indicam que espécies de vida longa evoluíram mecanismos robustos de manutenção celular, como conservação de vias de reparo de DNA e de supressão tumoral. Analisaremos se os marcadores genômicos encontrados também aparecem nessas espécies, o que pode reforçar sua importância na longevidade.

4 - Metas e Indicadores da Proposta

Para alcançar o objetivo geral e objetivos específicos associados propostos, foram definidas as seguintes metas, acompanhadas de indicadores de sucesso mensuráveis:

- M1. Protótipo de Pipeline de Análise Genômica Baseado em IA Neuro-Simbólica: Desenvolver um pipeline computacional unindo deep learning genômico e raciocínio simbólico até o 12º mês. Indicadores: código-fonte funcional; documento de arquitetura; demonstração por meio da aplicação do pipeline em dados de teste (este protótipo já está em desenvolvimento – <https://github.com/albertodesouza/genomics>).
- M2. Identificação de Marcadores Genéticos ou Epigenéticos de Longevidade: Aplicar o pipeline a dados reais e obter um conjunto de 2 ou mais marcadores candidatos (genes/variantes) associados à longevidade até o 18º mês. Indicadores: relatório técnico listando marcadores descobertos, com evidências estatísticas (e.g., p-values de associação).
- M3. Validação Cruzada com Conhecimento Biológico: Verificar se os marcadores candidatos têm suporte em bases de conhecimento ou em outros conjuntos de dados até o 21º mês. Indicadores: tabela comparativa citando estudos/entradas relevantes de bancos de dados que corroboram cada marcador e métricas de validação.
- M4. Disseminação Científica: Publicar 2 trabalhos científicos resultantes do projeto (um em conferência internacional de IA/Bioinformática qualificada e um em periódico qualificado) até o 24º mês. Indicadores: submissão/aceite de artigos; apresentação em evento internacional; disponibilização de software na forma de código aberto (repositório público) – o código já se encontra em desenvolvimento (<https://github.com/albertodesouza/genomics>).
- M5. Formação de Recursos Humanos: Orientação de um aluno de iniciação científica, um aluno de mestrado e um aluno de doutorado. Indicadores: lista de alunos sob orientação trabalhando em temas do projeto; iniciação científica concluída (já há bolsa disponível vinculada a projeto paralelo); experiência internacional dos alunos brasileiros por meio de visitas às instituições às quais estão vinculados os pesquisadores colaboradores estrangeiros, para atividades de intercâmbio acadêmico e científico.

O cumprimento dessas metas será monitorado trimestralmente. Os indicadores quantitativos acima permitirão avaliar objetivamente o progresso e impacto do projeto.

5 - Metodologia

O problema de identificar marcadores genéticos e epigenéticos de longevidade apresenta alta complexidade biológica e computacional, pois a dimensão do espaço de busca (genomas inteiros, com bilhões de bases e múltiplas combinações de variantes e modificações epigenéticas) torna inviável uma abordagem manual. Trata-se de um caso típico em que ferramentas de Inteligência Artificial (IA) podem contribuir. De fato, este problema enquadra-se nas características de um problema apropriado para IA, conforme enunciado por Demis Hassabis (Nobel de Química 2024) em sua *Nobel Lecture “Accelerating Scientific Discovery with AI”* (https://www.youtube.com/watch?v=yxAJohmOl_g), pois ele envolve: (1) um espaço de busca combinatorial massivo (as combinações de bases do DNA e as marcações epigenéticas possíveis sobre o mesmo), (2) uma função objetivo clara (identificar marcadores genéticos ou epigenéticos associados à longevidade, medidos por métricas de predição da longevidade) e (3) abundância de dados para aprendizado (temos abundância de dados biológicos e possibilidades amplas de realizar simulações, isto é, experimentos com animais modelo).

Hassabis argumenta que a essência da solução de muitos problemas científicos está em aprender um modelo do escopo do problema e usar esse modelo para guiar uma busca otimizada segundo uma função objetivo de interesse, o que se aplica desde à jogos complexos (por exemplo, encontrar as melhores jogadas

em Go [87]) até à descoberta de moléculas de interesse em espaços bioquímicos. Nossa metodologia segue esse paradigma: a partir de grandes volumes de dados biológicos, treinaremos modelos neuro-simbólicos para que aprendam representações (modelos do escopo do problema), internas aos modelos, da genômica e sua associação a diversos marcadores conhecidos (de características positivas, negativas, de doenças, ou mesmo de longevidade saudável). Em seguida, utilizaremos esses modelos para explorar o enorme espaço de possíveis variantes e combinações genotípicas, realizando predições guiadas por critérios de maximização da longevidade saudável (ou minimização de riscos de doenças associadas à idade).

Em uma etapa puramente computacional, pode-se, por exemplo: separar dados para treino, validação e teste do sistema; treinar o sistema com os dados de treino; e examinar sua capacidade de explorar e fazer predições sobre os dados de validação; e finalmente avaliar o sistema sobre os dados de teste. Caso o sistema obtenha bom desempenho na etapa computacional, suas predições poderão futuramente servir de base para estudos experimentais específicos, inclusive em modelos animais, sujeitos às autorizações éticas e regulatórias próprias. Tais experimentos não fazem parte do escopo do presente protocolo.

Em termos práticos, neste projeto examinaremos apenas a abordagem puramente computacional. Adotaremos um enfoque neuro-simbólico em três etapas principais para atingir o objetivo geral e os objetivos específicos do projeto. As etapas são: (i) Seleção e Preparação de Dados e Conhecimento, (ii) Desenvolvimento do Pipeline Neuro-Simbólico, e (iii) Experimentos e Validação dos resultados. A seguir detalhamos cada etapa buscando mostrar, técnica e cientificamente, como elas permitirão alcançar nossos objetivos e metas (M1–M5, Seção 4 deste projeto) estabelecidas.

(i) - Seleção e Preparação de Dados e Conhecimento

Nesta etapa, reuniremos e organizaremos os dados genômicos e o conhecimento biológico associado que darão suporte ao treinamento do modelo neural do pipeline neuro-simbólico, e à validação do mesmo. As principais fontes de dados incluem:

- Genomas e variantes humanos: Desenvolveremos código para acesso a dados genéticos humanos de alta qualidade e ampla cobertura (já estamos desenvolvendo, na verdade – ver <https://github.com/albertodesouza/genomics>). Em especial, usaremos bases públicas de variação genética humana, como a do *1000 Genomes Project* [88] e coortes de indivíduos longevos.
- Dados epigenéticos (marcadores de envelhecimento): Integraremos, também, perfis epigenéticos associados ao envelhecimento. Em particular, utilizaremos dados de metilação de DNA correlacionada à idade, como os obtidos pelos “*epigenetic clocks*” de Steve Horvath [7]. Esses “relógios epigenéticos” fornecem uma estimativa da idade biológica de tecidos/células baseada em padrões de metilação do DNA. Ao incluir tais perfis, esperamos capturar assinaturas epigenéticas de envelhecimento acelerado ou retardado, enriquecendo o conjunto de características a serem aprendidas pelo modelo. Assim, o pipeline poderá considerar marcas epigenéticas dinâmicas além da sequência estática do DNA.
- Conhecimento biológico estruturado: Paralelamente à obtenção dos dados brutos, buscaremos desenvolver uma ontologia dos *Hallmarks of Aging* (marcos biológicos do envelhecimento, [4], [5]) e/ou empregaremos ontologias existentes (por exemplo, a gene ontology [50], [51]) para categorizar genes/variantes e marcadores epigenéticos conforme processos do envelhecimento aos quais estejam ligados. Essas ontologias permitirão anotar, por exemplo, se um gene ou variante X está envolvida em estabilidade genômica, senescência celular, proteostase, etc. Além disso, integraremos grafos de conhecimento associados a bases curadas, como a HAGR [81], [82] (que inclui bases como a GenAge [84] e a LongevityMap [83]). Essas bases de conhecimento estruturado permitirão anotações automatizadas de dados de entrada de nosso pipeline de IA e validação das predições de saída do Módulo Neural (ver etapa (ii)).

Destacamos que esta pesquisa não realizará recrutamento direto de participantes, nem coleta prospectiva de material biológico, nem contato com indivíduos. O estudo utilizará exclusivamente dados genômicos secundários provenientes de bases já existentes, com acesso obtido de acordo com as regras

próprias de cada repositório. Restringiremos os datasets de genomas humanos a serem utilizados a apenas dois: (i) o dataset “Welllderly Study”, como grupo experimental, e (ii) o dataset 1000 Genomes, como grupo controle.

O dataset Welllderly Study, identificado como EGAD00001003941, vinculado ao estudo EGAS00001002306, intitulado *Whole-Genome Sequencing of a Healthy Aging Cohort*, está disponível no European Genome-phenome Archive (EGA): <https://ega-archive.org/studies/EGAS00001002306>. O acesso a esse conjunto de dados foi solicitado por meio de troca de e-mails com o Prof. Ali Torkamani, do *Scripps Research Institute*, pesquisador responsável pela concessão do acesso ao referido dataset. O Prof. Ali Torkamani está vinculado ao *Scripps Research* e sua página institucional é: <https://www.scripps.edu/faculty/torkamani/>. O *Scripps Research Institute* tem como página institucional: <https://www.scripps.edu/>.

O estudo de referência do dataset Welllderly é: Erikson, G. A. et al. *Whole-Genome Sequencing of a Healthy Aging Cohort*. *Cell*, v. 165, n. 4, p. 1002–1011, 2016. Conforme descrito neste trabalho, os indivíduos da coorte Welllderly apresentam idade entre 80 e 105 anos, tendo sido a maioria recrutada entre 80 e 85 anos; o critério básico de inclusão envolvia idade superior a 80 anos; além disso, o fenótipo “Welllderly” foi caracterizado pela ausência de doenças crônicas relevantes e pelo não uso contínuo de medicamentos para condições crônicas. O conjunto específico de 511 indivíduos analisado no estudo genômico decorreu de um processo adicional de filtragem de uma coorte inicial maior, preservando apenas indivíduos com ancestralidade europeia igual ou superior a 95% e com grau de parentesco máximo de 12,5%, conforme o desenho analítico adotado pelos autores.

No caso do grupo controle, será utilizado o dataset 1000 Genomes Project, repositório aberto e amplamente utilizado em genômica humana, acessível em <https://www.internationalgenome.org/>. O recorte empregado nesta pesquisa corresponde aos dados de alta cobertura do projeto, disponíveis em https://www.internationalgenome.org/data-portal/data-collection/1000genomes_30x. Trata-se de uma base com 3.202 indivíduos de diferentes populações do mundo, cuja seleção original não foi orientada por critérios de longevidade saudável, o que a torna adequada, no presente projeto, como base de dados do grupo controle.

Dessa forma, os critérios de inclusão do presente estudo passam a ser os seguintes: serão incluídos no grupo experimental exclusivamente indivíduos pertencentes ao dataset Welllderly Study efetivamente autorizados e disponibilizados para análise no âmbito do acesso concedido ao pesquisador coordenador, observando-se as características da própria coorte original, isto é, indivíduos com idade superior a 80 anos e fenótipo de envelhecimento saudável, conforme definidos no estudo de origem. Serão incluídos no grupo controle exclusivamente indivíduos pertencentes ao dataset 1000 Genomes Project, no recorte selecionado para esta pesquisa, proveniente da base pública correspondente. Os participantes do 1000 Genomes Project não foram originalmente escolhidos com base em características de longevidade saudável.

Os critérios de exclusão passam a ser: exclusão de quaisquer indivíduos que não pertençam a um desses dois datasets; exclusão de amostras destes datasets sem qualidade técnica mínima para processamento computacional; e exclusão de duplicidades ou de amostras com inconsistências de identificação nestes datasets.

Vale ressaltar que, as bases de dados HAGR (Human Ageing Genomic Resources) [81], [82], LongevityMap [83] e GenAge [84] não contém genomas de indivíduos, mas sim listas de ferramentas computacionais, de genes, de referências a trabalhos científicos, etc., sendo de livre acesso na Internet.

Empregaremos técnicas de pré-processamento, incluindo filtros de qualidade de dados genômicos (remoção de variantes de baixa cobertura ou potencialmente artefatuais) e armazenaremos os dados coletados em servidores do Laboratório de Computação de Alto Desempenho (LCAD – www.lcad.inf.ufes.br) da Universidade Federal do Espírito Santo (UFES). O LCAD/UFES já possui supercomputador com grande capacidade computacional e de armazenamento, o Atena (<https://ufes.br/conteudo/evento-debate-inteligencias-artificiais-aplicadas-e-inaugura-supercomputador>).

Empregaremos técnicas de pré-processamento, incluindo filtros de qualidade de dados genômicos (remoção de variantes de baixa cobertura ou potencialmente artefatuais) e armazenaremos os dados coletados em servidores do Laboratório de Computação de Alto Desempenho (LCAD – www.lcad.inf.ufes.br) da Universidade Federal do Espírito Santo (UFES). O LCAD/UFES já possui supercomputador com grande

capacidade computacional e de armazenamento, o Atena (<https://ufes.br/conteudo/evento-debate-inteligencias-artificiais-aplicadas-e-inaugura-supercomputador>), que permitirá um adequado início de projeto. Buscaremos, também, recursos adicionais para pesquisa que viabilizem ampliar a capacidade computacional e de armazenamento disponíveis para este projeto. Toda a manipulação de dados genômicos será feita segundo o “Plano de Governança, Segurança e Tratamento de Dados” do Anexo 1.

Reconhece-se que, em estudos genômicos, existe risco residual de reidentificação dos participantes, ainda que os dados sejam pseudonimizados, bem como risco associado a eventual vazamento ou acesso não autorizado aos dados, o que pode comprometer a confidencialidade das informações genéticas. Tais riscos decorrem da natureza altamente informativa dos dados genômicos e são amplamente reconhecidos na literatura científica e nas diretrizes éticas aplicáveis.

No presente estudo, não há interação direta com os participantes nem coleta prospectiva de dados; contudo, considera-se que a inclusão dos dados individuais nas análises computacionais implica risco indireto relacionado à privacidade e à proteção das informações.

Para mitigação desses riscos, serão adotadas as seguintes medidas:

- utilização de dados exclusivamente provenientes dos repositórios e bases previstos neste projeto, respeitando-se suas respectivas modalidades e políticas de acesso, incluindo o acesso controlado ao Welllderly Study por meio do European Genome-phenome Archive (EGA) e o acesso aberto ao 1000 Genomes Project;
- cumprimento integral dos termos de uso e das restrições de acesso impostas pelos respectivos repositórios e bases de origem;
- armazenamento e processamento dos dados em ambiente computacional seguro (supercomputador Atena / LCAD / UFES), protegido por controles institucionais de segurança da informação (ver Apêndice 1);
- controle rigoroso de acesso, restrito aos pesquisadores e bolsistas autorizados, com autenticação individual (ver Apêndice 1);
- proibição de cópia dos dados para ambientes não controlados (ver Apêndice 1);
- divulgação apenas de resultados agregados e não identificáveis, sem qualquer tentativa de reidentificação.

Assim, embora o risco seja considerado baixo, ele não é inexistente, sendo devidamente reconhecido e tratado por meio de medidas técnicas e administrativas compatíveis com a natureza sensível dos dados, em conformidade com a Resolução CNS nº 466/2012.

(ii) - Desenvolvimento do Pipeline Neuro-Simbólico

A segunda etapa de nossa metodologia corresponde ao desenvolvimento de um pipeline de IA neuro-simbólico capaz de aprender padrões genômicos e integrar conhecimento biológico. Nosso pipeline possui dois módulos principais que operam em conjunto: o Módulo Neural e o Módulo Simbólico. A seguir detalhamos cada módulo e como eles interagem.

Módulo Neural

O núcleo do Módulo Neural será um modelo próprio para sequências genômicas de grande comprimento. Examinaremos modelos tais como: o Enformer, proposto por Avsec et al. (DeepMind) [21] e projetado para integrar informações de até cerca de 200 mil bases e prever a expressão gênica e estados da cromatina com grande acurácia; ou o AlphaGenome, proposto também por Avsec et al. (DeepMind) [27] e projetado para integrar informações de até 1 milhão de bases e prever a expressão gênica e estados da cromatina com acurácia no nível de bases individuais. O emprego do Enformer, AlphaGenome ou outras arquiteturas de última geração, permitirá que nosso modelo leia segmentos genômicos relevantes inteiros

(regiões cromossômicas até 1 milhão de bases) de uma só vez, capturando elementos distantes como promotores, *enhancers* e outros elementos regulatórios no contexto de genes de interesse.

Para treinar os modelos de nosso Módulo Neural, adotaremos um procedimento de duas fases. Primeiro, realizaremos um pré-treinamento (*pre-training*) auto supervisionado dos *transformers* genômicos em grandes corpora de sequências de DNA, para que ele aprenda padrões genômicos básicos. Por exemplo, poderemos expor o modelo a todo o genoma humano (3 bilhões de bases) de vários indivíduos de modo equivalente ao pré-treinamento de modelos de linguagem natural. Em seguida, realizaremos ajuste fino supervisionado (*fine-tuning*) direcionado ao problema da longevidade (caso o código e os parâmetros de um modelo do porte de AlphaGenome esteja disponível, consideraremos seguir direto para esta fase de *fine-tuning*). Para isso, utilizaremos conjuntos de dados de treinamento anotados com dados sobre longevidade. Esse conjunto incluirá exemplos positivos, isto é, humanos longevos, versus controles, além de marcadores conhecidos de longevidade para o modelo aprender. Sabe-se, por exemplo, de variantes genéticas associadas à longevidade graças a estudos anteriores presentes na literatura – por exemplo, alelos em genes como o APOE (como o alelo $\epsilon 2$ protetor, ou ausência do alelo de risco, $\epsilon 4$ [89]) e variantes em genes como o FOXO3 [89].

A saída principal do Módulo Neural será um conjunto de *features* associados à longevidade. Essas *features* podem assumir várias formas: por exemplo, variantes, motivos regulatórios em regiões promotoras, assinaturas de metilação de DNA peculiares, ou até variações estruturais de DNA que o modelo considere relevantes (por exemplo, inversões de trechos, ausências de trechos, etc.). Importante destacar que nosso enfoque não é apenas preditivo, mas também interpretativo: junto com as predições, buscaremos, por exemplo, extrair pontuações de atenção do *transformer* e empregar métodos de interpretabilidade (como *saliency maps*) para identificar por que o modelo está focando em certos trechos das sequências de entrada. Ou seja, se o modelo destaca uma região genética, buscaremos saber se é devido a uma mutação específica, a um motivo na sequência, ou alguma outra característica aprendida. Essas saídas servirão de ponto de partida para o Módulo Simbólico realizar validações e explicações.

Grande esforço será dedicado à modelagem dos dados de entrada e saída do Módulo Neural. No caso dos dados de entrada, precisaremos transformá-los para que se adaptem à entrada do modelo neural a ser desenvolvido/empregado, e, no caso dos dados de saída, precisaremos transformá-los do formato típico de uma saída de modelo neural para um formato que se adapte à entrada do Módulo Simbólico. Buscaremos inspiração em trabalhos anteriores, tais como os que apresentaram o AlphaFold3 [19], o Enformer [21] e AlphaGenome [27].

Módulo Simbólico

Em paralelo ao Módulo Neural, desenvolveremos um Módulo Simbólico, que irá consumir as *features* detectadas pelo Módulo Neural e confrontá-las com o conhecimento biológico existente. Este módulo tem duas funções principais e complementares: (a) validação biológica das predições do modelo; e (b) descrição dessas predições.

Para a função de validação, o sistema simbólico atuará como um filtro de consistência com o estado da arte. Cada predição do Módulo Neural será verificada contra as bases de conhecimento e literatura disponíveis. Podemos empregar, ainda, o AlphaMissense [26] para indentificar a possibilidade da predição (se for predição de variante) ser provável benigna, provável patogênica ou incerta. Suponha, por exemplo, que o Módulo Neural apontou uma variante no gene *IGF-1R* como associada à longevidade. Neste caso, o Módulo Simbólico fará consultas às bases de dados, usando as ontologias do envelhecimento e grafos de conhecimento associados a serem desenvolvidos, para checar evidências conhecidas envolvendo o *IGF-1R*. Descobrirá que *IGF-1R* é parte da via de sinalização insulina/IGF-1, a qual é um caminho metabólico clássico ligado à regulação da longevidade em múltiplos organismos (por exemplo, mutações de *IGF-1R* prolongam a vida em modelos animais [90]). Essa informação de validação da predição, associada eventualmente à uma indicação de benignidade por AlphaMissense, pode, então, ser empregada em ajuste fino adicional do Módulo Neural ou em um loop de *reinforced learning*.

Caso o Módulo Neural aponte para *features* que o Módulo Simbólico possua conhecimento para descartar, esta informação de erro de predição pode ser usada do mesmo modo. Por outro lado, se o Módulo Neural apontar para uma região genômica ou variante em um gene sem registro prévio de ligação com envelhecimento, o Módulo Simbólico a sinalizará como um “potencial novo achado”, exigindo maior escrutínio e testes (mas sem descartá-lo de imediato, pois pode ser uma descoberta genuína). Assim, a validação simbólica atua tanto para confirmar e negar predições, quanto para priorizar aquelas totalmente novas que merecem investigação. A forma de comunicação entre o Módulo Neural e o Módulo Simbólico é um dos mais importantes tópicos de pesquisa deste projeto.

Para a função de descrição das predições, o Módulo Simbólico irá agregar metadados e contexto biológico a cada *feature* reportada. Estes metadados serão apresentados junto com o resultado do Módulo Neural, explicando por que aquela variante pode ser relevante. Esse enriquecimento semântico torna as previsões do Modelo Neural mais interpretáveis para pesquisadores humanos e a forma de implementá-lo é também um importante tópico de pesquisa deste projeto.

Iteração Neuro-Simbólica (Fechando o Laço de Aprendizado)

Um aspecto importante da nossa metodologia é a integração interativa entre o Módulo Neural e o Módulo Simbólico. Inspirados por técnicas de aprendizado por reforço como RLHF [85] e alternativas mais recentes [86], implementaremos um loop de realimentação onde as saídas do Módulo Simbólico influenciam o ajuste dos parâmetros do modelo neural do Módulo Neural em múltiplas iterações. Em termos práticos, funcionará assim: inicialmente, o Módulo Neural gera um conjunto de predições. O Módulo Simbólico, então, avalia e anota essas predições, conferindo a algumas um valor de confiança (por estarem alinhados com conhecimento prévio) e marcando outras como duvidosas ou desconhecidas. Essas avaliações serão usadas para refinar o treinamento do Módulo Neural. Implementaremos um mecanismo de aprendizado por reforço, ou *fine-tuning* supervisionado, onde predições confirmadas pelo conhecimento recebem uma recompensa, enquanto que predições que contrárias ou sem explicação recebem uma penalização na função objetivo (*loss*). Isso incentivará o Módulo Neural, em novas épocas de treinamento, a focar em padrões que tenham mais plausibilidade biológica.

Ao longo de múltiplos ciclos, esperamos que o desempenho preditivo melhore junto com a interpretabilidade. Em última instância, essa estratégia de fechar o laço cria um sistema de IA cujo critério de sucesso não é apenas acurácia preditiva, mas acurácia aliada à consistência com dados simbólicos. Conceitualmente, é análogo a uma abordagem de *reinforcement learning* onde um agente de IA aprende com recompensas fornecidas por conhecimento biológico consolidado. Essa etapa iterativa coloca o conhecimento simbólico no ciclo de treinamento, fazendo o sistema proposto aproximar-se do ideal de IA explicável.

(iii) - Experimentos e Validação

Com o pipeline desenvolvido conforme descrito nas etapas anteriores, partiremos para os experimentos e avaliação de resultados em diferentes cenários. Dividimos essa fase em sub-fases para uma validação incremental do nosso pipeline de análise genômica baseado em IA neuro-simbólica:

Fase de Experimentos Controlados (Simulações)

Inicialmente, executaremos o pipeline em cenários controlados ou semi-sintéticos cujo *ground truth* é conhecido e contém uma quantidade pequena de dados, a fim de verificar se o pipeline funciona conforme esperado antes do seu uso em um volume maior de dados. Por exemplo, podemos simular genomas onde inserimos manualmente “assinaturas de longevidade”, isto é, podemos selecionar genomas reais e introduzir neles variantes artificiais de efeito conhecido em longevidade – por exemplo, adicionar alelos benéficos em *APOE* [89] ou *FOXO3* [89] em algumas amostras e rotulá-las como longevas – para examinar se o pipeline consegue detectá-las.

A expectativa nessa fase é ajustar hiperparâmetros e identificar eventuais falhas em ambiente controlado – se o pipeline não conseguir detectar uma variante óbvia inserida no processo de simulação, isso indica a necessidade de ajustes (seja nos limiares estatísticos, na implementação da arquitetura do Módulo Neural ou do Módulo Simbólico, na comunicação entre, nas regras simbólicas, etc.). Essa etapa de *sanity check* garantirá que o pipeline esteja operando corretamente antes da próxima fase.

Fase de Descoberta

Terminados os testes controlados, realizaremos um treinamento mais amplo do Módulo Neural e aplicaremos o pipeline completo em dados reais preparados na etapa (i). Nesse processo, esperamos realizar descobertas de novos biomarcadores de longevidade. Os experimentos desta fase de descoberta incluem experimentos com dados genéticos dos grupos experimental e controle. Utilizaremos, ainda, dados de domínio público, como os do estudo de Timmers et al. [28] e os em LongevityMap [81], [82], [85], por exemplo. A ideia é que nosso modelo identifique variantes que correlacionam com longevidade.

O resultado esperado dessa fase de descoberta é a obtenção de uma lista priorizada de marcadores candidatos (genes/variantes/assinaturas epigenéticas) associadas à longevidade, com suporte preditivo do Módulo Neural e sinalização de relevância biológica por parte do Módulo Simbólico.

Validação Estatística e Biológica dos Achados

Identificados os marcadores candidatos, empregaremos testes estatísticos para avaliar a significância das associações encontradas. Por exemplo, utilizaremos testes de hipótese para verificar se a frequência de uma variante candidata difere de forma significativa entre grupos longevos e controle. Adicionalmente, avaliaremos a capacidade preditiva global do conjunto de marcadores via métricas como AUC-ROC em dados *hold-out* para quantificar quão bem nosso modelo distingue longevos.

Em paralelo às análises estatísticas, buscaremos a validação qualitativa junto a pesquisadores da área biológica. Para isso, organizaremos seminários e reuniões de projeto com nossos parceiros pesquisadores das áreas de fisiologia e biotecnologia para discutir os achados.

6 - Cronograma de Execução

No desenvolvimento deste projeto, seguiremos o cronograma apresentado na Tabela 2 e Tabela 3. Na Tabela 2, são listadas as atividades principais; enquanto que, na Tabela 3, estas atividades são apresentadas distribuídas ao longo da duração do projeto. Este cronograma está de acordo com a Seção 4 - Metas e Indicadores da Proposta.

Tabela 2: Lista de Atividades Principais

Atividade	Descrição	Bimestre Início	Bimestre Término
1	Protótipo de Pipeline de Análise Genômica Baseado em IA Neuro-Simbólica	1	12
2	Identificação de Marcadores Genéticos ou Epigenéticos de Longevidade	7	12
3	Validação Cruzada com Conhecimento Biológico	10	12
4	Disseminação Científica (Elaboração de artigos)	5, 11	6, 12
5	Formação de Recursos Humanos	1	12

Tabela 3: Cronograma de Execução das Atividades Principais

Atividade	Bimestre											
	1	2	3	4	5	6	7	8	9	10	11	12
1												
2												
3												
4												
5												

7 - Resultados Esperados

Esperamos atingir resultados práticos quantitativos e qualitativos com este projeto. Esperamos os seguintes resultados quantitativos:

- P1. Pipeline Neuro-Simbólico: Código-fonte aberto implementando o pipeline completo, com scripts de treinamento e inferência (uma versão bem preliminar já se encontra em desenvolvimento: <https://github.com/albertodesouza/genomics>). *Entrega até o fim do projeto (o projeto terá duração de 2 anos).*
- P2. Conjunto de Marcadores de Longevidade Identificados: Lista dos principais genes/variantes/assinaturas epigenéticas associados à longevidade descobertos no projeto, acompanhada de documentação descrevendo cada item. *Entrega até o fim do projeto.*
- P3. Artigos Científicos: Pelo menos dois artigos completos: (i) um artigo em conferência internacional enfocando a metodologia neuro-simbólica e resultados preliminares; (ii) um artigo de revista com resultados associados às descobertas de marcadores. *Entrega de (i) até o fim do primeiro ano, e de (ii) até o fim do projeto.*
- P4. Artigos de Divulgação Científica: Dois artigos de divulgação para público leigo. *Entrega até o fim do projeto.*
- P5. Formação de Recursos Humanos: Orientação de um aluno de iniciação científica, um aluno de mestrado e um aluno de doutorado. *Até o fim do projeto.*

O principal resultado qualitativo deste projeto será estender o estado da arte na área de IA neuro-simbólica aplicada à genômica, o que impactará positivamente na qualidade dos Programas de Pós-Graduação nas áreas de Informática, Fisiologia e Biotecnologia da UFES, e na qualidade dos cursos de graduação em Engenharia de Computação, Ciência da Computação, Medicina e Biologia da UFES. Além disso, o projeto possibilitará a ampliação da interlocução e colaboração acadêmica de caráter consultivo com pesquisadores vinculados à *City St George's, University of London* (a *City, University of London* e a *St George's, University of London* se juntaram em 1o. de agosto de 2024) e à *University of Twente*, por meio de discussões técnico-científicas, intercâmbio de conhecimento, palestras, seminários e reuniões científicas.

A *City St George's, University of London* é uma universidade global, com mais de 27.000 estudantes de mais de 170 países. Em avaliações nacionais (Reino Unido), *City St George's* ostenta classificação “gold”, tendo alcançado o 14º lugar no Reino Unido para “graduate prospects”. O professor Artur d'Avila Garcez, pesquisador colaborador acadêmico de caráter consultivo neste projeto e vinculado à instituição, é atualmente Diretor do *Research Centre for Machine Learning* da *City St George's*, sendo figura central do grupo de IA da nova instituição resultante da fusão da *City, University of London* com a *St George's, University of London*. Sua trajetória inclui contribuições fundamentais na área de IA neuro-simbólica. Garcez é presidente do comitê diretor da *Neural-Symbolic Learning and Reasoning Association* e é coautor de livros fundamentais da área, como *Neural-Symbolic Learning Systems* (2002) e *Neural-Symbolic Cognitive Reasoning* (2009).

A *University of Twente* é uma das principais universidades técnicas holandesas, com um histórico de grandes contribuições para a computação – a universidade aparece na 5ª posição em computação frente demais as universidades holandesas (126-150 no mundo em computação) e em primeiro lugar no mundo em “Industry, Innovation and Infrastructure” (<https://www.timeshighereducation.com/world-university-rankings/university-twente>). Em particular, o *Department of Semantics, Cybersecurity & Services* (SCS) é reconhecido como um dos principais departamentos internacionais na área de ontologias. Atualmente, esse departamento (dirigido pelo professor Giancarlo Guizzardi, pesquisador colaborador acadêmico de caráter consultivo neste projeto e vinculado à instituição) está em processo de formação de um novo *cluster* com um departamento de Ciência de Dados para formar um centro de competência na área de Engenharia de Dados e IA Neuro-simbólica. Este projeto permitirá o intercâmbio de alunos e professores de ambas as instituições. A parceria entre o PPGI/UFES e a *University of Twente* tem longo histórico, incluindo a titulação dupla de alunos de doutorado.

Outro importante resultado qualitativo deste projeto será a disponibilização local, nacional e internacional do código, tecnologias e sistemas eventualmente desenvolvidos – sempre que for apropriado e vantajoso para o cumprimento dos objetivos do projeto, serão empregadas linguagens, ferramentas, bibliotecas e ambientes livres de desenvolvimento de software (software livre). Além destas contribuições, este projeto viabilizará a formação de especialistas em *Deep Learning*, IA generativa e neuro-simbólica nos níveis de doutorado, mestrado e graduação. Por fim, buscaremos fazer com que os resultados técnico-científicos provenientes da execução deste projeto levem à criação de novas *startups* (projetos anteriores com o porte deste levaram membros da equipe a, juntamente com estudantes, criar as *startups* AUMO S.A., www.aumo.ai, Lume Robotics S.A., www.lumerobotics.ai, e Motora Technologies S.A., www.motora.ai) e ao fortalecimento do ecossistema brasileiro, britânico e holandês de inovação. Esperamos, ainda, que os resultados do projeto produzam riqueza, renda e a geração de novos empregos de qualidade no Brasil, no Reino Unido e na Holanda.

8 - Recursos Financeiros de Outras Fontes

O pesquisador proponente é Professor Emérito da Universidade Federal do Espírito Santo (UFES) e atua de forma voluntária no Programa de Pós-Graduação em Informática (membro permanente), conforme a Resolução nº 26/99 do Conselho Universitário da UFES. Assim, não há custos de pessoal financiados por instituição externa no âmbito deste projeto.

Os recursos de outras fontes envolvidos na execução do estudo são classificados como indiretos e de pequena monta, sendo cobertos da seguinte forma:

- Infraestrutura computacional (LCAD / supercomputador Atena): utilização de recursos institucionais já disponíveis para atividades de pesquisa, sem custo adicional específico para o projeto;
- Equipamentos de informática e software: utilização de equipamentos e ferramentas já disponíveis ao pesquisador e equipe;
- Material de escritório e apoio administrativo: custos mínimos, cobertos com recursos próprios do pesquisador;
- Custos operacionais gerais (energia, rede, armazenamento): absorvidos pela infraestrutura institucional da UFES, no âmbito de suas atividades regulares de ensino, pesquisa e extensão.

Dessa forma, não há previsão de despesas significativas adicionais para a execução do projeto.

9 - Infraestrutura e Apoio Técnico

O Laboratório de Computação de Alto Desempenho (LCAD – www.lcad.inf.ufes.br) da UFES, fundado em 2002 e atualmente membro do Instituto de Inteligência Computacional Aplicada (I2CA – www.i2ca.ai), será a base operacional do projeto no Brasil. Atualmente, o laboratório dispõe de cinco salas: LCAD4 (prédio

CT-13, Sala 5, com 122 m²), LCAD3 (prédio da antiga Gráfica da UFES, com 99 m²), LCAD2 (prédio CT-9, com 58 m²), LCAD5 (prédio CT-13, Sala 6, com 49 m²), e a Sala de Servidores (prédio CT-13, Sala 21, com 22 m²).

O LCAD4 possui 14 estações de trabalho equipadas com placas gráficas de alto desempenho, incluindo NVIDIA GeForce RTX (4060, 3080, 3060), NVIDIA Titan (X, Xp, V), e Tesla K40, e 14 notebooks especializados para pesquisa em IA e visão computacional. Conta ainda com 2 robôs móveis, incluindo 1 robô quadrúpede Unitree Go2, 1 robô móvel Ackerman, além de 2 braços robóticos OpenMANIPULATOR-X, utilizados em pesquisas que envolvem manipulação e interação. Recentemente foi adquirido um robô humanoide (Unitree G1 EDU - *Humanoid agent AI avatar*). O LCAD3 é especializado em robótica de grande porte (caminhões e empilhadeiras), possuindo pé direito alto e bancadas especializadas para trabalho em mecânica. Destacam-se entre os equipamentos avançados do LCAD2, 2 carros autônomos (denominados SARA e IARA), baseados em veículos comerciais e equipados com sensores avançados como LiDAR Velodyne HDL-32E, câmeras e IMUs. O LCAD5 é voltado para o desenvolvimento de software na área de IA, robótica móvel assistiva e realidade virtual e aumentada, possuindo 1 robô móvel com câmera e sensores, e 2 robôs de assistência. A Sala de Servidores conta com 1 supercomputador (Atena), equipado com 3 placas gráficas NVIDIA A100 de 80GB, 512 GB de RAM e dezenas de terabytes de armazenamento para processamento intensivo de aplicações de IA e aprendizado profundo, e parque computacional com 17 servidores para processamento e armazenamento de dados científicos. O LCAD conta também com diversos dispositivos para pesquisa adicionais, como câmeras estereoscópicas, óculos de rastreamento ocular (SMI *Eye Tracking Glasses*), sistemas de captura de movimento, impressora 3D e óculos de realidade mista (Meta Quest 2 e Meta Quest 3).

Em termos de pessoal de apoio, o LCAD conta com dois analistas de TI concursados, providos pelo Departamento de Informática da UFES, e um bolsista de apoio técnico (I2CA). A UFES oferece ainda suporte de sua Superintendência de Tecnologia da Informação para questões de rede e segurança computacional. Essa infraestrutura e suporte técnico já existentes viabilizam a execução do projeto no que concerne às suas necessidades de infraestrutura computacionais básicas.

O Centro de Investigação ELSA (Estudo Longitudinal de Saúde do Adulto) do Espírito Santo é uma das seis unidades nacionais que integram o ELSA-Brasil (<https://elsabrasil.org/sobre>). O ELSA-Brasil é uma investigação multicêntrica de coorte composta por 15 mil funcionários de seis instituições públicas de ensino superior e pesquisa das regiões Nordeste, Sul e Sudeste do Brasil. O Centro ELSA-ES é vinculado ao Departamento de Ciências Fisiológicas e ao Programa de Pós-Graduação em Ciências Fisiológicas (PPGCF) da UFES e é coordenado pelo Prof. José Geraldo Mill, pesquisador deste projeto. Sua infraestrutura é dedicada à condução de estudos epidemiológicos de larga escala voltados à investigação dos determinantes biológicos, comportamentais e sociais das doenças crônicas em adultos.

O Centro dispõe de espaço físico próprio no Centro de Ciências da Saúde da UFES, com laboratórios de coleta e processamento biológico, salas de entrevista e exame físico, consultórios, setor de antropometria, laboratório de bioquímica e salas de apoio administrativo e de informática. A estrutura inclui equipamentos para exames cardiovasculares (ecocardiografia, eletrocardiografia, tonometria de pulso, doppler de carótidas), aferição de pressão arterial automatizada, bioimpedância, espirômetro, analisadores bioquímicos, refrigeradores e ultrafreezers para armazenamento de amostras biológicas, além de sistemas de gerenciamento de dados clínicos e epidemiológicos.

O ELSA-ES abriga o biobanco estadual de amostras biológicas do ELSA-Brasil, contendo DNA, sangue, plasma, soro e urina de milhares de participantes. O centro conta ainda com equipe multidisciplinar formada por médicos, fisiologistas, enfermeiros, técnicos e analistas de dados, que dão suporte a projetos de pesquisa locais e nacionais. A infraestrutura instalada, aliada à expertise consolidada do grupo em fisiologia, epidemiologia e saúde pública, proporciona suporte robusto às atividades do presente projeto, em especial por propiciar acesso a dados de DNA local, e a desejada futura validação clínica e fisiológica dos biomarcadores que almejamos identificar por meio da IA neuro-simbólica que será desenvolvida.

O Núcleo de Genética Humana e Molecular (NGHM – <https://nghm.ufes.br>) da UFES agrega um grupo de pesquisadores e um conjunto de laboratórios multiusuário dedicados à pesquisa em Genética Humana e Molecular. O NGHM é derivado do antigo SAG (Serviço de Aconselhamento Genético da UFES) e, nos seus

quase 30 anos de existência, possibilitou a publicação de inúmeros artigos científicos, além de projetos de extensão, dissertações de mestrado e doutorado, e monografias de graduação.

O NGHM possui 5 laboratórios de pesquisa destinados a protocolos de biologia molecular (pré-PCR e pós-PCR), cultura celular e esterilização de materiais, além de 5 escritórios de apoio para professores e alunos. Dispõe de equipamentos que incluem: sequenciador de DNA Illumina MiSeq, 3500 Series Genetic Analyzer, QuantStudio 3 e 5, fotômetro de microplacas, fluorímetro, nanoespectrofotômetro, cubas/fontes para eletroforese de ácidos nucleicos, centrífugas, agitadores, ultrafreezer, freezers, geladeiras, balanças, pHmetro, banhos-maria, estufas, fotodocumentador, termocicladores, incubadoras CO₂, microscópio invertido, fluxos laminares, shaker orbital, cabine de exaustão de gases, autoclaves e nobreaks.

10 - Perspectivas de Colaboração Internacional durante a Execução do Projeto

Durante a execução do projeto, pesquisadores vinculados à *City St George's, University of London*, e à *University of Twente*, incluindo os Profs. Artur Garcez e Giancarlo Guizzardi, atuarão como colaboradores acadêmicos de caráter consultivo, por meio de discussões técnico-científicas, intercâmbio de conhecimento, palestras, seminários e reuniões científicas. Essa interlocução permitirá à equipe brasileira trocar experiências sobre aprendizado profundo, IA neuro-simbólica, ontologias e temas correlatos, contribuindo para a formação de pesquisadores e para o fortalecimento dos programas de pós-graduação da UFES envolvidos no projeto (Informática, Fisiologia e Biotecnologia). Vale destacar que não haverá compartilhamento ou acesso aos dados genômicos utilizados no projeto, execução de etapas do protocolo pelas instituições estrangeiras, desenvolvimento de módulos sob responsabilidade dessas instituições, decisão metodológica ou responsabilidade institucional estrangeira.

11 - Parcerias já Estabelecidas

Este projeto baseia-se em relações de colaboração acadêmica já existentes entre o proponente e os professores Artur d'Ávila Garcez (*City St George's, University of London*) e Giancarlo Guizzardi (*University of Twente*).

A cooperação científica com o professor Artur Garcez remonta a 1999, quando o proponente deste projeto elaborou o artigo científico abaixo em parceria com Garcez:

- GARCEZ, Artur d'Ávila ; BRODA, K. ; GABAY, D. ; De Souza, Alberto Ferreira . Knowledge Extraction from Trained Neural Networks: A Position Paper. In: 6th International Conference on Neural Information Processing, 1999, Perth - Australia. Proceedings of the 6th International Conference on Neural Information Processing. Los Alamitos, CA - USA: IEEE Computer Society, 1999. p. 685-691.

Mais tarde, o proponente e Garcez publicaram o artigo:

- CARNEIRO, Raphael Vivacqua ; DIAS, Stiven Schwanz ; FARDIM JÚNIOR, Dijalma ; OLIVEIRA, Hallysson ; GARCEZ, Artur d'Ávila ; De Souza, Alberto F. . Improving VG-RAM Neural Networks Performance Using Knowledge Correlation. In: 13th International Conference on Neural Information Processing - ICONIP 2006, 2006, Hong Kong, China. Lecture Notes in Computer Science (ICONIP 2006). Berlin: Springer, 2006. v. 4232. p. 427-436.

Em 2008, o proponente realizou uma visita técnico-científica de 30 dias ao *Department of Computing da City, University London*, integrando-se ao grupo *Autonomous Intelligent Systems*. Essa visita ocorreu no âmbito do projeto CNPq FACADIC (Fortalecimento das Áreas de Computação de Alto Desempenho e Inteligência Computacional do PPGI/UFES), e foi dedicada à discussão de propostas de pesquisa conjuntas

com o Prof. Artur Garcez. Desde então, diversas iniciativas foram conduzidas em parceria, tais como o doutorado sanduíche de um aluno de doutorado do proponente, agora Dr. Avelino Forechi, que culminou em publicações conjuntas:

- Forechi, Avelino ; De Souza, Alberto F. ; D'AVILA GARCEZ, ARTUR . Self-driving cars could be the answer to congested roads. *The Conversation*, 21 de novembro 2014.
- Forechi, Avelino ; De Souza, Alberto F. ; Oliveira Neto, Jorcy de ; AGUIAR, EDILSON DE ; BADUE, Claudine ; D'AVILA GARCEZ, ARTUR ; Oliveira-Santos, Thiago . Fat-Fast VG-RAM WNN: A high performance approach. *Neurocomputing (Amsterdam)*, v. 183, p. 56-69, 2016.

Paralelamente, a cooperação com o professor Giancarlo Guizzardi, da *University of Twente*, vem sendo fortalecida pela execução do projeto CNPq ICAGPT (2024–2026: Inteligências Computacionais Autônomas Baseadas em GPT), coordenado pelo proponente. Esse projeto viabilizou a intensificação das atividades conjuntas com o Prof. Giancarlo Guizzardi, incluindo visitas do professor Guizzardi à UFES, e resultou em alinhamento estratégico de pesquisa para este projeto. A parceria UFES–Twente já conta com histórico de titulação conjunta de doutores do PPGI/UFES–Twente.

Em síntese, a presente proposta se ancora em uma rede de cooperações ativa e madura, com experiências prévias de sucesso em publicações, orientações, mobilidade, e projetos de pesquisa financiados, criando um terreno propício para a execução e impacto científico internacional deste projeto.

12 - Relevância da Cooperação Internacional e Natureza da Participação dos Colaboradores

A relevância da interlocução acadêmica internacional com pesquisadores vinculados à *City St George's, University of London*, e à *University of Twente* está fundamentada nos seguintes aspectos:

1. Complementaridade de Expertise

A UFES aporta experiência em aprendizado profundo, fisiologia humana, análise de dados genômicos e computação de alto desempenho. O Prof. Artur Garcez, vinculado à *City St George's*, possui expertise reconhecida internacionalmente em IA neuro-simbólica, integração entre redes neurais e raciocínio lógico, além de vasta rede de contatos globais. O Prof. Giancarlo Guizzardi, vinculado à *University of Twente*, possui reconhecida experiência em ontologias fundamentais e engenharia conceitual, incluindo o desenvolvimento pelo Prof. Guizzardi e outros da *Unified Foundational Ontology (UFO)*.

2. Formação e Mobilidade Internacional

A interlocução acadêmica poderá contribuir para a formação dos estudantes brasileiros por meio de seminários, palestras, workshops, discussões técnico-científicas e eventuais visitas acadêmicas às instituições às quais estão vinculados os pesquisadores colaboradores. Essas atividades terão caráter de intercâmbio científico e acadêmico, sem implicar coorientação formal no âmbito deste projeto, execução de etapas do projeto pelas instituições estrangeiras ou transferência dos dados genômicos utilizados na pesquisa. A experiência anterior de cooperação acadêmica entre pesquisadores da UFES e da *University of Twente*, inclusive em iniciativas de titulação conjunta, constitui antecedente relevante para essa interlocução científica.

A execução e a responsabilidade científica e institucional pelo presente projeto permanecerão na UFES. A participação dos pesquisadores estrangeiros terá natureza acadêmica e consultiva, conforme descrito a seguir:

- UFES (coordenação e execução): liderança do projeto, preparação e análise de dados genômicos, desenvolvimento e treinamento de modelos de *deep learning*, execução de experimentos em larga

escala, interpretação biológica, além da gestão administrativa e divulgação de resultados.

- Prof. Artur Garcez, vinculado à *City St George's, University of London*: colaboração acadêmica de caráter consultivo por meio de discussões técnico-científicas e intercâmbio de conhecimento em temas relacionados à IA neuro-simbólica, além de eventual participação em palestras, seminários e reuniões científicas.
- Prof. Giancarlo Guizzardi, vinculado à *University of Twente*: colaboração acadêmica de caráter consultivo por meio de discussões técnico-científicas e intercâmbio de conhecimento em temas relacionados a ontologias, engenharia conceitual e IA neuro-simbólica, além de eventual participação em palestras, seminários e reuniões científicas.

As decisões científicas e metodológicas relativas à condução do projeto serão tomadas pela equipe da UFES, sob coordenação do pesquisador responsável. Os colaboradores estrangeiros poderão contribuir com comentários, discussões e aconselhamento científico, sem poder de decisão sobre o desenho metodológico, os critérios de análise ou a condução do estudo. O resultado é um projeto interdisciplinar de alto impacto, que posiciona o Brasil no contexto mundial de IA neuro-simbólica e genômica.

15 - Plano de Divulgação Científica

A divulgação científica é parte integrante deste projeto e será orientada para a sociedade em geral, incluindo públicos não especializados. O plano proposto busca maximizar a divulgação pública dos avanços da pesquisa em inteligência artificial aplicada à longevidade, além de inspirar jovens para carreiras científicas.

a) Descrição das ações e cronograma

As ações previstas serão distribuídas ao longo dos 24 meses do projeto:

- **Produção de artigos de divulgação:** Até o mês 24, elaboração de pelo menos dois textos em português, em linguagem acessível, a serem publicados em portais de ciência, como a Revista Universidade (<https://blog.ufes.br/revistauniversidade/>) ou a Revista FAPESP (<https://revistapesquisa.fapesp.br/>), e pelo menos um em inglês na The Conversation - UK (<https://theconversation.com/uk>) ou veículo equivalente.
- **Palestras públicas e seminários:** Ao longo do projeto, o coordenador participará de seminários de extensão em espaços comunitários, escolas, faculdades ou universidades, realizando pelo menos duas palestras voltadas ao público leigo.
- **Conteúdo digital:** Entre os meses 6 e 24, divulgação de vídeos curtos, posts explicativos (LinkedIn, X/Twitter, Instagram e YouTube) e podcasts sobre marcos da pesquisa (como este já realizado - https://youtu.be/F_RI44V6-YA?si=3rCgkcYBI6DVR_km).
- **Material educativo:** Até o mês 18, produção de um infográfico bilíngue (“Como a IA pode ajudar você a viver mais e melhor”) em formato digital a ser distribuído pelas redes sociais do LCAD (<https://www.instagram.com/lcad.ufes/>, <https://www.youtube.com/user/lcadufes>, etc.) e I2CA (https://www.instagram.com/instituto_i2ca/).
- **Participação em feiras e eventos de ciência:** Participação anual em feiras regionais de divulgação científica, como as que ocorrem na Semana Nacional da Ciência e Tecnologia (SNCT), do Ministério da Ciência, Tecnologia e Inovação (MCTI).

b) Público-alvo

O público-alvo é composto por:

- Estudantes de ensino médio, graduação e pós-graduação, potenciais futuros pesquisadores;
- Público geral interessado em IA, saúde, longevidade e tecnologia;

- Jornalistas e comunicadores científicos;
- Comunidade acadêmica ampliada (via divulgação bilíngue).

c) Meios a serem empregados

- Portais de notícias e revistas de divulgação científica, como a Revista Universidade (<https://blog.ufes.br/revistauniversidade/>), a Revista FAPESP (<https://revistapesquisa.fapesp.br/>) e o Jornal da Ciência da SBPC (<https://portal.sbpcnet.org.br/jornal-da-ciencia/>);
- Redes sociais populares e acadêmicas (Instagram, YouTube, LinkedIn, X/Twitter, etc.);
- Materiais gráficos (infográficos) e vídeos;
- Feiras, exposições e eventos de extensão;
- Palestras presenciais e transmissões ao vivo em plataformas digitais como o Youtube.

d) Responsáveis pela execução

- **Coordenação:** Prof. Alberto Ferreira de Souza (UFES).
- **Equipe LCAD/I2CA:** apoio na produção de materiais digitais, manutenção de redes sociais e organização de palestras locais.
- **Pesquisadores colaboradores vinculados à City St George's, University of London, e à University of Twente:** eventual participação, em caráter acadêmico e consultivo, em discussões, palestras, seminários ou outras atividades de divulgação científica relacionadas aos temas do projeto.
- **Alunos bolsistas:** participação direta na produção de conteúdo e interação com o público em eventos.

e) Resultados esperados

- Ampliação da compreensão pública sobre o impacto da IA na saúde e ampliação da longevidade saudável;
- Maior inserção da UFES no cenário de divulgação científica nacional e internacional, favorecida também pela interlocução acadêmica com pesquisadores estrangeiros;
- Engajamento de jovens em carreiras científicas;
- Disponibilização pública de materiais bilíngues de alta qualidade (artigos, infográficos, vídeos);
- Fortalecimento da percepção social de que ciência e tecnologia contribuem para a qualidade de vida e o envelhecimento saudável.

O planejamento acima visa tornar acessíveis os resultados científicos alcançados com o projeto, inspirando jovens para carreiras científicas e informando a população em geral sobre avanços na área de IA aplicada à longevidade saudável. O impacto esperado é a ampliação da compreensão pública de que envelhecimento saudável é um objetivo científico relevante, e que a pesquisa brasileira, enriquecida por interlocuções acadêmicas internacionais, pode contribuir de forma relevante para esse esforço científico global.

16 - Informações sobre os Membros da Equipe

E-mail: alberto@lcad.inf.ufes.br

Nome completo: Alberto Ferreira de Souza

País de nascimento: Brasil

Função: Coordenador

Instituição: Universidade Federal do Espírito Santo

Departamento: Departamento de Informática

Titulação: Doutorado

Dedicação ao Projeto: no Brasil, 20 horas por semana, no exterior, 40 horas por semana

Responsabilidade no projeto (Português):

As responsabilidades do Prof. Alberto compreendem coordenar todas as atividades administrativas, de cooperação científica, divulgação científica, pesquisa e desenvolvimento do projeto, além de conduzir pessoalmente parte significativa das atividades de pesquisa e desenvolvimento do pipeline de IA neuro-simbólica para a análise de dados genômicos a fim de identificar biomarcadores de longevidade de forma precisa e interpretável. Para isso, vai desenvolver pesquisa na direção de (i) criar modelos computacionais capazes de prever a predisposição à longevidade a partir de dados genômicos brutos, e de (ii) explicar biologicamente essas previsões, vinculando-as a processos biológicos conhecidos relacionados à longevidade saudável.

Responsabilidade no projeto (Inglês):

The responsibilities of Prof. Alberto include coordinating all administrative, scientific cooperation, scientific dissemination, research, and development activities of the project, as well as personally conducting a significant part of the research and development of the neuro-symbolic AI pipeline for genomic data analysis in order to identify longevity biomarkers in an accurate and interpretable way. To this end, he will conduct research aimed at (i) creating computational models capable of predicting predisposition to longevity from raw genomic data, and (ii) biologically explaining these predictions by linking them to known biological processes of health longevity.

E-mail: a.garcez@citystgeorges.ac.uk

Nome completo: Artur d'Ávila Garcez

País de nascimento: Brasil

Função: Colaborador acadêmico de caráter consultivo

Instituição: City St George's, University of London

Departamento: Computer Science Department

Titulação: Doutorado

Dedicação ao Projeto: 2 horas por semana

Responsabilidade no projeto (Português):

O Prof. Garcez atuará como colaborador acadêmico de caráter consultivo, contribuindo por meio de discussões técnico-científicas, intercâmbio de conhecimento, palestras, seminários e reuniões científicas em temas relacionados à IA neuro-simbólica e à integração entre aprendizado profundo e raciocínio simbólico. Sua participação não envolverá acesso aos dados genômicos utilizados no projeto, execução de etapas de processamento ou análise, responsabilidade pelo desenvolvimento de módulos do pipeline, decisão metodológica ou responsabilidade institucional pela condução do estudo.

Responsabilidade no projeto (Inglês):

Prof. Garcez will act as an academic collaborator in a consultative capacity, contributing through technical-scientific discussions, knowledge exchange, lectures, seminars, and scientific meetings on topics related to neuro-symbolic AI and the integration of deep learning and symbolic reasoning. His participation will not involve access to the genomic data used in the project, execution of data processing or analysis stages, responsibility for the development of pipeline modules, methodological decision-making, or institutional responsibility for the conduct of the study.

E-mail: g.guizzardi@utwente.nl

Nome completo: Giancarlo Guizzardi

País de nascimento: Brasil

Função: Colaborador acadêmico de caráter consultivo

Instituição: University of Twente

Departamento: Department of Semantics, Cybersecurity & Services Group

Titulação: Doutorado

Dedicação ao Projeto: 5 horas por semana

Responsabilidade no projeto (Português):

O Prof. Giancarlo Guizzardi atuará como colaborador acadêmico de caráter consultivo, contribuindo por meio de discussões técnico-científicas, intercâmbio de conhecimento, palestras, seminários e reuniões

científicas em temas relacionados a ontologias, engenharia conceitual, grafos de conhecimento e IA neuro-simbólica. Sua experiência com a *Unified Foundational Ontology* (UFO) poderá contribuir para discussões científicas sobre representação e interpretabilidade. Sua participação não envolverá acesso aos dados genômicos utilizados no projeto, execução de etapas de processamento ou análise, responsabilidade pelo desenvolvimento do módulo simbólico, ontologias, grafos de conhecimento ou outros componentes do pipeline, coorientação formal no âmbito deste projeto, decisão metodológica ou responsabilidade institucional pela condução do estudo.

Responsabilidade no projeto (Inglês):

Prof. Giancarlo Guizzardi will act as an academic collaborator in a consultative capacity, contributing through technical-scientific discussions, knowledge exchange, lectures, seminars, and scientific meetings on topics related to ontologies, conceptual modeling, knowledge graphs, and neuro-symbolic AI. His experience with the Unified Foundational Ontology (UFO) may contribute to scientific discussions concerning representation and interpretability. His participation will not involve access to the genomic data used in the project, execution of data processing or analysis stages, responsibility for the development of the symbolic module, ontologies, knowledge graphs, or other pipeline components, formal co-supervision within this project, methodological decision-making, or institutional responsibility for the conduct of the study.

E-mail: josegmill@gmail.com

Nome completo: José Geraldo Mill

País de nascimento: Brasil

Função: Pesquisador

Instituição: Universidade Federal do Espírito Santo

Departamento: Departamento de Ciências Fisiológicas

Titulação: Doutorado

Dedicação ao Projeto: 2 horas por semana

Responsabilidade no projeto (Português):

As responsabilidades do Prof. José Geraldo Mill compreendem avaliar os aspectos médicos e fisiológicos associados aos marcadores genéticos e epigenéticos identificados pelo pipeline de IA neuro-simbólica. Ele atuará na interpretação clínica dos achados, especialmente no que tange à sua correlação com doenças crônicas, cardiovasculares e metabólicas, contribuindo para a validação dos resultados do projeto. Sua experiência como coordenador do ELSA-Brasil no Espírito Santo e sua reconhecida liderança em estudos longitudinais de saúde permitirão contextualizar os biomarcadores descobertos em termos de relevância clínica e impacto sobre a saúde populacional. Além disso, apoiará a interpretação dos resultados do projeto à luz do conhecimento fisiológico, clínico e epidemiológico disponível, fortalecendo as chances de transferência dos achados para a prática médica e de saúde pública.

Responsabilidade no projeto (Inglês):

The responsibilities of Prof. José Geraldo Mill include evaluating the medical and physiological aspects associated with the genetic and epigenetic markers identified by the neuro-symbolic AI pipeline. He will contribute to the clinical interpretation of the findings, particularly regarding their correlation with chronic, cardiovascular, and metabolic diseases, thereby supporting the validation of the project's results. His experience as the coordinator of ELSA-Brasil in Espírito Santo and his recognized leadership in longitudinal health studies will enable the discovered biomarkers to be contextualized in terms of clinical relevance and population health impact. Furthermore, he will support the interpretation of the project's results in light of available physiological, clinical, and epidemiological knowledge, strengthening the potential for translating the findings into medical and public health practice.

E-mail: debora.meira@ufes.br

Nome completo: Débora Dummer Meira

País de nascimento: Brasil

Função: Pesquisadora

Instituição: Universidade Federal do Espírito Santo

Departamento: Departamento de Ciências Biológicas

Titulação: Doutorado

Dedicação ao Projeto: 5 horas por semana

Responsabilidade no projeto (Português):

As responsabilidades da Profa. Débora compreendem fornecer suporte científico aos aspectos biológicos do projeto, especialmente na análise e interpretação de dados genéticos e epigenéticos. Ela atuará na validação biológica dos resultados gerados pela plataforma de IA, garantindo rigor científico na integração dos dados biológicos ao pipeline de IA. Além disso, contribuirá para a discussão dos achados à luz da biologia do envelhecimento.

Responsabilidade no projeto (Inglês):

The responsibilities of Prof. Débora includes providing scientific support to the biological aspects of the project, particularly in the analysis and interpretation of genomic and epigenetic data. She will be responsible for establishing biological validation of the results produced by the AI platform, ensuring scientific rigor in the integration of biological data into the IA pipeline. Additionally, she will contribute to the interpretation of findings in the context of aging biology.

E-mail: todsantos@inf.ufes.br

Nome completo: Thiago Oliveira dos Santos

País de nascimento: Brasil

Função: Pesquisador

Instituição: Universidade Federal do Espírito Santo

Departamento: Departamento de Informática

Titulação: Doutorado

Dedicação ao Projeto: 1 hora por semana

Responsabilidade no projeto (Português):

As responsabilidades do Prof. Thiago compreendem a colaboração na investigação científica de modelos de *deep learning* voltados à análise de dados genômicos. Ele atuará na proposição de algoritmos e técnicas de ajuste e experimentação de arquiteturas neurais avançadas, incluindo *transformers* genômicos, com foco na predição de biomarcadores de longevidade. Também contribuirá para as discussões de integração desses modelos ao pipeline neuro-simbólico, podendo participar da orientação ou coorientação de alunos envolvidos no projeto.

Responsabilidade no projeto (Inglês):

Professor Thiago's responsibilities include collaborating on the scientific investigation of deep learning models aimed at genomic data analysis. He will work on proposing algorithms and techniques for tuning and experimenting with advanced neural architectures, including genomic transformers, with a focus on the prediction of longevity biomarkers. He will also contribute to discussions on integrating these models into the neuro-symbolic pipeline, which may include advising or co-advising students involved in the project.

Referências Bibliográficas

- [1] T. Niccoli and L. Partridge, "Ageing As a Risk Factor for Disease," *Curr. Biol.*, vol. 22, pp. R741–R752, 2012, doi: 10.1016/j.cub.2012.07.024.
- [2] Y. Hou *et al.*, "Ageing As the Most Important Risk Factor for Neurodegenerative Disease," *Nat. Rev. Neurol.*, vol. 15, pp. 565–581, 2019, doi: 10.1038/s41582-019-0244-7.
- [3] B. J. North and D. A. Sinclair, "The Intersection Between Aging and Cardiovascular Disease," *Circ. Res.*, vol. 110, pp. 1097–1108, 2012, doi: 10.1161/CIRCRESAHA.111.246876.
- [4] C. López-Otín, M. A. Blasco, L. Partridge, M. Serrano, and G. Kroemer, "The Hallmarks of Aging," *Cell*, vol. 153, no. 6, pp. 1194–1217, 2013, doi: 10.1016/j.cell.2013.05.039.
- [5] C. López-Otín and others, "Hallmarks of Aging: An Expanding Universe," *Cell*, vol. 186, pp. 243–278, 2023, doi: 10.1016/j.cell.2022.12.038.
- [6] B. Alberts *et al.*, *Molecular Biology of the Cell: Seventh Edition*. W. W. Norton, 2022.

- [7] S. Horvath, "DNA methylation age of human tissues and cell types," *Genome Biol.*, vol. 14, no. 10, p. R115, 2013, doi: 10.1186/gb-2013-14-10-r115.
- [8] D. E. Harrison, R. Strong, Z. Sharp, and others, "Rapamycin fed late in life extends lifespan in genetically heterogeneous mice," *Nature*, vol. 460, no. 7253, pp. 392–395, 2009, doi: 10.1038/nature08221.
- [9] D. J. Baker *et al.*, "Naturally Occurring p16Ink4a-positive Cells Shorten Healthy Lifespan," *Nature*, vol. 530, pp. 184–189, 2016, doi: 10.1038/nature16932.
- [10] K. Takahashi and S. Yamanaka, "Induction of Pluripotent Stem Cells from Mouse Embryonic and Adult Fibroblast Cultures by Defined Factors," *Cell*, vol. 126, no. 4, pp. 663–676, 2006, doi: 10.1016/j.cell.2006.07.024.
- [11] A. Ocampo *et al.*, "In Vivo Amelioration of Age-Associated Hallmarks by Partial Reprogramming," *Cell*, vol. 167, no. 7, pp. 1719–1733.e12, 2016, doi: 10.1016/j.cell.2016.11.052.
- [12] J. Yang, K. Mills, T. A. Rando, D. A. Sinclair, and others, "Loss of Epigenetic Information as a Cause of Mammalian Aging," *Cell*, vol. 186, no. 2, pp. 305–326, 2023, doi: 10.1016/j.cell.2023.01.055.
- [13] B. Bernardes de Jesus *et al.*, "Telomerase gene therapy in adult and old mice delays aging and increases longevity without increasing cancer," *EMBO Mol. Med.*, vol. 4, no. 8, pp. 691–704, 2012, doi: 10.1002/emmm.201200245.
- [14] H. Cui *et al.*, "Towards multimodal foundation models in molecular cell biology," *Nature*, vol. 640, no. 8059, pp. 623–633, Apr. 2025, doi: 10.1038/s41586-025-08710-y.
- [15] J. Jumper *et al.*, "Highly accurate protein structure prediction with AlphaFold," *Nature*, vol. 596, no. 7873, pp. 583–589, Aug. 2021, doi: 10.1038/s41586-021-03819-2.
- [16] M. Baek *et al.*, "Accurate prediction of protein structures and interactions using a three-track neural network," *Science*, vol. 373, no. 6557, pp. 871–876, Aug. 2021, doi: 10.1126/science.abj8754.
- [17] Z. Lin *et al.*, "Language models of protein sequences at the scale of evolution enable accurate structure prediction," *bioRxiv*, 2022, doi: 10.1101/2022.07.20.500902.
- [18] T. Hayes *et al.*, "Simulating 500 million years of evolution with a language model," *Science*, vol. 387, no. 6736, pp. 850–858, 2025, doi: 10.1126/science.ads0018.
- [19] J. Abramson *et al.*, "Accurate structure prediction of biomolecular interactions with AlphaFold 3," *Nature*, vol. 630, no. 8016, pp. 493–500, June 2024, doi: 10.1038/s41586-024-07487-w.
- [20] A. Vaswani *et al.*, "Attention Is All You Need," in *Advances in Neural Information Processing Systems*, 2017, pp. 5998–6008. doi: 10.5555/3295222.3295349.
- [21] Ž. Avsec *et al.*, "Effective gene expression prediction from sequence by integrating long-range interactions," *Nat. Methods*, vol. 18, pp. 1196–1203, 2021, doi: 10.1038/s41592-021-01252-x.
- [22] H. Cui *et al.*, "scGPT: toward building a foundation model for single-cell multi-omics using generative AI," *Nat. Methods*, vol. 21, no. 8, pp. 1470–1480, Aug. 2024, doi: 10.1038/s41592-024-02201-0.
- [23] C. V. Theodoris *et al.*, "Transfer learning enables predictions in network biology," *Nature*, vol. 618, no. 7965, pp. 616–624, June 2023, doi: 10.1038/s41586-023-06139-9.
- [24] F. Yang *et al.*, "scBERT as a large-scale pretrained deep language model for cell type annotation of single-cell RNA-seq data," *Nat. Mach. Intell.*, vol. 4, no. 10, pp. 852–866, Oct. 2022, doi: 10.1038/s42256-022-00534-z.
- [25] N. S. Detlefsen, S. Hauberg, and W. Boomsma, "Learning meaningful representations of protein sequences," *Nat. Commun.*, vol. 13, no. 1, p. 1914, Apr. 2022, doi: 10.1038/s41467-022-29443-w.
- [26] J. Cheng *et al.*, "Accurate proteome-wide missense variant effect prediction with AlphaMissense," *Science*, vol. 381, no. 6664, p. eadg7492, 2023, doi: 10.1126/science.adg7492.
- [27] Ž. Avsec *et al.*, "AlphaGenome: advancing regulatory variant effect prediction with a unified DNA sequence model," *bioRxiv*, 2025, doi: 10.1101/2025.06.25.661532.
- [28] P. R. H. J. Timmers and others, "Genomics of 1 Million Parent Lifespans Implicates Novel Pathways and Common Diseases and Distinguishes Survival Chances," *Nat. Commun.*, vol. 10, p. 3367, 2019, doi: 10.1038/s41467-019-11358-2.
- [29] M. Moqri, C. Herzog, J. Poganik, J. Justice, and and others, "Biomarkers of aging for the identification and evaluation of longevity interventions," *Cell*, vol. 186, no. 18, pp. 3758–3775, 2023, doi: 10.1016/j.cell.2023.08.003.
- [30] R. J. Sternberg and K. Sternberg, *Cognitive Psychology*, 7th ed. Cambridge University Press, 2025.

- [31] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. The MIT Press, 2016.
- [32] T. Diwan, G. Anirudh, and J. V. Tembhurne, "Object detection using YOLO: challenges, architectural successors, datasets and applications," *Multimed. Tools Appl.*, vol. 82, no. 6, pp. 9243–9275, Mar. 2023, doi: 10.1007/s11042-022-13644-y.
- [33] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997, doi: 10.1162/neco.1997.9.8.1735.
- [34] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust Speech Recognition via Large-Scale Weak Supervision." 2022. [Online]. Available: <https://arxiv.org/abs/2212.04356>
- [35] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving Language Understanding by Generative Pre-Training." 2018. [Online]. Available: https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf
- [36] T. B. Brown and others, "Language Models Are Few-Shot Learners," in *Advances in Neural Information Processing Systems*, 2020, pp. 1877–1901.
- [37] T. Brants, A. C. Popat, P. Xu, F. J. Och, and J. Dean, "Large Language Models in Machine Translation," in *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, 2007, pp. 858–867.
- [38] S. Reed *et al.*, "A Generalist Agent." 2022. [Online]. Available: <https://arxiv.org/abs/2205.06175>
- [39] D. Lopez-Paz and M. Ranzato, "Gradient Episodic Memory for Continual Learning," in *Advances in Neural Information Processing Systems*, 2017. [Online]. Available: <https://arxiv.org/abs/1706.08840>
- [40] S. Y. Lee, S. Choi, and S.-Y. Chung, "Sample-Efficient Deep Reinforcement Learning via Episodic Backward Update," in *Advances in Neural Information Processing Systems*, 2019. [Online]. Available: <https://papers.nips.cc/paper/2019/hash/e6d8545daa42d5ced125a4bf747b3688-Abstract.html>
- [41] C. de M. d'Autume, S. Ruder, L. Kong, and D. Yogatama, "Episodic memory in lifelong language learning," in *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, Red Hook, NY, USA: Curran Associates Inc., 2019.
- [42] D. M. Mistry, A. Bajaj, Y. Aggarwal, S. S. Maini, and Z. Tiganj, "Emergence of Episodic Memory in Transformers: Characterizing Changes in Temporal Structure of Attention Scores During Training," in *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, L. Chiruzzo, A. Ritter, and L. Wang, Eds., Albuquerque, New Mexico: Association for Computational Linguistics, Apr. 2025, pp. 8893–8911. doi: 10.18653/v1/2025.naacl-long.448.
- [43] G. I. Parisi, R. Kemker, J. L. Part, C. Kanan, and S. Wermter, "Continual Lifelong Learning with Neural Networks: A Review," *Neural Netw.*, vol. 113, pp. 54–71, 2019, doi: 10.1016/j.neunet.2019.01.012.
- [44] R. Kemker, M. Abitino, T. McClure, and C. Kanan, "Measuring Catastrophic Forgetting in Neural Networks," *Proc. AAAI Conf. Artif. Intell.*, vol. 32, no. 1, 2018, doi: 10.1609/aaai.v32i1.11651.
- [45] G. Guizzardi, "Ontological foundations for structural conceptual models," PhD Thesis - Research UT, graduation UT, University of Twente, 2005.
- [46] G. Guizzardi, A. B. Benevides, C. M. Fonseca, D. Porello, J. P. A. Almeida, and T. P. Sales, "UFO: Unified Foundational Ontology," *Appl. Ontol.*, vol. 17, no. 1, pp. 167–210, 2022, doi: 10.3233/AO-210256.
- [47] A. Hogan *et al.*, "Knowledge Graphs," *ACM Comput Surv*, vol. 54, no. 4, July 2021, doi: 10.1145/3447772.
- [48] N. L. Podkolodnyy, O. A. Podkolodnaya, V. A. Ivanisenko, and M. A. Marchenko, "Ontologies in Modelling and Analysing of Big Genetic Data," *Vavilov J. Genet. Breed.*, vol. 28, no. 8, pp. 940–949, 2024, doi: 10.18699/vjgb-24-101.
- [49] M. Ashburner *et al.*, "Gene Ontology: tool for the unification of biology," *Nat. Genet.*, vol. 25, no. 1, pp. 25–29, May 2000, doi: 10.1038/75556.
- [50] B. Smith, J. Williams, and S. Schulze-Kremer, "The ontology of the gene ontology," *AMIA Annu. Symp. Proc. AMIA Symp.*, vol. 2003, pp. 609–613, 2003.
- [51] T. G. O. Consortium, "The Gene Ontology resource: enriching a GOld mine," *Nucleic Acids Res.*, vol. 49, no. D1, pp. D325–D334, Dec. 2020, doi: 10.1093/nar/gkaa1113.
- [52] P. Chandak, K. Huang, and M. Zitnik, "Building a knowledge graph to enable precision medicine," *Sci. Data*, vol. 10, no. 1, p. 67, Feb. 2023, doi: 10.1038/s41597-023-01960-3.
- [53] A. d'Avila Garcez, K. Broda, D. Gabbay, and A. F. De Souza, "Knowledge extraction from trained neural

- networks: a position paper,” in *ICONIP’99. ANZIS’99 & ANNES’99 & ACNN’99. 6th International Conference on Neural Information Processing. Proceedings (Cat. No. 99EX378)*, IEEE, 1999, pp. 685–691.
- [54] A. S. d’Avila Garcez, K. Broda, and D. M. Gabbay, “Symbolic Knowledge Extraction from Trained Neural Networks: A Sound Approach,” *Artif. Intell.*, vol. 125, no. 1–2, pp. 155–207, 2001, doi: 10.1016/S0004-3702(00)00077-1.
- [55] A. d’Avila Garcez, M. Gori, L. C. Lamb, L. Serafini, M. Spranger, and S. N. Tran, “Neural-symbolic computing: An effective methodology for principled integration of machine learning and reasoning,” *ArXiv Prepr. ArXiv190506088*, 2019.
- [56] A. d’Avila Garcez and L. C. Lamb, “Neurosymbolic AI: The 3rd Wave.” 2020. [Online]. Available: <https://arxiv.org/abs/2012.05876>
- [57] H. Kim and J.-S. Kim, “A Guide To Genome Engineering With Programmable Nucleases,” *Nat. Rev. Genet.*, vol. 15, pp. 321–334, 2014, doi: 10.1038/nrg3686.
- [58] T. Gaj, C. A. Gersbach, and C. F. Barbas, “ZFN, TALEN, and CRISPR/Cas-based methods for genome engineering,” *Trends Biotechnol.*, vol. 31, no. 7, pp. 397–405, July 2013, doi: 10.1016/j.tibtech.2013.04.004.
- [59] J. A. Doudna and E. Charpentier, “The New Frontier of Genome Engineering with CRISPR-Cas9,” *Science*, vol. 346, no. 6213, p. 1258096, 2014, doi: 10.1126/science.1258096.
- [60] L. Cong *et al.*, “Multiplex Genome Engineering Using CRISPR/Cas Systems,” *Science*, vol. 339, no. 6121, pp. 819–823, 2013, doi: 10.1126/science.1231143.
- [61] B. Zetsche *et al.*, “Cpf1 Is a Single RNA-Guided Endonuclease of a Class 2 CRISPR-Cas System,” *Cell*, vol. 163, no. 3, pp. 759–771, 2015, doi: 10.1016/j.cell.2015.09.038.
- [62] A. C. Komor, A. A. Y. Kim, M. S. Packer, J. A. Zuris, and D. R. Liu, “Programmable editing of a target base in genomic DNA without double-stranded DNA cleavage,” *Nature*, vol. 533, no. 7603, pp. 420–424, 2016, doi: 10.1038/nature17946.
- [63] A. V. Anzalone *et al.*, “Search-and-replace genome editing without double-strand breaks or donor DNA,” *Nature*, vol. 576, no. 7785, pp. 149–157, 2019, doi: 10.1038/s41586-019-1711-4.
- [64] N. Marino *et al.*, “Towards AI-driven longevity research: An overview,” *Front. Aging*, vol. Volume 4-2023, 2023, doi: 10.3389/fragi.2023.1057204.
- [65] M. Jinek, K. Chylinski, I. Fonfara, M. Hauer, J. A. Doudna, and E. Charpentier, “A Programmable Dual-RNA-Guided DNA Endonuclease in Adaptive Bacterial Immunity,” *Science*, vol. 337, no. 6096, pp. 816–821, 2012, doi: 10.1126/science.1225829.
- [66] A. Bhardwaj and V. Nain, “TALENs—an indispensable tool in the era of CRISPR: a mini review,” *J. Genet. Eng. Biotechnol.*, vol. 19, no. 1, p. 125, Dec. 2021, doi: 10.1186/s43141-021-00225-z.
- [67] H. Frangoul, D. Altshuler, M. D. Cappellini, and others, “CRISPR-Cas9 Gene Editing for Sickle Cell Disease and beta-Thalassemia,” *N. Engl. J. Med.*, vol. 384, no. 3, pp. 252–260, 2021, doi: 10.1056/NEJMoa2031054.
- [68] C. Sheridan, “The World’s First CRISPR Therapy Is Approved: Who Will Receive It?,” *Nat. Biotechnol.*, vol. 41, pp. 1298–1300, 2023.
- [69] E. A. Stadtmauer, J. A. Fraietta, M. M. Davis, and others, “CRISPR-engineered T cells in patients with refractory cancer,” *Science*, vol. 367, no. 6481, p. eaba7365, 2020, doi: 10.1126/science.aba7365.
- [70] J. D. Gillmore, E. Gane, J. Taubel, and others, “CRISPR-Cas9 In Vivo Gene Editing for Transthyretin Amyloidosis,” *N. Engl. J. Med.*, vol. 385, no. 6, pp. 493–502, 2021, doi: 10.1056/NEJMoa2107454.
- [71] S. M. Fadul, A. Arshad, and R. Mehmood, “CRISPR-Based Epigenome Editing: Mechanisms and Applications,” *Epigenomics*, vol. 15, no. 21, pp. 1137–1155, 2023, doi: 10.2217/epi-2023-0281.
- [72] H. Park, J. Shin, Y. Kim, T. Saito, T. C. Saido, and J. Kim, “CRISPR/dCas9-Dnmt3a-mediated targeted DNA methylation of APP rescues brain pathology in a mouse model of Alzheimer’s disease,” *Transl. Neurodegener.*, vol. 11, no. 1, p. 41, Sept. 2022, doi: 10.1186/s40035-022-00314-0.
- [73] F. Tremblay *et al.*, “A potent epigenetic editor targeting human PCSK9 for durable reduction of low-density lipoprotein cholesterol levels,” *Nat. Med.*, vol. 31, no. 4, pp. 1329–1338, Apr. 2025, doi: 10.1038/s41591-025-03508-x.
- [74] S. Huang, C.-H. Lau, C. Tin, and R. H. W. Lam, “Extended replicative lifespan of primary resting T cells by CRISPR/dCas9-based epigenetic modifiers and transcriptional activators,” *Cell. Mol. Life Sci.*, vol. 81, no.

- 1, p. 407, Sept. 2024, doi: 10.1007/s00018-024-05415-9.
- [75] R. Abujarour *et al.*, “A chemical approach facilitates CRISPRa-only human iPSC generation and minimizes the number of targeted loci required,” *Future Sci. OA*, vol. 10, no. 1, p. FSO964, Feb. 2024, doi: 10.2144/fsoa-2023-0257.
- [76] J. Huang, Y. Zhou, J. Li, A. Lu, and C. Liang, “CRISPR/Cas systems: Delivery and application in gene therapy,” *Front. Bioeng. Biotechnol.*, vol. Volume 10-2022, 2022, doi: 10.3389/fbioe.2022.942325.
- [77] A. Cavazza *et al.*, “Advanced delivery systems for gene editing: A comprehensive review from the GenE-HumDi COST Action Working Group,” *Mol. Ther. Nucleic Acids*, vol. 36, no. 1, p. 102457, 2025, doi: <https://doi.org/10.1016/j.omtn.2025.102457>.
- [78] T. Li *et al.*, “CRISPR/Cas9 therapeutics: progress and prospects,” *Signal Transduct. Target. Ther.*, vol. 8, no. 1, p. 36, Jan. 2023, doi: 10.1038/s41392-023-01309-7.
- [79] D. C. Luther, Y. W. Lee, H. Nagaraj, F. Scaletti, and V. M. Rotello, “Delivery approaches for CRISPR/Cas9 therapeutics in vivo: advances and challenges,” *Expert Opin. Drug Deliv.*, vol. 15, no. 9, pp. 905–913, 2018, doi: 10.1080/17425247.2018.1517746.
- [80] E. Nguyen *et al.*, “Sequence modeling and design from molecular to genome scale with Evo,” *Science*, vol. 386, no. 6723, p. eado9336, 2024, doi: 10.1126/science.ado9336.
- [81] R. Tacutu *et al.*, “Human Ageing Genomic Resources: new and updated databases,” *Nucleic Acids Res.*, vol. 46, no. D1, pp. D1083–D1090, Nov. 2017, doi: 10.1093/nar/gkx1042.
- [82] J. P. de Magalhães *et al.*, “Human Ageing Genomic Resources: updates on key databases in ageing research,” *Nucleic Acids Res.*, vol. 52, no. D1, pp. D900–D908, Nov. 2023, doi: 10.1093/nar/gkad927.
- [83] A. Budovsky *et al.*, “LongevityMap: a database of human genetic variants associated with longevity,” *Trends Genet.*, vol. 29, no. 10, pp. 559–560, 2013, doi: <https://doi.org/10.1016/j.tig.2013.08.003>.
- [84] J. P. de Magalhães and O. Toussaint, “GenAge: a genomic and proteomic network map of human ageing,” *FEBS Lett.*, vol. 571, no. 1–3, pp. 243–247, 2004, doi: <https://doi.org/10.1016/j.febslet.2004.07.006>.
- [85] P. F. Christiano, J. Leike, T. Brown, M. Martic, S. Legg, and D. Amodei, “Deep reinforcement learning from human preferences,” *Adv. Neural Inf. Process. Syst.*, vol. 30, 2017.
- [86] D. Guo *et al.*, “DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning,” *Nature*, vol. 645, no. 8081, pp. 633–638, Sept. 2025, doi: 10.1038/s41586-025-09422-z.
- [87] D. Silver *et al.*, “Mastering the game of Go with deep neural networks and tree search,” *Nature*, vol. 529, no. 7587, pp. 484–489, Jan. 2016, doi: 10.1038/nature16961.
- [88] R. M. Durbin *et al.*, “A map of human genome variation from population-scale sequencing,” *Nature*, vol. 467, no. 7319, pp. 1061–1073, Oct. 2010, doi: 10.1038/nature09534.
- [89] J. Deelen *et al.*, “A meta-analysis of genome-wide association studies identifies multiple longevity genes,” *Nat. Commun.*, vol. 10, no. 1, p. 3669, Aug. 2019, doi: 10.1038/s41467-019-11558-2.
- [90] J. Xu, G. Gontier, Z. Chaker, P. Lacube, J. Dupont, and M. Holzenberger, “Longevity effect of IGF-1R+/- mutation depends on genetic background-specific receptor activation,” *Aging Cell*, vol. 13, no. 1, pp. 19–28, 2014, doi: <https://doi.org/10.1111/accel.12145>.

Anexo 1

Plano de Governança, Segurança e Tratamento de Dados

1. Origem e natureza dos dados

Os dados genômicos utilizados neste projeto serão obtidos a partir do repositório controlado European Genome-phenome Archive (EGA - <https://ega-archive.org/>), que é um repositório internacional especializado no armazenamento e distribuição segura de dados genéticos humanos sensíveis.

O EGA é projetado para distribuir conjuntos de dados derivados de humanos sob acesso controlado, onde o acesso geralmente requer procedimentos formais de aprovação (ou seja, acesso permissionado em vez de acesso totalmente aberto - <https://ega-archive.org/submission/metadata/submission/FAQ/>). Ele documenta publicamente diversas propriedades de segurança relevantes para a alegação de “download seguro”: verificações de integridade durante a transmissão (*hash*), criptografia de ponta a ponta em trânsito e criptografia forte em repouso, com chaves de criptografia não armazenadas no mesmo sistema que os dados criptografados (<https://ega-archive.org/about/security/>).

Para o download dos dados, o EGA suporta mais de um caminho operacional, e este protocolo seguirá a abordagem documentada pelo EGA para o conjunto de dados específico e a autorização concedida:

- **Cliente de download PyEGA3 (download baseado em API).** O EGA documenta o PyEGA3 como o cliente de download principal, que realiza chamadas HTTPS à API de Dados do EGA (acessível pela porta 8443) e realiza verificação de integridade pós-download usando checksums (md5 - <https://ega-archive.org/access/download/files/pyega3/>). O EGA também documenta que, neste fluxo específico do PyEGA3, os arquivos são transferidos via HTTPS seguro e são recebidos descriptografados no destino (ou seja, nenhuma etapa separada de descriptografia é necessária após o download - <https://ega-archive.org/access/download/files/pyega3/>).
- **Live distribution via “Outbox” (SFTP + arquivos criptografados).** O EGA também documenta um fluxo de *Live Distribution* no qual o usuário se conecta via SFTP e se autentica usando um arquivo de chave, e os arquivos arquivados são criptografados com base no Crypt4GH (exigindo uma etapa de descriptografia usando ferramentas Crypt4GH - <https://ega-archive.org/access/download/files/live-outbox/>).

Como o EGA pode distribuir arquivos criptografados usando diferentes sistemas de criptografia dependendo do local de arquivamento (por exemplo, Crypt4GH e outros tipos de criptografia gerenciados pelo EGA), o projeto tratará todos os arquivos Welllderly como dados genômicos humanos sensíveis de acesso controlado, usando o método de distribuição do EGA aprovado para o conjunto de dados, e aplicando controles de acesso rigorosos e salvaguardas de armazenamento, independentemente de um fluxo específico de download do EGA resultar em um artefato criptografado em repouso no ponto de download (<https://ega-archive.org/check-encryption-type/>).

O Crypt4GH é um padrão comunitário reconhecido pela *Global Alliance for Genomics and Health* (GA4GH) para manter dados genômicos seguros ao longo de todo o seu ciclo de vida (armazenamento, transferência e análise), permitindo acesso aleatório (<https://www.ga4gh.org/product/genetic-data-encryption-crypt4gh/>). O EGA fornece orientações de ferramentas para uso do Crypt4GH em sua própria documentação (<https://ega-archive.org/submission/data/file-preparation/crypt4gh/>).

2. Armazenamento e infraestrutura computacional

Os dados da base Wellderly, de acesso controlado, serão armazenados e processados no supercomputador Atena, que é um sistema computacional de alto desempenho projetado para suportar cargas de trabalho de aprendizado profundo em larga escala (<https://ufes.br/conteudo/evento-debate-inteligencias-artificiais-aplicadas-e-inaugura-supercomputador>).

O Atena é parte de uma infraestrutura criada em um contexto diretamente relevante para segurança da informação e hospedagem física. Ele o data center recentemente reformado da Superintendência de TI da UFES atendem a padrões elevados de segurança de rede e armazenamento (<https://www.ufes.br/conteudo/representantes-do-consulado-geral-dos-eua-visitam-ufes-para-discutir-parcerias-nas-areas>).

O ambiente computacional deste projeto está associado ao Laboratório de Computação de Alto Desempenho (LCAD/DI/UFES - <https://www.lcad.inf.ufes.br/>) do Departamento de Informática (DI/UFES - <https://informatica.ufes.br/>) da UFES, que abriga o Atena, e cujo propósito é fornecer um ambiente computacional para aplicações que se beneficiam de computação paralela/distribuída e apoiar projetos que requerem alta capacidade de processamento, memória e armazenamento.

No nível do Departamento de Informática da UFES (DI/UFES - <https://informatica.ufes.br/>), o Laboratório de Administração de Redes (LAR/DI/UFES - <https://informatica.ufes.br/pt-br/lar-laboratorio-de-administracao-de-redes>) mantém a rede corporativa do Departamento de Informática e de seus laboratórios, incluindo o LCAD.

Do ponto de vista da governança institucional, o ambiente de rede e data center no qual o Atena/LCAD/DI opera é apoiado pela Superintendência de Tecnologia da Informação (STI/UFES - <https://npd.ufes.br/>), que inclui explicitamente “segurança” entre seus valores e é responsável por fornecer soluções/serviços de TI à comunidade da UFES (<https://npd.ufes.br/sobre-superintendencia-de-tecnologia-da-informacao-sti>). A estrutura organizacional da STI também documenta responsabilidades alinhadas com controles de segurança de perímetro e infraestrutura, incluindo a coordenação da infraestrutura/equipamentos de rede do data center, gerenciamento de serviços de rede, garantia de segurança da informação e implementação de políticas de backup (<https://npd.ufes.br/estrutura-organizacional-e-equipes>). O DI, LAR e LCAD aderem às diretrizes institucionais estabelecidas pela STI.

3. Governança e controle de acesso

O acesso aos dados será restrito e controlado conforme os seguintes princípios:

3.1. Perfis de acesso

Terão acesso aos dados:

- pesquisadores listados no projeto aprovado;
- bolsistas formalmente vinculados ao projeto e previamente autorizados e cadastrados nos sistemas do LCAD/LAR.

3.2. Controle de acesso

- autenticação individual (login e senha institucionais);
- controle baseado em permissões (least privilege);
- revisão periódica dos acessos concedidos.

3.3. Compromissos formais

Todos os usuários deverão:

- assinar termo de confidencialidade;
 - cumprir as políticas do EGA e da UFES/STI/DI/LAR e /LCAD;
 - utilizar os dados exclusivamente para os fins científicos aprovados.
-

4. Colaboração internacional

Por se tratar de estudo com dados provenientes de repositório internacional, a colaboração seguirá os seguintes princípios:

- os dados não serão redistribuídos para instituições externas;
 - eventuais colaborações científicas ocorrerão por meio de:
 - compartilhamento de resultados agregados (não identificáveis);
 - modelos computacionais treinados (sem dados brutos);
 - qualquer transferência de dados dependerá de:
 - autorização da coordenação do projeto;
 - aderência às normas de segurança de dados da UFES/STI/DI/LAR e /LCAD.
-

5. Medidas técnicas de segurança

Serão adotadas as seguintes medidas:

- criptografia durante transferência de dados (protocolos seguros do EGA);
 - armazenamento em ambiente protegido por rede institucional;
 - isolamento lógico dos dados no cluster;
 - backups controlados conforme política institucional;
 - ausência de identificação direta dos indivíduos (dados pseudonimizados);
 - proibição de download para máquinas locais.
-

6. Medidas administrativas

Serão adotadas as seguintes medidas:

- definição formal de responsáveis pelos dados (PI e equipe);
 - capacitação dos usuários em segurança da informação;
 - auditorias internas periódicas;
 - conformidade com a LGPD (Lei Geral de Proteção de Dados – Lei nº 13.709/2018);
 - cumprimento dos termos de uso do EGA.
-

Mapa de Fluxo de Dados

1. Fluxo geral

O fluxo de dados seguirá as etapas abaixo:

1. **Solicitação de acesso ao EGA**
↓
 2. **Aprovação pelo Data Access Committee (DAC)**
↓
 3. **Download seguro dos dados (criptografado)**
↓
 4. **Transferência para o supercomputador Atena (LCAD/UFES)**
↓
 5. **Armazenamento em ambiente seguro e controlado (Atena)**
↓
 6. **Processamento computacional (modelos de IA que rodarão no Atena)**
↓
 7. **Geração de resultados agregados**
↓
 8. **Publicação científica (sem dados identificáveis)**
↓
 9. **Descarte dos dados**
-

2. Localização dos dados

Etapa	Localização
Origem	EGA (repositório internacional)
Transferência	Canal seguro criptografado
Armazenamento	Supercomputador Atena (LCAD/UFES)
Processamento	Atena (LCAD/UFES)
Resultados	UFES / publicações

3. Acesso por instituição

Instituição	Acesso aos dados brutos
UFES (LCAD)	Sim
Pesquisadores do projeto	Sim (controlado)
Instituições externas	Não (apenas resultados agregados)

Política de Retenção e Descarte

1. Retenção

Os dados serão mantidos:

- durante a vigência do projeto;
- pelo período adicional necessário para reprodutibilidade científica;
- sempre respeitando as restrições do EGA.

2. Descarte

Ao término do projeto:

- os dados serão:
 - removidos dos sistemas de armazenamento, ou
 - mantidos apenas se permitido pelos termos do EGA;
- o descarte seguirá procedimentos seguros (deleção controlada).

3. Encerramento de acessos

Ao término do projeto:

- com exceção do coordenador, todos os acessos aos dados serão revogados;
 - com exceção do coordenador, todas as credenciais serão desativadas;
 - logs serão arquivados para auditoria.
-

Conclusão

O projeto adota uma abordagem robusta de governança e segurança de dados, alinhada a padrões internacionais (EGA), à legislação brasileira (LGPD) e às políticas institucionais da UFES, garantindo:

- proteção de dados genéticos sensíveis;
- controle rigoroso de acesso;
- uso ético e responsável das informações;
- rastreabilidade e auditabilidade.