Alejandro Buendia

https://abuendia.github.io/

Email: abuen@stanford.edu
Mobile: 1-718-687-3030

Research Interests

Single-cell sequencing foundation models, large language models, viral evolution and immune escape, computational genomics, clinical natural language processing.

Education

Massachusetts Institute of Technology — Cambridge, MA
Advanced Study Program — *Sept. 2021 - Dec. 2021*

Coursework: Advanced Computational Biology: Genomes, Networks, Evolution

Columbia University — New York, NY
B.A. in Computer Science & Mathematics [Double Major] — *Feb. 2019*
GPA: 3.9/4.0 (Cum Laude)

Selected Coursework: Machine Learning, Natural Language Processing, Intro to Computational Complexity, Intro to Computational Learning Theory, Analysis of Algorithms I, Linear Algebra, Intro to Modern Analysis I & II, Senior Projects in Computer Science

Honors: Dean's List, Summer Funding Program Recipient, Kluge Scholar

Experience

Columbia University Department of Systems Biology — New York, NY
Computational Research Lab Manager — *Sept. 2023 – July 2024*
Program for Mathematical Genomics [Advisor: Raul Rabadan]

- Research co-lead for foundation model of cellular transcription trained on ATAC-seq data, with focus on application to cancer-specific cell types.
- Research lead and co-authored NSF grant for modeling evolution and immune escape of influenza and SARS-CoV-2 viral protein sequences by integrating protein language models.
- Developed website for visualizing results of a CRISPR genetic interaction screen of the DNA damage response.
- Managed data analyst and provided engineering and infrastructure support to graduate students and postdoctoral researchers.

Massachusetts Institute of Technology CSAIL — Cambridge, MA
Research Engineer / Junior Research Scientist — *Feb. 2022 – June 2023*
Clinical Machine Learning Group [Advisor: David Sontag]

- Research lead for autoregressive language modeling over longitudinal healthcare data with downstream clinical predictive applications.
- Assisted research for language modeling over single-cell RNA-seq data for learning representations of patients.
- Assisted research for few-shot tabular classification using large language models.
- Development lead for open-source package omop-learn for predictive machine learning modeling over longitudinal healthcare data.

Microsoft — Cambridge, MA
Data Scientist II — *July 2020 – Jan. 2022*
Azure Global Commercial Industry AI

- Member of core team developing a natural language processing platform for news ingestion, entity extraction, summarization, question answering, and text classification on financial documents.
- Owned entity extraction feature end-to-end and pretrained domain-specific language model backing the product.

Software Engineer — *Jan. 2019 – June 2020*
AI Development Acceleration Program

- Developed gradient-boosted trees model for intelligent command prediction within Microsoft Word, Excel, and PowerPoint.
- Developed inference infrastructure for personalized language model for sentence autocomplete in Microsoft Outlook.
- Developed multi-object tracking computer vision model for soccer player tracking during customer engagement.

Columbia University New York, NY
Undergraduate Researcher *May 2018 – Dec. 2018*
Data Science Institute [Advisor: Daniel Hsu]

- Proposed statistical test for Gaussian mixture detection with hypothesized sample complexity bound lower than state-of-the-art methods.
- Ran empirical evaluations of proposed test to compare against theoretical complexity.

Undergraduate Researcher *May 2017 – Aug. 2017*
Mathematics Department REU [Advisor: Daniel Litt]

- Studied representations of Lie algebras, using applications of Ramsey theory in graphs.
- Built an open-source Python/MATLAB library for symbolic computations involving Lie algebras.

University of Minnesota Minneapolis, MN
Undergraduate Researcher *June 2016 – Aug. 2016*
Computer Science & Engineering Department REU [Advisor: Daniel Boley]

- Optimized time complexity of an algorithm to compute hitting times of random walks over graphs.
- Applied algorithm to efficient node centrality measure computation on graphs.
- Talks presented jointly with Daniel Boley at 2016 REU Symposium, Householder Symposium XX, 2017 AAMAS International Workshop on Trust in Agent Societies, and 2019 IJCAI International Workshop on Big Social Media Data Management and Analysis.

Publications

Xi Fu*, Shentong Mo*, **Alejandro Buendia***, Anouchka Laurent, Anqi Shao, Maria del Mar Alvarez-Torres, Tianji Yu, Jimin Tan, Jiayu Su, Romella Sagatelian, Adolfo A. Ferrando, Alberto Ciccia, Yanyan Lan, David M. Owens, Teresa Palomero, Eric P. Xing, Raul Rabadan. GET: A foundation model of transcription across human cell types. *bioRxiv 2023.09.24.559168*, July 2024.

Rebecca Boiarsky, Nalini Singh, **Alejandro Buendia**, Gad Getz, David Sontag. A Deep Dive into Single-Cell RNA Sequencing Foundation Models. In *Machine Learning in Computational Biology* (MLCB 2023), Nov. 2023. (Oral Presentation, Top 15%)

Stefan Hegselmann, **Alejandro Buendia**, Hunter Lang, Monica Agrawal, Xiaoyi Jiang, and David Sontag. TabLLM: Few-shot classification of tabular data with large language models. In *26th International Conference on Artificial Intelligence and Statistics* (AISTATS 2023), April 2023.

Liqun Shao, Sahitya Mantravadi, Tom Manzini, **Alejandro Buendia**, Manon Knoertzer, Soundar Srinivasan, and Chris Quirk. Examination and extension of strategies for improving personalized language modeling via interpolation. In *First Workshop on Natural Language Interfaces* (ACL 2020), July 2020.

Alejandro Buendia and Daniel Boley. Random walk fundamental tensor and graph importance measures. In *International Workshop on Big Social Media Data Management and Analysis* (IJCAI 2019), Aug. 2019.

Daniel Boley, **Alejandro Buendia**, and Golshan Golnari. Random walk Laplacian and network centrality measures. *arXiv 1808.02912*, Aug. 2018.

Alejandro Buendia and Daniel Boley. Optimized graph-based trust mechanisms using hitting times. In *International Workshop on Trust in Agent Societies* (AAMAS 2017), May 2017.

Teaching & Service

Massachusetts Institute of Technology
Clinical Machine Learning Group

- With postdoctoral fellow, mentored one master's thesis and two undergraduates in clinical natural language processing research and product deployment.

Microsoft
TEALS Outreach Program

- Led team of six employees to develop a three-module introductory course on AI for high school students.
- Presented talk about working at Microsoft to disadvantaged high school students in Boston area.

Diversity and Early-in-Career Recruiting

- Founded and co-led Diversity & Inclusion Initiative for the Artificial Intelligence Development Program to recruit underrepresented minority graduates, targeting academic and job search conference recruitment.
- Partnered with Black in AI, LatinX in AI, and Women in ML at AI conferences for student recruitment.

GLEAM New England

- Board member for Microsoft's LGBTQ+ employee group, with focus on recruiting LGBTQ+ students and organizing employee morale events.

Columbia University
Computer Science & Mathematics Departments

- Teaching assistant for Discrete Mathematics and Calculus IV [4.8/5.0 average rating].

References

Raul Rabadan
Professor of Systems Biology and Biomedical Informatics
Columbia University
rr2579@cumc.columbia.edu
212-851-5141

Daniel Hsu
Associate Professor of Computer Science
Columbia University
djhsu@cs.columbia.edu
212-853-8473

Said Bleik
Principal Applied Machine Learning Scientist Manager
Microsoft
bleik@microsoft.com
781-539-3032