Proceedings of CONF-MLA 2025 Symposium: Intelligent Systems and Automation: AI Models, IoT, and Robotic Algorithms
DOI: 10.54254/2755-2721/2025.LD26945

From First Principles: A Statistical Mechanics Framework for Machine Learning

Xiaochen Liu

University of California, Santa Barbara, USA
xiaochen_liu@ucsb.edu

Abstract. Modern machine learning has achieved remarkable empirical success, yet its theoretical underpinnings remain incomplete. This paper develops a unified framework for understanding learning through the first principles of statistical physics. Within this view, a model's loss function corresponds to an energy landscape, the training process to stochastic dynamics, and generalization to an entropy-driven preference for flat minima and phase-transition behavior. Energy-Based Models (EBMs) are analyzed as a central case study. This study shows how their mathematical form arises naturally from the Principle of Maximum Entropy, and how their latent-variable structures can be interpreted as learning an effective theory of the data. The learning dynamics are also given a direct physical interpretation, with training algorithms like Contrastive Divergence representing non-equilibrium relaxation processes. In particular, we highlight that the Hessian-dependent gradient noise in Stochastic Gradient Descent emerges as the key mechanism driving an exponential preference for the flat minima associated with good generalization. Beyond explanation, this framework offers constructive principles. Symmetry, conservation laws, and duality provide powerful inductive biases for building data-efficient and interpretable models, as exemplified by Physics-Informed Neural Networks and compositional generative architectures. We conclude by pointing toward future research directions, including exploring the limits of compositionality through the lens of spin glass theory and developing a non-equilibrium thermodynamic framework for learning dynamics.

Keywords: Energy-Based Models, Statistical Mechanics of Learning, Physics-Informed Machine Learning

1. Introduction

The 2024 Nobel Prize in Physics, awarded to John Hopfield and Geoffrey Hinton, serves as a formal recognition of the deep, intrinsic connection between physics and machine learning [1]. For many years, this relationship has been characterized as a "two-way street," where physics provides inspiration for new models and machine learning offers powerful tools for physical inquiry [2]. However, the growing complexity of modern machine learning models [3,4] invites a stronger thesis: the two fields are converging toward a profound theoretical unification. Many of machine learning's core challenges are, in essence, systematic reappearances of fundamental problems long studied in statistical physics.

© 2025 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2025.LD26945

This paper aims to construct a theoretical framework from the first principles of physics to systematically articulate this unity. We posit that the energy function provides a unifying language between the two fields [2,5,6]. The core concepts of machine learning can be mapped to the fundamental principles of statistical physics: a model's Loss Functioncan be precisely analogized to a physical system's Energy Landscape [4,7]; the training and optimization processcan be modeled as the stochastic dynamical evolution of a system across this landscape [4]; and the central puzzle of generalization is deeply connected to the geometry of this landscape and to the concepts of Phase Transitions in physical systems [2,4].

The historical roots of this unification run deep. The earliest neural networks were not merely inspired by physics but were direct translations of physical models. Hopfield networks encode memory as attractors in an energy landscape, with dynamics analogous to gradient descent toward local energy minima, making them structurally similar to spin glasses [1]. The Boltzmann machine, in turn, introduced probabilistic sampling by explicitly adopting the Boltzmann-Gibbs distribution, creating a direct correspondence with the Ising model in physics: spins are replaced by binary neurons, couplings by synaptic weights, and the Hamiltonian by the model's energy function [8]. The later introduction of hidden variables in Restricted Boltzmann Machines (RBMs) further deepened this analogy, mirroring the use of auxiliary fields or latent variables in statistical mechanics [4,8].

From this statistical mechanics perspective, the process of training an Energy-Based Model (EBM) is equivalent to minimizing a free energy functional. Maximum likelihood estimation, a cornerstone of machine learning, can be reformulated as the minimization of the cross-entropy between the model distribution ρθ and the data distribution ρ^* . The gradient of this objective function is formally identical to the expression for the derivative of free energy in statistical physics:

$$\partial_\theta H\left(\rho^*, \rho_\theta\right) = \underbrace{\mathbb{E}_{\rho^*}\left[\partial_\theta U_\theta\right]}_{\text{data expectation}} - \underbrace{\mathbb{E}_{\rho_\theta}\left[\partial_\theta U_\theta\right]}_{\text{model expectation}} .$$

This equation reveals that the optimization of model parameters is a problem of statistical ensemble matching: the model must be adjusted until it reproduces the expectations observed in the data distribution. The principal difficulty in this process—the calculation of the model expectation—is identical to the problem of evaluating ensemble averages under intractable high-dimensional partition functions in statistical mechanics [7,9,10]. This intractability necessitates the use of sampling methods like Markov Chain Monte Carlo (MCMC), bringing thermodynamic simulation directly into the heart of the learning algorithm [9,11,12]. Algorithms such as Contrastive Divergence, which approximate gradients via limited-step MCMC, can thus be understood as non-equilibrium physical processes [8,9].

By reinterpreting machine learning in these physical terms, we can unify diverse phenomena under a shared framework. Concepts like the partition function, free energy, phase transitions, and even renormalization group theory find direct counterparts in the study of machine learning models [4,5]. This perspective not only deepens our understanding but also inspires practical innovations, such as Physics-Informed Neural Networks (PINNs) that embed physical laws directly into the learning objective [6,10] and physically motivated interatomic potentials that improve both accuracy and interpretability in materials modeling [13].

DOI: 10.54254/2755-2721/2025.LD26945

Building on this unified framework, this paper will use Energy-Based Models as a "theoretical standard model" to systematically explore the physical mechanisms of machine learning. We will analyze the "statics" (representation) and "dynamics" (optimization) of these models, and demonstrate how physical principles like symmetry and duality inspire novel, powerful machine learning architectures [5,14]. Ultimately, this work argues that the phrase "from statistical mechanics to machine learning" is not a loose metaphor but a rigorous mathematical and algorithmic correspondence, offering a principled path toward more robust, interpretable, and controllable artificial intelligence.

2. The statistical mechanics framework of machine learning

The conceptual unification introduced in the previous chapter can be formalized by systematically reframing the core problems of machine learning within the language of statistical mechanics. This framework allows us to analyze machine learning models using the well-established tools of physics, dividing the problem into two parts: the "statics", which concerns the representational properties of a model in equilibrium, and the "dynamics", which describes the optimization process that allows a model to learn from data.

2.1. The "statics" of machine learning: equilibrium states and model representation

At its core, a generative machine learning model seeks to learn a probability distribution $p(x)$ that best describes a given dataset. From a physics standpoint, this is equivalent to defining a statistical ensemble. The first-principles approach to deriving the functional form of this distribution is the Principle of Maximum Entropy. This principle states that, given a set of constraints derived from data (e.g., the observed mean and correlations of features), the most unbiased and objective choice for the underlying probability distribution is the one that maximizes its information entropy [15]. The resulting distribution is uniquely given by the Boltzmann-Gibbs form, $\mathrm{p(x)} \propto \exp(-\mathrm{E(x)})$, where the energy function $E(x)$ is a linear combination of the constrained features [15,16]. This derivation establishes the Boltzmann distribution not merely as an analogy, but as the natural and necessary form for any model that aims to be maximally non-committal about information not present in the data. This provides a fundamental justification for the structure of Energy-Based Models (EBMs).

A central challenge in modeling complex data is capturing the intricate, often high-order, correlations between observed variables. In physics, a powerful strategy for dealing with complex interactions is the development of an effective theory, where the influence of unobserved, high-energy degrees of freedom is integrated out, resulting in effective interactions among the remaining low-energy variables. Machine learning has independently discovered a parallel concept in the form of latent variables [4]. In a Restricted Boltzmann Machine (RBM), for instance, the introduction of a layer of hidden (latent) variables, which interact directly with the visible data variables but not with each other, allows the model to represent complex, high-order correlations among the visible variables once the hidden units are marginalized over [4,8]. This process is mathematically analogous to using a Hubbard-Stratonovich transformation in theoretical physics, where auxiliary fields are introduced to simplify pairwise interactions [1]. Thus, the hierarchical abstraction in deep learning, particularly the role of hidden layers, can be understood as a physical mechanism for learning an effective theory of the data [3].

DOI: 10.54254/2755-2721/2025.LD26945

2.2. The "dynamics" of machine learning: optimization on the energy landscape

The process of training a machine learning model is fundamentally a problem of optimization. Within our physical framework, this optimization is best understood as a dynamical process on a high-dimensional energy landscape, where the loss function defines the potential energy [4,7]. The dominant algorithm for training deep neural networks, Stochastic Gradient Descent (SGD), can be modeled as a Langevin dynamics process, describing the motion of a particle in this potential under the influence of random forces [4,17]. The SGD update rule, $\Delta\theta_t = -\eta\nabla L(\theta_t) + \text{noise}$, is a discrete-time approximation of the Langevin equation.

Crucially, the noise in SGD, which arises from sampling mini-batches of data, is not simple white noise. Its covariance is highly structured, parameter-dependent, and intrinsically linked to the geometry of the loss landscape itself [17,18]. Theoretical and empirical work has shown that the covariance of the stochastic gradient noise is approximately proportional to the Hessian of the loss function, $C\left(\theta\right) \approx \frac{\eta}{B} H\left(\theta\right)$, where η is the learning rate and B is the batch size [17,18]. This Hessian-dependent noise structure is a key feature of SGD's learning dynamics.

This dynamical perspective provides a powerful physical explanation for the remarkable generalization performance of deep learning models. It is now widely believed that solutions that generalize well correspond not to the sharpest minima of the loss landscape, but to the "wide" or "flat" minima [18,19]. The stochastic dynamics of SGD naturally favor such solutions. The landscape-dependent noise in SGD acts as an effective thermal bath, allowing the system to escape sharp, narrow valleys (which have low entropy, or small volume in parameter space) and settle into wide, flat valleys (which have high entropy) [19]. This phenomenon can be understood through a diffusion theory of the learning dynamics, which shows that SGD favors flat minima exponentially more than sharp minima. This exponential preference is a direct consequence of the Hessian-dependent gradient noise [17]. In contrast, a simple Gradient Descent process with injected isotropic white noise (often called Stochastic Gradient Langevin Dynamics, or SGLD) only favors flat minima polynomially [17,20]. This highlights that the data-driven, anisotropic nature of the noise in SGD is a critical, emergent property that biases the learning process towards more generalizable solutions.

3. Energy-based models—the "ising model" of generative learning

To systematically investigate the statistical mechanics of machine learning, it is instructive to select a canonical model system, much like the Ising model serves as a theoretical laboratory for statistical physics [1,4]. For generative learning, Energy-Based Models (EBMs) are the natural choice for this role. Their explicit energy function and direct correspondence to the Boltzmann-Gibbs ensemble allow for a rigorous application of physical principles, providing a clear window into the core challenges of representation, optimization, and sampling. This chapter analyzes EBMs through the lens of a physicist examining a many-body system, focusing on its equilibrium properties, its relaxation dynamics, and methods for probing its internal structure.

3.1. "Equilibrium" properties of the model

The "equilibrium state" of a trained EBM is defined by its probability distribution, $P_\theta(x)$, which is determined entirely by the energy function $E_\theta(x)$ and the partition function Z_θ . The partition function, $Z_\theta = \sum_x e^{-E_\theta(x)}$, encapsulates the global properties of the model's equilibrium state and

DOI: 10.54254/2755-2721/2025.LD26945

is mathematically equivalent to the partition function in statistical mechanics. However, for any non-trivial EBM with many degrees of freedom (i.e., high-dimensional data), this quantity is computationally intractable [9,21]. The difficulty stems from the exponential scaling of the configuration space, a challenge familiar to any physicist working with many-body systems. This intractability has profound consequences: for instance, recent theoretical work has demonstrated that for sufficiently expressive sequence models, the partition function can be formally uncomputable, rendering tasks like model selection and direct likelihood evaluation generally undecidable [22]. This is not a mere technical inconvenience; it is a fundamental barrier that necessitates the use of approximate, sampling-based training methods.

Beyond normalization, the equilibrium perspective also reveals deep insights into model capacity and its relationship with generalization. The number of parameters in an EBM, which controls its expressive power, plays a role analogous to temperature or an external field in a physical system: it governs the phase structure of learning. An under-parameterized model, with insufficient capacity, will fail to capture the true structure of the data distribution. It remains in a disordered "paramagnetic" phase, characterized by high bias, where the learned features are weak and uncorrelated [4]. Conversely, an over-parameterized model trained on limited data risks entering a "spin-glass" phase. In this regime, the model overfits by memorizing spurious, sample-specific configurations, creating a rugged energy landscape fractured into an exponential number of local minima [1,4]. This analogy powerfully reframes the problem of generalization as a statistical phase transition problem. The goal of learning is to find a model in the "ferromagnetic" phase—one that captures the true underlying order of the data—which requires a delicate balance between expressive power (low bias) and data sufficiency to avoid glassy behavior (low variance) [1,2].

3.2. "Relaxation dynamics"—training algorithms

If equilibrium defines the target distribution, then training is the relaxation process that guides the model toward that state. The guiding principle of this dynamic, derived from maximizing the data likelihood, is moment matching: the model parameters θ are updated to ensure that the statistical correlations generated by the model match those measured from the data [8,16]. However, since the model-side expectation, $E_{\rho_\theta}\left[\partial_\theta U_\theta\right]$, is intractable, it must be approximated via sampling.

The most common training algorithm, Contrastive Divergence (CD-k), is best understood from a non-equilibrium physics perspective [8,9]. Instead of allowing the MCMC sampler to run until it reaches equilibrium, CD-k halts after a small, fixed number of steps, k. It thereby approximates the equilibrium statistics using short-time dynamics initiated from data points. This procedure is a computationally efficient but systematically biased, interrupted relaxation process [23]. The efficacy and behavior of the training depend critically on the relationship between k and the intrinsic mixing time of the model's dynamics—the characteristic time required for the MCMC chain to generate independent samples from the equilibrium distribution [23].

Recent work has shown that the learning process itself can be divided into two distinct regimes based on this relationship [23]. When k is much larger than the mixing time, the model is trained in an equilibrium regime. In this case, the gradient approximation is accurate, and the learned parameters converge to values that correctly describe the target Boltzmann distribution. However, as is common in practice, when k is small and the mixing time is long (which it often is for complex models), the system operates in a non-equilibrium regime. Here, the model does not learn the true equilibrium state but instead "memorizes" the short-time relaxation dynamics dictated by the sampler. This can lead to a model that generates visually convincing samples quickly (by learning to

DOI: 10.54254/2755-2721/2025.LD26945

refine data points over k steps) but whose underlying probability distribution is incorrect. This distinction is crucial for scientific applications where accurate likelihood estimation, and not just sample quality, is required [9,23].

3.3. "Probing" the model—interpretability and feature extraction

Just as physicists probe a material's properties using scattering experiments or by measuring response functions, we can interrogate a trained EBM to understand its internal structure. This provides a powerful route to interpretability, a significant challenge for many "black-box" machine learning models. For a RBM, the learned weights, $W_{i\mu}$, can be interpreted as the effective interaction coupling strengths between the visible data units vi and the hidden feature units h_μ , directly analogous to exchange interactions in spin systems [4,8].

Through the training process, the hidden units self-organize to represent abstract, high-level features present in the data. By visualizing the weights connected to a specific hidden unit, one can often identify the feature that it has learned to detect. For example, when an RBM is trained on the MNIST dataset of handwritten digits, individual hidden units often develop weight patterns that correspond to elemental components like strokes, loops, or corners [3,4]. Similarly, an RBM trained on spin configurations from the Ising model can learn weights that correctly identify the nearest-neighbor interactions of the physical Hamiltonian [4]. This process of visualizing weights is a direct method for probing the model's internal representations. It is analogous to measuring an order parameter or a correlation function to characterize the state of a physical system. This physics-inspired approach provides a powerful and intuitive tool for achieving model interpretability, turning the abstract parameters of a neural network into meaningful, observable features [7].

4. Frontiers of convergence—systematic application of physical principles

The statistical mechanics perspective on machine learning does not merely reframe existing algorithms but also provides a generative methodology for constructing new ones. By embedding fundamental physical principles directly into model architectures and training paradigms, one can design learning systems that are more data-efficient, generalizable, and interpretable. This chapter highlights two particularly fertile points of contact: the role of symmetry and conservation laws as inductive biases, and the principle of duality as a blueprint for compositional and equivalent representations.

4.1. Symmetry, conservation laws, and inductive bias

In physics, symmetry principles and their associated conservation laws (via Noether's theorem) are first principles that constrain the space of possible theories and dynamics. A physical model that violates energy conservation, for example, is fundamentally flawed. In machine learning, the analogous concept is inductive bias: the set of assumptions a model uses to generalize from finite data [4]. Incorporating physical laws as strong inductive biases is a powerful strategy for creating data-efficient and robust models.

This idea is most directly realized in the field of Physics-Informed Machine Learning (PIML) [6]. A prominent example is the Physics-Informed Neural Network (PINN), which embeds a partial differential equation (PDE) directly into the neural network’s loss function [6,10]. The total loss becomes a weighted sum of a data-fitting term and a "residual" term that penalizes violations of the PDE. By embedding conservation laws—such as mass, momentum, or energy—into the

DOI: 10.54254/2755-2721/2025.LD26945

optimization objective, PINNs can learn from remarkably sparse data, a regime where traditional deep learning often fails [2,10]. This approach has proven highly effective for developing transferable and physically plausible interatomic potentials in materials science [13] and for solving complex inverse problems.

A more architecturally fundamental approach arises in Lagrangian and Hamiltonian neural networks. These models hard-code the variational principles of classical mechanics into their structure by parameterizing a system's Lagrangian or Hamiltonian rather than its dynamics directly [4]. By construction, the learned dynamics are guaranteed to conserve energy and other symplectic invariants over long time horizons. This explicit incorporation of physical structure is not just an engineering convenience; it represents a principled method for ensuring long-term stability, interpretability, and fidelity to physical law—attributes often missing in "black-box" neural architectures. The ability to control overfitting by enforcing such physical constraints is a key advantage of this approach [20].

4.2. Duality, equivalent representations, and compositionality

Another central organizing principle of physics is duality: the existence of mathematically equivalent yet distinct descriptions of a single system. In the Kramers–Wannier duality of the 2D Ising model, for example, the low-temperature, strong-coupling phase is mapped onto the high-temperature, weak-coupling regime of its dual formulation [4,5]. Such correspondences illustrate that a single physical reality can have multiple valid mathematical representations, each useful for understanding different aspects of its behavior.

In the context of machine learning, EBMs exhibit a natural form of duality through composition. Because probabilities are defined via exponentiated energy functions, combining concepts amounts to summing their energies. For instance, the concept of a "smiling male" can be represented as:

$$E(x) = E_{smile}(x) + E_{male}(x)$$

which induces a joint probability distribution that is the product of the individual distributions:

$$p(x) \propto e^{-E_{smile}(X)} \cdot e^{-E_{male}(X)}$$

This operation is analogous to finding the intersection of two distributions in feature space, enabling a compositional and modular way of constructing complex generative models [11,21]. Just as duality in physics provides multiple equivalent formulations of a theory, compositionality in EBMs offers alternative yet consistent ways to build high-level concepts from simpler components. This approach not only enables controllable, zero-shot generation of composite concepts [11,21] but also reinforces the deep structural link between the two fields. Remarkably, recent work has shown that neural networks can be trained to automatically discover the mathematical mappings of physical dualities, such as the Kramers-Wannier duality, directly from lattice configuration data, further cementing this profound connection [5].

Together, symmetry-based inductive biases and duality-based compositionality illustrate how physical principles can do more than inspire analogies—they can serve as constructive rules for

DOI: 10.54254/2755-2721/2025.LD26945

designing new machine learning paradigms. This systematic application of physics provides a blueprint for the next generation of robust, interpretable, and controllable AI models.

5. Research outlook

The convergence of physics and machine learning opens up a vast and fertile ground for new research. As a student with a background in physics, I am particularly interested in applying the rigorous analytical tools of my discipline to address fundamental open questions in machine learning theory. Based on the literature surveyed in this paper, I propose two potential directions for my future graduate research:

5.1. The geometry of compositionality and "conceptual frustration"

The ability to compose concepts by combining their respective EBMs is a powerful feature for generative modeling [11], yet its boundaries remain unexplored. My research interest lies in applying tools from spin glass theory to study conceptual frustration: when the energy functions of two concepts interfere destructively (e.g., combining "smiling" with "frowning"), the resulting energy landscape may become intractable. I hypothesize that high conceptual frustration will lead to a glassy energy landscape, characterized by a proliferation of metastable states [1,4]. I plan to investigate this by analyzing the Hessian spectrum of the combined energy function to quantify its curvature [18-20] and its complexity(the number of local minima). The goal is to establish a quantitative link between the degree of frustration and the practical failure modes of generative models, such as increased MCMC mixing times [9,23] or a collapse in sample diversity. A successful outcome of this research would be a principled method to predict which conceptual combinations are stable, thereby preventing the generation of nonsensical or incoherent outputs and guiding the design of more robust compositional systems.

5.2. A non-equilibrium thermodynamic framework for learning dynamics

The training of deep neural networks is an inherently non-equilibrium process [17,23], yet analyses of SGD often focus on near-equilibrium properties, such as its preference for minima with low free energy [19,20]. I am interested in developing a more complete non-equilibrium thermodynamic description of the learning process itself. My research aims to define and measure thermodynamic quantities like work and dissipated heat for the SGD trajectory by applying tools from modern stochastic thermodynamics, such as the Jarzynski equality and Crooks fluctuation theorem. For instance, the work done on the model during a parameter update could be related to the change in the loss function, while the heat dissipated could be linked to the entropy production from the stochastic gradient noise. I anticipate that key events in training, such as escaping a sharp local minimum or the onset of overfitting, will be associated with distinct signatures in the distribution of these thermodynamic quantities. This framework could yield novel, physically-grounded metrics for monitoring training in real-time, leading to more effective early-stopping criteria and principled schedules for tuning hyperparameters like the learning rate.

6. Conclusion

This paper has sought to construct a unified theoretical framework for understanding machine learning through the first principles of statistical physics. We have argued that the profound connection between these two fields is not merely one of analogy, but of deep structural and

DOI: 10.54254/2755-2721/2025.LD26945

mathematical equivalence. The core challenges of machine learning—optimization, representation, and generalization—can be systematically reframed as problems of dynamics, equilibrium states, and phase transitions in complex, high-dimensional statistical systems. Our analysis was centered on Energy-Based Models (EBMs), which serve as a "theoretical standard model" for generative learning. We established the "statics" of this framework, demonstrating how the Principle of Maximum Entropy naturally gives rise to the Boltzmann-Gibbs distribution that defines EBMs, and how latent variables can be understood as a physical mechanism for learning an effective theory of the data. We then explored the "dynamics," modeling the Stochastic Gradient Descent (SGD) algorithm as a non-equilibrium Langevin process. This perspective reveals that the structured, Hessian-dependent nature of the stochastic gradient noise is a critical feature, enabling SGD to exponentially favor the wide, flat minima of the loss landscape that are associated with good generalization. Building on this foundation, we demonstrated how fundamental physical principles are not just explanatory but also constructive. Symmetries and conservation laws provide powerful inductive biases for designing data-efficient and robust models like Physics-Informed Neural Networks (PINNs), while the concept of duality inspires novel compositional architectures in generative modeling. This systematic application of physics provides a principled blueprint for developing the next generation of robust and interpretable machine learning systems.

References

[1] Baiesi, M., & Suweis, S. (2025). The enduring relevance of simple machine learning models. Europhysics News, 56(1), 20–23.
[2] Levine, H., & Tu, Y. (2024). Machine learning meets physics: A two-way street. Proc. Natl. Acad. Sci. USA, 121(27), e2403580121.
[3] Carbone, D. (2024). Hitchhiker's guide on the relation of Energy-Based Models with other generative models, sampling and statistical physics: a comprehensive review. arXiv preprint arXiv: 2406.13661.
[4] Mehta, P., Bukov, M., Wang, C. H., Day, A. G., Richardson, C., Fisher, C. K., & Schwab, D. J. (2019). A high-bias, low-variance introduction to machine learning for physicists. Physics Reports, 810, 1–124.
[5] Gupta, P., Ferrari, A. E., & Iqbal, N. A. (2025). A machine learning approach to duality in statistical physics. In Proc. 42nd Int. Conf. Machine Learning.
[6] Meng, C., Griesemer, S., Cao, D., Seo, S., & Liu, Y. (2025). When physics meets machine learning: A survey of physics-informed machine learning. Mach. Learn. Comput. Sci. Eng., 1(1), 20.
[7] Niroomand, M. P., & Wales, D. J. (2023). Physics-inspired interpretability of machine learning models. arXiv preprint arXiv: 2304.02381.
[8] Ackley, D. H., Hinton, G. E., & Sejnowski, T. J. (1985). A learning algorithm for Boltzmann machines. Cognitive Science, 9(1), 147–169.
[9] Du, Y., Li, S., & Mordatch, I. (2020). Compositional visual generation with energy based models. In Advances in Neural Information Processing Systems, 33 (pp. 6637–6647).
[10] Raissi, M., & Karniadakis, G. E. (2018). Hidden physics models: Machine learning of nonlinear partial differential equations. Journal of Computational Physics, 357, 125–141.
[11] Hinton, G. E., & Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. Science, 313(5786), 504–507.
[12] Liu, W., Wang, X., Owens, J., & Li, Y. (2020). Energy-based out-of-distribution detection. In Advances in Neural Information Processing Systems, 33 (pp. 21464–21475).
[13] Pun, G. P., Batra, R., Ramprasad, R., & Mishin, Y. (2019). Physically informed artificial neural networks for atomistic modeling of materials. Nature Communications, 10(1), 2339.
[14] Domingo-Enrich, C., Bietti, A., Vanden-Eijnden, E., & Bruna, J. (2021). On energy-based models with overparametrized shallow neural networks. In Proc. Int. Conf. Machine Learning (pp. 2771–2782). PMLR.
[15] De Martino, A., & De Martino, D. (2018). An introduction to the maximum entropy approach and its application to inference problems in biology. Heliyon, 4(4).
[16] LeCun, Y., Chopra, S., Hadsell, R., Ranzato, M., & Huang, F. (2006). A tutorial on energy-based learning. Predicting Structured Data, 1(0).

DOI: 10.54254/2755-2721/2025.LD26945

[17] Xie, Z., Sato, I., & Sugiyama, M. (2020). A diffusion theory for deep learning dynamics: Stochastic gradient descent exponentially favors flat minima. arXiv preprint arXiv: 2002.03495.
[18] Jastrzębski, S., Kenton, Z., Arpit, D., Ballas, N., Fischer, A., Bengio, Y., & Storkey, A. (2017). Three factors influencing minima in sgd. arXiv preprint arXiv: 1711.04623.
[19] Chaudhari, P., Choromanska, A., Soatto, S., LeCun, Y., Baldassi, C., Borgs, C., … & Zecchina, R. (2019). Entropy-sgd: Biasing gradient descent into wide valleys. Journal of Statistical Mechanics: Theory and Experiment, 2019(12), 124018.
[20] Kozyrev, S. V., Lopatin, I. A., & Pechen, A. N. (2024). Control of overfitting with physics. Entropy, 26(12), 1090.
[21] Du, Y., & Mordatch, I. (2019). Implicit generation and modeling with energy based models. Advances in Neural Information Processing Systems, 32.
[22] Lin, C. C., & McCarthy, A. (2022). On the uncomputability of partition functions in energy-based sequence models. In Proc. Tenth Int. Conf. Learning Representations (ICLR).
[23] Decelle, A., Furtlehner, C., & Seoane, B. (2021). Equilibrium and non-equilibrium regimes in the learning of restricted Boltzmann machines. Advances in Neural Information Processing Systems, 34, 5345–5359.