Jakob Paul Zimmermann

jakob.paul.zimmermann@gmail.com

Website | LinkedIn | Google Scholar | FU Discrete Geometry

Last updated: June 2026

Research Profile

My research develops mathematically grounded interpretability methods for neural networks, with a focus on attribution, transformer interpretability, circuit discovery, and internal-representation monitoring. I combine tools from polyhedral geometry, optimization, reinforcement learning, and representation learning to study how evidence flows through neural networks.

Research Interests

Mechanistic interpretability; attribution methods; transformer interpretability; circuit discovery; internal-representation monitoring; robustness and safety-critical failure detection.

Selected Publications & Preprints

- Jakob Paul Zimmermann and Georg Loho. *"Hidden Monotonicity: Explaining Deep Neural Networks via their DC Decomposition."* **CVPR 2026 (Highlight)**, 2026.
  Introduces a DC-decomposition-based view of ReLU networks and derives the SplitCAM/SplitLRP attribution methods.

- Jakob Paul Zimmermann, Jim Berend, Georg Loho, Sebastian Lapuschkin, and Wojciech Samek. *"Playing the Network Backward: A Game Theoretic Attribution Framework."* Preprint, arXiv:2605.06212, 2026.
  Recasts backpropagation as a two-player game on an extended computational graph, inducing distributions over computation paths.

- Jakob Paul Zimmermann, Gerrit Holzbach, and David Lerch. *"Knowledge-Guided Failure Prediction: Detecting When Object Detectors Miss Safety-Critical Objects."* **SAIAD Workshop at CVPR 2026**, 2026.
  Uses internal representations and foundation-model embeddings to detect missed safety-critical objects.

- Maria Axenovich and Jakob Zimmermann. *"Induced Turán problem in bipartite graphs."* *Discrete Applied Mathematics*, Volume 360, Pages 497–505, 2025.

- Jakob Zimmermann. *"Bipartite Turán problem on cographs."* Preprint, arXiv:2601.07406, 2026.

Research Experience

Working Student, Fraunhofer HHI Berlin, 2025 – present

- *Transformer interpretability with professor Wojciech Samek. Developing "Playing the Network Backward", a game-theoretic attribution framework that recasts the backward pass as a two-player game on an extended computational graph and identifies computation trajectories stable under input perturbations.*

HiWi Researcher, Discrete Geometry Group (Free University Berlin) Berlin, July 2025 – present

- *With professor Georg Loho, investigating the geometry of the Newton polytopes of neural networks and Difference-of-Convex decompositions. This structure underpins our CVPR 2026 Highlight "Hidden Monotonicity".*

HiWi Researcher / Intern, ML4Safety (Fraunhofer IOSB) Karlsruhe, Fall 2024

- *Runtime monitoring of image-recognition models and representation learning. Developed a knowledge-guided monitoring approach operating on internal representations and foundation-model embeddings such as DINO.*

Education

Technical University Berlin Berlin

- M.Sc. in **Computer Science** Fall 2024 – present

Karlsruhe Institute of Technology (KIT) Karlsruhe

- B.Sc. in **Computer Science** October 2020 – March 2024
- B.Sc. in **Mathematics** April 2021 – March 2024
  - Relevant coursework: Statistical Learning, Convex Geometry, Random Graphs and Networks, Graph Theory, Combinatorics, Planar Graphs, Elementary Geometry.
- Bachelor thesis *"Induced Turán problems"* under supervision of *Maria Axenovich* and *Thorsten Ueckerdt*, resulting in the published paper *"Induced Turán problem in bipartite graphs"* (Discrete Applied Mathematics, 2025).

Haus Lindenhof Foundation Schwäbisch Gmünd
Voluntary social year in a residential group for people with disabilities March 2020 – August 2020

Talks & Presentations

- Talk: *"Bipartite Turán problem on cographs"* at Szabó's Research Seminar, Free University Berlin January 2026
- Poster: *"Hidden Monotonicity"* at Workshop on Polyhedral Geometry for Neural Networks, Nuremberg March 2026
- *"Knowledge-Guided Failure Prediction"* (talk) at SAIAD Workshop at CVPR 2026, Denver 04.06.2026
- *"Hidden Monotonicity"* (Highlight, poster) at CVPR 2026, Denver 07.06.2026
- Talk: *"From DC Decompositions to Two-Player Games: A Structural Approach to Explainability in Deep Learning"* at Gottschalk's Research Seminar, TU Berlin (IMOS) 01.07.2026

Teaching

Tutor & Grader, Graph Theory (KIT Karlsruhe) Fall 2023
Tutor & Grader, Stochastics (KIT Karlsruhe) Spring 2023

Skills

- **Programming:** Python (ML: NumPy, pandas, scikit-learn, PyTorch), C++, Java, LaTeX, SQL, Lean 4 (formal verification). **Familiar with**: Kotlin, R, MATLAB, Haskell, Lisp.
- **Other:** Version Control (Git), Linux Command Line, parallel computing.

Activities & Honors

- *Scheffelpreis* (literary prize for top German graduates), Literarische Gesellschaft (Literary Society) Karlsruhe 2019
- Art exhibition on the subject of mass merchandise, *Haus der Begegnung* (House of Encounter), Ulm 2018
- First prize, *Jugend musiziert* ("Youth Makes Music" national competition), singing duet 2013
- Boys choir, *Collegium Iuvenum* Stuttgart 2011 – 2014