Preface of UniReps: the Second Workshop on Unifying Representations in Neural Models

Clémentine C. J. Dominé
Gatsby Computational Neuroscience Unit
University College London
London, UK
clementine.domine.20@ucl.ac.uk

Marco Fumero
Dept. of Computer Science
IST Austria
Klosterneuburg, Austria
marco.fumero@ist.ac.at

Donato Crisostomi
Dept. of Computer Science
Sapienza University of Rome
Rome, IT
crisostomi@di.uniroma1.it

Kim Stachenfeld
Google DeepMind
New York, USA
stachenfeld@google.com

Luca Moschella
Apple
Zurich, Swizerland
luca.moschella94@gmail.com

Zorah Lähner
Dept. of Computer Science
University of Bonn and Lamarr Institute
Bonn, Germany
laehner@uni-bonn.de

Editors: Marco Fumero, Clementine Domine, Donato Crisostomi, Zorah Lähner, Luca Moshella, Kim Stachenfeld

Abstract

Discover why, when and how distinct learning processes yield similar representations, and the degree to which these can be unified.

https://unireps.org

Introduction

We are pleased to introduce this preface to the proceedings of the 2nd edition of the UniReps workshop on unified representations, held at NeurIPS (you can find here the workshop recording). For the proceedings of the first edition, please refer to [10]. In the following sections, we summarize the workshop's organization, present key statistics, and discuss its future direction.

Workshop Summary

Neural models tend to learn similar representations when subject to similar stimuli; this behavior has been observed both in biological [11, 20] and artificial settings [23, 19, 24]. The word *similar* here plays a fundamental role: under different conditions and assumptions on the observed data and the neural model (for instance, two distinct individuals exposed to the same stimulus [28] or different initializations of the same neural architecture [37]), inner representations of distinct models can be related to each other, e.g. up to a linear transformation [30]. The similarities in the observational space can refer to settings where data are acquired in a multimodal environment, for instance textual and image representations of the same entity [27], or in a multiview setting [35] where observations in a single modality are acquired under different conditions.

The emergence of similar representations is a ubiquitous but mysterious phenomenon, which is igniting a growing interest in the fields of Neuroscience, Artificial Intelligence and Cognitive Science. It is central to the question of understanding how mechanisms relate between different models, and between models and brains. This workshop will facilitate a rich exchange of ideas and insights among experts in these fields, with the goal of addressing the following points:

- (*When*): To explore the specific conditions under which these similarities emerge in different neural models. Modelling the transformations, symmetries and invariances between similar representations is key to measure if these can be unified [19, 18, 21]. In the previous edition, an outstanding paper introduced Soft Matching Distance—a metric for neural representations that captures single-neuron tuning, further advancing this field [16].
- (*Why*): To investigate the underlying causes of these similarities in neural representations, with a focus on both artificial and biological models, as well as across them. Promising directions include analyzing the learning dynamics of neural models [1, 3, 32], studying model identifiability in the functional and parameter space [33, 14, 30, 15] and investigating the relations between different local minima reached by the optimization process [9, 7, 22]. Finally, an important question to consider is: What mechanisms does knowledge distillation distill? as explored in [39] published in the first edition of the workshop.
- (*What for*): To explore and showcase applications in modular deep learning ranging from model merging [2], reuse [6, 17] and stitching [4, 26] to efficient strategies for fine-tuning and knowledge transfer between models [38] even in out-of-distribution settings [29], or to exploit cross-domain representation similarities (e.g. comparisons between images and neural data [34]). As well as application to Multimodal decoding of human brain activity into images and text also published in the previous edition [8].

The workshop provides an exciting, timely, and diverse environment for discussing theoretical findings, empirical evidence, and practical applications of the emergence of similar representations across models, benefiting from the cross-pollination of different fields (ML, Neuroscience, Cognitive Science) to foster the exchange of ideas and encourage collaborations. The suggested *topics* include:

- Model merging, stitching and reuse [2]
- Identifiability in neural models [30]
- Learning dynamics [32]
- Convergent learning [31, 13]
- Similarity based learning [36, 40]
- Representational alignment [25]
- Symmetry and equivariance in NNs [12]
- Synergy of biological & artificial NNs[5]
- Multiview representation learning [35]
- Linear mode connectivity [9]
- Multimodal learning [27]

Workshop Format We designed a dynamic workshop program integrating *invited talks* (i) with a *panel discussion* (ii), a *mentorship & brainstorming program*, (iii) a *poster session* (iv) and a *social event* (v). In the *panel discussion*, we gathered renowned experts from the fields of AI, Neuroscience, and Cognitive Sciences to engage in a dynamic round table discussion on key topics explored in the workshop. We aimed to establish a cohesive understanding of the emergence of similar representations in neural models and pave the way for a new interdisciplinary community and research area, fixing the relevant research questions to be addressed. The *mentoring & brainstorming program* took place during the sponsored coffee breaks and lunch, along with casual discussions. This time served as an opportunity to conduct research discussions, engage in informal conversations among peers, and offer a mentoring initiative for junior and senior researchers. The *poster session* provided the chance to showcase recent work, share findings, and engage in meaningful discussions among peers. Finally, a *social event* took place at the end of the workshop in collaboration with other workshops with comment topics such as NeurReps and Neuro AI, fostering informal connections among participants and favoring the establishment of long term relationships and collaborations within the workshop community.

Schedule				
08.15 AM	Opening Remarks		12.45 AM	Lunch (Mentorship)
08.30 AM	Invited Talk: E.Grant		2.00 PM	Invited Talk: M.Cuturi
09.00 AM	Invited Talk: S.Chung		2.30 PM	Invited Talk: N. Nanda
09.30 AM	Invited Talk: P.Isola		3.00 PM	Invited Talk: S. Jegelka
10.00 AM	Coffee Break (Mentorship)		3:30 PM	Closing Remarks
10.30 AM	Contributed talks		3.45 PM	Poster Session
11.45 AM	Panel Discussion		5:00 PM	Social Event

Multiple submission tracks Submissions to the workshop were organized in three technical tracks, both requiring novel and unpublished results: we received 49 submissions for the *extended abstract* track, which addresses early-stage results, insightful negative findings, position papers, and 27 submissions for the *proceedings* track, which address complete papers to be published in a dedicated workshop proceedings volume. Both tracks were be included in the workshop poster session to allow authors to present their work. A subset of the submissions have been selected for a contributed talks session during the workshop and award, which are gathered below.

Best Paper Awards Proceedings Track

Authors: Sarah Harvey, David Lipshutz, and Alex H. Williams
Title: "What Representational Similarity Measures Imply about Decodable Information."

Best Paper Awards Extended Abstracts Track

Authors: Richard Antonello and Emily Shana Cheng
Title: "Evidence from fMRI Supports a Two-Phase Abstraction Process in Language Models."

Honorable Mentions Proceedings Track

Author: Alex H. Williams
Title: "Equivalence between Representational Similarity Analysis, Centered Kernel Alignment, and Canonical Correlations Analysis."

Honorable Mentions Extended Abstracts Track

Authors: Chenyu Wang, Sharut Gupta, Xinyi Zhang, Sana Tonekaboni, Stefanie Jegelka, Tommi Jaakkola, and Caroline Uhler
Title: "An Information Criterion for Controlled Disentanglement of Multimodal Data."

We introduced the *Conference-to-Workshop track* where we invited a selection of relevant papers from the NeurIPS2024 main track to be presented in our poster session. 18 papers were accepted on a first-come, first-served basis, provided that they aligned with the workshop's topics. Additionally, in this edition we introduced a *blogpost track*, which received one submission, in which participants were encouraged to present comprehensive guides to known methods and results, opinion pieces, challenges, deep dives, or informal presentations of their own technical submissions in a dynamic format, suitable to host interactive visualization content. These are showcased on the workshop website (following ICLR format).

Program Committee & Chairs

We thank our Program Chairs Irene Cannistraci (Sapienza, University of Rome) and Valentino Maiorca (Sapienza, University of Rome). We would also like to thank Riccardo Marin for serving as COI chair, and Fabian M. Mager and Andrea Santilli for their support during the event. We are proud to introduce our esteemed reviewing committee, comprised of 181 dedicated reviewers who have collectively contributed 293 reviews. Their expertise and commitment have been instrumental in ensuring the high quality and rigor of the discussions and findings presented at our workshop.

Best Reviewer Award A special Best Reviewer Award was given to Abhi Kamboj for their exceptional feedback.

- Abhi Kamboj (University of Illinois at Urbana-Champaign)
- Aishwarya Gupta (Indian Institute of Technology, Kanpur)
- Ajay Subramanian (New York University)
- Akshata Kishore Moharir (Microsoft)
- Akshay Malhotra (Interdigital)
- Albert Manuel Orozco Camacho (Concordia University)
- Aldo Glielmo (Banca d'Italia)
- Alejandro Garca-Castellanos (University of Amsterdam)
- Aleksander Piotr Skorupa (University of Edinburgh)
- Alex H Williams (New York University)
- Alexander Huth (The University of Texas at Austin)
- Alice Bizeul (ETH Zurich)
- Alish Dipani (Georgia Institute of Technology)
- Ananya Passi (Johns Hopkins University)
- Andy T. Liu (ASUS)
- Anirudh Govil (International Institute of Information Technology Hyderabad)
- Anugunj Naman (Purdue University)
- Arif Dnmez (IUF Leibniz Research Institute for Environmental Medicine)
- Austin Meek (University of Delaware)
- Avisha Das (University of Houston)
- Ayyce Begm Bekta (Memorial Sloan Kettering Cancer Centre)
- Beatrix Miranda Ginn Nielsen (Technical University of Denmark)
- Berfin Inal (University of Amsterdam)
- Berivan Isik (Google)
- Berker Demirel (Institute of Science and Technology)
- Binxi Xie (Emory University)
- Biswarup Bhattacharya (Citadel)
- Bo Zhao (University of California, San Diego)
- Bogdan-Ionut Cirstea (Tlcom ParisTech)
- Bridget Leonard (University of Washington)
- Bruna Junqueira Lopes (École des Ponts ParisTech)
- Chandrasekhar Karnam (Hudson River Trading)
- Changqing Fu (Universit Paris-Dauphine (Paris IX))
- Charles Camboulin (CY Cergy Paris Université)
- Chen Bo Calvin Zhang (UC Berkeley)
- Chenyu Wang (MIT)
- Chi-Ning Chou (Center of Computational Neuroscience, Flatiron Institute)
- Ching Fang (Columbia University)
- David Lipshutz (Flatiron Institute)
- Davit Soselia (University of Maryland, College Park)
- Denis A Gudovskiy (Panasonic Corp)
- Deval Mehta (Monash University)
- Dimitra Maoutsa (Technische Universität München)
- EHSAN KIAKOJOURI (Sapienza University of Rome)
- Edgar E Robles (University of California, Irvine)
- Elom Amematsro (Columbia University)
- Emilian Postolache (Ca' Foscari University of Venice)
- Emily Cheng (Universitat Pompeu Fabra)
- Erin Grant (University College London)
- Etienne Gay (Sorbonne Universit - Facult des Sciences (Paris VI))
- Fabian Mager (Technical University of Denmark)
- Fahimeh Arab (University of California, Riverside)
- Gabor Lengyel (University of Rochester)
- Gasser Elbanna (Harvard University)
- Hamed Karimi (Boston College)
- Hamidreza Jamalabadi (Philipps-Universität Marburg)
- Hanlin Yu (University of Helsinki)
- Hannah Small (Johns Hopkins University)
- Haoyu Zhang (The Chinese University of Hong Kong, Shenzhen)
- Harshay Shah (MIT)
- Huichi Zhou (Imperial College London)
- Hyewon Willow Han (University of Western Ontario)
- Inwoo Hwang (Seoul National University)
- Irene Tallini (Sapienza University of Rome)
- Itay Evron (Facebook)
- Jacob S. Prince (Harvard University)
- Jan Finkbeiner (Research Center Juelich)
- Jan Philipp Bauer (University College London, University of London)
- Jenelle Feather (Flatiron Institute)
- Jerry Ngo (MIT)
- Jiajing Chen (New York University)
- Jiancheng Pan (Tsinghua University)
- Jianyu Wu (Beijing University of Aeronautics and Astronautics)

- Jie Mei (IT:U Interdisciplinary Transformation University Austria)
- Jingxiao Tian (University of California, San Diego)
- Jinyung Hong (Arizona State University)
- Joo Abrantes (Sakana AI)
- Juan Miguel Navarro Carranza (Stanford University)
- Juanxi Tian (University of Illinois at Urbana-Champaign)
- Jrg Schltterer (Universität Mannheim)
- Kasper Vinken (Fujitsu Research of America)
- Keun Hee Park (Arizona State University)
- Konrad Karanowski (Technical University of Wroclaw)
- Konstantin Hemker (University of Cambridge)
- Krishna Sri Ipsit Mantri (Purdue University)
- Kusumakumari Vanteru (Westcliff University)
- Kyle Daruwalla (Cold Spring Harbor Laboratory)
- Leyla Isik (Johns Hopkins University)
- Lorenzo Basile (University of Trieste)
- Lorenzo Giusti (CERN)
- Luca Zhou (Sapienza University of Rome)
- Luigi Capogrosso (University of Verona)
- Luigi Gresele (University of Copenhagen)
- Luke McDermott (University of California, San Diego)
- Maciej Zieba (Wroclaw University of Science and Technology)
- Marco Pegoraro (Sapienza University of Rome)
- Martin Ester (Simon Fraser University)
- Maryam Hoseini Behbahani (Sharif University of Technology)
- Maryam Mirian (University of British Columbia)
- Maryam Mirian (University of British Columbia)
- Matteo Alleman (Columbia University)
- Matteo Ferrante (Università di Roma Tor Vergata)
- Maxime Di Folco (Helmholtz Zentrum München)
- Mohamed Shawky Sabae (Faculty of Engineering Cairo University, Cairo University)
- Mohammadamin Tavakoli (Caltech)
- Mohammed Adnan (University of Calgary / Vector Institute)
- Morteza Mahdiani (Université de Montréal)
- Moshe Eliasof (University of Cambridge)
- Mycal Tucker (MIT)
- Nanda H Krishna (Université de Montréal)
- Nicola Toschi (Università di Roma Tor Vergata)
- Nicolas Zilberstein (Rice University)
- Niharika S. D'Souza (International Business Machines)
- Nikolas McNeal (Georgia Institute of Technology)
- Omkar Joglekar (Bosch)
- Oriol Caudevilla (eBay Inc.)
- Osamu Hirose (Kanazawa University)
- Paris Giampouras (University of Warwick)
- Phuong Quynh Le (Philipps-Universität Marburg)
- Qiang Li (GSU)
- Qingqing Yang (Ohio State University, Columbus)
- Raghav Singhal (Mohamed bin Zayed University of Artificial Intelligence)
- Raja Kumar (CNRS)
- Riccardo Cadei (Institute of Science and Technology)
- Riccardo Renzulli (University of Turin)
- Richard Antonello (University of Texas, Austin)
- Rohan Jain (University of Calgary)
- Ruchira Dhar (University of Copenhagen)
- Rylan Schaeffer (Computer Science Department, Stanford University)
- Saaketh Medepalli (CMU)
- Sarah E Harvey (Flatiron Institute)
- Sayed Soroush Daftarian (Universitätsklinikum Marburg)
- Shashwat Singh (International Institute of Information Technology Hyderabad)
- Shiwen Zhang (Alibaba Group)
- Shreya Kapoor (Friedrich-Alexander-Universität Erlangen-Nürnberg)
- Shreya Kapoor (Friedrich-Alexander-Universität Erlangen-Nürnberg)
- Shubham Shukla (Nordstrom)
- Shuman Peng (Simon Fraser University)
- Siddhartha Gairola (Saarland Informatics Campus, Max-Planck Institute)
- Sining Huang (UC Berkeley)
- Stefan Horoi (Université de Montréal)
- Sujin Jeon (Seoul National University)
- Sukanya Moorthy (Credit Karma)
- Sukanya Moorthy (Credit Karma)

- Suklav Ghosh (Indian Institute of Technology, Guwahati)
- Tahmineh A. Koosha (Philipps-Universität Marburg)
- Tanmoy Mukherjee (Vrije Universiteit Brussel)
- Teresa Dorszewski (Technical University of Denmark)
- Thanh-Tung Nguyen (asus)
- Thibault Malherbe (Inetum)
- Thomas Edward Yerxa (New York University)
- Thu Bui (Purdue University)
- Tom White (Victoria University of Wellington)
- Udita Patel (Amazon)
- Valeria Ruscio (Sapienza University of Rome)
- Vatsala Nema (Indian Institute of Technology, Hyderabad, Dhirubhai Ambani Institute of Information and Communication Technology)
- Vighnesh Subramaniam (MIT)
- Vijay Prakash Dwivedi (Computer Science Department, Stanford University)
- Viktor Schlegel (Imperial College London)
- Wangjiaxuan Xin (University of North Carolina at Charlotte)
- Weisi Liu (University of Memphis)
- Weizhi Zhang (University of Illinois Chicago)
- William Yang (Flatiron Institute)
- Xiequn Wang (South University of Science and Technology of China)
- Xingyu Zheng (Cold Spring Harbor Laboratory)
- Xinyi Yang (Salesforce Research)
- Yalda Mohsenzadeh (University of Western Ontario)
- Yani Ioannou (University of Calgary)
- Ye Zhang (University of Pittsburgh)
- Yicheng Fu (Stanford University)
- Yixiao Kang (Facebook)
- Yixiao Yuan (Columbia University)
- Yulong Zhang (Zhejiang University)
- Yuxin Qiao (Northern Arizona University)
- Zaigham Abbas Randhawa (West Virginia University)
- Zhan Zhuang (City University of Hong Kong)
- Ziao Zhang (School of Engineering and Applied Sciences, Harvard University)
- Ziqing Yang (CISPA Helmholtz Center for Information Security)
- Ziyi Zhu (Deloitte Consulting)
- Arvind Saraf (Microsoft)
- Geraldin Nanfack (Concordia University)

Attendance We surpassed our expectations and last year's edition count by drawing in a diverse crowd of 1,200 attendees in person throughout the day. The audience was a rich tapestry of students, researchers, and industry practitioners from a variety of communities and cultures. The welcoming nature of our event was further enhanced by the thoughtful room setup and environment we created, which fostered a sense of inclusion and engagement among all attendees.

Feedback We ran a feedback survey for the workshop, and below is a summary of the responses. In general, the event received highly positive feedback, with attendees praising the quality of the talks (rated 4.31/5), engaging poster sessions (rated 4/5), and valuable discussions both in person and online. Many appreciated the organization and depth of research presented, with some highlighting the usefulness of platforms like Discord for remote participation (rated to 4.8/5). However, there were areas for improvement, including better spacing and navigation for poster sessions, more structured mentorship opportunities, and increased interactivity in sessions given as feedback comments. Some also suggested shortening or revising panels, improving room setups, and incorporating broader topic representation. Despite these points, many participants expressed gratitude to the organizers for a well-executed and insightful event, rating the workshop 4.54/5).

Diversity and inclusivity Our workshop upheld diversity and inclusivity as fundamental principles for fostering a balanced and productive environment. To achieve this, we strived for diversity in various aspects, including seniority, gender balance, and nationality. Our organizers and invited speakers ranged from PhD students to junior and senior researchers, reflecting a broad spectrum of experience levels. We made a conscious effort to ensure gender balance among both our organizers and keynote speakers, and included participants from different regions, covering Europe, the United States, and Middle Eastern Asia. To promote an inclusive environment, we actively sought participation from the BlackInAI, Women In Machine Learning (WiML), QueerInAI, and LatinxInAI communities by sending Program Committee calls and invitations to attend the workshop through their mailing lists and communication channels. In this regard, with the generous contribution in funding from the G-research for UniReps and Google Deepmind, we were able to establish a travel and registration assistance program for attendees. This program was designed to provide financial aid to researchers, students, or individuals who encountered financial obstacles when trying to attend NeurIPS and UniReps. Thanks to this financial support, we directly offset expenses such as the registration fee, which typically amounts to around $500, making it more feasible for a wider range of participants to attend and contribute to our workshop.

Sponsors

We extend our deepest gratitude to our sponsors and G-research for their generous support and commitment to advancing research and innovation. Their contributions have been invaluable in making our event a success, enabling us to create a platform for sharing knowledge, fostering collaborations, and promoting the latest advancements in the field. We are truly thankful for their support and look forward to continuing our partnership in the future.

Future directions

We find it both crucial and timely to establish a research forum and a supportive community that encourages knowledge exchange at the intersection of machine learning and neuroscience, with a particular emphasis on unified representations. In this spirit, our blog post track remains open for submissions throughout the year, and we actively welcome innovative contributions. Additionally, we are excited to announce the launch of the ELLIS UniReps Speaker Series—stay tuned for updates and announcements. As we move forward, we remain dedicated to facilitating meaningful discussions on these topics at NeurIPS and other key events.

Community

To strengthen our sense of community, we have also established an active network of students and researchers. This network is a central hub for coordinating activities such as seminars and hackathons, further enriching the UniReps workshop experience. Join us to stay up-to-date with the latest workshop news, connect with a vibrant community, display your latest projects, and remain informed about exciting opportunities, events, and research. Our aim is to foster an engaging and inclusive environment, allowing each participant to contribute, learn, and maintain lasting connections beyond the workshop. Check out the UniReps Website! In addition, you can follow the last updates on the UniReps community on our Twitter profile and Discord!

References

[1] Madhu S Advani, Andrew M Saxe, and Haim Sompolinsky. High-dimensional dynamics of generalization error in neural networks. *Neural Networks*, 132:428–446, 2020.

[2] Samuel K. Ainsworth, Jonathan Hayase, and Siddhartha Srinivasa. Git re-basin: Merging models modulo permutation symmetries, 2023.

[3] Shun-ichi Amari, Tomoko Ozeki, Ryo Karakida, Yuki Yoshida, and Masato Okada. Dynamics of learning in mlp: Natural gradient and singularity revisited. *Neural computation*, 30(1):1–33, 2017.

[4] Yamini Bansal, Preetum Nakkiran, and Boaz Barak. Revisiting model stitching to compare neural representations, 2021.

[5] David G. T. Barrett, Ari S. Morcos, and Jakob H. Macke. Analyzing biological and artificial neural networks: challenges with opportunities for synergy? *Current Opinion in Neurobiology*, 55:55–64, 2018.

[6] Róbert Csordás, Sjoerd van Steenkiste, and Jürgen Schmidhuber. Are neural nets modular? inspecting functional modularity through differentiable weight masks, 2021.

[7] Rahim Entezari, Hanie Sedghi, Olga Saukh, and Behnam Neyshabur. The role of permutation invariance in linear mode connectivity of neural networks, 2022.

[8] Matteo Ferrante, Tommaso Boccato, Furkan Ozcelik, Rufin VanRullen, and Nicola Toschi. Multimodal decoding of human brain activity into images and text. In Marco Fumero, Emanuele Rodolá, Clementine Domine, Francesco Locatello, Karolina Dziugaite, and Caron Mathilde, editors, *Proceedings of UniReps: the First Workshop on Unifying Representations in Neural Models*, volume 243 of *Proceedings of Machine Learning Research*, pages 87–101. PMLR, 15 Dec 2024.

[9] Jonathan Frankle, Gintare Karolina Dziugaite, Daniel M. Roy, and Michael Carbin. Linear mode connectivity and the lottery ticket hypothesis, 2020.

[10] Marco Fumero, Emanuele Rodolá, Clementine Domine, Francesco Locatello, Karolina Dziugaite, and Caron Mathilde. Preface of unireps: the first workshop on unifying representations in neural models. In *Proceedings of UniReps: the First Workshop on Unifying Representations in Neural Models*, pages 1–10. PMLR, 2024.

[11] James V Haxby, M Ida Gobbini, Maura L Furey, Alumit Ishai, Jennifer L Schouten, and Pietro Pietrini. Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, 293(5539):2425–2430, 2001.

[12] Irina Higgins, Sébastien Racanière, and Danilo Rezende. Symmetry-based representations for artificial and biological general intelligence, 2022.

[13] Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. The platonic representation hypothesis, 2024.

[14] Aapo Hyvarinen, Hiroaki Sasaki, and Richard Turner. Nonlinear ica using auxiliary variables and generalized contrastive learning. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 859–868. PMLR, 2019.

[15] Ilyes Khemakhem, Diederik Kingma, Ricardo Monti, and Aapo Hyvarinen. Variational autoencoders and nonlinear ica: A unifying framework. In *International Conference on Artificial Intelligence and Statistics*, pages 2207–2217. PMLR, 2020.

[16] Meenakshi Khosla and Alex H Williams. Soft matching distance: A metric on neural representations that captures single-neuron tuning. In Marco Fumero, Emanuele Rodolá, Clementine Domine, Francesco Locatello, Karolina Dziugaite, and Caron Mathilde, editors, *Proceedings of UniReps: the First Workshop on Unifying Representations in Neural Models*, volume 243 of *Proceedings of Machine Learning Research*, pages 326–341. PMLR, 15 Dec 2024.

[17] Louis Kirsch, Julius Kunze, and David Barber. Modular networks: Learning to decompose neural computation, 2018.

[18] Max Klabunde, Tobias Schumacher, Markus Strohmaier, and Florian Lemmerich. Similarity of neural network models: A survey of functional and representational measures, 2023.

[19] Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International Conference on Machine Learning*, pages 3519–3529. PMLR, 2019.

[20] Aarre Laakso and Garrison Cottrell. Content and cluster analysis: assessing representational similarity in neural systems. *Philosophical psychology*, 13(1):47–76, 2000.

[21] Giovanni Luca Marchetti, Christopher Hillar, Danica Kragic, and Sophia Sanborn. Harmonics of learning: Universal fourier features emerge in invariant networks. *ArXiv preprint*, abs/2312.08550, 2023.

[22] Seyed Iman Mirzadeh, Mehrdad Farajtabar, Dilan Gorur, Razvan Pascanu, and Hassan Ghasemzadeh. Linear mode connectivity in multitask and continual learning, 2020.

[23] Ari Morcos, Maithra Raghu, and Samy Bengio. Insights on representational similarity in neural networks with canonical correlation. *Advances in Neural Information Processing Systems*, 31, 2018.

[24] Luca Moschella, Valentino Maiorca, Marco Fumero, Antonio Norelli, Francesco Locatello, and Emanuele Rodolà. Relative representations enable zero-shot latent space communication, 2023.

[25] Lukas Muttenthaler, Jonas Dippel, Lorenz Linhardt, Robert A. Vandermeulen, and Simon Kornblith. Human alignment of neural network representations, 2023.

[26] Antonio Norelli, Marco Fumero, Valentino Maiorca, Luca Moschella, Emanuele Rodolà, and Francesco Locatello. Asif: Coupled data turns unimodal models to multimodal without training, 2023.

[27] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.

[28] Rajeev D. S. Raizada and Andrew C. Connolly. What makes different people's representations alike: Neural similarity space solves the problem of across-subject fmri decoding. *Journal of Cognitive Neuroscience*, 24:868–877, 2012.

[29] Alexandre Ramé, Kartik Ahuja, Jianyu Zhang, Matthieu Cord, Léon Bottou, and David Lopez-Paz. Model ratatouille: Recycling diverse models for out-of-distribution generalization, 2023.

[30] Geoffrey Roeder, Luke Metz, and Durk Kingma. On linear identifiability of learned representations. In *International Conference on Machine Learning*, pages 9030–9039. PMLR, 2021.

[31] Karsten Roth, Lukas Thede, Almut Sophia Koepke, Oriol Vinyals, Olivier Hénaff, and Zeynep Akata. Fantastic gains and where to find them: On the existence and prospect of general knowledge transfer between any pretrained model, 2024.

[32] Andrew M Saxe, James L McClelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. *arXiv preprint arXiv:1312.6120*, 2013.

[33] Peter Sorrenson, Carsten Rother, and Ullrich Köthe. Disentanglement by nonlinear ica with general incompressible-flow networks (gin). *arXiv preprint arXiv:2001.04872*, 2020.

[34] Yu Takagi and Shinji Nishimoto. High-resolution image reconstruction with latent diffusion models from human brain activity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14453–14463, 2023.

[35] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding, 2020.

[36] Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning? *Advances in neural information processing systems*, 33:6827–6839, 2020.

[37] Liwei Wang, Lunjia Hu, Jiayuan Gu, Yue Wu, Zhiqiang Hu, Kun He, and John Hopcroft. Towards understanding learning representations: To what extent do different neural networks learn the same representation, 2018.

[38] Mitchell Wortsman, Gabriel Ilharco, Samir Yitzhak Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S. Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, and Ludwig Schmidt. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time, 2022.

[39] Cindy Wu, Ekdeep Singh Lubana, Bruno Kacper Mlodozeniec, Robert Kirk, and David Krueger. What mechanisms does knowledge distillation distill? In Marco Fumero, Emanuele Rodolá, Clementine Domine, Francesco Locatello, Karolina Dziugaite, and Caron Mathilde, editors, *Proceedings of UniReps: the First Workshop on Unifying Representations in Neural Models*, volume 243 of *Proceedings of Machine Learning Research*, pages 60–75. PMLR, 15 Dec 2024.

[40] Roland S Zimmermann, Yash Sharma, Steffen Schneider, Matthias Bethge, and Wieland Brendel. Contrastive learning inverts the data generating process. In *International Conference on Machine Learning*, pages 12979–12990. PMLR, 2021.