Workshop on the Science of Training: Emergence, Prediction, and Steering (STEPS)

Isabelle Lee USC | **Emmy Liu** CMU | **Kaiser Sun** JHU | **Stella Biderman** Eleuther | **David Alvarez-Melis** Harvard

Tagline. Toward actionable training dynamics: describe trajectories, predict outcomes, steer training.
Website. `https://science-of-training.github.io/`

1 Workshop Description

Although current models are extensively evaluated after training, far less work has focused on characterizing how a training run unfolds. A substantial body of prior work predicts model performance using aggregate statistical relationships, such as scaling laws [19, 18, 1, 8, 23] and architecture-based performance estimation [32, 12, 20]. Recent studies of training dynamics shed light on how such patterns may emerge, but typically remain descriptive [9] or post-hoc [31], documenting *when* representations or behaviors emerge, often via manually crafted or heuristic evaluations. While these approaches have yielded valuable high-level insights, they do not capture the dynamics within the training trajectories.

A trajectory-level view of training is useful precisely because it surfaces phenomena that final perplexity, accuracy, or win-rates cannot. For example, models may perform well on downstream evaluations while becoming overtrained, making them difficult to fine-tune subsequently [26, 11]. If model flexibility could be quantified *during* training, such failure modes could be anticipated before they manifest downstream. In a similar vein, recent work on the interaction between fine-tuning and pretraining suggests that training dynamics can predict when adaptation will succeed, degrade, or induce forgetting [24, 27, 2]. Bridging the gap between understanding and practice requires moving beyond conceptual descriptions toward diagnosing failures that outcome-only evaluation misses and toward guiding interventions during training [5].

This workshop aims to encourage such types of work that can provide insights toward a dynamical and actionable understanding of training [5, 22]. We organize this around three escalating goals:

1. **Descriptive**: characterizing how representations, circuits, gradients, and behaviors evolve over the course of training.
2. **Predictive**: identifying signals early in training that reliably forecast later capabilities, failures, generalization, or structural changes.
3. **Prescriptive**: using predictive insight to actively steer, control, and repair training.

Prediction is the critical bridge between insight and action: to act on training, we should be able to anticipate how models will evolve with data, scale, or optimization. While descriptive analyses have yielded valuable conceptual insights that aided interpretations, description alone cannot enable control. The central goal, then, is to move from descriptive to predictive understanding–and to lay the groundwork for eventually translating predictive understanding into prescriptive actions.

We organize the workshop's topics of interest around the three goals detailed below and welcome work that shares similar interests:

1. Descriptive

- Representation formation, phase transitions, reorganization, and interaction across training
- Interpretability across time, scale, gradients, and optimizer dynamics
- Characterizing random variations and initial conditions [30, 25]

Submitted to 40th Conference on Neural Information Processing Systems (NeurIPS 2026). Do not distribute.

2. Predictive

- Early monitoring of training failures and criteria for stopping
- Stress-testing of early checkpoints and identifying misrepresentations that lead to downstream malformation
- Pretraining evaluations and invariants that predict post-training behavior across scale
- Generalization prediction: what transfers from which data, when, and why
- Training outcome prediction, including negative results

3. Prescriptive

- Training repair, such as reverting to and correcting from earlier checkpoints
- Causal and interventional methods for training [4]
- Phase specific data selection strategies [21, 28]
- Controlling model behavior during training

Relevant Workshops Our proposed workshop is closely related to the High-dimensional Learning Dynamics (HiLD) workshop series [14, 15, 16, 17], overlapping with our *descriptive* layer: representation formation, phase transitions, and interpretability across training. However, we substantially broaden this scope to include predicting from early failure signals and possible training repairs —what could make the science of training actionable [22]. Beyond descriptive analysis, we emphasize *predictive* and *prescriptive* dimensions of training dynamics, seeking work in both theory as well as how the theoretical insights inform practice, beyond HiLD's theoretical focus. We highlight the need for methods that *anticipate* training dynamics and outcomes (e.g., early detection of failure modes, generalization prediction, and checkpoint stress-testing) and that enable actions such as training repair, causal manipulation, and phase-aware data selection. A central goal is to bring together academic researchers and practitioners to identify which analyses are useful in practice, where current methods fall short, and what predictive or prescriptive tools would best support real-world model development.

2 Invited Speakers and Panelists

SPEAKERS

Andrew Saxe <a.saxe@ucl.ac.uk> is a professor at UCL. His lab develops mathematical and experimental tools to understand how learning emerges in brains and artificial neural networks. (Confirmed for Paris, remote talk for other locations due to extenuating circumstances)

Angelica Chen <angelica.chen5@gmail.com> is a senior research scientist working on Gemini training at Google DeepMind. During her PhD, she studied natural language understanding and generalization. (Confirmed)

Gautam Reddy <greddy@princeton.edu> is an assistant professor at Princeton, whose lab develops physics-inspired theory and tools for understanding learning and decision-making in biological and artificial systems. (Confirmed)

Shikai Qiu <sq2129@nyu.edu> is a PhD student at New York University whose research focuses on the science of scaling neural networks and optimization. (Confirmed)

PANELISTS

Martin Ziqiao Ma <marstin@umich.edu> is a Member of Technical Staff at Thinking Machines Lab. His research focuses on continual multimodal learning with minimal and natural supervision. (Confirmed)

Misha Belkin <mbelkin@ucsd.edu> is a professor at UCSD whose research develops theory for modern machine learning and deep learning, including high-dimensional learning, interpolation, kernel methods, and double descent. (Tentatively yes, deciding on NeurIPS attendance)

Naomi Saphra <nsaphra@nsaphra.net> is a research fellow at the Kempner Institute at Harvard and incoming faculty at BU whose research focuses on empirical NLP training dynamics and interpretability. (Confirmed)

3 Tentative schedule and logistics

Tentative important dates for submissions, reviews, and final materials: We invite short (5-page) or long (9-page) paper submissions. Submissions will be managed via OpenReview.

- Paper submission deadline: August 29, 2026
- Reviewer bidding and assignment period: August 30 – September 2, 2026
- Reviewer deadline: September 18, 2026
- Acceptance/rejection notification: September 29, 2026
- Camera-ready deadline: October 20, 2026
- Final workshop program and accepted papers due: October 27, 2026

Location preference in order: Sydney, Atlanta, Paris.

Concurrent submissions. There are no concurrent workshop proposals submitted to other venues.

Tentative schedule. We propose a full-day workshop consisting of invited talks, contributed oral presentations selected from accepted submissions, two poster sessions, and a combined panel and interactive discussion designed to foster community engagement.

Time	Event	Time	Event
9:00–9:10	Opening Remarks	13:30–14:10	Invited Talk 4
9:10–9:50	Invited Talk 1	14:10–14:50	Contributed Orals
9:50–10:30	Invited Talk 2	14:50–15:30	Poster Session 2
10:30–11:15	Poster Session 1	15:30–16:00	Panel Discussion
11:15–11:55	Invited Talk 3	16:00–17:00	Interactive "Debugging Training" Session
11:55–12:45	Contributed Orals	17:00–17:10	Closing Remarks
12:45–13:30	Lunch		

The closing hour combines a moderated panel discussion on open challenges in the field with an interactive "debugging training runs" segment, where participants bring real training bugs—instabilities, divergences, scaling pathologies—for the panelists and audience to diagnose together. We will collect training bug examples through a short survey shortly before the workshop, then split into small-group discussions led by organizers and panelists. In the final 20 minutes, each group will share key takeaways and highlights from their discussion. Following the workshop, we will host an informal social for organizers, speakers, and participants.

3.1 Participation and Accessibility Plan

At NeurIPS 2026, we anticipate receiving approximately 100 paper submissions, with an expected acceptance rate of around 35–45%. Of the accepted papers, approximately 20–30% will be invited for contributed oral presentations, with the remainder presented as posters. As this is a first-time workshop, prior submission, acceptance, and attendance statistics are not available.

Attendees. We expect 100–200 in-person attendees. Our goal is to attract a diverse range of participants, including: 1) graduate students and early-career researchers, through accessible workshop materials, mentorship-oriented programming, and opportunities for contributed presentations; 2) researchers from underrepresented and interdisciplinary communities, through broad outreach, inclusive call-for-papers language, and targeted dissemination through relevant Slack and Discord communities, EleutherAI channels, social media platforms such as X/Twitter and Bluesky; and 3) senior researchers and practitioners, through invited talks, panel discussions, and opportunities for collaboration.

Virtual Access and Outreach. To broaden participation, we will maintain an online presence through the workshop website, social channels, academic mailing lists, and relevant research communities. Where permitted, accepted papers, posters, and presentation materials will be posted online, along with summaries and key discussions, to support participants unable to attend physically and engage the broader community.

Diversity of Organizers and Speakers. Our organizing team reflects this commitment, spanning diverse career stages, disciplines, institution types, geographies, and lived experiences, including nontraditional academic pathways and underrepresented gender identities. Our proposed speakers and panelists similarly represent a range of seniority levels, research communities, and institutional contexts. As we finalize invitations, we will continue prioritizing diversity across lived experiences and career.

Special Requirements & Technical Needs. No special needs beyond the standard NeurIPS workshop setup. We request poster space for two poster sessions and standard AV support, including projector/display, microphones, reliable internet, live streaming, and recording where supported by NeurIPS.

4 Organizers and Program Committee

Below are the organizers bio and a list of confirmed program committee. To the best of our knowledge, none of the organizers are listed on any other NeurIPS 2026 workshop proposals. Organizers will not assess submissions from their own institutions or from individuals with whom they hold personal conflicts of interest, and any such submissions will be handled solely by non-conflicted organizers and reviewers.

Isabelle Lee <lee.isabelle.g@gmail.com> is a PhD student at the University of Southern California. Her work focuses on interpretability, training, and reasoning. Specifically, her research aims to (1) predict training and (2) predict failures. To this aim, she uses dynamical systems and physics-inspired approaches to analyze small-scale toy learning problems, as well as analyzing larger-scale training from developmental perspectives. She is a recipient of the Viterbi School of Engineering Graduate Fellowship and Coefficient Giving's Technical AI Safety Research Grant. Previously, she was a researcher at Carnegie Mellon University and Amazon Alexa, and her background is in Plasma Physics and Complex Systems.

Emmy Liu <emmy@cmu.edu> is a PhD Candidate at Carnegie Mellon University in the Language Technologies Institute. Her current work is on understanding pretraining from a cognitive perspective. She is a recipient of a Natural Sciences and Engineering Research Council of Canada Doctoral Scholarship, the SoftBank-ARM fellowship, an NEC fellowship as well as a CMU Presidential Fellowship. Her work has received the Best Resource Paper award at ACL 2023. She has co-organized the workshop on Figurative Language Processing at NAACL 2024, the overall Student Research Workshop at NAACL 2025, as well as the Computational Developmental Linguistics workshop at ACL 2026.

Kaiser Sun <hsun74@jhu.edu> is a Ph.D. student in Computer Science at the Data Science and AI Institute and the Center for Language and Speech Processing at Johns Hopkins University. Their research focuses on the training dynamics, evaluation, and interpretability of Large Language Models. Generally, Kaiser is interested in understanding how language models function and exploring how we can change them. Their work has been recognized with an honorable mention at CoNLL2023, an outstanding paper award at MLRC2022, and coverage by a few media outlets such as MIT Technology Review and WIRED. Previously, they obtained their B.S. and M.S. from the University of Washington and were a researcher at Meta AI, Together AI, and AWS AI Labs.

Stella Biderman <stella@eleuther.ai> is the Executive Director of EleutherAI, a non-profit research institute and open science community that does research on large language models, works to make research on large-scale AI technologies more widely accessible around the world, and works to grow and promote the open source AI community. She pioneered the development of open and transparent language models that make the type of research advocated for in this workshop possible, including the first state-of-the-open LLMs to be released alongside their training data [6, 7, 13] and the Pythia model suite [4, 30] which introduced the idea that releasing partially trained model checkpoints was scientifically valuable and set the standard for maximally open releases adopted by projects such as LLM360, OLMo and Marin. Her lab has done influential research on learning dynamics and interpretability, including introducing Sparse Autoencoders as a tool for interpretability [10], introducing the problem of forecasting which specific sequences are memorized by a large language model during training [3], performing the first study of the stability of circuit analysis over the course of training [29], and demonstrating that it is possible to use data filtering to eliminate undesirable capabilities in large language models [21]. Her recent position paper at ICML outlines the need for the type of work solicited in this workshop [5].

Stella co-leads the EvalEval Coalition, a multi-stakeholder initiative bringing methodological rigor and accountability to AI evaluation, is on the advisory board to the Machine Learning Reproducibility Challenge, is organizing an Open-Weight Pre-training Safety Accelerator as part of the Seoul Alignment Workshop (co-located with ICML), and is an AC for NeurIPS. Previously, she was an organizer of the AI Village @ DEF CON (2018-2021), working group lead for the BigScience Research Workshop (2022), and organized Mozilla and EleutherAI's Dataset Convening. Stella received her M.S. in Computer Science from Georgia Tech, her S.B. (honors) in Mathematics from the University of Chicago, and her A.B. in Philosophy from the University of Chicago. Her work regularly appears at NeurIPS, ICML, ICLR, and ACL.

David Alvarez-Melis <dam@seas.harvard.edu> is an Assistant Professor of Computer Science at Harvard University, where he leads the Data-Centric Machine Learning (DCML) Lab, with appointments as Associate Faculty at the Kempner Institute and Research Scientist at Microsoft Research. His group develops principled methods for characterizing, curating, and optimizing training data, work that bears directly on the workshop's prescriptive axis—in particular, treating data selection, ordering, and composition as levers for shaping training trajectories and downstream behavior. He has organized NeurIPS-scale workshops at the intersection of machine learning and applied mathematics, most directly comparable being the Optimal Transport and Machine Learning Workshop at NeurIPS 2023, along with the Optimal Transport Workshop at the Institut d'Études Scientifiques de Cargèse (2024) and the LatinX in AI Workshop at ICML (2023, 2024). David received his PhD in Computer Science from MIT, his MS in Mathematics from NYU's Courant Institute, and his BSc in Mathematics from ITAM in Mexico City. His work appears regularly at NeurIPS, ICML, ICLR, AISTATS, and UAI.

4.1 Program Committee

We are continuing to recruit program committee members with expertise across the workshop themes and will aim for diversity across institutions, geographies, career stages, and research areas. We anticipate approximately 100 submissions. Each paper will receive 3 reviews, with no reviewer assigned more than 3 papers, requiring a program committee of approximately 100 members. **To manage conflicts of interest**, submissions will be assigned so that no reviewer or organizer evaluates a paper from their own institution or from anyone with whom they have a personal conflict (e.g., advisor–advisee, close collaborators, or family), and organizers with a conflict on a given submission will be recused from its decision. Our current program committee (22 confirmed) includes:

Sarah Liaw (Harvard), Millicent Li (Northeastern), Leonardo Blas Urrutia (USC), Terry Zhou (Harvard), Indranil Halder (Harvard), Jennifer Lu (Brown/MIT), Ruochen Zhang (Brown/Microsoft), Rachit Bansal (Harvard), Logan Riggs-Smith (AISST), Alice Rigg (NDIF), Jeffrey Andrade (Harvard), Jay Huang (JHU), Bernal Jimenez Gutierrez (JHU), Niyati Bafna (JHU), Bingbin Liu (Harvard), Marc Marone (JHU), Elisabeth Fittschen (JHU), Leo Du (JHU), Glenn Matlin (Georgia Tech), Xiaochuan Li (CMU), Atharva Naik (CMU), Dabashish Chakraborty (JHU)

We also plan to host an after-social for organizers, program committee, participants and speakers, organized by our social chair, Glenn Matlin (Georgia Tech).

References

[1] Yasaman Bahri, Ethan Dyer, Jared Kaplan, Jaehoon Lee, and Utkarsh Sharma. Explaining neural scaling laws. *Proceedings of the National Academy of Sciences*, 121(27), June 2024. ISSN 1091-6490. doi: 10.1073/pnas.2311878121. URL http://dx.doi.org/10.1073/pnas.2311878121.

[2] Louis Bethune, David Grangier, Dan Busbridge, Eleonora Gualdoni, Marco Cuturi, and Pierre Ablin. Scaling laws for forgetting during finetuning with pretraining data injection, 2025. URL https://arxiv.org/abs/2502.06042.

[3] Stella Biderman, Usvsn Prashanth, Lintang Sutawika, Hailey Schoelkopf, Quentin Anthony, Shivanshu Purohit, and Edward Raff. Emergent and predictable memorization in large language models. *Advances in Neural Information Processing Systems*, 36:28072–28090, 2023.

[4] Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. Pythia: A suite for analyzing large language models across training and scaling, 2023. URL https://arxiv.org/abs/2304.01373.

[5] Stella Biderman, Mohammad Aflah Khan, Niloofar Mireshghallah, Catherine Arnett, Fazl Barez, and Naomi Saphra. Position: Don't just "fix it in post": A science of ai must study training dynamics. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, volume 205 of *Proceedings of Machine Learning Research*, 2026.

[6] Sid Black, Gao Leo, Phil Wang, Connor Leahy, and Stella Biderman. Gpt-neo: Large scale autoregressive language modeling with mesh-tensorflow. *Zenodo*, 2021.

[7] Sidney Black, Stella Biderman, Eric Hallahan, Quentin Anthony, Leo Gao, Laurence Golding, Horace He, Connor Leahy, Kyle McDonell, Jason Phang, et al. Gpt-neox-20b: An open-source autoregressive language model. In *Proceedings of BigScience Episode# 5–Workshop on Challenges & Perspectives in Creating Large Language Models*, pages 95–136, 2022.

[8] Ethan Caballero, Kshitij Gupta, Irina Rish, and David Krueger. Broken neural scaling laws, 2023. URL https://arxiv.org/abs/2210.14891.

[9] Angelica Chen, Ravid Shwartz-Ziv, Kyunghyun Cho, Matthew L. Leavitt, and Naomi Saphra. Sudden drops in the loss: Syntax acquisition, phase transitions, and simplicity bias in mlms, 2025. URL https://arxiv.org/abs/2309.07311.

[10] Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. *arXiv preprint arXiv:2309.08600*, 2023.

[11] Shibhansh Dohare, J Fernando Hernandez-Garcia, Qingfeng Lan, Parash Rahman, A Rupam Mahmood, and Richard S Sutton. Loss of plasticity in deep continual learning. *Nature*, 632 (8026):768–774, 2024.

[12] Thomas Elsken, Jan Hendrik Metzen, and Frank Hutter. Neural architecture search: a survey. *J. Mach. Learn. Res.*, 20(1):1997–2017, January 2019. ISSN 1532-4435.

[13] Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2020.

[14] HiLD Workshop Organizers. Workshop on high-dimensional learning dynamics (hild). https://sites.google.com/view/hidimlearning/home, 2023. ICML Workshop.

[15] HiLD Workshop Organizers. Workshop on high-dimensional learning dynamics (hild). https://sites.google.com/view/hidimlearning/home, 2024. ICML Workshop.

[16] HiLD Workshop Organizers. Workshop on high-dimensional learning dynamics (hild). https://sites.google.com/view/hidimlearning/home, 2025. ICML Workshop.

[17] HiLD Workshop Organizers. Workshop on high-dimensional learning dynamics (hild). https://sites.google.com/view/hidimlearning/home, 2026. ICML Workshop.

[18] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. Training compute-optimal large language models, 2022. URL https://arxiv.org/abs/2203.15556.

[19] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models, 2020. URL https://arxiv.org/abs/2001.08361.

[20] Joseph Mellor, Jack Turner, Amos Storkey, and Elliot J. Crowley. Neural architecture search without training, 2021. URL https://arxiv.org/abs/2006.04647.

[21] Kyle O'Brien, Stephen Casper, Quentin Anthony, Tomek Korbak, Robert Kirk, Xander Davies, Ishan Mishra, Geoffrey Irving, Yarin Gal, and Stella Biderman. Deep ignorance: Filtering pretraining data builds tamper-resistant safeguards into open-weight llms, 2025. URL https://arxiv.org/abs/2508.06601.

[22] Hadas Orgad, Fazl Barez, Tal Haklay, Isabelle Lee, Marius Mosbach, Anja Reusch, Naomi Saphra, Byron Wallace, Sarah Wiegreffe, Eric Wong, et al. Interpretability can be actionable. *arXiv preprint arXiv:2605.11161*, 2026.

[23] Tim Pearce and Jinyeop Song. Reconciling kaplan and chinchilla scaling laws, 2024. URL https://arxiv.org/abs/2406.12907.

[24] Yi Ren and Danica J. Sutherland. Learning dynamics of llm finetuning, 2025. URL https://arxiv.org/abs/2407.10490.

[25] Thibault Sellam, Steve Yadlowsky, Jason Wei, Naomi Saphra, Alexander D'Amour, Tal Linzen, Jasmijn Bastings, Iulia Turc, Jacob Eisenstein, Dipanjan Das, Ian Tenney, and Ellie Pavlick. The multiberts: Bert reproductions for robustness analysis, 2022. URL https://arxiv.org/abs/2106.16163.

[26] Jacob Mitchell Springer, Sachin Goyal, Kaiyue Wen, Tanishq Kumar, Xiang Yue, Sadhika Malladi, Graham Neubig, and Aditi Raghunathan. Overtrained language models are harder to fine-tune, 2025. URL https://arxiv.org/abs/2503.19206.

[27] Kaiser Sun and Mark Dredze. Amuro & char: Analyzing the relationship between pre-training and fine-tuning of large language models. In Vaibhav Adlakha, Alexandra Chronopoulou, Xiang Lorraine Li, Bodhisattwa Prasad Majumder, Freda Shi, and Giorgos Vernikos, editors, *Proceedings of the 10th Workshop on Representation Learning for NLP (RepL4NLP-2025)*, pages 131–151, Albuquerque, NM, May 2025. Association for Computational Linguistics. ISBN 979-8-89176-245-9. doi: 10.18653/v1/2025.repl4nlp-1.11. URL https://aclanthology.org/2025.repl4nlp-1.11/.

[28] Cameron Tice, Puria Radmard, Samuel Ratnam, Andy Kim, David Africa, and Kyle O'Brien. Alignment pretraining: Ai discourse causes self-fulfilling (mis)alignment, 2026. URL https://arxiv.org/abs/2601.10160.

[29] Curt Tigges, Michael Hanna, Qinan Yu, and Stella Biderman. Llm circuit analyses are consistent across training and scale. *Advances in Neural Information Processing Systems*, 37:40699–40731, 2024.

[30] Oskar van der Wal, Pietro Lesci, Max Muller-Eberstein, Naomi Saphra, Hailey Schoelkopf, Willem Zuidema, and Stella Biderman. Polypythias: Stability and outliers across fifty language model pre-training runs, 2025. URL https://arxiv.org/abs/2503.09543.

[31] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. Emergent abilities of large language models, 2022. URL https://arxiv.org/abs/2206.07682.

[32] Colin White, Arber Zela, Binxin Ru, Yang Liu, and Frank Hutter. How powerful are performance predictors in neural architecture search?, 2021. URL https://arxiv.org/abs/2104.01177.