Anirban Sarkar

Computational Postdoctoral Fellow, Cold Spring Harbor Laboratory, USA
+1 617 821 5834 / +91 7675 909695 • anirbansarkar-cs.github.io
asarkar@cshl.edu / anirbansarkar.cs@gmail.com

Research Interests

Core Areas: Machine Learning, Deep Learning, Computer Vision, Computational Biology

Focus: Deep Generative Modeling, Explainable AI, Inference-time Control, Trustworthy & Robust Machine Learning

Applications: Model interpretability and robustness under distribution shifts; deep generative models for biological sequence design, inference-time optimization, and evaluation; emerging directions in concept-aligned, cross-modal counterfactual generation

Academic Positions

Cold Spring Harbor Laboratory — **Cold Spring Harbor, NY, USA**
Computational Postdoctoral Fellow — *2023–Present*
Simons Center for Quantitative Biology · Advisor: Peter Koo (Koo Lab)
Focus: Developing deep generative models tailored with biological priors and interpreting them to understand how they build representations to push the boundaries of basic science and cancer biology

Massachusetts Institute of Technology — **Cambridge, MA, USA**
Postdoctoral Associate — *2022–2023*
Department of Brain and Cognitive Sciences · Advisors: Xavier Boix and Pawan Sinha (Sinha Lab)
Focus: Bridging human and machine intelligence through neuroscience-inspired approaches to improve interpretability and address the lack of robustness outside training distributions

Education

Indian Institute of Technology Hyderabad — **India**
Ph.D. in Computer Science — *2016–2022*
Advisor: Vineeth N Balasubramanian
Research Focus: Gradient-based and causal explainability techniques, self-explaining neural networks, adversarial and attributional robustness
Awards: IKDD Best Doctoral Dissertation in Data Science Award (Winner)

National Institute of Technology Rourkela — **India**
M.Tech in Computer Science — *2014–2016*
Specialization: Information Security

IIEST Shibpur — **India**
Master of Computer Applications (MCA) — *2007–2010*

Selected Publications

Google Scholar Citations: 5400+ · h-index: 8

Recent High-Impact Papers

MLCB 2024: **A. Sarkar**, Z. Tang, C. Zhao, P. Koo. "**Designing DNA With Tunable Regulatory Activity Using Discrete Diffusion**". Machine Learning in Computational Biology. *Long Oral Presentation · [41 citations]*

CVPR 2022: **A. Sarkar**, D. Vijaykeerthy, A. Sarkar, VN Balasubramanian. "**A Framework for Learning Ante-hoc Explainable Models via Concepts**". IEEE Conference on Computer Vision and Pattern Recognition. *[104 citations]*

WACV 2022: **A. Sarkar**, A. Sarkar, VN Balasubramanian. "**Leveraging Test-Time Consensus Prediction for**

Robustness against Unseen Noise". IEEE Winter Conference on Applications of Computer Vision. *[8 citations]*

NeurIPS 2021: **A. Sarkar**, A. Sarkar, S. Gali, VN Balasubramanian. "**Adversarial Robustness without Adversarial Training: A Teacher-Guided Curriculum Learning Approach**". *[16 citations]*

AAAI 2021: A. Sarkar, **A. Sarkar**, VN Balasubramanian. "**Enhanced Regularizers for Attributional Robustness**". Association for the Advancement of Artificial Intelligence. *[22 citations]*

ICML 2019: A. Chattopadhyay, P. Manupriya, **A. Sarkar**, VN Balasubramanian. "**Neural Network Attributions: A Causal Perspective**". *Long Oral Presentation · [220 citations]*

WACV 2018: **A. Sarkar**, A. Chattopadhyay, P. Howlader, VN Balasubramanian. "**Grad-CAM++: Generalized Gradient-based Visual Explanations for Deep Convolutional Networks**".
[4992 citations] *- Highly cited paper*

Preprint: **A. Sarkar**, M. Groth, I. Mason, T. Sasaki, X. Boix. "**Deephys: Deep Electrophysiology - Debugging Neural Networks under Distribution Shifts**".

Workshop Papers

ICML 2026: **A. Sarkar**, A. Duran, P. Koo. "GPA: Generative Population Annealing for Test-Time Sequence Design". Workshop on Generative & Agentic AI for Biology

ICLR 2025: **A. Sarkar**, Y. Kang, N. Somia, P. Koo. "Understanding DNA Discrete Diffusion for Engineering Regulatory DNA Sequences". Workshop on AI for Nucleic Acids

NeurIPS 2024: **A. Sarkar**, Z. Tang, C. Zhao, P. Koo. "Designing DNA With Tunable Regulatory Activity Using Discrete Diffusion". Workshop on AI for New Drug Modalities

ICCV 2021: **A. Sarkar**, A. Sarkar, VN Balasubramanian. "Leveraging Test-Time Consensus Prediction for Robustness against Unseen Noise". Workshop on Adversarial Robustness In The Real World

Featured Research Projects

Generative Population Annealing for Test-Time Sequence Design *2025–Present*
Inference-time control of generative models to design sequences beyond the training activity range
- Model-agnostic method that steers frozen generators without retraining or model-specific hooks
- Population annealing-based test-time sampling that pushes generation past activities seen during training
- Accepted at GenBio 2026 (ICML Workshop on Generative & Agentic AI for Biology)

DNA Sequence Design with Discrete Diffusion Models *2023–Present*
Developing generative models for designing regulatory DNA sequences with tunable activity
- Novel application of discrete diffusion models to biological sequence design
- Enables controllable generation of DNA sequences with specific regulatory properties
- Published at MLCB 2024 (Long Oral) and featured in ICLR 2025 and NeurIPS 2024 workshops

Deephys: Deep Electrophysiology *2022–2023*
Analyzing artificial neuron behaviors under distribution shifts inspired by neuroscience
- Interactive visualization tool for debugging deep models
- Novel framework connecting neuroscience principles to deep learning interpretability
- Enables understanding of model failures through neuron-level analysis

IBM Research Collaboration
Ante-hoc Explainable Models via Concepts *2021–2022*
Learning to predict and explain jointly through concepts during training
- Supports multiple supervision levels: unsupervised, weakly-supervised, and fully-supervised
- Semantic segmentation of latent concepts for human-understandable explanations
- Achieves significant performance improvements while maintaining interpretability

Teacher-Guided Adversarial Robustness *2020–2021*

Non-iterative robust training method without generating adversarial examples
- 10x faster training compared to traditional adversarial training methods
- Maintains high natural accuracy (minimal trade-off) while improving robustness
- Enforces attribution alignment to actual objects, reducing attack surface

Enhanced Attributional Robustness *2019–2021*

Novel regularizers for training models with robust attribution maps
- Attribution-based contrastive regularizer enforcing realistic pixel importance distributions
- Attribution change-based regularizer for stability under imperceptible perturbations
- Achieved state-of-the-art results across multiple robustness benchmarks

Causal Attribution Methods *2018–2019*

Interpreting neural networks through Structural Causal Models (SCMs)
- Efficient calculation of interventional expectations and causal attributions
- Provides both global understanding and local justifications of model decisions
- ICML 2019 Long Oral presentation with 200+ citations

Grad-CAM++: Visual Explanations for Deep Networks *2017–2018*

Generalized gradient-based visual explanation method for CNN architectures
- Improved localization and faithfulness compared to existing CAM-based methods
- Handles multiple objects and better coverage of object regions
- Most cited work with ∼5000 citations, widely adopted in computer vision community

Visiting Positions

University of Tokyo **Japan**
Visiting Research Student *Jun–Jul 2017*
Machine Intelligence Lab · Advisor: Prof. Tatsuya Harada
- Sakura Science Plan Award recipient
- Explored causal inference applications in machine learning for model explainability
- Developed foundational understanding of causality in deep learning interpretability

Industry Experience

IBM India Research Lab **Bangalore, India**
Research Intern *May–Jul 2019*
Project: Self-explaining neural networks with meaningful concepts
- Developed interpretable AI systems with human-understandable building blocks
- Led to CVPR 2022 publication on ante-hoc explainable models

IBM India Pvt. Ltd. **Kolkata, India**
System Engineer *Jun 2010–Dec 2012*
Technical consultant for Oracle Applications 11i and Oracle Financials
- Oracle e-Business Suite Certification from Oracle University

Honors & Awards

2022: IKDD Best Doctoral Dissertation Award in Data Science · *Winner*

2022: **NASSCOM AI Gamechanger Award** · AI Research (DL Algorithms) · *Runner-up*

2022: CVPR travel Grant

2021: Selected for WACV Doctoral Consortium

2019: ICML Travel Grant

2019: IBM Research Internship on AI Explainability and Active Learning

2018: Machine Learning Summer School · Universidad Autónoma de Madrid, Spain

2017: Sakura Science Plan Award · University of Tokyo Research Internship

Teaching Experience

IIT Hyderabad
Teaching & Research Assistant *2016–2022*

- **Deep Learning for Computer Vision** (NPTEL National Platform) - Course Link
- Deep Learning for Vision (CS5370) · 100+ graduate and undergraduate students
- Optimization Methods in Machine Learning (CS6230) · 25+ students
- Deep Learning (CS5480) · 90+ students
- Applied Machine Learning (CS6510) · 120+ students

Professional Service

Conference Reviewing

ICML (2025), NeurIPS (2022-2026), CVPR (2021-2023), ICLR (2022), ICCV (2021), WACV (2022)

Community Service

ICML 2020 Student Volunteer

ACM India Student Chapter (Former Member)

Technical Skills

Languages: Python, C/C++, R, MATLAB, SQL/PL-SQL **ML/DL**: PyTorch, TensorFlow, Keras

Tools: Git, LaTeX, Linux, Jupyter **Libraries**: NumPy, Pandas, Scikit-learn, Matplotlib

Selected Talks & Presentations

2025: "Designing DNA With Tunable Regulatory Activity" · QB/AI Seminar · Cold Spring Harbor Laboratory

2024: "Designing DNA With Tunable Regulatory Activity" · MLCB 2024 · Long Oral

2022: "Deephys: Deep Electrophysiology" · Internal Workshop · Fujitsu Research, Kawasaki, Japan

2022: "IKDD Best Doctoral Dissertation Award Talk" · Data Science Symposium

2022: "Learning Ante-hoc Explainable Models" · CVPR 2022 · New Orleans, USA

2021: "Adversarial Robustness without Adversarial Training" · NeurIPS 2021 · Virtual

2021: "Doctoral Consortium Presentation" · WACV 2021

2019: "Neural Network Attributions: A Causal Perspective" · ICML 2019 · Long Beach, CA

2018: "Grad-CAM++: Visual Explanations for Deep Networks" · WACV 2018 · Lake Tahoe, NV

References

Available upon request