Jatin Nainani

NainaniJatinZ | jsnainani@gmail.com | nainanijatinz.github.io | Google Scholar

Mechanistic interpretability for protein language models and large language models, focusing on turning internal circuits into scientifically meaningful hypotheses.

Education

University of Massachusetts Amherst **Aug 2023 – May 2025**
Master of Science in Computer Science *GPA: 4.0/4.0*

K. J. Somaiya College of Engineering, Mumbai, India **Aug 2019 – May 2023**
Bachelor of Technology in Electronics and Telecommunication Engineering *CGPA: 9.51/10*

Publications & Preprints

- **Mechanistic evidence that motif-gated domain recognition drives contact prediction in protein language models**
  Jatin Nainani, Bryn Marie Reimer, Connor Watts, David Jensen, Anna G. Green
  Mechanistic Interpretability Workshop @ NeurIPS 2025, under review @ PNAS (MS Thesis) [bioRxiv]
- **Detecting and Characterizing Planning in Language Models**
  Jatin Nainani, Sankaran Vaidyanathan, Connor Watts, Andre N. Assis, Alice Rigg
  Mechanistic Interpretability Workshop @ NeurIPS 2025 [arXiv]
- **CS4: Measuring the Creativity of Large Language Models Automatically by Controlling the Number of Story-Writing Constraints**
  Atmakuru, Anirudh*, Jatin Nainani*, Rohith Siddhartha Reddy Bheemreddy, Anirudh Lakkaraju, Zonghai Yao, Hamed Zamani, and Haw-Shiuan Chang
  Narrative Understanding Workshop @ EMNLP 2024 [arXiv]
- **Adaptive Circuit Behavior and Generalization in Mechanistic Interpretability**
  Jatin Nainani*, Sankaran Vaidyanathan*, A. J. Yeung, Kartik Gupta, David Jensen [arXiv]
- **Smartphone Based Tactile Feedback System Providing Navigation and Obstacle Avoidance to the Blind and Visually Impaired**
  Anish Pawar, Jatin Nainani, Priyanka Hotchandani, Gayatri Patil
  IEEE ICAST 2022 [IEEE]
- **Evaluating Brain-Inspired Modular Training in Automated Circuit Discovery for Mechanistic Interpretability**
  Jatin Nainani [arXiv]

Research & Industry Experience

NVIDIA **Jun 2024 – Present**
CAD / AI Engineer *Santa Clara, CA*

- **Building** LLM-driven root-cause analysis tools for static timing analysis.
- **Evaluating** LLMs on large netlists for reading, tracing, and making safe edits.

Machine Learning Alignment & Theory Scholars (MATS) **Oct 2024 – Nov 2024**
Research Trainee (Mechanistic Interpretability) *Berkeley, CA* *LessWrong Blog Post*

- **Proposed** binary masking optimization over SAE latents to find circuits sparser and better at recovering performance than attribution patching.
- **Found** circuits in Gemma 9B by placing residual SAEs at intervals throughout the model, rather than at every layer and type, scaling interpretability.
- **Redteamed** a vulnerability of the model using insights gained from the error detection circuit.

NVIDIA **May 2024 – Sep 2024**

VLSI CAD Intern *Santa Clara, CA*

- **Developed** a hierarchical multi-agent framework using LLMs to automate analysis of complex hardware timing reports, reducing engineer debugging time across multiple reports.
- **Implemented** an agentic RAG pipeline to retrieve timing information and distill it into a timing debug relation graph, achieving a 90% pass rate on multi-report benchmarks.

Technical Skills

- Programming: Python, PyTorch, NumPy, Pandas
- Interp Tooling: TransformerLens, nnsight, PyTorch hooks, sparse autoencoders
- Comp Bio: py3Dmol, biotite, esm (ESM-style protein LMs), contact maps
- Domains: Mechanistic Interpretability, Protein Language Models, Computational Biology, LLM Agents, Transformers, NLP.