

# Viewing Neural Networks Through a Statistical-Physics Lens

Statistical physics is shedding light on how network architecture and data structure shape the effectiveness of neural-network learning.

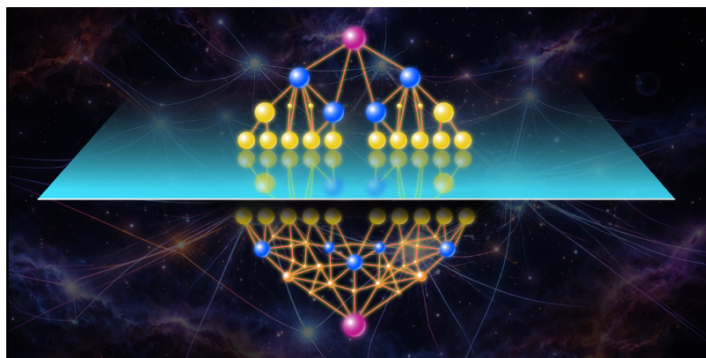
By **Hugo Cui**

Machine-learning technologies have profoundly reshaped many technical fields, with sweeping applications in medical diagnosis, customer service, drug discovery, and beyond. Central to this transformation are neural networks (NNs), models that learn patterns from data by combining many simple computational units, or neurons, linked by weighted connections. Acting collectively, these neurons can process data to learn complex input–output relationships. Despite their practical success, the fundamental mechanisms by which NNs learn remain poorly understood at a theoretical level. Statistical physics offers a promising framework for exploring central questions in machine-learning theory, potentially clarifying how learning depends on the

layout of the network—the NN architecture—and on statistics of the data—the data structure (Fig. 1).

Three recent papers in a special *Physical Review E* collection (See [Collection: Statistical Physics Meets Machine Learning - Machine Learning Meets Statistical Physics](#)) provide significant insights into these questions. Francesca Mignacco of City University of New York and Princeton University and Francesco Mori of the University of Oxford in the UK derived analytical results on the optimal fraction of neurons that should be active at a given time [1]. Abdulkadir Canatar and SueYeon Chung of the Flatiron Institute in New York and New York University investigated the influence of the precision with which a network is “trained” on the amount of data the NN can reliably decode [2]. Francesco Cagnetta at the International School for Advanced Studies in Italy and colleagues showed that NNs whose structure mirrors that of the data learn faster [3].

To understand why analyzing NNs poses unique mathematical challenges, one needs to delve into their inner workings. NNs are complex algorithms that predict properties, or “labels,” of data points—individual items in a dataset. For example, they can identify the type of object represented in an image or the topic of a text. To accomplish these tasks, neurons are arranged into a large network, and the strength of their mutual connections is adjusted so that, for any given data point in a known dataset, the output of the NN is as close as possible to the correct label. This so-called training process is carried out by minimizing an energy function—in a way reminiscent of deriving spin dynamics in an Ising model from a Hamiltonian.



**Figure 1:** Three studies apply statistical-physics approaches to understand neural networks. This artistic visualization renders a situation where the hierarchical structure of the data (above) is mirrored by the structure of the neural network (below).

Credit: APS/Carin Cain

Learning therefore results from the evolution of neuron interactions as the NN processes the data. While each individual interaction follows simple, elementary rules, their combination, which gives rise to the remarkable learning ability of NNs, still largely eludes mathematical understanding.

The sheer number of interneuron connections makes analyzing their collective interaction a daunting task. GPT-3, for example, contains a staggering  $175 \times 10^9$  connections [4]. It is therefore unsurprising that statistical physics—the branch of physics that studies the collective behavior of large collections of elementary units, such as particles in a gas—offers a natural and powerful framework for studying NNs. The application of statistical physics to machine learning has a long history, starting with the foundational works of Elizabeth Gardner and Bernard Derrida in the 1980s [5, 6] and culminating in the recognition of Giorgio Parisi’s work with the 2021 Nobel Prize in Physics. This line of research adopts a guiding principle from statistical physics: studying simple toy models that are amenable to detailed analysis yet capture essential aspects of NNs. The three papers highlighted here share that philosophy.

In some cases, the network may become too fine-tuned to the training data. This overadaptation can lead the NN to perform poorly on new data. To mitigate this problem, certain neurons can be turned off during training, incentivizing them to learn independently and thereby improving the network’s robustness. In practice, however, the fraction of deactivated neurons is selected heuristically, by trial and error rather than through a principled theoretical framework. Bridging this gap, Mignacco and Mori provided a theoretical analysis of this mechanism. Building on a technique introduced by David Saad and Sara Solla [7], the researchers described the network’s behavior in terms of just three variables, deriving equations governing their evolution during training. Such a compact description is emblematic of statistical-physics approaches—think of how the properties and dynamics of a gas containing countless molecules can be subsumed by just pressure and temperature. Leveraging these reduced equations, Mignacco and Mori were able to determine in remarkable detail how neuron deactivation affects learning and to derive an explicit expression for the optimal deactivation rate.

Another strategy to mitigate overadaptation is to introduce a

tolerance in the NN prediction—that is, to require agreement with the training labels only up to a prescribed accuracy. Canatar and Chung uncovered a critical value of this tolerance that separates two regimes: one in which the NN perfectly fits the training data and another in which overadaptation is avoided. In physical terms, this regime crossover corresponds to a first-order phase transition. Canatar and Chung tackled this problem by harnessing the “replica method” from statistical physics—a standard tool in the study of spin glasses [8]. Using this approach, the duo derived an exact analytical expression for the critical tolerance and showed how it depends on hidden statistical structures in the data [2].

In a striking result, Cagnetta and collaborators elucidated how the statistical distribution, or structure, of the data shapes NN learning [3]. To model structure in a tractable yet realistic way, they took inspiration from the “compositionality” of natural language: Sentences can be broken into their constitutive clause, which in turn can be broken into individual words. Analyzing data with such a hierarchical structure, the researchers showed that NNs with matching compositional architectures can learn significantly faster than architectures that are nominally more powerful but less suited to this situation. Remarkably, the team could precisely predict the speed of learning for these highly complex NNs from the correlation structure of the data. The result highlights the central role of data structure in NN performance.

Looking ahead, the physicist’s path in machine-learning theory leads into terrain that is both exciting and challenging. Progress will depend on confronting several core questions: How can the structure of real data be modeled analytically? More specifically, which statistical correlations must be included in the theory, and which can be safely neglected? And how can the daunting complexity of modern NN architectures—often comprising thousands of different building blocks—be theoretically tackled? The three studies highlighted here offer partial answers and chart the way forward. Striking the right balance between tractability and realism requires careful choices of abstraction, calling for both the refinement of existing tools, and the development of new ones. In seeking a better theory of NNs, physicists may uncover new ideas that may, in turn, feed back into and enrich the toolbox of statistical physics.

**Hugo Cui:** Mathematics Laboratory of Orsay, University of Paris-Saclay and French National Centre for Scientific Research (CNRS), Gif-sur-Yvette, France

## REFERENCES

1. F. Mori and F. Mignacco, “Analytic theory of dropout regularization,” **Phys. Rev. E** **112**, 045301 (2025).
2. A. Canatar and S. Chung, “Statistical mechanics of support vector regression,” **Phys. Rev. E** **112**, 025301 (2025).
3. F. Cagnetta *et al.*, “Scaling laws and representation learning in simple hierarchical languages: Transformers versus convolutional architectures,” **Phys. Rev. E** **112**, 065312 (2025).
4. T. B. Brown *et al.*, “Language models are few-shot learners,” in **Proc. 34th Int’l Conf. Neural Information Processing Systems (NIPS ‘20)** (Curran Associates, Red Hook, 2020), p. 1877.
5. E. Gardner and B. Derrida, “Optimal storage properties of neural network models,” **J. Phys. A: Math. Gen.** **21**, 271 (1988).
6. E. Gardner and B. Derrida, “Three unfinished works on the optimal storage capacity of networks,” **J. Phys. A: Math. Gen.** **22**, 1983 (1989).
7. D. Saad and S. A. Solla, “On-line learning in soft committee machines,” **Phys. Rev. E** **52**, 4225 (1995).
8. G. Parisi, “Toward a mean field theory for spin glasses,” **Phys. Lett. A** **73**, 203 (1979).