Stefano Sarao Mannelli

Assistant Professor,
Data Science and AI, Computer Science and Engineering,
Chalmers University of Technology & Gothenburg University.
stefano.sarao@gmail.com
stefsmlab.github.io
scholar.google.com/citations?user=Kq272_MAAAAJ

Research Experience

- Sept. 2024 – Present: WASP-AI **Assistant Professor Tenure Track** in the Department of Computer Science and Engineering, Chalmers University of Technology and University of Gothenburg, Sweden. I use statistical physics to study how neural networks learn. I ask which choices about data and training cause a model to make more errors for one group than another, and when changing the training procedure can reduce that gap.
- Feb. 2025 – Present: **Visiting Lecturer** in the School of Computer Science and Applied Mathematics, University of the Witwatersrand, South Africa. I designed and teach Modern Theories of Machine Learning and co-supervise Gabriel Drew's master's research. With Devon Jarvis and colleagues, I study what happens to poorly represented languages when models train on text generated by earlier models.
- Jun. 2021 – Aug. 2024: **Research Fellow → Senior Research Fellow** in Gatsby Computational Neuroscience Unit and Sainsbury Wellcome Centre, University College London, UK. With Andrew Saxe and colleagues, I developed models of how animals learn complex tasks in stages and how neural networks forget earlier tasks as they learn new ones.
- Oct. 2020 – Jun. 2021: **Research Fellow** in the Department of Experimental Psychology, University of Oxford, UK. I began working with Andrew Saxe on the mathematics of curriculum learning: when does seeing easier examples first help a network learn?
- Mar. 2020 - Apr. 2020: **Visiting researcher** with **Eric Vanden-Eijnden** at NYU, USA.
- Jan. 2019 - Mar. 2019: **Visiting researcher** at KITP, UCSB, U.S.A.
- Feb. 2018: **Visiting researcher** at Duke University, U.S.A.
- Mar. 2017 - Jul. 2017: **Research intern** with **Lenka Zdeborová** at IPhT, CEA, France.
- Jan. 2016 - Apr. 2016: **Research intern** with **Nicolas Vayatis** at CMLA, ENS Cachan, France.

Education

- Oct. 2017 – Oct. 2020: **Ph.D. in Theoretical Physics** at IPhT, CEA, France - Supervisor: **Lenka Zdeborová**;
- Sep. 2016 - Oct. 2017: **M.Sc. in Electronic Engineering** at Politecnico di Torino, Italy, grade: *110/110 cum laude*;
- Sep. 2014 - Jul. 2016: **M.Sc. in Physics of Complex Systems** at Politecnico di Torino-SISSA, Italy, grade: *110/110 cum laude*;
- Sep. 2014 - Jul. 2016: **M2, Physique Théorique** at Paris Diderot, UPMC, and ENS Cachan, France;
- Sep. 2014 - Jul. 2016: **Alta Scuola Politecnica Diploma** at Politecnico di Torino, Politecnico di Milano, Italy;
- Sep. 2014 - Jun. 2017: **M.Sc. in Engineering Physics** at Politecnico di Milano, Italy, grade: *110/110 cum laude*;
- Sep. 2011 - Jul. 2014: **B.Sc. in Mathematics for Engineering** at Politecnico di Torino, Italy, grade: *110/110 cum laude*.

Organisation of Scientific Events

- (future event) A.Y. 2026-27: KITP program on *Psychophysics and Statistical Physics of Learning in Higher Level Cognition*, Santa Barbara, USA [link];
- Aug. 2026: School on Analytical Connectionism, Gothenburg, Sweden [link][Secured 500K SEK];
- Aug. 2025: School on Analytical Connectionism, London, UK [link][Secured 25K GBP];
- Aug. 2024: School on *Analytical Connectionism*, New York, USA [link][Secured 152K USD];
- Aug. 2023: School on Analytical Connectionism, London, UK [link][Secured 49K GBP];
- Mar. 2024: Workshop on *Bridging Analytical and Experimental Insights into Representational Change* as part of CoSyNe, Lisbon, Portugal [link];
- Oct. 2023: Workshop *Analytical Approaches for Neural Network Dynamics*,Paris,FRA [link][Secured 10K€];
- Aug.-Sep. 2023: School and workshop on *Analytical Connectionism*, London, UK [link][Secured 42K GBP];
- Jun. 2022: Workshop *Communication Across Communities in Machine Learning Research and Practice* as part of FAccT 2022, Seoul, South Korea [link];
- May 2021: Workshop *Science and Engineering of Deep Learning* as part of ICLR 2021, virtual [link].

Ongoing Projects

- **Bias Generation and Amplification**: I study how group frequency and feature geometry affect learning over time. I track when classification errors diverge between groups and when a generative model begins to memorise examples from each group. I use these results to test whether changing how often each group is sampled can narrow a performance gap. (*AI Alignment Grant, 2026*)
- **Connecting Dynamics and Landscapes in High Dimensions:** I study how much the geometry of a loss landscape can tell us about the solution a learning algorithm will find. I compare the dynamics of gradient descent with the curvature near saddle points. In a model of signal recovery, this explains how the algorithm can recover the signal even when the landscape contains exponentially many poor local minima. (*WASP Starting Pkg 2025-*)
- **Bridging Biological and Artificial Neural Networks:** I study how people, animals and artificial networks learn a sequence of related tasks. I model how networks reuse what they learned on an earlier task to explain when knowledge transfers and when it is forgotten. By giving people and networks the same tasks in different orders, I test which aspects of human learning these models capture. (*WASP Starting Pkg 2025-*)

List of current PhD students and postdocs

- **Flavio Nicoletti**: Postdoc Researcher;
- **Loek van Rossem**: 4th-year Ph.D. student;
- **Chenxiao Ma**: 2nd-year PhD student.

Grants & Awards

- May 2026: Vetenskapsfond, Wilhelm och Martina Lundgrens stiftelser , 50,000 SEK.
- Jan. 2026: AI Alignment Grant, AISI, 969,811.36 GBP.
- 2025: Top Reviewer Award, NeurIPS.
- Jan. 2025: Academic Grant, CM Lerici Foundation, 60,000 SEK.
- Jul. 2024: Guarantor of Brains Travel grant, 1000 GBP.
- Feb. 2024: G-Research Grant, 1400 GBP.
- Nov. 2023: UK-IT Trustworthy AI Exchange Programme, 4500 GBP, The Alan Turing Institute;
- Mar. 2023: SCGB Conference Awards, 3000 GBP, Simons Foundation;
- Sep. 2017: Ph.D. scholarship from CEA;

Positions of Responsibility

- *Founding team member and Principal Investigator* at Principia, helping shape its AI safety research programme;
- *Leader* at CHAIR Mathematical Foundation of AI theme, Chalmers & GU;
- Vice-leader of the WASP Mathematical Foundations of AI cluster;
- Co-founder and co-director of the Analytical Connectionism schools (since 2023);
- *Organiser* Colloquia at CSE department, Chalmers;
- Part of the Technical Program Committee CCN 2026;
- Action Editor for TMLR;
- Area Chair for NeurIPS, ICML, ICLR;
- Reviewer for NeurIPS, ICML, ICLR, SciPost, JSTAT, JPhysA, Physica A, PRX, TMLR.

Publications and preprints

1. F. Kovačević, E. Cyffers, S. Sarao Mannelli and M. Mondelli. High-Dimensional Asymptotics and Dataset Selection for Private Transfer Learning. arXiv:2610.02578 (2026), preprint.

2. A. O'Nils, S. Sarao Mannelli, O. Engkvist, J. Sjölund and A. M. Westerlund. Potency Prediction via Probabilistic Modeling of Dose–Response Curves. ChemRxiv (2026), preprint. doi:10.26434/chemrxiv.15008694/v1.

3. F. Nicoletti, C. Ma, E. Ventura, L. Saglietti and S. Sarao Mannelli. The Interplay of Data Structure and Imbalance in the Learning Dynamics of Diffusion Models. NeurIPS 2026, forthcoming; arXiv:2605.06367.

4. D. Jarvis, R. Klein, B. Rosman, S. James and S. Sarao Mannelli. Position: the Stochastic Parrot in the Coal Mine. Model Collapse is a Threat to Low-Resource Communities. ICML 2026 (Spotlight); arXiv:2605.04127.

5. J. Huang, B. Loureiro and S. Sarao Mannelli. Sharp description of local minima in the loss landscape of high-dimensional two-layer ReLU neural networks. ICML 2026; arXiv:2604.09412.

6. K. J. Sandbrink, S. Sarao Mannelli and F. Krzakala. Thinking of Neural Networks Like a Physicist: The Statistical Physics of Machine Learning. Proceedings of the Analytical Connectionism Schools 2023-2024, PMLR 320:15-41 (2026).

7. Y. Strittmatter, S. Sarao Mannelli, M. Ruiz-Garcia, S. Musslick and M. W. H. Spitzer. Curriculum Learning in Humans and Neural Networks. Open Mind 10:739-753 (2026); earlier version in Proceedings of the Annual Meeting of the Cognitive Science Society 47 (2025).

8. S. Sarao Mannelli, F. Gerace, N. Rostamzadeh and L. Saglietti. Bias-inducing geometries: An exactly solvable data model with fairness implications. Physical Review E 112, 025304 (2025).

9. D. Jarvis, S. Lee, C. C. J. Dominé, A. M. Saxe and S. Sarao Mannelli. A Theory of Initialisation's Impact on Specialisation. ICLR 2025; Journal of Statistical Mechanics: Theory and Experiment 2025, 114001.

10. F. Mori, S. Sarao Mannelli and F. Mignacco. Optimal Protocols for Continual Learning via Statistical Physics and Control Theory. ICLR 2025; Journal of Statistical Mechanics: Theory and Experiment 2025, 084004.

11. N. Patel, S. Lee, S. Sarao Mannelli, S. Goldt and A. Saxe. RL Perceptron: Generalization Dynamics of Policy Learning in High Dimensions. Physical Review X 15, 021051 (2025).

12. F. Gerace, D. Doimo, S. Sarao Mannelli, L. Saglietti and A. Laio. How to choose the right transfer learning protocol? A qualitative analysis in a controlled set-up. Transactions on Machine Learning Research (2024).

13. Y. Li, R. Carrasco-Davis, S. Sarao Mannelli, Y. Strittmatter and S. Musslick. A Meta-Learning Framework for Rationalizing Cognitive Fatigue in Neural Systems. Proceedings of the Annual Meeting of the Cognitive Science Society 46 (2024).

14. A. Jain, R. Nobahari, A. Baratin and S. Sarao Mannelli. Bias in Motion: Theoretical Insights into the Dynamics of Bias in SGD Training. NeurIPS 2024.

15. S. Sarao Mannelli, Y. Ivashinka, A. Saxe and L. Saglietti. Tilting the Odds at the Lottery: the Interplay of Overparameterisation and Curricula in Neural Networks. ICML 2024.

16. J. H. Lee, S. Sarao Mannelli and A. Saxe. Why Do Animals Need Shaping? A Theory of Task Composition and Curriculum Learning. ICML 2024.

17. L. Saglietti, S. Sarao Mannelli and A. Saxe. An Analytical Theory of Curriculum Learning in Teacher-Student Networks. NeurIPS 2022; Journal of Statistical Mechanics: Theory and Experiment 2022, 114014.

18. S. Lee, S. Sarao Mannelli, C. Clopath, S. Goldt and A. Saxe. Maslow's Hammer in Catastrophic Forgetting: Node Re-Use vs. Node Activation. ICML 2022.

19. F. Gerace, L. Saglietti, S. Sarao Mannelli, A. Saxe and L. Zdeborová. Probing transfer learning with a model of synthetic correlated datasets. Machine Learning: Science and Technology 3, 015030 (2022).

20. S. Sarao Mannelli and P. Urbani. Analytical Study of Momentum-Based Acceleration Methods in Paradigmatic High-Dimensional Non-Convex Problems. NeurIPS 2021.

21. A. Baker, I. Biazzo, A. Braunstein, G. Catania, L. Dall'Asta, A. Ingrosso, F. Krzakala, F. Mazza, M. Mézard, A. P. Muntoni, M. Refinetti, S. Sarao Mannelli and L. Zdeborová. Epidemic mitigation by statistical inference from contact tracing data. Proceedings of the National Academy of Sciences (2021).

22. S. Sarao Mannelli, E. Vanden-Eijnden and L. Zdeborová. Optimization and Generalization of Shallow Neural Networks with Quadratic Activation Functions. NeurIPS 2020.

23. S. Sarao Mannelli, G. Biroli, C. Cammarota, F. Krzakala, P. Urbani and L. Zdeborová. Complex Dynamics in Simple Neural Networks: Understanding Gradient Flow in Phase Retrieval. NeurIPS 2020.

24. S. Sarao Mannelli and L. Zdeborová. Thresholds of descending algorithms in inference problems. Journal of Statistical Mechanics: Theory and Experiment 2020, 034004.

25. S. Sarao Mannelli, G. Biroli, C. Cammarota, F. Krzakala, P. Urbani and L. Zdeborová. Marvels and Pitfalls of the Langevin Algorithm in Noisy High-Dimensional Inference. Physical Review X 10, 011057 (2020).

26. S. Sarao Mannelli, G. Biroli, C. Cammarota, F. Krzakala and L. Zdeborová. Who is Afraid of Big Bad Minima? Analysis of gradient-flow in spiked matrix-tensor models. NeurIPS 2019 (Spotlight).

27. S. Sarao Mannelli, F. Krzakala, P. Urbani and L. Zdeborová. Passed & Spurious: Descent Algorithms and Local Minima in Spiked Matrix-Tensor Models. ICML 2019.