ICTP-INdAM-SLMath Summer Graduate School for Machine Learning | (SMR 4222)

15 Jun 2026 - 26 Jun 2026
ICTP, Trieste, Italy

P01 - ALADRAH Nicola

Implicit Bias as a Gauge Correction: Theory and Inverse Design

P02 - ANNESI Brandon Livio

Overparametrization bends the landscape: BBP transitions at initialization in simple Neural Networks

P03 - BADER Roger Michel

Time-Series Forecasting with Restricted Boltzmann Machines via Conditional Gibbs Sampling

P04 - BELAROUI Khaoula

Beyond Surface Similarities: Graph Neural Networks for More Natural Modeling in Recommender Systems

P05 - BERETTI Thomas

Recovering geometric features via the Fourier transform

P06 - BERGAMIN Eleonora

Memorization, convergence and generalization in first-order Markov chains

P07 - CESTOLA Pietro

Flow-aware training strategy for Physics-Informed Neural Networks

P08 - CRUZ Madelyn Esther Chua

Spatially-Informed Biologically Interpretable Machine Learning Approaches for Analyzing Neural Data

P09 - DELL'ORTO Marco

Algebraic issues arising in training neural networks

P10 - FLORIO Federico

Inference in continuous-space Markov processes on graphs

P11 - FOGLIANI Luigi Kentaro

Annealing in variational inference mitigates mode collapse: A theoretical study on Gaussian mixtures

P12 - FORSTATER Jonathan David

E(3)-Equivariant Fragment-based Graph Neural Networks for Biomolecules

P13 - FRAIMAN O'CONNELL Demian

Universal approximation of semantic couplings via Sinkhorn Transformers

P14 - FRATTEGIANI Luca

Machine Learning modeling of cancer dynamic

P15 - FUERTE PEREZ Angel Cuauhtemoc

Discrete Analytic Functions on a Rhombic Lattice

P16 - GIORLANDINO Alessio

Two Failure Modes of Deep Transformer Models and How to Avoid Them

P17 - GRECO Giacomo

A Malliavin-Gamma calculus approach to Score Based Diffusion Generative models for random fields

P18 - GUEVARA Ignacio

DYNAMICAL MEAN FIELD THEORY FOR MOMENTUM-BASED GRADIENT DESCENT IN DEEP NEURAL NETWORKS

P19 - HURINENKO Karyna

USING MACHINE LEARNING TO ANALYZE USER BEHAVIOR PATTERNS AND FORECAST REVENUE IN DIGITAL PRODUCTS

P20 - JANKOLA Marek

Dynamics of Performative Classification in High-Dimensional Gaussian Mixtures

P21 - KORPE Suhanee Sunil

High-Rank Matrix Denoising Beyond Rotational Invariance

P22 - LADIANA Andrea

Federated Associative-Memory Framework for Continual Learning via Spectral Inference

P23 - LAICHE Nabil

Estimation and Simulation in nonlinear time series models samples driven by applications on topological data analysis tools

P24 - LEPRE Andrea

Emergent Cooperativeness in Three-Directional Associative Memories: A Statistical Mechanics Perspective on Supervised and Unsupervised Learning.

P25 - LOPEZ AMADO Cristina

Understanding Oversmoothing in Causal Transformers: The Role of Attention Sparsity

P26 - LOPEZ Kimberly Andrea

The Geometry of Compositional Abstraction in Large Language Models

P27 - LOUP FOREST Clement

Theoretical insights on Post-Training in Generalised Linear Models

P28 - MAGRINI ALUNNO Stefano

Adaptive Collocation via Stochastic Solvers in PINNs

P29 - MAHARANA Piyush Ranjan

Can a Large Language Model do Chemistry?

P30 - MARCHETTI Alessandro

Bias Mitigation in Quantum Machine Learning: A Fairness Analysis of the COMPAS Dataset

P31 - MARTIN II Patrick Michael

Torsion Phenomena Hypergraph Chromatic Homology

P32 - MAZZUCCO Luca

Large Deviations for Transformer-like Models

P33 - MEHRAB Aneeqa

Sheaf Neural Networks and Biomedical Applications

P34 - MIRZAEI Erfan

Generalization of Gibbs and Langevin Monte Carlo Algorithms in the Interpolation Regime

P35 - MOEN Kristina Violet

Shape and Texture Analysis of Satellite Imagery for Weather Prediction

P36 - MOODY Merijn

Markov Random Field Information Bottleneck for Latent Conditional Diffusion

P37 - NGUYEN Minh Toan

Compressibility and scaling laws in linear-width neural networks with hierarchical features: an information-theoretic perspective

P38 - NICOLAI Francesco

On the Role of Activation Functions in the Storage Capacity of Neural and Associative Memory Models

P39 - ORIENTALE CAPUTO Carlo

Categorical coding beyond binary neuronal states enables memory transitions: a DMFT analysis of Potts associative network

P40 - PASTORE Mauro

Critical capacity of shallow neural networks in the interpolation asymptotics: a statistical mechanics analysis

P41 - PELLEGRINELLI Ilaria

Masked Consensus-based Bayesian factorization for a multi-omics tensor.

P42 - PEROSA Marta

Controlling and understanding representations in state-space models of biological sequences

P43 - QAWASMEH Aminah Husein

Phase Transition in the Ising Model on (m,k)-Ary Trees

P44 - RICCI Fabiola

A Fourier perspective on the learning dynamics of neural networks: from sample complexities to mechanistic insights

P45 - RICCI Gabriele

Epidemic thresholds and transient dynamics in multi-type SIR and SIS models

P46 - SAWAYA Kazuma

Provable False Discovery Rate Control for Deep Feature Selection

P47 - SAXENA Rajat

Escaping plasticity-memory dilemma in intelligent robotic systems

P48 - SHARMA Nisha

Phase behavior of a locally disordered non-conserving exclusion process with finite resources

P49 - SIETSEMA Alexander Nicholas

Harmful Overfitting in Sobolev Spaces

P50 - SKERK Rudy

Statistical physics of deep learning: Optimal learning of a multi-layer perceptron near interpolation

P51 - SUCCESS Peace Udoka

OPTIMAL SOLUTIONS TO STOCHASTIC DIFFERENTIAL EQUATIONS

P52 - TALEBLOU Mina

A practical examination of uncertainty quantification approaches in machine learning for materials science

P53 - TOSELLO Francesco

Does Data Structure Affect Learning in Restricted Boltzmann Machines?

P54 - TSHIANGOMBA Reagan Kasonsa

Stabilization system for solar loading thermography applied on cultural heritage objects exposed outdoors: the contribution of advanced algorithms and dual-branch U-Net

P55 - UMAR Ali Hussaini

The Effect of Label Noise on the Information Content of Neural Representations

P56 - VARAKUNAN Saranya

GFlowNets + PINNs for Multi-Candidate Symbolic Model Discovery in Dynamical Systems from Sparse Data

P57 - WEBER Emma Caroline

Reconstructing Incomplete Surface Temperature Fields using Graph Attention Networks

P58 - ZHANG Kaining

Maximizing Memory Capacity in Heterogeneous Networks

P01

Implicit Bias as a Gauge Correction: Theory and Inverse Design

Nicola Aladrah [1] , Emanuele Ballarin [1] , Matteo Biagetti [2] , Alessio Ansuini [2] , Alberto d'Onofrio [1] , Fabio Anselmi [1,3]

[1]Università di Trieste, Dipartimento di Matematica, Informatica e Geoscienze, Trieste, Italy
[2] Area Science Park, Trieste, Italy
[3] Massachusetts Institute of Technology, Cambridge, USA

A central problem in machine learning theory is to explain how stochastic training selects specific solutions among the many parameter configurations that achieve identical loss, a phenomenon known as implicit bias.
We identify a general mechanism for implicit bias arising from the interaction between stochastic optimization and continuous symmetries of model parameterizations. In the diffusion-limit description of stochastic gradient descent, parameter symmetries create manifolds of equivalent solutions whose volume affects the stationary distribution of the dynamics.
We show that when redundant symmetry directions are removed through gauge fixing, the stationary Gibbs distribution acquires a geometric correction proportional to the log-determinant of a Gram matrix associated with the symmetry constraints. This correction acts as an effective regularizer that we interpret as the implicit bias of the model.
The resulting framework provides a constructive method to compute implicit biases induced by model symmetries and recovers several known phenomena, including sparsity and spectral biases arising from scalar and matrix factorizations. Moreover, it enables an inverse-design principle: by introducing suitable redundant parameterizations, one can engineer desired implicit biases.
We validate these predictions through numerical experiments on controlled synthetic models, demonstrating quantitative agreement between theory and stochastic training dynamics.

P02

Overparametrization bends the landscape: BBP transitions at initialization in simple Neural Networks

Brandon Livio Annesi [1] , **Dario Bocchi** [1] , **and Chiara Cammarota** [1,2]

[1] *Physics Department, University of Rome 'La Sapienza', Piazzale Aldo Moro 5, Rome, 00185, Italy*

[2] *INFN sezione Roma1, Piazzale Aldo Moro 5, Rome, 00185, Italy*

High-dimensional non-convex loss landscapes play a central role in the theory of Machine Learning. Gaining insight into how these landscapes interact with gradient-based optimization methods, even in relatively simple models, can shed light on this enigmatic feature of neural networks. In this work, we will focus on a prototypical simple learning problem, which generalizes the Phase Retrieval inference problem by allowing the exploration of overparametrized settings. Using techniques from field theory, we analyze the spectrum of the Hessian at initialization and identify a Baik-Ben Arous-Péché (BBP) transition in the amount of data that separates regimes where the initialization is informative or uninformative about a planted signal of a teacher-student setup. Crucially, we demonstrate how overparameterization can bend the loss landscape, shifting the transition point, even reaching the information-theoretic weak-recovery threshold in the large overparameterization limit, while also altering its qualitative nature. We distinguish between continuous and discontinuous BBP transitions and support our analytical predictions with simulations, examining how they compare to the finite-N behavior. In the case of discontinuous BBP transitions strong finite-N corrections allow the retrieval of information at a signal-to-noise ratio (SNR) smaller than the predicted BBP transition. In these cases we provide estimates for a new lower SNR threshold that marks the point at which initialization becomes entirely uninformative.

P03

Time-Series Forecasting with Restricted Boltzmann Machines via Conditional Gibbs Sampling

Roger Bader[1] **and Giuseppe Genovese**[1]

[1] *University of Freiburg, Freiburg im Breisgau, Germany*

Restricted Boltzmann machines (RBMs) are energy-based models with well-established training and sampling procedures. While RBMs are widely used for generative modelling, their use for time-series forecasting is less explored, and existing temporal variants often modify the architecture and require specialized training [1, 2, 3]. This poster presents a simple forecasting procedure that uses the original RBM structure with binary variables and remains compatible with standard RBM learning algorithms.

We consider a discrete-valued time series $(X_t)_{t\geq 1}$ with $X_t \in \mathcal{X}$, where $\mathcal{X}$ is finite. Under an order-k dependence assumption, one-step forecasting amounts to estimating the distribution $\mathbb{P}(X_{t+1} = \cdot \mid X_{t-k+1:t} = x_{t-k+1:t})$, where $X_{t-k+1:t} := (X_{t-k+1}, \ldots, X_t)$. With an additional stationarity assumption, this conditional is time-invariant, so it suffices to learn

$$\mathbb{P}(X_{k+1} = \cdot \mid X_{1:k} = x_{1:k}), \qquad x_{1:k} \in \mathcal{X}^k.$$

To use a binary RBM, we fix a binary encoding $\phi : \mathcal{X} \to \{+1, -1\}^d$ and concatenate $(k+1)$ consecutive encoded symbols into a visible vector $v \in \{+1, -1\}^{(k+1)d}$, with blocks $v_i = \phi(X_i)$. From an observed sequence $(x_t)_{t=1}^T$, we construct a training set of encoded sliding windows $(\phi(x_{t-k+1}), \ldots, \phi(x_{t+1}))$ and train an RBM with joint density $p_\theta(v, h) \propto \exp(-E_\theta(v, h))$ (e.g. via Contrastive Divergence or Parallel Tempering). After training, the visible marginal $p_\theta(v)$ approximates the distribution of these observed encoded windows, i.e. a data-driven estimate of the joint distribution of $(X_1, \ldots, X_{k+1})$.

Given an observed past context $x_{1:k}$, we generate forecasts by running a conditional (blocked) Gibbs sampler: the first k visible blocks are clamped to $\phi(x_1), \ldots, \phi(x_k)$, while the final block is updated by alternately sampling $h \sim p_\theta(h \mid v)$ and $v_{k+1} \sim p_\theta(v_{k+1} \mid h, v_{1:k})$. We have shown that this Markov chain is ergodic with stationary distribution $p_\theta(v_{k+1} \mid v_{1:k})$, yielding a principled probabilistic forecast without modifying the RBM architecture or training procedure.

[1] G. W. Taylor, G. E. Hinton, S. T. Roweis, Adv. Neural Inf. Process. Syst. **19**, 1345–1352 (2006).
[2] G. W. Taylor, G. E. Hinton, Proc. 26th Int. Conf. Mach. Learn. (ICML), 1025–1032 (2009).
[3] I. Sutskever, G. E. Hinton, G. W. Taylor, Adv. Neural Inf. Process. Syst. **21**, 1601–1608 (2008).

P04

Beyond Surface Similarities: GNN For More Natural Modeling in Recommender Systems

Khaoula BELAROUI

PhD Student, University Hassan 2 – ENSET

Abstract:

Many real-world systems, including recommender systems, are governed by complex interaction structures that cannot be fully captured through surface-level similarities alone. Traditional approaches such as matrix factorization or classical collaborative filtering primarily model pairwise similarities, often overlooking the relational organization underlying user–item interactions. This poster presents my recent work, **"Towards Natural Representation of User Preferences: A Graph Neural Network Approach to Recommender Systems"**, which investigates how Graph Neural Networks (GNNs) enable more expressive and structure-aware representations of recommendation data. By modeling users and items as nodes within a graph and leveraging message-passing mechanisms, the proposed approach captures higher-order relational dependencies and propagates information through the underlying graph topology. Through empirical evaluation on benchmark datasets, the study highlights how incorporating structural information leads to richer representations of user preferences compared to traditional recommendation methods. While the work is grounded in recommender systems, it also illustrates how graph-based learning provides a principled framework for modeling complex systems, where global behavior emerges from local interactions. By emphasizing structure-aware learning, this poster contributes to the broader discussion on how Graph Neural Networks can move beyond surface similarities and support more natural representations in modern machine learning.

[1] W. Hamilton, Z. Ying, and J. Leskovec, "*Inductive Representation Learning on Large Graphs*," in Advances in Neural Information Processing Systems (NeurIPS), vol. 30, 2017.

[2] M. M. Bronstein, J. Bruna, Y. LeCun, A. Szlam, and P. Vander- gheynst. *Geometric deep learning : going beyond euclidean data.* IEEE Signal Processing Magazine, 34(4) :18–42, July 2017.

[3] H. Wang, J. Zhang, M. Zhao, W. Zhang, F. X. Xie, and M. Wang, "*Neural Graph Collaborative Filtering,*" in Proceedings of the 42nd International ACM SIG.

P05

Recovering geometric features via the Fourier transform

T. Beretti[1]

[1] *SISSA - International School for Advanced Studies*

Modern sensing and learning pipelines often access an object via spectral information (e.g., Fourier transform, or frequency-domain surrogates). A natural question is which geometric properties can be inferred from limited data in the frequency domain, and how this relates to the design of the sampling.

In [1], a weighted Plancherel identity for BV functions shows that the high-frequency energy of the Fourier transform localises on the discontinuity set. When specialising to indicator functions, this yields a Fourier characterisation of sets of finite perimeter and an identity that recovers the perimeter from high frequencies. We connect these identities to the sampling side of discrepancy theory. Specifically, we consider the quadratic error in sampling with an N-point set (in d dimensions) and averaging over affine transformations of test functions. By considering the respective optimal design problem, we find that the best achievable average squared error decays at a rate $N^{(1-d)/d}$, with the leading term determined by the jump-energy. The optimal value in such a discrepancy problem is obtained by Fourier-analytic methods and determined by geometric complexity, suggesting a route to recover discontinuity sets from spectral data.

[1] T. Beretti, L. Gennaioli, **Fourier transform of BV functions and applications** *Preprint* (2024).

P06

Memorization, convergence and generalization in first-order Markov chains

Eleonora Bergamin[1], **Enrico Ventura**[1] **and Sebastian Goldt**[1]

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

What does it mean for a generative model to generalize? In supervised learning, interpolation could be benign, but in generative modeling, overfitting is pathological: a model that memorizes reproduces its training samples during generation. Recent experiments by Kadkhodaie, Guth, Simoncelli, and Mallat [1] demonstrated a transition in diffusion models: two denoisers trained on disjoint datasets produce identical outputs above a critical sample size, suggesting a phase transition from memorization to generalization. However, a theoretical understanding of this phenomenon in autoregressive models remains open.

We study this question in a simplified setting that still allows us to characterize an autoregressive model: a first-order Markov chain encoding the statistical information in the data.

To quantify similarity between two such chains, we introduce a notion of overlap based on the survival probability under maximum coupling, following Levin and Peres [2]. The survival probability is defined as the probability that two chains, started from the same initial state and coupled maximally, generate identical sequences up to time t. Its exponential decay rate is governed by the spectral radius of a kernel matrix formed from the entrywise minimum of the two estimated transition matrices.

We then study the transition from memorization to generalization through a random matrix perspective. Specifically, we analyze a high-dimensional teacher–student toy model in which the teacher generates a Markov chain whose transition matrix consists of a low-rank signal component and noise, and the student observes samples and estimates the transition matrix from data. We show that signal retrieval is controlled by a phase transition in the spectrum of the estimated transition matrix: below a critical sample complexity, estimation noise dominates; above it, the leading eigenstructure aligns with the teacher model. This transition can be characterised in analogy with Baik–Ben Arous–Péché (BBP) transitions in spiked random matrix models.

Our results provide an analyzable toy model for convergence in generative models. They model how memorisation, convergence, and spectral recovery are related yet distinct phenomena, and offer a minimal theoretical framework for understanding generalisation transitions in autoregressive generative models.

[1] Z. Kadkhodaie, F. Guth, E. P. Simoncelli, and S. Mallat, Proc. Int. Conf. Learn. Represent. (2024).

[2] D. A. Levin and Y. Peres, *Markov Chains and Mixing Times*, 2nd ed. (American Mathematical Society, 2017).

P07

A FLOW-AWARE TRAINING STRATEGY FOR PHYSICS-INFORMED NEURAL NETWORKS

Pietro Cestola[1*], **Luciano Teresi**[2], **and Antonio De Simone**[3]

[1]*Department of Mathematics and Physics, Roma Tre University, Largo San Leonardo Murialdo 1, Roma, 00146, Italy*

[2]*Department of Industrial Engineering Electronics and Mechanics, Roma Tre University, Via Vito Volterra 62, Roma, 00184, Italy*

[3]*Mathematics Area, International School for Advanced Studies (SISSA), Via Bonomea 265, Trieste, 34136, Italy*

Physics-Informed Neural Networks (PINNs) typically update parameters at all collocation points from the first epoch, even in regions still unreachable from boundary or initial conditions. In these zones, early updates may be uninformative or harmful, as residual gradients often correlate poorly with the true error. We introduce a flow-aware training strategy that delays supervision until the governing physics can propagate meaningful information. The computational mesh is decomposed into geodesic subdomains ranked by distance from an information boundary, where initial or boundary conditions are applied, and progressively activated according to an epoch schedule. This selective exposure concentrates optimization where updates are most effective, preventing wasted capacity and misleading gradients in causally disconnected regions. The method requires no architectural changes and uses the standard PINN loss; only the sampling mask evolves over time. Benchmarks on seven PDE problems show that flow-aware training matches or improves baseline accuracy while reducing computational cost.

[1] P. Cestola, L. Teresi, A. De Simone, A Flow-Aware Training Strategy for Physics-Informed Neural Networks, *Comput. Methods Appl. Mech. Eng.*, Manuscript CMAME-D-25-03873R1.

P08

Spatially-Informed Biologically Interpretable Machine Learning Approaches for Analyzing Neural Data

Madelyn Esther C. Cruz [1,2] , **Daniel B. Forger** [1,2]

[1] *Department of Mathematics, University of Michigan, Ann Arbor, Michigan, USA*
[2] *Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, Michigan, USA*

Modified Hodgkin-Huxley neuron models have previously been used in biological neural networks (BNNs) to classify neural data. This study extends the use of such BNNs, where we apply backpropagation to networks of biophysically-accurate model neurons to generate EEG-like signals to predict brain states and analyze the neurophysiology of EEG signals using parameters derived from the BNNs. In particular, we trained our BNNs to exhibit different frequencies observed in EEG recordings and found that the variability of synaptic weights and applied currents increased with the target frequency range. However, while the present BNNs provide biological interpretability, they treat each electrode independently without considering spatial relationships as in real measurements. Consequently, we extend the current architecture to incorporate spatial information and connectivities as additional hyperparameters. By incorporating these metrics, our adapted BNNs extrapolate activity into deeper brain regions, with neurons in hidden layers remaining biologically interpretable and visualizable. Their classification performance on single-subject data remains comparable to previous models, while significantly improving cross-subject classification. Additionally, using physical distances between electrodes further enhances accuracy. These findings highlight the significance of spatial relationships in biological networks. Analyses of the learning mechanism of the BNNs reveal spatiotemporal patterns across the network, including traveling waves, offering novel insights into the dynamics of brain activity and improving interpretability through spatially-informed learning.

P09

Algebraic issues arising in training neural networks

Marco Dell'Orto[1], **Fabio Marcuzzi**[1]

[1] *Università di Padova, Dipartimento di Matematica "Tullio Levi-Civita"*

It is a well-documented phenomenon [1, 2] that, under certain conditions, neural networks trained with gradient descent tend to concentrate learned information in a relatively small number of leading singular components of their weight matrices. The study of the associated large singular values, commonly referred to as *spikes*, has therefore attracted attention in recent research.

For one-hidden-layer linear networks, some insight into the formation of spikes is provided by [3], which characterizes their number as a function of the model rank and initialization scale.

For even simple nonlinear networks, however, the mechanism underlying spike formation and the factors determining their number are not yet fully understood.

In this ongoing work we investigate this phenomenon in ReLU networks using a teacher–student setup, with the goal of developing a deeper understanding. Moreover we examine what happens when a model transitions from linear to nonlinear.

Finally, we aim to apply these insights to the study of VarMiONs (Variationally Mimetic Operator Networks) [4] for solving parameterized PDEs. Thanks to their highly interpretable structure, VarMiONs provide a promising setting in which to explore connections between spike formation, continuum physics, and model parsimony.

[1] M.Thamm, M.Staats, and B.Rosenow, 'Random matrix analysis of deep neural network weight matrices,' Phys. Rev. E **106**, 054124 (2022).

[2] C.H.Martin and M.W.Mahoney, 'Implicit self-regularization in deep neural networks: Evidence from random matrix theory and implications for learning', J.Mach.Learn.Res. **22** (2021).

[3] A. Varre, M.L. Vladarean, L. Pillaud-Vivien, and N. Flammarion, 'On the spectral bias of two-layer linear networks', *Advances in Neural Information Processing Systems* **36**, 2811 (2023).

[4] D.Patel, D.Ray, M.R.A.Abdelmalik, T.J.R.Hughes, and A.A.Oberai, 'Variationally mimetic operator networks', Comput. Methods Appl. Mech. Eng. **419**, 116536 (2024).

P10

Inference in continuous-space Markov processes on graphs

Florio F.[1], **and Braunstein A.**[1,2]

[1]*Institute of Condensed Matter Physics and Complex Systems, Department of Applied Science and Technology, Politecnico di Torino, Corso Duca degli Abruzzi 24, 10129 Torino, Italy*
[2]*Italian Institute for Genomic Medicine, Candiolo Cancer Institute, Fondazione del Piemonte per l'Oncologia (FPO), Candiolo, 10060 Torino, Italy*

Recent advances in Statistical Physics, such as the Dynamic Cavity (DC) and Dynamics Belief Propagation (DBP) methods, have enabled semi-analytical estimation of observables in discrete stochastic processes over large, sparse networks. However, these methods are limited by their reliance on succinct parametrizations of pair-of-neighbors trajectories. To address this limitation, the Matrix-Product Belief Propagation (MPBP) method was introduced, offering linear computational complexity in time for a large family of models with discrete degrees of freedom. Moreover, efficient calculation of the MPBP update relies on a recursive computation involving intermediate variables that also need to be discrete. Here, a generalization of MPBP to continuous degrees of freedom is presented, using a tunable expansion in a Hilbert function basis. The method's efficacy is demonstrated by employing a Fourier basis for the Kinetic Ising dynamics with real-valued random couplings, where intermediate "local fields" are effectively continuous. This approach reduces the computational cost of the MPBP update step from exponential to linear in the node's degree, enabling the study of previously intractable graph classes. Results are compared to Monte-Carlo simulations.

P11

Annealing in variational inference mitigates mode collapse: A theoretical study on Gaussian mixtures

Luigi Fogliani[1,2], **Bruno Loureiro**[1], **and Marylou Gabrié**[2]

[1]*(Presenting author underlined) Departement d'Informatique, École Normale Supérieure, Université PSL, CNRS*

[2]*Laboratoire de Physique de l'École Normale Supérieure, Université PSL, CNRS*

Mode collapse — the failure to capture one or more modes when targetting a multimodal distribution — is a central challenge in modern variational inference. In this work, we provide a mathematical analysis of annealing-based strategies for mitigating mode collapse in a tractable setting: learning a Gaussian mixture, where mode collapse is known to arise. Leveraging a low-dimensional summary statistics description, we precisely characterize the interplay between the initial temperature and the annealing rate, and derive a sharp formula for the probability of mode collapse. Our analysis shows that an appropriately chosen annealing scheme can robustly prevent mode collapse. Finally, we present numerical evidence that these theoretical trade-offs qualitatively extend to neural network–based models, RealNVP normalizing flows, providing guidance for designing annealing strategies mitigating mode collapse in practical variational inference pipelines.

[1] Luigi Fogliani, Bruno Loureiro, Marylou Gabrié, *Annealing in variational inference mitigates mode collapse: A theoretical study on Gaussian mixtures*, arXiv preprint arXiv:2602.12923, (2026).

P12

$\mathrm{E}(3)$-Equivariant Fragment-based Graph Neural Networks for Biomolecules

Shih-Hsin Wang[3], Yuhao Huang[3], Taos Transue[3], Justin Baker[2], Jonathan Forstater[1], Thomas Strohmer[1], and Bao Wang[3]

[1]*Department of Mathematics, UC Davis, Davis, CA 95616, USA*
[2]*Department of Mathematics, UCLA, Los Angeles, CA 90095, USA*
[3]*Department of Mathematics and Scientific Computing and Imaging (SCI) Institute University of Utah, Salt Lake City, UT 84102, USA*

Fragment-based graph neural networks (GNNs) have emerged as a class of effective models to overcome the limitations of traditional GNNs in learning meaningful graph substructures. These models utilize a pre-designed knowledge dictionary of fragments – such as functional groups in organic molecules or residues in proteins – to identify key substructures within data and learn features from these knowledge-based components. However, existing fragment-based models often overlook the encoding of crucial geometric information required to accurately capture the fragments' spatial arrangement. To bridge this gap, we propose $\mathrm{E}(3)$-equivariant fragment-based GNNs, a new class of fragment-based GNNs that integrate equivariant frames of fragments to ensure enhanced performance with theoretical guarantees. Furthermore, we augment this framework with a knowledge dictionary of protein secondary structures to design tailored models for biomolecular modeling. Empirical results demonstrate the efficiency and advantages of our proposed framework over existing approaches.

P13

Universal approximation of semantic couplings via Sinkhorn Transformers

Demián Fraiman[1]

[1]*Calculus institute, University of Buenos Aires*

Transformer models have achieved outstanding performance capturing contextual relationships, yet a rigorous mathematical explanation of their expressive power is still missing. In this work, we propose a measure-theoretic framework to analyze the attention mechanism of Transformers through optimal transport theory and establish a universal approximation theorem for a broad class of measure-valued mappings arising in contextual representation learning.

Texts of arbitrary length are modeled as probability measures on a compact embeddign metric space X, allowing variable-size inputs to be treated in a unified functional-analytic setting. Contextual relations between two texts are represented as probability measures on the product space $X \times Y$, interpreted as semantic couplings with prescribed marginals. Within this framework, a semantic coupling is defined as a continuous map

$$F : \mathcal{P}(X) \times \mathcal{P}(Y) \to \mathcal{P}(X \times Y)$$

such that $F(\mu, \nu) \in \Pi(\mu, \nu)$ for every pair (μ, ν).

We show that the space of semantic couplings is convex and closed in the topology induced by Wasserstein metrics. We then introduce a class of architectures, called *Sinkhorn Transformers*, obtained by combining measure-theoretic attention operators with entropic optimal transport layers. A key observation is that entropic transport plans admit continuous densities with respect to the product measure and satisfy stability and Lipschitz properties under perturbations of the cost function.

Using approximation arguments in optimal transport and stability results for Schrödinger potentials, we prove that transport plans with continuous densities are dense in the space of all couplings. This leads to our main theorem: for compact metric spaces X and Y, the family of Sinkhorn Transformers is dense in the space of semantic couplings under the W_p metric. In particular, for every semantic coupling F, every $p > 1$, and every $\varepsilon > 0$, there exists a Sinkhorn Transformer T^* such that

$$\sup_{(\mu,\nu)\in\mathcal{P}(X)\times\mathcal{P}(Y)} W_p(T^*(\mu, \nu), F(\mu, \nu)) < \varepsilon.$$

This result provides a rigorous characterization of the expressive power of a structured attention mechanism in a fully measure-theoretic setting, connecting deep learning expressivity with entropic optimal transport and geometric properties of probability measures.

[1] C. Léonard, From the Schrödinger Problem to the Monge–Kantorovich Problem, *Journal of Functional Analysis*, 262(4):1879–1920, 2012.

[2] M. Nutz and J. Wiesel, Quantitative Stability of Regularized Optimal Transport and Convergence of Sinkhorn's Algorithm, *SIAM Journal on Mathematical Analysis*, 54(4):4609–4640, 2022.

[3] M. Nutz and J. Wiesel, Stability of Schrödinger Potentials and Convergence of Sinkhorn's Algorithm, *Probability Theory and Related Fields*, 184:1355–1406, 2022.

[4] T. Furuya, M. V. de Hoop, and G. Peyré, Transformers Are Universal In-Context Learners, arXiv:2408.01367, 2024.

P14

Machine Learning modeling of cancer dynamic

Luca Frattegiani[1,2], **Guido Sanguinetti**[1,2], **and Andrea Sottoriva**[2]

[1] *Scuola Internazionale Superiore di Studi Avanzati, Via Bonomea 265, Trieste, Italy*
[2] *Fondazione Human Technopole, Viale Rita Levi Montalcini 1, Milano, Italy*

Understanding rules that drive tumor tissue growth is a fundamental challenge in computational biology, which can be addressed through machine learning. Here we aim to describe the concrete problem we face and introduce the mathematical/methodological framework one can use to infer such dynamics from observables. The fundamental feature we want to take into account is the role played by local interactions between cells, which are in general believed to have a remarkable influence on the cancer evolutionary process.

Spatial transcriptomics

The experimental procedure that generates data containing both positional and biological information at a single cell resolution is *spatial transcriptomics*. After sequencing, the output is a collection of time-indexed datasets $\{\mathcal{D}_t\}_t$, each of which contains a sample of independent *organoids* (i. e. groups of cells sharing the same environment) measured at the same physical time. Each organoid is then modeled as a graph where nodes represent cells which are labeled with a state (the single-cell gene expression) while edges represent cell-to-cell connections.

Neural Cellular Automata

The tool chosen for the purpose is the one of Cellular Automaton (CA), a model frequently used in different fields of natural sciences to provide a mathematical representation of self-reproducing systems [1]. In particular, CA require two crucial ingredients to reproduce the temporal evolution of the system:

- *Neighborhood of interaction:* Each node evolves under the influence of its local environment (surrounding cells). In our case, the neighborhood is given by the connections within the organoid graph.
- *Transition rule:* Either a function or a conditional probability $x_t \mapsto x_{t+1} = \mathcal{R}(x_t \mid \mathcal{N}(x_t))$ that pushes forward the system's state and applies node-wise.

Differently from standard CA theory, here the transition rule is unknown and we aim to parametrize it with a neural network [2]. The final output is then a generative process of cells organoids whose temporal dynamic is explicitly shaped in a CA fashion, allowing for a specific interpretation of how interaction forces operate.

Distribution matching problem

Due to the destructive nature of spatial transcriptomics, our data is not *longitudinal* in the sense that we can't track the whole temporal trajectory of an organoid, but just its snapshot at time t. As a consequence, we deal with multi-marginal distributions ($\mathcal{D}_t \sim \mathcal{P}_t$ and is totally independent fromm $\mathcal{D}_{t-1}$), a scenario in which is extremely difficult to train an automata rule which instead would benefit of observing direct transitions $x_t \mapsto x_{t+1}$. Thus, our strategy is to initially learn a continuous normalizing flow through flow matching, which will allow us to produce synthetic organoids as a proper Markov process and train the Neural CA on these.

[1] Deutsch, Dormann, "Cellular Automaton Modeling of Biological Pattern Formation" (2005).
[2] Mordvintsev, Randazzo, Niklasson, Levin, "Growing Neural Cellular Automata" (2020).
[3] Lipman, Chen, Ben-Hamu, Nickel, Le, "Flow Matching for Generative Modeling" (2023).

P15

Discrete Analytic Functions on a Rhombic Lattice

Daniel Alpay[1] , Angel Fuerte[2] , and Dan Volok[2]

[1]Chapman University
[2] Kansas State University

The notion of discrete analytic functions on rhombic lattices, originating in the work of Ferrand[1] and Duffin[2][3], provides a natural discrete counterpart to classical complex analysis. In this work, we explore the structure of reproducing kernel Hilbert spaces of discrete analytic functions that remain invariant under suitably defined shift operators. This is joint work with Daniel Alpay and Dan Volok.

[1] Jacqueline Ferrand. Fonctions préharmoniques et fonctions préholomorphes, Bull. Sci. Math., vol 68(1994), second series, pp. 152-180.
[2] R.J. Duffin. Basic properties of discrete analytic functions, Duke Math. J. 23 (1956), 335-363. MR 17, 1193.
[3] R.J. Duffin. Potential theory on a rhombic lattice. Journal of Combinatorial Theory, 5: 258-272, 1968.

P16

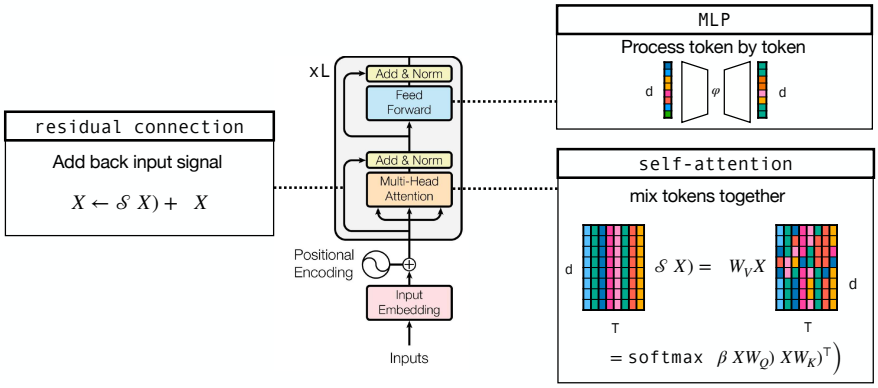
Two Failure Modes of Deep Transformer Models and How to Avoid Them

a unified theory of signal propagation at initialisation

A. Giorlandino

S. Goldt

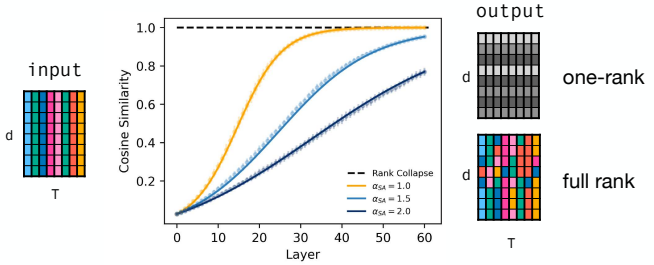
SISSA



TL;DR

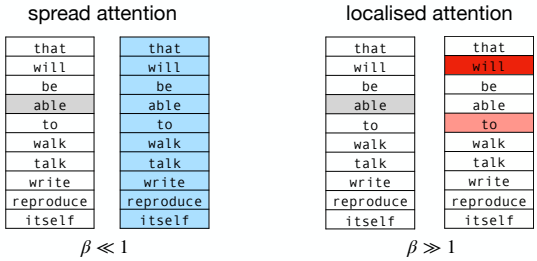
We provide a complete analysis of signal propagation in a vanilla Transformer block—including the self-attention, residual connections, LayerNorm, and MLP—making it possible to predict the trainability of Transformers by avoiding both Rank and Entropy Collapse.

Rank Collapse



All tokens become the same.

Entropy Collapse



Attention is frozen and random.

Sequence Geometry

The input sequence has a simplex random geometry.

The geometry is preserved in the limits $d \to \infty$ and $T \to \infty$.

The only evolving quantities are the average squared token norm q and the average pairwise overlap p.

Forward

Let $\sigma_Q^2 = \sigma_K^2 = \beta d^{-1}\sqrt{\log T}$, the critical threshold for entropy collapse is

$$\beta_c = \sqrt{\frac{2}{q\,q-p)}}$$

and the cosine similarity $\rho = p/q$ evolves as:

$$\mathcal{S}\,\rho) = \begin{cases} 1, & \beta < \beta_c, \\ \dfrac{\rho}{1-\beta^{-1}\sqrt{2\,1-\rho)}}, & \beta > \beta_c\,. \end{cases}$$

Backward

Under the same assumption, the norm of the query/key gradient takes the form of:

$$\frac{T}{d^2\sqrt{\log T}}\mathbb{E}\left\|\frac{\partial \mathcal{L}}{\partial W_Q}\right\|_F^2 = \begin{matrix} 0, & \beta < \beta_c, \\ g\;q,p) > 0, & \beta > \beta_c\,. \end{matrix}$$

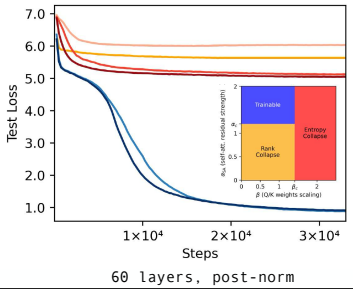
Spread attention = Vanishing gradients

Residual connections are vital !

Avoid both Collapses

Avoid Entropy Collapse by controlling β.

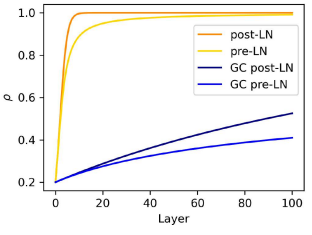
Avoid Rank Collapse by controlling the strength of the skip connection.



Modified Self-Attention

To avoid Entropy Collapse one should start with a spread attention, which effectively takes the mean of the Values; collapsing the representations.

Simply remove the average Value after the self-attention layer.



$$\mathcal{S}_{GC}\,X)_t = \quad {}_{t\tau} W_V X_\tau - W_V \bar{X}$$

P17

A Malliavin-Gamma calculus approach to Score Based Diffusion Generative models for random fields

Giacomo Greco[1]

[1]*(Presenting author underlined) Università degli Studi di Roma Tor Vergata*

We adopt a Gamma and Malliavin Calculi point of view in order to generalize Score-based diffusion Generative Models (SGMs) to an infinite-dimensional abstract Hilbertian setting. Particularly, we define the forward noising process using Dirichlet forms associated to the Cameron-Martin space of Gaussian measures and Wiener chaoses; whereas by relying on an abstract time-reversal formula, we show that the score function is a Malliavin derivative and it corresponds to a conditional expectation. This allows us to generalize SGMs to the infinite-dimensional setting. Moreover, we extend existing finite-dimensional entropic convergence bounds to this Hilbertian setting by highlighting the role played by the Cameron-Martin norm in the Fisher information of the data distribution. Lastly, we specify our discussion for spherical random fields, considering as source of noise a Whittle-Matérn random spherical field.

[1] G. Greco, *A Malliavin-Gamma calculus approach to Score Based Diffusion Generative models for random fields*, arXiv:2505.13189 (2025), accepted in *Indam-Springer volume: Analysis and Geometry of Random Fields*.

P18

DYNAMICAL MEAN FIELD THEORY FOR MOMENTUM-BASED GRADIENT DESCENT IN DEEP NEURAL NETWORKS

Ignacio Guevara[1], **Germán Mato**[1], **Gonzalo Torroba**[2]

[1]*Medical Physics Department, Centro Atómico Bariloche and Instituto Balseiro, CNEA–UNCuyo, San Carlos de Bariloche, Argentina*
[2]*Particle Physics Group, Centro Atómico Bariloche, CONICET, San Carlos de Bariloche, Argentina*

Understanding how deep neural networks learn internal representations during training remains a central challenge in machine learning theory. We address this problem by extending Dynamical Mean Field Theory (DMFT) to networks trained with momentum-based gradient descent. In the standard gradient flow setting, recent work has shown that the training dynamics of wide networks can be reduced, via a path integral formulation and a saddle-point approximation, to a closed set of self-consistent equations for kernel order parameters: correlation functions of hidden unit activations and backpropagated error signals across data points and training times [1]. We extend this framework to the case where the weights evolve according to a second-order dynamical system that incorporates a momentum variable. The momentum carries an independent random initialization, which requires introducing additional stochastic fields per layer and new response functions coupling the momentum initialization to both the forward and backward passes. After applying a Hubbard–Stratonovich transformation and taking the saddle point at infinite width, we obtain a generalized set of DMFT equations in which the momentum timescale appears as an exponential memory kernel, modifying the causal structure of the dynamics relative to gradient flow. The standard DMFT equations are recovered exactly in the limit of vanishing momentum, providing a nontrivial consistency check. Our results offer an analytical framework to study how momentum shapes the evolution of feature kernels throughout training, with direct implications for the theory of representation learning in wide networks.

[1] B. Bordelon and C. Pehlevan, J. Stat. Mech. **2023**, 114009 (2023).

P19

USING MACHINE LEARNING TO ANALYZE USER BEHAVIOR PATTERNS AND FORECAST REVENUE IN DIGITAL PRODUCTS

Karyna Hurinenko

Tarasa Shevchenko National University of Kyiv, Faculty of Computer Science and Cybernetics, Kyiv, Ukraine

This document presents the abstract for the poster presentation at the Summer Graduate School on Mathematics for Machine Learning to be held in Trieste, Italy, from 15 June 2026 to 26 June 2026.

The research explores a framework for analyzing user behavior patterns in digital environments to predict financial outcomes. Using a dataset of user events and session timestamps, the study employs unsupervised learning techniques, specifically K-means and DBSCAN, to segment users into distinct behavioral profiles. The analysis identified high-value clusters where specific actions, such as A3, A5, A9, and A27, served as strong predictors of conversion, with one target cluster showing a 47% subscription rate.

To understand the dynamics of user interaction, Markov Chains were used to analyze transition probabilities between events. The findings indicate that the cumulative frequency and execution of core actions are more significant drivers of payment than a fixed linear sequence of events. Furthermore, the study compares SimpleRNN and Long Short-Term Memory (LSTM) architectures for multi-parameter revenue forecasting.

The results demonstrate that LSTM models, enhanced with Dropout layers and optimized Learning Rates refined to 10^{-5}, significantly outperform basic recurrent models. The final multivariate model achieved a Mean Absolute Error (MAE) of 0.94, confirming the effectiveness of integrating behavioral triggers into predictive financial modeling. This approach provides a robust mathematical foundation for optimizing user experience and increasing company revenue.

[1] U. Kamps, E. Cramer, Grunzuge der Stochastik.

[2] J. Assfalg et al., Knowledge Discovery in Databases, KDD (2001)

P20

Dynamics of Performative Classification in High-Dimensional Gaussian Mixtures

M. Jankola[1], **E. Cyffers**[1,2], **and M. Mondelli**[1]

[1]*(Presenting author underlined) Institute of Science and Technology Austria (ISTA)*
[2] *CNRS Researcher at Dauphine*

The deployment of decision rules can alter the data distribution through strategic responses of affected agents. This feedback mechanism between the model and the data-generating process, known as the performative effect, challenges the standard assumption of a fixed data distribution and can lead to suboptimal or unstable outcomes if ignored. A natural mitigation strategy is repeated risk minimization (RRM), in which a model is iteratively retrained on data reshaped by its own predictions.

In this work, we analyze the performative effect in a binary classification setting based on Gaussian mixture models, focusing on the high-dimensional proportional regime, where the number of samples and features scale comparably. Using the Convex Gaussian Min–Max Theorem, we derive a deterministic characterization of the resulting dynamics. We identify conditions for convergence to fixed points and analyze the evolution of the classification risk along the RRM trajectory. Our analysis reveals a range of behaviors driven by the interplay between model updates and performatively induced distribution shifts, including regimes in which performative effects can improve classification performance.

P21

High-Rank Matrix Denoising Beyond Rotational Invariance

Jean Barbier[1], **Francesco Camilli**[2], **and Suhanee Korpe** [3]

[1]*The Abdus Salam International Centre for Theoretical Physics Strada Costiera 11, 34151 Trieste, Italy*

[2]*Alma Mater Studiorum - Università di Bologna, Dipartimento di Matematica Piazza di Porta S. Donato 5, 40126 Bologna, Italy*

[3] *Indian Institute of Science Education and Research, Bhopal 462066, India*

Matrix denoising is a central inference problem in information theory, with direct relevance to representation learning, particularly dictionary learning, when the matrix exhibits a non-symmetric factorised structure. We study matrix denoising in the high-rank setting, focusing on the challenging case where the signal is generated from non-rotationally invariant priors. We propose a unified theoretical technique that combines replica theory from statistical physics with tools from random matrix theory, in particular HCIZ integrals, to analyse this general problem. We apply this framework to both symmetric and non-symmetric high-dimensional settings. We derive expressions for the free entropy, mutual information, and minimum mean squared error (MMSE), and characterise the associated phase transitions, identifying distinct inference regimes. We demonstrate excellent agreement between theoretical predictions and the optimal denoising algorithm, namely the rotationally invariant estimator (RIE) and with Markov Chain Monte Carlo (MCMC) simulations. Overall, this work bridges techniques from statistical physics and random matrix theory to provide a unified approach for matrix denoising beyond rotationally invariant priors, yielding results consistent with RIE-based optimal denoising.

[1] J. Barbier, F. Camilli, J. Ko, and K. Okajima, Phys. Rev. X **15**, 021085 (2025).
[2] Y. Kabashima, F. Krzakala, M. Mézard, A. Sakata, and L. Zdeborová, IEEE Trans. Inf. Theory **62**(7), 4228–4265 (2016).
[3] F. Pourkamali, J. Barbier, and N. Macris, IEEE Trans. Inf. Theory **70**(11), 8133–8163 (2024).
[4] F. Pourkamali and N. Macris, Inf. Inference **14**(3), iaaf015 (2025).
[5] A. Maillard, F. Krzakala, M. Mézard, and L. Zdeborová, J. Stat. Mech.: Theory Exp. **2022**(8), 083301 (2022).
[6] J. Bun, R. Allez, J.-P. Bouchaud, and M. Potters, IEEE Trans. Inf. Theory **62**(12), 7475–7490 (2016).

P22

Abstract template for Federated Associative-Memory Framework for Continual Learning via Spectral Inference

Andrea Alessandrelli[1,4,5], **Fabrizio Durante**[1], **Andrea Ladiana**[2,5], **Andrea Lepre**[3,5]

[1] *Università del Salento, Dipartimento di Matematica e Fisica , Lecce, Italy*

[2] *Sapienza Università di Roma, Dipartimento di Scienze di Base e Applicazioni all'Ingegneria, Roma, Italy*

[3] *Sapienza Università di Roma, Dipartimento di Matematica, Roma, Italy*

[4] *Istituto Nazionale di Fisica Nucleare, Sezione di Lecce, Italy*

[5] *Istituto Nazionale di Alta Matematica Francesco Severi, Roma, Italy*

Federated learning enables collaborative training without sharing raw data, but struggles under client heterogeneity and streaming distribution shifts, where drift and novel data can impair convergence and cause forgetting. We propose a federated associative-memory framework that learns shared archetypes in heterogeneous, continual settings, where client data are independent but not necessarily balanced. Each client encodes its experience as a low-rank Hebbian operator, sent to a central server for aggregation and factorization into global archetypes. This approach preserves privacy, avoids centralized replay buffers, and is robust to small, noisy, or evolving datasets. We cast aggregation as a low-rank-plus-noise spectral inference problem, deriving theoretical thresholds for detectability and retrieval robustness. An entropy-based controller balances stability and plasticity in streaming regimes. Experiments with heterogeneous clients, drift, and novelty show improved global archetype reconstruction and associative retrieval, supporting the spectral view of federated consolidation.

P23

Estimation and Simulation in nonlinear time series models samples driven by applications on topological data analysis

Nabil Laiche [1]

[1]*Applied department Mathematics. IMUS. Seville University. Spain*

Abstract

Our research introduces a novel framework that leverages topological data analysis (TDA) to extract complex, hidden properties from time series data, with a specific focus on economic and financial contexts. While traditional time series analysis often relies on statistical moments or spectral properties, it frequently fails to capture the underlying shape and structure of the data. To address this, we apply TDA tools, specifically persistent homology, to characterize multi-dimensional data landscapes. This allows us to quantify invariant topological features such as connected components and loops that persist across different scales, providing a robust signature of the underlying dynamical system. A core component of our contribution is the extraction of topological entropy from these persistence landscapes. This entropy metric serves as a measure of complexity and disorder within the system, offering a unique lens through which to analyze market regimes or economic cycles. By transforming raw time series into these topological summaries, we create rich, feature-rich representations that go beyond conventional indicators. To operationalize these insights, we have developed specialized algorithms designed to seamlessly integrate topological features with both deep learning and machine learning architectures. For instance, persistence diagrams are vectorized using templates like persistence images, making them compatible with convolutional neural networks (CNNs) or gradient-boosting models. This hybrid approach allows us to enhance predictive performance and risk characterization in financial data. Our experiments demonstrate that incorporating persistence and entropy features significantly improves the accuracy of regime detection and price movement forecasting, validating the power of combining algebraic topology with modern computational methods for big data analytics in economics.

Keywords

Big data; Deep learning; Networks; Nonlinear Model; Stochastic Data. Persistent entropy.

References

[1] Lim, Bryan, and Stefan Zohren. "Time-series forecasting with deep learning: a survey." *Philosophical transactions of the royal society a: mathematical, physical and engineering sciences* 379.2194 (2021).

[2] Busseti, Enzo, Ian Osband, and Scott Wong. "Deep learning for time series modeling." *Technical report, Stanford University* (2012): 1-5.

[3] Fischer, Thomas, Christopher Krauss, and Alex Treichel. *Machine learning for time series forecasting-a simulation study*. No. 02/2018. FAU Discussion Papers in Economics, 2018.

[4] Hensel, Felix, Michael Moor, and Bastian Rieck. "A survey of topological machine learning methods." *Frontiers in Artificial Intelligence* 4 (2021): 681108.

[5] Atienza, Nieves, Rocío González-Díaz, and Manuel Soriano-Trigueros. "On the stability of persistent entropy and new summary functions for topological data analysis." *Pattern Recognition* 107 (2020): 107509.

P24

Emergent Cooperativeness in Three-Directional Associative Memories: A Statistical Mechanics Perspective on Supervised and Unsupervised Learning.

A. Alessandrelli[1], A. Barra[3], A. Ladiana[3], **A. Lepre**[2], F. Ricci-Tersenghi[4]

[1]*Dipartimento di Matematica e Fisica, Università del Salento, Lecce.*
[2]*Dipartimento di Matematica, Sapienza Università di Roma.*
3 *Dipartimento di Scienze di Base Applicate all'Ingengeria, Sapienza Università di Roma.*
4 *Dipartimento di Fisca, Sapienza Università di Roma.*

We introduce a unified learning framework for Three-Directional Associative Memory (TAM) models, extending the classical Hebbian paradigm to multi-layer architectures under both supervised and unsupervised protocols. By leveraging techniques from the statistical mechanics of disordered systems, specifically Guerra's interpolation and replica theory, we provide an analytical description of the network's storage capacity and computational phase transitions [1].

A central result is the emergence of "cooperativeness" among interconnected layers [2]. Our analysis reveals that retrieval performance is not isolated to local layer noise; instead, layers trained on low-entropy datasets actively stabilize those exposed to higher noise, synergistically balancing retrieval boundaries across the network. Validated by Monte Carlo simulations, these findings offer a robust blueprint for designing resilient hetero-associative systems capable of pattern completion and disentanglement.

[1] Andrea Alessandrelli and Adriano Barra and Andrea Ladiana and Andrea Lepre and Federico Ricci-Tersenghi, Physica A: Statistical Mechanics and its Applications **676**, 130871 (2025), doi: 10.1016/j.physa.2025.130871.

[2] Andrea Alessandrelli and Adriano Barra and Andrea Ladiana and Andrea Lepre and Federico Ricci-Tersenghi, arXiv preprint arXiv:2503.04454 (2025).

P25

Understanding Oversmoothing in Causal Transformers: The Role of Attention Sparsity

Cristina López Amado[1], **Peter Súkeník**[1], **Marco Mondelli**[1], **and Christoph H. Lampert**[1]

[1]*Institute of Science and Technology (ISTA)*

Oversmoothing in transformers refers to the tendency of token representations to become increasingly similar as information is propagated through self-attention layers. Although prior work has studied this phenomenon under various architectural assumptions [1, 2, 3, 4, 5], a general characterization of causal self-attention beyond initialization remains lacking. In particular, existing analyses do not fully account for the combined effects of normalization and residual connections under weak assumptions on the learnable matrices. As a result, the mechanisms governing token mixing in realistic causal transformers remain only partially understood. In this work, we investigate whether sparse attention patterns are necessary to prevent oversmoothing in causal transformers, or whether other components of the transformer block can counteract the homogenizing effect of attention.

We study vanilla single-head causal self-attention with RMSNorm and residual connections. To isolate the role of attention sparsity, we analyze a single attention step in which the attention matrix continuously interpolates between the identity map and uniform causal attention. Under a distributional model for the token representations, we derive an explicit approximation for the expected pairwise cosine similarity after the attention update. This approximation yields sufficient conditions under which increasing attention density either increases or decreases token similarity. In particular, our analysis shows that the value-output transformation can, in principle, counteract the mixing induced by uniform attention when it satisfies certain conditions.

We complement the theory with experiments on trained LLMs, including LLaMA3-8B, Gemma-7B, GPT2-XL, and Mistral-7B. Our results show that our theoretical approximation closely matches the empirical average cosine similarity across models and context lengths, supporting its use as a proxy for studying token mixing. Empirically, we find that head-specific value-output transformations do not, on average, counteract the mixing effect of attention. Moreover, the conditions identified by our theory under which mixing decreases with attention uniformity hold for only a small fraction of heads. The dominant trend in trained transformers is instead that token similarity increases with the degree of attention uniformity.

These findings contribute to a finer understanding of the role of the attention matrix in causal transformers and motivate further study of why trained models develop sparse attention patterns, such as attention sinks[6] and self-attending heads.

[1] Y. Dong, J.-B. Cordonnier, A. Loukas, in *Proceedings of the International Conference on Machine Learning*, 2793–2803 (2021).
[2] A. Giorlandino, S. Goldt, arXiv:2505.24333 (2025).
[3] X. Wu, A. Ajorlou, Y. Wang, S. Jegelka, A. Jadbabaie, Adv. Neural Inf. Process. Syst. **37**, 14774–14809 (2024).
[4] N. Karagodin, Y. Polyanskiy, P. Rigollet, Adv. Neural Inf. Process. Syst. **37**, 115652–115681 (2024).
[5] B. Geshkovski, C. Letrouit, Y. Polyanskiy, P. Rigollet, Bull. Am. Math. Soc. **62**, 427–479 (2025).
[6] X. Gu, T. Pang, C. Du, Q. Liu, F. Zhang, C. Du, Y. Wang, M. Lin, arXiv:2410.10781 (2024).

P26

The Geometry of Compositional Abstraction in Large Language Models

Kimberly Lopez[1], **Ziwen Pan**[1], **Pierce Zhang**[1], **Salma Ziadi**[1], **Samuel Lippl**[2], **and Ida Momennejad**[2]

[1]*Institute of Pure and Applied Mathematics (IPAM)*
[2]*Microsoft Research NYC*

While transformer-based Large Language Models (LLMs) excel at generating fluent text, they often struggle to complete tasks requiring multi-step reasoning. This work investigates the geometry of compositional abstraction in LLMs by conceptualizing solutions to a task as compositions of reusable reasoning steps. We analyze the performance of Llama-3-8B on a graph navigation task and characterize how steps of its reasoning manifest within hidden-state geometry. We then investigate whether algorithmic primitives correspond to linear directions in the activation space. Based on model performance, we estimate primitive-specific vectors from labeled contexts and intervene additively at an intermediate depth. Our results suggest that composing the last-node primitive and comparison primitive vectors steers Llama-3-8B toward correct decisions and improves accuracy relative to either intervention alone. This result supports a geometric understanding of reasoning where algorithmic steps occupy a salient direction in the latent space and can be operationalized to direct reasoning.

[1] O. Eberle et al., Proc. Mach. Learn. Res. **267**, 81292–81314 (2025).
[2] E. Todd et al., Int. Conf. Learn. Represent. (2024).
[3] S. Lippl et al., Int. Conf. Learn. Represent. (2026).

P27

Theoretical insights on Post-Training in Generalized Linear Models

Clément Loup-Forest[1], **and Bruno Loureiro**[1],

[1]*(Presenting author underlined) Département d'Informatique, École Normale Supérieure - PSL, CNRS, France*

Post-training has emerged as a critical phase in the development of Large Language Models (LLMs) [1, 2]. However, a rigorous theoretical understanding of this stage and its associated trade-offs remains limited. In this work, we address this gap by studying post-training within a mathematically tractable model of a mixture of linear regressions. In particular, we examine the role of the reward function and how its specification shapes the optimisation landscape during post-training.

A central focus of our analysis is the phenomenon of reward hacking, whereby a model maximises the reward signal independently of the underlying objective [3]. This issue is closely related to the broader problem of alignment, which similarly lacks a solid theoretical foundation.

Our results show that, even when pre-training exhibits mode-covering behaviour, post-training can still drive such collapse. To better understand this effect, we develop a phenomenological framework that identifies the key mechanisms enabling recovery of a specific teacher direction. Finally, we extend our analysis to characterise how imperfections in pre-training influence the dynamics of this collapse.

References

[1] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. Christiano, J. Leike, and R. Lowe, "Training language models to follow instructions with human feedback," 2022.

[2] R. Rafailov, A. Sharma, E. Mitchell, S. Ermon, C. D. Manning, and C. Finn, "Direct preference optimization: Your language model is secretly a reward model," 2024.

[3] L. Gao, J. Schulman, and J. Hilton, "Scaling laws for reward model overoptimization," 2022.

P28

Adaptive Collocation via Stochastic Solvers in PINNs

Stefano Magrini Alunno[1,2], **Lorenzo Pareschi**[2,3,4]

[1] *University of Studies of Modena and Reggio-Emilia,*
[2] *ADAMUS Project,*
[3] *University of Studies of Ferrara,*
[4] *Heriot Watt University*

Physics-Informed Neural Networks (PINNs) typically rely on uniformly sampled collocation points to enforce the governing PDE, which can be inefficient when solutions concentrate on low-measure regions, as often happens in kinetic and Fokker–Planck-type problems. We propose an adaptive collocation strategy in which training points are generated by a stochastic (Monte Carlo) solver associated with the PDE, exploiting the natural SDE representation to sample where the solution "lives". This yields a mass-aware estimate of the PDE residual and can be interpreted as importance sampling of the physics loss. To mitigate undersampling in low-density but potentially relevant regions (tails, boundary layers, normalization constraints), we introduce a hybrid scheme mixing stochastic-solver samples with uniform coverage points. We present preliminary experiments on Fokker–Planck equations, focusing on accurately capturing stationary states and comparing uniform, stochastic-driven, and hybrid collocation in terms of error on uniform test grids, moment accuracy, and distributional metrics. The approach is designed as a stepping stone toward more complex kinetic equations (e.g., Landau/Boltzmann), where particle-based solvers offer a pathway to scalable sampling.

P29

Can a Large Language Model do Chemistry?

Piyush Ranjan Maharana [1,2] , **Kavita Joshi** [1,2]

[1]***P****hysical and Materials Chemistry, CSIR-National Chemical Laboratory, Pune, 411008*
[2] *Academy of Scientific and Innovative Research (AcSIR), Ghaziabad 201002, India*

The rapid advancement in capabilities of Large Language Models (LLMs) has led to the exploration of their remarkable potential in other tasks that go beyond language processing. In this work, we try to understand LLM's performance on complex scientific tasks, particularly in chemistry. We are investigating the application of quantized LLMs for retrosynthesis, a critical challenge in organic chemistry. This includes fine-tuning LLMs on text based representations of reactant-product pairs from Chemical Reaction Database. The key focus being elucidating how LLMs solve retrosynthetic tasks and the impact of fine-tuning on base LLMs. Recognizing that text-based data alone is often insufficient for reproducing synthesis protocols, we augment our model with DFT-trained MLPs. This allows us to ground the LLM's predictions in quantum mechanics through rapid and accurate potential energy surface evaluations. Also we are exploring the use of these models, to autonomously navigate and explore the vast space of possible reaction pathways. Ongoing evaluations of model performance across diverse retrosynthetic challenges are currently being benchmarked against established baselines. By employing interpretability techniques to analyze internal representations of chemical structures, we are actively investigating how fine-tuning refines the model's underlying chemical intuition. Preliminary results are beginning to show measurable improvements, suggesting an evolving capacity for LLMs to internalize complex chemical transformations. This research continues to explore how the integration of LLMs with high-fidelity computational tools can ultimately accelerate automated synthesis and provide a more systematic means of navigating the vast, uncharted chemical space.

References:

[1] Jablonka, Kevin Maik, et al. "Leveraging large language models for predictive chemistry." *Nature Machine Intelligence* 6.2 (2024): 161-169.

[2] M. Bran, Andres, et al. "Augmenting large language models with chemistry tools." *Nature Machine Intelligence* 6.5 (2024): 525-535.

[3] Wu, Zhaofeng, et al. "The semantic hub hypothesis: Language models share semantic representations across languages and modalities." *arXiv preprint arXiv:2411.04986* (2024).

[4] Edamadaka, S., Yang, S., Li, J., & Gómez-Bombarelli, R. (2025). Universally converging representations of matter across scientific foundation models. *arXiv preprint arXiv:2512.03750*.

[5] van der Lingen, Rik (2025). Reaction SMILES CRD 1.37M dataset. figshare. Dataset https://doi.org/10.6084/m9.figshare.28230053.v1

P30

Bias Mitigation in Quantum Machine Learning: A Fairness Analysis of the COMPAS Dataset

Alessandro Marchetti[1], **Ludovica Ilari**[1], **Paolo Marino**[1], **Simona Tiribelli**[2], **and Filippo Caruso**[1]

[1]*Department of Physics and Astronomy, University of Florence, Sesto Fiorentino (FI), Italy*
[2]*Department of Political Science, Communication and International Relations, University of Macerata, Macerata (MC), Italy*

The deployment of machine learning (ML) systems in high-stakes domains—most notably criminal justice—has raised serious concerns about algorithmic fairness and bias. Tools such as the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) system have been shown to exhibit racially disparate error rates [1]. While classical bias-mitigation strategies address this at the pre-, in-, and post-processing levels, they remain constrained by classical computational paradigms. Quantum Machine Learning (QML) has emerged as a promising alternative, exploiting superposition, entanglement, and high-dimensional feature spaces to potentially regularize decision boundaries in novel ways [2].

This work presents, to our knowledge, the first systematic integration of classical bias-mitigation techniques—Reweighing (RW), Adversarial Debiasing (AD), and Equalized Odds (EO)—into a hybrid classical–quantum ML pipeline, evaluated on the COMPAS dataset for the `race`, `sex`, and intersectional `(race, sex)` attributes. We benchmark Variational Quantum Classifiers (VQC) and Quantum Support Vector Machines (QSVM) against classical baselines (logistic regression, linear and kernel SVMs). Fairness and performance are jointly assessed via empirical Pareto front analysis using balanced accuracy, statistical parity difference, disparate impact, and average odds difference. Statistical significance is evaluated through pairwise Wilcoxon signed-rank tests with Bonferroni correction across all model–mitigation combinations.

Our results reveal three key findings: (i) quantum models exhibit intrinsic fairness advantages without mitigation—QSVM achieves more favorable accuracy–fairness trade-offs than classical SVMs, and the VQC consistently yields the lowest fairness disparity; (ii) classical mitigation strategies transfer effectively to quantum models: EO and RW+EO produce fairness improvements comparable to classical baselines, with only marginal differences in balanced accuracy ($p \leq 0.05$) and no significant differences in fairness metrics; (iii) the intersectional setting reveals the strongest quantum advantage: under EO alone, the QSVM with amplitude encoding is the *only* model simultaneously Pareto-optimal and within the admissible fairness–performance region, whereas classical models require the stronger RW+EO strategy—which carries additional explainability costs [3].

These findings establish that an "ethics-by-design" approach is both feasible and advantageous for QML systems. The consistent Pareto-dominance of quantum models, particularly in intersectional scenarios, positions QML as a promising framework for bias-aware algorithmic decision-making. Future work will extend this analysis to healthcare and credit scoring, and investigate the theoretical mechanisms underlying quantum models' implicit fairness regularization.

[1] J. Angwin et al., *Machine Bias*, ProPublica (2016).
[2] J. Biamonte et al., Nature **549**, 195–202 (2017).
[3] D. Pessach and E. Shmueli, ACM Comput. Surv. **55**(3), 51 (2022).

P31

To be submitted at a later date.

P32

Large Deviations for Transformer-like Models

Luca Mazzucco[1], **Andrea Agazzi**[2]

[1] *SISSA*
[2] *University of Berna*

The subject of this work is the study of the Large Deviation Principle (LDP) for *Transformers*, a widely studied class of Deep Neural Networks. Large deviations theory provides a rigorous mathematical framework to describe the probabilities of rare events, and in the context of neural networks it offers a way to quantify fluctuations around their deterministic Gaussian process limits. Understanding these rare fluctuations is crucial, as they provide insights into the stability of learning dynamics and the universality properties of large models.

We consider the transformer-like architecture introduced in [1]. By assuming the query and key weights to be fixed or pre-trained, the architecture can be expressed as a Deep Linear Neural Network. The model is studied in the Bayesian framework, with a Gaussian prior on the weights and gaussian additive noise affecting the data. The analysis is performed in the large-scale limit, where the number N of neurons at each layer, the number of training samples P and the input dimension N_0 diverge, while their ratios P/N and P/N_0 converge to finite constants. With formal, non rigorous techniques, the covariance kernel under the posterior distribution is expressed through the minimizer of a functional called *action* in the previously mentioned large scale limit.

Our contribution consists in the partial framing of the results of [1] within the formalism of Large Deviations. In the specific setting where N, the width of each hidden layer, goes to infinity while the input dimension N_0 and the data set size P remain finite, we frame the result above in a large deviations context, identifying the action introduced as a rate function. Through this identification, we connect with the literature on large deviations for the covariance process.

[1] L. Tiberi, F. Mignacco, K. Irie, and H. Sompolinsky, *Dissecting the Interplay of Attention Paths in a Statistical Mechanics Theory of Transformers*, Adv. Neural Inf. Process. Syst. **37**, 72710–72753 (2024).

P33

Sheaf Neural Networks and Biomedical Applications

A. Merhab[1], **J.-W. Van Looy**[2], **P. Demurtas**[2], **S. Iotti**[2], **E. Malucelli**[2], **F. Rossi**[2], **F. Zanchetta**[2], **R. Fioresi**[2]

[1] *University of Ferrara, Italy*
[2] *University of Bologna, Italy*

Graph Neural Networks (GNNs) provide an effective framework for learning from relational data, yet their expressivity is constrained by the classical graph Laplacian $L = D - A$. This work introduces Sheaf Neural Networks (SNNs), which generalize GNNs by equipping a graph $G = (V, E)$ with a sheaf $\mathcal{F}$ assigning vector spaces $\mathcal{F}(v)$ to nodes and linear restriction maps $\mathcal{F}_{v \leq e} : \mathcal{F}(v) \to \mathcal{F}(e)$ to edges [1]. The resulting sheaf Laplacian $L_{\mathcal{F}} = \delta^{\top}\delta$, where δ is the coboundary operator, enables a richer message passing mechanism that captures complex node interactions beyond adjacency structure [2]. We validate this framework on a biomedical dataset of synchrotron-based XANES spectroscopic measurements from osteosarcoma patients, modeled as a cosine similarity graph with PCA reduced node features. The SNN model achieves a majority vote accuracy of 98.66%, outperforming GCN (76.34%), GAT (78.12%) and GraphSAGE (92.86%), demonstrating how sheaf theoretic structure can yield more expressive and robust learning on complex biological networks.

[1] C. Bodnar et al., *Neural Sheaf Diffusion: A Topological Perspective on Heterophily and Oversmoothing in GNNs*, ArXiv abs/2202.04579 (2022).

[2] J. Hansen, R. Ghrist, *Toward a Spectral Theory of Cellular Sheaves*, J. Appl. Comput. Topology **3**, 315 (2019).

P34

Generalization of Gibbs and Langevin Monte Carlo Algorithms in the Interpolation Regime

Andreas Maurer[1], Erfan Mirzaei[1], Massimiliano Pontil[1]

[1](Presenting author underlined) Istituto Italiano di Tecnologia

Modern learning algorithms possess the capacity to achieve near-zero training errors even on arbitrary or completely impossible data, such as datasets with randomly assigned labels. In this overparameterized interpolation regime, observing a small training error within a massive hypothesis space is no longer sufficient to predict test error. The key to generalization is buried deeper in the data. This mystery extends to algorithms that approximate the Gibbs posterior, a distribution where probabilities decrease exponentially with the training error, governed by an inverse temperature parameter β.

For practical stochastic algorithms like Stochastic Gradient Langevin Dynamics (SGLD) and broader Langevin Monte Carlo (LMC) methods, the Gibbs measure serves as a limiting distribution. Existing theoretical literature [1] provides generalization bounds for these methods that typically scale with β/n or $\sqrt{\beta/n}$, where n is the sample size. However, these bounds are only effective in the high-temperature regime ($\beta < n$). In the low-temperature regime ($\beta > n$), where the algorithms successfully interpolate the data, existing bounds apply equally to true and random labels, thereby becoming entirely vacuous.

This research addresses this critical gap by providing high-probability, data-dependent bounds on the expected error of the Gibbs algorithm that remain nontrivial in the low-temperature interpolation regime. The central message of this research is that generalization at lower temperatures is already fundamentally signaled by the behavior of the model at high temperatures. We uncover a simple but novel analytical mechanism: by leveraging the PAC-Bayesian theorem [2], we show that the log-density of the Gibbs posterior can be explicitly expressed as an integral from 0 to β of the mean training errors. Consequently, the generalization bound is proportional to the area under this curve. Because the area accumulates across all temperatures, a smaller mean training error in the high-temperature regime directly translates to a tighter generalization bound in the low-temperature regime, neatly separating meaningful learning from mere memorization.

We outline three primary contributions. First, we establish rigorous bounds on the true error for both a drawn hypothesis and the posterior mean across the entire temperature spectrum. Second, we prove that these bounds are stable under approximations of the Gibbs posterior in relative entropy, allowing them to be applied to practical LMC algorithms. Finally, to bridge the gap between theoretical guarantees and practical computation, we introduce a heuristic calibration step based exclusively on the training data. Applying this method to overparameterized neural networks trained via SGLD on the MNIST and CIFAR-10 datasets, we demonstrate that our bounds yield remarkably close predictions of the test error for true-labeled data, while correctly maintaining a valid, high upper bound for randomly labeled data.

[1] M. Raginsky, A. Rakhlin, and M. Telgarsky. Non-convex learning via Stochastic Gradient Langevin Dynamics: a nonasymptotic analysis. Conference on Learning Theory (COLT), 2017.

[2] D. A. McAllester. PAC-Bayesian Model Averaging. Conference on Computational Learning Theory (COLT), 1999.

P35

Shape and Texture Analysis of Satellite Imagery for Weather Prediction

Kristina Moen [1] , Emily J. King [1] , and Imme Ebert-Uphoff [2,3]

[1]Colorado State University Department of Mathematics
[2] Cooperative Institute for Research in the Atmosphere (CIRA)
[2] Colorado State University Department of Electrical and Computer Engineering

Weather satellite imagery is critical for severe weather prediction. The increasing amount and complexity of satellite data, together with growing societal demand for timely and accurate weather information, drive the need for automated methods that identify and analyze weather events in satellite imagery. This research develops well-defined and theoretically-grounded mathematical methods that quantify meaningful shapes and textures in satellite imagery for use in weather prediction models. Working closely with atmospheric scientists at Colorado State University and NOAA, we focus on two severe weather applications: (1) detecting convective cloud regions in high-resolution visible imagery using texture analysis tools such as the Gray Level Co-Occurrence Matrix method; and (2) characterizing tropical cyclone convective structure and eye formation in a newly developed dataset of synthetic passive microwave imagery using tools from geometry and topology. By combining mathematics with atmospheric science domain knowledge, this work aims to provide interpretable and practical tools for satellite image analysis that will ultimately support deeper understanding of severe weather processes and accurate prediction.

P36

Markov Random Field Information Bottleneck for Latent Conditional Diffusion

Merijn Moody[1], **Marco Federici**[2] **Patrick Forré**[1]

[1]*University of Amsterdam, Amsterdam, The Netherlands* [2]Qualcomm AI Research, Amsterdam, The Netherlands

Accurately simulating large-scale dynamic systems at long time scales is computationally expensive due to the short time steps required for integration [2]. Applying standard Information Bottleneck (IB) techniques directly to large spatial systems such as fluid simulations and weather data is computationally intractable due to the sheer dimensionality of the data. To address this, we introduce the spatiotemporal Markov Random Fields Information Bottleneck (MRFIB), which leverages the spatial structure of the data to make IB optimization feasible. For an MRF defined over a spatiotemporal grid, we formalize autoinformation as $AI(x_v;\tau) = I(x_v; x_{\partial_\tau v})$, where $\partial_\tau v$ denotes the Markov blanket of node v at a time lag τ. We define the autoinformation gap for a latent representation z_v as $AIG(z_v;\tau) := AI(x_v;\tau) - AI(z_v;\tau)$.

We prove that if a latent representation z_v preserves this autoinformation (i.e., $AIG(z_v;\tau) = 0$), it carries rigorous theoretical guarantees: it acts as a minimally sufficient representation for all downstream predictive tasks at that scale, and the latent encoding itself structurally forms a Markov Random Field. An optimal representation minimizes this gap, preserving essential system dynamics while actively discarding superfluous, high-frequency information.

Building on these theoretical guarantees, we implement this latent simulation inference scheme for the probabilistic prediction of turbulent fluid flows using latent diffusion models. We leverage the MRFIB objective to map complex observations into a simplified representation space Z. We decompose the conditional score function as:

$$\nabla_z \log p(z_{t+1}|x_t) = \nabla_z \log p(z_{t+1}) + \nabla_z f(x_t, z_{t+1}) \tag{1}$$

where f is a mutual information critic trained via contrastive estimation [1]. Contrastive methods, such as InfoNCE, provide a flexible and tractable upper-bound for maximizing time-lagged mutual information [3]. This latent diffusion scheme promises *robust* evolution in the face of noise and *efficient* compression allowing for fast latent evolution.

[1] J. Ho, A. Jain, P. Abbeel, NeurIPS **33**, 6840 (2020).
[2] M. Federici, P. Forré, B. S. Veeling, R. Tomioka, ICLR (2024).
[3] A. van den Oord, Y. Li, O. Vinyals, arXiv:1807.03748 (2018).

P37

Compressibility and scaling laws in linear-width neural networks with hierarchical features: an information-theoretic perspective

Minh-Toan Nguyen[1], **Jean Barbier**[1],

[1] *ICTP, Trieste, Italy*

We study the transferring of knowledge from a teacher model to a student model of equal or smaller size, trained on input-output pairs generated by the teacher. Using a teacher with hierarchical features, we first explore experimentally how data size and model size jointly affect the student's performance under SGD training. The experiments reveal three main findings: a critical model size beyond which performance degrades, a scaling law that holds up to this threshold, and two distinct scaling laws in the feature learning regime and fine-tuning regime. We explain these phenomena from an information-theoretic perspective, introducing a key concept – the teacher's *effective size* – which unifies the two scaling regimes and provides a principled account of the empirical scaling behavior observed in SGD-trained students.

P38

On the Role of Activation Functions in the Storage Capacity of Neural and Associative Memory Models

F. Nicolai[1], **Y. Roudi**[1], **and E. Agliari**[2]

[1] *(Presenting author underlined) Department of Mathematics, King's College London, London, UK*

[2] *Dipartimento di Matematica, Sapienza Università di Roma, Rome, Italy*

Recent advances in machine learning have highlighted the need for a deeper theoretical understanding of the mechanisms governing the performance of neural networks. In particular, the role of activation functions in determining learning and memory properties remains only partially understood.

In this work, we study the storage capacity of neural networks using a Gardner-type approach. Starting from [1], which considers feedforward architectures with threshold-linear input–output relations, we extend the analysis to general transfer functions, including both thresholded and non-thresholded cases. In addition, we incorporate output error tolerance and input noise, following the framework introduced in [2]. Building on this setting, we then consider recurrent neural networks, where memories are associated with fixed points of the dynamics. We extend the capacity calculation to this case, focusing on situations in which fixed points lie within a small tolerance ϵ of prescribed activity patterns.

Our theoretical framework allows us to compute the storage capacity for generic transfer functions and pattern distributions, while explicitly accounting for tolerance effects. We show how these factors influence the capacity and stability of stored patterns, and we characterize the dependence of the capacity on the choice of activation function. In particular, our analysis provides a framework to identify transfer functions that optimize storage capacity under given constraints. The analytical predictions are compared with numerical simulations in order to test the theoretical assumptions and assess the validity of the approach.

These results provide insight into the role of activation functions in high-dimensional learning systems and represent a step toward a more systematic theoretical understanding of memory and learning in neural networks.

[1] F. Schönsberg, Y. Roudi, and A. Treves, Physical Review Letters **126.1** (2021): 018301.

[2] D. Bollé, R. Kuhn, and J. Van Mourik, Journal of Physics A: Mathematical and General **26.13** (1993): 3149-3158..

P39

Categorical coding beyond binary neuronal states enables memory transitions: a DMFT analysis of Potts associative network

Carlo Orientale Caputo[1], **Anindita Basu**[1], **Fernando Lucas Metz**[2], **Alessandro Treves**[1]

[1]*SISSA - Institute of Advanced Studies, Trieste, Italy*
[2]*Physics Institute – Federal University of Rio Grande do Sul, Porto Alegre, Brazil*

Spontaneous transitions between memories—observed in spontaneous cognition such as mind-wandering and creative free associations—suggest the existence of neural mechanisms capable of autonomously destabilizing a retrieved memory state and drifting toward another. We ask what minimal representational complexity is required for such transitions to emerge in associative neural networks.

Using dynamical mean-field theory (DMFT) [1], we analyze multi-state Potts networks with slow adaptive thresholds [2] in the extensive-load regime, where the number of stored patterns scales with system size. To isolate transitions between two memories, we fix two patterns in the connectivity matrix and perform the disorder average over the remaining ones. This yields an effective dynamical theory in which the high-dimensional network reduces to a small number of representative degrees of freedom.

We show that, in this limit, network units organize into five distinct dynamical classes. The dynamics of each class evolves independently once the mean field is specified, which must be determined self-consistently. Analyzing the resulting equations in the presence of two fixed patterns, we derive the conditions under which a system initialized near one memory can transition to the other.

Importantly, we find that this transition mechanism is suppressed in the binary case, indicating that multi-state Potts representations are dynamically required to sustain transitions between memory states. Our results provide a theoretical link between representational complexity and the emergence of spontaneous cognitive dynamics in associative networks.

[1] A. C. C. Coolen, in *Handbook of Biological Physics*, vol. 4, eds. F. Moss and S. Gielen, North-Holland, pp. 619–684 (2001).
[2] E. Russo, A. Treves, Phys. Rev. E **85**, 051920 (2012).

P40

Critical capacity of shallow neural networks in the interpolation asymptotics: a statistical mechanics analysis

ICTP-INdAM-SLMath Summer Graduate School for Machine Learning (smr 4222)

Mauro Pastore [1] , **Jean Barbier** [1]

[1] *(Presenting author underlined) ICTP*

Statistical learning theory suggests that neural networks with many parameters should be able to fit almost arbitrary data. But how much of this expressive power is actually realized in typical high-dimensional settings? In this contribution, we revisit this question through the lens of statistical mechanics. We consider a shallow neural network with width comparable to the input dimension, trained on Gaussian data with random labels. In the so-called interpolation regime, the number of parameters and constraints scale together, placing the model at the edge of feasibility. We show that the problem undergoes a sharp SAT/UNSAT transition, analogous to classical results for perceptrons and committee machines. Using the replica method, we estimate the critical capacity and study how it depends on architectural choices. Moreover, a replica symmetry breaking transition occurs in the SAT phase, signalling that the solution space splits into typically non-connected clusters. This anlysis highlights the importance of typical-case analysis in understanding neural network expressivity.

P41

MASKED CONSENSUS-BASED BAYESIAN FACTORIZATION FOR A MULTI-OMICS TENSOR

Ilaria Pellegrinelli[1]

[1] Department of Mathematics, University of Trento, Italy

The analysis of multi-dimensional biological datasets presents significant challenges, both from a mathematical and a biological perspective: sparsity interferes with the convergence of optimization procedures, while the inherent difficulty of collecting biological measurements introduces systematic missingness.

The goal of this work is to find a robust mathematical procedure to stratify clinical patients into clusters based on type of disease and severity, handling the problems related to the nature of data.

Standard tensor factorizations offer a powerful framework to model multi-dimensional interactions and express them in a computationally efficient and easily interpretable way; however, their performance degrades rapidly as sparsity increases, highlighting the need for a new approach tailored for the structural sparsity. To address this limitation, we propose a Masked Consensus-based Bayesian tensor factorization.

Rather than relying on noise-inducing data imputation or information-losing sample filtering, our approach introduces a binary tensor mask to evaluate the likelihood strictly on the true observed values. To properly account for structural zeros and the bounded nature of the similarity scores, we model the generative process of the data via a Zero-Inflated Truncated Gaussian likelihood, formulated as a mixture of a Bernoulli point mass at zero and a continuous Truncated Gaussian distribution. To select empirically the optimal number of factors and overcome the instability caused by possible convergence to local optima, we employ a consensus-based approach, aggregating multiple restarts via K-means clustering on the latent factors of each modality of the tensor.

Applying this framework to a COVID-19 multi-omics clinical dataset, we robustly identify three distinct clinical phenotypes. The proposed tensor factorization effectively captures a global cytotoxic antiviral response and isolates a severe thrombo-inflammatory signal from a functional, standard inflammatory response. By mathematically disentangling healthy immune mobilization from pathological coagulopathy, this model facilitates the independent targeting of specific disease mechanisms, suggesting a broader applicability of the framework to any sparse, zero-inflated multi-omics setting.

[1] D. Chafamo, V. Shanmugam, N. Tokean, arXiv:2308.08060 (2023).
[2] D. J. Ahern et al., Cell 185, 1-22 (2022).

P42

Laboratory name: Institut de Physique Théorique (IPhT). CNRS id code: UMR 3681
Directors: J. FERNANDEZ DE COSSIO DIAZ & P. URBANI
e-mail: jorge.fdcd@ipht.fr, pierfrancesco.urbani@ipht.fr
web page: jorgefdcd, pierfrancesco-urbani
Thesis possibility after internship: Yes

Controlling and understanding representations in state-space models of biological sequences

State-space models process sequence data one token at a time. A state vector is updated with each consumed token, and stores a compressed “memory” of past tokens. To be effective, the model must be able to recognize and exploit long-range correlations between tokens along the sequence. This is a particularly important issue in applications of these models to biological sequences, such as proteins or RNA, where sites may be far-apart along the sequence, yet interact closely in the folded 3D structure of the molecule. Since the latent state vector encodes all the information that the model needs to construct future tokens of the sequence, one can try to find ways of controlling this latent vector during generation to modify aspects of the sequence, *e.g.*, to manipulate desired properties of an artificial protein.

We will attack these issues from both a computational and a theoretical side. On a computational side, we will apply state-space models (such as MAMBA [3]) to biological sequence data, and develop algorithms to manipulate state representation vectors to design sequences with desired properties (*e.g.* through disentangled representations [1] or guidance [5]). Designed molecules can be tested experimentally with collaborators [4]. On a theoretical side, we will formulate simplified models and exploit methods from statistical physics (e.g., dynamical mean-field theory [2]), aiming at understanding how these models learn about long-range interactions and how token representations can be controlled during generation to modify properties of designed sequences.

This interdisciplinary internship and PhD project are aimed at students with a strong background and interest in statistical physics, machine learning, biology, and coding.

References

[1] JFdCD, S. Cocco, R. Monasson. "Disentangling representations in restricted Boltzmann machines without adversaries." Physical Review X 13.2 (2023): 021003
[2] P.J. Kamali, PU. (2023) “Dynamical mean field theory for models of confluent tissues and beyond,” *SciPost Physics*, 15(5), p. 219. Available at: https://doi.org/10.21468/SciPostPhys.15.5.219.
[3] D. Sgarbossa, et al. (2025) “ProtMamba: a homology-aware but alignment-free protein state space model,” Bioinformatics, 41(6), p. btaf348.
[4] JFdCD, et al. “Designing molecular RNA switches with Restricted Boltzmann machines” Nat. Comm. (2025, in Press). bioRxiv 2023.05.10.540155.
[5] A. Bansal, et al. "Universal guidance for diffusion models." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023.

P43

Phase Transition in the Ising Model on (m,k)-Ary Trees

Aminah Qawasmeh[1], **Farrukh Mukhamedov**[1]

[1]*Department of Mathematical Sciences, College of Science,*
United Arab Emirates University, P.O. Box 15551, Al Ain, Abu Dhabi, UAE

The Ising model is a fundamental framework in statistical mechanics for understanding phase transitions, such as the change from a magnetized state to a non-magnetized one. It is also relevant in machine learning, where related models appear in probabilistic graphical models and energy-based learning. Although the Ising model has been widely studied on regular trees, its behaviour on trees with varying branching patterns remains less explored. In this poster, we study the Ising model on (m, k)-Ary trees, where the number of branches changes across levels. We present new results on phase transitions in this setting and highlight how the structure of the tree affects the model's behaviour. Our work offers further insight into how hierarchical geometry shapes collective behaviour in both physical and learning systems.

[1] H.O. Georgii, Gibbs measures and phase transitions, de Gruyter studies in mathematics, Vol. **9**, (1988).
[2] U. Rozikov, Gibbs measures on Cayley trees, World Scientific, Hackensack, IJ, (2013).

P44

A Fourier perspective on the learning dynamics of neural networks: from sample complexities to mechanistic insights

F. Ricci, C. Merger, and S. Goldt

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

Neural networks trained with gradient-based methods exhibit a strong simplicity bias, learning simpler statistical features of their data before moving to more complex features. In this work, we study this bias from a Fourier perspective, motivated by the approximate translation-invariance and the characteristic power spectra of natural images. We first show experimentally that simple neural networks trained on image classification tasks first rely on amplitude information – related to pair-wise correlations between pixels – before exploiting phase information, which encodes edges and higher-order correlations. To explain this phenomenon, we introduce a synthetic data model for translation-invariant inputs that allows precise control over the amplitudes and phases while remaining tractable. We rigorously establish that for isotropic and high-dimensional inputs, classifying them by relying only on phase information is a genuinely hard task: online stochastic gradient descent (SGD) cannot distinguish the structured inputs from noise within $n \ll N^3$ steps, but needs at least $n \gg N^3 \log^2 N$ steps. In contrast, we show both experimentally and theoretically that, for non-isotropic inputs with power-law spectra, the existence of a dominant principal subspace can dramatically accelerate the speed of learning, even if the Fourier amplitudes are shared among classes and do not help with classification. Simulations with two-layer networks trained on textures and with deep convolutional networks on ImageNet confirm this non-trivial interaction between amplitudes and phases, providing mechanistic insights into how deep neural networks can learn natural image distributions efficiently.

P45

Epidemic thresholds and transient dynamics in multi-type SIR and SIS models

Gabriele Ricci[1]

[1](*Presenting author underlined) University of Rome Tor Vergata, Rome (Italy)*

Behavioral heterogeneity (e.g., different levels of caution, risk perception, or exposure to misinformation) induces type-dependent susceptibility and infectivity, while network effects (degree heterogeneity and homophily) reshape both epidemic thresholds and early-time spreading patterns. Motivated by these mechanisms, we study two-type spreading processes on random graphs and derive analytical conditions that predict the onset of large outbreaks and explain transient phenomena.

We first consider a multi-type SIR model on a two-type network with heterogeneous mixing [1]. By mapping the final outbreak size to bond percolation, we express the dynamics in terms of type-dependent edge transmissibilities, i.e., transmissibilities that depend on the types at the two ends of each edge, extending the percolation framework in [2]. We then derive a self-consistent percolation fixed-point equation for the probability of reaching the giant infected component. This yields an exact epidemic threshold condition, and a linearized approximation that is accurate near criticality. To study the initial phase, we derive two early-time growth thresholds: one based on the prevalence and one based on an effective prevalence, which accounts for the presence of a more infectious node type. The region between these two thresholds defines a "resurgence zone", where the prevalence can initially decrease while the effective prevalence grows. After a short transient, the prevalence starts increasing, driven by the growth of the effective prevalence of the more infectious type.

We then study the multi-type SIS process starting from the standard QMF framework and introduce two reduced spectral thresholds to separate macroscopic from non-macroscopic endemic regimes. PIB is a model-based reduction that requires an assumption on the underlying network ensemble. In contrast, our compressed-quenched-mean-field criterion is model-free: we take the weighted adjacency matrix B of the quenched network and compress it onto the typed-degree space, obtaining $C(B)$. The leading eigenvalue of $C(B)$ provides a practical threshold estimate for macroscopic versus non-macroscopic endemic regimes. These results can be extended to m different types.

Finally, we validate the theoretical predictions using stochastic simulations on random graphs. For the SIS process, we use quasi-stationary simulations to estimate the stationary prevalence and its variability, and to assess how well the spectral thresholds predict the observed transition. Overall, the results provide a unified picture of how heterogeneity and network structure shape threshold behavior and transient effects in multi-type SIR/SIS models, suggesting that low-dimensional spectral compressions can capture threshold and transient behavior in heterogeneous diffusion processes on networks.

[1] F. Mazza, G. Ricci, F. Colaiori, S. Guarino, S. Meloni, F. Saracco, Phys. Rev. E **113**, 014306 (2026).
[2] J. C. Miller, Phys. Rev. E **76**, 010101(R) (2007).

P46

Provable False Discovery Rate Control for Deep Feature Selection

Kazuma Sawaya

The University of Tokyo

We develop a flexible feature selection framework based on deep neural networks that approximately controls the false discovery rate (FDR), a measure of Type-I error. The method applies to architectures whose first layer is fully connected. From the second layer onward, it accommodates multilayer perceptrons (MLPs) of arbitrary width and depth, convolutional and recurrent networks, attention mechanisms, residual connections, and dropout. The procedure also accommodates stochastic gradient descent with data-independent initializations and learning rates. To the best of our knowledge, this is the first work to provide a theoretical guarantee of FDR control for feature selection within such a general deep learning setting.

Our analysis is built upon a multi-index data-generating model and an asymptotic regime in which the feature dimension n diverges faster than the latent dimension q^*, while the sample size, the number of training iterations, the network depth, and hidden layer widths are left unrestricted. Under this setting, we show that each coordinate of the gradient-based feature-importance vector admits a marginal normal approximation, thereby supporting the validity of asymptotic FDR control. As a theoretical limitation, we assume $\boldsymbol{B}$-right orthogonal invariance of the design matrix, and we discuss broader generalizations. We also present numerical experiments that underscore the theoretical findings.

This paper (Sawaya, 2025) is accepted for AISTATS 2026 in Tangier, Morocco.

References

Sawaya, K. (2025) Provable fdr control for deep feature selection: Deep mlps and beyond, *arXiv preprint arXiv:2512.04696*.

P47

Escaping plasticity-memory dilemma in intelligent robotic systems

Rajat Saxena [1], Onurcan Bektaş [1] and Steffen Rulands [1]

[1] Ludwig-Maximilians-Universität München, Arnold-Sommerfeld-Center for Theoretical Physics, Theresienstraße 37, 80333 München, Germany

Artificial intelligence systems are increasingly deployed in settings that require continual learning, particularly for robots operating in complex and evolving environments. Such systems must balance the acquisition of new skills with the retention of previously learned knowledge. In standard single-neural-network-based robots, this balance is constrained by the plasticity-memory dilemma, in which learning new tasks often leads to forgetting already learned ones; this is known as catastrophic forgetting. Here, we study role specialisation in interacting multi-robot systems as an emergent form of collective memory, introducing a stochastic interaction mechanism, inspired by [1], that drives symmetry breaking and enables robots to dynamically differentiate into specialised roles, thereby providing a distributed mechanism to overcome the dilemma. I also present the hyperparameter phase space, which reveals different learning regimes and configurations in which specialisation emerges. These results suggest that memory and plasticity in learning systems can emerge from collective interactions rather than individual models, offering a statistical physics framework for overcoming fundamental limitations such as catastrophic forgetting.

[1] Patalano, Solenn, et al. "Self-organization of plasticity and specialization in a primitively social insect." Cell Systems 13.9 (2022): 768-779.

P48

Phase behavior of a locally disordered non-conserving exclusion process with finite resources

Nisha Sharma[1], **Bipasha Pal**[2], **Ankita Gupta**[3], **and Arvind Kumar Gupta**[1]

[1]*(Presenting author underlined) Department of Mathematics, Indian Institute of Technology Ropar, Rupnagar-140001, Punjab, India*

[2]*Indian Institute of Information Technology Design and Manufacturing, Kancheepuram, Chennai - 600127, Tamil Nadu, India*

[3]*Mathematical Biophysics Group, Max Planck Institute for Multidisciplinary Sciences, D-37077 Göttingen, Germany*

Motivated by stochastic transport processes in biological systems [1], we investigate a locally disordered totally asymmetric simple exclusion process with Langmuir kinetics under finite resources. The disorder is represented by a dynamic defect that can attach to or detach from a specific site and reduces particle hopping when present. Using mean-field analysis together with numerical computations, we characterize steady states through density profiles and phase diagrams in the α–β plane, revealing diverse phase regimes.

We study the influence of finite resources and Langmuir kinetics [3, 4] by varying the filling factor and binding constant, which express threshold values where the phase structure changes qualitatively. To describe the impact of dynamic disorder, we define an obstruction factor that combines defect density and slowdown rate [2]. Larger obstruction values lead to simpler phase diagrams, suppressing Meissner phases, while favoring high density and shock states. All theoretical findings are confirmed through Monte Carlo simulations implemented via the Gillespie algorithm, emphasizing links between stochastic modeling, statistical physics, and computational methods relevant to machine learning.

[1] C. T. MacDonald, J. H. Gibbs, and A. C. Pipkin, Biopolymers **6**, 1 (1968).
[2] B. Pal and A. K. Gupta, Chaos, Solitons & Fractals **162**, 112471 (2022).
[3] A. Parmeggiani, T. Franosch, and E. Frey, Phys. Rev. E **70**, 046101 (2004).
[4] B. Pal and A. K. Gupta, Chaos, Solitons & Fractals **153**, 111517 (2021).

P49

Harmful Overfitting in Sobolev Spaces

Kedar Kardharkar[*1], **Alexander Sietsema**[*1], **Deanna Needell**[1], **and Guido Montufar**[1,2]

[1]*(Presenting author underlined) Department of Mathematics, University of California, Los Angeles, Los Angeles, CA, USA*

[2]*Department of Statistics & Data Science, University of California, Los Angeles, Los Angeles, CA, USA*

[*] equal contribution

Motivated by recent work on benign overfitting in overparameterized machine learning, we study the generalization behavior of functions in Sobolev spaces $W^{k,p}(\mathbb{R}^d)$ that perfectly fit a noisy training data set. Under assumptions of label noise and sufficient regularity in the data distribution, we show that approximately norm-minimizing interpolators, which are canonical solutions selected by smoothness bias, exhibit *harmful overfitting*: even as the training sample size $n \to \infty$, the generalization error remains bounded below by a positive constant with high probability. Our results hold for arbitrary values of $p \in [1, \infty)$ in contrast to prior results studying the Hilbert space case $(p = 2)$ using kernel methods. Our proof uses a geometric argument which identifies harmful neighborhoods of the training data using Sobolev inequalities.

P50

Statistical physics of deep learning: Optimal learning of a multi-layer perceptron near interpolation

Jean Barbier[1], **Francesco Camilli**[2], **Minh-Toan Nguyen**[1], **Mauro Pastore**[1] **Rudy Skerk**[3]

[1] *The Abdus Salam International Centre for Theoretical Physics*
Strada Costiera 11, 34151 Trieste, Italy
[2] *Alma Mater Studiorum - Università di Bologna, Dipartimento di Matematica*
Piazza di Porta S. Donato 5, 40126 Bologna, Italy
[3]*(Presenting author underlined) International School for Advanced Studies*
Via Bonomea 265, 34136 Trieste, Italy

For four decades statistical physics has been providing a framework to analyse neural networks. A long-standing question remained on its capacity to tackle deep learning models capturing rich feature learning effects, thus going beyond the narrow networks or kernel methods analysed until now. We positively answer through the study of the supervised learning of a multi-layer perceptron. Importantly, (i) its width scales as the input dimension, making it more prone to feature learning than ultra wide networks, and more expressive than narrow ones or ones with fixed embedding layers; and (ii) we focus on the challenging interpolation regime where the number of trainable parameters and data are comparable, which forces the model to adapt to the task. We consider the matched teacher-student setting. Therefore, we provide the fundamental limits of learning random deep neural network targets and identify the sufficient statistics describing what is learnt by an optimally trained network as the data budget increases. A rich phenomenology emerges with various learning transitions. With enough data, optimal performance is attained through the model's "specialisation" towards the target, but it can be hard to reach for training algorithms which get attracted by sub-optimal solutions predicted by the theory. Specialisation occurs inhomogeneously across layers, propagating from shallow towards deep ones, but also across neurons in each layer. Furthermore, deeper targets are harder to learn. Despite its simplicity, the Bayes-optimal setting provides insights on how the depth, non-linearity and finite (proportional) width influence neural networks in the feature learning regime that are potentially relevant in much more general settings.

P51

OPTIMAL SOLUTIONS TO STOCHASTIC DIFFERENTIAL EQUATIONS

Peace Udoka Success
Department of Mathematics, University of Ibadan

Stochastic differential equations (SDEs) provide a natural framework for modeling dynamical systems influenced by randomness. In my undergraduate thesis, I studied the analytical characterization of optimal solutions to SDEs within a rigorous mathematical framework. The project focused on establishing existence and optimality conditions for controlled stochastic systems using tools from measure theory, probability theory, and functional analysis.

The analysis considered Itô type stochastic processes and examined admissible control strategies under uncertainty. Emphasis was placed on proof-based derivations of necessary optimality conditions and stability properties of solutions. This work strengthened my understanding of stochastic processes, variational principles, and mathematical structures underlying decision-making in uncertain environments.

The study of SDEs is closely related to modern machine learning, particularly in areas involving probabilistic modeling, stochastic optimization, and uncertainty quantification. My work has motivated a broader interest in the mathematical foundations of learning systems, where principled treatment of randomness plays a central role. I am particularly interested in how stochastic analysis informs Bayesian methods and optimization techniques used in contemporary machine learning.

P52

A practical examination of uncertainty quantification approaches in machine learning for materials science

Mina Taleblou[1], Stefano de Gironcoli[1]

[1] *Scuola Internazionale Superiore di Studi Avanzati (SISSA)*

The adoption of Machine Learning (ML) models across various fields has been expanded rapidly. Remarkable improvements in ML models have been done on accuracy, efficiency, and scalability. However, the uncertainty of prediction introduced by stochastic nature of ML models and training processes, has been addressed to a lesser extent. Hence, reliable measurement of a model's confidence, or Uncertainty Quantification (UQ), remains an active area of research. This critical quantity must be quantified and managed to ensure reliable conclusions [1].

In this study, we trained and tested multiple ML models on diverse datasets, focusing on materials simulation tasks. Then, we evaluated several leading UQ approaches systematically, assessing their ability to capture true model uncertainty. We analyzed a range of common statistical and information-theoretic methods, benchmarking their performance against the Query By Committee (QBC) [2] as a reference standard.

Our findings suggest a moderate correlation between the statistical estimates and true uncertainty values. Further, the results underscore the inconsistencies among UQ metrics, a challenge previously identified in literatures [3]. This study monitors these inconsistencies across different models and under data shift, and confirms the limitations in current scoring frameworks and their interpretability in comparing UQ metrics.

Despite their imperfect correlation with true uncertainty, statistical UQ methods still prove effective as selection tools for active learning (AL) and fine-tuning, enabling the identification of informative and out-of-distribution data points. This targeted data inquiry enhances model generalization and confidence compared to random sampling.

Our work highlights the need for more robust, standardized UQ methodologies in ML for materials science and provides practical insights for improving model reliability in real-world applications via AL.

[1] Grasselli F, Chong S, Kapil V, Bonfanti S, Rossi K. Digit. Discovery. **4**, 2654 (2025).
[2] Parker, Wendy S, Wiley Interdiscip. Rev. Clim. Change. **4**, 213 (2013).
[3] Rasmussen, Duan, Kulik, Jensen. J. Cheminform. **15**, 121 (2023).

P53

Does Data Structure Affect Learning in Restricted Boltzmann Machines?

Robin Thériault[2], **Francesco Tosello**[1], **and Daniele Tantari**[3]

[1]*Department of Mathematics, The University of British Columbia*
[2]*Scuola Normale Superiore di Pisa*
[3]*Department of Mathematics, University of Bologna*

Restricted Boltzmann Machines (RBM) are generative models capable of learning data with a rich underlying structure. To understand how does data structure affect the learning dynamics, we study the teacher–student setting, where a student RBM learns data generated by a teacher RBM. The amount of structure in the data is controlled by adjusting the number of hidden units of the teacher and the correlations in the rows of the weights, a.k.a. patterns. Previous studies considered at most two hidden units, we explored the setting with finitely-many hidden units.

In the absence of correlations, we validate the conjecture that the performance is independent of the number of teacher patterns and hidden units of the student RBMs. Beyond this regime, we find that the critical amount of data required to learn the teacher patterns decreases with both their number and correlations. In both regimes, we find that, even with a relatively large dataset, it becomes impossible to learn the teacher patterns if the inference temperature used for regularization is kept too low.

In summary, our study clarifies the effect of data modes and correlations on the learning process of RBMs.

[1] Robin Thériault, Francesco Tosello, Daniele Tantari, Neural Networks **189**, 107542 (2025).

P54

Stabilization system for solar loading thermography applied on cultural heritage objects exposed outdoors: the contribution of advanced algorithms and dual-branch U-Net

Yinuo Ding[1], Gilda Russo[2], Reagan Kasonsa Tshiangomba[3], Enza Pellegrino[2], Antonio Cicone[3], Stefano Sfarra[2]

[1]*(Centre for Composite Materials and Structures (CCMS), Harbin Institute of Technology, Harbin 150001, China)*

[2]*Department of Industrial and Information Engineering and Economics, University of L'Aquila, 67100 L'Aquila, Italy ,*

[3]*Department of Information Engineering, Computer Science and Mathematics, University of L'Aquila, 67100 L'Aquila, Italy*

This study investigates the use of solar loading thermography (SLT) for thermal non-destructive testing (TNDT) and image stabilization of cultural heritage objects, specifically focusing on a century-old ancient book. The irregular contours and deteriorated areas of the book posed significant challenges for feature extraction due to non-uniform temperature varia- tions. To address these challenges, a convolutional neural network (CNN) based dual-branch network of U-Net was used to stabilize the dataset across three degrees of freedom with the ancient book. The stabilization process involved tracking feature lines across each frame of the time-domain datasets, correcting for frame misalignment caused by sample movement during prolonged data acquisition. The effectiveness of this stabilization technique was evaluated by comparing the results of principal component analysis (PCA), fast Fourier transform (FFT), and fast iterative filtering (FIF) algorithms before and after stabilization. Significant improvements were observed, particularly in the clarity and accuracy of defect detection, indicating that this technique provides a robust foundation for further analysis and processing of SLT datasets in cultural heritage preservation. This research demonstrates the potential of combining advanced image processing techniques with SLT to enhance the quality and reliability of NDT in preserving valuable historical artifacts.

P55

The Effect of Label Noise on the Information Content of Neural Representations

Ali Hussaini Umar[1], **Franky Kevin Nando Tezoh**[1], **Jean Barbier**[2], **Santiago Acevedo**[1] **and Alessandro Laio**[1,2]

[1] *Scuola Internazionale Superiore di Studi Avanzati (SISSA), Trieste, Italy*
[2]*International Centre for Theoretical Physics (ICTP), Trieste, Italy*

In supervised classification tasks, models are trained to predict a label for each data point. In real-world datasets, these labels are often noisy due to annotation errors. While the impact of label noise on the performance of deep learning models has been widely studied, its effects on networks' hidden representations remain poorly understood. We address this gap by systematically comparing hidden representations using the Information Imbalance, a computationally efficient proxy of conditional mutual information. We analyze representations learned by networks of varying sizes, trained on datasets with controlled levels of label noise. Through this analysis, we observe that the information content of the hidden representations follows a double descent as a function of the number of network parameters, akin to the behavior of the test error. We show that in the underparameterized regime, representations learned with noisy labels are more informative than those with clean labels, while in the overparameterized regime they are equally informative. We also show that label noise decreases the information content between the penultimate and pre-softmax layers, mirroring the increase in the test error. Finally, representations learned from random labels perform worse than random features when the number of network parameters and training samples are scaled proportionally with a fixed ratio. Overall, our results suggest that representations of overparameterized networks are robust to label noise. The relationship between the information imbalance (between the penultimate and pre-spftmax layers) and the test error offers a new perspective on understanding generalization and highlights how training objectives shape internal representations. In addition, the poor performance of the representations learned from random labels—compared to random features—indicates that training in this setting goes beyond lazy learning. The network weights adapt to encode label-specific information even when labels are uninformative.

P56

GFlowNets + PINNs for Multi-Candidate Symbolic Model Discovery in Dynamical Systems from Sparse Data

Saranya Varakunan, Mohammad Kohandel

Department of Applied Mathematics, University of Waterloo, Canada

Inverse problems in biology often require inferring unknown dynamical relationships from sparse and noisy observations. Traditional symbolic regression methods can recover candidate equations from data, but frequently rely on numerical derivative estimation and typically return a single best-fitting model, limiting their robustness and ability to represent uncertainty. In this work, we propose a framework that combines Generative Flow Networks (GFlowNets) with Physics-Informed Neural Networks (PINNs) for symbolic model discovery in dynamical systems. Candidate equations are generated sequentially using a GFlowNet and evaluated through a PINN-based reward that measures consistency with both observed data and the governing differential equation.

As a proof of concept, we consider a mutual-inhibition gene regulatory network with cell-cell communication and investigate recovery of an unknown nonlinear interaction term from sparse trajectory data. We show that the proposed approach identifies multiple plausible candidate expressions while successfully recovering the underlying Hill-type activation function from as few as five time point observations. In contrast to a symbolic regression baseline based on genetic programming, the GFlowNet–PINN framework more accurately recovers the mechanistic structure of the interaction despite limited data availability. These results demonstrate the potential of combining generative modeling and physics-informed learning for interpretable equation discovery in scientific machine learning and computational biology.

P57

Reconstructing Incomplete Surface Temperature Fields using Graph Attention Networks

Emma C. Weber[1], **and Dominique Guillot**[1],

[1]*Mathematical Sciences, University of Delaware*

Incomplete and irregularly sampled climate records, such as the HadCRUT surface temperature dataset [1], motivate the need for robust data reconstruction methods. We present a Graph Attention Network (GAT) framework [2] for climate field infilling that models Earth's surface as a geospatial graph, with edges encoding geodesic distance, latitudinal, and elevational relationships. The attention mechanism adaptively learns spatial dependencies to better reconstruct surface temperature. Preliminary experiments demonstrate improved accuracy over Kriging [3] and greater computational efficiency than GraphEM [4].

[1] C. P. Morice, J. J. Kennedy, N. A. Rayner, and P. D. Jones, "Quantifying uncertainties in global and regional temperature change using an ensemble of observational estimates: the HadCRUT4 dataset," *Journal of Geophysical Research: Atmospheres*, vol. 117, p. D08101, 2012.

[2] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph Attention Networks," in *International Conference on Learning Representations (ICLR)*, 2018.

[3] N. Cressie, *Statistics for Spatial Data*, rev. ed., John Wiley & Sons, 1993.

[4] D. Guillot, B. Rajaratnam, and J. Emile-Geay, "Statistical paleoclimate reconstructions via Markov random fields," *Annals of Applied Statistics*, vol. 9, no. 1, pp. 324–352, 2015.

P58

Maximizing Memory Capacity in Heterogeneous Networks

Kaining Zhang [1,2] , **Gaia Tavoni** [2,3,4]

[1] *Quantitative Life Sciences Section, The Abdus Salam International Centre for Theoretical Physics, Trieste, 34151, Italy*

[2] *Department of Neuroscience, Washington University in St. Louis, St. Louis, MO, 63130, USA*

[3] *Department of Physics, Washington University in St. Louis, St. Louis, MO, 63130, USA*

[4] *Department of Electrical and Systems Engineering, Washington University in St. Louis, St. Louis, MO, 63130, USA*

A central question in neuroscience is which neuronal and connectivity properties determine a network's ability to store information. The theory of Hopfield models has provided powerful tools to investigate this question. Here, we generalize classical results on the memory capacity of homogeneous Hopfield models to a broader class of heterogeneous networks. We derive an analytical formula for the maximal capacity of networks with arbitrary, and generally heterogeneous, activation rates (coding levels) of neurons in memory patterns, as well as arbitrary connectivity architectures. We use this result to make normative predictions about the properties that maximize capacity in brain-inspired networks. We show that, although heterogeneity in neuron coding levels and inward connection counts (in-degrees) generally reduces capacity, maximal capacity is retained when these two parameters are correlated. This prediction holds across various biologically relevant scenarios, including the storage of independent patterns in both classical and dendritic networks, as well as the storage of example patterns clustered around prototypes representing concepts. In the latter case, heterogeneity similarly affects the capacity for both examples and concepts. Finally, we analyze bipartite models of the CA3-DG circuit in the hippocampus, known to be crucial for memory encoding. In these networks, capacity is maximized by a quasi-indexing encoding scheme, where each neuron in the DG binds subsets of features from a few memory patterns stored in CA3. Compared to a complete-indexing scheme, where each DG unit binds all the features of a single pattern, quasi-indexing significantly improves both network capacity and memory robustness to neuron ablation. These findings introduce new dimensions to the hippocampal index theory of memory encoding and retrieval, while suggesting specific neural substrates for these functions.