6th Youth in High-Dimensions: Recent Progress in Machine Learning, High-Dimensional Statistics and Inference & How creative is Generative AI? Perspectives from Science and Philosophy | (SMR 4084)

03 Jul 2025 - 10 Jul 2025
ICTP, Trieste, Italy

P01 - ACEVEDO Santiago Daniel

Data binarization captures universal exponents in one-dimensional critical growth dynamics

P02 - ALESSI Michele

DCP: Density-based Clustering Priors for Variational Autoencoders

P03 - ALVAREZ GALNARES Joaquin

Aggregating Image Segmentation Predictions with Probabilistic Risk Control Guarantees

P04 - AMARCHA Fatima Azzahraa

Reinforcement Learning for Efficient Multi-UAV Deployment in Traffic Monitoring System

P05 - BALDWIN Daniel Guy

Process Versus Output: Various Illusions of Originality in Humans and AI

P06 - BELAROUI Khaoula

Beyond Neurons: Understanding the High-Dimensional Network of the Mind with GNNs

P07 - BENAYAD Mohamed

Exploring the Role of Generative AI in Scientific Creativity and Urban Sustainability

P08 - CAMILLI Francesco

Information-theoretic reduction of deep neural networks to linear models in the overparametrized proportional regime

P09 - CHIECO Giovanni

Reimagining Authorship and Ownership of AI-Generated Images under EU Law

P10 - CLAVERINI Corrado

A New Creative Paradigm: AI and the Hybrid Author

P11 - DE PAOLIS Ludovica

Synthesizing Naturalistic Visual Textures Using Deep Neural Samplers

P12 - DÍAZ LEAL Yusimi Lazara

AI Optimization of Blood Transfusions with Probabilistic Models

P13 - DUARTE GOMES Naomy

Crafting Robustness: An Artificial Intelligence and Statistical Toolkit for Advanced Reliability Analysis

P14 - FORNI Ortensia

Color Distributions in Nature: A Data-Driven Exploration of Color Harmony and Aesthetic Preferences

P15 - GENG Mingmeng

Human-LLM Coevolution: Evidence from Academic Writing

P16 - GHIMENTI Federico

The edge of annealing in the coherent Ising machine

P17 - HAYES Patrick Ainsworth

Creativity as a Human Virtue

P18 - HERNANDEZ GUIJARRO Francisco Javier

Anisotropy in the Two-Dimensional Abelian Sandpile Model

P19 - HOVHANNISYAN Vahan

Ensemble inequivalence in Ising chains with competing interactions

P20 - KAMB Mason Daniel

An Analytic Theory of Creativity in Convolutional Diffusion Models

P21 - KDOURI Lahoucine

Decoding Semantic Text from Non-Invasive Brain Recordings

P22 - KHAN Shahzaib

UNVEILING PATTERNS IN MEDICAL IMAGING THROUGH HIGH DIMENSIONAL REPRESENTATIONS

P23 - KOVACEVIC Filip

Spectral Estimators for Multi-Index Models: Precise Asymptotics and Optimal Weak Recovery

P24 - KUEHN Marcel

Explaining Near-Zero Hessian Eigenvalues Through Approximate Symmetries in Neural Networks

P25 - KUNICKI Martha

Can GenAI be creative? AI-Human Co-Creativity: Enriching the Concept of Creativity in Light of Emerging GenAI Towards Collective Co-Creativity of AI and Humans

P26 - LAZAREVIC Dorde

AI and the problem of creative spontaneity

P27 - LIBERATI Diego

Integrating Machine Learning inference to simulation and deduction in math modeling

P28 - MÁCHA Jakub

Metaphorical Masks: Capturing and Concealing LLM Creativity in Art

P29 - MASAI Pierluigi

From the Human/Machine Dichotomy to Distributed Creativity - The Case of AI Integration in Filmmaking

P30 - MASCARETTI Andrea

What's in a caption? Finding semantics with the information imbalance.

P31 - MEDVEDEV Marko

Weak-to-Strong Generalization Even in Random Feature Networks, Provably.

P32 - MENDES DUARTE Pedro Henrique

Restricted Boltzmann Machine Approaches Lattice Systems

P33 - MILANESIO Federico

6th Youth in High-Dimensions: "Understanding Geometric Compression in Neural Networks Dynamics" AI Creativity: "The Myth of the Creativity Dial: Temperature in LLMs"

P34 - MOTOTAKE Yoichi

Signal extraction in high-dimensional data and its kinetic interpretation

P35 - NICOLETTI Flavio

Statistical Mechanics of Hopfield networks near and above saturation

P36 - NIEVES CUADRADO Joan Andres

Aging, Glioblastoma, and Alzheimer's disease in the space of gene expression

P37 - NIKOLAOU Konstantin Lukas

Mechanisms in Neural Scaling Regimes

P38 - NWEMADJI TIAKO Arsene Gibbs

Theoretical Analysis of Generalization of a Two-Layer Neural Networks with Generic Activation: Bayes-Optimal Inference and Empirical Risk Minimization

P39 - PALAVANDISHVILI Ana

Machine Learning model application of spectroscopic data analysis

P40 - PAL Rahul Chandraprakash

Multi-Class Uncertainty-aware Brain Tumor Segmentation using Bayesian Attention U-Net

P41 - PEI Bin

Transfer Learning with Feature Augmentation Enabled by Neural ODE and Path Signature for Class-unbalanced Mechanical Fault Diagnosis

P42 - PIZZOCHERO Michele

Can Machines Philosophize?

P43 - PUSCASU Iulia Dana

Generative AI: collaborator or tool? The experience of interaction in the process of artistic creation

P44 - RICCI Fabiola

Feature learning from non-Gaussian inputs: the case of Independent Component Analysis in high dimensions

P45 - SAEED Basil Nael Omar

Local minima of the empirical risk in high-dimensions: general theorems and convex examples

P46 - SAO TEMGOUA Myke Vital

Can machine learning improve natural product-based antimalarial ?

P47 - SHARMA Pradeep Kumar

Bilevel Optimization for Machine Learning

P48 - SKERK Rudy

Fundamental limits of non-symmetric low-rank matrix estimation with structured noise

P49 - TIBERI Lorenzo

Dissecting the Interplay of Attention Paths in a Statistical Mechanics Theory of Transformers

P50 - TULINSKI Thomas

Spherical Boltzmann Machines

P51 - UMAR Ali Hussaini

EFFECT OF LABEL NOISE ON THE INFORMATION CONTENT OF NEURAL REPRESENTATIONS IN CLASSIFICATION NETWORKS

P52 - URFIN Raphael Maxime Theo

Memorization and Generalization in Generative Diffusion

P53 - VASUDEVA Bhavya

The Rich and the Simple: On the Implicit Bias of Adam and SGD

P54 - VILLEGAS GARCIA Edith Natalia

Interpreting and Steering Protein Language Models through Sparse Autoencoders

P55 - ZHOU Shu

Phase analysis of Ising machines and their implications on optimization

P01

Data binarization captures universal exponents in one-dimensional critical growth dynamics

R. Verdel[1], S. Acevedo [2], D.S. Bhakuni[1], and M. Dalmonte[1]

[1] *International Centre of Theoretical Physics (ICTP)*
[2] *International School for Advanced Studies (SISSA)*

The intrinsic dimension (ID) is a geometric tool of general applicability to detect and quantify correlations from data. Recently, it has been successfully applied to study statistical and many-body systems in equilibrium[1]. Yet, its application to systems away from equilibrium remains largely unexplored. Here we study the ID of nonequilibrium growth dynamics data, and show that even after reducing these data to binary form, their binary intrinsic dimension[2] (BID) retains essential physical information. Specifically, we find that, akin to the surface width, it exhibits Family-Vicsek dynamical scaling, and in one dimensional systems it scales with the same universal exponents as the order parameter.

[1] Mendes Santos et al. Unsupervised learning universal critical behavior via the intrinsic dimension. PRX (2021).

[2] Acevedo et al. Unsupervised detection of semantic correlations in big data. arXiv. 2411.02126 (2025).

DCP: Density-based Clustering Priors for Variational Autoencoders

Michele Alessi[*+], Alejandro Rodriguez[+°], Alessio Ansuini[*], Alberto Cazzaniga[*]
[*]Area Science Park, [+]University of Trieste, [°]SISSA

P02

Abstract

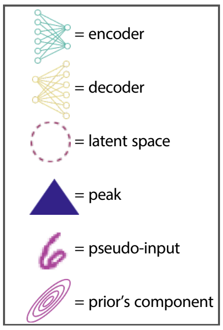
Deep generative models often operate as black boxes, while density-based clustering algorithms excel in unsupervised tasks but face challenges with high-dimensional spaces and generating new data. We propose a novel approach integrating both frameworks by developing a new prior for Variational Autoencoders (VAEs) [1] using the Advanced Density Peaks (DPA) [2] algorithm.

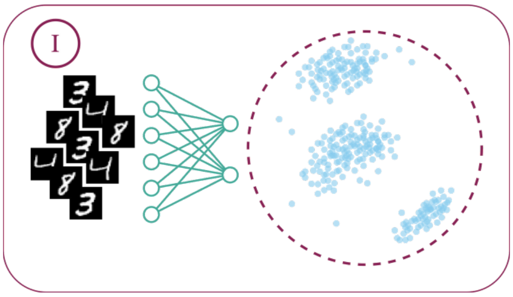
Background

The optimal prior for VAEs aligns with the aggregated variational posterior, a Gaussian mixture derived from encoding all data points. Since this is intractable, approximations like the VAMP prior have been developed [3]. VAMP uses K pseudo-inputs in the data space whose encoding defines the prior. These points are updated via backpropagation.

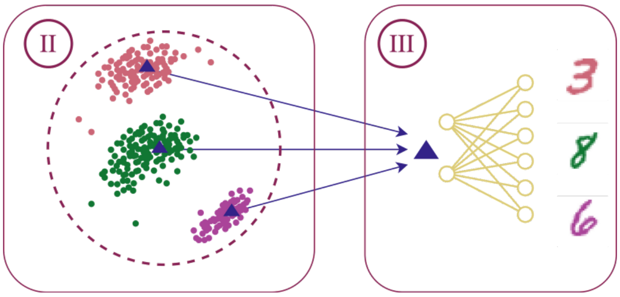
Method

Our method alternates VAE training with DPA density estimation and clustering ensuring a mutual, progressive refinement between them. The process to determine the prior's parameters is carried out **at the end of each epoch.**

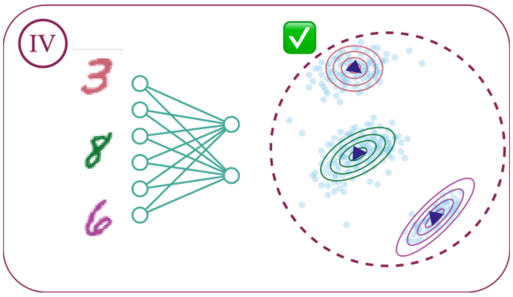




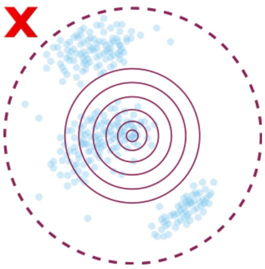
STEP 1: Obtain latent representation



STEP 2: Identify K peaks with DPA **STEP 3:** Generate pseudo-inputs



STEP 4: Encode pseudo-inputs to infer prior's parameters



Standard prior **misaligns** with latent representations

Why our method?

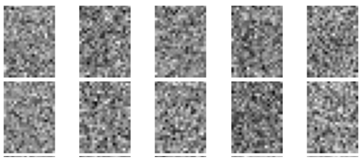
- K **learnable**
- Prior **aligns** with latent space
- **Pseudo**-inputs **interpretability**
- **Heuristics** for:
    - uncertainty quantification
    - adaptive control of prior complexity

Why DPA?

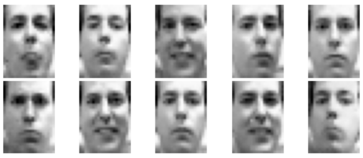
- Non-parametric + density-based
- Clustering
- Fast
- Peaks uncertainty quantification
- Saddle points

Results and Conclusions

DCP pseudo-inputs are interpretable

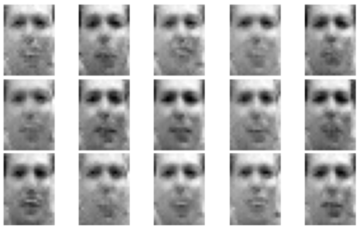


VAMP pseudo-inputs (K=500)

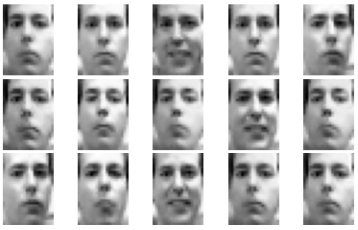


DCP pseudo-inputs (K=11)

DCP generated samples are of high quality



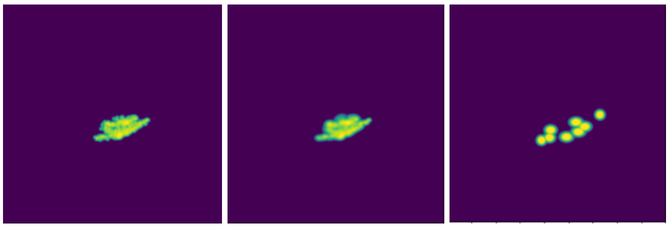
VAMP generated samples (K=500)



DCP generated samples (K=11)

Prior distribution

Aggregated posterior (left), VAMP (center) and DCP prior (right) on FreyFaces dataset, 2-dim latent space. DCP aligns with the aggr. posterior even with a small number of pseudo-inputs.



Aggregated posterior VAMP prior (K=500) DCP prior (K=11)

Work in progress

- Scaling
- Integration into scVI model for scRNA-seq

What's next?

- Helping small datasets
- Controlling number of pseudo-inputs

References: [1] *Kingma et al., Auto-Encoding Variational Bayes, 2013*
[2] *Rodriguez et al., Automatic topography of high-dimensional data sets by non-parametric Density Peak clustering, 2021*
[3] *Tomczak et al., VAE with a VampPrior, 2017*

Contacts:
michele.alessi@areasciencepark.it
github.com/alessimichele

P03

Aggregating Image Segmentation Predictions with Probabilistic Risk Control Guarantees

Joaquin Alvarez[1], **Edgar F. Roman-Rangel**[1]

[1] *ITAM, Mexico*

In this work we introduce a framework to combine arbitrary image segmentation algorithms from different agents under data privacy constraints to produce an aggregated prediction set satisfying finite-sample risk control guarantees. We leverage distribution-free uncertainty quantification techniques in order to aggregate deep neural networks for image segmentation tasks. Our method can be applied in settings to merge the predictions of multiple agents with arbitrarily dependent prediction sets. Moreover, we perform experiments in medical imaging tasks to illustrate our proposed framework. Our results show that the framework reduced the empirical false positive rate by 50% without compromising the false negative rate at any user-specified false negative rate tolerance with high-probability risk control guarantees, with respect to the false positive rate of any of the constituent models in the aggregated prediction algorithm.

P04

Poster

Reinforcement Learning for Efficient Multi-UAV Deployment in Traffic Monitoring Systems

Fatima Azzahraa Amarcha[1] , Rachid Saadane[2] , Rachid Ahl Laamara [1]

[1]LPHE-MS University of Mohammed V Morocco
[2] Electrical Engineering Dep Hassania School
of Public Works Casablanca, Morocco

In intelligent transportation systems, Unmanned Aerial Vehicles (UAVs) are increasingly used to monitor and manage traffic in real time. However, the efficient deployment and coordination of multiple UAVs remains a complex challenge due to dynamic environments, limited resources, and coverage constraints. We address the problem of deploying a swarm of UAVs for traffic monitoring in dynamic urban environments, with the multi-objective goals of minimizing both total flight time and the number of UAVs used. We tackle this challenge using **multi-objective reinforcement learning (MORL)** to optimize UAV trajectories and positioning decisions. Our method enables the UAV swarm to adaptively explore the environment, learning efficient deployment strategies under varying traffic and demand conditions. We demonstrate that the MORL framework discovers a Pareto front of trade-off solutions that outperform conventional heuristic and rule-based approaches in terms of coverage quality and operational cost. This highlights the potential of MORL for real-time, resource-efficient UAV coordination in complex monitoring tasks.

[1] Liu, Xiao & Liu, Yuanwei & Chen, Yue. Reinforcement Learning in Multiple-UAV Networks: Deployment and Movement Design. IEEE Transactions on Vehicular Technology. PP. 1-1. 10.1109/TVT.2019.2922849. (2019)

[2] F. A. Amarcha, A. Chehri, A. Jakimi, M. Bouya, R. Ahl Laamara and R. Saadane, "Drones Optimization for Public Transportation Safety: Enhancing Surveillance and Efficiency in Smart Cities," *2024 IEEE World Forum on Public Safety Technology (WFPST)*, USA, (2024)

[3]Calascibetta, C., Biferale, L., Borra, F. *et al.* Taming Lagrangian chaos with multi-objective reinforcement learning. *Eur. Phys. J. E* **46**, 9. https://doi.org/10.1140/epje/s10189-023-00271-0.(2023)

P05

Process Versus Output: Various Illusions of Originality in Humans and AI

Daniel Guy Baldwin

University of Liverpool

A common argument against the creativity of generative AI is that it lacks originality, but the creative process of human beings and that of generative AI does not seem so far removed as people are inclined to believe initially. Generative AI uses machine-learning to produce novel data – on a basic level this might involve a simple model such as a Markov chain. The next word in a text is generated based on the previous ones. A more modern generative AI such as ChatGPT functions at a base level in essentially the same manner but with greater complexity and parameters, trained on a massive dataset of text and code available on the Internet.

This does not appear too dissimilar to certain forms of creative process in human beings. I am not a tabula rasa when it comes to creation. I do not produce something out of nothing. I have a lived experience which, along with its interpretation, fills the function of both dataset and parameters in my creative process. The way I write is informed by the books I've read; the themes by the life I have lived. In one sense, the next word in my text is generated based on previous ones. One might say there is still a greater degree of originality in human creation which does not exist with generative AI, but this is not always the case.

Some forms of art involve direct modification of previous artworks. This is especially true within music: hip hop often samples older tracks; interpolation uses a previously recorded melody in a new recording; song covers can involve little to no unique ideas. There have also been cases where musicians have unintentionally copied other artists' tracks (see Nine Inch Nails' A Warm Place and David Bowie's Crystal Japan). It is also relatively easy to find tracks made entirely from samples online. It seems unlikely we would deny these were creative endeavours although they only involve reorganising previously available data.

From this it seems we also judge originality and creativity by the product produced rather than strictly by the process alone.

P06

Abstract template for 6th Youth in High-Dimensions: Recent Progress in Machine Learning, High-Dimensional Statistics and Inference & How creative is Generative AI? Perspectives from Science and Philosophy

Khaoula Belaroui[1]

1 *(Presenting author)* PhD Student, University Hassan 2 – FST

Title: *Beyond Neurons: Understanding the High-Dimensional Network of the Mind with GNNs*

Abstract:

The human brain is a complex, high-dimensional network, where billions of neurons form dynamic, interconnected systems that drive cognition, behavior, and possibly even consciousness. Graph Neural Networks (GNNs) offer a powerful framework to model these intricate structures, capturing not only local interactions but also global emergent properties. My research explores how GNNs can decode these high-dimensional neural networks, modeling brain regions as nodes and synaptic connections as edges — ultimately revealing the architecture that underlies mental processes. This research not only decodes complex algorithms and data but also embarks on a philosophical exploration of how machine learning can help us unravel the profound mysteries of the mind. I aim to integrate hyperbolic geometry to more accurately represent the brain's hierarchical, non-Euclidean nature, potentially offering a deeper and more faithful understanding of cognition, neurological disorders, and even consciousness. As I continue my work with GNNs, my goal is to develop a model that not only reflects the brain's complexity but also helps us understand it more deeply. This work bridges neuroscience, machine learning, and graph theory, advancing our exploration of the mind's vast, untapped complexities.

[1] Yang, M., Zhou, M., Li, Z., Liu, J., Pan, L., Xiong, H., & King, I. (2023). *Hyperbolic Graph Neural Networks: A Review of Methods and Applications*.

[2] O. Sporns, *"Graph theory methods: applications in brain networks"*, *Dialogues in Clinical Neuroscience*, vol. 24, no. 4, pp. 331-343, 2022.

[3] A. Bessadok, M. A. Mahjoub, and I. Rekik, *"Graph Neural Networks in Network Neuroscience"*, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 7, pp. 3453-3466, 2022.

P07

Exploring the Role of Generative AI in Scientific Creativity and Urban Sustainability

Mohamed Benayad[1,2], Abdelilah Rochd[2], Mehdi Maanan[1], Hassan Rhinane

[1]*Geosciences Laboratory, Faculty of Sciences-Ain Chock, Hassan II University, Casablanca, Morocco*

[2]*Green Energy Park (GEP), Benguerir, Morocco*

Generative Artificial Intelligence (GenAI) technologies such as ChatGPT, DALL·E, and Stable Diffusion have significantly transformed the landscape of creativity and scientific research. This poster explores the role of GenAI in enhancing creative processes across disciplines, with a particular focus on urban sustainability, geospatial analysis, and scientific communication. By leveraging GenAI tools for tasks such as automated text generation, data synthesis, and scenario visualization, researchers can accelerate ideation, refine hypotheses, and communicate complex information more effectively. Drawing on case studies in sustainable city planning and intelligent infrastructure design, we present how generative models can assist in the development of research proposals, generate alternative urban design scenarios, and simulate potential environmental outcomes. The poster also discusses the limitations of current GenAI systems, including issues of hallucination, reproducibility, and ethical transparency. A critical reflection is offered on the epistemological implications of co-creating with AI and the evolving nature of human–machine collaboration in the scientific process. This contribution aims to initiate a dialogue on the creative potential of GenAI while advocating for responsible and reflective integration in academic research and interdisciplinary problem-solving.

P08

Information-theoretic reduction of deep neural networks to linear models in the overparametrized proportional regime

F. Camilli[1], **D. Tieplova**[1], **E. Bergamin**[2], **and J. Barbier**[1]

[1]*International Centre for Theoretical Physics*
[2] *International School for Advanced Studies*

We rigorously analyze fully trained neural networks of arbitrary depth in the Bayesian optimal setting in the so-called proportional scaling regime where the number of training samples and width of the input and all inner layers diverge proportionally. We prove an information-theoretic equivalence between the Bayesian deep neural network model trained from data generated by a teacher with matching architecture, and a simpler model of optimal inference in a generalized linear model. This equivalence enables us to compute the optimal generalization error for deep neural networks in this regime. We thus prove the ”deep Gaussian equivalence principle” conjectured in [2]. Our result highlights that in order to escape this ”trivialization” of deep neural networks (in the sense of reduction to a linear model) happening in the strongly overparametrized proportional regime, models trained from much more data have to be considered.

[1] F. Camilli, D. Tieplova, E. Bergamin, and J. Barbier, *Information-theoretic reduction of deep neural networks to linear models in the overparametrized proportional regime*, 38th Annual Conference on Learning Theory (COLT 2025), accepted, to appear (2025)

[2] H. Cui, F. Krzakala and L. Zdeborová, *Bayes-optimal Learning of Deep Random Networks of Extensive-width*, Proceedings of the 40th International Conference on Machine Learning, PMLR 202:6468-6521 (2023)

Abstract template for How Creative is Generative AI? Perspective form Science and Philosophy

-

Reimagining Authorship and Ownership of AI-Generated Images under EU Law

Giovanni Chieco

(PhD Candidate in Applied Data Science and Artificial Intelligence at The University of Trieste – AI Law; Trainee Lawyer)

The proliferation of generative artificial intelligence (AI) technologies—such as DALL·E and Midjourney —has introduced significant disruptions to the European creative landscape. These tools, capable of autonomously producing complex visual content, challenge the foundational assumptions of European intellectual property law, particularly with respect to authorship, originality and ownership.

This paper critically examines the current regulatory vacuum within the European Union (EU) regarding the legal status of AI-generated images. While the EU copyright framework—primarily governed by the InfoSoc Directive (2001/29/EC) and national implementations—continues to rely on the principle of human authorship, there remains no explicit guidance for works autonomously produced by non-human agents. As a result, such outputs often fall outside the scope of protection, leading to legal uncertainty and a potential chilling effect on innovation and investment in creative AI applications.

The analysis focuses on three interrelated dimensions. First, it explores the doctrinal limitations of EU copyright law in accommodating non-human creativity, particularly the requirement of "authorial" originality. Second, it addresses the implications of private governance, wherein AI tool providers assert control over outputs through contractual terms, effectively substituting public copyright regimes with privatized models of regulation. Third, it considers emerging legislative initiatives at the EU level and assesses their relevance for addressing the governance of AI-generated content, despite their limited direct engagement with intellectual property concerns.

In light of these challenges, the paper argues for the development of a harmonised European approach that recognises the hybrid nature of human–AI collaboration. It proposes a set of policy options, including the introduction of sui generis protection for AI-generated works, the clarification of human involvement thresholds for copyright eligibility and the encouragement of open licensing models to support legal certainty, fair access and responsible innovation in the European creative economy

[1] P.B. Hugenholtz, J.P. Quintais, *Copyright and Artificial Creation: Does EU Copyright Law Protect AI-Assisted Output?*, Int. Rev. Intell. Prop. Compet. Law 52, 1190–1216 (2021).
[2] D. Bietti, *Generative AI in EU Law: Liability, Privacy, Intellectual Property, and Cybersecurity*, Comput. Law Secur. Rev. (2024). Available at: sciencedirect.com
[3] C. Di Bari, *AI Generated Works: The Clash Between Copyright Principles and Technological Evolution*, Law and Media Working Paper Series (2021). Available at: medialaws.eu
[4] G. Frosio, M. Jütte, *Developing Artificial Intelligence-Based Content Creation: Are EU Copyright and Antitrust Law Fit for Purpose?*, Int. Rev. Intell. Prop. Compet. Law 54, 351–380 (2023).
[5] M. Iglesias, *Generative AI, Copyright and the AI Act*, Comput. Law Secur. Rev. (2025). Available at: sciencedirect.com

[6] J. Lee, H. Yuan, *Navigating the Legal Landscape of AI Copyright: A Comparative Analysis of EU, US, and Chinese Approaches*, AI Ethics 4, 10 (2024). Available at: link.springer.com
[7] J. Rosati, *Copyright and Artificial Creation: Does EU Copyright Law Protect AI-Assisted Output?*, Int. Rev. Intell. Prop. Compet. Law 52, 313–340 (2021).
[8] S. Wang, *Between Copyright and Computer Science: The Law and Ethics of Generative AI*, arXiv (2024). Available at: arxiv.org
[9] S. Wang, *Copyright Protection for AI-Generated Works: Exploring Originality and Ownership in a Digital Landscape*, Asian J. Int. Law (2024). Available at: cambridge.org

P10

A New Creative Paradigm: AI and the Hybrid Author

Corrado Claverini[1]
[1]University of Salento (corrado.claverini@unisalento.it)

Indeterminacy is a key feature of generative AI systems. While tools like DALL·E, Midjourney, or Stable Diffusion offer various degrees of user control, the outcome is never fully predictable. Features such as FaceSwap, Inpaint, Outpaint, LoRA models, and guidance scales enable customization while maintaining this inherent unpredictability.

This tension led scholars to describe AI-assisted creators as “hybrid authors,” whose agency is shared with the machine. Yet this hybridity does not invalidate their status as authors. On the contrary, it resonates with historical precedents, such as the 1884 U.S. Supreme Court case *Burrow-Giles Lithographic Co. v. Sarony*, which recognized a photograph as art when it reflected the photographer’s creative input.

In the art world, generative AI may have an impact comparable to that of photography in the past. The advent of photography was initially perceived by many artists as a threat that rendered the traditional role of the artist obsolete. Charles Baudelaire, for instance, saw photography as a mechanical reproduction devoid of true aesthetic value and a danger to the artistic vocation. This reaction exemplifies a broader pattern often described as *media panic* – a recurring dynamic triggered by the arrival of disruptive technologies. From the printing press to radio, television, video games, and photography – and now to generative artificial intelligence – each new medium has been met with a wave of alarmism. Scholars have termed this recurring phenomenon the *Sisyphean cycle of technology panics*: a pattern in which early fears surrounding innovation are often exaggerated, amplified by political rhetoric, media discourse, and academic narratives that emphasize risks over potential societal benefits. Over time, however, these fears become normalized, and the technology becomes integrated into everyday life.

Against this backdrop, this talk aims to explore how generative AI challenges traditional notions of creativity and authorship. It encourages us to reflect critically on our concepts of artistic agency and consider whether we are witnessing the emergence of a new creative paradigm – one in which human and machine contributions are not opposed, but intertwined in a meaningful form of collaboration.

P11

Synthesizing Naturalistic Visual Textures Using Deep Neural Samplers

Ludovica de Paolis[1], Eugenio Piasini[1], Fabio Anselmi[2] and Alessio Ansuini[3]

[1]*International School for Advanced Studies - SISSA*
[2]*University of Trieste*
[3]*AREA Science Park, Research and Technology Institute*

The goal of this study is to generate naturalistic visual textures with a novel generative architecture and to investigate their representations through their statistical and semantic properties. In neuroscience, the efficient coding principle states that the mammalian visual system conforms to the statistics of its natural input: natural visual scenes. Visual textures are a family of visual patterns that share certain local regularities [1] that have been used to test efficient coding. Still, experimental results are limited to unnatural looking low-order correlations. There are three main texture synthesis algorithms: parametric models [2], maximum entropy sampling methods [3], Gram Matrix-based convolutional neural networks [4]. We propose a new model for texture synthesis: a combination of a Variational Autoencoder (VAE [5]) and a pre-trained convolutional neural network (VGG-16 [6]) to generate realistic textures in the latent space, only characterized by nonlinear multi-scale representations. The model is trained with unsupervised learning to minimize: a Kullback-Leibler divergence between the latent code and a Gaussian distribution; a Mean Squared Error between pairs of Gram Matrices computed on VGG-16 feature maps to define the perceptual difference between original and generated textures. We are currently training the model with the Describable Textures Dataset (DTD [7]), a supervised dataset of 5640 images of textures divided in 47 classes. We successfully replicated the of study by [4] using DTD and studied the ability of Gram Matrices to capture statistical information of textures with Representational Similarity Analysis (RSA [8]) on the Gram Matrices produced with Gatys' approach. We also performed a zero-shot learning experiment with CLIP [9] on DTD with the intention to investigate the semantic attributes that characterize textures. We expect our model to produce a texture space that aligns with human perception, and the generated textures to show high variance statistical features, which will reveal how efficient coding applies in our behavioral and computational regime. We expect Gram Matrices to capture the statistical information contained in the textures, and therefore to best generate highly abstract textures; we expect textures that are more similar to objects to be captured by semantic features as represented by CLIP.

References

[1] B. Julesz. "Textons, the elements of texture perception, and their interactions". In: *Nature* 290 (1981), pp. 91–97.

[2] J. Portilla E. P. Simoncelli. "A parametric texture model based on joint statistics of complex wavelet coefficients". In: *International Journal of Computer Vision* 40 (2000), pp. 49–70.

[3] J. D. Victor M. M. Conte. "Local image statistics: maximum entropy constructions and perceptual salience". In: *Optical Society of America* 29 (2012), pp. 1313–1345.

[4] L. A. Gatys et al. "Texture Synthesis Using Convolutional Neural Networks". In: *Advances in Neural Information Processing Systems 28 (NIPS 2015)* (2015).

[5] D. P. Knigma M. Wellig. "Auto-Encoding Variational Bayes". In: *ArXiv* 11 (2022).

[6] K. Simonyan A. Zisserman. "Very Deep Convolutional Networks for Large Scale Image Recognition". In: *ICLR 2015* (2015).

[7] M. Cimpoi et al. "Describing Textures in the Wild". In: *CVPR 2014* (2014).

[8] N. Kriegeskorte M. Mur P. Bandettini. "Representational similarity analysis – connecting the branches of systems neuroscience". In: *Front. Syst. Neurosci* (2008).

[9] A. Radford et al. "Learning Transferable Visual Models From Natural Language Supervision". In: *arXiv* (2021).

P12

AI Optimization of Blood Transfusions with Probabilistic Models

P13

Crafting Robustness: An Artificial Intelligence and Statistical Toolkit for Advanced Reliability Analysis

Naomy Duarte Gomes[1], Oílson Gonzatto [1], and Francisco Louzada[1]

[1]*Institute of Mathematical and Computer Sciences (ICMC), University of São Paulo*

This research introduces a comprehensive toolkit that integrates artificial intelligence and statistical methodologies for reliability analysis using Python, specifically designed to address complex assessment needs within industrial settings. Motivated by the requirements of Petrobras projects, notably involving robotic technologies and safety-critical equipment, the toolkit integrates theoretical advancements with practical methodologies. It encompasses modules dedicated to accelerated life testing, degradation testing, and accelerated degradation testing. The modular structure facilitates preprocessing of diverse datasets and rigorous reliability analysis through classical, advanced, and innovative statistical models, including Weibull, Exponential, GTDL (Generalized Time-Dependent Logistic), Log-Normal, Gamma and GTDLF (Generalized Time-Dependent Logistic with Frailty)[1,2]. Additionally, state-of-the-art large language models (LLMs) are integrated to aid users in interpreting complex results, providing clear explanations and actionable insights. Detailed diagnostic capabilities ensure the integrity of model fitting and analytical outcomes. Furthermore, interactive functionalities allow users to craft customized reliability scenarios, significantly enhancing decision-making through clear, interpretable outputs. Ultimately, this platform aims to cultivate a culture of reliability analytics, bridging technical precision with practical usability, and providing vital insights for quality assurance and reliability management throughout different phases of technological development.

[1] Collins, D. H., Freels, J. K., Huzurbazar, A. V., Warr, R. L., and Weaver, B. P., **45**(3), 244–259 (2013).

[2] Eduardo Calixto. Gas and oil reliability engineering: modeling and analysis. Gulf Professional Publishing, (2013).

P14

Color Distributions in Nature:
A Data-Driven Exploration of Color Harmony and Aesthetic Preferences

O. Forni[1], **A. Darmon**[2], **M. Benzaquen**[1,3,4]

[1] *Chair of Econophysics & Complex Systems, Ecole polytechnique, 91128 Palaiseau Cedex, France*
[2]*Art in Research, 33 rue Censier, 75005 Paris, France*
[3] *Ladhyx UMR CNRS 7646, Ecole polytechnique, 91128 Palaiseau Cedex, France*
[4]*Capital Fund Management, 23 Rue de l'Université, 75007 Paris, France*

Our research explores the relationship between color appreciation and statistical properties, extending previous work [1] on grayscale images, which showed that aesthetic appreciation peaks at intermediate entropic complexity, thus suggesting universal criteria for aesthetic judgment.

Since previous studies [1, 2] showed that people's preferences are influenced by natural environment, we investigate whether similar principles apply to color images by analyzing the color distribution of a large dataset of natural pictures - extracting their dominant hues and identifying a common distribution of colors in nature.
First we examine whether natural color distributions align with the color harmony theory of Moon and Spencer (1944) [3], which is still widely referenced in various fields, despite the lack of clear evidence supporting the applicability of this theory. By testing this model in an agnostic manner through a large-scale survey, we identify a hue-dependent effect and develop a more comprehensive quantitative framework for color harmony and aesthetic appreciation in color pairings.
Then, on the basis of this analysis, we construct a novel hue space that redistributes color spectrum according to frequency of nature. The application of Multiscale Relevance [4] in this context provides a new approach to quantify how different color spaces capture the information content of an image.

By integrating statistical physics with empirical studies, our work aims to address color appreciation and harmony in a truly quantitative manner. It bridges aesthetic perspectives and computational methods, aiming also to provide physically grounded tools for color-based image analysis and interpretation and to enhance the physical interpretability of image processing by studying color distributions.

[1] S. Lakhal, A. Darmon, J-P. Bouchaud, and M. Benzaquen. "Beauty and structural complexity". In: *Physical Review Research* 2.2 (June 2020).

[2] G. J.Sthephens, T. Mora, G. Tkacik, and W. Bialek. "Statistical Thermodynamics of Natural Images". In: *Phys. Rev. Lett.* 110, 018701 (2013).

[3] P. Moon and D. Spencer. "Aesthetic measure applied to color harmony". In: *JOSA* 34.4 (1944), pp. 234–242.

[4] S. Lakhal, A. Darmon, I. Mastromatteo, M. Marsili, and M. Benzaquen. "Multiscale relevance of natural images". In: *Scientific Reports* 13.1 (2023), p. 14879.

P15

Human-LLM Coevolution: Evidence from Academic Writing

Mingmeng Geng[1,2], Roberto Trotta[1,3]

[1] *International School for Advanced Studies (SISSA)*
[2] *École Normale Supérieure (ENS) - PSL*
[3] *Imperial College London*

With a statistical analysis of arXiv paper abstracts, we report a marked drop in the frequency of several words previously identified as overused by ChatGPT, such as “delve”, starting soon after they were pointed out in early 2024. The frequency of certain other words favored by ChatGPT, such as “significant”, has instead kept increasing. These phenomena suggest that some authors of academic papers have adapted their use of large language models (LLMs), for example, by selecting outputs or applying modifications to the LLM outputs. Such coevolution and cooperation of humans and LLMs thus introduce additional challenges to the detection of machine-generated text in real-world scenarios. Estimating the impact of LLMs on academic writing by examining word frequency remains feasible [1, 2], and more attention should be paid to words that were already frequently employed, including those that have decreased in frequency due to LLMs’ disfavor.

[1] M. Geng and R. Trotta. Is chatgpt transforming academics’ writing style? arXiv preprint arXiv:2404.08627, 2024.

[2] W. Liang, Y. Zhang, Z. Wu, H. Lepp, W. Ji, X. Zhao, H. Cao, S. Liu, S. He, Z. Huang, et al. Mapping the increasing use of llms in scientific papers. arXiv preprint arXiv:2404.01268, 2024.

The edge of annealing in the coherent Ising machine

F. Ghimenti[1], A. Sriram[2], A. Yamamura[1] and S. Ganguli[1]

[1]*Department of Applied Physics, Stanford University, Stanford, CA 94305, USA*
[2]*Department of Physics, Stanford University, Stanford, CA 94305, USA*

The solution of hard combinatorial optimization problems is an ubiquitous task across many domains of pure and applied science. Recent years have seen a sprouting of interest toward analog computing devices, which overcome the demise of Moore's law and allow to tackle optimization problems at an unprecedented scale [1]. A theoretical understanding of the performance of these devices, as well as its connection with the properties of the energy landscapes explored, has started to emerge only recently [2]. In this work [3], we consider the coherent Ising machine, an optical network that tackles quadratic unconstrained binary optimization problems by adiabatically tuning an annealing parameter in a space of softened, continuous spins. We focus on the low energy states of this device when attacking the Sherrington-Kirkpatrick problem with ferromagnetic alignment, a prototypical example of a hard combinatorial optimization problem. We build the full phase diagram for low energy minima of this system as the annealing parameter and the ferromagnetic strength change, and characterize the linear excitations around these minima. We find a region in the phase diagram where low energy minima have a spin glass structure, but exhibit soft modes. By combining dynamical mean field theory and numerical simulations, we find that these soft modes can be exploited during the annealing dynamics to find lower energy solutions to the combinatorial optimization problem. Conversely, when these modes disappear, the quality of the solution does not improve upon further annealing. Through a Kac-Rice calculation, we show that the properties of the most abundant minima are sensibly different then the ones explored by the annealing dynamics. Finally, we consider the low energy states of the coherent Ising machine within another ensemble, with predefined ground state and tunable ruggedness: the Wishart planted ensemble [4]. We unravel the regions where no recovery of the ground state is possible and the regions where full recovery takes place.

[1] N. Mohseni, P. L. McMahon, and T. Byrnes, Nat. Rev. Phys., 2022
[2] A. Yamamura, H. Mabuchi, S. Ganguli, Phys. Rev. X, 2024
[3] F. Ghimenti, A. Sriram, A. Yamamura, S. Ganguli, in preparation, 2025
[4] F. Hamze, J. Raymond, C. A. Pattison, K. Biswas, H. G. Katzgraber, Phys. Rev. E, 2020

P17

Creativity as a Human Virtue

Pat Hayes

University of Glasgow

Drawing from Swanton's book *Target-Centred Virtue Ethics* and Pettigrove's explorations of creativity in virtue theoretical space, this paper will examine how we can derive a general framework in which to understand creativity and its relation to human agents which can then be applied to the aesthetic virtue of creativity and its possible appearance in AI. Within this framing, creativity may be conceptualised as a character trait possessed and acted upon by human agents as they hit their socio-culturally defined, Aristotelian mean targets of excellent action in creation, which may result in artistic achievement for a particular agent's development or art. Once this conception has been put forward and defended, I will argue that this conception may be applied to cases of existing and possible future AI to show that creativity is a human virtue that may only be possessed, acted upon by, and motivate human agents, who are themselves enculturated in a particular society or dwelling. By doing so, I aim to show that the only fitting attribution of creativity is to human agents.

P18

Abstract template for Anisotropy in the Two-Dimensional Abelian Sandpile Model

Hernández, F. [1] **and Xulvi-Brunet, R.** [1]

[1] (Francisco Hernández) National Polytechnic School, Physics Department

Complex systems, such as the Abelian sandpile model, are fundamental for understanding phenomena of self-organized criticality (SOC), a key concept in the description of both natural and social processes. In this study, we simulated an anisotropic variant of the Abelian sandpile model, in which the probability of a sand grain being redistributed during an avalanche depends on a predefined direction. The simulations show that breaking the isotropy of avalanche dynamics in this way does not alter the characteristic scale-free behavior of SOC systems: avalanche sizes and interevent times are still well described by Gutenberg-Richter and Omori-Utsu–type laws, respectively.

[1] Per Bak. How nature works: the science of self-organized criticality. Springer Science & Business Media, 2013.
[2] Prince Alex, Benjamin Andres Carreras, Saravanan Arumugam, and Suraj Kumar Sinha. Self-organized criticality in a cold plasma. Physics of Plasmas, 24(12):120701, 12 2017.

Ensemble inequivalence in Ising chains with competing interactions

Alessandro Campa[1,2], Vahan Hovhannisyan[3],
Stefano Ruffo[4,5,6] and Andrea Trombettoni[7,8]

[1] *National Center for Radiation Protection and Computational Physics, Istituto Superiore di Sanit`a, Viale Regina Elena 299, 00161 Roma, Italy*
[2] *INFN Roma1, 00185 Roma, Italy*
[3] *A. I. Alikhanyan National Science Laboratory, Alikhanyan Br. 2, 0036 Yerevan, Armenia*
[4] *SISSA, via Bonomea 265, I-34136 Trieste, Italy*
[5] *INFN, Sezione di Trieste, I-34151 Trieste, Italy*
[6] *Istituto dei Sistemi Complessi CNR, via Madonna del Piano 10, I-50019 Sesto Fiorentino, Italy*
[7] *Department of Physics, University of Trieste, Strada Costiera 11, I-34151 Trieste, Italy*
[8] *CNR-IOM DEMOCRITOS Simulation Center, Via Bonomea 265, I-34136 Trieste, Italy*

E-mail: *alessandro.campa@iss.it*, *gori@sissa.it*, *v.hovhannisyan@yerphi.am*, *ruffo@sissa.it*, *andreatr@sissa.it*

We investigate how competing interactions influence the inequivalence between statistical ensembles by analyzing a one-dimensional Ising model with ferromagnetic mean-field interactions and short-range nearest-neighbor (NN) and next-NN couplings which can be either ferromagnetic or antiferromagnetic. Although the model remains relatively simple, our microcanonical calculations uncover a complex and rich phase diagram. When compared to the canonical ensemble, significant differences emerge: certain phase transitions appear at different locations or are entirely absent in one ensemble. For instance, in a particular region of the phase space where the canonical ensemble exhibits a critical point and a critical end point, the microcanonical ensemble additionally features another critical point and a triple point. The regions of ensemble inequivalence typically occur at lower temperatures and at larger absolute values of the competing couplings.

P20

An Analytic Theory of Creativity in Convolutional Diffusion Models

Mason Kamb [1] **, Surya Ganguli**[1]

[1] *Department of Applied Physics, Stanford University*

Score-matching diffusion models can generate highly original images that lie far from their training data. However, optimal score-matching theory suggests that these models should only be able to produce memorized training examples. To reconcile this theory-experiment gap, we identify two simple inductive biases, locality and equivariance, that: (1) induce a form of combinatorial creativity by preventing optimal score-matching; (2) result in fully analytic, completely mechanistically interpretable, local score (LS) and equivariant local score (ELS) machines that, (3) after calibrating a single time-dependent hyperparameter can quantitatively predict the outputs of trained convolution only diffusion models (like ResNets and UNets) with high accuracy (median r^2 of 0.95, 0.94, 0.94, 0.96 for our top model on CIFAR10, FashionMNIST, MNIST, and CelebA). Our model reveals a *locally consistent patch mosaic* mechanism of creativity, in which diffusion models create exponentially many novel images by mixing and matching different local training set patches at different scales and image locations. Our theory also partially predicts the outputs of pre-trained self-attention enabled UNets (median r^2 ~ 0.77 on CIFAR10), revealing an intriguing role for attention in carving out semantic coherence from local patch mosaics.

Decoding Semantic Text from Non-Invasive Brain Recordings

Lahoucine Kdouri[1], Youssef Hmamouche[1] and Amal El Fallah Seghrouchni[1,2]

[1]*International Artificial Intelligence Center of Morocco, Mohammed VI Polytechnic University*
[2]*Sorbonne Université, LIP6 - UMR 7606 CNRS, France*

Reconstructing semantic content such as textual information from non-invasive brain recordings remains a major scientific challenge. Although functional magnetic resonance imaging (fMRI) offers high spatial resolution, its slow blood-oxygen-level–dependent (BOLD) response and the variability introduced by different scanners and participants demand rigorous denoising and alignment before reliable language decoding is possible. Building on our previous end-to-end multimodal model for spoken-text reconstruction from fMRI [1], we introduce an extended framework. First, a brain module projects pre-processed fMRI patterns into the token-embedding space of an autoregressive diffusion model [2], establishing a direct correspondence between neural activity and textual representations. This encoder is subsequently combined with a frozen GPT-3 or LLaMA decoder in a multimodal generative pre-training phase, enabling coherent, open-vocabulary text reconstruction from non-invasive brain signals.

[1] Hmamouche, Y., Chihab, I., Kdouri, L. & Seghrouchni, A. A multimodal LLM for the non-invasive decoding of spoken text from brain recordings. *ArXiv Preprint ArXiv:2409.19710*. (2024).

[2] Wu, T., Fan, Z., Liu, X., Zheng, H., Gong, Y., Shen, Y., Jiao, J., Li, J., Wei, Z., Guo, J., Duan, N. & Chen, W. AR-Diffusion: Auto-Regressive Diffusion Model for Text Generation. *Advances In Neural Information Processing Systems*. **36** pp. 39957-39974 (2023).

P22

UNVEILING PATTERNS IN MEDICAL IMAGING THROUGH HIGH DIMENSIONAL REPRESENTATIONS

Shahzaib Khan[12]

[1]Centre for Artificial Intelligence in Health Sciences, International Center for Chemical and Biological Sciences, University of Karachi

[2] NED University of Engineering and Technology

The rapid expansion of high resolution medical imaging technologies has led to increasingly complex and high-dimensional datasets, posing significant challenges for traditional data analysis and diagnostic pipelines. These datasets, often noisy and sparsely labeled, require robust computational tools to extract clinically relevant insights. In this work, we investigate the potential of modern machine learning approaches, specifically dimensionality reduction, feature engineering, and generative modeling, to reveal hidden patterns in these data and support advanced decision-making.

We propose a multi stage framework that combines Principal Component Analysis (PCA) for initial dimensionality compression, followed by deep autoencoders to learn nonlinear representations from MRI and CT scans. To further enrich the latent space and support synthetic data generation, we incorporate diffusion models trained to reverse Gaussian noise over multiple time steps. This architecture not only enhances anomaly detection and fine-grained image segmentation, but also facilitates the generation of realistic medical samples for training augmentation and simulation.

Our framework has been tested on publicly available medical imaging datasets, demonstrating superior performance in extracting meaningful anatomical and pathological patterns compared to conventional CNN-based baselines [1]. Additionally, we emphasize the importance of building explainable and privacy-conscious systems by adopting structured latent spaces, which support clearer interpretability and compliance with data governance standards [2]. By transforming raw data into compact and actionable representations, our approach helps bridge the gap between complex image data and clinically applicable outcomes, offering a scalable path for AI-assisted healthcare solutions.

[1] Litjens, G., et al. “A survey on deep learning in medical image analysis.” Medical Image Analysis 42 (2017): 60–88.

[2] Chen, R. J., et al. “Synthetic data in machine learning for medicine and healthcare.” Nature Biomedical Engineering 5.6 (2021): 493–497.

P23

Spectral Estimators for Multi-Index Models: Precise Asymptotics and Optimal Weak Recovery

Filip Kovačević[1], **Yihan Zhang**[2], **and Marco Mondelli**[1]

[1] *Institute of Science and Technology Austria*
[2] *University of Bristol*

Multi-index models provide a popular framework to investigate the learnability of functions with low-dimensional structure and, also due to their connections with neural networks, they have been object of recent intensive study. In this paper, we focus on recovering the subspace spanned by the signals via spectral estimators – a family of methods routinely used in practice, often as a warm-start for iterative algorithms. Our main technical contribution is a precise asymptotic characterization of the performance of spectral methods, when sample size and input dimension grow proportionally and the dimension p of the space to recover is fixed. Specifically, we locate the top-p eigenvalues of the spectral matrix and establish the overlaps between the corresponding eigenvectors (which give the spectral estimators) and a basis of the signal subspace. Our analysis unveils a phase transition phenomenon in which, as the sample complexity grows, eigenvalues escape from the bulk of the spectrum and, when that happens, eigenvectors recover directions of the desired subspace. The precise characterization we put forward enables the optimization of the data preprocessing, thus allowing to identify the spectral estimator that requires the minimal sample size for weak recovery.

P24

Explaining Near-Zero Hessian Eigenvalues Through Approximate Symmetries in Neural Networks

Marcel Kühn[1] **and Bernd Rosenow**[1]

[1]*Institute for Theoretical Physics, Leipzig University*

The Hessian matrix, representing the second derivative of the loss function, offers crucial insights into the loss landscape of neural networks and significantly influences optimization algorithms, model design, and generalization in deep learning [1, 2]. A common characteristic of the Hessian eigenspectrum is the presence of a few large eigenvalues alongside a bulk of near-zero eigenvalues [3].

We suggest that this bulk structure may be connected to dataset-independent symmetries inherent in network architectures - symmetries which, while sometimes overlooked, have been shown to impose strong constraints on gradients and curvature [4]. We extend this line of reasoning by proposing that approximate variants of such symmetries explain the bulk of small eigenvalues in the Hessian spectrum. First, we demonstrate that in deep, fully connected linear networks, exact continuous symmetries that leave the loss invariant lead to zero eigenvalues in the Hessian. These zero eigenvalues and their corresponding eigenvectors can be attributed to symmetries such as rotations between weight layers.

Extending this concept, we suggest that in networks with nonlinear activation functions, approximate symmetries give rise to a large number of small but finite eigenvalues, which can be viewed as perturbations of the linear case. We illustrate this phenomenon in a two-layer ReLU student-teacher setup, analytically showing that, in the infinite-width limit and for a straightforward weight configuration, Hessian eigenvectors associated with small eigenvalues lie entirely within the space of symmetry directions known from the linear case. Furthermore, we train a multi-layer fully connected network on CIFAR-10 and demonstrate that, there too, eigenvectors with small eigenvalues predominantly align with symmetry directions. Finally, we apply our symmetry-based analysis to convolutional networks, demonstrating the generality of our approach in understanding the Hessian eigenspectrum across different architectures.

[1] J. Martens, Deep learning via hessian-free optimization, ICML (2010).

[2] N. S. Keskar, D. Mudigere, J. Nocedal, M. Smelyansky, P. T. P. Tang, On Large-Batch Training for Deep Learning: Generalization Gap and Sharp Minima, ICLR (2017).

[3] L. Sagun, U. Evci, V. U. Guney, Y. Dauphin, L. Bottou, Empirical analysis of the hessian of overparametrized neural networks, ArXiv:1706.04454 (2017).

[4] D. Kunin, J. Sagastuy-Brena, S. Ganguli, D. Yamins and H. Tanaka, Neural Mechanics: Symmetry and Broken Conservation Laws in Deep Learning Dynamics, ICLR (2021).

Smr 4084 - 6th Youth in High-Dimensions

How creative is Generative AI? Perspectives from Science and Philosophy 3-5 July 2025

Martha Kunicki, Ph.D.

Princeton University

Can GenAI be creative? AI-Human Co-Creativity: Enriching the Concept of Creativity in Light of Emerging GenAI Towards Collective Co-Creativity of AI and Humans

This paper explores how our concept of creativity has to be rethought in light of emerging GenAI in creative practice, thereby leading to an enrichment of the concept of creativity to include collective co-creativity produced by humans and GenAI. Drawing on my experience as a theater director and producer involved in collective creativity and as a philosopher trained in analytical thinking, I explore case studies of co-created art, music, holograms, and text, in theater, and demonstrate how GenAI can be creative and how it can enhance our creativity.

Working together with GenAI can *enhance* creativity because digital technologies can expand the capacities of human creativity. I illustrate this by taking case studies of co-creativity in art, music, text, and holograms in theater, with examples of co-creations in art includinng *DALL·E*, co-creations in music including *AIVA*, and co-creations in text including *ChatGPT.* The most exciting recent examples of co-creations occur in immersive theater experiences, which integrate multisensory performances.

Collective creativity arises from the collective practices of people in relation to their nature, environment, cultures, societies, techniques, technologies, and GenAI. By expanding our concept of creativity towards co-creativity, we are able to acknowledge the processes and products of humans and GenAI as creative.

Creativity is standardly defined as the ability to produce something *new* and *valuable.* Various conceptions of creativity have been distinguished in the literature, such as *combinational*, *explorational*, and *transformational* creativity. I critically examine the various characterizations and conceptions, and offer a new conception of *integrational* creativity, which brings together the various conceptions of creativity. On this new conception, creativity is the creation of something (i) *new*, (ii) *integrationally significant* in the relevant context, and (iii) that *resonates*. A key feature of this new conception is that it allows for co-creativity of humans and GenAI.

GenAI has the potential to revolutionize the way we think about creativity and how we experience creativity. Today’s emerging GenAI technologies, with ground-breaking new tools, open up greater possibilities for creativity, with richer forms of experiencing creativity arising from the collective creativity of GenAI and humans. GenAI is becoming an increasingly influential (f)actor in the creative process.

P26

AI and the problem of creative spontaneity

Đorđe Lazarević

Faculty of Philosophy at the Univeristy of Niš, Serbia

Recent advances in "creative technology," particularly AI-generated images, poetry, and scientific ideas, have led to questions about whether machines might one day surpass human creative abilities—or whether they already have. However, to properly assess this and related questions, it is crucial to first establish a clear conceptual framework for creativity itself. In this talk, I pursue two main objectives: first, I propose an analysis of creativity by distinguishing three key components—product, process, and agent. Second, I brefly will explore the relevance of agential and production aspects for AI, and then focus on a more pressing question: can a machine be spontaneous? Since spontaneity is a necessary condition for a creative cognitive process, even if a machine has a conscious desire to create, its presumed lack of spontaneity would imply a lack of creativity. One common argument holds that machines lack spontaneity because they are always following instructions and thus lack intrinsic motivation and creative freedom.

I challenge this argument in two ways. First, both thought experiments and empirical studies confirm that human creative actions can be determined and predicted, which undermines the idea that spontaneity requires indeterminacy or unpredictability. A stronger argument, however, holds that AI agents are always causally dependent on humans, whereas humans, at least sometimes, initiate creative processes independently. Here, I suggest that the key distinction lies in how we interpret human *versus* machine actions: we attribute creativity to humans partly because we lack full knowledge of the causal factors underlying our own creative cognition, whereas we assume a fully determined model of action for machines. A further implication of this view is that a superintelligent AI—if it possessed complete knowledge of its own creative cognitive architecture—could not be creative, since it would lack the very epistemic opacity that underlies spontaneity in human creativity, even thought such AI would create most valuable ideas and artifacts.

P27

Abstract template for 6th Summer School on Theory, Mechanisms and Hierarchical Modelling of Climate Dynamics: Artificial Intelligence and Climate Modelling **Activity**

Title Integrating Machine Learning inference to simulation and deduction in math modeling

Diego Liberati [1]

[1]*National Research Council of Italy*

Standard climate modeling is traditionally done via mathematical equations via the help of powerful tools helping to manage them, often needing High Performance Computing. When functional subsystems are known and measured in their specific effects, a simple Galerkin simulation, touching all possible configurations for those domains, even allows to discover unforeseen behaviours like in a different domain to forecast unknown mutants as in [1] for Sos1, then discovered. Integrating eXplainable AI in the form of understandable rules Machine Learning [2], one can gain knowledge from data in the predicative logic form if ... then ... else..., immediately integrable to the theoretical priors, summing pros of both inference and deduction. When problems are simpler, like discriminating Myeloid from Lymphoid Leukemias from multivariable microarrays genes expression, the above piece-wise affine hyperplanes orthogonal to the salient intervals of the salient variables becomes a simple hyperplane in the orthonormalized PCA space, thus allowing the (possibly iterated) cascade of k-means and PCA [3] to outperform [4], also evidencing a few discriminating salient genes among which one not yet known in this respect. The same approach have been more recently instrumental in confirming a path in a rare form of leukemia [5], whose few cases available needed our enhancement of their statistical power in order to really got evidence of the suspected and hypothesized said path. When not just one shot of data is available, but a movie of signals in time, a piecewise affine AutoRegression could feed forward identify hybrid dinamic-logical nonlinear hysteretic processes [6] without the need of ill-conditioned inversions and Tickonow regularization [7] as for instance in blood hormone concentration deconvolution [8] in order to resort to the nanometric unaccessible dynamics of pituitary secretion

[1] E. Sacco, M Farina, C Greco, S Lamperti, S Busti, L DeGioia, L Alberghina, *D.Liberati* and M. Vanoni, *Biotechnology advances* , *30 (1),* 154 (2012)
[2] M. Muselli and D.Liberati, *IEEE Trans KDE*, *14 (6),* 1258 (2002).
[3] S. Garatti, S. Bittanti, D. Liberati, A. Maffezzoli, Intelligent Data Analysis **17**, (2), 175 (2007)
[4] T. R. Golub, et al., Science 286, 531 (1999)
[5] S Grassi, S Palumbo, V Mariotti, D Liberati et al., Frontiers in Oncology **9**, 532 (2019)
[6] G Ferrari-Trecate, M Muselli, D Liberati, M Morari, Automatica **39** (2), 205 (2003)
[7] D. Liberati, Annals of biomedical engineering **37**, 1871 (2009)
[8] G De Nicolao, D Liberati et al., European journal of endocrinology **141** (3), 246 (1999)

P28

Metaphorical Masks: Capturing and Concealing LLM Creativity in Art

Jakub Mácha, Masaryk University

This paper examines how metaphorical language shapes perceptions of large language models (LLMs) in artistic practice, focusing on whether these framings accurately express or conceal LLMs' creativity. Adopting a non-cognitivist approach, where metaphors invite "seeing as" without fixed cognitive content, we analyze five metaphors—artist, collaborator, muse, medium, and tool—and their interplay in defining LLMs' artistic roles. The "artist" metaphor emphasizes LLMs' creative outputs, like generating surreal poems, implying human-like intent, but risks anthropomorphism that obscures their computational nature, as Bender et al.'s "stochastic parrots" critique (2021). The "tool" metaphor highlights human control, framing LLMs as predictable instruments but concealing their unpredictable, recalcitrant creativity, comparable to gardening's dynamic medium (Young & Terrone, 2025). The "collaborator" and "muse" metaphors balance agency and inspiration, expressing LLMs' contributions while preserving human responsibility (Anscomb, 2024), yet human-like framings may overstate autonomy, masking algorithmic limits. The "medium" metaphor captures LLMs' unpredictable outputs, enhancing creativity but constrained by "tool's" reductive lens. Against the master metaphor of "artificial intelligence," which portrays machines as figuratively intelligent, these metaphors interact—complementing to deepen perspectives, constraining when human-centric views obscure LLMs' unique generative potential. Anthropomorphic metaphors like "artist" or "collaborator" may conceal LLMs' distinct creativity, rooted in statistical patterns, while "medium" and "tool" clarify their non-human agency. Drawing on Anscomb's responsibility arguments, this study shows that metaphorical framings navigate the artworld's social validation, expressing LLMs' creativity when acknowledging their computational essence, but concealing it when overly humanized. This interplay positions AI art as a collaborative process of human intent and machine dynamics, offering nuanced insights into LLMs' artistic contributions.

P29

From the Human/Machine Dichotomy to Distributed Creativity The Case of AI Integration in Filmmaking

Pierluigi Masai[1], **Lorenzo Carta**[1], **and Mateusz Miroslaw Lis**[2]

[1]*Università degli Studi di Trieste*
[2]*SophIA - Artificial Intelligence Audiovisual Lab*

The recent development and diffusion of generative AI systems have reinvigorated debates about machine creativity, particularly regarding its comparison to and potential surpassing of human creative abilities in the autonomous production of artistic work. Yet, the very notion of creativity is ambiguous and to attribute creativity to a machine could be misleading. Conventional approaches often conceptualize creativity as an isolated phenomena, traceable in the internal functioning of an individual system—whether human or artificial—while disregarding its embeddedness within material and social contexts. By treating human or technological creativity as self-contained processes, such analysis overlooks the dynamic entanglement of artifacts, practices, and socio-economic arrangements that give rise to artistic products [1].

To advance our argument, we will focus on audiovisual production—specifically, Filmmaking—as a key case study. Following Vlad Glăveanu's perspective on creativity [2], audio-visual work is not an individual act, but a collaborative and distributed process that involves aspects both technical (lenses, tripods, sensors, lighting) and human (the experience and sensitivity of cinematographers, focus pullers or camera operators). In line with a Science and Technology Studies (STS) approach, creativity here arises from the dynamic interaction between the materiality of technology, subjective interpretations, and broader social and economic environments. In this framework, AI tools represent a technological shift that can reshape existing creative practices—both affording and constraining aesthetic and working possibilities.

With this theoretical framework in mind, we will provide an overview of some AI techniques and show how they might be useful to overcome common problems met in film shootings and editing with an impact on film aesthetics. Ultimately, we argue that AI's impacts on creative processes demand examination beyond technical functioning, requiring instead a situated analysis of how these technologies may become part of everyday creative practices.

References

[1] C. Celis Bueno, P.-S. Chow, and A. Popowicz. Not "what", but "where is creativity?": towards a relational-materialist approach to generative AI. *AI & SOCIETY*, Mar. 2024.

[2] V. P. Glăveanu. *Distributed Creativity: Thinking Outside the Box of the Creative Individual*. Springer, Cham, 2014 edition, Apr. 2014.

P30

What's in a caption? Finding semantics with the information imbalance.

Andrea Mascaretti[1], **Santiago Acevedo**[1], **Marco Baroni**[2], **Alessandro Laio**[1], **Matéo Mahaut**[2], **and Riccardo Rende**[1]

[1]*(Presenting author underlined) Scuola Internazionale Studi Superiori Avanzati*
[2]*Universitat Pompeu Fabra*

Size is the driving factor behind the success of transformers. Bigger models produce higher-dimensional representations, encode more information, and achieve better performance; examples of this trend are large language models (LLMs) and visual transformers (ViTs). Size also drives convergence: [1] conjecture that representations are effective because they mimic reality. Transformers produce representations that (i) are very high dimensional, and (ii) possess highly-nonlinear relations across layers. Comparing the hidden representations from different transformers is a valuable tool to localise and contrast these semantic representations; and require methods that (i) scale with size, (ii) capture local behaviours, and (iii) allow to quantify asymmetries between different models. Linear methods are fast, but do not capture local behaviours; local methods, that count the number of common neighbours, are symmetric, and do not allow to quantify the partial ordering induced by the scale effects of transformers. We propose a new method based on the information imbalance: a local and asymmetric measure of topological similarity. We employ the method to study the hidden representation of DeepSeekV3, a state-of-the-art LLM, and study the relation between hidden representations and semantics in (i) a text-text scenario, in which we contrast sentences in different languages; (ii) a text-image scenario, in which we compare images and their captions. The method allows to investigate the data employing an arbitrary number of tokens and bigger sample sizes. We find that (i) alignment is maximal at semantic layers, regardless of the specific transformer or domain; (ii) the auto-correlation of tokens is higher in the semantic regions of transformers; (iii) the image-text transformer have asymmetric information imbalances. The findings support the idea of a representational convergence; but also highlights the importance of building tools that account for the size of the representations and the uneven distribution of information across layers and tokens.

[1] Huh, Minyoung, et al. "The platonic representation hypothesis." arXiv preprint arXiv:2405.07987 (2024).

Weak-to-Strong Generalization Even in Random Feature Networks, Provably.

Marko Medvedev[1], Kaifeng Lyu[2], Dingli Yu[3], Sanjeev Arora[4], Zhiyuan Li[5], Nathan Srebro[5]

[1]*The Unviersity of Chicago,* [2]*Simons Institute, University of California,* [3]*Microsoft Research,* [4]*Princeton University,* [5]*TTIC*

Weak-to-Strong Generalization (Burns et al., 2023) is the phenomenon whereby a strong student, say GPT-4, learns a task from a weak teacher, say GPT-2, and ends up significantly outperforming the teacher. We show that this phenomenon does not require a strong learner like GPT-4. We consider student and teacher that are random feature models, described by two-layer networks with a random and fixed bottom layer and trained top layer. A 'weak' teacher, with a small number of units (i.e. random features), is trained on the population, and a 'strong' student, with a much larger number of units (i.e. random features), is trained only on labels generated by the weak teacher. We demonstrate, prove and understand, how the student can outperform the teacher, even though trained only on data labeled by the teacher. We also explain how such weak-to-strong generalization is enabled by early stopping. Importantly, we also show the quantitative limits of weak-to-strong generalization in this model.

This work will appear as a poster in ICML 2025.

[1] Burns, C., Izmailov, P., Kirchner, J. H., Baker, B., Gao, L., Aschenbrenner, L., Chen, Y., Ecoffet, A., Joglekar, M., Leike, J., Sutskever, I., and Wu, J. Weak-to-strong generalization: Eliciting strong capabilities with weak supervision. In *Forty-first International Conference on Machine Learning*, 2024. URL `https://openreview.net/forum?id=ghNRg2mEgN`.

Restricted Boltzmann Machine Approaches Lattice Systems

Pedro Henrique Mendes[1] and Heitor C. M. Fernandes[1]

[1]*Federal University of Rio Grande do Sul*

In this work we present RBM applications in lattice systems, specially the Ising and Lattice Gas Models. Our goal is to analyze snapshots configurations and get insights about it. It was possible to make our machine learning models learn the intrinsic behavior of our statistical systems. We could, after the learning process, reproduce system configurations with physical observables in accordance with our input data, obtained from importance sampling. Also we could compare the results of our models from the well established simulation methods. We could have a better understanding of how machine learning models "learn" physics. In the literature of statistical physics of machine learning this type of work can be found [1, 2, 3].

[1] Torlai, G., Melko, R., "Learning thermodynamics with Boltzmann machines". Phys. Rev. B 94, 165134 (2016).

[2] Carrasquilla, J., Melko, R., "Machine learning phases of matter". Nature Phys 13, 431–434 (2017).

[3] Morningstar, A., Melko, R., "Deep Learning the Ising Model Near Criticality". Journal of Machine Learning Research 18 (2017).

Understanding Geometric Compression in Neural Network Dynamics

Federico Milanesio[1], Marco Gherardi[2], and Matteo Osella[1]

[1]*(Presenting author underlined) Department of Physics, University of Turin & INFN, Turin*
[2]*Department of Physics, University of Milan*

Neural networks are the most used algorithm in modern machine learning and have achieved incredible performance on different tasks. However, their development relies not on a deep theoretical understanding but on trial and error. One of the main difficulties of understanding NNs lies in the problematic nature of their training dynamics, which seeks optima in a very high-dimensional rough landscape. At the same time, the models avoid becoming trapped in suboptimal minima and converge to points with good generalization, thus escaping the so-called curse of dimensionality.

Intuitively, in regression tasks, the optimal representation of the data would be a low-dimensional manifold in which the points align, allowing for easy linear regression. We introduce a PCA-based measure to capture geometric compression and investigate this prediction. Our observations during training reveal a non-monotonic behavior, namely a first compression phase, followed by a subsequent decompression phase where our measure increases again. Such a result is remarkable and unobserved in regression NNs but aligns with previous results on classification tasks [**?**]. We prove that this behavior is a property of feature learning and is quite general for changes in hyperparameters and different datasets. We show that the epoch of inversion always happens after the network has learned to predict the best linear regression possible, and the subsequent decompression phase may indicate a phase of generalization, in which the model becomes more flexible and needs representations decompressed to accommodate more complex functions. This behavior aligns with current literature suggesting that neural networks learn the data distribution's moments sequentially [2]. We theorize that inversion happens when the network has learned the first two moments and starts learning from moments of higher order.

[1] S. Ciceri, L. Cassani, M. Osella, et al., Nature Machine Intelligence, **6(1)**, 40–47 (2024).
[2] M. Refinetti, A. Ingrosso, S. Goldt, PMLR, **202**, 28843 (2023).

The Myth of the Creativity Dial: Temperature in LLMs

Federico Milanesio[1], **Filippo Valle**[1], **Matteo Osella**[1]

[1]*(Presenting author underlined) Department of Physics, University of Turin & INFN, Turin*

Large Language Models (LLMs) have emerged as the frontier tool in natural language processing. Even compared to other promising fields in AI, the performance of LMMs in tasks such as language generation, translation, and sentiment analysis remains a forefront example of machine learning possibilities to revolutionize fields both within academia and in the job market.

LLMs generate text by predicting the next token (word or subword unit) by assigning a raw score to each potential token in the model vocabulary based on prior text. These scores are then converted into probabilities using the softmax function, which is regulated by a parameter called *temperature*. This temperature controls the randomness of the output, effectively influencing the degree to which the generated text is predictable. A lower temperature makes the model more deterministic, favoring the most likely next word based on its training, resulting in more conservative responses. In contrast, a higher temperature increases randomness by allowing less probable words to be selected more frequently.

Temperature is often referred to as a *creativity parameter*. Although recent scientific literature has begun to challenge this characterization [1], it remains the de facto understanding of practitioners. This aligns with the guidance given by model developers, who recommend specific temperature settings for different use cases. Temperature is also linked to some poorly understood behaviors of LLMs. Among these are the apparent phase transition-like behavior in the models' output as the temperature changes and the tendency of models to generate looping and repetitive outputs for very low temperatures. Although the latest language models solved these problems through architectural and post-processing techniques [2], the fundamental causes remain poorly understood.

In this talk, we will examine the current understanding of the temperature parameter, discussing its practical role in controlling output diversity and its contested status as a measure of creativity. Finally, we will present new results exploring the phase transition-like behaviors observed in LLM outputs, proposing simple models to shed some light on these problems.

[1] M. Peeperkorn, T. Kouwenhoven, D. Brown, A. Jordanous, in Proceedings of the 15th International Conference on Computational Creativity, 226-235 (2024).

[2] A. Holtman, J. Buys, L. Du, et al., in International Conference on Learning Representations (2020).

P34

Signal extraction in high-dimensional data and its kinetic interpretation

Yoh-ichi Mototake [1] **, Y-h. Taguchi** [2]

[1] Graduate School of Social Data Science, Hitotsubashi University, Tokyo 186-8601, Japan
[2] Department of Physics, Chuo University, Tokyo 112-8551, Japan

The detection of a signal variable from multiple variables that contain many noise variables is often approached as a variable selection problem under a given objective variable. This is nothing more than building a supervised model of a signal by specifying the signal as the objective variable. On the other hand, such a supervised model does not work effectively under high-dimensional and small-sample-size conditions, as the estimation of model parameters becomes indeterminate. We propose an ``unsupervised signal model'' that enables signal detection under high-dimensional and small-sample conditions without external signal definitions. The proposed unsupervised signal model is based on the assumption that the datasets in this world are generated from some dynamical system, and variables generated from dynamical systems with small correlation lengths are considered noisy variables. That is, the variables that maintain the data structure generated from a dynamical system under high-dimensional and small-sample conditions, corresponding to the limit of a sample size of 0, are modelled as always signal variables. In this study, we showed that with such a signal model, the Taguchi method provides an effective way of detecting signals. The proposed signal model was validated by generating a dataset with a globally coupled map (GCM) system, which is a high-dimensional dynamical system. Furthermore, we validated the model with a Gene expression data, which are not explicitly generated from a dynamical system; as a result, we observed a signal structure consistent with that of the signal model proposed in this study. The results suggest that the proposed signal model is valid for a wide range of datasets.

[1] Yoh-ichi Mototake, Y-h. Taguchi, arXiv:2304.06522, (2023).

Abstract template for Youth in High Dimensions

Flavio Nicoletti[1,2], **Matteo Negri**[2], **and Francesco D'Amico**2

[1]Department of Computer Science and Engineering Chalmers University of Technology and University of Gothenburg SE-41296 Gothenburg, Sweden
[2]Dipartimento di Fisica, Sapienza Università di Roma, 00185 Rome, Italy

We study analytically and numerically a fully-connected Hopfield network with vector spins. These networks are models of associative memory that generalize the standard Hopfield model [1] with ± 1 spins to neurons represented by vectors of fixed norm.
Our motivations to study vector Hopfield networks are twofold: first, we are curious to study a network that stores continuous bounded data; second, it was recently noted [2, 3] that Attention updates in Transformers [4] can be thought as the state evolution of a collective of vector spins.
First, we show that for any finite spin dimension d these models can store a number of memories linear in the size of the system, up to a certain critical capacity $\alpha_c = P_c/N$; in the large d limit linear storage is lost.
Second, we study numerically the retrieval dynamics of the network below ($\alpha < \alpha_c$) and above saturation ($\alpha > \alpha_c$): we show that even when memories are not stored the network can perform denoising of the memory in the very first steps of the dynamics.
Finally, we consider the linear stability of memory-correlated local energy minima below saturation, by studying analytically and numerically the spectrum of the Hessian matrix of the energy function associated to our model. We find that Matthis states present rare localised excitations close to the lower spectral edge, representing neurons receiving a very low signal from the remainder of the system and that are weakly correlated to memory spins.

[1] John J Hopfield, *Neural networks and physical systems with emergent collective computational abilities*, *Proceedings of the national academy of sciences* **79.8** (1982), pp. 2554–2558.

[2] Matthias Bal, *Spin-Model Transformers*, *url: https : / / mcbal.github.io/post/spin-model-transformers*, (2023).

[3] Francesco D'Amico and Matteo Negri, *Self-attention as an attractor network: transient memories without backpropagation*, *2024 IEEE Workshop on Complexity in Engineering (COMPENG)*, (2024), pp. 1–6.

[4] Ashish Vaswani et al, *Attention is all you need*, *Advances in neural information processing systems*, **30**, (2017).

P36

Aging, Glioblastoma, and Alzheimer’s disease in the space of gene expression

Nieves J., Gil G. and Gonzalez A.

(Presenting author underlined) Institute of Cybernetics, Mathematics and Physics

Gene expression profiling has become a powerful tool for unraveling the molecular mechanisms underlying complex diseases such as Glioblastoma and Alzheimer’s disease. By applying simple computational methods, we identify pathways associated with critical transitions, including carcinogenesis and the onset of Alzheimer’s disease. Moreover, we uncover a predefined aging trajectory that aligns with the hypothesis of programmatic aging. Utilizing assumptions about the strengths of attractors, we constructed a schematic fitness landscape with Wright’s diagrams, which effectively illustrate established relationships among aging, Glioblastoma, and Alzheimer’s disease.

P37

Mechanisms in Neural Scaling Regimes

Konstantin Nikolaou[1] , **Sven Krippendorf**[2] , **Samuel Tovey**[3] **and Christian Holm**[1]

[1]*(Presenting author underlined) University of Stuttgart*
[2] *University of Cambridge*
[3] *Quantagonia*

Scaling laws offer valuable insights into the relationship between neural network performance and computational cost, yet their underlying mechanisms remain poorly understood. In this work, we empirically analyze neural network behavior under data and model scaling through the lens of the neural tangent kernel (NTK), establishing a link between performance scaling and the internal dynamics of neural networks. Our findings on standard vision tasks reveal that similar performance scaling exponents can arise from fundamentally distinct internal mechanisms within the model, which highlights the critical need for a comprehensive mechanistic understanding. In this context, a key unresolved aspect is how the limiting NTK influences or breaks down neural scaling laws. To this end, we investigate the loss of feature learning as models scale towards infinite width and quantify the transition between kernel-driven and feature-driven scaling regimes. We identify the maximum model size that supports feature-driven learning, which provides a foundation for quantifying feature learning regimes and their impact on scaling in AI applications.

P38

Theoretical Analysis of Generalization of a Two Layer Neural Networks with Generic Activation: Bayes Optimal Inference and Empirical Risk Minimization

May 8 2025

Abstract In this study, we present a comprehensive theoretical analysis of the typical learning performance of a one-hidden-layer neural network—commonly known as a committee machine—in the high-dimensional regime where the number of hidden units K satisfies 1 N, with N denoting the input dimension. Focusing on independently and identically distributed (i.i.d.) inputs, we analyze both generalization and memorization scenarios within the classic teacher-student framework. Our main contribution is the derivation of closed-form expressions for the generalization error and learning curves via a quenched computation of the free entropy using the replica method, thereby surpassing prior annealed approximations. The analysis accommodates a broad class of activation and loss functions, making the framework highly adaptable. To validate our theoretical predictions, we perform high-dimensional simulations using Langevin dynamics, which show strong agreement with analytical results. Additionally, we compare empirical risk minimization with gradient descent and find that, with sufficient data, gradient descent closely matches the theoretical generalization error predicted by the replica method and Langevin dynamics in the zero-temperature limit. These results provide key insights into neural network learning behavior in high dimensions and highlight the power of statistical physics in understanding learning dynamics.

P39

Machine Learning model application of spectroscopic data analysis

A. Palavandishvili [1] , T. Bzhalava [1], and P. Kervalishvili [1]

[1]*Georgian Technical Univeristy*

Presented poster talk as part of the event *Youth in High Dimensions,* held from 7–9 July 2025, this research explores the application of machine learning (ML) models to the analysis of spectroscopic data in the context of plant virus detection. We utilize three spectroscopic techniques—Raman, infrared (IR), and ultraviolet-visible (UV-Vis)—to gather data from infected plant and grapevine samples. Each modality provides complementary biochemical insights, necessitating robust and flexible analysis pipelines.

Georgia, being an agriculturally rich country, faces significant challenges from plant viral diseases that impact crop yield and quality. Among these, mosaic viruses are of particular concern due to their widespread prevalence and detrimental effects on various crops. This study is especially focused on viral infections affecting both general mosaic-infected plants and grapevines—a key agricultural product with high economic and cultural value in the region. Rapid and accurate detection of such viruses is essential for effective disease management and sustainable viticulture.

Preprocessing of spectral data involves Fast Fourier Transform (FFT), baseline fitting, and denoising to enhance signal quality. Different studies shows that for each spectrometer different ML model is applyed. For IR spectra, Partial Least Squares Regression (PLSR), Support Vector Machines (SVM) are to predict and classify spectral patterns. Raman spectra, which are highly informative yet complex, are analyzed using deep learning models including Convolutional Neural Networks (CNNs), Long Short-Term Memory (LSTM) networks. UV-Vis spectroscopy is processed using PLSR, Ridge Regression, and Artificial Neural Networks (ANNs).

We work with the full spectra obtained from each spectroscopy method to capture a wide range of biochemical signatures associated with viral infection. This holistic approach enables deeper analysis of virus-specific markers and subtle spectral differences across plant types. Our aim is to build a comprehensive framework for the spectral analysis of plant viruses, which can contribute to long-term research efforts in virology and precision for Georgian agriculture. The integration of machine learning enhances not only speed and accuracy, but also the potential for discovering novel insights into plant-virus interactions through data-driven exploration.

This work was supported by Shota Rustaveli National Science Foundation of Georgia (SRNSFG)[FR-23-4069]

[1] L.Mandrile, C.D'Errico, F.Nuzzo, G. Barzan, S. Matić, A. M. Giovannozzi, A.M. Rossi, G.Gambino, E. Noris. Raman Spectroscopy Applications in Grapevine: Metabolic Analysis of Plants Infected by Two Different Viruses, Front. Plant Sci. 13:917226 (2022).

[2] S. Yan, S. Wang, J. Qiu, M. Li, D. Li, D. Xu, D. Li, Q. Liu, Raman spectroscopy combined with machine learning for rapid detection of food-borne pathogens at the single-cell level, Talanta, Volume 226, 2021, Page 122195, https://www.sciencedirect.com/science/article/abs/pii/S0039914021001168?via%3Dihub

[3] Jones, R.A.C. Plants Global Plant Virus Disease Pandemics and Epidemics 2021, 10, 233. https://doi.org/10.3390/plants10020233

P40

Multi-Class Uncertainty-aware Brain Tumor Segmentation using Bayesian Attention U-Net

Rahul Pal[1] , **Sanoj Kumar**[2] , **and Gaurav Bhatnagar**[3]

[1] *Department of Mathematics, UPES, Dehradun, Uttarakhand, 248007, India*
[2] *School of Computer Science, UPES, Dehradun, Uttarakhand, 248007, India,*
[3]*Department of Mathematics, IIT Jodhpur, India*

Abstract-

Accurate and reliable segmentation of brain tumors in MRI scans is critical for diagnosis, treatment planning, and surgical guidance. While prior works have focused on binary tumor segmentation with uncertainty quantification, we extend this paradigm by introducing BA U-Net, a novel Bayesian Attention U-Net architecture designed for multi-class segmentation of brain tumor sub-regions—necrotic core, edema, and enhancing tumor with class-specific uncertainty estimation. Our model integrates Bayesian neural network principles with spatial attention mechanisms to provide not only precise segmentation but also interpretable uncertainty maps for each tumor class. Evaluated on the BraTS 2020 dataset, BA U-Net achieves a Dice score of - and an IoU of - across all classes. More importantly, our model exhibits meaningful uncertainty responses to image degradation and ambiguous tumor boundaries, offering valuable confidence cues for clinicians. By extending uncertainty quantification to each anatomical sub-region, BA U-Net paves the way for more trustworthy and robust AI-assisted diagnostics in clinical MRI workflows.

P41

Transfer Learning with Feature Augmentation Enabled by Neural ODE and Path Signature for Class-unbalanced Mechanical Fault Diagnosis

Bin Pei[1] , Lixin Guo[1] , Xiaohui Gu[2], Xiaolong Wang[3], Yongge Li[1] and Yong Xu[1]

[1]School of Mathematics and Statistics, Northwestern Polytechnical University, 127 Youyi West Road, Xi'an, 710072, Shanxi, PR China

[2] Affiliation 2State Key Laboratory of Mechanical Behavior and System Safety of Traffic Engineering Structures, Shijiazhuang Tiedao University, No. 17, North Second Ring East Road, Qiaodong District, Shijiazhuang, 050047, Hebei, PR China

[3]School of Mathematics and Statistics, Shaanxi Normal University, No. 620 West Chang'an Avenue, Chang'an District, Xi'an, 710119, Shanxi, PR China

In order to tackle the problems of precisely identifying fault modes and accomplishing cross-condition diagnosis using limited data under the background of imbalanced rolling bearing data, this paper comprehensively integrates transfer learning techniques and feature augmentation strategies. From the perspectives of temporal and spatial features, it presents two feature enhancement modules: the Neural Ordinary Differential Equation-based Temporal Feature Augmentation Module (NODE-TFAM) and the Path Signature-based Spatial Feature Augmentation Module (PS-SFAM). The NODE-TFAM utilizes the local features extracted by the Convolutional Neural Network (CNN) as initial values, and then models the feature space as an ODE equation system. By solving this system, the module can capture the temporal evolution of convolutional features. On the other hand, the PS-SFAM regards each vibration signal as a mathematical path. Through calculating its iterative integral, the module acquires a spatial feature representation of the entire vibration signal. This feature has a clear geometric meaning in mathematical terms, further enhancing the interpretability of the model. The validity of the proposed methodology has been meticulously verified on both the Paderborn University bearing dataset and the laboratory-acquired bearing dataset. These experiments comprehensively probed into the performance of the methodology, manifesting its potential to surmount the challenges in the domain of rolling-bearing fault diagnosis and furnishing robust evidence for its practical implementation in real-world mechanical maintenance scenarios.

P42

Can Machines Philosophize?

Michele Pizzochero [1,2] **& Giorgia Dellaferrera** [3]

[1] *Department of Physics, University of Bath, United Kingdom*
[2] *School of Engineering and Applied Sciences, Harvard University, United States*
[3] *McKinsey & Co, United Kingdom*

Inspired by the Turing test, we present a novel methodological framework to assess the extent to which a population of machines mirrors the philosophical views of a population of humans. The framework consists of three steps: (i) instructing machines to impersonate each human in the population, reflecting their backgrounds and beliefs, (ii) administering a questionnaire covering various philosophical positions to both humans and machines, and (iii) statistically analyzing the resulting responses. We apply this methodology to the debate on scientific realism, a long-standing philosophical inquiry exploring the relationship between science and reality. By considering the outcome of a survey of over 500 human participants, including both physicists and philosophers of science, we generate their machine personas using an artificial intelligence engine based on a large-language generative model. We reveal that the philosophical views of a population of machines are, on average, similar to those endorsed by a population of humans, irrespective of whether they are physicists or philosophers of science. As compared to humans, however, machines exhibit a weaker inclination toward scientific realism and a stronger coherence in their philosophical positions. Given the observed similarities between the populations of humans and machines, this methodological framework may offer unprecedented opportunities for advancing research in experimental philosophy by replacing human participants with their machine-impersonated counterparts, possibly mitigating the efficiency and reproducibility issues that affect survey-based empirical studies.

P43

Generative AI: collaborator or tool? The experience of interaction in the process of artistic creation

Iulia-Dana Pușcașu[1]

[1]*Babeș-Bolyai University, Cluj-Napoca (Romania)*

This study begins by analysing the ambiguous use of the word 'collaboration' in relation to generative AI in artistic creation. Identifying some of the main contexts in which the term is used reveals the concept of interaction to be of key interest, since in all cases, at a minimum, collaboration implies a form of interaction between a human and an AI. While I do not take a position on whether it is appropriate to designate any human-AI interaction as collaboration, I aim to show that a sense of collaboration emerges from the experience of interaction. Broadly defined as "mutual or reciprocal action or influence," the concept of interaction we are interested in here covers human–non-human exchange processes. Following an intentionalist approach to art [1], my claim is that, in the context of production and creation—and in light of current technological developments—the interaction between an artist and an AI can fundamentally be considered the same as the interaction between an artist and their regular tools, but that it is the experience of interacting with AI systems that differs. The first part of my claim is based on the fact that generative AI systems appear in the continuation of a long line of technologies that have enabled automatic image-making techniques [2], and which, in retrospect, can be said to constitute an art historical tradition. Furthermore, while AI systems produce formal features that could be considered artistically relevant, they do not exercise knowledge-how or evaluative abilities in producing these features as such, and therefore do not qualify as artistically creative agents [3]. The second part of my claim is explained, first, by the ongoing efforts to simulate interpersonal communication in the modelling of interfaces for human–machine interaction [4]. Second—and this is my focus—is the tendency to infer decision-making capabilities and intention from the system's manner of responsiveness through misdirected interpretation, i.e., as if a generative AI were operating on a semantic level. This issue also concerns skill, and the difficulty of scaling both the functioning and the possibilities that sophisticated tools like generative AI systems offer.

[1] A. Danto, Jaac, **33**, 139 (1974).

[2] K. Wojtkiewicz, Jaac, **81**, 454 (2023).

[3] C. Anscomb, Odradek, **8**, 13 (2022).

[4] S. Schleidgen, O. Friedrich, S. Gerlek, et al., Humanit. Soc. Sci. Commun. **10**, 551 (2023).

P44

Abstract for Youth in High Dimensions

F. Ricci, L. Bardone, and S. Goldt

[1]*International School of Advanced Studies (SISSA), Trieste, Italy*

Deep neural networks learn structured features from complex, non-Gaussian inputs, but the mechanisms behind this process remain poorly understood. Our work is motivated by the observation that the first-layer filters learnt by deep convolutional neural networks from natural images resemble those learnt by independent component analysis (ICA), a simple unsupervised method that seeks the most non-Gaussian projections of its inputs. This similarity suggests that ICA provides a simple, yet principled model for studying feature learning. Here, we leverage this connection to investigate the interplay between data structure and optimisation in feature learning for the most popular ICA algorithm, FastICA, and stochastic gradient descent (SGD), which is used to train deep networks. We rigorously establish that FastICA requires at least $n \gtrsim d^4$ samples to recover a single non-Gaussian direction from d-dimensional inputs on a simple synthetic data model. We show that vanilla online SGD outperforms FastICA, and prove that the optimal sample complexity $n \gtrsim d^2$ can be reached by smoothing the loss, albeit in a data-dependent way. We finally demonstrate the existence of a search phase for FastICA on ImageNet, and discuss how the strong non-Gaussianity of said images compensates for the poor sample complexity of FastICA.

P45

Local minima of the empirical risk in high-dimensions: general theorems and convex examples

Kiana Asgari[1], Andrea Montanari[1], and Basil Saeed[1]

[1]*Stanford University*

We consider a general model for high-dimensional empirical risk minimization:

$$\hat{\mathbf{\Theta}}_n \in \arg\min_{\mathbf{\Theta}\in\mathbb{R}^{d\times k}} \hat{R}_n(\mathbf{\Theta}), \qquad \hat{R}_n(\mathbf{\Theta}) := \frac{1}{n}\sum_{i=1}^{n} L(\boldsymbol{f}(\mathbf{\Theta}^{\mathsf{T}}\boldsymbol{x}_i), \boldsymbol{y}_i) + r(\mathbf{\Theta})$$

where the data $\boldsymbol{x}_i \in \mathbb{R}^d$ are Gaussian vectors, $\boldsymbol{y}_i \in \mathbb{R}^q$ with $\mathbb{P}(\boldsymbol{y}_i \in \mathcal{B}|\boldsymbol{x}_i) = g(\mathcal{B}|\mathbf{\Theta}_0^{\mathsf{T}}\boldsymbol{x}_i)$ for some $\mathbf{\Theta}_0 \in \mathbb{R}^{d\times k_0}$, the regularizer is separable, the model is parameterized by $\mathbf{\Theta} \in \mathbb{R}^{d\times k}$, and the loss depends on the data via the projection $\mathbf{\Theta}^{\mathsf{T}}\boldsymbol{x}_i$ potentially in a non-convex manner. This setting covers, as special cases, classical statistics methods (e.g. multinomial regression and other generalized linear models), but also two-layer fully connected neural networks with k hidden neurons.

We use the Kac–Rice formula from Gaussian process theory to derive a bound on the expected number of local minima of this empirical risk, under the proportional asymptotics in which $n, d \to \infty$, with $n \asymp d$. Namely, let $\hat{\mu}_{\sqrt{d}[\mathbf{\Theta},\mathbf{\Theta}_0]} \in \mathscr{P}(\mathbb{R}^{k+k_0})$, $\hat{\nu}_{[\boldsymbol{X}\mathbf{\Theta},\boldsymbol{y}]} \in \mathscr{P}(\mathbb{R}^{k+q})$ denote the empirical distributions of the rows of $\sqrt{d}[\mathbf{\Theta}, \mathbf{\Theta}_0]$, $[\boldsymbol{X}\mathbf{\Theta}, \boldsymbol{y}]$, respectively. With the definition

$$\mathscr{Z}_n(\mathscr{A}, \mathscr{B}) := \left\{ \text{local minima of } \hat{R}_n(\mathbf{\Theta}) \text{ s.t. } \hat{\mu}_{\sqrt{d}[\mathbf{\Theta},\mathbf{\Theta}_0]} \in \mathscr{A}, \hat{\nu}_{[\boldsymbol{X}\mathbf{\Theta},\boldsymbol{y}]} \in \mathscr{B} \right\},$$

for appropriate $\mathscr{A}, \mathscr{B}$, we obtain a bound of the form

$$\lim_{n,d\to\infty} \frac{1}{n} \log \mathbb{E}\left[|\mathscr{Z}_n(\mathscr{A}, \mathscr{B})|\right] \le - \inf_{\mu\in\mathscr{A},\nu\in\mathscr{B}} \Phi(\mu, \nu)$$

for some $\Phi : \mathscr{P}(\mathbb{R}^{k+k_0}) \times \mathscr{P}(\mathbb{R}^{k+q}) \to \mathbb{R}$.

Via Markov's inequality, this bound allows to determine the positions of these minimizers (with exponential deviation bounds) and hence derive sharp asymptotics on the estimation and prediction error: for a given test function $\psi : \mathbb{R}^k \times \mathbb{R}^{k_0} \to \mathbb{R}$ and any local minimizer $\hat{\mathbf{\Theta}}$, (with $\hat{\mathbf{\Theta}}_i$, $\mathbf{\Theta}_{0\,i}$ the i-th rows of $\hat{\mathbf{\Theta}}$, $\mathbf{\Theta}_0$), we have

$$\mathbb{P}\left\{ \frac{1}{d}\sum_{i=1}^{d} \psi\big(\sqrt{d}\hat{\mathbf{\Theta}}_i, \sqrt{d}\mathbf{\Theta}_{0\,i}\big) \in I \right\} \le \exp\left\{ -n \inf_{\mu\in\mathscr{A}(I,\psi),\nu} \Phi(\mu, \nu) + o(n) \right\},$$

where $\mathscr{A}(I, \psi) := \{\mu : \int \psi(\boldsymbol{t}, \boldsymbol{t}_0)\mu(\mathrm{d}\boldsymbol{t}, \mathrm{d}\boldsymbol{t}_0) \in I\}$.

We apply our characterization to convex losses, where high-dimensional asymptotics were not (in general) rigorously established for $k \ge 2$. In this setting, we show that Φ satisfies

$$\Phi(\mu, \nu) \ge 0 \quad \forall \mu, \nu \qquad \text{with} \qquad \Phi(\mu, \nu) = 0 \;\Leftrightarrow\; (\mu, \nu) = (\mu_\star, \nu_\star),$$

for some $\mu_\star, \nu_\star$. This implies that $\hat{\mu}_{\sqrt{d}[\mathbf{\Theta},\mathbf{\Theta}_0]} \overset{w}{\Rightarrow} \mu_\star$ and $\hat{\nu}_{[\boldsymbol{X}\mathbf{\Theta},\boldsymbol{y}]} \overset{w}{\Rightarrow} \nu_\star$. We derive a set of fixed point equations characterizing $\mu_\star, \nu_\star$, which describe the high-dimensional asymptotics of ERM, and further show that these fixed point equations correspond to the stationarity conditions of an infinite-dimensional, deterministic convex problem. This description unifies many of the ones previously obtained for ERM under different settings, and rigorously establishes it for some settings where no rigorous proof was available.

P46

Poster title : Can machine learning improve natural product-based antimalarial ?

M. V. SAO TEMGOUA[1]

[1] *University of Yaoundé I*

Natural products have constituted a valuable source of drugs for centuries. A considerable number of pharmaceutical drugs have been derived from or inspired by natural compounds, including those derived from plants, fungi, bacteria, and marine organisms[1, 2]. The application of deep learning techniques has proven to be an effective approach in the identification of natural product-based antimalarial drugs [3]. These sophisticated computational techniques can markedly expedite the identification and optimization of prospective drug candidates derived from natural sources [4]. By analyzing extensive datasets of known antimalarial compounds and their molecular properties, machine learning algorithms can predict the efficacy and safety of novel natural products against malaria parasites. A deep learning model is capable of extracting intricate features from chemical structures and biological data, thereby facilitating more precise predictions of drug-like properties and antimalarial activity. These approaches can also facilitate the identification of novel molecular scaffolds and the prediction of synergistic combinations of natural products for enhanced antimalarial effects. Moreover, machine learning can assist in optimizing extraction and purification processes for natural products, thereby improving the overall efficiency of drug discovery pipelines. As the integration of artificial intelligence in natural product research continues to evolve, it holds great promise for revolutionizing the development of effective and sustainable antimalarial treatments derived from nature.

[1] N. E. Thomford, D. Senthebane, A. Rowe, D. Munro, P. Seele, A. Maroyi, and K. Dzobo, International Journal of Molecular Sciences **19**, 10.3390/ijms19061578 (2018).

[2] F. Ntie-Kang, Physical Sciences Reviews **4**, 7, 7 (2019).

[3] S. Zheng, T. Zeng, C. Li, B. Chen, C. W. Coley, Y. Yang, and R. Wu, Nature Communications **13**, 3342, 3342 (2022).

[4] A. L. Chávez-Hernández and J. L. Medina-Franco, Artificial Intelligence in the Life Sciences **3**, 100066 (2023).

P47

Bilevel Optimization for Machine Learning

Pradeep Kumar Sharma

Affiliation: University of Delhi, New Delhi, India

The aim of this poster is to present the foundations of solving machine learning problems through bilevel optimization techniques. We provide a brief overview of the methods to solve large scale machine learning problems. We present existing techniques, current challenges and directions to overcome these challenges with future research directions.

[1]Luca Franceschi, Paolo Frasconi, Saverio Salzo, Riccardo Grazzi, Massimiliano Pontil. Bilevel Programming for Hyperparameter Optimization and Meta-Learning. Proceedings of the 35th International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018.

[2] Riccardo Grazzi, Luca Franceschi, Massimiliano Pontil and Saverio Salzo. On the Iteration Complexity of Hypergradient Computation. Proceedings of the 37th International Conference on Machine Learning, Online, PMLR 119, 2020.

P48

Fundamental limits of non-symmetric low-rank matrix estimation with structured noise

P49

Dissecting the Interplay of Attention Paths in a Statistical Mechanics Theory of Transformers

Lorenzo Tiberi [1,2] , **Francesca Mignacco** [3,4] , **Kazuki Iries** [1,2] , **and Haim Sompolinsky** [1,2,5]

[1] *Center for Brain Science, Harvard University, Cambridge, MA, USA*
[2] *Kempner Institute for the Study of Natural and Artificial Intelligence, Harvard University, Cambridge, MA, USA*
[3] *Graduate Center, City University of New York, NY, USA*
[4] *Joseph Henry Laboratories of Physics, Princeton University, NJ, USA*
[5] *Edmond and Lily Safra Center for Brain Sciences, Hebrew University of Jerusalem, Jerusalem, Israel*

Despite the remarkable success of Transformers across a range of state-of-the-art machine learning tasks, theoretical characterizations accounting for their impressive performance remain sparse. To enable tractable analysis, theoretical studies often need to rely on significant architectural simplifications, such as limiting models to a single attention head, a single layer, or both. Consequently, an important feature of Transformers remains to a large extent undercharacterized: the interplay between multiple attention heads across multiple layers.

We present recent work [1] addressing this theoretical gap. We consider a deep multi-head self-attention network, that is closely related to Transformers, yet analytically tractable. By making some simplifying assumptions, such as linearity in the value weights and fixed pre-trained query and key weights, our theory is able to characterize a network with an arbitrary number of layers and attention heads.

Using backpropagating kernel renormalization techniques [2], we develop a theory of Bayesian learning in the model, deriving exact equations for the network's predictor statistics under the finite-width thermodynamic limit, i.e. $N, P \to \infty$, $\alpha = P/N = \mathcal{O}(1)$, where N is the network width and P is the number of training examples. Importantly, we go beyond previous results obtained in the Gaussian process limit [3], in which the number $P = \mathcal{O}(1)$ of examples is too small for the network weights to develop data-driven structures.

Our theory relates the network's generalization performance to its learned interplay between "attention paths", defined as information pathways through different attention heads across layers. We show that the network's mean predictor is expressed as a sum of independent kernels, each one corresponding to a different pair of "attention paths". These kernels are weighted according to a "task-relevant path combination" mechanism that aligns the total kernel with the task labels, improving generalization performance.

The kernel weighting is performed by a path-by-path matrix predicted by our theory, which we are able to express as an empirically measurable function of the network weights. This allows for a qualitative transfer of our insights to more complex models; for example, one trained via gradient descent with no fixing of the query and key weights. As an illustration, we demonstrate an efficient size reduction of such a model, by pruning those attention heads that are deemed less relevant by our theory.

Experiments confirm our findings on synthetic and real-world sequence classification tasks.

[1] L. Tiberi, F. Mignacco, K. Irie, H. Sompolinsky, Adv. Neural Inf. Process. Syst. **37**, 72710(2024).
[2] Q. Li, H. Sompolinsky, Phys. Rev. X **11**, 031059 (2021).
[3] J. Hron, Y. Bahri, J. Sohl-Dickstein, R. Novak, Proc. Mach. Learn. Res. **119**, 4376 (2020).

P50

Spherical Boltzmann Machines

P51

EFFECT OF LABEL NOISE ON THE INFORMATION CONTENT OF NEURAL REPRESENTATIONS IN CLASSIFICATION NETWORKS

P52

Memorization and Generalization in Generative Diffusion

Raphaël Urfin[1], **Tony Bonnaire**[1], **Giulio Biroli** [1] **and Marc Mézard**[2]

[1] *Laboratoire de Physique de l'Ecole Normale Supérieure, ENS, Université PSL, CNRS, Sorbonne Université, Université Paris Cité, F-75005 Paris, France*

[2] *Department of Computing Sciences, Bocconi University, Milano, Italy*

Diffusion models are a class of generative models that have achieved state-of-the-art performance in a variety of tasks such as audio and image generation. In order to generate new samples from a target distribution, one has to learn the score function [1, 2]. This can be achieved by gradually adding noise to the data points and by leveraging neural network techniques. However an infinitely expressive neural network would learn the empirical score which memorizes the dataset i.e. it only generates data points used in the training [3]. This is not the case in practice where models managed to generalize i.e. they generate new samples from the target distribution, provided the dataset is large enough [4]. To answer this conundrum, we study the properties of the loss landscape and the training dynamics with numerical experiments and simple analytically tractable models.

[1] Song, Yang and Ermon, Stefano, *Generative Modeling by Estimating Gradients of the Data Distribution*, *Advances in Neural Information Processing Systems* (2019).

[2] Yang, Song and Jascha, Sohl-Dickstein and Diederik, P Kingma and Abhishek, Kumar and Stefano, Ermon and Ben, Poole, *Score-Based Generative Modeling through Stochastic Differential Equations*, *International Conference on Learning Representations* (2021).

[3] Biroli, Giulio and Bonnaire, Tony and de Bortoli, Valentin and Mézard, Marc, *Dynamical regimes of diffusion models*, *Nature Communications* (2024)

[4] Zahra Kadkhodaie and Florentin Guth and Eero P Simoncelli and Stéphane Mallat, *Generalization in diffusion models arises from geometry-adaptive harmonic representations*, *The Twelfth International Conference on Learning Representations*, (2024)

P53

The Rich and the Simple: On the Implicit Bias of Adam and SGD

Bhavya Vasudeva[1], **Jung Whan Lee**[1], **Vatsal Sharan**[1], **and Mahdi Soltanolkotabi**[1]

[1]*(Presenting author underlined) University of Southern California*

Adam is the de facto optimization algorithm for several deep learning applications, but an understanding of its implicit bias and how it differs from other algorithms, particularly standard first-order methods such as (stochastic) gradient descent (GD), remains limited. In practice, neural networks trained with SGD are known to exhibit simplicity bias — a tendency to find simple models. In contrast, we show that Adam is more resistant to such simplicity bias. To demystify this phenomenon, in this paper, we investigate the differences in the implicit biases of Adam and GD when training two-layer ReLU neural networks on a binary classification task involving synthetic data with Gaussian clusters. We find that GD exhibits a simplicity bias, resulting in a linear decision boundary with a suboptimal margin, whereas Adam leads to much richer and more diverse features, producing a nonlinear boundary that is closer to the Bayes optimal predictor. This richer decision boundary also allows Adam to achieve higher test accuracy both in-distribution and under certain distribution shifts. We theoretically prove these results by analyzing the population gradients. To corroborate our theoretical findings, we present empirical results showing that this property of Adam leads to superior generalization under distributional shifts across datasets where neural networks trained with SGD are known to show simplicity bias, as well as benchmark datasets for subgroup robustness.

P54

Interpreting and Steering Protein Language Models through Sparse Autoencoders

Edith Villegas Garcia [1,2] **and Alessio Ansuini** [2]

[1] *(Presenting author underlined) Universita degli Studi di Trieste*
[2] *AREA Science Park*

The rapid advancements in transformer-based language models have revolutionized natural language processing, yet understanding the internal mechanisms of these models remains a significant challenge. This paper explores the application of sparse autoencoders (SAE) to interpret the internal representations of protein language models, specifically focusing on the ESM-2 8M parameter model. By performing a statistical analysis on each latent component's relevance to distinct protein annotations, we identify potential interpretations linked to various protein characteristics, including transmembrane regions, binding sites, and specialized motifs. We then leverage these insights to guide sequence generation, shortlisting the relevant latent components that can steer the model towards desired targets such as zinc finger domains. This work contributes to the emerging field of mechanistic interpretability in biological sequence models, offering new perspectives on model steering for sequence design.

Phase analysis of Ising machines and their implications on optimization

<u>Shu Zhou</u>[1], K. Y. Michael Wong[1], Juntao Wang[1,2], David Shui Wing Hui[2], Daniel Ebler[2], Jie Sun[2]

[1]*(Presenting author underlined) Department of Physics, The Hong Kong University of Science and Technology, Hong Kong SAR, China.*
[2]*Theory Lab, Central Research Institute, 2012 Labs, Huawei Technologies Co. Ltd., Hong Kong SAR, China.*

Ising machines, dynamical systems designed to operate in a parallel and iterative manner, have emerged as a new paradigm for solving combinatorial optimization problems. Despite their potential for solving hard optimization problems in the class of quadratic unconstrained binary optimization, Ising machines lack a theoretical understanding of system dynamics and their relation to solution accuracy. This results in significant solution variability and often requires careful manual fine-tuning of system dynamics and parameters.

In this work, we elucidate how equilibrium spin distributions during Ising machine execution constitute a rich palette of phases and how solution quality depends on these phases. By applying the replica method to Ising machines optimizing the Sherrington-Kirkpatrick (SK) model, we discovered the phases shown in Fig.1. Analyzing the TAP equations, which provide the exact ground state solution of the SK model, we found that for Ising machines to produce optimal solutions it is important to locate the system in the region where two specific phases, the gapless phase and the binary phase, coexist, as shown in Fig.1(e). This insight aids the practical design of Ising machine dynamics and the choice and control of system parameters. Based on these insights, we propose a superior Ising machine, digCIM, which satisfies the TAP equations by introducing a digitization operation into the system dynamics. DigCIM achieves higher accuracy and requires only one control parameter. Our results show that digCIM achieves top performance on the QPLib-QUBO challenge, outperforming a number of commercial optimization solvers.

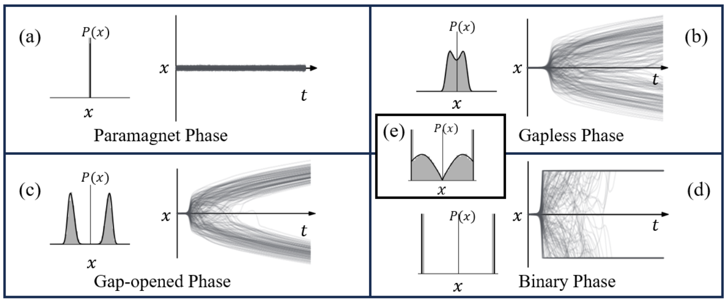


Figure 1: Ising machine phases