Gonzalo Benegas

New York, NY • gsbenegas@gmail.com • gonzalobenegas.github.io • Google Scholar • gonzalobenegas

Summary

ML researcher with a PhD in Computational Biology. Deep expertise in genomic language models, broadly interested in applied ML research.

Experience

Open Athena, Research Scientist – New York, NY 2025 – present

- Leading an open research program on genomic language models.
- Developing scalable data curation strategies and training autoregressive transformers, leveraging NLP infrastructure and scaling recipes.

University of California, Berkeley, Postdoctoral Researcher, EECS – Berkeley, CA 2024 – 2025

- Co-developed GPN-Star, a phylogeny-informed genomic language model achieving SOTA variant effect prediction across both rare and common variants in the human genome (bioRxiv, 2025).
- Co-authored a review on genomic language models, featured on the cover of Trends in Genetics (2025).
- Created TraitGym, a benchmark for evaluating DNA sequence models on variant effect prediction (bioRxiv, 2025). Used as evaluation benchmark by DeepMind's AlphaGenome (Nature, 2026).

University of California, Berkeley, Graduate Student Researcher – Berkeley, CA 2018 – 2023

- Developed GPN-MSA, the first alignment-based genomic language model, achieving SOTA variant effect prediction for rare variants and scalable inference across all 9B possible SNVs in the human genome (Nature Biotechnology, 2025).
- Developed GPN, the first genomic language model to exhibit zero-shot variant effect prediction capabilities across the entire genome (PNAS, 2023).
- Developed scQuint, a variational auto-encoder for analysis of alternative splicing in single-cell RNA-seq (eLife, 2022).

Autodesk, Software Engineer – Buenos Aires, Argentina 2015 – 2016

Synthicity, Software Engineer – Buenos Aires, Argentina 2014 – 2015

Education

PhD **University of California, Berkeley**, PhD in Computational Biology 2023
- Advisor: Yun S. Song

BS+MS **University of Buenos Aires**, BS+MS in Computer Science 2018
- GPA: 9.70/10

Selected Publications

bioRxiv (Co-first author) 2025
Predicting functional constraints across evolutionary timescales with phylogeny-informed genomic language models

Nature Biotechnology (First author) 2025
A DNA language model based on multispecies alignment predicts the effects of genome-wide variants

Trends in Genetics (Co-first author) 2025
Genomic Language Models: Opportunities and Challenges

bioRxiv (First author) 2025
Benchmarking DNA Sequence Models for Causal Regulatory Variant Prediction in Human Genetics

Proceedings of the National Academy of Sciences (First author) 2023
DNA language models are powerful predictors of genome-wide variant effects

eLife (First author) 2022
Robust and annotation-free analysis of alternative splicing across diverse cell types in mice