Breno Livio Silva de Almeida

View short CV ↗

Computational biology · Biological sequence representation · Machine learning

brenoslivio@pm.me São Carlos, Brazil +55 11 96839 2259

Website Google Scholar 0000-0002-2041-9170 GitHub LinkedIn

Research interests: Biological sequence representation and functional inference in microbial ecology, with a focus on handcrafted descriptors, foundation models and automated machine learning for sequence classification and annotation, complemented by metagenomic metadata curation.

Education

2024 – Present

PhD in Computer Science — University of São Paulo / Leipzig University
Double-degree programme

- Researching biological sequence representations through automated machine learning and foundation models, using orthologous groups from over 23,000 high-quality metagenome-assembled genomes to connect functional annotation with shared evolutionary ancestry. Work also includes metagenomic metadata curation and submission to public archives. Currently investigating whether sequence representations capture evidence of small-protein evolutionary origins.
- **Thesis:** Decoding the Language of Life: Towards the Limits of Biological Sequence Representation.
- **Supervisors:** André C. P. L. F. de Carvalho and Peter F. Stadler.

2020 – 2024

BSc in Computer Science, with specialisations in Data Engineering and Data Science — University of São Paulo
Five-year programme · GPA: 9.4/10 · Rank: 7/116
Research-intensive final year with 600 hours of thesis work. Accelerated completion through credit transfer from prior Computer Engineering coursework.

- **Thesis:** BioDeepFuse: A Hybrid Deep Learning Approach with Handcrafted Features for Non-coding RNA Classification. DOI.
  Research conducted with UFZ, supported by the São Paulo Research Foundation (FAPESP).
- **Supervisor:** André C. P. L. F. de Carvalho.

Employment History

07.2025 – 06.2026

Research Assistant — Helmholtz Centre for Environmental Research (UFZ).
Microbial Data Science group, Department of Computational Biology and Chemistry.

- Helped develop a machine-learning approach to classify data types across millions of runs from the Sequence Read Archive (SRA) within EReMiD.
- Contributed to metagenomic metadata standardisation in related projects and to the development of teaching materials.

07.2022 – 09.2022

Research Intern — Big Data Biology Lab, Fudan University.
Research internship in Prof. Luis Pedro Coelho's group, contributing to the Global Microbial Small ORFs Catalog (GMSC).

- Developed a rarefaction workflow for nearly one billion unique small open reading frames (smORFs) across 63,402 samples.
- Compared smORF richness and sampling saturation across habitats, identifying undersampled soil-associated diversity.
- Contributed to *A catalog of small proteins from the global microbiome* (*Nature Communications*, 2024). Lab blog post · Paper.

Research Funding and Fellowships

Awarded approximately **€110,181** in competitive fellowships, research scholarships, and mobility/travel grants across the award periods.

2025

Google PhD Fellowship in Health Research — USD 30,000. International fellowship recognising outstanding doctoral research in computing and related fields, with USD 15,000 annually for two years and mentorship from a Google researcher. Selected among 15 fellows in Latin America in 2025, including eight from Brazil. Recipients · USP award report.

2025

ISME Scholar Mobility Fund — EUR 2,300. Awarded by the International Society for Microbial Ecology for a research visit to UFZ, Leipzig, in July 2026. The programme supports international research and training visits by graduate students and postdoctoral researchers to advance microbial ecology and collaboration with host laboratories. Programme · USP visit report.

2024 **FAPESP Doctoral Scholarship** — BRL 453,168. *Decoding the language of life: Towards protein classification and design through large language models.* São Paulo Research Foundation (FAPESP); competitive, peer-reviewed; grant #2024/10958-1. FAPESP grant record.

2024 **FAPESP Technical Training Scholarship** — BRL 10,553.20. *Democratization of machine learning in life sciences and annotation of protein families.* Grant #2024/03304-5. FAPESP grant record.

2024 **Prototypes for Humanity Travel Grant** — USD 760. Travel support to present BioAutoML at Prototypes for Humanity in Dubai, November 2024. The project was selected for an international showcase of university innovations addressing social and environmental challenges, providing an opportunity to present accessible machine-learning tools for life-science research. Selected project.

2023 **FAPESP Scientific Initiation Scholarship** — BRL 7,920.99. *Machine learning applied to multiomics analysis for exploration of microbial communities.* Competitive, peer-reviewed; grant #2023/00264-0. FAPESP grant record.

2023 **CNPq Scientific Initiation Scholarship** — BRL 1,500. *Automated machine learning for life sciences.* National Council for Scientific and Technological Development (CNPq); grant #101243/2023-5. CNPq scholarship category.

2021 – 2022 **FAPESP Scientific Initiation Scholarship** — BRL 7,652.70. *Classification and analysis of bacterial non-coding RNA sequences using machine-learning techniques.* Competitive, peer-reviewed; grant #2021/08561-8. FAPESP grant record.

Indicative total at fixed reporting rates: BRL 1 = EUR 0.17; USD 1 = EUR 0.85. These are indicative conversions, not current market rates. Total excludes cash prizes, estimated training values, and consortium budgets; it represents award values, not payments received.

Honours and Awards

2026 **IRCAI Global Top 100 — Outstanding (2025 edition).** Team recognition for BioAutoML from the International Research Centre on Artificial Intelligence (IRCAI), under the auspices of UNESCO. Recognises accessible machine learning for life sciences supporting UN Sustainable Development Goals 3 (health) and 10 (reduced inequalities). Contributing developer of BioAutoML; currently lead developer of BioAutoML-FAST. Official recognition.

2026 **HIDA-sponsored Carpentries Instructor Training.**

Selected for one of 15 annually sponsored places for Helmholtz researchers; estimated training value: USD 1,600. The programme prepares researchers to teach practical computing and data-science workshops using evidence-based instruction, supporting reproducibility and open science. Participation leads towards Carpentries instructor certification. HIDA programme and selection details.

2024 **J. F. Marar Artificial Intelligence Prize for Undergraduate Research.** Winner for *BioDeepFuse: Empowering Researchers in Life Sciences with Deep Learning*; cash prize: **BRL 10,500**. Inaugural award for ICMC-USP undergraduate AI research (2023 edition, presented in February 2024), selected unanimously by the committee on academic relevance, social relevance, alignment with the UN Sustainable Development Goals, and the student's academic performance. BioDeepFuse combines deep learning and handcrafted sequence features to make non-coding RNA classification more accessible to life-science researchers. USP award announcement.

2024 **Prototypes for Humanity.** BioAutoML selected among 100 projects from nearly 3,000 submissions worldwide. Presented in Dubai in November 2024, showcasing automated machine learning that enables life-science researchers to analyse biological sequences with less programming expertise. Travel support is reported in Research Funding and Fellowships. Selected project.

Publications

127	78	4	3
Citations	Citations	h-index	i10-index
Google Scholar	Scopus	Google Scholar	Google Scholar

Google Scholar as of 2026-10-24 · Scopus as of 2026-10-24.

† Equal contribution.

Manuscripts Under Review

- **de Almeida, B. L.**, Bonidia, R. P., Bole, M., Avila-Santos, A. P., Stadler, P. F., Nunes da Rocha, U., & de Carvalho, A. C. P. L. F. (2026). BioAutoML-FAST: an automated machine-learning platform for reusable and benchmarked biological sequence models. *bioRxiv*. DOI. **Under revision at *Bioinformatics* following peer review (major revision).**

Published Journal Articles

- Tavares, D., Barreto, M. C. A., **de Almeida, B. L. S.**, Florentino, B., Tetzner, F., Goedert, G., Struchiner, C., Bonidia, R., & de Carvalho, A. (2026). Dominique: An AI-Powered Fact-Checking Chatbot for Democratizing Access to Reliable Information. Oxford Open Digital Health, oqag045. DOI.

- Duan, Y., Santos-Júnior, C. D., Schmidt, T. S., Fullam, A., **de Almeida, B. L.**, Zhu, C., ... & Coelho, L. P. (2024). A catalog of small proteins from the global microbiome. Nature Communications, 15(1), 7563. DOI.
- **de Almeida, B. L.**†, Avila Santos, A. P.†, Bonidia, R. P., Stadler, P. F., Stefanic, P., Mandic-Mulec, I., ... & de Carvalho, A. C. (2024). BioDeepfuse: a hybrid deep learning approach with integrated feature extraction techniques for enhanced non-coding RNA classification. RNA Biology, 21(1), 410-421. DOI.
- Bonidia, R. P., Avila Santos, A. P., **de Almeida, B. L.**, Stadler, P. F., Nunes da Rocha, U., Sanches, D. S., & De Carvalho, A. C. (2022). Information theory for biological sequence classification: A novel feature extraction technique based on Tsallis entropy. Entropy, 24(10), 1398. DOI.
- Bonidia, R. P., Santos, A. P. A., **de Almeida, B. L.**, Stadler, P. F., da Rocha, U. N., Sanches, D. S., & de Carvalho, A. C. (2022). BioAutoML: automated feature engineering and metalearning to predict noncoding RNAs in bacteria. Briefings in Bioinformatics, 23(4), bbac218. DOI.

Accepted Journal Articles

- **de Almeida, B. L. S.**†, Bonidia, R.†, Florentino, B.†, Barreto, M. C. A.†, Constantino, J. L. R.†, Tavares da Silva, D.†, Zwicker, G. G., Rocha, U., Goedert, G., Struchiner, C., & de Carvalho, A. AutoAI Hub: From data to society, democratizing machine learning for everyone. Oxford Open Digital Health. **Accepted for publication**.

Conference Proceedings

- **de Almeida, B. L. S.**†, Queiroz, A. P.†, Santos, A. P. A.†, Bonidia, R. P., da Rocha, U. N., Sanches, D. S., & de Carvalho, A. C. P. D. L. F. (2021, November). Feature importance analysis of non-coding DNA/RNA sequences based on machine learning approaches. In Brazilian Symposium on Bioinformatics (pp. 81-92). Cham: Springer International Publishing. DOI.

International Collaborations

Dates indicate my participation; budgets refer to projects or institutions, not individual awards.

2024 – Present **AutoAI-Pandemics.** AI4PEP collaboration across Brazil, Germany, Canada, and the UK developing accessible AI for epidemic preparedness and biological-data analysis. **CAD 362,500 awarded in 2023** for four years, supported by IDRC and UK International Development. I contribute to BioAutoML-FAST and publications on AutoAI Hub and Dominique. AI4PEP Brazil hub · Funding announcement.

2024 – Present **NFDI4Microbiota.** German consortium providing FAIR microbial data, analysis tools, and training. Initial DFG funding: **approximately EUR 3 million annually for five years from October 2021**, including **approximately EUR 1 million for UFZ** over that period. I contribute to microbial genome and metadata deposition and workshop teaching. Funding and UFZ role · DFG project record.

2025 – Present **AI MareExplore.** Helmholtz Innovation Pool project (2025–2027), led by GEOMAR with UFZ, Forschungszentrum Jülich, and GFZ, using AI to discover marine microbial enzymes for plastic degradation and carbon capture. I contribute protein-language-model approaches for microbial protein annotation. Funder, dates, and partners · Project announcement.

Research Software and Resources

BioAutoML-FAST **Automated machine learning for biological sequences.** Lead developer; expanded sequence types, prediction tasks, and evaluation within the BioAutoML framework. Code · Web application · Preprint. Provides reusable trained models for genomics, transcriptomics, and proteomics.

BioAutoML **Automated feature engineering and model selection.** Contributing developer. Automates sequence-feature extraction and selection, algorithm recommendation, and hyperparameter tuning for biological sequence classification, including bacterial non-coding RNAs. Code · Paper.

BioDeepFuse **Hybrid deep learning for non-coding RNA classification.** Co-first author of the methods paper. Combines convolutional or bidirectional LSTM networks with handcrafted sequence features to capture complementary sequence patterns. Includes training and evaluation scripts, datasets, and experimental results for bacterial RNA classification. Code · Paper.

MAG2ENA **Metagenome-assembled genome submission to the European Nucleotide Archive.** Lead developer; developed within NFDI4Microbiota. Provides web and command-line workflows for metadata curation, validation, and genome submission using the MIMAG checklist, plus metadata updates for existing ENA records. Code · Web application.

Invited Talks and Conference Presentations

Invited Talk

- **de Almeida, B. L.** Da bancada ao teclado: ferramentas gratuitas para a pesquisa científica [From bench to keyboard: free tools for scientific research]. **Invited talk**, 36th FisFar Summer Course, Federal University of Minas Gerais (UFMG), Brazil. January 2027. *Scheduled.*

Oral Presentations and Flash Talks

- **de Almeida, B. L.**, Kamath, S., Santos, A. P. A., Bonidia, R. P., Espíndola-Hernández, P., Stadler, P. F., Wierckx, N., Borchert, E., Rother, D., Oliveira, A., de Carvalho, A. C. P. L. F., & Rocha, U. N. Combining protein language models with different architectures expands protein family annotation 14.2-fold across 16 million open reading frames from 3,802 environmental metagenomes. **Oral presentation**, 20th International Symposium on Microbial Ecology (ISME20), Auckland, New Zealand, 18 August 2026 (MS13, Computational Approaches in Microbial Ecology). Conference programme.
- **de Almeida, B. L. S.**, Kamath, S., Santos, A. P. A., Bonidia, R. P., Espíndola-Hernández, P., Stadler, P. F., Wierckx, N., Borchert, E., Rother, D., Oliveira, A., de Carvalho, A. C. P. L. F., & Rocha, U. N. Different protein language models uncover 23.21% of the functional dark matter across 16 million open reading frames recovered from thousands of environmental metagenomes. **Oral presentation**, VAAM Annual Conference, Berlin, Germany, 22–25 March 2026 (SL-CM-01, Computational Microbiology). Abstract book.
- **de Almeida, B. L.**, Kamath, S., Santos, A. P. A., Bonidia, R. P., Espíndola-Hernández, P., Stadler, P. F., Oliveira, A., de Carvalho, A. C. P. L. F., & Rocha, U. N. Exploration of multiple protein language models and over 16 million open reading frames recovered from thousands of terrestrial metagenomes uncover protein domains and families in 13,486 previously unknown orthologs. In: 17th Symposium on Genetics and Bacterial Ecology (BAGECO 2025), 2025, Graz, Austria. **Flash talk**, 4 July 2025 (FT043). Abstract.
- **de Almeida, B. L. S.**, Kamath, S., Santos, A. P. A., Bonidia, R. P., Espíndola-Hernández, P., Stadler, P. F., Oliveira, A., de Carvalho, A. C. P. L. F., & Rocha, U. N. Unveiling protein domains and families in 183,044 novel orthologs exploring multiple protein language models and over 16 million open reading frames recovered from thousands of terrestrial metagenomes. **Oral presentation**, Helmholtz AI Conference 2025, Messe Karlsruhe, Rheinstetten, Germany, 5 June 2025. Conference archive

Conference Abstracts and Posters

- **de Almeida, B. L.**, Bonidia, R. P., Bole, M., Santos, A. P. A., Stadler, P. F., Rocha, U. N., & de Carvalho, A. C. P. L. F. From sequences to traits: An AI-driven AutoML framework for genome-resolved microbial ecology. In: 20th International Symposium on Microbial Ecology (ISME20), 2026, Auckland, New Zealand.
- Bonidia, R. P., Florentino, B. R., Barreto, M. C. A., **de Almeida, B. L.**, Constantino, J. L. R., Silva, D. T., Rocha, U. N., Goedert, G. T., Struchiner, C. J., & de Carvalho, A. C. P. L. F. AutoAI Hub: From Data to Society, Democratizing Machine Learning for Everyone. In: Human Genome Meeting 2026, 2026, Athens, Greece.
- Santos, A. P. A., **de Almeida, B. L.**, Bole, M., Kamath, S., & Rocha, U. N. A FAIR-by-Design and Standard-Compliant System for Automated Submission of Microbial Genomes to the European Nucleotide Archive. In: VAAM Annual Conference, 2026, Berlin, Germany.
- Bonidia, R. P., Santos, A. P. A., **de Almeida, B. L.**, Stadler, P. F., Rocha, U. N., Sanches, D. S., & de Carvalho, A. C. P. L. F. BioAutoML: End-to-End Machine Learning Package for Life Sciences. In: 10th FEMS Congress of European Microbiologists, 2023, Hamburg, Germany.

Additional conference contributions and abstracts.

Teaching and Mentoring

2025 – 2026 **Instructor — UFZ Summer School on Trends in Multi-Omics Data Analysis for Microbial Ecology and Biotechnology**, Leipzig.

Taught Unix/Bash/Git, machine learning, and sequence annotation to approximately 30 participants across two editions, delivering approximately 20 hours of instruction. Teaching materials (GitHub).

09.2025 **Instructor — NFDI4Microbiota workshop, 13th International Symposium on Anaerobic Microbiology (ISAM)**, Leipzig. Taught submission of research data and metadata to public repositories to approximately 20 students.

06.2025 **Co-organiser and instructor — Metagenomics workshop, 17th Symposium on Bacterial Genetics and Ecology (BAGECO)**, Graz.

Taught experimental design, computer experiments, and end-to-end sequence analysis with BioAutoML to approximately 20 students. Workshop programme.

05.2021 – 11.2022 **Teaching Assistant — University of São Paulo.**
Eight hours/week across Object-Oriented Analysis and Design, Statistics, Fundamentals of Data Science, and Computational Visualization, serving approximately 150 students in total. Graded assignments, prepared exams, and delivered practical lectures.

Selected Professional Development

2025 **Summer School in Trends in Multi-omics Data Analysis for Microbial Ecology and Biotechnology**, UFZ, Leipzig.
2024 **15th Summer Course in Bioinformatics**, FMUSP, São Paulo.
2020 – 2021 Data science specialisations: University of Michigan (*Applied Data Science with Python*) and Johns Hopkins University (*Data Science*), Coursera.
2020 *Nutrition and Health: Human Microbiome*, Wageningen University, edX.

Academic Service

Journal Peer Review

2026 – Present **Computational and Structural Biotechnology Journal.** One manuscript reviewed.
2026 – Present **Journal of Thermal Biology.** One manuscript reviewed.
2025 – Present **Briefings in Bioinformatics.** One manuscript reviewed across two rounds.

Research Symposium Evaluation

2026 **Evaluator — 34th SIICUSP**, University of São Paulo. Simpósio Internacional de Iniciação Científica e Tecnológica da Universidade de São Paulo.

Thesis Examination Committees

Member of six undergraduate thesis examination committees.

Community and Student Leadership

2018 – 2021 **Co-founder and coordinator — efestus**, University of São Paulo. Student organisation addressing computing, data science, and technology ethics for social good. Funded participation in the TechShift Summit, Chapel Hill, USA.
2017 – 2019 **Software division member — Warthog Robotics**, University of São Paulo. Participated in robotics competitions and taught Git and object-oriented programming to new members.

Technical Skills

Methods	Biological sequence representation; handcrafted descriptors and learned representations from foundation models; automated machine learning; functional annotation; metagenomic sequence analysis.
Programming	Python, R, Bash, Nix, SQL, Java, C, C++.
Modelling and data	Scikit-learn, TensorFlow, pandas, Polars, NumPy, SciPy, NetworkX.
Sequence analysis	Biopython, BLAST, Prodigal, OrthoFinder, Anvi'o, HMMER, Infernal, DIAMOND, MMSeqs2.
Research tools	Git, Linux, Jupyter, R Markdown, LaTeX, Typst, Matplotlib, Plotly, Streamlit.
Languages	Portuguese (native), English (C1), German (A1.2).

Professional Memberships

International Society for Microbial Ecology (ISME), since October 2025; Federation of European Microbiological Societies (FEMS), since November 2025.