2ND WORKSHOP ON SCIENTIFIC METHODS FOR UNDERSTANDING DEEP LEARNING (SCI4DL)

ABSTRACT

While deep learning continues to achieve impressive results on an ever-growing range of tasks, our understanding of the principles underlying these successes remains largely limited. This problem is usually tackled from a mathematical point of view, aiming to prove rigorous theorems about optimization or generalization errors of standard algorithms, but so far they have been limited to overly-simplified settings. The main goal of this workshop is to promote a complementary approach that is centered on the use of the scientific method, which forms hypotheses and designs controlled experiments to test them. More specifically, it focuses on empirical analyses of deep networks that can validate or falsify existing theories and assumptions, or answer questions about the success or failure of these models. This approach has been largely underexplored, but has great potential to further our understanding of deep learning and to lead to significant progress in both theory and practice. The secondary goal of this workshop is to build a community of researchers, currently scattered in several subfields, around the common goal of understanding deep learning through a scientific lens.

1 MOTIVATION AND DESCRIPTION

Over the last decade, deep learning has brought about astonishing improvements in computer vision, signal processing, language modeling and beyond. This unprecedented success is due to a huge collective endeavor in building large-scale models. But the remarkable performance of these models has largely surpassed our fundamental understanding of them.

An established approach to analyzing these deep learning models is to build a theory from the bottom-up, with rigorous proofs explaining all aspects including architecture, datasets, objectives and regularizers, and optimization. However, such mathematical models are usually only tractable when they are oversimplified. Although valuable progress has been made on that front, the complexity of real high-dimensional data and deep network architectures used in practice makes these models resistant to traditional mathematical analysis. Hence, many aspects remain mysterious, and our understanding of their success and failure modes remains very limited.

This workshop aims to promote a complementary approach to further our understanding of deep learning through the lens of the scientific method. This approach uses carefully designed experiments to answer precise questions about how and why deep learning works. The scientific method has been used successfully in the past to validate or falsify hypotheses (e.g., deep networks generalize because they cannot fit random labels (Zhang et al., 2016)), challenge common assumptions (e.g., deep networks are robust to imperceptible input perturbations (Szegedy et al., 2014)), or reveal empirical regularities (e.g., discovering scaling laws (Kaplan et al., 2020)). These well-known examples all crucially rely on controlled experiments and constitute an important part of our current understanding of deep learning.

These endeavors aim neither to improve the state-of-the-art nor to prove theorems, but they have had a profound impact, spurring many follow-up theoretical and applied studies. Indeed, such results serve the theory of deep learning by grounding it with empirical observations (e.g., training occurs on the edge of stability (Cohen et al., 2020)) or formulating conjectures (e.g., the lottery ticket hypothesis (Frankle and Carbin, 2019)). Simultaneously, they lead to practical improvements by informing engineering decisions, e.g., compute-optimal scaling (Hoffmann et al., 2022), and spurring new research directions, e.g., adversarial robustness (Gu and Rigazio, 2014). Thus, we believe that this workshop will be of interest to both theoretical and applied communities.

Despite their significant impact, scientific approaches have been largely underexplored and underappreciated. While the criteria for assessing numerical contributions such as improving state-of-the-art performance or for proving rigorous theorems are more clear-cut, assessing the significance of contributions within this category is still developing. Our workshop offers a venue for studies that fall outside the standard acceptance criteria yet have a high impact potential. Thus, our workshop significantly differs from and complements past workshops accepted at main machine learning conferences.

The scientific study of deep learning is currently scattered across several subfields, including in-context learning in transformers, generalization properties of generative models, inductive biases of learning algorithms, (mechanistic) interpretability, empirical studies of loss landscapes, training dynamics, and learned weights and representations. This workshop aims to facilitate communication and collaboration across subfields by building a community centered around a common approach.

The first edition of this workshop at NeurIPS 2024 laid the foundation for realizing these goals. It was enthusiastically endorsed by workshop reviewers, participants, invited speakers, and panelists. Both reviewers emphasized that the topic is not only timely but arguably overdue. We received 85 submissions, 73 of which were accepted, leading to a vibrant and engaging poster session. The turnout exceeded our expectations, with approximately 700 unique attendees throughout the day (1,773 conference attendees registered interest on Whova, more than all workshops in 2024). Notably, the workshop also inspired one of our panelists to write a follow-up review article (Wilson, 2025). We are excited to build on this momentum. Given the rapid rate of development in the field, we believe the topic remains highly relevant and holds significant potential for impact. As for the first edition, we have carefully chosen speakers and panelists to represent a diverse range of perspectives and approaches to understanding deep neural networks in both vision and language domains, drawing from expertise across academia and industry.

Comparison with related workshops. We wish to highlight the differences between our proposal and some previously accepted workshops at major ML conferences:

- Theory-focused workshops aim to understand deep learning either through formal analysis (e.g., Mathematics of Modern Machine Learning at NeurIPS 2024, Theoretical Foundations of Foundation Models at ICML 2024) or by drawing on theoretical physics (e.g., Machine Learning and the Physical Sciences series at NeurIPS, Physics for Machine Learning at ICLR 2023). In contrast, we emphasize experimental validation of theories and assumptions, following the methodology of the experimental sciences.
- Other workshops focus on empirical methods, such as the I Can't Believe It's Not Better series (most recently ICLR 2025) or Mathematical and Empirical Understanding of Foundation Models at ICLR 2023–2024. These workshops are centered on empirical failures and improving performance (with a recent focus on foundation models) rather than building a fundamental understanding of why deep learning models work.
- Mechanistic Interpretability (e.g., at ICML 2024 and NeurIPS2025) aims to reverse-engineer network behaviors into human-interpretable mechanisms. While valuable, this work is often descriptive, rather than driven by falsifiable hypotheses.
- There have been a series of workshops focused on using AI for science at NeurIPS or ICML 2021–2025. Our workshop rather aims to use the scientific method for understanding AI.
- Bridging the Gap Between Practice and Theory in Deep Learning at ICLR 2024 shares our goal in contributing to both theory and practice, by bringing together researcher from both areas. We propose that scientific experimentation is the missing piece in bridging this gap.

2 Schedule and Logistics

We include a tentative schedule below. **All invited speakers and panelists have been confirmed.** Contributed papers are presented in two poster sessions, and the four best submissions will be showcased in 15-minute talks, for which we will prioritize works by junior authors. In addition, submitted papers can enter our *"novel phenomenon"* challenge and win monetary prizes. We will moderate a panel discussion, in which we will reserve time for audience questions. Finally, we will organize a *mentorship program*, in which participants, especially early stage students, will be

Time	Session	Notes
9:00	**Invited talk 1**	Jascha Sohl-Dickstein (Anthropic)
9:30	**Invited talk 2**	Preetum Nakkiran (Apple)
10:00	**Invited talk 3**	Jeremy Cohen (Flatiron Institute)
10:30	**Coffee Break**	(mentorship program)
11:00	**Contributed talks 1 & 2**	15 minutes each
11:30	**Poster Session 1**	
12:30	**Lunch Break**	
13:30	**Contributed talks 3 & 4**	15 minutes each
14:00	**Invited talk 4**	Julia Kempe (NYU & Meta)
14:30	**Invited talk 5**	Matthieu Wyart (Johns Hopkins University)
15:00	**Coffee Break**	(mentorship program) Lenka Zdeborová (EPFL) Jascha Sohl-Dickstein (Anthropic)
15:30	**Panel Discussion**	Julia Kempe (NYU & Meta) Richard Baraniuk (Rice University & OpenStax) Matthieu Wyart (Johns Hopkins University)
16:30	**Competition winners**	
17:00	**Poster Session 2**	

matched with speakers and panelist who serve as mentors for conversation around main areas of the workshop.

Submissions. We expect around 150 submissions (in short form, 4 pages) and 700 participants. The call for papers will be worded to welcome work-in-progress submissions that may fall short of full conference paper requirements. Similarly, reviewer guidelines will emphasize checking for topical fit and correctness, as well as providing useful feedback to authors. Submissions will be handled on OpenReview, with a deadline of January 30th, 2026. The organizers will oversee the reviewing process and handle potential conflicts of interest using OpenReview's reviewer matching. The reviewing period will be February 3rd-February 24th, with final decisions released to authors on March 1st. All accepted papers are non-archival, will receive a poster presentation (on-site) and be made public on OpenReview, with the four best papers selected for contributed talks.

The Novel Phenomenon Challenge. We will hold a competition for identifying new empirical phenomena in deep learning systems. It focuses on identifying previously unknown behaviors, patterns, or regularities in deep learning through careful experimentation. Discoveries may span any aspect of deep learning, from emergent representations and optimization dynamics (Cohen et al., 2020) to generalization patterns (Zhang et al., 2016) and mechanistic insights (Lindsey et al., 2025), as long as they reveal something interesting about how models work. Strong submissions will demonstrate that the phenomenon is robust, e.g., appearing across different seeds, architectures, scales, or modalities rather than being artifacts of specific implementations. Examples of impactful discoveries include scaling laws (Kaplan et al., 2020), double descent (Nakkiran et al., 2021), and the lottery ticket hypothesis (Frankle and Carbin, 2019). Participants can enter the competition by answering a series of questions we provide in an extra page attached to their submitted article framing their work in light of the challenge. Authors of the winning papers will receive a prize and an oral presentation. We held a similar competition in the first iteration of the workshop, focused on challenging commonly-held assumptions. 14 papers entered the competition. The winners showed that reusing data during training can generalize better than only using fresh data, for the same number of training steps. **We have secured $2,000 in funding from the Simons Foundation for competition prizes and coffee breaks.**

Diversity efforts By intentionally engaging with diverse perspectives, we aim to cultivate an inclusive and open space for all participants. This is reflected in our diverse group of organizers, invited speakers.

Organizers. We have a well-balanced composition of organizers, considering various factors such as seniority level, geographical diversity, and gender. Ethnically we originate from four continents:

Africa, Asia, Europe and North America, with three female organizers. Half of the organizers are from last year's team, while the other half is new.

Speakers and panelists. Our speakers and panelists were invited from academia and industry to maximize the diversity of perspectives on understanding deep learning in particular. They come from diverse research backgrounds which cover theory, practice, as well as physical and natural sciences. Additionally, two of our speakers and panelists are prominent female researchers.

Participants. To foster diverse participation, some of our organizers volunteered to give their registration voucher to participants from underrepresented countries based on demonstrated need. Although we require at least one author per accepted poster to attend, we recognize that funding or visa constraints may prevent attendance. In such cases, we will accommodate the authors by displaying their slides during the poster session, ensuring their research remains accessible to workshop participants.

3 Workshop Organizers

Zahra Kadkhodaie, Flatiron Institute (Google Scholar) zk388@nyu.edu
Zahra Kadkhodaie has a background in solid state physics and is currently a research fellow at the Center for Computational Neuroscience and Mathematics at the Flatiron institute. She completed her PhD at the Center for Data Science at New York University under the supervision of Eero Simoncelli. She has worked on learning high dimensional image densities from data, understanding and utilizing them. Specifically, she studied generalization versus memorization in densities embedded in diffusion models, low-dimensionality of conditional density models, interpreting effective dimensionality of image manifolds, and using learned image densities to solve inverse problems. Her recent work on generalization in diffusion models received an outstanding paper award in ICLR 2024. **She previously served as an organizer for the 2024 iteration of this workshop**.

Florentin Guth, NYU and Flatiron Inst. (Google Scholar) florentin.guth@nyu.edu
Florentin Guth is a Faculty Fellow in the Center for Data Science at NYU and a Research Fellow in the Center for Computational Neuroscience at the Flatiron Institute. Prior to this, he completed his PhD at École Normale Supérieure in Paris under the supervision of Prof. Stéphane Mallat, where he co-organized the seminar of the Center for the Science of Data. His research focuses on improving our scientific understanding of deep learning, e.g., understand how it escapes the curse of dimensionality, particularly in the context of image generative models. His work received an outstanding paper award at ICLR 2024. **He previously served as an organizer for the 2024 iteration of this workshop.**

Sanae Lotfi, New York University (Google Scholar) sl8160@nyu.edu
Sanae Lotfi is an incoming Research Scientist at Meta. She completed her PhD at NYU, working with Professor Andrew Gordon Wilson. Sanae works on the science of deep learning and focuses on understanding the generalization properties of deep neural networks using notions that relate to generalization such as model compression and loss surface analysis. Using insights about generalization, her goal is to build improved, scalable and robust deep learning models. Sanae's research has been recognized with numerous accolades including the ICML Outstanding Paper Award and the Microsoft and Google DeepMind Fellowships. **Sanae previously served as an organizer for the NeurIPS 2021 competition on Approximate Inference in Bayesian Deep Learning and co-organized the NeurIPS 2023 Muslims in ML affinity workshop. She also served as an organizer for the 2024 iteration of this workshop.**

Davis Brown, UPenn and National Lab (Google Scholar) davisrbr@seas.upenn.edu
Davis Brown is a first-year PhD student at the University of Pennsylvania advised by Eric Wong and Hamed Hassani, with broad research interests in AI safety, the science of deep learning, and their intersection. He is also a research scientist at Pacific Northwest National Laboratory (PNNL) on the math of AI assuredness team. At UPenn, he co-organizes a reading group focusing on mechanistic interpretability and through PNNL, he helped organize a long-running mathematics and data science

seminar. Davis has papers published in ICML, NeurIPS, EMNLP, etc. **He previously served as an organizer for the 2024 iteration of this workshop.**

ANTONIO SCLOCCHI, U. COLLEGE LONDON (Google Scholar) a.sclocchi@ucl.ac.uk
Antonio Sclocchi is a research fellow in the Gatsby Computational Neuroscience Unit at UCL, advised by Andrew Saxe. Before, he was a postdoctoral researcher at the École Polytechnique Fédérale de Lausanne, where he co-organized a weekly seminar of the Physics of Complex Systems Lab. His research interests lie in machine learning theory, where he studies models of structured data, diffusion models, and optimization dynamics. His background is in theoretical physics, and during his PhD at Université Paris-Saclay he studied critical phenomena in complex systems. His work is published in machine learning conferences (ICML, ICLR) and journals, including PNAS and PRL.

SHARVAREE VADGAMA, UNIVERSITY OF AMSTERDAM (Google Scholar) sharvaree.vadgama@gmail.com
Sharvaree Vadgama is a PhD student at Amsterdam Machine Learning Lab advised by Erik Bekkers and Jakub Tomczak. Her research interests lie in structured representation learning and generative modeling. **She previously served as organizer of the GRaM workshop at ICML 2024, co-organizer of Generative Modeling Summer School (GeMSS) 2023/2024, as well as of Women in AI meetups (Amsterdam chapter).**

JAMIE SIMON, IMBUE AND UC BERKELEY (Google Scholar) jsi@berkeley.edu
James (Jamie) Simon is a research fellow at Imbue and a postdoc at UC Berkeley who just completed his PhD under Mike DeWeese. He wants to work out fundamental, empirically grounded theory for deep learning. He is a physicist by training, and this background informs his approach to the science of ML. He received an NSF graduate fellowship and the Berkeley physics department's outstanding thesis award. **He now leads a small research team at UC Berkeley. His previous organizational experience includes creating and running several years of a campus-wide puzzlehunt for thousands of participants.**

EERO SIMONCELLI, NYU AND FLATIRON INST. (Google Scholar) eero.simoncelli@nyu.edu
Eero Simoncelli is a Silver Professor of Neural Science, Mathematics, Data Science, and Psychology at NYU and the Director of the Center for Computational Neuroscience at the Flatiron Institute of the Simons Foundation. He received his Ph.D. in Electrical Engineering and Computer Science from MIT in 1993, and served as a Howard Hughes Medical Institute Investigator from 2000 to 2020. His honors include election to the American Academy of Arts and Sciences (2019), the Minerva Foundation Golden Brain Award (2017), and an Engineering Emmy Award (2015). In 2024, he was awarded the prestigious Swartz Prize for contributions to Theoretical and Computational Neuroscience by the Society for Neuroscience.

Program Committee We include below an initial list of 41 program committee members. Additionally, we will ask authors of each submission to the workshop to nominate a reviewer among them. We believe this will fulfill the goal of having at least 3 reviews per submission, and keeping a maximum load of 3 papers per reviewer. Organizers will also serve as emergency reviewers if needed (while avoiding any conflicts of interest that may arise).

Brice Ménard (JHU), Rudy Morel (Flatiron Institute), Etienne Lempereur (ENS), Nathanaël Cuvelle-Magar (ENS), Michaël Sander (ENS), Raphaël Barboni (ENS), Valérie Castin (ENS), Sybille Marcotte (ENS), Zaccharie Ramzi (Meta), Anton Xue (UPenn), Ezra Edelman (UPenn), Alex Robey (CMU), Henry Kvinge (PNNL), Eric Michaud (MIT), Nicholas Konz (Duke), Lauro Langosco (Cambridge), Nicolas Fishman (Oxford University), Sahra Ghalebikesabi (Oxford University), Ya-Ping Hsieh (ETH Zurich), Michael Hutchinson (Oxford University), Leo Klarner (Oxford University), Guan-Horng Liu (Georgia Tech), Emile Mathieu (Cambridge University), Maxence Noble-Bourillot (Ecole Polytechnique), Angus Phillips (Oxford University), Yuyang Shi (Oxford University), Gowthami Somepalli (University of Maryland, College Park), Ella Tamir (Aalto University), Brian Trippe (Columbia University), Alexander Tong (MILA), Francisco Vargas (Cambridge University), Pierre-Etienne Fiquet (NYU), Thomas Yerxa (NYU), Isabel Garon (NYU), Edoardo Balzani (Flatiron Institute), David Lipshutz (Flatiron Institute), Jenelle Feather (Flatiron Institute), Nikhil Parthasarathy (Google DeepMind), Julie Xueyan Niu (NYU), Billy Broderick (Flatiron Institute), Ling-Qi Zhang (HHMI-Janelia).

REFERENCES

Cohen, J., Kaur, S., Li, Y., Kolter, J. Z., and Talwalkar, A. (2020). Gradient descent on neural networks typically occurs at the edge of stability. In *International Conference on Learning Representations*.

Frankle, J. and Carbin, M. (2019). The lottery ticket hypothesis: Finding sparse, trainable neural networks. In *International Conference on Learning Representations*.

Gu, S. and Rigazio, L. (2014). Towards deep neural network architectures robust to adversarial examples. *arXiv preprint arXiv:1412.5068*.

Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., Casas, D. d. L., Hendricks, L. A., Welbl, J., Clark, A., et al. (2022). Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. (2020). Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Lindsey, J., Gurnee, W., Ameisen, E., Chen, B., Pearce, A., Turner, N. L., Citro, C., Abrahams, D., Carter, S., Hosmer, B., Marcus, J., Sklar, M., Templeton, A., Bricken, T., McDougall, C., Cunningham, H., Henighan, T., Jermyn, A., Jones, A., Persic, A., Qi, Z., Thompson, T. B., Zimmerman, S., Rivoire, K., Conerly, T., Olah, C., and Batson, J. (2025). The biology of language models. *Transformer Circuits Thread*.

Nakkiran, P., Kaplun, G., Bansal, Y., Yang, T., Barak, B., and Sutskever, I. (2021). Deep double descent: Where bigger models and more data hurt. *Journal of Statistical Mechanics: Theory and Experiment*, 2021(12):124003.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. (2014). Intriguing properties of neural networks. In *International Conference on Learning Representations*.

Wilson, A. G. (2025). Deep learning is not so mysterious or different. *arXiv preprint arXiv:2503.02113*.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. (2016). Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations*.