Samir Char

sc.samirchar@gmail.com | samirchar.com | linkedin.com/in/samir-char | github.com/samirchar
Boston, Massachusetts | 646-262-8466

Research Interests

Self-supervised & multimodal learning; Artificial Intelligence (AI) for biomedicine and science.

Education

Columbia University New York, NY, US
Master of Science in Data Science (GPA: 4.08 / 4.0) Jan. 2021 – May 2022

- Relevant Courses: Machine Learning (ML), Deep Learning, Advanced Deep Learning, Natural Language Processing

Universidad de Los Andes Bogotá, Bogotá D.C., CO
Bachelor of Science in Electrical Engineering (GPA: 4.1 / 5.0) Aug. 2013 – Dec. 2018

- Relevant Courses: Intelligent Analysis of Signals and Systems, Signals, Optimization

Publications

Equal contribution indicated with *

Peer-reviewed

1. **Char, S**., Corley, N., Alamdari, S., Yang, K.K., Amini, A.P. (2025). ProtNote: A Multimodal Method for Protein-Function Annotation. *Bioinformatics*. DOI: doi.org/10.1093/bioinformatics/btaf170

Under Review

1. Yang, K.K.*, Alamdari, S.*, Lee, A.J.*, Kaymak-Loveless, K., **Char, S**., Brixi, G., Domingo-Enrich, C., Wang, C., Lyu, S., Fusi, N., Tenenholtz, N., Amini, A.P. (2025). The Dayhoff Atlas: Scaling Sequence Diversity Improves Protein Design. bioRxiv. DOI: doi.org/10.1101/2025.07.21.665991. In consideration at *Cell*
2. **Char, S**., Bruce, S., Tiao, Jonathan., Shah, N., Adeuyan, O., Williams, O., Elhadad, N., Noble, J. (2025). Smartphone Artificial Intelligence For Face-Arm-Speech-Time (FAST) Stroke Detection. Under review at *Neurology: Clinical Practice*

Research Experience

Microsoft Research, BiomedicalML Lab (Nicolo Fusi's Group) Boston, MA, US
Independent Researcher Jul. 2023 – Mar. 2025 (~15 hours/week)

- Led research and development of a multimodal protein function prediction model and contributed to protein generative model research. Used Distributed Data Parallel (DDP) with Transformers and Convolutional Neural Nets (CNNs). Advisors: Dr. Kevin K. Yang and Dr. Ava Amini. Outcomes: 1 *Bioinformatics* publication; 1 preprint under consideration at *cell*.

Columbia University, Data Science Institute New York, NY, US
Data Science Institute Scholar Jan. 2022 – Jul. 2022

- Selected for the competitive Data Science Institute Scholars program to advance research at the intersection of ML and clinical neurology. Advisors: Dr. James Noble and Dr. Olajide Williams.
- Spearheaded a smartphone-based multimodal (video, speech, tabular) stroke prediction model using deep learning (CNN, ML). Achieved 89% sensitivity and 58% specificity. Outcomes: Columbia's BiomedX competition finalist; paper under review at *Neurology Clinical Practice*.

Universidad de Los Andes, Department of Electrical Engineering Bogotá, Bogotá D.C., CO
Undergraduate Researcher Jun. 2017 – Dec. 2017

- Thesis: Pioneered an ML ensemble of Extreme gradient boosting (XGBoost) and Random Forest to predict patients' hospital length of stay at admission, improving how low-income hospitals manage workforce, facilities, and resources. Algorithm surpassed physicians' accuracy by 15% (absolute). Advisor: Dr. Luis Felipe Giraldo. [Thesis PDF]

Industry Experience

Microsoft, Azure Core Remote – Boston, MA, US
Applied Scientist 2 Jun. 2022 – Present

- Created a hierarchical forecast and Monte Carlo simulation to detect customer capacity risks over a six-month horizon. Tool deployed across 65 regions, identifying 400K+ vCore gaps that drove procurement of new servers to mitigate global risk.
- Built an ML model to predict the additional servers required in a data center to avoid capacity-related failures.
- Applied Double ML to estimate the causal effect of a customer management program. Found that managed customers have ~20% less quota backlog and ~0.3% lower capacity failure rate.

- Developed a recommendation system to prioritize global quota requests, enabling triage of over 2M vCore requests to date.
- Improved XGBoost model to forecast Azure global quota requests, helping allocate capacity across Data Centers globally.
- Developed a forecasting ensemble (ARIMA, ETS, regression) for Azure support tickets, reducing Mean Absolute Percentage Error (MAPE) by 20% (absolute) from baseline. Used for estimating annual staffing and budget (~$40M/year).

Microsoft, Azure Core Remote – New York, NY, US
Applied Scientist Intern Jul. 2021 – Sep. 2021

- Improved Azure quota forecasting using ML. Used clustering to identify workload and churn patterns of Azure Data Explorer.

Mercado Libre Bogotá, Bogotá D.C., CO
Applied Scientist Jun. 2020 – Dec. 2020

- Improved gradient-boosting demand forecasting across multiple warehouses, increasing prediction accuracy by 5%.

Grupodot Bogotá, Bogotá D.C., CO
Applied Scientist Oct. 2018 – Jun. 2020

- Built a Long Short-Term Memory (LSTM) model forecasting fuel sales for 200+ gas stations of a major energy company, achieving an 11% MAPE improvement (absolute) over the client's model.
- Created an XGBoost churn model for Latin America's largest telecom; attained a 5.2 uplift on 200K monthly customers.
- Developed a credit default prediction model for a major Colombian bank to secure resources in case of client default.

Young and Rubicam Bogotá, Bogotá D.C., CO
Junior Applied Scientist Jun. 2018 – Oct. 2018

- Automated an ML tool to measure brands' social media success using Python, reducing manual tasks by 85%.

International Business Machines (IBM) Bogotá, Bogotá D.C., CO
Intern Jan. 2018 – Jun. 2018

- Designed and delivered ML workshops (150+ attendees). Advised startups on building ML solutions with IBM Cloud.

Teaching Experience

Grupodot Santiago, Santiago Metropolitan Region, CL
Instructor Aug. 2019

- Designed and delivered a week-long ML & Google Cloud workshop for 50 engineers.

Universidad de Los Andes, Department of Electrical Engineering Bogotá, Bogotá D.C., CO
Teaching Assistant, Control Systems Jan. 2017 – Jun. 2017

- Delivered lectures and led review sessions for a class of 60 students. Wrote, administered, and graded exams.

Talks & Presentations

- **Invited Talk** – Universidad de Los Andes, *ProtNote: A Multimodal Method for Protein-Function Annotation.* Jun. 2025
- **Invited Talk** – Universidad de Los Andes, *ProtNote: A Multimodal Method for Protein-Function Annotation.* Dec.2024
- **Lecturer** – Microsoft Azure Learning Academy, *Correlation; Casualty; Quantifying and Visualizing Patterns.* Jun. 2024
- **Student Lecturer** – Columbia University (Deep Learning Seminar), self-supervised learning & object detection. Mar. 2022

Leadership, Outreach & Volunteering

- **Microsoft Hispanic & Latinx Group** – organize and speak at events; review scholarship applications. Jan. 2024 – Present
- **Pro-bono Mentor** – provide guidance to students and early-career professionals pursuing ML careers. Jun 2021–Present
- **Microsoft** – mentor early-career employees on technical development and career progression. Jun. 2024 – Present
- **Microsoft Give Campaign** – launched initiative pairing 1:1 mentorship session with donations; mentored 10 individuals and raised $500 for ABACO, an organization fighting hunger and food waste in Colombia. Oct. 2025
- **Columbia University** – conducted mock technical and behavioral interviews for M.S. in Data Science students. May 2025
- **Team for Kids Foundation** – ran the Brooklyn Half Marathon, raising $1010 to combat childhood obesity. May 2025
- **NMDP Foundation** – assembled bone-marrow donor kits to support transplant matching. Jan. 2025
- **National Federation of the Blind** – helped create 1,000 sensory kits for visually-impaired children. May 2024

Skills

- **ML & AI**: Deep Learning, Machine Learning, Time Series Forecasting
- **Frameworks & Libraries**: PyTorch, XGBoost, Hugging Face, PyTorch Lightning, scikit-learn, NumPy, Pandas
- **Distributed Training**: Distributed Data Parallel (DDP), Fully Sharded Data Parallel
- **Programming, Systems:** Python, SQL, R, UNIX, Linux, JavaScript

- **ML Ops & Data Engineering**: Weights & Biases, Apache Spark (PySpark)
- **Cloud**: Microsoft Azure, Amazon Web Services
- **Data Visualization**: Matplotlib, Seaborn, Power BI
- **Databases**: Azure Data Explorer, MySQL, BigQuery
- **Languages**: English (fluent), Spanish (native)

Personal Interests

- Running (completed four six marathons), weightlifting, martial arts (karate black belt), books, fine dining