

Alessa Carbo | acarbol1@jh.edu

EDUCATION

JOHNS HOPKINS UNIVERSITY, Baltimore MD

Bachelor of science in Computer Science

Start: 2023|Expected Graduation: 2027

PUBLICATIONS

A. Carbo and E. Nalisnick. Improving Handshape Representations for Sign Language Processing: A Graph Neural Network Approach. In *Proceedings of Empirical Methods in Natural Language Processing (EMNLP), Main Conference, 2025*.

A Carbo and E. Nalisnick Which Sparse Code? Identifiability Failures in SAE Inference. *ICLR 2026 Sci4DL workshop*

V. Javorek, T. Železný, **A. Carbo**, M. Hruš, I. Gruber. Beyond a Single Reference: Training and Evaluation with Paraphrases in Sign Language Translation.* Under review for ACL 2026

RESEARCH EXPERIENCE

On the linear identifiability of Sparse Autoencoder decompositions

(December 2025-January 2026) Advisor: Eric Nalisnick, Johns Hopkins University

-Demonstrated that sparse autoencoder inference is underdetermined, with alternative algorithms finding equally valid but causally distinct feature decompositions

Understanding and Predicting Latent Finetuning Behavior via the Neural Tangent Kernel

(August 2025-Present). Advisor: Eric Nalisnick, Johns Hopkins University

--Investigating emergent misalignment (cross-domain behavioral drift after finetuning) in large language models
-Applying NTK theory to characterize and develop predictive and preventive tools when finetuning operates in the “lazy” regime and how this relates alignment brittleness

Systematic Discovery and Mechanistic Interpretability of Gradient-Opposed Fine-Tuning Dynamics

(February 2025-Present), Advisor: Eric Nalisnick, Johns Hopkins University

-Developing an evolutionary framework to identify task pairs with maximal gradient conflict in fine-tuning
- Using sparse autoencoders to understand how gradient opposition alters internal representations and capabilities

Mechanistic Interpretability and Steering of Persona Prompting and Role-Playing LLMs, Brno University of Technology (June 2025-August 2025) Advisor: Ricard Maxer, Université de Toulon

-Worked on data attribution in the context of Sparse Autoencoders
-Used activation steering for role-playing LLM dynamics and persona control

Improving Handshape Representations for Sign Language Processing: A Graph Neural Network Approach (September 2024-January 2025), Advisor: Eric Nalisnick, Johns Hopkins University

-Developed novel dual graph neural network architecture for sign language handshape recognition, establishing the first benchmark for structured handshape recognition in continuous signing sequences.

-Created anatomically-informed graph edge configuration and incorporated biomechanical metrics for improved handshape analysis.

JSALT (Frederick Jelinek Memorial Summer Workshop) Junior Researcher, Johns Hopkins University (June-August 2024), Advisor: Marek Hruz, University of West Bohemia

- Intensive 8-week workshop in language and speech technologies.
- Worked on a LLaMa-based system for Sign Language Translation, utilizing sign language instruction tuning inspired by a LLaVa approach in processing visual data through an LLM.

PRESENTATIONS

- Moderator:** "ML in food Systems Panel", Food4Thought Conference, Stanford University (April 2026)
- Interviewer:** "A Conversation with Bruce Friedrich, founder and president of the Good Food Institute"
- **Panelist**, "Alternative Protein Educational Initiatives Panel," Good Food Institute (October 2025)
- Guest Lecturer**, "Introduction to Mechanistic Interpretability," JSALT 2025 Human Language Technology Summer School , Brno University of Technology (June 2025)
- **Moderator**, "Machine Learning & AI Approaches in Alternative Proteins Panel," Johns Hopkins University (April 2025)
- "Improving Handshape Representations for Sign Language Processing: A Graph Neural Network Approach," MATRX Conference, Johns Hopkins University (March 2025)
- "A Game Theory Analysis of The Frog and Scorpion Fable," MATRX Conference, Johns Hopkins University (March 2024)

TEACHING EXPERIENCE

Future Food Manufacturing, Johns Hopkins University (Fall 2024, Fall 2025)

- Served as the sole teaching assistant and coordinator for the course.
- Delivered lectures to the students on alternative protein topics, and coordinated all guest lecturers.

LEADERSHIP & ACTIVITIES

Alternative Protein Project: President (May 2025- Present), Director of Education(January 2024 - May 2025), Johns Hopkins University

-Lead all organization initiatives, such as our course, organizing events with industry guests, connecting with faculty in relevant research areas, organizing panels and seminars.

Machine Learning Projects Lead, Artificial Intelligence Society, Johns Hopkins University (October 2023 - May 2024)

- Curator of machine learning reading guides and leading members for data science and machine learning development in projects. Delivered learning workshops (linear regression, OpenAI API, etc).

PROJECTS

The Bayesian (June 2023 - Present)

- Personal Substack featuring writings on diverse topics such as The Free Energy Principle, Machine Learning walkthroughs, Bayesian inference, and agricultural technology.

TECHNICAL SKILLS

- Programming Languages: Python, Java, Bash, and C

-Machine Learning: Pytorch, huggingface, LLMs, multimodal systems, vision transformer, mediapipe, graph neural networks, sparse autoencoders, transformerlens.

AWARDS

Pistritto Fellowship

Research award of 5k from the Johns Hopkins computer science department given to 1-3 students per year.