# Ann Huang

PhD student in AI and computational neuroscience at Harvard University
Cambridge, MA | annhuang@fas.harvard.edu | ann-huang-0.github.io | Google Scholar | LinkedIn

## Research Statement

My research lies at the intersection of representation learning, interpretability, and dynamical systems theory, guided by three recurring themes:

(1) **System identifiability**: I develop methods and theory to identify and steer the computations implemented by artificial and biological neural networks. In particular, I study **under-specification**: when a task or training objective admits many models with the same performance but different internal mechanisms. I characterize and traverse these spaces of functionally equivalent networks, separating what the task requires from what optimization selects, and steering models toward solutions with desired internal properties.

(2) **Representational alignment**: I design metrics that compare computations across systems by their **dynamics**: how internal states evolve in time and how they respond to inputs.

(3) **Training dynamics**: I combine theory and experiment to study how curriculum learning speeds up learning and steers networks away from bad minima, and how training shapes models' representations and selects among the many solutions a task admits.

Ultimately, I aim to translate our mechanistic understandings of deep learning models into more efficient training and safer models that are robust to adversarial attacks and reward hacking.

## Education

**Harvard University** — Aug 2023 – May 2028 (expected)
*Ph.D. in Neuroscience; Advisor: Kanaka Rajan; Kempner Institute Graduate Fellow* — *Cambridge, MA*

**McGill University** — Sept 2019 – May 2023
*B.Sc. Honours Neuroscience, Minor Mathematics, Minor Computer Science* — *Montreal, QC*

- Cumulative GPA: 4.00/4.00; Dean's Honours List & First Class Honours in Neuroscience.

## Current Projects

**Traversing the Solution Space of Neural Networks with Hessian Null-Space Continuation**
*With Ostrow M., Lu H., Redman W., Kozachkov L., and Rajan K.*
Developing Hessian null-space continuation (HNC), a scalable method for traversing functionally equivalent networks with distinct internal mechanisms. Applied across RNNs, vision transformers, RL agents, and language models.

**Training is Less Consistent in Mixture of Experts than Dense Models**
*With Saphra N.*
Showing that sparse mixture-of-experts models are systematically less consistent across training seeds than performance-matched dense models. Developing stability-aware scaling guidelines for MoE transformer architectures.

## Selected Publications

* denotes equal contribution.

**Huang A.**, Ostrow M., Lu H., Redman W.*, Kozachkov L.*, Rajan K.*, Traversing the Solution Space of Neural Networks with Hessian Null-Space Continuation, *In preparation.*

Vastola J., **Huang A.**, Singh S., Gershman S., Rajan K., Effectiveness of curriculum learning depends on reward sparsity and competing optima: analysis of a tractable model of policy learning, *Under review.*

**Huang A.**, Singh S., Rajan K., Measuring and Controlling Solution Degeneracy across Task-Trained Recurrent Neural Networks, **NeurIPS 2025 (Spotlight, 3%).** *Establishes a multi-level framework to measure inter-model variability and identifies four control factors governing solution degeneracy.*

**Huang A.***, Ostrow M.*, Singh S., Kozachkov L., Fiete I., Rajan K., InputDSA: Demixing then Comparing Recurrent and Externally Driven Dynamics, **ICLR 2026 (Top 2% scores).** *Interpretability tool to separate intrinsic vs. input-driven dynamics in neural networks, and compare across models on either or both components.*

**Huang A.**, Singh S., Rajan K., Learning Dynamics and the Neural Geometry of Recurrent Neural Controllers, *Interpretable Policy Workshop @* ***RLC 2024*** **(Oral).**

Chiappa A., Vegas A., **Huang A.**, Mathis A., Latent Exploration for Reinforcement Learning, **NeurIPS 2023.**

Latin D.*, **Huang A.***, Mavor-Parker A.*, Ghosh A., Richards B.A., Zador A., Stochastic Wiring of Cell Types Enhances Fitness by Generating Phenotypic Variability, *bioRxiv.*

Lai L.*, **Huang A.***, Gershman S.J., Action Chunking as Policy Compression, *Cognition.*

Lin D., **Huang A.**, Richards B.A., Temporal Encoding in Deep Reinforcement Learning Agents, *Scientific Reports.*

## RESEARCH EXPERIENCE

**Rajan Lab, Harvard University** Mar 2024 – Present
*Supervisor: Kanaka Rajan* *Cambridge, MA*

- Developed a systematic framework for measuring and controlling **solution degeneracy** in task-trained RNNs, identifying four factors that mediate inter-seed variability at the level of behavior, neural dynamics, and weights, which serves as a principled guideline for when a mechanism found in one model transfers to another (NeurIPS 2025 Spotlight).
- Built **InputDSA**, an interpretability tool for separating and comparing intrinsic vs. externally driven dynamics in complex systems, applicable to both neural data and model activations (ICLR 2026).
- Currently developing **Hessian Null-Space Continuation**, a scalable method that traverses the space of functionally equivalent networks and steers models toward mechanisms with desired internal properties at fixed behavior.

**Richards Lab, Mila** Sept 2021 – June 2024
*Supervisor: Blake Richards* *Montreal, QC*

- Developed **Stochastic Genomic Bottleneck**, a generative method that compresses task-relevant priors into RNN weights, enabling rapid learning from minimal data — a structured-prior analog of pretraining (bioRxiv).
- Interpreted the memory mechanisms and navigation strategies implemented by deep RL agents via decoding, dynamical systems analysis, dimensionality reduction, and ablation (Scientific Reports).

**Mathis Lab, EPFL** June 2022 – May 2023
*Supervisor: Alexander Mathis* *Geneva, Switzerland*

- Designed a **goal-conditioned hierarchical deep RL** algorithm for musculoskeletal control, combining a pretrained language model that proposes behavioral subgoals with a distributed actor-critic architecture that executes the low-level actions.
- Characterized and quantified RL agents' exploration behavior and proposed an improved exploration strategy that leverages the covariance across actuators for coordinated exploration and control (NeurIPS 2023).

**Gershman Lab, Harvard University** May 2021 – Sept 2021
*Supervisor: Samuel Gershman* *Cambridge, MA*

- Developed a theory that explains action chunking as **policy compression** under the rate-distortion framework; designed cost-sensitive actor-critic RL models with information-theoretic penalties on policy complexity.
- Validated models against human behavior via Bayesian model selection and simulation (Cognition).

**Brandon Lab, Douglas Research Institute** Nov 2019 – May 2021
*Supervisor: Mark Brandon* *Montreal, QC*

- Studied long-term stability of spatial context representations in the hippocampus despite neural drift
- Built a Bayesian decoder for animal location from neural recordings in mice; used representational similarity analysis and generalized linear models to factor the neural representations of animal location into orthogonal contextual and temporal-drift components.

## SELECTED TALKS

**New England Mechanistic Interpretability Workshop (NEMI)** (Talk, top 4%), August 2026, Boston, MA.
**Emergence of Intelligent Phenomenon Journal Club @ Harvard**, April 2026, Cambridge, MA.
**Computational and Systems Neuroscience (Cosyne)** (Talk, top 2.5%), March 2026, Lisbon, Portugal.
**Perich Lab, Mila & Université de Montréal**, February 2026, Montreal, QC.
**Harvard Medical School Friday Seminar Series**, February 2026, Boston, MA.
**NeurReps Workshop @ NeurIPS 2025**, December 2025, San Diego, CA.
**Kempner Institute All-Hands**, October 2025, Cambridge, MA.
**Kempner Spring into Science**, March 2025, Cambridge, MA.
**Interpretable Policy Workshop @ Reinforcement Learning Conference (RLC)** (Oral), August 2024, Amherst, MA.
**From Neuroscience to Artificially Intelligent Systems (NAISys)**, September 2024, Cold Spring Harbor, NY.
**RL at Harvard Workshop**, August 2023, Cambridge, MA.
**Cognitive Computational Neuroscience (CCN)** (Oral, top 5%), August 2022, San Francisco, CA.
**Canadian Computational Neuroscience Symposium** (Spotlight Talk), June 2022.
**"Maps in RL" Workshop @ Reinforcement Learning and Decision Making (RLDM)**, June 2022, Providence, RI.

## SELECTED AWARDS & HONORS

**ICLR 2026 Top 2% Scores** (InputDSA paper)
**Cosyne Presenter Travel Grant** ($1,000), 2026
**NeurIPS 2025 Spotlight** (top 3% of accepted papers)
**Kempner Institute Professional Development Fund** ($2,000), 2025
**Kempner Institute Graduate Fellowship**, Harvard (full tuition + $47K/yr stipend), 2023–2029
**NSERC Canada Graduate Scholarship** ($17,500), 2023–2024
**EPFL Summer Research Fellowship** ($8,000; selected 20/843), 2022

## TECHNICAL SKILLS

Python, PyTorch, SLURM, Hugging Face, Distributed training (FSDP/DDP), Dynamical systems analysis, Training dynamics