Dear Director, I would like to present my Teaching and Research Statement below: Computer Science and Engineering Department of the Instituto Superior Técnico - University of Lisbon

Teaching and research statement

Bioinformatics for Functional Annotation and Phylogeny

My teaching philosophy emphasizes fostering curiosity and critical thinking through an engaging learning environment. I aim to make complex biological concepts, like enzyme function and evolution, accessible through practical bioinformatics examples. By incorporating real-world applications and promoting collaborative learning, I hope to inspire students to apply interdisciplinary approaches to solve biological challenges. I propose a course titled "Bioinformatics for Functional Annotation and Phylogenetics," covering essential bioinformatics concepts (Homology and Analogy, Galperin & Koonin, 2012), tools and databases (Mount, 2004; Lesk, 2020), with a focus on functional annotation of proteins, identification of homologous sequences, and phylogenetic analysis. The course will introduce bioinformatics basics, such as exploring databases like NCBI and UniProt, and will teach practical skills using tools like BLASTp (Altschul et al., 1990) and InterProScan (Jones et al., 2014) for protein function prediction. Students will explore sequence alignment algorithms like MAFFT (Katoh & Standley, 2013) and the Needleman-Wunsch (Needleman & Wunsch, 1970) and Smith-Waterman (Smith & Waterman, 1981) algorithms, with workshops on multiple sequence alignment and sequence conservation. The course will also cover phylogenetic tree construction using methods like maximum likelihood and Bayesian inference (Felsenstein, 1981; Yang, 2014), and introduce AI-driven tools such as AlphaFold2 (Jumper et al., 2021) for protein structure prediction. Students will apply these concepts to global challenges, including enzymes related to climate change and pollution. They will formulate hypotheses and mini-projects on enzyme functions (Yoshida et al., 2016; Kumar & Bharadvaja, 2019) that can be utilized through synthetic biology to address emerging issues such as global warming, environmental pollution from metals, plastics, textiles, etc. They will design *in silico* experiments and work collaboratively on projects. Teaching methods will include lectures and seminars to introduce theoretical concepts (30%), hands-on workshops for practical experience (40%), group projects to encourage teamwork (30%) with presentations and discussions to develop communication skills and critical thinking. Assessment will consist of in-class assignments (30%) involving short reports, a group project (40%) culminating in a detailed report and presentation, and examinations (30%) to assess understanding of theoretical concepts. By the end of the course, students will gain proficiency in bioinformatics tools, develop scientific hypotheses, enhance teamwork and communication skills, and understand biotechnology's role in solving environmental challenges. The course aims to bridge the gap between theory and practical application, preparing students for future careers in science and biotechnology.

Integration of Artificial Intelligence (AI) and Bioinformatics

My teaching philosophy centers on fostering curiosity and developing critical thinking skills in postgraduate students. I aim to create an engaging learning environment where complex concepts—such as the integration of Artificial Intelligence (AI) and Bioinformatics—become accessible through practical examples and hands-on experiences. By emphasizing real-world applications and collaborative learning, my goal is to inspire students to think creatively and apply interdisciplinary approaches to solve complex

biological challenges. I propose to offer a foundational postgraduate course titled **"Integrating AI and Bioinformatics for Biotechnological Applications (Postgraduate Level)."** This course will cover recent advancements in AI and Bioinformatics integration, focusing on Natural Language Processing (NLP) models like Word2Vec (Mikolov et al., 2013) and neural networks such as the Perceptron (Rosenblatt, 1958), Multilayer Perceptron (Rumelhart et al., 1986), and transformers (Vaswani et al., 2017). Students will learn how these algorithms can extract protein sequence embeddings to identify hidden patterns in protein sequences and integrate them with traditional machine learning methods such as Support Vector Machines (SVM) and Random Forest for functional classification of proteins. The course will also explore advanced models based on transformer algorithms like ESM2 (Lin et al., 2023) for enzyme annotation and ProtGPT2 (Ferruz et al., 2022) for protein design. A significant component of the course will be practical workshops using Python notebooks exploring the methods presented. Students will engage in hands-on coding sessions to implement the algorithms and models discussed, allowing them to apply AI techniques to real bioinformatics problems. During the course, students will be divided into groups, with each group assigned to present one of the methods covered and lead a corresponding Python workshop showing the use of the methods presented during the class session. This collaborative approach will enhance peer learning and ensure a deep understanding of each technique. Throughout the course, we will demonstrate the integration of these methods with both supervised learning algorithms (e.g., SVM, Random Forest) used to train classifiers aimed at the functional annotation of sequences, and unsupervised learning algorithms (e.g., K-means clustering, Principal Component Analysis (PCA), Non-metric Multidimensional Scaling (NMDS)) for the analysis of embeddings generated from protein sequences. Topics will include the history of Natural Language Processing and large language models, foundational AI concepts like the Perceptron and Multilayer Perceptron, advanced AI models applied to functional annotation of proteins and protein prediction (3D structures), practical applications of AI in biomedical research—including drug discovery—and short discussions on ethical and security considerations. Assessment will consist of in-class assignments (20%), Python workshops (20%), and group presentations (30%) where students present on the assigned methods and propose how these methods can be used for other applications, participation (20%), and a final exam (10%). By the end of the course, students will have a solid understanding of foundational AI concepts and advanced AI models, equipping them with practical skills to apply AI in predicting protein functions, designing novel proteins, and addressing real-world challenges in molecular biology and biotechnology—all while considering ethical and security implications.

Outreach Vision

Outreach is a fundamental part of my academic vision, as I believe that scientific knowledge should extend beyond the classroom and laboratory. I plan to engage with different communities by organizing free workshops presented by students, showcasing the programs they developed during their training and internships in the lab, and hosting small talks organized with the students on topics involving bioinformatics and artificial intelligence to demystify complex subjects such as enzyme engineering and AI. These events will be tailored to various audiences, including school students, business leaders, and the general community. The goal of this outreach is to stimulate interest in science, especially among underrepresented groups, by demonstrating how AI can lead to environmentally friendly industrial solutions. Additionally, I see great potential in using online platforms and social media (Instagram, Facebook, and LinkedIn and Youtube) to share research findings in an accessible format, ensuring the general public understands the societal impact of biotechnological innovations. Through collaboration with schools and educational

programs, I plan to introduce young students to biotechnology and bioinformatics concepts, aiming to foster future interest in STEM fields.

Technology Transfer (Valorization) Vision

My project on Non-Homologous Isofunctional Enzymes (NISEs) has significant potential for technology transfer, particularly in industries such as biofuels, pharmaceuticals, and environmental sustainability. I envision the development of a comprehensive database of NISEs, along with AI programs based on transformer algorithms like ProtGPT2, becoming a vital tool for both academic and industrial partners. This platform will facilitate the creation of custom-designed enzymes and genetic circuits with enhanced efficiency and specificity for industrial applications. My goal is to establish collaborations with industry leaders to explore the commercialization of synthetic enzymes generated through this project, and to better understand market needs and the existing gaps in tools that need to be developed to meet societal demands. This could lead to patents and licensing opportunities. Additionally, I plan to work closely with technology transfer offices to ensure that research outcomes are translated into tangible products and processes, ultimately contributing to sustainable industrial practices and economic growth.

Other potential disciplines/courses:

Python Programming, Algorithms

Introduction to Bioinformatics

Microbiome and Metagenomic Analyses

Functional annotation of Genomes

RNAseq and Transcriptome Analyses

All topics in which I have extensive experience in bioinformatics analyses.

My Previous experience with teaching (Classrooms and Courses):

- **2017: Preceptor in "NGS and Bioinformatics" for the Biology and Biomedicine Programs, Multiprofessional Residency in Cancer - COREMU. 16-hour workload on July 20th and 21st at Barretos Cancer Hospital.**
- **2018: Preceptor in "NGS and Bioinformatics" for the Biology and Medicine Programs, Multiprofessional Residency in Cancer - COREMU. 16-hour workload on August 30th and 31st at Barretos Cancer Hospital.**
- **2018: Instructor for the Bioinformatics Course during the IV Winter Course in Molecular Oncology. 40-hour workload from July 23rd to 27th at Barretos Cancer Hospital.**
- **2017: Extension Course in Microbiome. 44-hour workload at UFRGS – Clinical Hospital of Porto Alegre (HCPA).**

- **2016: Instructor for the Bioinformatics Course: "Genome Assembly, Functional Annotation, and Analysis of Biological Networks." 20-hour workload at PUC-RS.**

- **2013: Coordinator of the Bioinformatics Winter Course. 40-hour workload at Oswaldo Cruz-Fiocruz Institute, RJ.**

- **2011: Professor of the Bioinformatics Winter Course. 35-hour workload at Oswaldo Cruz Institute - Fiocruz, RJ.**

- **2008: Theoretical and Practical Course in Molecular Biology for the Biological Sciences course. 78-hour workload at UENF, Campos-RJ.**

- **2007: Volunteer Professor in a social course for university admission: UENF discipline: General Biology. 1-year workload.**

References:

- Galperin, M. Y., & Koonin, E. V. (2011). Divergence and convergence in enzyme evolution. *Journal of Biological Chemistry, 287*(1), 21–28. https://doi.org/10.1074/jbc.R111.241976
- Mount, D. (2004). *Bioinformatics: Sequence and Genome Analysis* (2nd ed.). Cold Spring Harbor Laboratory Press.
- Lesk, A. (2020). *Introduction to Bioinformatics* (5th ed.). Oxford University Press.
- Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology, 215*(3), 403-410. https://doi.org/10.1016/S0022-2836(05)80360-2
- Jones, P., Binns, D., Chang, H. Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., & Nuka, G. (2014). InterProScan 5: genome-scale protein function classification. *Bioinformatics, 30*(9), 1236–1240. https://doi.org/10.1093/bioinformatics/btu031
- Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution, 30*(4), 772-780. https://doi.org/10.1093/molbev/mst010
- Needleman, S. B., & Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology, 48*(3), 443-453. https://doi.org/10.1016/0022-2836(70)90057-4
- Smith, T. F., & Waterman, M. S. (1981). Identification of common molecular subsequences. *Journal of Molecular Biology, 147*(1), 195-197. https://doi.org/10.1016/0022-2836(81)90087-5
- Felsenstein, J. (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution, 17*(6), 368-376. https://doi.org/10.1007/BF01734359
- Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., ... Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature, 596*(7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2
- Yoshida, S., Hiraga, K., Takehana, T., Taniguchi, I., Yamaji, H., Maeda, Y., Toyohara, K., Miyamoto, K., Kimura, Y., & Oda, K. (2016). A bacterium that degrades and assimilates

poly(ethylene terephthalate). *Science, 351*(6278), 1196-1199. https://doi.org/10.1126/science.aad6359

- Kumar, L., & Bharadvaja, N. (2019). **Enzymatic bioremediation: A smart tool to fight environmental pollutants**. In *Smart Bioremediation Technologies*, *Microbial Enzymes* (pp. 99–118).
- Mikolov, T., Chen, K., Corrado, G. S., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.
- Rosenblatt, F. (1958). **The perceptron: A probabilistic model for information storage and organization in the brain**. *Psychological Review*, *65*(6), 386–408. https://doi.org/10.1037/h0042519
- Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, 323, 533–536. https://doi.org/10.1038/323533a0.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). **Attention is all you need**. *arXiv preprint arXiv:1706.03762v7*. https://arxiv.org/abs/1706.03762
- Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., Dos Santos Costa, A., Fazel-Zarandi, M., Sercu, T., Candido, S., & Rives, A. (2023). **Evolutionary-scale prediction of atomic-level protein structure with a language model**. *Science*, *379*(6637), 1123–1130. https://doi.org/10.1126/science.ade2574

Research Statements:

Mission, Goals, and Vision for the Computer Science and Engineering Department of the Instituto Superior Técnico - Universidade de Lisboa: To contribute to the Computer Science and Engineering Department of the Instituto Superior Técnico - Universidade de Lisboa by being a leader in the integration of bioinformatics and AI, specifically in the use of language models, AI agents, and generative AI applied to the design of new enzymes. Additionally, I aim to collaborate with other researchers and focus on the following research areas: Comparative Genomics, Gene Mining with biotechnological potential, Functional Genomics: transcriptome analyses, Phylogenomics, Microbiome analyses, and Metagenomics—areas in which I have been actively involved over the past 15 years.

My research statements for the coming years in AI in Biology: (Proposal combining fundamental and applied science)

Title: *Revealing and Exploring the Biotechnological Potential and Evolutionary Significance of Non-Homologous Isofunctional Enzymes (NISEs)*

The aim of my research plan is to investigate NISEs, i.e., enzymes that catalyze the same biochemical reactions, classified under the same four-digit Enzyme Commission (EC) number. Despite this common functional classification, these enzymes have evolved independently, leading to significant differences in their primary sequences and tertiary structures. The EC

numbering system categorizes enzymes based on the chemical reactions they catalyze, with each unique four-digit EC number representing a specific enzymatic activity. Thus, NISEs perform identical biochemical functions, as indicated by their EC numbers, but their evolutionary pathways have led to different structural solutions (primary and tertiary) to achieve the same biochemical transformation. These enzymes are products of convergent evolution, offering alternative biochemical solutions to the same catalytic functions (Galperin et al., 1998; Omelchenko et al., 2010; Tschoeke et al., 2014; Silva et al., 2019, Gherardini et al. et al., 2007). The interesting evolutionary feature is that the structural differences in NISEs at the secondary and tertiary fold levels can result in distinct physical (thermostability, solubility, purification, and more) and chemical/biochemical parameters (such as active site affinity and specificity, cofactor requirements, enzymatic kinetics, substrate turnover rate (Kmol/mM), pH stability, general stability, oxidation stability, and others). Additionally, these enzymes have distinct genomic contexts, differences in gene expression (Piergiorge et al., 2017), and distinct codon usage, which could be explored to select the most efficient genes encoding enzymes for cloning and with biotechnological potential. To date, all these characteristics of NISEs have not been cataloged. The fundamental question in the biology of the project is how enzymes with distinct structural properties can perform the same biochemical reaction. The project aims to annotate and catalog NISEs across all predicted proteomes obtained from genomes and metagenomes from the NR – NCBI database, MGnify databases, and animal venom data, while exploring their evolutionary significance. After the identification and annotation of the enzymes according to their EC numbers, we will create a reference database, grouping analogous enzymes into distinct clusters. Each cluster will represent a set of enzymes with different secondary and tertiary structures. The project has the potential to increase the understanding of enzyme structure and function in prokaryotes (bacteria and archeas) and eukaryotes (including Human pangenomes, parasites and others). The database created will be used in the final stage to train advanced language models by fine tuning, such as ProtGPT2, capable of generating synthetic enzymes tailored for specific biochemical activities with potential of application in different industries (health, agricultural and others). By leveraging the EC number classification system, the model will be trained to understand the relationships between enzyme structures and the biochemical reactions they catalyze, providing a foundational framework for generating novel enzymes with tailored biochemical properties, designed for specific catalytic functions. For example, alcohol dehydrogenases (EC 1.1.1.1), commonly used in the production of bioethanol and other biotechnological applications, can be modeled and synthesized using ProtGPT2 by providing different structural enzymes and label classes of distinct NISEs during the fine tuning of the model. The creation of new alcohol dehydrogenases could enhance enzyme performance in different industrial processes conditions and improve specificity for different alcohol substrates. Modifying alcohol substrates enables the creation of compounds with diverse functionalities, which impacts industrial applications such as bioethanol production, where substrate specificity and enzyme efficiency are critical for optimizing yield and process efficiency. NISEs Alcohol dehydrogenases have been identified in bacteria and eukaryotes (Omelchenko et al., 2010). This approach could utilize structtural difference observed between different enzyme able to perform the same function to design new enzymatic solutions combining distinct folds or catalytic sites, drawing from the diverse structural adaptations seen in NISEs. The central hypothesis of this research is that NISEs, due to their distinct evolutionary features such as tertiary structures and

secondary folds, present alternative biochemical solutions that can be more efficient for industrial processes. However, up to this point, this has not been fully studied or explored, and there is no work that has cataloged these enzymes on a large scale, including metagenomic data. These enzymes may offer unique advantages like enhanced thermostability, solubility, and substrate specificity, making them more suitable for industrial applications in areas such as biodiesel and bioethanol production, food and beverage processing, and pharmaceutical development. The project will contribute significantly to our understanding of enzyme function and evolution, particularly how convergent evolutionary processes give rise to enzymes with similar functions but different structures. This project aligns strongly with VIB's interest in artificial intelligence (AI) methods and mechanistic models to address fundamental biological questions. After functional annotation of NISEs the the project will use AI-driven tools like AlphaFold2 (Jumper et al., 2021) to predict the three-dimensional structures of NISEs, combining this approach with traditional structural biology techniques such as comparative molecular modeling, molecular dynamics (MD) simulations, and molecular docking. Furthermore, ProtGPT2 (Ferruz et al., 2022) will be employed to generate synthetic enzyme models based on the reference EC number database, enabling the design of enzymes optimized for specific industrial tasks. The hybrid approach of using AI and mechanistic models will provide insights into how these enzymes function at the molecular level and how they can be optimized for industrial use. The methodology begins with the collection of predicted proteomes from global genomic projects such as the Earth Microbiome Project (Thompson et al., 2017), Mgnify (https://www.ebi.ac.uk/metagenomics/), and the Animal Toxin Project (https://www.uniprot.org/help/Toxins). These proteomes will be analyzed computationally using a previously developed approach to identify NISEs by comparing enzyme sequences that share the same catalytic functions but exhibit significant structural divergence using the AnenPi-based methodology (Silva, R, Mattos, LP., et al., 2019). In our methodology, the Superfamily database program will be used to evaluate and annotate the distinct folds in each EC number. The main premise is that enzymes classified under the same EC number but with distinct structural folds, as annotated by Superfamily, are considered true analogous enzymes or NISEs. Once NISEs are identified, their physical and biochemical properties, such as thermostability and pH stability, will be analyzed using databases like BRENDA to assess their industrial potential (BRENDA Enzyme Database, 2021). The project will also use cutting-edge AI tools like AlphaFold2 to predict the three-dimensional structures of these enzymes, particularly those that do not have known homologous sequences in the Protein Data Bank (PDB). Comparative modeling using MODELLER (A. Šali and T. L. Blundell, 1993) will be applied to generate multiple models for each NISE, and the best structural models will be validated using stereochemical quality checks performed through the MOLPROBITY (https://github.com/rlabduke/MolProbity). These predictions will be further validated through molecular dynamics simulations with GROMACS (Abraham et al., 2015) and molecular docking using AutoDock to study enzyme-substrate interactions (Morris et al., 2009). For experimental validation, selected NISEs (identified in the public data or created by ProtGPT2) with promising industrial potential will be cloned and expressed in bacterial vectors, purified, and characterized biochemically to confirm their catalytic activity and efficiency. Enzymes with unknown three-dimensional structures will undergo crystallographic analysis to further validate AI predictions and study their structural and functional properties in detail in collaboration with CNPEM (https://cnpem.br/) at Brazil. This experimental validation will help confirm the viability of using AI predictions, particularly those generated by

AlphaFold, in real-world applications. A key output of the project will be the development of a web-based platform that will host the catalog of identified NISEs, along with detailed structural, functional, and genomic data. This platform will also include the EC number database used to fine-tune ProtGPT2, offering researchers and industries a tool to generate synthetic enzymes with tailored biochemical activities. By focusing on the structural and functional diversity of enzymes evolved through convergent evolution, the project will contribute significantly to the development of novel enzymes for industrial use, particularly in areas where current enzymes do not perform optimally. This project has a strong potential to impact biotechnology by introducing NISEs as alternative solutions for biocatalysis in industrial processes. The AI-driven prediction of enzyme structures, combined with experimental validation, will open new avenues for designing enzymes with enhanced properties tailored for specific industrial needs. These advancements are particularly relevant in industries such as biofuels, pharmaceuticals, and environmental sustainability, where the efficiency and specificity of enzymes are critical for optimizing processes and reducing environmental impact. The integration of AI methods with mechanistic models makes this project particularly relevant to VIB's strategic goals. By advancing our understanding of enzyme evolution and functional convergence, the project not only addresses fundamental biological questions but also has practical implications for synthetic biology and enzyme engineering. The identification and characterization of NISEs will provide a new perspective on how diverse structural folds can perform the same catalytic activities, contributing to the growing body of knowledge in structural biology and computational biology. In conclusion, this project bridges computational biology, AI, synthetic biology, and biotechnology to make significant contributions to the understanding of enzyme evolution and functionality. By identifying and validating NISEs with industrial potential, the project aims to revolutionize industrial processes by offering more efficient and sustainable enzymatic solutions. The use of AI tools like AlphaFold2 and ProtGPT2, combined with experimental techniques, will position the project at the cutting edge of biological research, aligning with VIB's focus on AI-driven advancements in biology and biotechnology. The project's outcomes will not only enhance our understanding of convergent evolution but also provide practical solutions for industries seeking to improve their production processes through innovative biocatalysis.

Methodology (Summary):

The predicted proteomes will be downloaded from NR, MGnify, and UniProtKB, and their respective EC numbers (enzymatic activities) will be functionally annotated using the methodology established by Otto et al., 2008. Once the data is collected and organized, the enzyme sequences are processed using the **BioPython** library, which reads and filters sequences based on metadata such as Biochemical features: EC numbers and clusters number (each cluster representing a set of homologous enzymes) and associated variables (others physical enzymes properties of interest). This metadata includes key information, all of which are essential for training the AI model. The **ProtGPT2** model, a pre-trained deep learning model for protein sequence generation, is fine-tuned using this dataset. The fine-tuning process involves tokenizing the enzyme sequences using the **GPT2Tokenizer** from the **transformers** library. Each sequence is paired with its associated biochemical data, and converted into tokenized inputs suitable for the model. The **GPT2LMHeadModel** from Hugging Face is then used to process these inputs. Fine-tuning is carried out with a customized version of Hugging Face's **Trainer** class,

allowing for flexible training across several epochs. The model learns the relationship between enzyme sequences and their biochemical properties, optimizing the ability to predict enzymatic activities and generate new sequences. Training is conducted on either a GPU or CPU to enhance performance, with a custom loss function ensuring accurate predictions of enzyme functions. After fine-tuning, the model generates synthetic enzyme sequences using a custom **SequenceGenerator** class. This class takes prompts related to associated biochemical data (e.g., alcohol dehydrogenase, EC 1.1.1.1), and generates tailored enzyme sequences.The generated sequences then undergo codon optimization to ensure efficient expression in the target organisms. Using a codon usage table built from the dataset, the program's `optimize_codon_sequence` function adjusts the sequences to match the most frequently used codons in the desired organism. Invalid amino acids are filtered during this process, and the optimized sequences are saved in a FASTA file. These optimized sequences are then prepared for experimental validation, where they are expressed in bacterial systems to test their catalytic activity, substrate specificity, and stability under laboratory conditions. The structural analyses will be done in collaboration with Luís Fernando Saraiva Macedo Timmers. Univates Brazil.

Project Collaborators:

Dr. Paulo de Oliveira: specialist in Synthetic Biology, Researcher at CIIMAR and i3S, Portugal.

Dra. Sandra Figueiredo is a specialist in Pharmaceutical Chemistry, Researcher at CIIMAR, Portugal.

Dra. Gisele Nunes, Head of Environmental Genomics, ITV- Institute Technologic Vale – Brazil

Luís Fernando Saraiva Macedo Timmers, Univates, Brazil, Structural Biologist Researcher.

Year	Research Focus	Key Deliverables
2025	Data collection, bioinformatics annotation of NISES, context genomic analyses, internal database construction	Initial NISEs catalog, database creation, first publication
2026	Evolutionary and structural analyses: 3D prediction and validation, simulations: Molecular dynamics and docking	Evolutionary and structural insights, second publication

2027	Model development (ProtGPT2), synthetic enzyme testing, experimental design	Synthetic enzyme testing results, third publication
2028	Cloning, expression, biochemical characterization, structure validation	Biochemical validation, optimization of models, fourth publication
2029 -2030	Web Platform development, industrial applications, final report, publication	Web Platform launch, industrial applications, final publication

References

- Galperin, M. Y., Walker, D. R., & Koonin, E. V. (1998). Analogous enzymes: independent inventions in enzyme evolution. *Genome Research*, *8*(8), 779–790. https://doi.org/10.1101/gr.8.8.779
- Omelchenko, M. V., Galperin, M. Y., Wolf, Y. I., & Koonin, E. V. (2010). Non-homologous isofunctional enzymes: a systematic analysis of alternative solutions in enzyme evolution. *Biology Direct*, 5, 31.
- Silva RA, Pereira LM, Silveira MC, Jardim R, Miranda AB. Mining of potential drug targets through the identification of essential and analogous enzymes in the genomes of pathogens of *Glycine max*, *Zea mays* and *Solanum lycopersicum*. *PLoS One*. 2018;13(5). doi: 10.1371/journal.pone.0197511.
- Tschoeke DA, Nunes GL, Jardim R, Lima J, Dumaresq AS, Gomes MR, Pereira LM, Loureiro DR, Stoco PH, Guedes HLM, Miranda AB, Ruiz J, Pitaluga A, Silva FP Jr, Probst CM, Dickens NJ, Mottram JC, Grisard EC, Dávila AM. The comparative genomics and phylogenomics of *Leishmania amazonensis* parasite. *Evol Bioinform Online*. 2014;10:131-53. doi: 10.4137/EBO.S13759.
- Gherardini, P. F., Wass, M. N., Helmer-Citterich, M., & Sternberg, M. J. E. (2007). Convergent evolution of enzyme active sites is not a rare phenomenon. *Journal of Molecular Biology*, *372*(3), 817–845. https://doi.org/10.1016/j.jmb.2007.06.017
- Piergiorge RM, Miranda AB, Guimarães AC, Catanho M. Functional analogy in human metabolism: enzymes with different biological roles or functional redundancy? *Genome Biol Evol*. 2017;9(6):1624-1636. doi: 10.1093/gbe/evx119.

- Ferruz N, Schmidt S, Höcker B. ProtGPT2 is a deep unsupervised language model for protein design. *Nat Commun*. 2022;13:4348. doi: 10.1038/s41467-022-32007-7.

- Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A, Bridgland A, Meyer C, Kohl SAA, Ballard AJ, Cowie A, Romera-Paredes B, Nikolov S, Jain R, Adler J, Back T, Petersen S, Reiman D, Clancy E, Zielinski M, Steinegger M, Pacholska M, Berghammer T, Bodenstein S, Silver D, Vinyals O, Senior AW, Kavukcuoglu K, Kohli P, Hassabis D. Highly accurate protein structure prediction with AlphaFold. *Nature*. 2021;596:583-589. doi: 10.1038/s41586-021-03819-2.

- Thompson LR, Sanders JG, McDonald D, Amir A, Ladau J, Locey KJ, Prill RJ, Tripathi A, Gibbons SM, Ackermann G, Navas-Molina JA, Janssen S, Kopylova E, Vázquez-Baeza Y, González A, Morton JT, Mirarab S, Xu ZZ, Jiang L, Haroon MF, Kanbar J, Zhu Q, Song SJ, Kosciolek T, The Earth Microbiome Project Consortium. *Nature*. 2017;551:457–463. doi: 10.1038/nature24621.

- Abraham, M. J., Murtola, T., Schulz, R., Páll, S., Smith, J. C., Hess, B., & Lindahl, E. (2015). GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX*, *1-2*, 19–25. https://doi.org/10.1016/j.softx.2015.06.001

- Morris GM, Huey R, Lindstrom W, Sanner MF, Belew RK, Goodsell DS, Olson AJ. AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *J Comput Chem*. 2009;30(16):2785-2791. doi: 10.1002/jcc.21256.

- A. Šali and T. L. Blundell. Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.* **234**, 779-815, 1993.