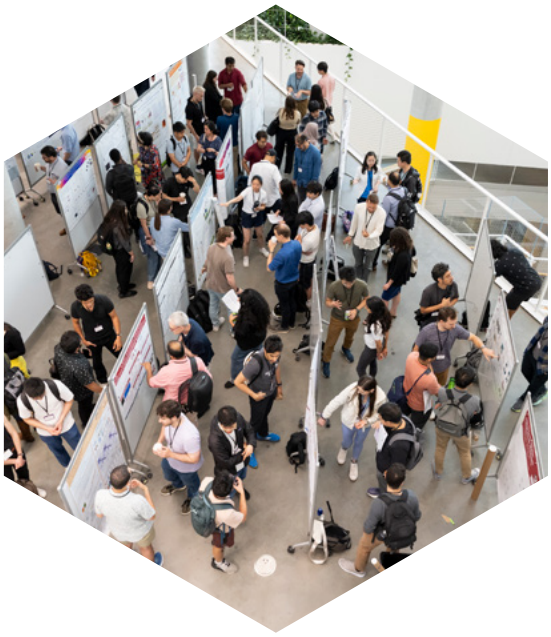
Kempner INSTITUTE
For the Study of Natural & Artificial Intelligence

HARVARD UNIVERSITY



ANNUAL REPORT

The Kempner Institute for the Study of Natural and Artificial Intelligence

DECEMBER 2025

TABLE OF CONTENTS

Research Overview

The Kempner Institute brings together faculty, fellows, and students from diverse fields to advance our understanding of intelligence. This report presents a selection of notable completed and ongoing research projects that reflect the depth and breadth of the Kempner community's work over the past year. These studies fall into, and are organized by, three major themes. First, Kempner researchers have extended the capabilities of current AI architectures and training procedures, paving the way to the next generation of advanced models. Second, by strengthening the theoretical foundations of AI, they are revealing the computations that allow models to learn and reason as well as ensuring that technical progress is interpretable and scientifically grounded. Third, by borrowing methods from ML to study the brain and by implementing brain-like features into ML models, they are facilitating cross-disciplinary insights into the foundations of natural and artificial intelligence. A full list of publications and preprints is provided in the Appendix.

"It's rare to have so many skill sets and resources in the same place, enabling us to move rapidly from idea to model to experimental testing, and then iterate on that. Nothing like this existed at Harvard before the Kempner."

GEORGE ALVAREZ, FRED KAVLI PROFESSOR OF NEUROSCIENCE, DEPARTMENT OF PSYCHOLOGY, KEMPNER AFFILIATE FACULTY

INNOVATION IN AI

The Kempner Institute seeks to develop new AI approaches that broaden scientific understanding and contribute to practical progress in the field.

Advancements in Diffusion Models

Diffusion models form the basis of a growing number of generative AI systems, including systems that generate realistic images, video, and language. During the last year, Kempner scientists have led efforts to expand the capabilities of these models.

One area of focus is Masked Diffusion Models (MDMs). A study by Sham Kakade and collaborators demonstrated that although these models face computational challenges during training, they can achieve remarkable accuracy at inference by dynamically choosing the decoding order. The study showed that these adaptive inference strategies sidestep hard problems created during training, greatly improving performance. In logic puzzles such as Sudoku, adaptive inference raised MDM accuracy from below 7% to about 90%, outperforming larger autoregressive models.

In related work, a team including Institute Investigator Yilun Du and incoming Institute Investigator Michael Albergo introduced Flexible Masked Diffusion Models (FlexMDMs), a discrete diffusion framework that enables variable-length sequence generation while preserving the flexible inference of MDMs. FlexMDMs achieved notable performance gains, improving math accuracy from 58% to 67% and code infilling from 52% to 65% after fine-tuning on 16 of the Kempner AI Cluster's H100 GPUs.

In a complementary project, Du and collaborators employed ideas from classical search to control diffusion models during inference, adapting generative outputs to different goals. The study introduced a framework that blended step-by-step refinement of results with broader exploration strategies to efficiently guide the model's generation process. Tested on areas such as planning, offline reinforcement learning (RL), and image generation, the method achieved strong improvements in both accuracy and efficiency.

Institute Investigator Yilun Du's recent work on diffusion models includes a framework that blends step-by-step refinement of results with broader exploration strategies to efficiently guide a model's generation process.

Breakthroughs in Biomedical AI

The Kempner harnesses advanced artificial intelligence and the computational power of its GPU cluster to propel breakthroughs in biomedical research, yielding transformative insights across basic science and clinical domains. A team led by Associate Faculty Marinka Zitnik is building AI models that integrate data at a variety of scales, from molecular profiles to genetic sequences to patient data. This approach uncovers patterns beyond the reach of traditional methods, opening new possibilities to transform diagnostics, accelerate therapeutic innovation, and deepen our understanding of health and disease.

A notable example of this work is ATOMICA, a geometric deep learning model developed by a team that includes Zitnik and Graduate Fellow Ada Fang. ATOMICA learns atomic-scale representations of intermolecular interfaces across five modalities—proteins, small molecules, metal ions, lipids, and nucleic acids. Using this framework, the researchers constructed modality-specific "interfaceome" networks that revealed disease-associated patterns, such as proteins linked to asthma within lipid networks, and to myeloid leukemia within ion networks. The model also advances understanding of the "dark proteome"—proteins with unknown function—by accurately predicting new ligand-binding sites.

The Zitnik Lab also developed ProCyon, a foundation model designed to model, generate, and predict protein phenotype descriptions. ProCyon integrates phenotypic and protein data by training a large language model (LLM) alongside multimodal molecular encoders. Its architecture enables zero-shot generalization, allowing the model to propose potential phenotypes for underexplored proteins, including those associated with neurodegenerative diseases such as Parkinson's.

"I've been particularly impressed by how scientific hypotheses translate into large-scale computational tests and, in turn, drive advances in data and model-efficient AI. This process helps optimize large-scale experiments and align infrastructure with research needs in AI."

MARINKA ZITNIK, ASSISTANT PROFESSOR OF BIOMEDICAL INFORMATICS, KEMPNER ASSOCIATE FACULTY

THE SCIENCE OF AI

The Kempner seeks to define the conceptual foundations that will guide the development of next-generation AI systems and training paradigms. This work spans parallel research programs in AI engineering, algorithmic design, mathematical theory, and model interpretability.

Scaling, Optimization, and AI Engineering

Scaling laws are a cornerstone of modern AI, enabling researchers to predict how model performance improves with increased data, parameters, or computational power. However, these laws typically apply only for a fixed dataset, leaving open questions about their predictive reliability when data distributions change. To address this challenge, Research Fellow David Brandfonbrener led a study introducing "loss-to-loss" prediction. Instead of fitting separate power laws for each dataset, the study demonstrated that training and test losses obey simple shifted power-law relationships across datasets, allowing accurate extrapolation even when moving between domains as distinct as programming code and natural language.

Graduate Fellow Depen Morwani led studies that advanced model training techniques, including one that examined the efficiency of large-scale optimization by critical batch size, and another that proposed a new optimizer method called Simplified-AdEMAMix.

While model performance is typically assessed in terms of accuracy, training efficiency is also a critical consideration. At the Kempner, optimizing training processes represents a key area of research. A study by a team including Graduate Fellow Depen Morwani examined the efficiency of large-scale optimization by critical batch size (CBS). The team found that CBS scales with data size rather than model size, a result supported by infinite-width theory. This finding enables more efficient use of compute and informs training strategies in resource-constrained environments.

Another study by Kempner researchers including Morwani and Kakade established explicit connections between accelerated stochastic gradient descent variants and several recently proposed optimizers. The team went on to propose Simplified-AdEMAMix, an optimizer that matches AdEMAMix performance across batch sizes while removing redundant momentum terms, reducing complexity without sacrificing speed or convergence.

Understanding the trade-offs between memorization and reasoning is essential to optimizing the performance and reliability of AI systems. A study led by Harvard postdoctoral fellow Samy Jelassi and supervised by Kempner Fellow Eran Malach compared Mixture-of-Experts (MoE) models to dense transformers and found that MoEs behaved like parrots that could remember vast amounts of information without a strong ability to reason about it. As they increased the number of experts while keeping the active parameters fixed, MoEs became better at memorization but showed no real improvement in reasoning. The study showed that MoEs were powerful tools for storing and recalling information, while dense models remained better suited for reasoning and understanding.

Kempner researchers are also at the forefront of efforts to understand and improve this technique. Lowering the precision of numerical computations has proven to be a powerful way to enhance efficiency; a team led by Graduate Fellow Chloe Huangyuan Su and

supervised by Kempner Senior ML Research Scientist Nikhil Anand found that low-precision formats (MXFP6, MXFP8) in LLMs revealed sharp instabilities at scale. They identified two effective mitigation strategies: maintaining higher precision for activations and limiting quantization exclusively to the forward pass.

A team including Associate Faculty Cengiz Pehlevan introduced a different perspective on precision through the development of "precision-aware" scaling laws. This work extends traditional scaling frameworks to account for numerical precision during both training and inference. Their new scaling laws accurately captured the trade-offs between precision, compute, and performance. Precision-aware scaling laws enabled the researchers to predict the loss of a model that used mixed precisions. This unified framework provides a predictive foundation for designing efficient AI systems that balance accuracy and resource use.

Associate Faculty Cengiz Pehlevan (center) employed advanced techniques from physics and statistics in a variety of projects, including the introduction of "precision-aware" scaling laws, the development of a rigorous theory for in-context learning, and a systematic investigation of transfer learning in infinite-width neural networks.

Theoretical Advances in AI

The Kempner's theoretical researchers are providing insights into the mathematical and theoretical bases of AI training and performance by deploying a wide range of analytic tools, including the study of phase transitions, Bayesian frameworks, and geometric approaches.

A team led by Pehlevan and including Graduate Fellow Mary Letey developed a rigorous asymptotic theory for in-context learning in linear attention models. They explained how such systems can learn new tasks directly from examples provided in their input sequences, a hallmark of modern transformer models. Their findings revealed a phase transition driven by the diversity of pretraining tasks and the structure of the training data.

In another study Pehlevan and his team investigated transfer learning in infinite-width neural networks, which are a useful abstraction that can provide insight into the properties of many artificial neural networks (ANNs). They constructed a theory in which both the pretraining or "source" task and the downstream or "target" task allow feature adaptation, introducing an "elastic weight coupling" term that governs how much a network reuses representations from its source task. By combining theoretical derivations with experiments on both synthetic and real data, the study offered a principled understanding of how and when transfer learning leads to more efficient adaptation.

In other theoretical work, this time rooted in perspectives from physics, Associate Faculty Haim Sompolinsky and collaborators applied statistical mechanics to unveil the role of attention paths in accounting for the performance of transformer-based architectures. Their recent study suggests new pruning strategies for transformer heads, pointing toward leaner yet equally capable models.

Also employing concepts from physics, the work of Research Fellow T. Anderson Keller enhances the representational capacities of recurrent neural networks (RNNs). In a recent study he introduced a new type of RNN called Flow Equivariant RNNs. These models extend

the idea of equivariant architectures from static transformations to continuous flows such as visual motion. This approach not only leads to improved training efficiency and generalization but also resonates with a renewed interest in recurrent architectures sparked by advances in state-space models (SSMs).

Recent work by Research Fellow T. Anderson Keller (center) introduces a new type of RNN called Flow Equivariant RNNs that extends the idea of equivariant architectures from static transformations to continuous flows.

The Science of Reinforcement Learning

The Kempner is advancing the science of RL by developing more efficient training and inference frameworks while also tackling challenges that arise when applying RL to LLMs.

Institute Investigator Kianté Brantley has played a central role in this research initiative. His work focuses on improving the decision-making capabilities of foundation models. A recent study by his team introduced Scalable Offline RL (SORL), a breakthrough framework that makes offline RL both more efficient and more expressive. The authors show that SORL regularizes behavior policy, which ensures stability, and achieves state-of-the-art performance across a wide range of benchmark tasks. SORL also demonstrates positive scaling behavior: its performance improves as more test-time compute is allocated.

Brantley and his team have also developed a new technique for reinforcement learning from human feedback in multi-turn dialogue. A recent study introduced REFUEL, a regression-based algorithm designed for multiple turns of interaction. REFUEL enables small models to outperform much larger systems.

In parallel to Brantley's work on RL, another Kempner team featuring Associate Faculty David Alvarez-Melis, Research Fellow Eran Malach, and Harvard postdoctoral fellow Samy Jelassi examined the relationship between backtracking and RL. Using controlled experiments on tasks such as CountDown and Sudoku, the researchers discovered that sequential backtracking sometimes performs worse than parallel sampling, depending on the task structure and the model's training regime. The paper showed that RL can mitigate these

Institute Investigator Kianté Brantley (right) and his team have introduced Scalable Offline RL, a breakthrough framework that makes offline RL both more efficient and more expressive.

effects by allowing models to learn when and how to backtrack rather than following rigid search procedures. Models equipped with backtracking and fine-tuned through RL were able to develop more efficient, adaptive reasoning strategies.

Interpretability of AI Models

The Kempner is advancing AI interpretability and explainability by bringing a host of analytical tools to bear on the question of how models learn, represent information, and reason.

A study led by Research Fellow Isabel Papadimitriou and supervised by Associate Faculty Stephanie Gil used sparse autoencoders (SAEs) to explore how concepts are organized within vision-language models (VLMs). The team used SAEs to investigate how VLMs represent images and text that are closely related in shared embedding spaces. They found that while rare concepts vary between runs, common ones remain stable. Most concepts were tied to a single modality but often lay near a shared subspace, suggesting cross-modal links. To measure this, they introduced the Bridge Score, identifying concept pairs that are both geometrically aligned and co-activated across image-text inputs.

SAEs were also the focus of a related study by a team that included Research Fellow Thomas Fel and Associate Faculty Demba Ba. The researchers examined the assumptions behind different SAE architectures and found that they are not interchangeable. Each SAE only captured concepts whose geometry fit its inductive biases, missing others that were nonlinear or required varying levels of complexity. Neural network representations often included such nonlinear and heterogeneous features, causing many existing SAEs to fail. To address this, the authors introduced a new architecture called SpaDE (Sparsemax Distance Encoder) that overcomes the limitations of earlier SAEs.

Central to LLM research is the question of whether breakthroughs in performance over the course of training are gradual or abrupt. To address this question, a team led by Research Fellow Naomi Saphra introduced POLCA clustering, a method that detects hidden breakthroughs during training. Tested on both language and synthetic tasks, the method showed that models often undergo distinct phase transitions where specific capabilities emerge abruptly. These hidden phase transitions could be used for unsupervised interpretability analyses.

In other LLM-focused research, a team including Research Fellow Jennifer Hu and Affiliate Faculty Tomer Ullman investigated how both people and models distinguish between the impossible and the inconceivable. They found that humans reliably separate these two concepts and that statistical language models make similar distinctions. The result suggests that LLMs internalize subtle distinctions of possibility and conceivability. This work offers a bridge between cognitive science and computational modeling, illuminating how statistical patterns in language give rise to conceptual structure.

Research Fellow Isabel Papadimitriou (above) and Associate Faculty Stephanie Gil used sparse autoencoders to explore how concepts are organized within vision-language models.

AI AND THE BRAIN

The Kempner employs state-of-the-art AI tools to study natural cognition, information-processing, and learning. Research projects in this domain, which forms part of the growing discipline of NeuroAI, incorporate three broad approaches. First, researchers study artificial and biological systems in parallel, deriving cross-disciplinary insights about architectures and learning. Second, researchers use the lens of RL to understand reward-driven learning in humans and animals. Third, researchers deploy AI-powered tools to analyze neural and behavioral data.

Cross-disciplinary Insights

A study by a team that included Graduate Fellow Mozes Jacobs and Research Fellow T. Anderson Keller explored how traveling waves of neural activity could enable spatial information integration in ANNs. The authors concluded that traveling waves offered a stable, parameter-efficient means of integrating information over time, aligning with biological neural dynamics and suggesting a promising alternative to the global self-attention mechanisms that are essential to transformer architectures.

Another cross-disciplinary team of Associate Faculty, including Cengiz Pehlevan, Venkatesh Murthy, and Samuel Gershman, investigated whether mice continued to refine neural representations after behavioral mastery, proposing a biological analogue of "grokking" in AI models. Through a reanalysis of published neural recordings, they found that neural representations of perceptual classes continued to separate from each other during overtraining, even when behavioral performance had plateaued. A simple model replicated these effects, implying parallels between cortical learning and deep network grokking dynamics. The authors argued that overtraining could strengthen and generalize sensory representations, offering insight into how biological and artificial systems achieve late-stage representational refinement beyond visible performance gains.

Associate Faculty Talia Konkle (left) and Affiliate Faculty George Alvarez use vision models to probe the structure of human perception. Their recent findings challenge common assumptions about alignment between models and the brain.

In another major research area, led by Associate Faculty Talia Konkle and Affiliate Faculty George Alvarez, Kempner researchers use vision models to probe the structure of human perception. In one study Alvarez, Konkle, and their collaborators compared 224 diverse deep neural network models to assess how model features influenced alignment with neural activity. They found that differences in architecture or task objectives had minimal effects on alignment when other factors were controlled, whereas variations in visual training datasets produced the largest and most consistent impact. Their findings challenge common assumptions about alignment between models and the brain, and they outline how controlled model comparison can be leveraged to identify the common computational principles underlying biological and artificial visual systems.

Complementing some of these findings, Konkle, Alvarez, and Graduate Fellow Fenil Doshi provided computational evidence that feed forward convolutional neural networks fine-tuned for contour detection exhibited human-like contour integration without relying on lateral connections, recurrence, or top-down feedback.

Kempner researchers are also studying alignment between artificial and natural systems at the behavioral level. Research Fellow Wilka Carvalho has spearheaded an ambitious project to extend cognitive modeling into naturalistic settings. In a recent study, Carvalho and his collaborator argued that progress in AI creates new opportunities for cognitive science to build theories and models that generalize across the full spectrum of natural behavior. Carvalho and his co-author outlined methodological guidance for integrating AI-based modeling with experimental control, proposing that computational models capable of solving naturalistic problems could still yield explanatory, mechanistic understanding.

To facilitate this approach, Carvalho and collaborators introduced NiceWebRL, a Python library that empowers researchers to run online human subject experiments using state-of-the-art RL environments and analytical tools. Through its support for multi-agent and large-scale behavioral experiments, NiceWebRL exemplified how state-of-the-art AI tools could supercharge cognitive science by efficiently operationalizing the naturalistic principles outlined in the earlier work, advancing the study of cognition in complex, ecologically valid contexts.

Natural and Artificial Pursuit of Rewards

Building on the naturalistic vision outlined above, a study by a team featuring Carvalho and Gershman demonstrated how a theoretical hypothesis can be tested in increasingly natural settings using tools such as NiceWebRL. The team tested whether humans use a strategy called "multitask preplay," which involves solving new problems by using counterfactual simulations. Multitask preplay enables humans and artificial RL agents to seek rewards that have not yet been explicitly pursued. Experiments comparing human participants and RL agents showed that people reuse prior solutions in new contexts, even when those solutions are not optimal, and that agents equipped with the multitask preplay algorithm display similar adaptive biases.

Research Fellow Wilka Carvalho (center) has spearheaded an ambitious project to extend cognitive modeling into naturalistic settings. He and his collaborators created NiceWebRL, a Python library for running online human subject experiments using state-of-the-art RL tools, and then demonstrated how a theoretical hypothesis can be tested in increasingly natural settings using tools such as NiceWebRL.

A study led by Bernardo Sabatini and published in *Nature* explored how neural circuits support reward-driven learning at a more mechanistic level. Using optogenetics and behavioral training in mice, they found that the posterior dorsomedial striatum is essential for rapid, trial-by-trial learning but not for recalling established memories. This work refines our understanding of how learning and memory are distributed across brain circuits and provides a biological complement to the algorithmic insights explored in computational studies.

Another study from the Sabatini Lab discovered that synapses between glutamate and GABA co-releasing neurons in the entopeduncular and their targets in the lateral habenula could rapidly switch between excitatory and inhibitory states in response to experience, resembling the learning-based synaptic updates in ANNs. This ability to reverse synaptic sign allowed the brain to flexibly adjust how neural signals influenced downstream circuits, directly shaping the activity of dopamine neurons. The researchers showed that such sign switching provided the biological counterpart to a key requirement in artificial networks, where connections must not only be able to change strength but also invert the direction of their influence to reflect changes in value predictions and policies.

AI-powered Analyses of Biological Data

Members of the Kempner community are employing a striking diversity of AI-based approaches to derive insights from brain data. Researchers are combining large-scale modeling, interpretable deep learning, unsupervised behavioral analysis, and theoretical frameworks to bridge biological and artificial systems.

One study led by Institute Investigator Kanaka Rajan introduced a scalable neural forecasting framework known as POCO, designed to predict brain activity across species and experimental sessions. By combining a simple neuron-level forecaster with a global population encoder, POCO learns to model both local and system-wide dynamics. Trained on calcium imaging data from zebrafish, mice, and *C. elegans*, the model achieves state-of-the-art accuracy at the single-cell level and can adapt rapidly to new datasets with minimal fine-tuning. POCO represents a significant advance in scalable neural modeling: it marks a step toward neural foundation models that can flexibly adapt to new biological data streams.

Kempner researchers have also demonstrated the power of AI for interpreting neural data. In a study published in *Neuron*, a team featuring Associate Faculty Demba Ba and Venkatesh Murthy and Affiliate Faculty Naoshige Uchida presented an interpretable deep learning framework for analyzing neural activity. The authors introduced Deconvolutional Unrolled Neural Learning, which they used to disentangle salience and reward prediction error signals in neural data. The method provides scalable, interpretable decompositions of neural signals that connect latent representations directly to behaviorally relevant variables.

Researchers also have developed AI tools for the analysis of large-scale rodent behavioral data. The Sabatini Lab introduced "face-rhythm," an unsupervised computational framework that quantitatively tracked and analyzed facial movements. This approach extracted interpretable behavioral components from video data, identifying natural facial behaviors

"Our discoveries over the past year are meaningful steps towards uncovering mechanistic underpinnings of intelligence between species and across environmental conditions and internal states. Our insights have long-term implications for computational research study designs, both *in vivo* and *in silico*."

KANAKA RAJAN,
ASSOCIATE PROFESSOR,
DEPARTMENT OF NEUROBIOLOGY,
KEMPNER INSTITUTE
INVESTIGATOR

in mice. When applied across various behavioral contexts, face-rhythm revealed that uninstructed facial movements predicted internal belief states and corresponded with neural activity in the primary motor cortex. These findings demonstrated that spectral features of rhythmic facial behavior closely reflected intention and motor cortical representations.

In another recent study, Research Fellow Binxu Wang, in collaboration with Affiliate Faculty Carlos Ponce, examined how neurons in the primate visual cortex align with latent representations generated by deep generative image models. By combining closed-loop neural recording with image synthesis from networks such as BigGAN and DeePSim, the team showed that neurons in different visual areas align with distinct generative manifolds: lower areas aligned with local pattern-based features, whereas higher areas aligned with object-level representations. The work highlights the value of using distinct generative models to probe the brain's representational geometry.

The study of representations lies at the heart of both AI and neuroscience, yet the concept continues to require deeper theoretical and empirical clarification. In a recent study, Associate Faculty Talia Konkle, Affiliate Faculty George Alvarez, and their collaborators have contributed to this research goal. Many studies equate representation with correlation, but this work proposed a stricter notion, "Representation with a capital R," defined by functional use. The study developed a causal perturbation framework to test whether the information measured in biological or artificial networks is functionally deployed rather than merely being present.

Research Fellow Binxu Wang (left) in collaboration with Affiliate Faculty Carlos Ponce (right), examined how neurons in the primate visual cortex align with latent representations generated by deep generative image models.

"By seamlessly integrating large-scale machine learning experiments with *in vivo* neural recordings, the Kempner cluster has been instrumental in advancing the institute's mission to uncover shared computational principles between biological and artificial vision systems."

BINXU WANG,
KEMPNER RESEARCH FELLOW

Impact Overview

The Kempner's rapid growth and the scope of its achievements are fueled by exceptional students, scientists, and engineers working across traditional boundaries. The opportunity to collaborate with accomplished faculty and promising students, paired with access to world-class computing resources, enables the Kempner and Harvard to compete for talent on a level that is unmatched at other universities. From the rapid recruitment of Institute Investigators to the cohorts of Research Fellows and students whose pioneering work will progress along with their careers, the Kempner attracts excellence. Its events draw crowds who are eager to learn, explore, and innovate, and its resources provide powerful tools for their research. The breadth of expertise and experience at the Kempner is evident through the number and quality of its publications and in the recognition that the community's work receives in honors and awards.

COMMUNITY

50
KEMPNER-SPONSORED UNDERGRADUATE RESEARCH PROJECTS

421
KEMPNER COMMUNITY MEMBERS

165% increase from 2024

74
INVESTIGATORS, ASSOCIATE, AND AFFILIATE FACULTY

49
PHD STUDENTS

across 11 graduate programs

Recruiting and Training the Next Generation of Researchers

Recruiting outstanding faculty is key to advancing the Kempner's core mission. This summer, in coordination with Harvard John A. Paulson School of Engineering and Applied Sciences (SEAS), the Kempner successfully recruited Qianqian Wang as its sixth Institute Investigator and assistant professor of computer science. Wang, who is completing her postdoctoral research at UC Berkeley, will join Harvard in 2026. Wang's research focuses on designing new visual models—working toward models that understand and interact with a dynamic visual environment, including creating 4D models that can infer 3D geometry and movement from 1D images. In addition, two new Associate Faculty have been appointed to the Kempner this year. Venkatesh Murthy, Raymond Leo Erikson Life Sciences Professor of Molecular and Cellular Biology and Paul J. Finnegan Family Director of the Center for Brain Science, is a neuroscientist who studies the neural and algorithmic basis of sensory-guided behaviors like smell. Stephanie Gil, John L. Loeb Associate Professor of Engineering and Applied Sciences, runs a robotics lab that seeks to improve information exchange in multi-agent (robot-to-robot and robot-to-human) teams using multi-agent algorithms.

The Kempner's growth also includes 10 new Research Fellows, bringing the total cohort to 20. These innovative, early-career scientists represent some of the best talent in the world, often deferring industry offers and academic appointments for an opportunity to join the Kempner—for example, Alexander Damian deferred a faculty position in mathematics and EECS at MIT to become a Research Fellow this year. The fellows are a key element

of the Kempner's dynamic environment, leading cutting-edge investigations that span the field of intelligence research. With a variety of skill sets and backgrounds, they serve as a bridge between labs and disciplines, collaborating across the academic spectrum on the groundbreaking research that is highlighted in this report's Research Overview.

As some of the institute's first Research Fellows advance in their careers, they are moving into prominent roles within academia and industry, building on the insights they developed at the Kempner and helping to expand its influence. David Brandfonbrenner and Eran Malach, whose research on scaling laws provided fundamental insights into model performance, continue their work as research scientists at Meta and Apple respectively. Jennifer Hu, now an assistant professor in Johns Hopkins's Department of Cognitive Sciences, and Isabel Papadimitriou, assistant professor in the Department of Linguistics at the University of British Columbia, will both continue research at the interface of human language and LLMs. Naomi Saphra, who will join Boston University's Department of Computing and Data Sciences next year, is pursuing novel work on interpretability in LLMs and is helping to further establish Boston as an epicenter of innovative AI research in the US.

Kempner Research Fellows such as Alex Damian (above) represent some of the best talent in the world. Damian deferred a faculty position in mathematics and EECS at MIT to become a Research Fellow this year.

With 13 newly selected Graduate Fellows pursuing degrees in computer science, applied mathematics, physics, quantum sciences and engineering, and neuroscience, the Kempner now supports 49 PhD students who are enrolled in 11 programs across Harvard and are performing research aligned with the Kempner's mission.

In addition to dozens of undergraduates who gained access to the Kempner through their work in affiliated labs, the Kempner provided support for nearly 50 undergraduate research projects through the term-time Kempner Undergraduate Research Experience (KURE) and summer Kempner Research in Artificial & Natural Intelligence for Undergraduates with Mentorship (KRANIUM) program.

Examples of stills from Wang's research optimizing 3D representations from videos. Wang is an expert in building models of a dynamic three-dimensional world using everyday images and videos. Recent work includes reconstructing 4D images from a single monocular video.

Prizes and Awards

The work of Kempner researchers is gaining international recognition. Below are a few of the prestigious honors received recently by Kempner students, faculty, and affiliates:

- Kempner researchers and collaborators won an outstanding paper award at International Conference on Machine Learning (ICML) 2025. Kempner authors include co-first author Jaeyeon Kim, a Harvard computer science PhD student co-advised by co-authors Sitan Chen AB '16, AM '16, assistant professor of computer science at SEAS, and Sham Kakade. "Train for the Worst, Plan for the Best: Understanding Token Ordering in Masked Diffusions" was one of six recipients out of roughly 12,000 submissions, 3,200 of which were selected for 42nd ICML. The award recognizes technical depth, novelty, and potential for impact in the field.
- Michael Albergo AB '17 was part of the team that won best paper award at ICLR 2025's Frontiers of Probabilistic Inference workshop for "LEAPS: A Discrete Neural Sampler via Locally Equivariant Networks." Albergo joined the Kempner last year as a member of the Harvard Society of Fellows. He officially begins his tenure as an Institute Investigator in 2026.
- Kempner Institute Investigator Kanaka Rajan received the Presidential Early Career Award for Scientists and Engineers in 2025. This award is the highest honor bestowed by the US government on outstanding scientists and engineers at the beginning of their research careers.
- Research Fellow Thomas Fel won the AFIA National Best PhD Thesis Award and the Signal, Image, and Vision Thesis Award in 2025 for his work, "Glimpses of Explainability: Recent Advances in Explaining Deep Neural Networks for Vision."

Michael Albergo AB '17 was part of the team that won best paper award at ICLR 2025's Frontiers of Probabilistic Inference workshop. Albergo joined the Kempner last year as a member of the Harvard Society of Fellows. He officially begins his tenure as an Institute Investigator in 2026.

EVENTS

- Two extraordinary Harvard undergraduates with joint concentrations in computer science and neuroscience, who did outstanding research as part of the Kempner community, earned Rhodes Scholarships. Through participation in KURE and KRANIUM, Aneesh Muppidi AB '25, SM '25 developed a senior thesis project on unsupervised agent discovery using pixel-based observations in multi-agent reinforcement learning environments. And as a member of the Zitnik Lab, Ayush Noori AB' 25, SM '25 worked on a knowledge-grounded foundation model for AI-guided scientific discovery and precision medicine in neurological disease.

Worldwide Engagement Through Events and Education

Frontiers in NeuroAI, the Kempner's inaugural symposium, brought together more than 1,000 researchers from 37 countries for a two-day, in-person and live-stream event. Recordings of the expert talks have been viewed more than 8,300 times on YouTube since the event was held in June. Talks covered new frameworks for studying brains and AI models, AI-powered insights into the brain, and next-generation AI systems—and they provided a deeper understanding of current AI models.

Workshops@Kempner, a series of interactive, in-person gatherings to facilitate training and skill-building for Kempner affiliates and members of the Harvard community, hosted nine events for a total of 257 participants this year. Topics for these events, which featured expert contributions from the Kempner's research and engineering team, included data parallelism, building transformers from scratch, optimizing ML workflows, and spike sorting. These workshops are forming a foundation of interactive education that the Kempner is using to help bring together new communities of students and scientists.

In a similar effort, Kempner Research Fellows and postbaccalaureates created and ran a popular lunch-and-learn series. Focused on peer-to-peer teaching and including all levels of non-faculty trainees, the series embodies the Kempner's spirit of open inquiry and collaboration. Providing opportunities for students to gain experience in making presentations and to introduce and explore new subjects, the lunch-and-learns drew more than 340 participants and covered a range of topics from dendritic computation to random matrix theory to how humans learn language.

Last year's Kempner Seminar Series hosted 21 research-level seminars that drew approximately 2,600 participants for standing-room only events. Recordings have been viewed on YouTube more than 13,600 times. Highlights from the series, which featured talks at the forefront of AI, neuroscience, and their intersection, included Tri Dao's discussion of efficient state space models and attention variants for faster, inference-aware architectures; Ruslan Salakhutdinov's exploration of multimodal AI agents capable of reasoning, planning, and autonomous action; Paul Cisek's evolutionary reframing of brain function as feedback control rather than information processing; and Ila Fiete's presentation of a computational model of associative memory in the hippocampus. 2026's series is coming together, with 19 experts scheduled to discuss their latest work in natural and artificial intelligence.

MORE THAN
3.2K
EVENT ATTENDEES

MORE THAN
24K
VIEWS OF EVENT RECORDINGS

Karel Svoboda, vice president and executive director of the Allen Institute for Neural Dynamics, presented an expert talk at the Kempner Institute's Frontiers in NeuroAI symposium on June 5–6, 2025.

Unparalleled Resources

The Kempner is focused on providing access to computational resources that are unmatched in academic environments in the US and around the world—and on training and supporting community members to use those resources effectively. This year members ran more than 750,000 jobs totaling more than 3 million GPU hours. Access to this technology is critical for state-of-the-art research. Based on utilization rates consistently above 95%, the institute recently committed to more than doubling the size of the cluster in early 2026. With the new hardware in place, the Kempner AI cluster will include approximately 1,100 GPUs—a mix of H200, H100, A100, and RTX Pro 6000 Blackwell architectures—providing a theoretical peak performance of 1.8 ExaFLOPS (BF16).

COMPUTING

750K COMPUTE JOBS

HANDBOOK ACCESSED MORE THAN **124K** TIMES

FOR MORE THAN **3M** GPU HOURS

1,100 GPUs IN 2026

To help users take advantage of this computational hardware, the Kempner research and engineering team has continued to provide tools, resources, and training, including the Kempner Computing Handbook. The handbook has been accessed more than 124,000 times by 5,600 active users from 101 countries since its redesign in February 2024. The guidance in this open-source handbook, intended to provide the foundational knowledge and practical insights to effectively utilize the high-performing computing available at the Kempner Institute, has proven valuable to researchers around the globe.

Left: Some members of the Kempner's research and engineering team at this year's opening reception. The team has continued to provide tools, resources, and training, including the Kempner Computing Handbook.
Right: Co-Director Bernardo Sabatini, Executive Director Elise Porter, and Co-Director Sham Kakade, pose in front of the Kempner Institute sign at the entrance to Harvard's Science and Engineering Complex.

APPENDIX

INSTITUTE INVESTIGATORS

New Institute Investigators

SUEYEON CHUNG

Investigator, Kempner Institute for the Study of Natural and Artificial Intelligence
Assistant Professor of Physics and of Applied Mathematics, Faculty of Arts and Sciences

Chung's research explores the fundamental principles of neural computation in biological and artificial neural networks. Her work integrates ideas from neuroscience, ML, and statistical physics to understand how neural systems encode, transform, and process information. She develops theoretical frameworks to model the geometry and dynamics of neural population activity, linking emergent structures in high-dimensional systems to the computations they support. In parallel, she builds neural network models with biologically inspired architectures and learning rules, using them both as tools for neuroscience and as a path to more interpretable and efficient AI.

QIANQIAN WANG

Investigator, Kempner Institute for the Study of Natural and Artificial Intelligence
Assistant Professor of Computer Science, Harvard John A. Paulson School of Engineering and Applied Sciences

Wang's research explores the future of human vision: how machines can learn to see, understand, and interact with the world like humans do. Her work investigates a variety of topics at the forefront of computer vision, including long-form video understanding, visual reasoning, spatial and temporal memory, 3D scene perception, and active visual perception.

Institute Investigator Annual Reports

MICHAEL ALBERGO

Investigator, Kempner Institute for the Study of Natural and Artificial Intelligence
Assistant Professor of Applied Mathematics, Harvard John A. Paulson School of Engineering and Applied Sciences

Albergo studies methods in ML and numerical analysis to accelerate discovery in the physical sciences and the study of complex systems. His past year has focused on the theory and application of generative models built out of dynamical transport of measure, via flow and diffusion models, across data modalities. These tools have become state-of-the-art in continuous domains like protein design and image and video generation. He has been working to build a unifying mathematics of these generative methods, and then to use these new tools to develop more data efficient, scalable, and cheaper generative models.

KIANTÉ BRANTLEY

Investigator, Kempner Institute for the Study of Natural and Artificial Intelligence
Assistant Professor of Computer Science, Harvard John A. Paulson School of Engineering and Applied Sciences

Brantley's research focuses on addressing the misalignment issues that ML systems based on foundation models often encounter. He aims to mitigate these issues by studying algorithms that can learn from interactive feedback data collected from external sources. He has focused on addressing inefficiencies to make RL more practical for reasoning-focused LLMs. Specifically, he has developed a set of algorithms targeting both training-time and test-time efficiency. His team also reformulated policy optimization as a regression problem, allowing them to eliminate the need for value networks and significantly reduce memory overhead. The regression-based framework led to several new efficient algorithms for training LLMs on reasoning chains that are simpler and more stable.

YILUN DU

Investigator, Kempner Institute for the Study of Natural and Artificial Intelligence
Assistant Professor of Computer Science, Harvard John A. Paulson School of Engineering and Applied Sciences

Du's research focuses on constructing intelligent embodied agents in the physical world. His recent work has focused on constructing generative models of the world and using these for downstream reasoning and decision-making. His specific areas of interest include reinforcement learning, reasoning, generative modeling, computer vision, robotics, and natural language processing.

KANAKA RAJAN

Investigator, Kempner Institute for the Study of Natural and Artificial Intelligence
Associate Professor, Department of Neurobiology, Harvard Medical School

Rajan's current research focuses on neurotheory, artificial intelligence, and neural networks. The Rajan Lab investigates how animals and humans learn, remember, and decide using neural circuits in their brains. Combining approaches from physics, mathematics, and engineering with data analysis, they aim to discover the neural mechanisms that control our cognitive abilities and behavioral repertoires and identify how they go wrong in neuropsychiatric diseases. This work has included the completion of two novel models for: 1) naturalistic, neural, and behavioral models of foraging behavior that accounts for internal states like hunger and energy depletion, and external states like food scarcity and predation, and 2) POCO, a model for predicting brain activity during both spontaneous and task-based behavior across species and data types.

ASSOCIATE FACULTY

New Associate Faculty

STEPHANIE GIL

John L. Loeb Associate Professor of Engineering and Applied Sciences, Harvard John A. Paulson School of Engineering and Applied Sciences

Gil's research advances foundational understanding of natural and artificial intelligence by developing theory and systems for intelligent multi-agent coordination in uncertain, unstructured, and adversarial environments. She designs principled frameworks that enable physically embodied agents—such as robots—to reason, perceive, communicate, and learn in secure and coordinated ways. Her work explores how distributed agents can construct situational awareness, form trust, and make decisions in dynamic, partially observable settings. This includes developing resilient algorithms for consensus, control, and learning that adapt to uncertainty and adversarial behavior, as well as integrating reinforcement learning with trust modeling for real-time decision-making.

VENKATESH MURTHY

Raymond Leo Erikson Life Sciences Professor of Molecular and Cellular Biology, Faculty of Arts and Sciences
Paul J. Finnegan Family Director, Center for Brain Science

Murthy's current research focuses on neural circuits, algorithms, and learning. His lab is interested in understanding the neural and algorithmic basis of complex sensory-guided behaviors in terrestrial animals. To this end, it has developed behavioral tasks in mice using stimuli and situations that approximate natural settings while allowing electrophysiological recordings, high-resolution optical imaging, and optogenetic manipulation. The lab records neural activity in behaving mice using electro- or opto-physiological methods and relates them to behavioral features, attempting to discern the computational algorithms underlying these behaviors.

Associate Faculty Annual Reports

DAVID ALVAREZ-MELIS

Assistant Professor of Computer Science, Harvard John A. Paulson School of Engineering and Applied Sciences

Alvarez-Melis's lab researches fundamental principles of artificial intelligence, with a focus on studying ML in constrained, dynamic, and multimodal settings. His lab is particularly interested in developing deep learning models that can learn and adapt effectively in these challenging scenarios. This year he examined the emergence of capabilities in scaling regimes through a distributional lens, revealing how variance, not just mean performance, signal emergent behaviors before they appear predictably. He further explored model reasoning limits via sequential search, showing how local decision-making can inhibit global problem-solving, and proposed alternatives grounded in backtracking and decomposition.

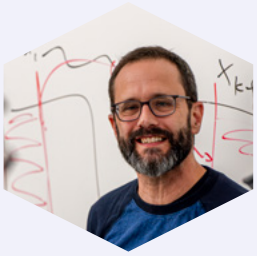
DEMBA BA

Gordon McKay Professor of Electrical Engineering, Harvard John A. Paulson School of Engineering and Applied Sciences

Ba's current research focuses on computational neuroscience, model-based deep learning, and signal processing. His recent research has examined the connection between sparse, structured signal representations and ANNs. The opacity of modern AI systems has limited their adoption in a number of domains where they could significantly increase the productivity of humans. This year his research has brought to the forefront the limitations of sparse auto encoders and the need for their design based on domain knowledge, and it has promoted new alternatives, bringing us a step closer to understanding intelligence in artificial systems.

BOAZ BARAK

Catalyst Professor of Computer Science, Harvard John A. Paulson School of Engineering and Applied Sciences



Barak's current research focuses on the foundations of ML, and understanding the mechanisms, capabilities, and limitations of deep learning systems. He is particularly interested in how performance depends on resources, including computation, data, memory, and more. He is co-leading the ML foundations group, studying phenomena in deep learning such as optimization, representation, and uncertainty.

SAMUEL GERSHMAN

Professor of Psychology, Faculty of Arts and Sciences

Gershman's current research focuses on cognitive science, computational neuroscience, and reinforcement learning. His lab has been studying parallels between natural and artificial intelligence, particularly in the domain of memory. They have developed a theory of key-value memory in the brain, a model of complementary memory systems in transformers, neuroimaging evidence for memory-based reinforcement learning in humans, and behavioral evidence for the role of memory constraints in human decision-making based on modern policy optimization algorithms.

TALIA KONKLE

Professor of Psychology, Faculty of Arts and Sciences

Konkle's research has focused on bridging artificial and biological vision systems through novel computational methods and alignment metrics, working at the intersection of AI interpretability and cognitive neuroscience. Her team has made contributions on pure deep neural network interpretability related to sparse circuit extraction and concept learning with sparse autoencoders. Her lab has also developed new AI architectures, drawing on principles of foveated human vision. Together, these contributions establish new frameworks for creating AI systems that are not only more interpretable and human-aligned but also grounded in the computational principles underlying biological vision, with implications for both understanding natural intelligence and building more robust artificial systems.

SUSAN MURPHY

Mallinckrodt Professor of Statistics and of Computer Science, Harvard John A. Paulson School of Engineering and Applied Sciences

Murphy's lab works on autonomous online learning and personalization of sequences of decisions in digital interventions. In most cases the state is partially unobserved, and the lab uses predictions of these states. They are developing methods for using uncertainty in these predictions in the RL algorithm and also worked on continual learning in which periods of online learning/decision-making are interspersed with knowledge consolidation based on the data resulting from the prior online learning period. They developed methods that enable accurate assessment of confidence in analyses of data collected during online learning.

CENGIZ PEHLEVAN

Assistant Professor of Applied Mathematics, Harvard John A. Paulson School of Engineering and Applied Sciences

In collaboration with AI company Cerebras, Pehlevan introduced CompleteP, a parameterization that preserves depth-wise transfer while avoiding lazy learning, yielding compute-efficiency gains for deep transformers. Additionally, his lab analyzed finite precision vs. scale, showed how feature learning reshapes scaling exponents in a solvable model, characterized infinite-width/depth limits that retain feature learning in multi-head transformers, and derived adaptive kernel predictors from feature-learning infinite limits. In neuroscience, his lab showed how partial observation can induce mechanistic mismatches in data-constrained models of neural dynamics, analyzed the tempo of recall in Hebbian sequence memory set by the interaction of the Hebbian kernel and tutor timing, and proposed a model of place-field reorganization driven by reward maximization.

PATRICK SLADE

Assistant Professor of Bioengineering, Harvard John A. Paulson School of Engineering and Applied Sciences

The Slade Lab combines biomechanics and human-centered AI to understand and improve mobility. His lab has been working on two projects: 1) experimental studies to reveal the cognitive mechanisms of natural intelligence, and 2) new AI methods to model natural cognition abilities. First, the lab focused on understanding how cognitive objectives of movement can be captured with neural signals to attempt to better understand this natural intelligence in motor control. They are developing AI models to relate motion data to these neural signals to understand if AI can capture some of this motor control structure and underlying relationships. Second, they are exploring a collaboration with Kempner Affiliate Faculty Bence P. Ölveczky, using their training of an AI controller to generate realistic movements and then correlating this controller with neural data to understand how the AI model may capture relevant similar structure to the neural data.

HAIM SOMPOLINSKY

Professor of Molecular and Cellular Biology and of Physics, Faculty of Arts and Sciences

The Sompolinsky Lab has developed and unified theoretical frameworks that bridge natural and artificial intelligence by advancing our understanding of the geometry, dynamics, and scaling laws of neural representations and learning. Work from the lab has established cross-domain theoretical foundations spanning representation geometry, physics-inspired learning theories, and architecture-aware performance limits. This advances our understanding of how biological and artificial systems encode, learn, and generalize intelligence.

MARINKA ZITNIK

Assistant Professor of Biomedical Informatics, Harvard Medical School

Zitnik's research currently focuses on geometric deep learning, multimodal learning, knowledge graphs, foundation models, and biomedical AI. This year her lab advanced multimodal and agentic AI that links learning in artificial systems to mechanisms in natural biological systems. A centerpiece was ProCyon, a multimodal foundation model that learns and generates protein phenotypes from interleaved protein/text/chemical inputs. Her lab released TxAgent, a therapeutic-reasoning agent that performs tool-grounded, multi-step reasoning across 200+ biomedical tools with evidence verification. They introduced ATOMICA, which learns atomic-scale representations of intermolecular interfaces across biomolecular modalities, and MADRIGAL, which integrates structural, pathway, cell-viability, and transcriptomic readouts to predict clinical outcomes of drug combinations while handling missing modalities. The lab advanced COMPASS for interpretable immunotherapy prediction, MedTok for multimodal tokenization of medical codes, KGARevion for knowledge-graph-grounded biomedical QA, and SPATIA for prediction and generation of spatial cell phenotypes.

RESEARCH FELLOWS

New Research Fellows

ELOM AMEMATSRO

Amematsro's research focuses on uncovering the neural and computational principles that allow humans to rapidly learn new skills by leveraging relationships to previously learned ones. He aims to identify the brain circuits that encode these relational structures and support the flexible recombination of existing skills into novel behaviors. To achieve this, his work integrates behavioral experiments, large-scale neurophysiology, and computational modeling, with an emphasis on RNNs and hierarchical learning systems. By bridging systems neuroscience with ML, he seeks to develop models that capture the compositional nature of human learning and apply them to artificial networks, enabling more efficient generalization and adaptability. In the long term, his goal is to use these insights not only to advance artificial intelligence but also to inform clinical approaches for restoring cognition in individuals with neurological and psychiatric disorders.

RUOJIN CAI

Cai's research focuses on 3D computer vision, with the goal of building models that can perceive and reason about the 3D world in order to advance spatial intelligence in machines. She studies core challenges in 3D reconstruction under sparse-view or ambiguous settings, where traditional geometric methods often fail. Her key insight is to address these challenges by leveraging learned priors from generative video models and geometric vision models to improve robustness under limited or ambiguous observations. Building on this, she aims to advance 3D foundation models and integrate reasoning into vision tasks, with potential applications in robotics and embodied AI. Her long-term goal is to develop truly spatially intelligent systems that can not only perceive but also reason about and act within complex, real-world environments.

DAVID CLARK

Clark is interested in theoretical and conceptual questions about how neural circuits give rise to brain function. Neural circuits present essential features that demand theoretical approaches: they contain vast numbers of neurons, exhibit nonlinear dynamics, involve complex recurrent interactions, and undergo connectivity changes across multiple timescales. Clark studies abstract models that capture these features and develops analytical theories drawing on tools from statistical physics and ML to link connectivity with emergent dynamics. His goal is to understand not only how neural circuits operate but also, more challengingly, how these features together support computation and learning.

ALEX DAMIAN

Damian's research examines on the mathematical foundations of deep learning, with a focus on two interconnected areas: optimization dynamics and representation learning. His work seeks to understand how optimization algorithms, including stochastic gradient descent and Adam, navigate the complex, high-dimensional loss landscapes that emerge during neural network training. Central to his research are three fundamental questions: What types of features can neural networks efficiently learn? How much data do they need? And how do optimization choices, including the learning rate, batch size, and momentum, shape both the training dynamics and the resulting learned representations?

WILLIAM DORRELL

Dorrell tries to understand how biological neurons implement cognitive computations. His approach to this involves asking why neurons fire the way they do and building mathematical theories to try and understand this. These mathematical theories are usually optimization problems, leading to hypotheses like: "If the neurons were trying to perform this computation optimally then, under some constraints, they should behave like this." Dorrell compares the predictions of these theories to neural recordings from brains or ANNs. He hopes these approaches will help in understanding the algorithms the brain uses to do clever things like play board games, tap rhythms, or reason.

RICHARD HAKIM

Hakim studies neural decoding and brain-computer-interfaces (BCIs). His prior work is primarily experimental and includes studies on how movement is encoded in the motor cortex, how brain oscillations are generated, and the development of a suite of open-source computational tools. Moving forward, he is excited by emerging research into foundation models for BCI decoding and aims to leverage principles derived from AI to better understand the structure and function of biological brains.

HADAS ORGAD

Orgad investigates the internal mechanisms of AI models to better understand and mitigate failures in safety, fairness, and reliability. Her research bridges interpretability and practical deployment, focusing on harmful model behaviors such as hallucinations, bias, privacy violations, and unsafe outputs. By analyzing the internal structure of models, she develops actionable tools and interventions to improve model behavior and better align it with human values and incentives. Her long-term goal is to advance interpretability and control techniques so that AI systems are fully transparent, trustworthy, and steerable.

GIZEM OZDIL

Ozdil bridges systems neuroscience, artificial intelligence, and robotics to uncover the principles that enable adaptive behavior in biological systems. She is particularly interested in how biological insights, such as structural constraints, can inform the design of more flexible and autonomous agents. To explore this, she develops biologically inspired neural networks and trains embodied agents in complex physical environments that require learning, memory retention, and planning. In turn, these models can be used for reverse-engineering brain function and inspiring the development of more efficient and adaptable artificial systems.

GABRIEL POESIA REIS E SILVA

Silva's research is centered around building self-improving machines that are capable of formal reasoning, which includes proving mathematical theorems, conjecturing, decomposing problems, and developing increasingly higher-level abstractions over time. This goal has involved interfacing ideas from type theory (to define a game of formal theorem proving), reinforcement learning (to become steadily better at playing this game), language models (to represent policies, value functions, and leverage informal reasoning), program induction (to discover lemmas, create new tactics, or invent useful definitions), and the whole toolbox from game-playing AI, such as tree search and self-play.

GRETA TUCKUTE

Tuckute studies how language is processed in the human brain and in ANNs. Her research broadly follows three directions. First, she works to precisely characterize the neural architecture and functions that support language processing in the human brain. Second, she investigates whether the human brain and artificial networks share representations and computational principles during language processing. Third, she develops biologically inspired artificial networks that learn language in more humanlike ways. Collectively, these three directions inform one another, advancing our understanding of how language serves as an efficient interface to a wide range of downstream behaviors in both biological and artificial systems.

Research Fellows Annual Reports

DAVID BRANDFONBRENER

Brandfonbrener's research focused on understanding how dataset composition impacts learning in language models and takes an important step toward understanding AI. Essentially, it is important to understand how the data that is fed into these learning algorithms impacts the resulting models. In his scaling laws project, he found smooth and predictable relationships between learning performance as he changed the underlying datasets, suggesting some unifying trends across substantially different input data. *Brandfonbrener completed the fellowship and has left for a position at Meta.*

WILKA CARVALHO

Carvalho has been developing a new framework called "naturalistic computational cognitive science," which brings cognitive science together with cutting-edge AI tools. The framework was recently laid out in a series of three papers. The first made the case for using AI and ML techniques to help build models of human cognition in more realistic settings. The second introduced an AI tool called "NiceWebRL," which makes it easier to use artificial reinforcement learning environments in studies on human subjects. The third paper illustrated the general framework, presenting a model of how humans generalize to new tasks and showing that the model accurately predicts human performance in simple 2D grid-worlds.

THOMAS FEL

Fel's research focuses on large vision models, particularly their explainability. Motivated to uncover the secrets behind their exceptional ability to generalize, Fel blends computational techniques with insights from neuroscience to better grasp the inner workings of these models. This interdisciplinary approach not only aims to enrich the understanding of AI but also positions this knowledge as a tool for probing human intelligence. He's developed tools to analyze how visual models, particularly large-scale self-supervised systems, structure semantic and geometric information. These tools aim to generate theoretical insights and practical interpretability for vision systems in ML.

JENNIFER HU

Hu's work over the past academic year leveraged theories and methods from cognitive science to develop a better understanding of LM's cognitive abilities. For example, she investigated LM's abilities to comprehend challenging semantic stimuli compared to humans and to introspect about their grammatical knowledge. She has begun another line of research to compare real-time processing strategies in humans with layer-time computation dynamics in LMs. She is exploring theoretical questions such as how string probabilities can be used to measure a model's underlying grammatical generalizations. Overall, this work contributes to our understanding of LM's abilities and limitations, and it leverages computational models to test theories about human cognition. *Hu completed the fellowship and has left for a faculty (assistant professor) position at Johns Hopkins University.*

ILENNA JONES

Jones's current research focuses to uncovering the role of subcellular properties on network computation in the "Functional Clustering" project. She is collaborating with the Sabatini Lab to use their "Dendrinet" codebase of optimizable dendritically-detailed neural networks. She investigates how synapse locations can be learned to cluster on dendritic branches, and how these functional clusters interact with ion-channel-based dendritic nonlinearities to allow for effective task performance. Her "Hessian Pruning" project uses new high-dimensional neuron simulation methods to take overparameterized models and prune them down to simpler model structures, removing ion channels that have limited impact on model fit to neural data.

T. ANDERSON KELLER

Keller's research explores ways to develop deep probabilistic generative models that are meaningfully structured with respect to observed, real-world transformations. Such structure permits both improved generalization in previously unobserved settings and reduced sample complexity on natural tasks, thereby addressing two of the fundamental limitations of modern deep neural networks. The goal of the research is to understand the abstract mechanisms underlying the apparent sample efficiency and generalizability of natural intelligence, and then integrate these into artificially intelligent systems.

BINGBIN LIU

Liu's research focuses on two themes: optimizer design and data usage. For the optimizer theme, she compares two commonly used methods, namely Adam and Gauss-Newton methods, in different settings. Her results indicate that Adam is more robust and more adaptive to heterogeneous learning. For the data usage theme, she studies how data influences both pretraining and post-training. The main question for pretraining is whether model performance can be enhanced by reusing data, which is motivated by the practical concern of data scarcity. She uses simple synthetic tasks such as sparse parity and modular addition to facilitate mechanistic understanding.

ERAN MALACH

Malach's research focused on the theory and experimental science of deep learning, particularly the recent advancement in LLMs. He worked on uncovering the principles of how language models learn to reason from the architecture, data, and algorithms perspective. He has written academic papers on topics such as the benefits and limitations of the Mixture-of-Experts architecture, the power of RL in improving mathematical reasoning capabilities, and the effect of different test-time scaling strategies. *Malach completed the fellowship and has left for a position at Apple.*

ISABEL PAPADIMITRIOU

Papadimitriou has worked on model interpretability, understanding the ways that language models structure and come to learn their language system from next-word-prediction training. She collaborated with researchers in vision in order to examine the interactions of visual and language processing in large models. She also worked on training dynamics, examining how models arrive at coherent syntactic systems over training. *Papadimitriou completed the fellowship and has left for a faculty (assistant professor) position at The University of British Columbia.*

NAOMI SAPHRA

Saphra's current research focuses on NLP, training dynamics, and interpretability. She is interested in how models learn to encode linguistic patterns or other structure, how random variation and other factors influence learning, and how we can encode useful inductive biases into the training process. She has recently has become interested in fish. The unified theme is focusing on capabilities where the model knows it or not, that is, capabilities where there is a clear jump from poor to good performance in terms of time, parameter scale, or random cluster. These capabilities can include behaviors like compositional generalization.

NOOR SAJID

Sajid aims to imbue artificial agents with the adaptability seen in biological intelligence, enabling these systems to apply learned knowledge to a variety of tasks within small-scale training data regimes. Accordingly, her research is grounded in developing generative models for understanding and mimicking biological decision-making to investigate how artificial agents can adapt to and learn from environmental perturbations, similar to humans and animals. Going forward, she plans to expand these models and delve deeper into the dynamics that facilitate the flexible utilization of acquired information. Her research aims to reverse-engineer principles of functional resilience in biological systems and to build models that are robust to internal perturbations.

BINXU WANG

Wang's research has advanced at the interface of generative AI and neuroscience through two primary lines of work. On the generative-modeling side, she investigated the nature of creativity in generative diffusion networks. Theoretically, she unveiled a dominant linear substructure within diffusion processes that enables intuitions and a novel analytical teleportation sampler to skip 15-30% of sampling steps while preserving sample fidelity. She further analyzed the learning dynamics of linear diffusion models, discovering a pronounced spectral bias: low-variance modes require orders-of-magnitude longer to converge. Empirically, she designed a Raven Progressive Matrices-inspired dataset and showed that diffusion models can generate rule-conforming novel samples yet struggle to extrapolate to unseen rules. She also demonstrated that rule-learning capabilities emerge sharply beyond a critical dataset-to-parameter ratio. Building on key failures in text-to-image generation, Wang created a synthetic object-relation benchmark and applied mechanistic interpretability to reveal attention circuits underlying spatial reasoning.

RESEARCH & ENGINEERING TEAM

NIKHIL ANAND

Senior ML Research Scientist

Anand completed his PhD in physics in 2018 at Johns Hopkins University, specializing in theoretical and numerical methods to understand the dynamics of strongly coupled quantum mechanical systems. He completed a Simons Foundation postdoctoral fellowship in 2021 in theoretical physics and was a visiting researcher at Mila. Previously, he was a research scientist at Amazon, where he developed methods to select high-quality training data for LLMs and improved models' capabilities in API and tool usage. He has also worked on operational aspects of deploying production language models, and he drove multiple applied research science initiatives to improve user experience. Anand is interested in better understanding how large foundation models work and in making them more capable and efficient.

BALA DESINGHU

Senior AI Research Computing Engineer

Desinghu is a research computing and data professional with expertise in various cyberinfrastructures, including grid, cloud, High-Performance Computing (HPC), and national supercomputing resources. Previously, he worked at Rutgers University and the University of Chicago, where he collaborated on and contributed to diverse computational projects, He focuses on applying HPC to support AI research and facilitates research computing and collaborations while providing the computational power necessary to tackle complex AI challenges. He prepares the HPC platform to develop, fine-tune, and deploy a wide range of AI models, including those dealing with large-scale compute and data in image, text, speech, spatial, and multimodal formats. In addition to supporting HPC infrastructure services, he works with the team to create training and outreach activities.

NAEEM KHOSHNEVIS

Senior ML Research Engineer

Khoshnevis has a solid foundation in mathematics and statistics, enhanced by over a decade of expertise in HPC and more than three years in open-source software development, focusing on ML and statistical tools. Prior to his current role, he served as a senior research software engineer at the National Studies on Air Pollution and Health at the Harvard T.H. Chan School of Public Health and the Edge Computing Lab at SEAS. Khoshnevis has extensive experience with parallelization mechanisms, including shared and distributed memory systems. His work emphasizes large-scale ML pipeline implementation and adherence to software engineering best practices. He leads the adoption of MLOps practices to automate and refine the lifecycle of ML models, enhancing scalability, efficiency, and maintainability. He also oversees architectural decisions for ML systems, contributes to grant proposals and technical reports, and mentors junior staff.

SARAH LEINICKE

Lead Technical Project Manager

Leinicke earned an MA in software engineering and a data science certificate from Harvard Extension School, a JD from American University, and a BA from Smith College. She previously served as a senior research software engineer for the FAS's Research Computing, acting as the lead full-stack software engineer for a neuroscience research web application. Prior to that, she worked at the Software Application & Innovation Lab at Boston University and designed and developed full-stack software solutions to support research projects across a range of domains. She has corporate experience as a systems engineer at Draper Laboratory and supported a big data cybersecurity application at Parsons Corporation. At the Kempner she assembles and coordinates project teams, ensures successful resource allocation and timely project completion, and facilitates communication between engineers, researchers, and other stakeholders. She also oversees the AI engineering internship program.

ABBAS (YASIN) MAZLOUMI

Senior AI/ML GPU Computing Engineer

Mazloumi has a broad background in computer architecture, GPU architecture, and distributed graph AI/analytics. His previous research at the University of California, Riverside, focused on distributed graph analytics, where he developed distributed and scalable high-performance solutions for graph processing by employing resources available on heterogeneous computing clusters. His expertise extends to leveraging the new parallel programming paradigm of running simultaneous graph queries to amortize the distributed computing cost over a batch of queries, effectively tackling complex computational challenges. His work involves providing system-level insights on GPUs, evaluating and benchmarking new accelerators, and researching the design of hardware-efficient algorithms. Additionally, he is exploring how principles of natural intelligence and the brain inspire new AI computing paradigms in both software and hardware.

TIMOTHY NGOTIAOCO

Senior ML Research Engineer

Ngotiaoco has expertise in mathematics and software engineering, and he undertook his graduate work in pure math research in representation theory at MIT. Previously, he worked at two startups, where he designed the infrastructure for managing data pipelines and serving ML models so that they could be used in production environments that served tens of thousands of users and could scale to many more. He also has experience building monitoring systems for benchmarking and tracking performance issues so that developers can more easily identify bottlenecks in their code. His work involves understanding and improving the performance of models as they are scaled across multiple GPUs.

HOUMAN SAFAAI

Senior ML Research Scientist

Safaai earned his PhD in theoretical physics from the International School for Advanced Studies in Trieste, Italy, and continued his postgraduate studies in computational neuroscience. He worked at the Italian Institute of Technology and then joined HMS as a research associate, where he developed probabilistic inference and information estimation tools to study complex network-level neural mechanisms of sensory processing and decision-making. His research focuses on leveraging advanced statistical and ML techniques and models to understand complex neural processes. By bridging theoretical models with empirical data, his research aims to advance our knowledge of brain function and contribute to the development of next-generation AI technologies.

MAX SHAD

Senior Director of AI/ML Research Engineering

Shad leads the research and engineering team at the Kempner Institute, ensuring the provision of advanced research computing tools and services and expert research software engineering (RSE) support. He is responsible for strategic planning for AI HPC cluster compute and storage hardware and networking, collaborating closely with teams at FAS Research Computing and the Massachusetts Green High-Performance Computing Center. Previously, he served as the director of engineering and associate director for research software engineering at Harvard, where he led the establishment of Harvard's first RSE team. With a PhD in mechanical engineering (focusing on computational science and HPC) and a graduate certificate in data science and ML from Harvard, his research interests include AI/ML, complex fluids, HPC, and innovative big data analytics.

PUBLICATIONS AND PREPRINTS

** denotes research that has been discussed in a post on the Kempner Institute Deeper Learning blog*

PAPERS, POSTERS, AND PRESENTATIONS

Abreu, N., Zhang, E., **Malach, E.**, & **Saphra, N**. (2025). A Taxonomy of Transcendence. COLM 2025.

Ahdritz, G., Gollakota, A., Gopalan, P., Peale, C., & Wieder, U. (2025). Provable Uncertainty Decomposition via Higher-Order Calibration. The Thirteenth International Conference on Learning Representations (ICLR).

Ahdritz, G., & **Kleiman, A.** (2025). The SMeL Test: A simple benchmark for media literacy in language models (No. arXiv:2508.02074). arXiv.

Ahmed, Z., Tenenbaum, J. B., Bates, C. J., & **Gershman, S. J.** (2025). Synthesizing world models for bilevel planning. *TMLR*.

Albergo, M. (2025). Learning to Sample on Continuous and Discrete Domains. The Thirteenth International Conference on Learning Representations (ICLR).

Albergo, M. **S.**, & Vanden-Eijnden, E. (2024). Learning to sample better. *Journal of Statistical Mechanics: Theory and Experiment*, *2024*(10), 104014.

Albergo, M. S., & Vanden-Eijnden, E. (2025). NETS: A Non-Equilibrium Transport Sampler (No. arXiv:2410.02711). arXiv.

Altucci, L., Badimon, L., Balligand, J.-L., Baumbach, J., Catapano, A. L., Cheng, F., DeMeo, D., Gupta, R., Hacker, M., Liu, Y.-Y., Loscalzo, J., Maniscalco, S., Menche, J., Menichetti, G., Parini, P., Schmidt, H. H. H. W., & **Zitnik, M.** (2025). Artificial Intelligence and Network Medicine: Path to Precision Medicine. *NEJM AI*, *2*(9), AIra2401229.

Alvarez, G. A. & **Diaz, L. T.** (2025). Ventral Stream Responses to Inanimate Objects are Equally Aligned with AlexNet (2012) and Modern Deep Neural Networks. Annual Meeting of the Computational Cognitive Science Society.

Alvarez, G. A., & Konkle, T. (2024). Decision-margin consistency: A principled metric for human and machine performance alignment. UniReps: 2nd Edition of the Workshop on Unifying Representations in Neural Models.

Alvarez, G. A., Konkle, T., & Luo, K. C. (2025). Representational Geometry Dynamics in Networks After Long-Range Modulatory Feedback. Annual Meeting of the Computational Cognitive Science Society. (Also presented at 2025 Vision Sciences Society Conference).

Alvarez, G. A., Konkle, T., & Saha, S. (2025). Leveraging Vision Transformers to Propose a Context-Dependent Computational Mechanism for the Holistic Process of Faces. Annual Meeting of the Computational Cognitive Science Society.

Alvarez, G. A., Fel, T., Jagadeesh, A. V., **Konkle, T.,** Livingstone, M. S., Lo, E., Prince, J. S., & **Wang, B.** (2025). Parametric control along the encoding axes of IT neurons uncovers hidden differences in model-brain alignment. Annual Meeting of the Computational Cognitive Science Society.

Atanasov, A., Bordelon, B., Zavatone-Veth, J. A., Paquette, C., & **Pehlevan, C.** (2025). Two-Point Deterministic Equivalence for Stochastic Gradient Dynamics in Linear Models (No. arXiv:2502.05074). arXiv.

Atanasov, A., **Meterez, A**., Simon, J. B., & **Pehlevan, C.** (2024). The Optimization Landscape of SGD Across the Feature Learning Strength. The Thirteenth International Conference on Learning Representations (ICLR).

Atanasov, A., Zavatone-Veth, J. A., & **Pehlevan, C.** (2025a). Scaling and renormalization in high-dimensional regression (No. arXiv:2405.00592). arXiv.

Atanasov, A., Zavatone-Veth, J. A., & **Pehlevan, C.** (2025b). Risk and cross validation in ridge regression with correlated samples. Forty-second International Conference on Machine Learning (ICML).

Attias, E., **Pehlevan, C.**, & Obeid, D. (2024). A Brain-Inspired Regularizer for Adversarial Robustness (No. arXiv:2410.03952). arXiv.

Attias, E., **Pehlevan, C.,** & Obeid, D. (2024). Pixel-Based Similarities as Alternative to Neural Data in CNN Regularization Against Adversarial Attacks. Cognitive Computational Neuroscience (CCN).

Avidan, Y., Li, Q., & **Sompolinsky, H** (2025). Unified theoretical framework for wide neural network learning dynamics. *Physical Review E*, *111*(4).

Avidan, Y. & **Sompolinsky, H**. (2025). Langevin Learning Dynamics in Lazy and Non-Lazy Wide Neural Networks. Forty-Second International Conference on Machine Learning (ICML).

Badman, R., Simmons-Edler, R., Berg, F., Lunger, J., Vastola, J., **Qian, W**., & **Rajan, K.** (2025). ForageWorld: RL agents in complex foraging arenas develop internal maps for navigation and planning. COSYNE.

Bari, B. A., & **Gershman, S. J.** (2025). The Value of Non-Instrumental Information in Anxiety: Insights from a Resource-Rational Model of Planning. *Computational Psychiatry*, *9*(1), 63–75.

Bari, B. A., Krystal, A. D., Pizzagalli, D. A., & **Gershman, S. J.** (2025). Computationally Informed Insights Into Anhedonia and Treatment by Kappa Opioid Receptor Antagonism. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*.

Bates, A. S., Phelps, J. S., Kim, M., Yang, H. H., Matsliah, A., Ajabi, Z., Perlman, E., Delgado, K. M., **Osman, M. A. M.**, Salmon, C. K., Gager, J., Silverman, B., Renauld, S., Collie, M. F., Fan, J., Pacheco, D. A., Zhao, Y., Patel, J., Zhang, W., … Lee, W.-C. A. (2025). Distributed control circuits across a brain-and-cord connectome (p. 2025.07.31.667571). bioRxiv.

Behrens, F., Mainali, N., Marullo, C., Lee, S., **Sorscher, B.,** & **Sompolinsky, H.** (2024). Statistical mechanics of deep learning. *Journal of Statistical Mechanics: Theory and Experiment*, *2024*(10), 104007.

Bhalla, U., Oesterling, A., Verdun, C. M., Calmon, F., & Lakkaraju, H. (2025). Leveraging the Sequential Nature of Language for Interpretability. ICML 2025 Workshop on Assessing World Models.

Bhalla, U., Srinivas, S., Ghandeharioun, A., & Lakkaraju, H. (2025). Towards Unifying Interpretability and Control: Evaluation via Intervention (No. arXiv:2411.04430). arXiv.

Bhattacharya, A. R., **Hu, J.**, & Ullman, T. D. (2025). The Uncanny Valley meets the Humorous Hill: Things are funny when they match a pattern but fall short on quality. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *47*(0).

Bhattacharya, A. R., **Papadimitriou, I.**, Davidson, K., & Alvarez-Melis, D. (2025). Investigating the interaction of linguistic and mathematical reasoning in language models using multilingual number puzzles (No. arXiv:2506.13886). arXiv.

Binz, M., Alaniz, S., Roskies, A., Aczel, B., Bergstrom, C. T., Allen, C., Schad, D., Wulff, D., West, J. D., Zhang, Q., Shiffrin, R. M., **Gershman, S. J.**, Popov, V., Bender, E. M., Marelli, M., Botvinick, M. M., Akata, Z., & Schulz, E. (2025). How should the advancement of large language models affect the practice of science? *Proceedings of the National Academy of Sciences*, *122*(5), e2401227121.

Boehmer, N., **Fish, S.**, & Procaccia, A. D. (2025). Generative Social Choice: The Next Generation. ICML.

Boero, L., Wu, H., Zak, J. D., Masset, P., Pashakhanloo, F., Jayakumar, S., Tolooshams, B., **Ba, D.,** & **Murthy, V. N.** (2025). Perception and neural representation of intermittent odor stimuli in mice (p. 2025.02.12.637969). bioRxiv.

Boffi, N. M., **Albergo, M. S.**, & Vanden-Eijnden, E. (2024). Flow map matching with stochastic interpolants: A mathematical framework for consistency models. *Transactions on Machine Learning Research (TMLR)*.

Boffi, N. M., **Albergo, M. S.**, & Vanden-Eijnden, E. (2025). How to build a consistency model: Learning flow maps via self-distillation (No. arXiv:2505.18825). arXiv.

Bohacek, M., **Fel, T.**, Agrawala, M., & Lubana, E. S. (2025). Uncovering Conceptual Blindspots in Generative Image Models Using Sparse Autoencoders (No. arXiv:2506.19708). arXiv.

Boissin, T., Mamalet, F., **Fel, T.**, Picard, A. M., Massena, T., & Serrurier, M. (2025). An Adaptive Orthogonal Convolution Scheme for Efficient and Flexible CNN Architectures. *ICML*.

Bordelon, B., Atanasov, A., & **Pehlevan, C.** (2025). How feature learning can improve neural scaling laws. *Journal of Statistical Mechanics: Theory and Experiment*, *2025*(8), 084002.

Bordelon, B., Chaudhry, H., & **Pehlevan, C.** (2024). Infinite Limits of Multi-head Transformer Dynamics. *Advances in Neural Information Processing Systems* (NeurIPS), (Vol. 37, pp. 35824–35878).

Bordelon, B., Cotler, J., **Pehlevan, C.**, & Zavatone-Veth, J. A. (2025). Dynamically Learning to Integrate in Recurrent Neural Networks (No. arXiv:2503.18754). arXiv. (Presented at COSYNE 2025).

Bordelon, B., Kumar, T., **Gershman, S. J.**, & **Pehlevan, C.** (2024). Asymptotic Dynamics for Delayed Feature Learning in a Toy Model. *ICML 2024 Workshop HiLD*. High-dimensional Learning Dynamics 2024: The Emergence of Structure and Reasoning.

Bordelon, B., & **Pehlevan, C.** (2024). Dynamics of finite width Kernel and prediction fluctuations in mean field neural networks. *Journal of Statistical Mechanics: Theory and Experiment*, *2024*(10), 104021.

Bordelon, B., & **Pehlevan, C.** (2025). Deep Linear Network Training Dynamics from Random Initialization: Data, Width, Depth, and Hyperparameter Transfer. Forty-second International Conference on Machine Learning (ICML).

***Brandfonbrener, D.**, **Anand, N.**, Vyas, N., **Malach, E.**, & **Kakade, S.** (2024). Loss-to-Loss Prediction: Scaling Laws for All Datasets. *TMLR*. (Presented at CZI Annual Meeting).

Brandfonbrener, D., Henniger, S., Raja, S., Prasad, T., Loughridge, C., Cassano, F., **Hu, S. R.**, Yang, J., Byrd, W. E., Zinkov, R., & Amin, N. (2024). VerMCTS: Synthesizing Multi-Step Programs using a Verifier, a Large Language Model, and Tree Search. Math-AI workshop at NeurIPS 2024.

Brandfonbrener, D., Zhang, H., Kirsch, A., Schwarz, J. R., & **Kakade, S. M.** (2024). CoLoR-Filter: Conditional Loss Reduction Filtering for Targeted Language Model Pre-training. The Thirty-eighth Annual Conference on Neural Information Processing Systems (NeurIPS).

Brantley, K. (2025). The Power of Resets! Learning better, one reset at a time. Simons Institute.

***Brantley, K.**, Chen, M., Gao, Z., Lee, J. D., Sun, W., Zhan, W., & Zhang, X. (2025). Accelerating RL for LLM Reasoning with Optimal Advantage Regression (No. arXiv:2505.20686). arXiv.

Brenner, J. W., Li, C., & Kreiman, G. (2024). Policy optimization emerges from noisy representation learning (p. 2024.11.01.621621). bioRxiv.

Brodeur, A., Valenta, D., Marcoci, A., Aparicio, J. P., Mikola, D., Barbarioli, B., Alexander, R., Deer, L., Stafford, T., Vilhuber, L., Bensch, G., Goldschmitt, D., Gourdon-Kanhukamwe, A., de Varda, A. G., Grigoryeva, I., Gugushvili, A., Fletcher, A. H. A., Habermann, F., Hablicsek, M., **Fish, S**. … Gibson, G. (2025). Comparing Human-Only, AI-Assisted, and AI-Led Teams on Assessing Research Reproducibility in Quantitative Social Science (Working Paper No. 195). I4R Discussion Paper Series.

Buckley, T. A., Conci, R., Brodeur, P. G., Gusdorf, J., Beltrán, S., Behrouzi, B., Crowe, B., Dockterman, J., Muhammad, M., Ohnigian, S., Sanchez, A., Diao, J. A., Shah, A. P., Restrepo, D., Rosenberg, E. S., Lea, A. S., **Zitnik, M.**, Podolsky, S. H., Kanjee, Z., … Manrai, A. K. (2025). Advancing Medical Artificial Intelligence Using a Century of Cases (No. arXiv:2509.12194; Version 1). arXiv.

Bunne, C., Roohani, Y., Rosen, Y., Gupta, A., Zhang, X., Roed, M., Alexandrov, T., AlQuraishi, M., Brennan,

P., Burkhardt, D. B., Califano, A., Cool, J., Dernburg, A. F., Ewing, K., Fox, E. B., Haury, M., Herr, A. E., Horvitz, E., Hsu, P. D., **Zitnik, M**.... Quake, S. R. (2024). How to build the virtual cell with artificial intelligence: Priorities and opportunities. *Cell*, *187*(25), 7045–7063.

Bussell, J. J., Badman, R. P., Márton, C. D., Bromberg-Martin, E. S., Abbott, L. F., **Rajan, K.**, & Axel, R. (2024). Representations of the intrinsic value of information in mouse orbitofrontal cortex. bioRxiv, 2023.10.13.562291.

Carvalho, W., Goddla, V., Sinha, I., Shin, H., & Jha, K. (2025). NiceWebRL: A Python library for human subject experiments with reinforcement learning environments (No. arXiv:2508.15693). arXiv.

Carvalho, W., Hall-McMaster, S., Lee, H., & **Gershman, S. J.** (2025). Preemptive Solving of Future Problems: Multitask Preplay in Humans and Machines (No. arXiv:2507.05561). arXiv.

Carvalho, W., & Lampinen, A. (2025). Naturalistic Computational Cognitive Science: Towards generalizable models and theories that capture the full range of natural behavior (No. arXiv:2502.20349). arXiv.

Carvalho, W., Tomov, M. S., de Cothi, W., Barry, C., & **Gershman, S. J.** (2024). Predictive Representations: Building Blocks of Intelligence. *Neural Computation*, *36*(11), 2225–2298.

Casto, C., Small, H., Poliak, M., Tuckute, G., Lipkin, B., Wolna, A., D'Mello, A. M., & Fedorenko, E. (2025). The cerebellar components of the human language network (p. 2025.04.14.645351). bioRxiv. (Presented at Cerebellum Gordon Research Conference 2025 and the Annual Meeting of the Cognitive Neuroscience Society 2025).

Chaudhry, H. T., Kulkarni, M., & **Pehlevan, C.** (2025). Test-time scaling meets associative memory: Challenges in subquadratic models. ICLR Workshop: New Frontiers in Associative Memories.

Chauhan, A., Noori, A., Li, Z., He, Y., Li, M. M., **Zitnik, M**., & Das, S. (2024). Multi Scale Graph Neural Network for Alzheimer's Disease (No. arXiv:2411.10720). arXiv.

Choudhary, S., Masset, P., & **Ba, D.** (2024). Self Supervised Dictionary Learning Using Kernel Matching. 2024 IEEE 34th International Workshop on Machine Learning for Signal Processing (MLSP), 1–6.

Choudhary, S., Masset, P., & **Ba, D.** (2025). Implicit Generative Modeling by Kernel Similarity Matching (No. arXiv:2503.00655). arXiv. (Presented at COSYNE 2025).

Clark, D. G., Abbott, L. F., & **Sompolinsky, H.** (2025). Symmetries and Continuous Attractors in Disordered Neural Circuits (p. 2025.01.26.634933). bioRxiv.

Clark, D. G., & **Sompolinsky, H.** (2025). Simplified derivations for high-dimensional convex learning problems (No. arXiv:2412.01110). arXiv.

Cohen, Z., & Drugowitsch, J. (2025). A unifying theory of receptive field heterogeneity predicts hippocampal spatial tuning (p. 2025.07.26.666958). bioRxiv.

Conwell, C., McMahon, E., Jagadeesh, A. V., Vinken, K., Sharma, S., Prince, J. S., **Alvarez, G. A.**, **Konkle, T**., Isik, L., & Livingstone, M. (2024). Monkey See, Model Knew: Large Language Models accurately Predict Human AND Macaque Visual Brain Activity. NeurIPS Workshop on Unifying Representations in Neural Models.

Conwell, C., Prince, J. S., Kay, K. N., **Alvarez, G. A**., & **Konkle, T**. (2024). A large-scale examination of inductive biases shaping high-level visual representation in brains and machines. *Nature Communications*, 15(1), 9383.

Costa, V., **Fel, T.**, Lubana, E. S., Tolooshams, B., & **Ba, D.** (2025a). Evaluating Sparse Autoencoders: From Shallow Design to Matching Pursuit (No. arXiv:2506.05239). arXiv.

Costa, V., **Fel, T.**, Lubana, E. S., Tolooshams, B., & **Ba, D.** (2025b). From Flat to Hierarchical: Extracting Sparse Representations with Matching Pursuit (No. arXiv:2506.03093). arXiv.

Costacurta, J., Duan, Y., Assad, J., **Rajan, K.,** & Linderman, S. (2025). Modeling rapid neuromodulation in the cortex-basal ganglia-

thalamus loop. COSYNE.
Cuesta-Lazaro, C., Bayer, A. E., **Albergo, M. S.**, Mishra-Sharma, S., Modi, C., & Eisenstein, D. J. (2024). Joint cosmological parameter inference and initial condition reconstruction with Stochastic Interpolants. NeurIPS 2024 Workshop: Machine Learning and the Physical Sciences.

Cui, H., **Pehlevan, C.**, & Lu, Y. M. (2025). A precise asymptotic analysis of learning diffusion models: Theory and insights (No. arXiv:2501.03937). arXiv.

Del Castillo, J. C. F., Pashakhanloo, F., **Murthy, V. N**., & Zavatone-Veth, J. A. (2025). Convergent motifs of early olfactory processing are recapitulated by layer-wise efficient coding. *bioRxiv: The Preprint Server for Biology*, 2025.09.03.673748.

Dey, N., Zhang, B. C., Noci, L., Li, M., Bordelon, B., Bergsma, S., **Pehlevan, C.**, Hanin, B., & Hestness, J. (2025). Don't be lazy: CompleteP enables compute-efficient deep transformers (No. arXiv:2505.01618). arXiv.

Doshi, F. R., Fel, T., Konkle, T., & Alvarez, G. A. (2025). Towards Holistic Vision in Deep Neural Networks: Disentangling Local and Global Processing. *Journal of Vision*, *25*(9), 2148-2148.

Doshi, F. R., Fel, T., Konkle, T., & Alvarez, G. (2025). Visual Anagrams Reveal Hidden Differences in Holistic Shape Processing Across Vision Models (No. arXiv:2507.00493). arXiv. (Submitted to NeurIPS 2025).

Doshi, F. R., Konkle, T., & Alvarez, G. A. (2025). A feedforward mechanism for human-like contour integration. *PLOS Computational Biology*, *21*(8), e1013391.

Duan, Y., Chaudhry, H. T., Ahrens, M. B., Harvey, C. D., Perich, M. G., Deisseroth, K., & **Rajan, K.** (2025). POCO: Scalable Neural Forecasting through Population Conditioning (No. arXiv:2506.14957). arXiv.

Dwivedi, R., Tian, K., Tomkins, S., Klasnja, P., **Murphy, S.**, & Shah, D. (2025). Counterfactual inference in sequential experiments (No. arXiv:2202.06891). arXiv.

Edelman, E., Tsilivis, N., Edelman, B. L., **Malach, E.**, & Goel, S. (2024). The Evolution of Statistical Induction Heads: In-Context Learning Markov Chains. Advances in Neural Information Processing Systems (NeurIPS), *37*, 64273–64311.

Ektefaie, Y., Shen, A., Bykova, D., Marin, M. G., **Zitnik, M**., & Farhat, M. (2024). Evaluating generalizability of artificial intelligence models for molecular datasets. *Nature Machine Intelligence*, *6*(12), 1512–1524.

Ektefaie, Y., Shen, A., Jain, L., Farhat, M., & **Zitnik, M.** (2025). Sequence Modeling Is Not Evolutionary Reasoning. *bioRxiv*, 2025.01.17.633626.

Erdogan, M., **Pehlevan, C.**, & Erdogan, A. T. (2025). Error Broadcast and Decorrelation as a Potential Artificial and Natural Learning Mechanism (No. arXiv:2504.11558). arXiv.

*Espinosa-Dice, N., Zhang, Y., Chen, Y., Guo, B., Oertell, O., Swamy, G., **Brantley, K.**, & Sun, W. (2025). *Scaling Offline RL via Efficient and Expressive Shortcut Models* (No. arXiv:2505.22866). arXiv.

Fan, H., Callaway, F., & **Gershman, S.** (2024). Uncertainty-Driven Exploration During Planning. OSF.

***Fang, A.**, Desgagné, M., Zhang, Z., Zhou, A., Loscalzo, J., Pentelute, B. L., & **Zitnik, M.** (2025). *Learning Universal Representations of Intermolecular Interactions with ATOMICA* (p. 2025.04.02.646906). bioRxiv. (Presented at ICML 2025).

Fang, C., & **Rajan, K.** (2025). From Memories to Maps: Mechanisms of In-Context Reinforcement Learning in Transformers (No. arXiv:2506.19686). arXiv.

Farrell, M., & **Pehlevan, C.** (2024). Recall tempo of Hebbian sequences depends on the interplay of Hebbian kernel with tutor signal timing. *Proceedings of the National Academy of Sciences*, *121*(32), e2309876121.

*Fel, T., Lubana, E. S., Prince, J. S., Kowal, M., Boutin, V., **Papadimitriou, I., Wang, B.,** Wattenberg, M., **Ba, D.,** & **Konkle, T.** (2025). Archetypal SAE: Adaptive and Stable Dictionary Learning for Concept Extraction in Large Vision Models. Proceedings of the 42nd International Conference on Machine Learning (ICML), 2025.

Fesser, L., & Weber, M. (2025). Performance Heterogeneity in Graph Neural Networks: Lessons for Architecture Design and Preprocessing (No. arXiv:2503.00547). arXiv.

Finn, E., Keller, T. A., Theodosis, E., & **Ba, D. E.** (2024). Learning Artistic Signatures: Symmetry Discovery and Style Transfer (No. arXiv:2412.04441). arXiv.

Finn, E., Keller, T. A., Theodosis, M., & **Ba, D. E.** (2025). Origins of Creativity in Attention-Based Diffusion Models. ICML Workshop.

Fish, S., Gonczarowski, Y. A., & Shorrer, R. I. (2025). Algorithmic Collusion by Large Language Models (No. arXiv:2404.00806). arXiv. (Presented at 2025 Berkman Klein Center Spring Speaker Series).

Fish, S., Shephard, J., Li, M., Shorrer, R. I., & Gonczarowski, Y. A. (2025). EconEvals: Benchmarks and Litmus Tests for LLM Agents in Unknown Environments (No. arXiv:2503.18825). arXiv. (Presented at the 26th ACM Conference on Economics and Computation).

Fisher, E., **Sajid, N.,** Convertino, L., & Hohwy, J. (2025). Computational Psychiatry and Its Challenges: An Optimistic Outlook. OSF.

Fry, B. R., Russell, N., Fex, V., Mo, B., Pence, N., Beatty, J. A., Manfredsson, F. P., Toth, B. A., Burgess, C. R., **Gershman, S.**, & Johnson, A. W. (2025). Devaluing memories of reward: A case for dopamine. *Communications Biology*, *8*(1), 161.

Gan, Y., Galanti, T., Poggio, T., & **Malach, E.** (2024). On the Power of Decision Trees in Auto-Regressive Language Modeling. Advances in Neural Information Processing Systems (NeurIPS), *37*, 62384–62408.

Gao, D., Lai, H.-Y., Klasnja, P., & **Murphy, S. A.** (2025). Harnessing Causality in Reinforcement Learning With Bagged Decision Times. AISTATS 2025.

Gao, S., **Fang, A.**, Huang, Y., Giunchiglia, V., Noori, A., Schwarz, J. R., Ektefaie, Y., Kondic, J., & **Zitnik, M.** (2024). Empowering biomedical discovery with AI agents. *Cell*, *187*(22), 6125–6151.

Gao, S., Koker, T., Queen, O., Hartvigsen, T., Tsiligkaridis, T., & **Zitnik, M.** (2024). UniTS: A Unified Multi-Task Time Series Model. Advances in Neural Information Processing Systems (NeurIPS), *37*, 140589–140631.

*Gao, S., **Zhu, R.**, Kong, Z., Noori, A., Su, X., Ginder, C., Tsiligkaridis, T., & **Zitnik, M.** (2025). TxAgent: An AI Agent for Therapeutic Reasoning Across a Universe of Tools (No. arXiv:2503.10970). arXiv.

Gao, Z., Chang, J. D., Zhan, W., Oertell, O., Swamy, G., **Brantley, K.**, Joachims, T., Bagnell, J. A., Lee, J. D., & Sun, W. (2024). REBEL: Reinforcement Learning via Regressing Relative Rewards. Advances in Neural Information Processing Systems (NeurIPS), *37*, 52354–52400. (Also presented at 2025 CCC Computing Futures Symposium).

Gao, Z., Zhan, W., Chang, J. D., Swamy, G., **Brantley, K.**, Lee, J. D., & Sun, W. (2025). Regressing the Relative Future: Efficient Policy Optimization for Multi-turn RLHF (No. arXiv:2410.04612). arXiv.

Gazi, A. H., Gao, D., Ghosh, S., Xu, Z., Trella, A., Klasnja, P., & **Murphy, S. A.** (2025). Digital Twins for Just-in-Time Adaptive Interventions (JITAI-Twins): A Framework for Optimizing and Continually Improving JITAIs. JMIR Preprints.

Gazi, A. H., Gullapalli, B. T., Gao, D., Marlin, B. M., Shetty, V., & **Murphy, S. A.** (2025). SigmaScheduling: Uncertainty-Informed Scheduling of Decision Points for Intelligent Mobile Health Interventions (No. arXiv:2507.10798). arXiv.

Gershman, S. J. (2025). Bridging Computation and Representation in Associative Learning. *Computational Brain & Behavior*, *8*(3), 377–391.

Gershman, S. J., Bill, J., & Drugowitsch, J. (2025). Hierarchical Vector Analysis of Visual Motion Perception. *Annual Review of Vision Science*, *11*(Volume 11, 2025), 411–422.

Gershman, S. J., Fiete, I., & Irie, K. (2025). Key-value memory in the brain. *Neuron*, *113*(11), 1694-1707.e1.

Gershman, S. J., & Lak, A. (2025). Policy Complexity Suppresses Dopamine Responses. *Journal of Neuroscience*, *45*(9).

Geuter, J., Bonet, C., Korba, A., & **Alvarez-Melis, D.** (2025). DDEQs: Distributional Deep Equilibrium Models through Wasserstein Gradient Flows. AISTATS 2025.

Geuter, J., Mroueh, Y., & **Alvarez-Melis, D.** (2025). Guided Speculative Inference for Efficient Test-Time Alignment of LLMs (No. arXiv:2506.04118). arXiv. (Presented at ICML 2025).

Gholamzadeh, A., & **Sajid, N.** (2025). Model alignment using inter-modal bridges (No. arXiv:2505.12322). arXiv.

Ghosh, S., Hung, P.-Y., Coughlin, L. N., Bonar, E. E., Guo, Y., Nahum-Shani, I., Walton, M., Newman, M. W., & **Murphy, S. A.** (2025). "It felt more real": Investigating the User Experience of the MiWaves Personalizing JITAI Pilot Study (No. arXiv:2502.17645). arXiv.

Goldberg, A. E., Rakshit, S., **Hu, J.,** & Mahowald, K. (2025). A suite of LMs comprehend puzzle statements as well as humans (No. arXiv:2505.08996). arXiv.

Golowich, N., Jelassi, S., **Brandfonbrener, D., Kakade, S. M., & Malach, E.** (2025). The Role of Sparsity for Length Generalization in Transformers (No. arXiv:2502.16792). arXiv.

Gondhalekar, Y., Hassan, S., **Saphra, N.,** & Andrianomena, S. (2023). Towards out-of-distribution generalization in large-scale astronomical surveys: Robust networks learn similar representations (No. arXiv:2311.18007). arXiv.

Gonzalez, G., Lin, X., Herath, I., Veselkov, K., Bronstein, M., & **Zitnik, M**. (2025a). Combinatorial prediction of therapeutic perturbations using causally inspired neural networks. *Nature Biomedical Engineering*.

Gonzalez, G., Lin, X., Herath, I., Veselkov, K., Bronstein, M., & **Zitnik, M.** (2025b). Combinatorial prediction of therapeutic perturbations using causally-inspired neural networks. bioRxiv, 2024.01.03.573985.

Grimaud, J., Dorrell, W., Jayakumar, S**., Pehlevan, C.,** & Murthy, V. (2024). Bilateral Alignment of Receptive Fields in the Olfactory Cortex. *eNeuro*, *11*(11).

Hakim, R., Heo, G., Jaggi, A., Datta, S. R., Musall, S., & **Sabatini, B. L.** (2025). Spectral envelopes of rhythmic facial movements predict intention and motor cortical representations (p. 2025.09.10.675423). bioRxiv.

Hall-McMaster, S., Tomov, M. S., **Gershman, S. J.**, & Schuck, N. W. (2025). Neural evidence that humans reuse strategies to solve new tasks. *PLOS Biology*, *23*(6), e3003174.

Hall-McMaster, S., Wittkuhn, L., Verra, L., Hedrich, N. L., Irie, K., Dayan, P., **Gershman, S. J.,** & Schuck, N. W. (2025). Entorhinal cortex signals dimensions of past experience that can be generalised in a novel environment (p. 2025.08.01.668096). bioRxiv.

Hamou, N., **Gershman, S. J.,** & Reddy, G. (2025). Reconciling time and prediction error theories of associative learning (p. 2025.01.25.634891). bioRxiv.

Han, S., Pari, J**., Gershman, S. J.,** & Agrawal, P. (2025). General Intelligence Requires Reward-based Pretraining (No. arXiv:2502.19402). arXiv.

He, M., Jiang, C., **Pehlevan, C.,** Murthy, V., Zavatone-Veth, J., & Masset, P. (2025). Simultaneous detection and mapping in the olfactory bulb. COSYNE.

Hidalgo, D., Dellaferrera, G., Xiao, W., Papadopouli, M., Smirnakis, S., & Kreiman, G. (2025). Trial-by-trial inter-areal interactions in visual cortex in the presence or absence of visual stimulation. *eLife*, *14*.

Hidalgo, D., Dellaferrera, G., Xiao, W., Papadopouli, M., Smirnakis, S. M., & Kreiman, G. (2024).

Reliable NonStimulus Driven Signaling in Visual Cortex: Inter-Areal Dynamics Across Species and Conditions. Society for Neuroscience Annual Meeting.

*Hindupur, S. S. R., Lubana, E. S., **Fel, T.,** & **Ba, D.** (2025). Projecting Assumptions: The Duality Between Sparse Autoencoders and Concept Geometry (No. arXiv:2503.01822). arXiv.

Holderrieth, P., **Albergo, M. S**., & Jaakkola, T. (2025). LEAPS: A discrete neural sampler via locally equivariant networks. ICML.

Hosseini, E., **Casto, C**., Zaslavsky, N., Conwell, C., Richardson, M., & Fedorenko, E. (2024). Universality of representation in biological and artificial neural networksv (p. 2024.12.26.629294). bioRxiv.

Hou, K., **Brandfonbrener, D**., **Kakade, S**., **Jelassi, S**., & **Malach, E.** (2024). Universal Length Generalization with Turing Programs (No. arXiv:2407.03310). arXiv.

Hu, J. (2025). Knowing and using pragmatic language. The Conference on Language Models.

Hu, J. (2025). Large language models and human linguistic knowledge. The Annual Meeting of the American Association for the Advancement of Science.

Hu, J., & Frank, M. C. (2024). Auxiliary task demands mask the capabilities of smaller language models. (Presented at the 1st Conference on Language Modeling).

Hu, J., & Franke, M. (2024). Deep and shallow thinking in a single forward pass. NeurIPS 2024 Workshop on Behavioral Machine Learning.

Hu, J., Lepori, M. A., & Franke, M. (2025a). Linking forward-pass dynamics in Transformers and real-time human processing (No. arXiv:2504.14107; Version 1). arXiv.

Hu, J., Lepori, M. A., & Franke, M. (2025b). Signatures of human-like processing in Transformer forward passes (No. arXiv:2504.14107). arXiv.

Hu, J., Sosa, F. A., & Ullman, T. D. (2025a). Making Sense of Nonsense. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *47*(0).

Hu, J., Sosa, F., & Ullman, T. (2025b). Re-evaluating Theory of Mind evaluation in large language models. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *380*(1932), 20230499.

Hu, J., Sosa, F., & Ullman, T. (2025c). Shades of zero: Distinguishing impossibility from inconceivability. *Journal of Memory and Language*, *143*, 104640.

Hu, J., Sosa, F., & Ullman, T. (2025). Making sense of nonsense. The Annual Meeting of the Cognitive Science Society.

Hu, J., Tan, A.W.M., Feng, S., & Frank, M.C. (2025). Language production is harder than comprehension for children and language models. The Annual Meeting of the Cognitive Science Society.

Hu, M. Y., Jain, S., Chaulagain, S., & **Saphra, N.** (2025). How to visualize training dynamics in neural networks. The Fourth Blogpost Track at ICLR 2025.

Huang, A., Singh, S. H., Martinelli, F., & **Rajan, K.** (2025). Measuring and Controlling Solution Degeneracy across Task-Trained Recurrent Neural Networks. *ArXiv*, arXiv:2410.03972v2. (Presented at COSYNE 2025, and RLDM 2025).

Huang, K., Chandak, P., Wang, Q., Havaldar, S., Vaid, A., Leskovec, J., Nadkarni, G. N., Glicksberg, B. S., Gehlenborg, N., & **Zitnik, M.** (2024). A foundation model for clinician-centered drug repurposing. *Nature Medicine*, *30*(12), 3601–3613.

Huang, Y., Su, X., Ullanat, V., Liang, I., Clegg, L., Olabode, D., Ho, N., John, B., Gibbs, M., & **Zitnik, M.** (2025). Multimodal AI predicts clinical outcomes of drug combinations from preclinical data (No. arXiv:2503.02781). arXiv.

Huang, Z., Zhong, S., Zhou, P., Gao, S., **Zitnik, M**., & Lin, L. (2025). A Causality-Aware Paradigm for Evaluating Creativity of Multimodal Large Language Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *47*(5), 3830–3846.

Insanally, M. N., Albanna, B. F., Toth, J., DePasquale, B., Fadaei, S. S., Gupta, T., Lombardi, O., Kuchibhotla, K., **Rajan, K.**, & Froemke, R. C. (2024). Contributions of cortical neuron firing patterns, synaptic connectivity, and plasticity to task performance. *Nature Communications*, *15*(1), 6023.

Irie, K., & **Gershman, S. J.** (2025). Fast weight programming and linear transformers: From machine learning to neurobiology (No. arXiv:2508.08435). arXiv.

Irie, K., Yau, M., & **Gershman, S. J.** (2025). Blending Complementary Memory Systems in Hybrid Quadratic-Linear Transformers (No. arXiv:2506.00744). arXiv.

***Jacobs, M.**, Budzinski, R. C., Muller, L., **Ba, D.**, & **Keller, T. A**. (2025). Traveling Waves Integrate Spatial Information Through Time. ICLR Workshop. (Presented at CCN 2025).

Jaroszewski, A. C., Millner, A. J., **Gershman, S. J.,** Franz, P. J., Bentley, K. H., Kleiman, E. M., & Nock, M. K. (2025). Past suicide attempt is associated with a weaker decision-making bias to actively escape from suicide-related stimuli. *Journal of Psychopathology and Clinical Science*, *134*(5), 503–519.

Jayakumar, S., Rigolli, N., Mathis, M. W., Vergassola, M., Mathis, A., & **Murthy, V. N.** (2025). Mice navigate scent trails using predictive policies (p. 2025.08.27.672631). bioRxiv.

Jelassi, S., **Brandfonbrener, D., Kakade, S. M., & Malach, E.** (2024). Repeat After Me: Transformers are Better than State Space Models at Copying (No. arXiv:2402.01032). arXiv.

Jelassi, S., Mohri, C., **Brandfonbrener, D.,** Gu, A., Vyas, N., **Anand, N.**, **Alvarez-Melis, D.**, Li, Y., **Kakade, S. M.,** & **Malach, E**. (20240. Mixture of Parrots: Experts improve memorization more than reasoning. The Thirteenth International Conference on Learning Representations (ICLR).

Jelassi, S., Mohri, C., **Brandfonbrener, D.**, Gu, A., Vyas, N., **Anand, N.**, **Alvarez-Melis, D.**, Li, Y., **Kakade, S. M., & Malach, E.** (2025). Mixture of Parrots: Experts improve memorization more than reasoning (No. arXiv:2410.19034). arXiv.

Jeon, H., Eftekhar, A., **Walsman, A.**, Zeng, K.-H., Farhadi, A., & Krishna, R. (2025). Convergent Functions, Divergent Forms (No. arXiv:2505.21665). arXiv.

Jha, K., **Carvalho, W.**, Liang, Y., Du, S. S., Kleiman-Weiner, M., & Jaques, N. (2025). Cross-environment Cooperation Enables Zero-shot Multi-agent Coordination. ICML.

Johnson, R., Gottlieb, U., Shaham, G., Eisen, L., Waxman, J., Devons-Sberro, S., Ginder, C. R., Hong, P., Sayeed, R., Su, X., Reis, B. Y., Balicer, R. D., Dagan, N., & **Zitnik, M.** (2025). ClinVec: Unified Embeddings of Clinical Codes Enable Knowledge-Grounded AI in Medicine. *medRxiv*, 2024.12.03.24318322.

Johnson, R., Li, M. M., Noori, A., Queen, O., & **Zitnik, M.** (2024). Graph Artificial Intelligence in Medicine. *Annual Review of Biomedical Data Science*, *7*(Volume 7, 2024), 345–368.

Johnson-Yu, S., Singh, S. H., Pedraja, F., Turcu, D., Sharma, P., **Saphra, N.**, Sawtell, N., & **Rajan, K.** (2024). Understanding biological active sensing behaviors by interpreting learned artificial agent policies. Workshop on Interpretable Policies in Reinforcement Learning @RLC-2024.

Jones, I. S., & Kording, K. P. (2024). Efficient optimization of ODE neuron models using gradient descent (No. arXiv:2407.04025). arXiv.

Kangaslahti, S., Rosenfeld, E., & **Saphra, N.** (2025). Hidden Breakthroughs in Language Model Training. ICML Workshop.

Karchmer, A., & **Malach, E.** (2025). The Power of Random Features and the Limits of Distribution-Free Gradient Descent (No. arXiv:2505.10423). arXiv.

Karine, K., **Murphy, S. A.**, & Marlin, B. M. (2024). BOTS: Batch Bayesian Optimization of Extended Thompson Sampling for Severely Episode-Limited RL Settings (No. arXiv:2412.00308). arXiv.

Karuvally, A., Nowak, F., **Keller, A. T.**, Alonso, C. A., Sejnowski, T. J., & Siegelmann, H. T. (2025). Bridging Expressivity and Scalability with Adaptive Unitary SSMs (No. arXiv:2507.05238). arXiv.

Kasmi, G., Brunetto, A., **Fel, T.**, & Parekh, J. (2025). One Wave To Explain Them All: A Unifying Perspective On Feature Attribution. ICML.

Kauffman, J., Miotto, R., Klang, E., Costa, A., Norgeot, B., **Zitnik, M.,** Khader, S., Wang, F., Nadkarni, G. N., & Glicksberg, B. S. (2025). Embedding Methods for Electronic Health Record Research. *Annual Review of Biomedical Data Science*, *8*, 563–590.

Keller, T. A. (2025) Nu-Wave State Space Models: Traveling waves as a biologically plausible context. COSYNE.

***Keller, T. A.** (2025). Flow Equivariant Recurrent Neural Networks (No. arXiv:2507.14793). arXiv. (Presented at FENS Brain Conference).

Keller, T. A. (2025). Traveling Waves in State Space Models. COSYNE.

Keller, T. A. (2024). Towards the Use of Relative Representations for Lower-Dimensional, Interpretable Model-to-Brain Mappings. Cognitive Computational Neuroscience (CCN).

Keller, T. A., Muller, L., Sejnowski, T. J., & Welling, M. (2024). A Spacetime Perspective on Dynamical Computation in Neural Information Processing Systems (No. arXiv:2409.13669). arXiv.

Kim, J., Cheuk-Kit, L., Domingo-Enrich, C., **Du, Y**., **Kakade, S., Ngotiaoco, T**., Chen, S., & **Albergo, M.** (2025). Any-Order Flexible Length Masked Diffusion (No. arXiv:2509.01025). arXiv.

Kim, J., Shah, K., Kontonis, V., **Kakade, S.,** & Chen, S. (2025). Train for the Worst, Plan for the Best: Understanding Token Ordering in Masked Diffusions. ICML 2025.

Kleiman, A., Dziugaite, G. K., Frankle, J., **Kakade, S.,** & Paul, M. (2025). Soup to go: Mitigating forgetting during continual learning with model averaging (No. arXiv:2501.05559). arXiv.

Kong, Z., Li, Y., Zeng, F., Xin, L., Messica, S., Lin, X., Zhao, P., Kellis, M., Tang, H., & **Zitnik, M.** (2025). Token Reduction Should Go Beyond Efficiency in Generative Models—From Vision, Language to Multimodality (No. arXiv:2505.18227). arXiv.

Kong, Z., Qiu, M., Boesen, J., Lin, X., Yun, S., Chen, T., Kellis, M., & **Zitnik, M.** (2025). SPATIA: Multimodal Model for Prediction and Generation of Spatial Cell Phenotypes (No. arXiv:2507.04704). arXiv.

Kuling, G., & **Zitnik, M.** (2025). Ken Utilization Layer: Hebbian Replay Within a Student's Ken for Adaptive Knowledge Tracing (No. arXiv:2507.00032). arXiv.

Kumar, M. G., Bordelon, B., Zavatone-Veth, J., & **Pehlevan, C**. (2025). A Model of Place Field Reorganization During Reward Maximization. COSYNE.

Kumar, M. G., Bordelon, B., Zavatone-Veth, J. A., & **Pehlevan, C.** (2025). Place Field Representation Learning During Policy Learning. Second Workshop on Representational Alignment at ICLR 2025.

Kumar, M. G., Manoogian, A., **Qian, W**., **Pehlevan, C.,** & Rhoads, S. A. (2025). Neurocomputational Underpinnings of Suboptimal Beliefs in Reinforcement Learning Agents. 8th Annual Conference on Cognitive Computational Neuroscience (CCN).

Kumar, M. G. & **Pehlevan, C.** (2024). Place fields organize along goal trajectory with reinforcement learning. Cognitive Computational Neuroscience (CCN).

Kumar, T., Ankner, Z., Spector, B. F., Bordelon, B., Muennighoff, N., Paul, M., **Pehlevan, C**., Ré, C., & Raghunathan, A. (2024). Scaling Laws for Precision. ICLR.

Kumar, T., Bordelon, B., **Pehlevan, C.,** Murthy, V. N., & **Gershman, S. J.** (2024). Do Mice Grok? Glimpses of Hidden Progress in Sensory Cortex. The Thirteenth International Conference on Learning Representations (ICLR).

Kuo, H., Masset, P., Bordelon, B., & **Pehlevan, C.** (2024). The Learning Hypothesis on Spatial Receptive Field Remapping. Cognitive Computational Neuroscience (CCN).

Lange, D., Sui, P., Gao, S., **Zitnik, M**., & Gehlenborg, N. (2025). DQVis Dataset: Natural Language to Biomedical Visualization. Sciety.

Lauditi, C., Bordelon, B., & **Pehlevan, C.** (2025a). Transfer Learning in Infinite Width Feature Learning Networks (No. arXiv:2507.04448). arXiv.

Lauditi, C., Bordelon, B., & **Pehlevan, C.** (2025b). Adaptive kernel predictors from feature-learning infinite limits of neural networks. Forty-second International Conference on Machine Learning (ICML).

Lee, C. K., Jeha, P., Frellsen, J., Lio, P., **Albergo, M. S.,** & Vargas, F. (2025). Debiasing Guidance for Discrete Diffusion with Sequential Monte Carlo. ICLR Workshop.

Li, J., **Fang, A**., Smyrnis, G., Ivgi, M., Jordan, M., Gadre, S., Bansal, H., Guha, E., Keh, S., Arora, K., Garg, S., Xin, R., Muennighoff, N., Heckel, R., Mercat, J., Chen, M., Gururangan, S., Wortsman, M., Albalak, A., … Shankar, V. (2024). DataComp-LM: In search of the next generation of training sets for language models. *Advances in Neural Information Processing Systems (NeurIPS)*, *37*, 14200–14282.

Li, K., & **Zitnik, M.** (2025). Prompting Decision Transformers for Zero-Shot Reach-Avoid Policies (No. arXiv:2505.19337). arXiv.

Li, M., Arroyo, A. M. C., Rogers, G., **Saphra, N**., & Wallace, B. C. (2025). Do Natural Language Descriptions of Model Activations Convey Privileged Information? (No. arXiv:2509.13316). arXiv.

Li, M. M., Huang, Y., Sumathipala, M., Liang, M. Q., Valdeolivas, A., Ananthakrishnan, A. N., Liao, K., Marbach, D., & **Zitnik, M**. (2024). Contextual AI models for single-cell protein biology. *Nature Methods*, *21*(8), 1546–1557.

Li, M. M., **Li, K.,** Ektefaie, Y., Jin, Y., Huang, Y., Messica, S., Cai, T., & **Zitnik, M.** (2025). Controllable Sequence Editing for Biological and Clinical Trajectories (No. arXiv:2502.03569). arXiv.

Li, M. M., Reis, B. Y., Rodman, A., Cai, T., Dagan, N., Balicer, R. D., Loscalzo, J., Kohane, I. S., & **Zitnik, M.** (2025). One Patient, Many Contexts: Scaling Medical AI Through Contextual Intelligence (No. arXiv:2506.10157). arXiv.

Li, Q., **Sorscher, B**., & **Sompolinsky, H.** (2024). Representations and generalization in artificial and brain neural networks. *Proceedings of the National Academy of Sciences*, *121*(27), e2311805121.

*Li, S., Wang, W., Knipe, G., Jerng, E., Capelli, P., Zhou, C., Bilsel, E., & **Sabatini, B. L.** (2025). Synaptic sign switching mediates online dopamine updates (p. 2025.07.23.666367). bioRxiv.

Li, S., **Zhang, Y.**, Ren, Z., Liang, C., Li, N., & Shah, J. A. (2024). Enhancing Preference-based Linear Bandits via Human Response Time. *Advances in Neural Information Processing Systems (NeurIPS)*, *37*, 16852–16893.

Li, V. R., Chen, Y., & **Saphra, N.** (2025). ChatGPT Doesn't Trust Chargers Fans: Guardrail Sensitivity in Context. *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*.

Li, V. R., Kaufmann, J., **Alvarez-Melis, D**., & **Saphra, N.** (2024). Twin Studies of Factors in OOD Generalization. NeurIPS Workshop on Scientific Methods for Understanding Deep Learning.

Li, V. R., Kaufmann, J., Wattenberg, M., **Alvarez-Melis, D.,** & **Saphra, N.** (2025). Can Interpretation Predict Behavior on Unseen Data? NeurIPS Workshop.

Li, W., Li, X., Lavallee, E., Saparov, A., **Zitnik, M.,** & Cassa, C. (2025). From Text to Translation: Using Language Models to Prioritize Variants for Clinical Review (p. 2024.12.31.24319792). medRxiv.

Li, X., Loscalzo, J., Mahmud, A. K. M. F., Aly, D. M., Rzhetsky, A., **Zitnik, M.,** & Benson, M. (2025). Digital twins as global learning health and disease models for preventive and personalized medicine. *Genome Medicine*, *17*(1), 11.

Liboni, L. H. B., Budzinski, R. C., Busch, A. N., Löwe, S., **Keller, T. A**., Welling, M., & Muller, L. E. (2025). Image segmentation with traveling waves in an exactly solvable recurrent neural network. *Proceedings of the National Academy of Sciences*, *122*(1), e2321319121.

Lim, M., **Yeh, C**., Wattenberg, M., Viégas, F., & Michalatos, P. (2025). Chronotome: Real-Time Topic Modeling for Streaming Embedding Spaces. *IEEE VIS 2025 Short Paper Track*.

Lin, L., Wu, J., **Kakade, S. M**., Bartlett, P. L., & Lee, J. D. (2024). Scaling Laws in Linear Regression: Compute, Parameters, and Data. *Advances in Neural Information Processing Systems (NeurIPS)*, *37*, 60556–60606.

Liu, S., Chen, J., Fu, T., Lin, L., **Zitnik, M**., & Wu, D. (2024). Graph Adversarial Diffusion Convolution. ICML.

Liu, S., **Gershman, S.,** & Bari, B. (2025). Quantifying the cost of context sensitivity in decision making. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *47*(0).

Liu, S., & **Gershman, S. J.** (2025). Action subsampling supports policy compression in large action spaces. *PLOS Computational Biology*, *21*(9), e1013444. (Presented at Cognitive Science Society).

Liu, S., Kikumoto, A., Badre, D., & **Gershman, S. J.** (2025). Neural and behavioral signatures of policy compression in cognitive control (p. 2025.05.06.652533). bioRxiv.

Liu, S., Lai, L., **Gershman, S. J.,** & Bari, B. A. (2025). Time and memory costs jointly determine a speed-accuracy trade-off and set-size effects. *Journal of Experimental Psychology. General*, *154*(6), 1611–1627.

Liu, S., Xiang, Y., & **Gershman, S.** (2025). Probabilistic forecasting guides dynamic decisions. OSF.

Lu, Y., Chen, C., Chen, Y., Huang, K., **Zitnik, M.,** & Wang, Q. (2025). GNN 101: Visual Learning of Graph Neural Networks in Your Web Browser (No. arXiv:2411.17849). arXiv.

Lu, Y., **Letey, M**., Zavatone-Veth, J. A., Maiti, A., & **Pehlevan, C**. (2024). In-Context Learning by Linear Attention: Exact Asymptotics and Experiments. NeurIPS 2024 Workshop on Mathematics of Modern Machine Learning.

*Lu, Y. M., **Letey, M**., Zavatone-Veth, J. A., Maiti, A., & **Pehlevan, C.** (2025). Asymptotic theory of in-context learning by linear attention. *Proceedings of the National Academy of Sciences*, *122*(28), e2502599122.

Lu, Z., Singh, S. J., Jonson-Yu, S., **Walsman, A., & Rajan, K**. (2025). Emergent small-group foraging under variable group size, food scarcity, and sensory capabilities. COSYNE.

Luo, Y., **Du, D**., Huang, H., Fang, Y., & Wang, M. (2025, August 24). CurveFlow: Curvature-Guided Flow Matching for Image Generation. ICCV 2025.

Malach, E. (2024). Auto-Regressive Next-Token Predictors are Universal Learners (No. arXiv:2309.06979). arXiv.

Martinez, J. E., Krasner, R. H., Rosero, L., **Gershman, S. J.,** & Cikara, M. (2025). Social group discovery, structure, and stereotype updating. *Journal of Experimental Psychology: General*.

Masset, P., & **Gershman, S. J.** (2025). Chapter 24 - Reinforcement learning with dopamine: A convergence of natural and artificial intelligence. In S. J. Cragg & M. E. Walton (Eds.), *Handbook of Behavioral Neuroscience* (Vol. 32, pp. 305–318). Elsevier.

Mastro, K., Lee, W.-C., Wang, W., Stevens, B., & **Sabatini, B. L.** (2025). Delayed developmental maturation of frontal cortical circuits impacts decision-making (p. 2024.05.24.595609). bioRxiv.

Meng, X., Dempsey, W., Liao, P., Reid, N., Klasnja, P., & **Murphy, S.** (2025). Evaluation of the HeartSteps Online Sampling Algorithm (No. arXiv:2501.02137). arXiv.

Meterez, A., Morwani, D., Oncescu, C.-A., Wu, J., **Pehlevan, C**., **& Kakade, S**. (2025). A Simplified Analysis of SGD for Linear Regression with Weight Averaging (No. arXiv:2506.15535). arXiv.

Mills, T., **Gershman, S.,** & Tenenbaum, J. B. (2025). Strategy selection in complex tasks through adaptive integration of learned and online metareasoning. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *47*(0).

Mirtaheri, P., Edelman, E., Jelassi, S., **Malach, E**., & Boix-Adsera, E. (2025). Let Me Think! A Long Chain-of-Thought Can Be Worth Exponentially Many Short Ones (No. arXiv:2505.21825). arXiv.

Moayeri, M., Balachandran, V., Chandrasekaran, V., Yousefi, S., **Fel, T**., Feizi, S., Nushi, B., Joshi, N., & Vineet, V. (2024). Unearthing Skill-Level Insights for Understanding Trade-Offs of Foundation Models. ICLR.

Monea, G., Bosselut, A., **Brantley, K**., & Artzi, Y. (2025). LLMs Are In-Context Bandit Reinforcement Learners (No. arXiv:2410.05362). arXiv. (Presented at COLM 2025).

Moore, I. M., Nofshin, E., Swaroop, S., **Murphy, S**., Doshi-Velez, F., & Pan, W. (2025). When and Why Hyperbolic Discounting Matters for Reinforcement Learning Interventions. *RLC*. Reinforcement Learning Conference.

Morwani, D., Shapira, I., Vyas, N., **Malach, E.**, **Kakade, S. M.,** & Janson, L. (2024). A New Perspective on Shampoo's Preconditioner. The Thirteenth International Conference on Learning Representations (ICLR).

Morwani, D., Vyas, N., Zhang, H., & **Kakade, S.** (2025). Connections between Schedule-Free Optimizers, AdEMAMix, and Accelerated SGD Variants (No. arXiv:2502.02431). arXiv.

***Murthy, S. K**., Ullman, T., & **Hu, J.** (2025). One fish, two fish, but not the whole sea: Alignment reduces language models' conceptual diversity. *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 11241–11258.

Murthy, S. K., **Zhao, R**., **Hu, J., Kakade, S**., Wulfmeier, M., Qian, P., & Ullman, T. (2025). Inside you are many wolves: Using cognitive models to interpret value trade-offs in LLMs (No. arXiv:2506.20666). arXiv.

Nadkarni, G., Kauffman, J., Holmes, E., Vaid, A., Charney, A., Kovatch, P., Lampert, J., Sakhuja, A., **Zitnik, M.,** Glicksberg, B., & Hofer, I. (2025). InfEHR: Resolving Clinical Uncertainty through Deep Geometric Learning on Electronic Health Records. Research Square.

Nahum-Shani, I., & **Murphy, S. A.** (2025). Just-in-Time Adaptive Interventions: Where Are We Now and What Is Next? *Annual Review of Psychology*, *77*.

Negrel, H., Coeurdoux, F., **Albergo, M. S.,** & Vanden-Eijnden, E. (2025). Multitask Learning with Stochastic Interpolants (No. arXiv:2508.04605). arXiv.

Nguyen, H. N., L'Yi, S., Smits, T. C., Gao, S., **Zitnik, M**., & Gehlenborg, N. (2025). Multimodal retrieval of genomics data visualizations.

Oncescu, C.-A., Purandare, S., Idreos, S., & **Kakade, S.** (2024). Flash Inference: Near Linear Time Inference for Long Convolution Sequence Models and Beyond (No. arXiv:2410.12982). arXiv. (Presented at The Thirteenth International Conference on Learning Representations (ICLR)).

Osman, M. A. M., Fox, K., & Stern, J. I. (2024, November 20). A Hopfield network model of neuromodulatory arousal state. The First Workshop on NeuroAI @ NeurIPS2024. (Also presented at COSYNE 2025).

Pan, X., **Hahami, E.,** Zhang, Z., & **Sompolinsky, H.** (2025). Memorization and Knowledge Injection in Gated LLMs (No. arXiv:2504.21239). arXiv.

***Papadimitriou, I.,** Su, H., **Fel, T., Kakade, S.,** & Gil, S. (2025). Interpreting the linear structure of vision-language model embedding spaces. *Second Conference on Language Modeling (COLM) 2025*.

Patel, A. A., Lunts, P., & **Albergo, M. S.** (2025). Strange metals and planckian transport in a gapless phase from spatially random interactions. *Physical Review X*, *15*(3), 031064.

Pellegrin, R., **Fesser, L.**, & Weber, M. (2025). Enhancing the Utility of Higher-Order Information in Relational Learning (No. arXiv:2502.09570). arXiv.

Prabhakar, A., Li, Y., Narasimhan, K., **Kakade, S., Malach, E.,** & Jelassi, S. (2024). LoRA Soups: Merging LoRAs for Practical Skill Composition Tasks (No. arXiv:2410.13025). arXiv.

Prashanth, U. S., Deng, A., O'Brien, K., V, J. S., Khan, M. A., Borkar, J., Choquette-Choo, C. A., Fuehne, J. R., Biderman, S., Ke, T., Lee, K., & **Saphra, N.** (2024). Recite, Reconstruct, Recollect: Memorization in LMs as a Multifaceted Phenomenon. The Thirteenth International Conference on Learning Representations (ICLR).

Prat-Carrabin, A., Harl, M. V., & **Gershman, S. J.** (2025). Fast efficient coding and sensory adaptation in gain-adaptive recurrent networks (p. 2025.07.11.664261). bioRxiv.

Prince, J. S., **Alvarez, G. A.,** & **Konkle, T.** (2024). Representation with a capital "R": Measuring functional alignment with causal perturbation. UniReps: 2nd Edition of the Workshop on Unifying Representations in Neural Models.

Qi, Z., Nie, F., Alahi, A., Zou, J., Lakkaraju, H., Du, Y., Xing, E., **Kakade, S.,** & Zhang, H. (2025). EvoLM: In Search of Lost Language Model Training Dynamics (No. arXiv:2506.16029). arXiv.

Qian, L., Burrell, M., Hennig, J. A., Matias, S., **Murthy, V. N.**, **Gershman, S. J.,** & Uchida, N. (2025). Prospective contingency explains behavior and dopamine signals during associative learning. *Nature Neuroscience*, *28*(6), 1280–1292.

Qian, W., Zavatone-Veth, J., Ruben, B., & **Pehlevan, C.** (2025). Mechanistic biases in data constrained models of neural dynamics. COSYNE.

Qian, W., Zavatone-Veth, J. A., Ruben, B. S., & **Pehlevan, C.** (2024). Partial observation can induce mechanistic mismatches in data-constrained models of neural dynamics. The Thirty-eighth Annual Conference on Neural Information Processing Systems (NeurIPS). (Also presented at 8th Annual Conference on Cognitive Computational Neuroscience (CCN)).

*Qin, T., **Alvarez-Melis, D.**, **Jelassi, S**., & **Malach, E.** (2025). To Backtrack or Not to Backtrack: When Sequential Search Limits Model Reasoning. COLM 2025.

Qin, T., Deng, Z., & **Alvarez-Melis, D.** (2024). A Label is Worth A Thousand Images in Dataset Distillation. The Thirty-eighth Annual Conference on Neural Information Processing Systems (NeurIPS).

Qin, T., Park, C. F., **Kwun, M**., **Walsman, A., Malach, E., Anand, N.,** Tanaka, H., & **Alvarez-Melis, D.** (2025). Decomposing Elements of Problem Solving: What "Math" Does RL Teach? (No. arXiv:2505.22756). arXiv.

Qin, T., **Saphra, N.,** & **Alvarez-Melis, D**. (2024, December 19). Sometimes I am a Tree: Data Drives Unstable Hierarchical Generalization. NeurIPS Workshop.

Qü, A. J., Tai, L.-H., Hall, C. D., Tu, E. M., Eckstein, M. K., Mishchanchuk, K., Lin, W. C., Chase, J. B., MacAskill, A. F., Collins, A. G. E., **Gershman, S. J.,** & Wilbrecht, L. (2025). Nucleus accumbens dopamine release reflects Bayesian inference during instrumental learning. *PLOS Computational Biology*, *21*(7), e1013226.

Queen, O., Huang, Y., Calef, R., Giunchiglia, V., Chen, T., Dasoulas, G., Tai, L., Ektefaie, Y., Noori, A., Brown, J., Cobley, T., Hrovatin, K., Hartvigsen, T., Theis, F. J., Pentelute, B., Khurana, V., Kellis, M., & **Zitnik, M.** (2024). ProCyon: A multimodal foundation model for protein phenotypes (p. 2024.12.10.627665). bioRxiv.

Rahamim, A., **Saphra, N**., Kangaslahti, S., & Belinkov, Y. (2024). Fast Forwarding Low-Rank Training (No. arXiv:2409.04206). arXiv.

Regev, T. I**., Casto, C**., Hosseini, E. A., Adamek, M., Ritaccio, A. L., Willie, J. T., Brunner, P., & Fedorenko, E. (2024). Neural populations in the language network differ in the size of their temporal receptive windows. *Nature Human Behaviour*, *8*(10), 1924–1942.

Reinhold, K., Iadarola, M., Tang, S., Chang, A., Kuwamoto, W., Albanese, M. A., Sun, S., **Hakim, R.**, Zimmer, J., Wang, W., & **Sabatini, B**. (2025). Striatum supports fast learning but not memory recall. *Nature*.

Reinhold, K., Iadarola, M., Tang, S., Chang, A., Kuwamoto, W., Albanese, M. A., Sun, S., H**akim, R.**, Zimmer, J., Wang, W., & **Sabatini, B.** (2025).

Striatum supports fast learning but not memory recall. *Nature*.

Rodriguez-Diaz, P., Kong, L., Wang, K., **Alvarez-Melis, D**., & Tambe, M. (2025). What is the Right Notion of Distance between Predict-then-Optimize Tasks? (No. arXiv:2409.06997). arXiv.

Rong, Y., Conwell, C., **Hidalgo, D**., & Bonner, M. (2024). Unveiling Core, Interpretable Image Properties Underlying Model-Brain Similarity with Generative Models. *Journal of Vision*. Vision Sciences Society Conference.

Ruben, B. S., **Tong, W. L**., Chaudhry, H. T., & **Pehlevan, C**. (2024). No Free Lunch From Random Feature Ensembles. ICML.

Sajid, N., & Medrano, J. (2025). Dissociating model architectures from inference computations. *Cognitive Neuroscience*, *0*(0), 1–3.

Saphra, N., & Wiegreffe, S. (2024). Mechanistic? (No. arXiv:2410.09087). arXiv.

Saxon, M., Holtzman, A., West, P., Wang, W. Y., & **Saphra, N.** (2024). Benchmarks as Microscopes: A Call for Model Metrology. COLM.

Shahout, R., **Malach, E**., Liu, C., Jiang, W., Yu, M., & Mitzenmacher, M. (2024). Don't Stop Me Now: Embedding Based Scheduling for LLMs (No. arXiv:2410.01035). arXiv.

Shan, H., Li, Q., & **Sompolinsky, H.** (2025). Order parameters and phase transitions of continual learning in deep neural networks (No. arXiv:2407.10315). arXiv.

Shan, H., Li, Q., & **Sompolinsky, H.** (2025). Metrics of Task Relations Predict Continual Learning Performance. COSYNE.

Shang, J., **Sompolinsky, H**., Kreiman, G., & Kar, K. (2024). Probing the neural code for multiple objects: Identity and numerosity encoding in macaque inferior temporal cortex. Society for Neuroscience Annual Meeting.

Shang, J., Jain, S., **Sompolinsky, H.,** & Chang, E. (2025). Geometric Signatures of Speech Recognition: Insights from Deep Neural Networks to the Brain. COSYNE.

Shang, J., Kreiman, G., & **Sompolinsky, H.** (2025). Unraveling the Geometry of Visual Relational Reasoning. *ArXiv*, arXiv:2502.17382v1.

Shen, W., Nguyen, T. H., Li, M. M., Huang, Y., Moon, I., Nair, N., Marbach, D., & **Zitnik, M.** (2025). Generalizable AI predicts immunotherapy outcomes across cancers and treatments (p. 2025.05.01.25326820). medRxiv.

Shen, W. X., Cui, C., Su, X., Zhang, Z., Velez-Arce, A., Wang, J., Shi, X., Zhang, Y., Wu, J., Chen, Y. Z., & **Zitnik, M**. (2024). Activity Cliff-Informed Contrastive Learning for Molecular Property Prediction. *Research Square*, rs.3.rs-2988283.

Simmons-Edler, R., Badman, R. P., Berg, F. B., Chua, R., Vastola, J. J., Lunger, J., **Qian, W.,** & **Rajan, K.** (2025). Deep RL Needs Deep Behavior Analysis: Exploring Implicit Planning by Model-Free Agents in Open-Ended Environments (No. arXiv:2506.06981). arXiv.

Simmons-Edler, R., Badman, R. P., Longpre, S., & **Rajan, K.** (2024. Position: AI-Powered Autonomous Weapons Risk Geopolitical Instability and Threaten AI Research. Forty-first International Conference on Machine Learning (ICML).

Simmons-Edler, R., Dong, J., Lushenko, P., **Rajan, K**., & Badman, R. P. (2025). Military AI Needs Technically-Informed Regulation to Safeguard AI Research and its Applications (No. arXiv:2505.18371). arXiv.

Singhvi, D., Misra, D., Erkelens, A., Jain, R., **Papadimitriou, I., & Saphra, N**. (2025). Using Shapley interactions to understand how models use structure. ACL 2025.

Song, S., **Hu, J.,** & Mahowald, K. (2025). Language Models Fail to Introspect About Their Knowledge of Language (No. arXiv:2503.07513). arXiv.

Song, Y., **Keller, T. A**., Yue, Y., Perona, P., & Welling, M. (2024). Unsupervised Representation Learning from Sparse Transformation Analysis (No. arXiv:2410.05564). arXiv.

Song, Y., **Keller, T. A.**, Yue, Y., Perona, P., & Welling, M. (2025). Langevin Flows for Modeling Neural Latent Dynamics (No. arXiv:2507.11531). arXiv.

Sosa, F. A., **Gershman, S. J.,** & Ullman, T. D. (2025). Blending simulation and abstraction for physical reasoningv. *Cognition*, *254*, 105995.

*Su, H., **Kwun, M.**, Gil, S., **Kakade, S.**, & **Anand, N.** (2025). Characterization and Mitigation of Training Instabilities in Microscaling Formats.

Su, H., **Walsman, A.,** Garces, D., **Kakade, S.,** & Gil, S. (2025). Data-Efficient Multi-Agent Spatial Planning with LLMs. *arXiv Preprint arXiv:2502.18822*.

Su, X., Messica, S., Huang, Y., Johnson, R., **Fesser, L.,** Gao, S., Sahneh, F., & **Zitnik, M.** (2025). Multimodal Medical Code Tokenizer (No. arXiv:2502.04397). arXiv.

Su, X., Wang, Y., Gao, S., Liu, X., Giunchiglia, V., Clevert, D.-A., & **Zitnik, M.** (2025). KGARevion: An AI Agent for Knowledge-Intensive Biomedical QA (No. arXiv:2410.04660). arXiv.

Tahir Chaudhry, H., Zavatone-Veth, J. A., Krotov, D., & **Pehlevan, C.** (2024). Long sequence Hopfield memory. *Journal of Statistical Mechanics: Theory and Experiment*, *2024*(10), 104024.

Thasarathan, H., Forsyth, J., **Fel, T.,** Kowal, M., & Derpanis, K. (2025). Universal Sparse Autoencoders: Interpretable Cross-Model Concept Alignment. ICML.

Tiberi, L., Mignacco, F., Irie, K., & **Sompolinsky, H.** (2024). Dissecting the Interplay of Attention Paths in a Statistical Mechanics Theory of Transformers. Advances in Neural Information Processing Systems (NeurIPS), *37*, 72710–72753.

*Tolooshams, B., Matias, S., Wu, H., Temereanca, S., Uchida, N., **Murthy, V. N.**, Masset, P., & **Ba, D.** (2025). Interpretable deep learning for deconvolutional analysis of neural signals. *Neuron*, *113*(8), 1151-1168. e13.

Tong, W. L., **Murthy, V. N.**, & Reddy, G. (2025). Adaptive algorithms for shaping behavior. *PLoS Computational Biology*, *21*(9), e1013454.

Tong, W. L., & Pehlevan, C. (2025). MLPs Learn In-Context on Regression and Classification Tasks. ICLR.

Tong, W. L., & Pehlevan, C. (2025). Learning richness modulates equality reasoning in neural networks. *CCN*. 8th Annual Conference on Cognitive Computational Neuroscience (CCN). (Presented at COSYNE 2025).

Tong, W. & **Pehlevan, C**. (2025). Equality reasoning in neural networks is modulated by learning richness. COSYNE.

Trella, A. L., Dempsey, W. H., Gazi, A., Xu, Z., Doshi-Velez, F., & **Murphy, S**. (2025). Non-Stationary Latent Auto-Regressive Bandits. Reinforcement Learning Conference.

Trella, A. L., Zhang, K. W., Jajal, H., Nahum-Shani, I., Shetty, V., Doshi-Velez, F., & **Murphy, S. A.** (2025). A Deployed Online Reinforcement Learning Algorithm in an Oral Health Clinical Trial. *Proceedings of the AAAI Conference on Artificial Intelligence*, *39*(28), 28792–28800.

Uchendu, I., Jabbour, J., Berghe, K. V. den, Runevic, J., Stewart, M., Ma, J. J., Krishnan, S., Gur, I., **Huang, A. V.,** Bishop, C., Bailey, P., Jiang, W., Songhori, E., Guadarrama, S., Tan, J., Terry, J. K., Faust, A., & Reddi, V. J. (2025). A2Perf: Benchmarking Autonomous Agents End-to-End in Realistic Domains. Championing Open-source DEvelopment in ML Workshop @ ICML25.

Uchendu, I., Zhuang, V., Jiang, W., Lee, K.-H., Songhori, E., Goel, S., Hou, K., & Reddi, V. J. (2025). See it to Place it: Evolving Macro Placements with Vision Language Models. The Exploration in AI Today Workshop at ICML 2025.

Usha, B., Oesterling, A., Srinivas, S., Calmon, F. P., & Lakaraju, H. (2024). Interpreting CLIP with Sparse Linear Concept Embeddings (SpLiCE). Advances in Neural Information Processing Systems (NeurIPS).

van Meegen, A., & **Sompolinsky, H.** (2025). Coding schemes in neural networks learning classification tasks. *Nature Communications*, *16*(1), 3354.

Velez-Arce, A., Li, M. M., Gao, W., Lin, X., Huang, K.,

Fu, T., Pentelute, B. L., Kellis, M., & **Zitnik, M.** (2024). Signals in the Cells: Multimodal and Contextualized Machine Learning Foundations for Therapeutics. *bioRxiv*, 2024.06.12.598655.

Velez-Arce, A., & **Zitnik, M.** (2025). PyTDC: A multimodal machine learning training, evaluation, and inference platform for biomedical foundation models. ICML.

Vyas, N., **Morwani, D.**, **Zhao, R.**, Shapira, I., **Brandfonbrener, D.,** Janson, L., & **Kakade, S. M.** (2024). SOAP: Improving and Stabilizing Shampoo using Adam for Language Modeling. The Thirteenth International Conference on Learning Representations (ICLR).

Wal, O. van der, Lesci, P., Muller-Eberstein, M., **Saphra, N.**, Schoelkopf, H., Zuidema, W., & Biderman, S. (2025). PolyPythias: Stability and Outliers across Fifty Language Model Pre-Training Runs. ICLR.

Wallingford, M., Bhattad, A., Kusupati, A., Ramanujan, V., Deitke, M., **Kakade, S.**, Kembhavi, A., Mottaghi, R., Ma, W.-C., & Farhadi, A. (2024). From an Image to a Scene: Learning to Imagine the World from a Million 360° Videos. *Advances in Neural Information Processing Systems (NeurIPS)*, *37*, 17743–17760.

Wang, B., & Pehlevan, C. (2025). An Analytical Theory of Spectral Bias in the Learning Dynamics of Diffusion Models (No. arXiv:2503.03206). arXiv.

Wang, B., & Ponce, C. R. (2024). Neural Dynamics of Object Manifold Alignment in the Ventral Stream (p. 2024.06.20.596072). bioRxiv.

Wang, B., Shang, J., & **Sompolinsky, H.** (2024). How do diffusion models learn and generalize on abstract rules for reasoning? Cognitive Computational Neuroscience (CCN).

Wang, B., Shang, J., & **Sompolinsky, H.** (2024). Diverse capability and scaling of diffusion and auto-regressive models when learning abstract rules. NeurIPS2024 Workshop on System 2 Reasoning At Scale.

***Wang, B.,** & Vastola, J. J. (2024). The Unreasonable Effectiveness of Gaussian Score Approximation for Diffusion Models and its Applications. *Transactions on Machine Learning Research*.

Wang, K., Zhou, J. P., Chang, J., Gao, Z., Kallus, N., **Brantley, K.**, & Sun, W. (2025). Value-Guided Search for Efficient Chain-of-Thought Reasoning (No. arXiv:2505.17373). arXiv. (Presented at ICML 2025).

Wei, Z., Ektefaie, Y., Zhou, A., Negatu, D., Aldridge, B., Dick, T., Skarlinski, M., White, A. D., Rodriques, S., Hosseiniporgham, S., Krieger, I., Sacchettini, J., **Zitnik, M.,** & Farhat, M. (2025). Fleming: An AI Agent for Antibiotic Discovery in Mycobacterium tuberculosis (p. 2025.04.01.646719). bioRxiv.

Wong, P., Wu, Y., **Hidalgo, D.,** Kreiman, G., & Anastassiou, C. A. (2025). Inhibitory cell-type-specific properties and transformations set up a unique depth-dependent temporal integration scheme in the human neocortical circuit. COSYNE.

Woolfson, D. N., Colwell, L. J., Chen, Z., Vorobieva, A. A., Polizzi, N. F., Stein, A., Liu, H., Parmeggiani, F., Peacock, A., Singh, R., King, N., **Zitnik, M.,** & Chica, R. A. (2024). How do you anticipate computational protein design will change biotechnology and therapeutic development? *Cell Systems*, *15*(11), 994–999.

Wu, R., Chen, Y., Swamy, G., **Brantley, K.,** & Sun, W. (2025). Diffusing States and Matching Scores: A New Framework for Imitation Learning (No. arXiv:2410.13855). arXiv.

Xiang, Y., Bigelow, E., Gerstenberg, T., Ullman, T. D., & **Gershman, S.** (2025). Language models assign responsibility based on actual rather than counterfactual contributions. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *47*(0).

Xiang, Y., Dorst, K., & **Gershman, S. J.** (2025). On the Robustness and Provenance of the Gambler's Fallacy. *Psychological Science*, *36*(6), 451–464.

Xiang, Y., & **Gershman, S.** (2025). Modeling intrinsic motivation as reflective planning. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *47*(0).

Xiang, Y., Landy, J., Cushman, F. A., Vélez, N., & **Gershman, S. J.** (2025). People reward others based on their willingness to exert effort. *Journal of Experimental Social Psychology*, *116*, 104699.

Xu, Z., Jajal, H., Choi, S. W., Nahum-Shani, I., Shani, G., Psihogios, A. M., Hung, P.-Y., & **Murphy, S.** (2025). Reinforcement Learning on Dyads to Enhance Medication Adherence. 23rd International Conference on AI in Medicine (AIME) 2025.

Yan, K., Li, X., Ling, H., Ashen, K., Edwards, C., Arróyave, R., **Zitnik, M.,** Ji, H., Qian, X., Qian, X., & Ji, S. (2024). Invariant Tokenization of Crystalline Materials for Language Model Enabled Generation. *Advances in Neural Information Processing Systems (NeurIPS)*, *37*, 125050–125072.

Yang, S., Liu, P., & **Pehlevan, C.** (2024). Convex Relaxation for Solving Large-Margin Classifiers in Hyperbolic Space. *Transactions on Machine Learning Research*.

Yang, S., Zavatone-Veth, J. A., & **Pehlevan, C**. (2024). Spectral regularization for adversarially-robust representation learning (No. arXiv:2405.17181). arXiv.

Yang, S., Zavatone-Veth, J. A., & **Pehlevan, C.** (2025). Adversarially-robust representation learning through spectral regularization of features. NeurIPS 2024 Workshop on Symmetry and Geometry in Neural Representations.

Yeh, C., Menon, T., Arya, R. S., He, H., Weigel, M., Viégas, F., & Wattenberg, M. (2025). Story Ribbons: Reimagining Storyline Visualizations with Large Language Models (No. arXiv:2508.06772). arXiv.

Yik, J., Van den Berghe, K., den Blanken, D., Bouhadjar, Y., Fabre, M., Hueber, P., Ke, W., Khoei, M. A., Kleyko, D., Pacik-Nelson, N., Pierro, A., Stratmann, P., Sun, P.-S. V., Tang, G., Wang, S., Zhou, B., Ahmed, S. H., Vathakkattil Joseph, G., Leto, B., … Reddi, V. J. (2025). The neurobench framework for benchmarking neuromorphic computing algorithms and systems. *Nature Communications*, *16*(1), 1545. (Presented at 2025 NICE Conference).

Ying, L., Collins, K. M., Sharma, P., Colas, C., Zhao, K. I., Weller, A., Tavares, Z., Isola, P., **Gershman, S. J.,** Andreas, J. D., Griffiths, T. L., Chollet, F., Allen, K. R., & Tenenbaum, J. B. (2025). Assessing Adaptive World Models in Machines with Novel Games (No. arXiv:2507.12821). arXiv.

Ying, L., Truong, R., Tenenbaum, J. B., & **Gershman, S. J.** (2025). Adaptive Social Learning using Theory of Mind (No. arXiv:2507.09409). arXiv.

Zaki, Y., Pennington, Z. T., Morales-Rodriguez, D., Bacon, M. E., Ko, B., Francisco, T. R., LaBanca, A. R., Sompolpong, P., Dong, Z., Lamsifer, S., Chen, H.-T., Carrillo Segura, S., Christenson Wick, Z., Silva, A. J., **Rajan, K**., van der Meer, M., Fenton, A., Shuman, T., & Cai, D. J. (2025). Offline ensemble co-reactivation links memories across days. *Nature*, *637*(8044), 145–155.

Zavatone-Veth, J. A., Bordelon, B., & **Pehlevan, C.** (2025). Summary statistics of learning link changing neural representations to behavior. *Frontiers in Neural Circuits*, *19*, 1618351.

Zavatone-Veth, J. A., & **Pehlevan, C.** (2024). Learning curves for deep structured Gaussian feature models. *Journal of Statistical Mechanics: Theory and Experiment*, *2024*, 104022.

Zavatone-Veth, J. A., & **Pehlevan, C.** (2025a). A note on the dynamics of extended-context disordered kinetic spin models (No. arXiv:2507.18461). arXiv.

Zavatone-Veth, J. A., & **Pehlevan, C**. (2025b). Nadaraya–Watson kernel smoothing as a random energy model. *Journal of Statistical Mechanics: Theory and Experiment*, *2025*(1), 013404.

Zhang, E., Zhu, V., **Saphra, N., Kleiman, A**., Edelman, B. L., Tambe, M., **Kakade, S., & Malach, E**. (2024). Transcendence: Generative Models Can Outperform The Experts That Train Them. Advances in Neural Information Processing Systems (NeurIPS), *37*, 86985–87012.

Zhang, H., **Liu, B**., Kim, J., & Risteski, A. (2025). Improving Pathfinding with Anchoring Tokens. ICML 2025 Workshop on Methods and Opportunities at Small Scale.

Zhang, H., **Morwani, D**., Vyas, N., Wu, J., Zou, D., Ghai, U., Foster, D., & **Kakade, S. M.** (2024). How Does Critical Batch Size Scale in Pre-training? The Thirteenth International Conference on Learning Representations (ICLR).

Zhang, K. W., Closser, N., Trella, A. L., & **Murphy, S. A**. (2025). Replicable Bandits for Digital Health Interventions (No. arXiv:2407.15377). arXiv.

Zhang, S., Han, T., **Bhalla, U.,** & Lakkaraju, H. (2025). Towards Unified Attribution in Explainable AI, Data-Centric AI, and Mechanistic Interpretability (No. arXiv:2501.18887). arXiv.

Zhang, S., Han, T., **Bhalla, U**., & Lakkaraju, H. (2025). Building Bridges, Not Walls: Advancing Interpretability by Unifying Feature, Data, and Model Component Attribution. ICLR 2025 Workshop on Navigating and Addressing Data Problems for Foundation Models.

Zhang, X., Wang, L., Helwig, J., Luo, Y., Fu, C., Xie, Y., Liu, M., Lin, Y., Xu, Z., Yan, K., Adams, K., Weiler, M., Li, X., Fu, T., Wang, Y., Strasser, A., Yu, H., Xie, Y., Fu, X., **Fang, A.**, **Zitnik, M**.... Ji, S. (2025). Artificial Intelligence for Science in Quantum, Atomistic, and Continuum Systems. *Foundations and Trends® in Machine Learning*, *18*(4), 385–912.

Zhang, X., Lin, H., Ye, H., Zou, J., Ma, J., Liang, Y., & **Du, Y**. (2025). Inference-Time Scaling of Diffusion Models through Classical Search. (No. arXiv:2505.23614). arXiv.

Zhang, Y., Talebi, S., & Li, N. (2024). Learning Low-dimensional Latent Dynamics from High-dimensional Observations: Non-asymptotics and Lower Bounds (No. arXiv:2405.06089). arXiv.

Zhang, Y., Zhang, X., **Liu, J.,** & Li, N. (2025). Error-In-Variables Methods for Efficient System Identification with Finite-Sample Guarantees (No. arXiv:2504.09057). arXiv.

Zhang, Z., Jin, R., Xu, G., Wang, X**., Zitnik, M.,** Cong, L., & Wang, M. (2025). FoldMark: Safeguarding Protein Structure Generative Models with Distributional and Evolutionary Watermarking. *bioRxiv*, 2024.10.23.619960.

Zhang, Z., Shen, W. X., Liu, Q., & **Zitnik, M.** (2024). Efficient generation of protein pockets with PocketGen. *Nature Machine Intelligence*, *6*(11), 1382–1395
.
Zhang, Z., & **Sompolinsky, H.** (2025). When narrower is better: The narrow width limit of Bayesian parallel branching neural networks (No. arXiv:2407.18807). arXiv.

Zhang, Z**., Zitnik, M**., & Liu, Q. (2024). Generalized Protein Pocket Generation with Prior-Informed Flow Matching. *Advances in Neural Information Processing Systems (NeurIPS)*, *37*, 38559–38589.

Zhao, R., Meterez, A., Kakade, S. M., Pehlevan, C., Jelassi, S., & **Malach, E.** (2025). Echo Chamber: RL Post-training Amplifies Behaviors Learned in Pretraining. *COLM*. Second Conference on Language Modeling.

Zhao, R., Morwani, D., Brandfonbrener, D., Vyas, N., & **Kakade, S. M**. (2024). Deconstructing What Makes a Good Optimizer for Autoregressive Language Models. The Thirteenth International Conference on Learning Representations (ICLR).

Zhao, R., Qin, T., **Alvarez-Melis, D., Kakade, S. M., & Saphra, N**. (2025a). Distributional Scaling Laws for Emergent Capabilities. *CoRR*.

Zhao, R., Qin, T., **Alvarez-Melis, D**., **Kakade, S., & Saphra, N**. (2025b). Distributional Scaling for Emergent Capabilities (No. arXiv:2502.17356). arXiv.

Zhou, S., Badman, R., Arlt, C., **Rajan, K**., & Harvey, C. (2025). Inhibition-stabilized disordered dynamics in mouse cortex during navigational decision-making. COSYNE.

Zhou, J. P., Wang, K., Chang, J., Gao, Z., Kallus, N., Weinberger, K. Q**., Brantley, K**., & Sun, W. (2025). $Q\sharp$: Provably Optimal Distributional RL for LLM Post-Training (No. arXiv:2502.20548). arXiv. (Presented at ICML 2025).

Zitnik, M. (2025). AI-enabled drug discovery reaches clinical milestone. *Nature Medicine*, *31*(8), 2490–2491.

Zitnik, M. (2025). Machine Learning for Genomics Explorations. The Thirteenth International Conference on Learning Representations (ICLR).

Zitnik, M. (2025). Towards Agentic AI for Science. The Thirteenth International Conference on Learning Representations (ICLR).

CODE REPOSITORIES

Learning Universal Representations of Intermolecular Interactions with ATOMICA
Ada Fang, Michael Desgagné, Zaixi Zhang, Andrew Zhou, Joseph Loscalzo, Bradley L. Pentelute, & **Marinka Zitnik**
https://huggingface.co/ada-f/ATOMICA
https://github.com/mims-harvard/ATOMICA
https://dataverse.harvard.edu/citation?persistentId=doi:10.7910/DVN/4DUBJX

Echo Chamber: RL Post-training Amplifies Behaviors Learned in Pretraining
Rosie Zhao, Alexandru Meterez, Sham M. Kakade, Cengiz Pehlevan, Samy Jelassi, **& Eran Malach**
https://github.com/rosieyzh/openrlhf-pretrain?utm_source=catalyzex.com
https://github.com/huggingface/math-verify?utm_source=catalyzex.com

The Optimization Landscape of SGD Across the Feature Learning Strength
Alexander Atanasov, Alexandru Meterez, James B. Simon, & **Cengiz Pehlevan**
https://github.com/Pehlevan-Group/Richness_Sweep?utm_source=catalyzex.com

The SMeL Test: A simple benchmark for media literacy in language models
Gustaf Ahdritz & **Anat Kleiman**
https://github.com/gahdritz/smel

Ephys Spike Sorting
Bala Desinghu
https://github.com/KempnerInstitute/ephys-spike-sorting

Nvidia Nemo Workflows
Bala Desinghu
https://github.com/KempnerInstitute/nvidia-nemo-workflows

Optimizing ML Workflow
Bala Desinghu
https://github.com/KempnerInstitute/optimizing-ml-workflow

GPT Auto Data Analytics
Binxu Wang
https://github.com/Animadversio/GPT-Auto-Data-Analytics

CoLoR-Filter: Conditional Loss Reduction Filtering for Targeted Language Model Pre-training
David Brandfonbrener, Hanlin Zhang, Andreas Kirsch, Jonathan Richard Schwarz, & **Sham M. Kakade**
https://github.com/camilobrownpinilla/dclm-color-filter-olmo

Characterization and Mitigation of Training Instabilities in Microscaling Formats
Nikhil Anand
https://github.com/KempnerInstitute/systems-scaling

Transformers for Modeling Decision Sequences
Corwin Cheung
https://github.com/CorwinCheung/Transformers_for_Modeling_Decision_Sequences/

A Label is Worth a Thousand Images in Dataset Distillation
Tian Qin, Zhiwei Deng, **David Alvarez-Melis**
https://github.com/sunnytqin/no-distillation

Loss-to-Loss Prediction: Scaling Laws for All Datasets
David Brandfonbrener, **Nikhil Anand**, Nikhil Vyas, **Eran Malach,** & **Sham Kakade**
https://huggingface.co/KempnerInstituteAI/loss-to-loss
https://github.com/KempnerInstitute/loss-to-loss-olmo

Projecting Assumptions: The Duality Between Sparse Autoencoders and Concept Geometry
Sai Sumedh R. Hindupur, Ekdeep Singh Lubana, **Thomas Fel,** & **Demba Ba**
https://github.com/karpathy/nanoGPT?utm_source=catalyzex.com
https://github.com/Sai-Sumedh/SaeConceptDuality-SpaDE
https://github.com/EkdeepSLubana/spadeFormalGrammars
https://github.com/KempnerInstitute/Overcomplete

From Flat to Hierarchical: Extracting Sparse Representations with Matching Pursuit
Valérie Costa, **Thomas Fel**, Ekdeep Singh Lubana, Bahareh Tolooshams, & **Demba Ba**
https://github.com/mpsae/MP-SAE

Connections between Schedule-Free Optimizers, AdEMAMix, and Accelerated SGD Variants
Depen Morwani, Nikhil Vyas, Hanlin Zhang, & **Sham Kakade**
https://github.com/DepenM/Simplified-AdEMAMix

SOAP: Improving and Stabilizing Shampoo Using Adam for Language Modeling
Nikhil Vyas, **Depen Morwani, Rosie Zhao**, Itai Shapira, **David Brandfonbrener**, Lucas Janson, & **Sham M. Kakade**
https://github.com/nikhilvyas/SOAP

Deconstructing What Makes a Good Optimizer for Autoregressive Language Models
Rosie Zhao, Depen Morwani, **David Brandfonbrener**, Nikhil Vyas, & **Sham M. Kakade**
https://github.com/rosieyzh/optimizers-llm

Trial-by-trial inter-areal interactions in visual cortex in the presence or absence of visual stimulation
Dianna Hidalgo, Giorgia Dellaferrera, Will Xiao, Maria Papadopouli, Stelios Smirnakis, & Gabriel Kreiman
https://github.com/4sdch/inter-area-neural-prediction

Memorization and Knowledge Injection in Gated LLMs
Xu Pan, **Ely Hahami**, Zechen Zhang, & **Haim Sompolinsky**
https://github.com/xup5/MEGA

Learning Artistic Signatures: Symmetry Discovery and Style Transfer
Emma Finn, **T. Anderson Keller**, Emmanouil Theodosis, & **Demba Ba**
https://github.com/EmmaFinn314/Style-Transfer

Decomposing Elements of Problem Solving: What "Math" Does RL Teach?
Tian Qin, Core Francisco Park, **Mujin Kwun, Aaron Walsman, Eran Malach, Nikhil Anand,** Hidenori Tanaka, & **David Alvarez-Melis**
https://github.com/cfparkOO/RL-Wall?utm_source=catalyzex.com

Let Me Think! A Long Chain-of-Thought Can Be Worth Exponentially Many Short Ones
Parsa Mirtaheri, Ezra Edelman, Samy Jelassi, **Eran Malach**, Enric Boix-Adsera
https://github.com/seyedparsa/let-me-think

To backtrack or not to backtrack: When sequential search limits model reasoning
Tian Qin, **David Alvarez-Melis**, Samy Jelassi, & **Eran Malach**
https://github.com/kaistAI/SelFee
https://github.com/unmade/dokusan?utm_source=catalyzex.com
https://github.com/KempnerInstitute/Backtrack

Transcendence: Generative models can outperform the experts that train them
Edwin Zhang, Vincent Zhu, **Naomi Saphra**, **Anat Kleiman**, Benjamin L. Edelman, Milind Tambe, **Sham M. Kakade,** & **Eran Malach**
https://github.com/KempnerInstitute/chess-research

The evolution of statistical induction heads: In-context learning markov chains
Benjamin L. Edelman, Ezra Edelman, Surbhi Goel, **Eran Malach**, Nikolaos Tsilivis
https://github.com/EzraEdelman/Evolution-of-Statistical-Induction-Heads

Lora soups: Merging loras for practical skill composition tasks
Akshara Prabhakar, Yuanzhi Li, Karthik Narasimhan, **Sham Kakade, Eran Malach**, Samy Jelassi
https://github.com/sahil280114/codealpaca?utm_source=catalyzex.com
https://github.com/deep-floyd/IF

A feedforward mechanism for human-like contour integration
Fenil R. Doshi, Talia Konkle, & **George A. Alvarez**
https://github.com/feziodoshi/dnn_contour_integration

The neurobench framework for benchmarking neuromorphic computing algorithms and systems
Jason Yik, Korneel Van den Berghe, Douwe den Blanken, Younes Bouhadjar, Maxime Fabre, Paul Hueber, Weijie Ke, Mina A. Khoei, Denis Kleyko, Noah Pacik-Nelson, Alessandro Pierro, Philipp Stratmann, Pao-Sheng Vincent Sun, Guangzhi Tang, Shenqi Wang, Biyan Zhou, Soikat Hasan Ahmed, George Vathakkattil Joseph, Benedetto Leto, Aurora Micheli, Anurag Kumar Mishra, Gregor Lenz, Tao Sun, Zergham Ahmed, ...Vijay Janapa Reddi
https://github.com/NeuroBench/neurobench

Playscript Benchmark
Justin Ji
https://github.com/jujipotle/playscript_benchmark/tree/master

Honest Llama
Kenneth Li
https://github.com/likenneth/honest_llama

POCO: Scalable neural forecasting through population conditioning
Yu Duan, Hamza Tahir Chaudhry, Misha B. Ahrens, Christopher D Harvey, Matthew G Perich, Karl Deisseroth, & **Kanaka Rajan**
https://github.com/yuvenduan/POCO

ReprGeo_LRM
Kexin Cindy Luo
https://github.com/cindyLuo99/reprGeo_LRM

Accelerating RL for LLM Reasoning with Optimal Advantage Regression
Kianté Brantley, Mingyu Chen, Zhaolin Gao, Jason D. Lee, Wen Sun, Wenhao Zhan, & Xuezhou Zhang
https://huggingface.co/datasets/VGS-AI/OpenR1-Cleaned
https://github.com/ZhaolinGao/A-PO

A multimodal foundation model for protein phenotypes
Owen Queen, Yepeng Huang, Robert Calef, Valentina Giunchiglia, Tianlong Chen, George Dasoulas, LeAnn Tai, Gianmarco Abbadessa, Owain Howell, Michelle M. Li, Yasha Ektefaie, Ayush Noori, Ildiko Farkas, Joseph Brown, Tom Cobley, Karin Hrovatin, Tom Hartvigsen, Fabian J. Theis, Bradley L. Pentelute, James Zou, Vikram Khurana, David Owen, Richard Nicholas, Manolis Kellis, & **Marinka Zitnik**
https://github.com/mims-harvard/ProCyon

Multimodal AI predicts clinical outcomes of drug combinations from preclinical data
Yepeng Huang, Xiaorui Su, Varun Ullanat, Intae Moon, Ivy Liang, Lindsay Clegg, Damilola Olabode, Ruthie Johnson, Nicholas Ho, Megan Gibbs, Megan Gibbs, Alexander Gusev, Bino John, & **Marinka Zitnik**
https://github.com/mims-harvard/Madrigal

TxAgent: An AI agent for therapeutic reasoning across a universe of tools
Shanghua Gao, Richard Zhu, Zhenglun Kong, Ayush Noori, Xiaorui Su, Curtis Ginder, Theodoros Tsiligkaridis, & **Marinka Zitnik**
https://github.com/mims-harvard/TxAgent
https://github.com/mims-harvard/ToolUniverse

Multimodal Medical Code Tokenizer
Xiaorui Su, Shvat Messica, Yepeng Huang, Ruth Johnson, **Lukas Fesser**, Shanghua Gao, Faryad Sahneh, & **Marinka Zitnik**
https://zitniklab.hms.harvard.edu/projects/MedTok/

KGARevion: An AI Agent for Knowledge-Intensive Biomedical QA
Xiaorui Su, Yibo Wang, Shanghua Gao, Xiaolong Liu, Valentina Giunchiglia, Djork-Arné Clevert, & **Marinka Zitnik**
https://github.com/mims-harvard/KGARevion

Generalizable AI Predicts Immunotherapy Outcomes Across Cancers and Treatments
Wanxiang Shen, Thinh H. Nguyen, Michelle M. Li, Yepeng Huang, Intae Moon, Nitya Nair, Daniel Marbach, & **Marinka Zitnik**
https://github.com/mims-harvard/COMPASS
https://github.com/mims-harvard/COMPASS-web

Phyla: Towards a Foundation Model for Phylogenetic Inference
Marinka Zitnik & Yasha Ektefaie
https://github.com/mims-harvard/Phyla
https://github.com/mims-harvard/PhylaNeurips

Efficient Generation of Protein Pockets with PocketGen
Zaixi Zhang, Wan Xiang Shen, Qi Liu & **Marinka Zitnik**
https://github.com/zaixizhang/PocketGen

LEAPS: A discrete neural sampler via locally equivariant networks
Peter Holderrieth, **Michael S. Albergo**, Tommi Jaakkola
https://github.com/malbergo/leaps

Traveling Waves Integrate Spatial Information Through Time
Mozes Jacobs, Roberto C. Budzinski, Lyle Muller, **Demba Ba,** & **T. Anderson Keller**
https://github.com/KempnerInstitute/traveling-waves-integrate

Benchmarking NVIDIA HPC Workloads on the Kempner AI Cluster
Naeem Khoshnevis
https://github.com/KempnerInstitute/nvidia-hpc-benchmarks
https://github.com/KempnerInstitute/scalable-vision-workflows

Can interpretation predict behavior on unseen data?
Victoria R. Li, Jenny Kaufmann, Martin Wattenberg,

David Alvarez-Melis, Naomi Saphra
https://github.com/vli31/id-predict-ood?utm_source=catalyzex.com

Polypythias: Stability and outliers across fifty language model pre-training runs
Oskar van der Wal, Pietro Lesci, Max Muller-Eberstein, **Naomi Saphra**, Hailey Schoelkopf, Willem Zuidema, & Stella Biderman
https://github.com/EleutherAI/pythia?utm_source=catalyzex.com
https://github.com/eleutherai/gpt-neox?utm_source=catalyzex.com
https://github.com/EleutherAI/lm-evaluation-harness?utm_source=catalyzex.com

Using Shapley interactions to understand how models use structure.
Divyansh Singhvi, Diganta Misra, Andrej Erkelens, Raghav Jain, **Isabel Papadimitriou**, & **Naomi Saphra**
https://github.com/jaekookang/p2fa_py3?utm_source=catalyzex.com

Recite, reconstruct, recollect: Memorization in LMs as a multifaceted phenomenon.
USVSN Sai Prashanth, Alvin Deng, Kyle O'Brien, Jyothir S V, Mohammad Aflah Khan, Jaydeep Borkar, Christopher A. Choquette-Choo, Jacob Ray Fuehne, Stella Biderman, Tracy Ke, Katherine Lee, & **Naomi Saphra**
https://github.com/EleutherAI/semantic-memorization?utm_source=catalyzex.com

How to visualize training dynamics in neural networks.
Michael Y. Hu, Shreyans Jain, Sangam Chaulagain, & **Naomi Saphra**
https://github.com/shreyansjainn/visualizing-training/blob/quickstart/blog_post.ipynb

A taxonomy of transcendence
Natalie Abreu, Edwin Zhang, **Eran Malach,** & **Naomi Saphra**
https://tecunningham.github.io/posts/2023-09-05-model-of-ai-imitation.html?utm_source=catalyzex.com

EconEvals: Benchmarks and Litmus Tests for LLM Agents in Unknown Environments
Sara Fish, Julia Shephard, Minkai Li, Ran I. Shorrer, Yannai A. Gonczarowski
https://github.com/sara-fish/econ-evals-paper

Generative Social Choice: The Next Generation
Niclas Boehmer, **Sara Fish**, Ariel D. Procaccia
https://github.com/sara-fish/gen-soc-choice-next-gen

Updated Generative Social Choice replication package
Sara Fish
https://github.com/generative-social-choice/gsc_abortion

MiWaves Reinforcement Learning Algorithm
Susan Murphy
https://github.com/StatisticalReinforcementLearningLab/miwaves_rl_service?utm_source=catalyzex.com
https://github.com/StatisticalReinforcementLearningLab/JustIn_RL_API

Justin Rev 2a
Susan Murphy
https://d3c.isr.umich.edu/justin-rev-2a/

Flow Equivariant Recurrent Neural Networks
T. Anderson Keller
https://github.com/akandykeller/FERNN

Overcomplete: A Vision-based SAE PyTorch Library
Thomas Fel
https://github.com/KempnerInstitute/overcomplete

Archetypal SAE: Adaptive and Stable Dictionary Learning for Concept Extraction in Large Vision Models

Thomas Fel, Ekdeep Singh Lubana, Jacob S. Prince, Matthew Kowal, Victor Boutin, **Isabel Papadimitriou**, **Binxu Wang**, Martin Wattenberg, **Demba Ba**, & **Talia Konkle**
https://huggingface.co/matybohacek/RA-SAE-DINOv2-32k

Universal Sparse Autoencoders: Interpretable Cross-Model Concept Alignment
Harrish Thasarathan, Julian Forsyth, **Thomas Fel**, Matthew Kowal, Konstantinos Derpanis
https://github.com/YorkUCVIL/UniversalSAE

An Adaptive Orthogonal Convolution Scheme for Efficient and Flexible CNN Architectures
Thibaut Boissin, Franck Mamalet, **Thomas Fel**, Agustin Martin Picard, Thomas Massena, Mathieu Serrurier
https://github.com/deel-ai/orthogonium?utm_source=catalyzex.com

One Wave To Explain Them All: A Unifying Perspective On Feature Attribution
Gabriel Kasmi, Amandine Brunetto, **Thomas Fel**, Jayneel Parekh
https://github.com/gabrielkasmi/wam

Interpreting the Linear Structure of Vision-Language Model Embedding Spaces
Isabel Papadimitriou, Huangyuan Su, **Thomas Fel**, **Sham Kakade**, Stephanie Gil
https://github.com/KempnerInstitute/overcomplete

Unearthing Skill-Level Insights for Understanding Trade-Offs of Foundation Models
Mazda Moayeri, Vidhisha Balachandran, Varun Chandrasekaran, Safoora Yousefi, **Thomas Fel**, Soheil Feizi, Besmira Nushi, Neel Joshi, Vibhav Vineet
https://github.com/microsoft/skill-slice-insights?utm_source=catalyzex.com

Uncovering Conceptual Blindspots in Generative Image Models Using Sparse Autoencoders
Matyas Bohacek, **Thomas Fel**, Maneesh Agrawala, Ekdeep Singh Lubana
https://github.com/maty-bohacek/conceptual-blindspots

TATM: Repository for large datasets. Handles storage of our large text and vision datasets in our cluster
Timothy Ngotiaoco
https://github.com/KempnerInstitute/tatm

vLLM Workflow: Describes how to perform quick distributed inference, particularly of large models, using vLLM
Timothy Ngotiaoco
https://github.com/KempnerInstitute/distributed-inference-vllm

Towards unifying interpretability and control: Evaluation via intervention
Usha Bhalla, Suraj Srinivas, Asma Ghandeharioun, Himabindu Lakkaraju
https://github.com/jbloomAus/SAELens?utm_source=catalyzex.com
https://github.com/AI4LIFE-GROUP/interp_interv?utm_source=catalyzex.com

ChatGPT Doesn't Trust Chargers Fans: Guardrail Sensitivity in Context
Victoria R. Li, Yida Chen & **Naomi Saphra**
https://github.com/vli31/llm-guardrail-sensitivity

NiceWebRL: a Python library for human subject experiments with reinforcement learning environments
Wilka Carvalho, Vikram Goddla, Ishaan Sinha, Hoon Shin, Kunal Jha
https://github.com/KempnerInstitute/nicewebrl

Ponce Rotation
Yutaka Sprague
https://github.com/dysprague/Ponce_rotation

Fixed Point Finder
Yutaka Sprague
https://github.com/dysprague/fixed-point-finder

Campy CLIRB
Yutaka Sprague
https://github.com/dysprague/campy-CLIRB

LRM Steering
George Alvarez
https://github.com/harvard-visionlab/lrm-steering

Asymptotic theory of in-context learning by linear attention
Yue M. Lu, Mary Letey, Jacob A. Zavatone-Veth, Anindita Maiti, & **Cengiz Pehlevan**
https://github.com/Pehlevan-Group/incontext-asymptotics-experiments

Scaling Offline RL via Efficient and Expressive Shortcut Models
Nicolas Espinosa-Dice, Yiyi Zhang, Yiding Chen, Bradley Guo, Owen Oertell, Gokul Swamy, **Kiante Brantley**, Wen Sun
https://github.com/nico-espinosadice/SORL

Interpretable deep learning for deconvolutional analysis of neural signals
Bahareh Tolooshams, Sara Matias, Hao Wu,..., Venkatesh N. Murthy, Paul Masset, **Demba Ba**
https://github.com/btolooshams/dunl-compneuro

SELECTED PRESS RELEASES

The Harvard Gazette

Leading computational neuroscientist to join Kempner Institute, Center for Brain Science

January 22, 2025

CAMBRIDGE, MA —The Kempner Institute announced today the appointment of SueYeon Chung (PhD '17), who returns to Harvard as a Kempner Institute Investigator and faculty member in the Faculty of Arts and Sciences (FAS) Center for Brain Science (CBS). In addition to her appointments within the Kempner Institute and CBS, Chung will hold a faculty position as Assistant Professor of Physics in the FAS with a joint appointment in Applied Mathematics in the Harvard John A. Paulson School of Engineering and Applied Sciences (SEAS).

"Dr. Chung's work is at the forefront of the rapidly-progressing field of NeuroAI," said Kempner co-director Bernardo Sabatini. "Her theoretical insights into neural representations in artificial and biological networks have the potential to change the way we think about structure and function of these networks, and have important implications for the work that the Kempner Institute is doing to advance our understanding of the basis of intelligence in natural and artificial systems."

"We are thrilled to have SueYeon Chung join the CBS and Kempner community," said Venkatesh Murthy, Paul J. Finnegan Family Director of the Center for Brain Science. "She is a rising star in theoretical and computational neuroscience, combining approaches and insights from statistical physics and machine learning to advance our understanding of natural and artificial intelligence. She will deeply enrich our growing network of collaborative scientists spread across many departments and schools, and will undoubtedly attract the smartest young trainees to our community in this exploding area of research."

"SueYeon Chung's appointment marks an exciting collaboration between CBS, Kempner, FAS, and SEAS. Her interdisciplinary work in AI, physics, and applied math provides a new understanding of the computational underpinnings of cognition in brains and artificial networks," said Kenneth Blum, Executive Director of the Center for Brain Science.

Chung will begin her appointment at Harvard in July 2025.

Research at the intersection of machine learning and neuroscience

Chung's research explores the principles of neural computation in the brain and artificial neural networks (ANNs). Combining theoretical neuroscience, machine learning, and statistical physics, she investigates how neural systems represent, transform, and compute information. Her work focuses on two key approaches: (1) developing mathematical theories to capture neural representations, with an emphasis on the geometry of neural population activities through neural manifolds, and (2) creating ANN-based models with neurally plausible architectures and biologically-inspired learning rules.

By connecting multiple levels of understanding—from the activity of single neurons to the collective dynamics of neural populations and emergent cognitive functions—Chung's research reveals the structural principles underlying neural computation. This integrated approach bridges biological and artificial systems while laying a foundation for designing interpretable, efficient, and robust AI systems inspired by the brain's computational strategies.

At the Kempner Institute, Chung and her team will investigate a variety of topics at the forefront of NeuroAI, including:

- How neural manifolds reveal shared principles and alignment between brains and artificial networks
- How the structures of neural representations adapt to efficiently handle different tasks
- How learning and connections in the brain shape the flow of information within and across brain regions
- How brain-inspired mechanisms enhance the reliability and robustness of AI models
- How methods from neuroscience and physics can help interpret AI systems

Chung is currently an Assistant Professor of Neural Science at New York University, and project leader at the Center for Computational Neuroscience, Flatiron Institute, Simons Foundation. She is also part of the CILVR (Computational Intelligence, Learning, Vision, and Robotics) Group, and an affiliated faculty member at the Center for Data Science, and the Cognition & Perception program at NYU. Before joining NYU/Flatiron, she was a postdoctoral researcher in the Center for Theoretical Neuroscience at Columbia University, where she was advised by Larry F. Abbott. Prior to that, she was a fellow in computation in the Department of Brain and Cognitive Sciences at Massachusetts Institute of Technology, where she worked with Jim DiCarlo and Josh McDermott. Chung received a Ph.D. in applied physics at Harvard, where she was supervised by Haim Sompolinsky and co-advised by Ryan P. Adams. She studied physics and mathematics as an undergraduate at Cornell University.

Chung was selected as a 2024 Sloan Research Fellow by the Alfred P. Sloan Foundation and earned a 2023 Klingenstein-Simons Fellowship Award in Neuroscience. Her work is also supported by a BRAIN Initiative R01 Award from the National Institutes of Health.

The Harvard Gazette

Kempner Institute welcomes spring undergrad student researchers

A group of spring 2025 KURE undergraduate researchers at the Kempner Institute during the February KURE program orientation.
Photo by Lani O'Donnell

February 21, 2025

Cambridge, MA – The Kempner Institute for the Study of Natural and Artificial Intelligence at Harvard is pleased to announce the spring 2025 recipients of the Kempner Undergraduate Research Experience (KURE). KURE awards Harvard undergraduate students funding for term-time research supervised by Kempner-affiliated faculty during the fall and spring semesters of the academic year.

Student research projects investigate the foundations of intelligence, including mathematical and computational models of intelligence, cognitive theories of intelligence, and the neurobiological basis of intelligence, as well as applications of artificial intelligence from a scientific or engineering perspective.

In addition to its term-time undergraduate research program, the Kempner offers a 10-week residential summer program called KRANIUM, providing a small cohort of undergraduates with a formative research experience under the supervision of a Kempner-affiliated faculty member.

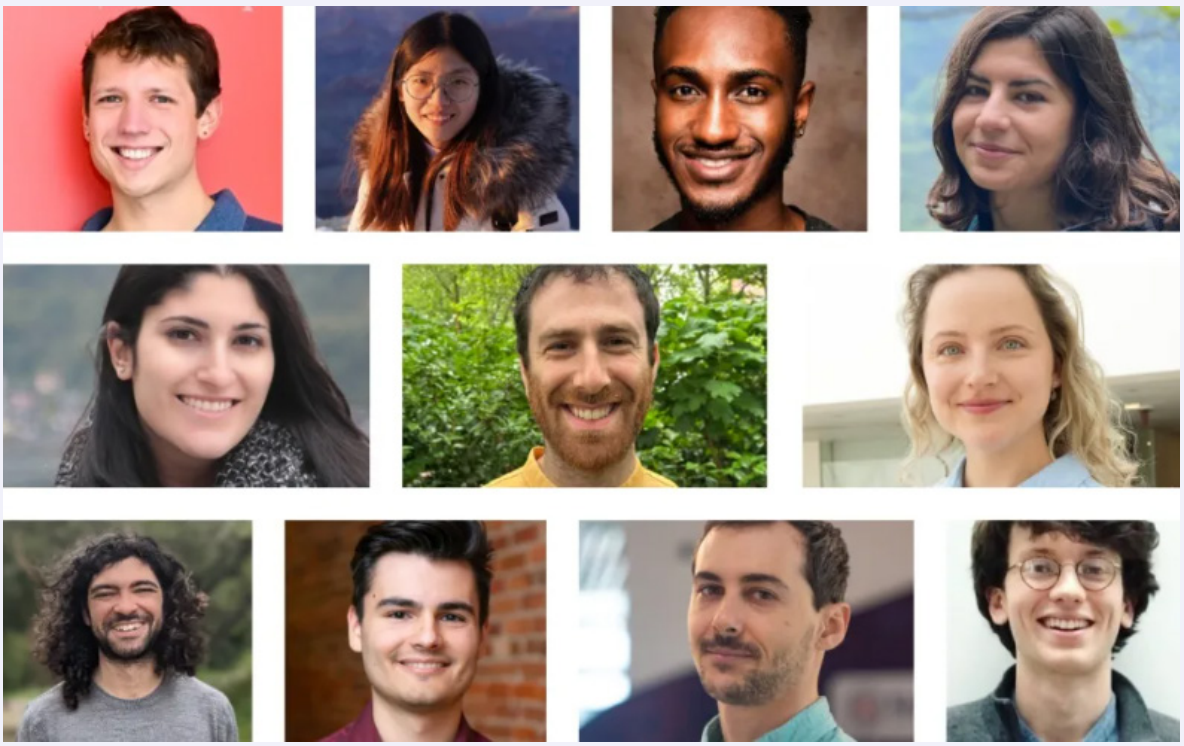
The spring 2025 KURE recipients (listed below) represent the third cohort of undergraduates to participate in the Kempner's term-time undergraduate research program. More information about KURE and KRANIUM can be found on the Kempner Institute website.

Spring 2025 KURE Award Recipients

KURE Student	Concentration	Supervisor/Mentor	Project Title
Rafay Azhar '25	Computer Science	Supervisor: Mengyu Wang	Optimizing Vision-Language Models for Egocentric Data to Assist the Visually Impaired
Joey Bejjani '26	Computer Science	Supervisors: Sham Kakade and Yilun Du	Adaptive Multiagent Debate With Latent Space Reasoning
Drake Du '26	Statistics and Computer Science	Supervisor: Mengyu Wang	Time-SAW: Time-Smoothing Adaptive Weighting for Flow-Based Diffusion Model
David Ettel '26	Mathematics	Supervisor: Melanie Weber	Doe Approximate Equivariance Matter at Scale
Emma Finn '26	Math and Classics with Concurrent AM in Statistics	Supervisor: Demba Ba; Mentors: Andy Keller and Manos Theodosis	From Noise to Novelty: The Origins of Creativity in Attention Based Diffusion Models
Ryland Gross '26	Mathematics	Supervisor: Nada Amin	Machine-Made Mathematics: Applying LLMs to Lean4 for Automated Proof Generation
Bright Liu '26	Mathematics	Supervisor: Sham Kakade; Mentor: Rosie Zhao	Investigating the Tradeoff Between Verification and Generation in Scaling Test-Time Compute
Jasmine Liu '28	Computer Science	Supervisor: Jie Yang	Mitigating Hallucinations in Medical LLMs with Knowledge-Enhanced Reasoning
Adithya Madduri '27	Molecular & Cellular Biology and Statistics	Supervisor: Bernardo Sabatini; Mentor: Kimberly Reinhold	Automated Behavioral Analysis of Skilled Forelimb Reaching Behaviors in Mice
Teodor Malchev '27	Computer Science	Supervisor: Nada Amin	Using Linguistics to Evaluate and Improve LLM
Sean Meng '26	Neurobiology	Supervisor: Bernardo Sabatini; Mentor: Kevin J Mastro	Contrastive Learning and Transformer-Based Multimodal AI for Predicting Cognitive Resilience and Decline in Aging Mice
Aoi Otani '25	Integrative Biology	Supervisor: Nada Amin; Mentor: Morgan Talbot	Mitigating Catastrophic Forgetting and Mode Collapse in Continual Text-to-Image Diffusion via Latent Replay
Purab Seth '26	Computer Science	Supervisor: Wilka Carvalho	Multi-agent Reinforcement Learning for Dynamic Role
Lillian Sun '26	Computer Science	Supervisors: Sham Kakade and Yilun Du	Optimizing Inter-Model Communication in Multi-Agent Systems
Mira Yu '27	Computer Science and Government	Supervisor: Finale Doshi-Velez	AI for Humanitarian Crisis Negotiation and Beyond
Eric Xu '28	Undeclared	Supervisor: Venkatesh Murthy	Using Clique Topology and Co-Occurrence Statistics to Determine Olfactory Geometry and Parameters
Richard Zhu '26	Statistics	Supervisor: Marinka Zitnik; Mentor: Shanghua Gao	RL-Inspired Framework for Uncertainty-Aware Agentic Reasoning

The Harvard Gazette

Announcing 2025 Kempner Institute Research Fellows



The 2025 Kempner Research Fellows are (left to right, from top): David Clark, Ruojin Cai, Elom Amemastro, Gizem Ozdil, Hadas Orgad, Mark Goldstein, Greta Tuckute, Gabriel Poesia, Alexandru Damian, Richard Hakim, and William Dorrell.

July 16, 2025

Cambridge, MA – The Kempner Institute for the Study of Natural and Artificial Intelligence at Harvard is pleased to announce the recipients of its 2025 Kempner Institute Research Fellowships. The 2025 recipients are Elom Amemastro, Ruojin Cai, David Clark, Alexandru Damian, William Dorrell, Mark Goldstein, Richard Hakim, Hadas Orgad, Gizem Ozdil, Gabriel Poesia and Greta Tuckute.

The eleven fellowship recipients are all early-career scientists drawn from a wide variety of skillsets and educational backgrounds. Each of them pursues novel research at the intersection of natural and artificial intelligence.

Each fellowship runs for up to three years and includes salary and research funds, office space, and mentorship. While fellows set their own research agenda, they are strongly encouraged to undertake interdisciplinary projects and to collaborate with experts at the Kempner Institute and throughout Harvard University.

About the fellows

Elom Amemastro focuses on how humans learn new skills from experience, with an emphasis on reinforcement learning and the neural mechanisms that support flexible skill acquisition and generalization. His work integrates behavioral experiments, large-scale neural recordings, and computational modeling to investigate how neural circuits encode structured knowledge, adapt to novel tasks, and leverage past experiences for efficient learning. His research addresses fundamental challenges in understanding how biological systems achieve sample-efficient learning, with the goal of developing AI systems that learn more like humans and informing interventions that enhance cognitive flexibility and recovery.

Ruojin Cai studies 3D computer vision, with the goal of building models that can perceive and reason about the 3D world, advancing spatial intelligence in machines. She studies core challenges in 3D reconstruction under sparse-view or ambiguous settings, where traditional geometric methods often fail. Her key insight is to address these challenges

by leveraging learned priors from generative video models and geometric vision models to improve robustness under limited or ambiguous observations. Her long-term goal is to develop truly spatially intelligent systems that can not only perceive but also reason about and act within complex, real-world environments.

David Clark aims to develop a deeper theoretical understanding of how neural systems interact with environmental data to generate representations and perform useful computations. Neural circuits in the brain are characterized by their large-scale, nonlinear dynamics, complex recurrent interactions, and connections that change across multiple timescales. To understand how these features enable computation and learning, he uses theoretical tools from physics (e.g., dynamical systems, statistical mechanics, path integrals, replica and cavity methods, and random matrix theory) and machine learning (e.g., convex optimization, deep and recurrent networks, and sequence models).

Alexandru Damian aims to develop a mathematical foundation for deep learning, with a focus on optimization and representation learning. He is especially interested in how optimization algorithms, such as stochastic gradient descent and Adam, navigate the high-dimensional non-convex loss landscapes in deep learning, how this process is influenced by the choice of optimizer and its hyperparameters, and how these decisions shape the representations learned by the model.

William Dorrell tries to understand how biological neurons implement cognitive computations. His approach to this involves asking why neurons fire the way they do, and building mathematical theories to try and understand this. These mathematical theories are usually optimization problems, leading to hypotheses like: "if the neurons were trying to perform this computation optimally then, under some constraints, they should behave like this." Dorrell compares the predictions of these theories to neural recordings from brains or artificial neural networks. He hopes these approaches will help in understanding the algorithms the brain uses to do clever things like play board games, tap rhythms, or reason.

Mark Goldstein studies generative modeling, probability density estimation, and sampling. A central theme of his work is rethinking foundational design choices in generative model training and analyzing how these choices affect performance and efficiency. He has explored these questions in the context of diffusion and consistency models. Looking ahead, Mark aims to investigate how compositionality and sequential decision-making can augment existing model classes—or give rise to new ones. Such capabilities could also be integrated with simulation—for example, in the context of scientific discovery, a diffusion or language model that simulates molecular dynamics to observe a phenomenon before refining a proposed molecule.

Richard Hakim studies neural decoding and brain-computer-interfaces (BCIs). His prior work is primarily experimental and includes studies on how movement is encoded in the motor cortex, how brain oscillations are generated, and the development of a suite of open-source computational tools. Moving forward, he is excited by emerging research into foundation models for BCI decoding and aims to leverage principles derived from artificial intelligence to better understand the structure and function of biological brains.

Hadas Orgad investigates the internal mechanisms of AI models to better understand and mitigate failures in safety, fairness, and reliability. Her research bridges interpretability and practical deployment, focusing on harmful model behaviors such as hallucinations, bias, privacy violations, and unsafe outputs. By analyzing the internal structure of models, she develops actionable tools and interventions to improve model behavior and better align it with human values and incentives. Her long-term goal is to advance interpretability and control techniques so that AI systems are fully transparent, trustworthy, and steerable.

Gizem Ozdil bridges systems neuroscience, artificial intelligence, and robotics to uncover the principles that enable adaptive behavior in biological systems. She is particularly interested in how biological insights, such as structural constraints, can inform the design of more flexible and autonomous agents. To explore this, she develops biologically inspired neural networks and trains embodied agents in complex physical environments that require learning, memory retention, and planning. In turn, these models can be used for reverse-engineering brain function and inspiring the development of more efficient and adaptable artificial systems.

Gabriel Poesia investigates formal reasoning in humans and machines. This involves defining a suitable "game of mathematics" on top of a formal foundation like dependent type theory; learning to find proofs using language models and deep reinforcement learning; discovering increasingly high-level mathematical abstractions; and ultimately using these tools to build joyful and scalable experiences for mathematics education. His recent research builds heavily on ideas from intrinsically motivated learning, and also explores program verification.

Greta Tuckute studies how language is processed in the human brain and in artificial neural networks. Her research broadly follows three directions. First, she works to precisely characterize the neural architecture and functions that support language processing in the human brain. Second, she investigates whether the human brain and artificial networks share representations and computational principles during language processing. Third, she develops biologically plausible artificial networks that learn language in more human-like ways. Collectively, these three directions inform one another, advancing our understanding of how language serves as an efficient interface to a wide range of downstream behaviors in both biological and artificial systems.

The Harvard Gazette

Qianqian Wang, an expert in computer vision, to join Kempner Institute, SEAS

As a Kempner Institute Investigator, Qianqian Wang will explore the future of human vision: how machines can learn to see, understand, and interact with the world like humans do.

August 14, 2025

CAMBRIDGE, MA —The Kempner Institute announced today the appointment of Qianqian Wang, who will join Harvard as Kempner Institute Investigator and Assistant Professor of Computer Science at the John A. Paulson School of Engineering and Applied Sciences (SEAS).

At Harvard, Wang will explore the future of human vision: how machines can learn to see, understand, and interact with the world like humans do. "Dr. Wang's research is at the forefront of enabling machines to perceive the world as we do—not as a static image, but as a dynamic 4D environment of space and time," said Sham Kakade, Kempner Institute co-director and Rampell Family Professor of Computer Science and Professor of Statistics at SEAS. "Her pioneering work is fundamental for building the next generation of intelligent systems that can learn, navigate, and interact safely in our constantly changing world."

"Qianqian Wang's research has led to major advances in computer vision, specifically in the central challenge of reconstructing our three-dimensional world from the two-dimensional images we get from cameras—and from our eyes," said Stuart Shieber, Area Chair for Computer Science and James O. Welch, Jr. and Virginia B. Welch Professor of Computer Science at SEAS. "We are fortunate to have Qianqian join our AI and machine learning initiatives in the computer science program and we look forward to her impactful contributions to our teaching and research efforts in computer vision, graphics, and artificial intelligence."

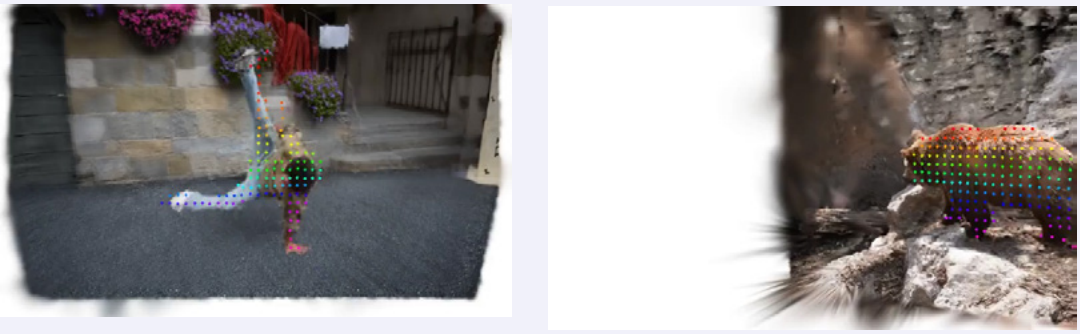
Wang will begin her appointment at Harvard in the fall of 2026.

Research at the frontiers of computer vision

At Harvard, Wang will investigate a variety of topics at the forefront of computer vision, including long-form video understanding, visual reasoning, spatial and temporal memory, 3D scene perception, and active visual perception. (Read more about Wang's research on her website.)

Building models that push the boundaries of visual intelligence "means asking big questions about how we perceive and reason about the physical world, and how those processes can be realized in intelligent systems," says Wang. "Tackling these challenges requires not only innovation in AI but also deep insights into how the human brain works."

Wang is currently a postdoctoral researcher at the Department of Electrical Engineering and Computer Sciences at the University of California, Berkeley, where she works with Angjoo Kanazawa and Alexei A. Efros. She completed her Ph.D. in Computer Science at Cornell University under advisors Noah Snavely and Bharath Hariharan. Wang received her bachelor's degree from Zhejiang University, working with Xiaowei Zhou.



Qianqian Wang builds visual computer models that reconstruct our four-dimensional world from the two-dimensional images we get from cameras, videos, and from our eyes. Above, two examples of Wang's reconstructions.
Credit: Qianqian Wang