

# Jakob Hansen

✉ [jakob@jakobhansen.org](mailto:jakob@jakobhansen.org) 🌐 [jakobhansen.org](https://jakobhansen.org) 🎧 [hansenjakob](https://hansenjakob.com) 📺 [jakobkhansen](https://www.youtube.com/channel/UCv33333333333333333333)

---

## ABOUT ME

I've spent the last five years working on mathematical and computational tools to understand AI models better. Most recently, I've trained cross-layer transcoders and sparse autoencoders for open source models, and built data-centric interpretability tools to audit benchmarks and RL environments, finding broken tasks and environment flaws. I'm interested in deepening our understanding of AI capabilities and behavior, including through interpretability and evaluations.

## EXPERIENCE

### BluelightAI | Head of Data Science (founding team)

October 2021–present

BluelightAI was founded to apply topological data analysis to AI, with a focus on interpretability. As the AI landscape has changed, the company and my role have evolved from work on an early set of academic research contracts to tools for exploratory analysis of ML models to a focus on applied mechanistic interpretability for LLMs.

- Developed an agent-based framework for data-centric mechanistic interpretability, using custom-trained sparse autoencoders to featurize data and identify large-scale patterns. Applied this framework to analyze evaluation and training datasets, identifying issues in RL environments and benchmark harnesses, including broken tasks in Olmo 3 and MiMo-v2.6 RL environments and missing harness sandboxing in BigFinanceBench.
- Led training and evaluation of cross-layer transcoders for Qwen 3 models (0.6B and 1.7B), replicating and extending Anthropic's original methodology. Implemented and optimized tensor-parallel distributed training infrastructure, reducing memory consumption by over 50% and accelerating training by a similar amount compared with a naive implementation. Developed and tested other improvements to the training process, including factored decoder structures, loss term scheduling, and precision handling. Used topological data analysis methods for structural analysis of learned features. See [qwen3.bluelightai.com](https://qwen3.bluelightai.com).
- Designed and led implementation of a [Python package](#) that uses topological data analysis tools for data exploration and ML model evaluation. Designed, implemented and optimized algorithms for graph transformations and analysis (e.g. hierarchical clustering, merge trees). Refined algorithms for interactive graph visualization and layout.
- Led a team of data scientists and engineers responsible for customer-driven research and product prototyping. Collaborated with a full-stack engineering team to specify, implement, and test the functionality of data analysis and visualization products.
- Developed PyTorch libraries implementing novel types of neural network layers (e.g. convolutional layers with kernels parameterized by coordinates on the Klein bottle or other spaces) for internal research projects.
- Developed novel machine learning models for biological data, including single-cell RNA-seq and spatial transcriptomic data.

### Ohio State University | Visiting Assistant Professor

August 2020–October 2021

Taught linear algebra to STEM undergraduates, focusing on empowering them to use mathematical tools in applications. Continued research on applications of cellular sheaves to problems in machine learning, data analysis, and network science.

### University of Pennsylvania | Graduate Research Associate

September 2016–July 2020

Studied cellular sheaves, a discrete algebraic structure that represents systems of constraints for data attached to a graph or cell complex. Developed the theory of sheaf Laplacians, which opened the door to spectral and approximate algorithms for cellular sheaves.

## SELECTED PUBLICATIONS

- Sedona Ewbank and J Hansen. [Why do Irrelevant Instructions Inhibit Refusal in LLMs?](#) ICML Workshop on Mechanistic Interpretability, 2026.  
Investigated the mechanistic basis for a surprising phenomenon: prefixing harmful requests with lists of irrelevant instructions led to increased compliance across a number of model families and size.

- Thomas Gebhart, J Hansen, and Paul Schrater. [Knowledge Sheaves: A Sheaf-Theoretic Framework for Knowledge Graph Embedding](#). AISTATS 2023.  
Formalizes the problem of knowledge graph embedding in algebro-topological terms and uses this framework to derive new architectures and loss functions.
- J Hansen and Robert Ghrist. [Opinion Dynamics on Discourse Sheaves](#). SIAM Journal on Applied Mathematics 81 (5), 2021.  
An exploration of various dynamical systems that can be associated with (the Laplacian of) a cellular sheaf, with a central motivating problem of modeling opinion dynamics in social networks.
- J Hansen and Thomas Gebhart. [Sheaf Neural Networks](#). NeurIPS Workshop: Topological Data Analysis and Beyond, 2020.  
Uses sheaf Laplacians to increase the expressive power of graph neural networks. This was a decent initial pass, but I'm most proud of this work because it inspired others to do even more (e.g. [Neural Sheaf Diffusion](#)).
- J Hansen and Robert Ghrist. [Toward a Spectral Theory of Cellular Sheaves](#). Journal of Applied and Computational Topology 3 (4), 2019.  
Introduces sheaf Laplacians as a way to add flexibility to the rigid algebraic structure of cellular sheaves. Sprouted a growing field of research and has been cited over 300 times.

## RECENT WORK & WRITING

- [Are Neural Networks Clean?](#): a response to Nick Cammarata's [Neural Networks are Already Clean](#), arguing that general neuron-level monosemanticity won't solve interpretability by itself
- [Covariance pooling works because it's nonlinear \(at least sometimes\)](#): testing a new probing method for LLMs
- [When you shouldn't trust einsum](#): debugging a memory blowup in CLT training
- [Something perplexing about UMAP](#): a subtle problem with the dimensionality reduction algorithm keeps it from handling tight clusters well

## EDUCATION

**University of Pennsylvania** | PhD, Applied Mathematics and Computational Science May 2020

Thesis: *Laplacians of Cellular Sheaves: Theory and Applications*

Honors: Carlitz-Zippin Prize (outstanding thesis in Mathematics)

**Arizona State University** | BS, Mathematics & Economics May 2015

Honors: Charles Wexler Mathematics Prize (top math graduate), Dean's Medal in Economics (top econ graduate)