Editors' Suggestion

Tensor networks and efficient descriptions of classical data

Sirui Lu,* Márton Kanász-Nagy, Ivan Kukuljan, and J. Ignacio Cirac
Max-Planck-Institut für Quantenoptik, Hans-Kopfermann-Strasse 1, D-85748 Garching, Germany
and Munich Center for Quantum Science and Technology (MCQST), Schellingstrasse 4, D-80799 München, Germany

(Received 28 November 2024; accepted 10 February 2025; published 6 March 2025)

We investigate the potential of tensor-network-based machine-learning methods to scale to large image and text datasets. We study how the mutual information between a subregion and its complement scales with the subsystem size L, similarly to approaches used in quantum many-body physics. Using various mutual information (MI) estimation methods, including an autoregressive network-based estimator, we find that simple image datasets exhibit area law scaling, suggesting efficient representation by two-dimensional tensor networks. More complex image datasets exceed the area law, indicating the need for generalized tensor network states or hybrid models. For Wikipedia text data, we observe power-law scaling $I(L) \propto L^{\nu}$ with $\nu \approx 0.82$, approaching a volume law. This implies that one-dimensional tensor networks with area law entanglement may not efficiently capture the structure of text. We introduce two models to reproduce this scaling: a quantum-inspired random pair toy model and a linguistically motivated Markovian dependency tree model. In the latter model, matching the observed MI scaling allows us to infer the word-word correlation length distribution in text.

DOI: 10.1103/PhysRevA.111.032409

I. INTRODUCTION

The past decade has witnessed remarkable advancements in machine learning, primarily driven by deep learning techniques [1,2]. Despite impressive practical successes in tasks such as text and image classification [3,4], generation [5–7], and representation learning [8], the theoretical foundations of deep learning remain an active area of research. Perspectives from information theory [9–14], statistical mechanics [15], and renormalization [16–19] are still being developed. A central question is how neural networks capture the relevant corners of the data space occupied by natural images and text [20].

In contrast, quantum many-body physics offers a mature theoretical framework for understanding quantum data and identifying suitable representations. Tensor networks, supported by rigorous theory based on entanglement [21] and mutual information (MI) [22], provide an efficient representation for quantum data. Analogous to how neural networks capture the manifold of natural images and text, tensor networks efficiently represent low-energy states of quantum many-body systems [23]. These states, while complex, occupy only a small corner of the exponentially large Hilbert space. Identifying this corner is essential for developing efficient numerical methods [24].

Characterizing low-energy states through entanglement or MI scaling has been instrumental in identifying suitable variational quantum states. For example, in one-dimensional systems, it is rigorously proven that the scaling of entanglement between a subsystem and the rest of the system must align with that of the tensor network states used to represent it. Otherwise, tensor network approximations become computationally intractable [25]. This can be explained in terms of the area law for entanglement: states satisfying an area law can be efficiently represented using matrix product states (MPSs), a one-dimensional family of tensor networks (TNs), as depicted in Fig. 1(c). In contrast, for quantum critical systems, where the entanglement entropy scales logarithmically with the subsystem size, a MPS description requires a bond dimension that grows polynomially with the system size [26]. In such cases, multilayer TNs such as the multiscale entanglement renormalization ansatz (MERA) [27] or tree tensor network (TTN) structures [28], shown in Fig. 1(c), are more suitable since they also exhibit critical scaling of entanglement entropy. This understanding also allows one to rule out certain architectures: for instance, TNs are not the appropriate choice for out-of-equilibrium systems that develop volume-law correlations over long times [25].

Given their success in quantum physics and connections to probabilistic graphical models [29,30] of machine learning, tensor networks have recently gained attention in classical machine learning. They have found applications in unsupervised [31–38] and supervised learning tasks [30,32,33,38,39], as well as in compressing neural networks [40,41]. Empirical studies suggest that using tensor networks can significantly reduce computational resources and speed up processes in text generation [40] and image processing [41]. However, a critical question remains: Are tensor networks capable of providing an adequate and efficient representation of real-world data distributions?

In this work, we investigate the scaling of mutual information in text and images, a measure that parallels quantum

*Contact author: sirui.lu@mpq.mpg.de

Published by the American Physical Society under the terms of the Creative Commons Attribution 4.0 International license. Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI. Open access publication funded by Max Planck Society.

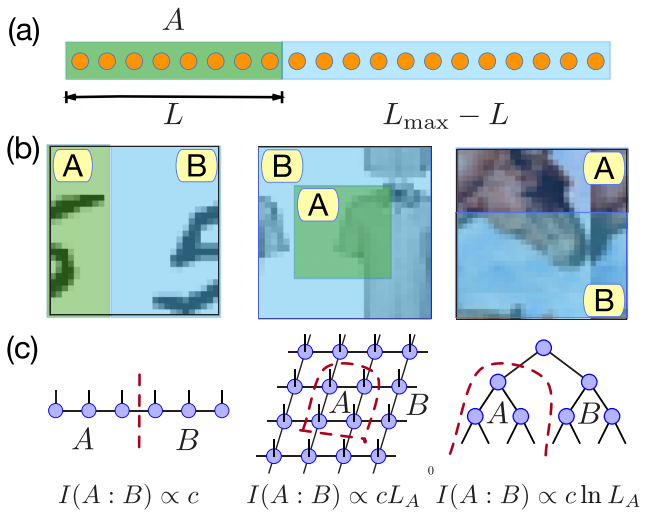


FIG. 1. (a) Schematic illustration of one-dimensional partitioning into regions A and B of lengths L and $L_{\max} - L$, where $L_{\max}$ is the total system size. (b) Three distinct partitioning schemes for two-dimensional image data: left-right (L : R), center-surroundings (C : S), and top-bottom (T : B). Example images are from MNIST (handwritten digit 5), Fashion-MNIST (T-shirt), and CIFAR-10 (horse head) datasets, with random displacements applied to ensure translational invariance. (c) Tensor network representations and their characteristic mutual information scaling: matrix product states (MPS) exhibit constant scaling $I(A:B) \propto c$, projected entangled pair states (PEPS) show linear scaling with boundary length $I(A{:}B) \propto cL_A$, and tree tensor network states display logarithmic scaling $I(A:B) \propto c \ln L_A$. When area laws hold, the mutual information $I(A{:}B)$ is bounded by the number of sites near the $A - B$ boundary.

entanglement in classical datasets [42]. Our analysis reveals that natural text exhibits power-law scaling of MI between subsets and their complements, with an exponent $\nu \approx 0.82$, approaching a volume law. This suggests that traditional one-dimensional tensor network approaches, such as MPS or tree tensor networks (TTN), may not scale efficiently to long texts. Surprisingly, despite the known presence of power-law-decaying correlations in natural text [43,44], the MI scaling in text approaches a volume law rather than the logarithmic scaling seen in quantum critical systems—a scaling that the hidden Markov tree model introduced by Lin and Tegmark [43] would predict, according to our analysis based on the connection between TNs and graphical models. To reconcile this discrepancy, we introduce two toy models: a quantum-inspired random pair model [22], which shows that algebraic correlations in classical probability distributions can coexist with power-law mutual information scaling, and a Markov generative model based on dependency parsing trees [45,46] from natural language processing that incorporates linguistic dependencies. This refined model also reproduces the observed MI scaling when dependency lengths between words follow a power-law distribution.

For image data, we examine widely used datasets such as MNIST [47], Fashion-MNIST [48], and CIFAR-10 [49] [see Fig. 1(b)]. Area-law scaling in these datasets would imply that two-dimensional tensor networks like projected entangled pair states (PEPS) [50], with moderately growing bond dimensions, could efficiently represent them. PEPS have proven effective in representing two-dimensional gapped local systems, which also follow area-law behavior [51–58]. Previous studies [59–61] have reported somewhat conflicting results on MI scaling in MNIST. While Ref. [60] suggested scaling stronger than the area law, Ref. [59] found no definitive evidence for area-law scaling in MNIST. A recent study [61] reported area-law scaling in the TinyImages dataset [62],[1] but not in MNIST. Our analysis, based on various MI estimation methods, shows close-to-area-law scaling for the simpler MNIST dataset, which consists of handwritten images. This supports the potential of PEPS-like tensor networks for representing MNIST [38]. However, our results for more complex datasets like Fashion-MNIST and CIFAR-10 indicate scaling beyond the area law.

Estimating mutual information from classical datasets is an active research area [63–67], with many current methods struggling with scalability and stability due to high-dimensional probability distributions. In this work, we develop a MI estimator for images based on an autoregressive network model [5] and benchmark it against a standard k-nearest-neighbor (kNN) density estimation method [64]. We also utilize the mutual information neural estimator (MINE) [65], enhancing it by using convolutional neural networks (CNNs) as variational functions. By increasing network complexity—from simple fully connected networks to CNNs—we systematically improve MI estimation accuracy. Despite previous challenges in training MINE estimators [68], we achieve stable and consistent results across image and text data.

The remainder of this paper is organized as follows: In Sec. II, we discuss how entanglement entropy and mutual information scaling empower or constrain the applicability of tensor network models for quantum states and classical data. In Sec. III, we present our numerical MI estimation methods, including our autoregressive network-based MI estimator, MINE with CNNs, and the kNN estimator as a reference. In Secs. IV and V, we analyze the mutual information scaling in images and text numerically and introduce toy models to interpret the power-law scaling of both MI and correlations in text. Finally, we summarize our findings and discuss their implications for future research in Sec. VI.

II. EFFICIENT DESCRIPTION OF QUANTUM SYSTEMS AND CLASSICAL DATA

A. Entanglement, tensor networks, and quantum many-body systems

Understanding quantum many-body systems on a lattice requires grappling with the exponential growth of the Hilbert space with the number of lattice sites, a consequence of the tensor product structure of local Hilbert spaces. This challenge

[1]Reference [61], which focuses on MI scaling in image data, finds area-law scaling in the Tiny Images dataset [62], but not in MNIST. We attribute the discrepancy with our work to our approach of making datasets translationally invariant to avoid featureless outer regions in MNIST. Our work now includes a discussion of the dependency tree model, which was not present in the initial version.

mirrors the difficulties in machine learning with large datasets: the space is too vast to fully process or store. In quantum physics, the resolution relies on the concept of entanglement entropy. Low-energy states of local Hamiltonians, particularly ground states, occupy a small subset of the full Hilbert space characterized by low entanglement entropy. This allows for their efficient numerical representation through tensor network states that reflect this entanglement scaling [23,24].

To clarify what is meant by "low entanglement," consider a quantum state described by the density matrix ρ_{AB}. For a subset A of a system and its complement B (see Fig. 1), the reduced density matrix $\rho_A = \mathrm{tr}_B(\rho_{AB})$ captures the state of A, and similarly for B. When the density matrix is pure, i.e., $\rho_{AB} = |\Psi\rangle\langle\Psi|$ for some Ψ, the entanglement entropy (EE) is defined as

$$S(\Psi) = S(\rho_A) = -\mathrm{tr}(\rho_A \ln \rho_A) = S(\rho_B), \tag{1}$$

measuring the entanglement between A and B. The scaling of EE with the size of A is of significant interest in quantum many-body physics. Systems with finite correlation lengths obey an area law for EE, where $S(\rho_A) \propto |\partial A|$, growing with the boundary of A rather than its volume [22]. This has been rigorously established for one-dimensional gapped systems with local Hamiltonians [69] and for two-dimensional systems under certain conditions [70–72]. In contrast, gapless systems like free theories, critical systems, and conformal field theories (CFTs) exhibit diverging correlation lengths and logarithmic EE scaling in one dimension, $S(\rho_A) \propto \ln |A|$, and area-law scaling $S(\rho_A) \propto |A|$ in two dimensions [73–78].

The logarithmic or area-law scaling of EE in ground states suggests that these states only occupy a restricted region of the Hilbert space, making them amenable to efficient characterization through variational states with matching entanglement scaling [22,23,25]. Tensor networks are particularly effective at representing ground states of many-body local Hamiltonians [23]. In one dimension, matrix product states (MPSs) comply with the area law and serve as effective variational states for gapped Hamiltonians [22]. The EE of the multiscale entanglement renormalization ansatz (MERA) and tree tensor networks (TTN) scales logarithmically.[2] In higher dimensions, projected entangled pair states (PEPSs) and MERA follow the area law, while TTN exhibit logarithmic scaling [23].

Empirical evidence and theoretical proofs demonstrate that tensor network methods can approximate states with desired accuracy when their entanglement scaling matches that of the tensor network. For instance, one-dimensional systems with area-law entanglement can be efficiently represented by MPSs with constant bond dimensions [26]. States with logarithmic entanglement corrections can also be described with MPSs with polynomially growing bond dimensions, although MERA and TTN may be more suitable for certain models. Moreover, algorithms based on MPS for ground states of one-dimensional gapped Hamiltonians are known to be efficient [80]. In higher dimensions, PEPS and MERA are proven to follow area laws, making them suitable for representing ground states of gapped local systems [22].

However, certain states violate the area law for entanglement. Notably, nonequilibrium states arising from long-time evolution under local Hamiltonians often exhibit entanglement entropy that grows linearly with time, $S(t) \propto t$ [25,76,81]. Such states eventually require exponentially large bond dimensions for accurate MPS representations, rendering them inaccurately describable by tensor networks.

B. Mutual information, generative models, and classical data

Mutual information (MI) is a key information-theoretic measure for quantifying interdependence between variables, applicable to both classical and quantum systems [22,82]. For mixed states like thermal states, which are quantum generalizations of classical Boltzmann distributions, the entanglement entropy (EE) accounts not only for quantum entanglement but also the degree of mixedness and thermal entropy, leading to mixed signals that cannot be diagnosed without looking at MI. The mutual information between subsystems A and B is defined as

$$I(A:B) = S(\rho_A) + S(\rho_B) - S(\rho_{AB}), \tag{2}$$

where $S(\rho)$ denotes the entropy of the state ρ. MI is advantageous as it provides an upper bound for all correlation functions, and it has been shown that thermal states with finite interaction ranges follow an area law for MI scaling [22]. Indeed, such thermal states can be efficiently represented using tensor networks [83,84].

For classical data, MI is defined using Shannon entropy:

$$S(A) = -\int_{\mathcal{A}} \mathbb{P}_A(a) \ln \mathbb{P}_A(a)\, \mathrm{d}a, \tag{3}$$

where $\mathbb{P}_{AB}$ is the joint probability distribution of the data, and $\mathbb{P}_A = \int_{\mathcal{B}} \mathbb{P}_{AB}(a,b)\, \mathrm{d}b$ is the marginal distribution of subsystem A. MI then quantifies the information shared between A and B.

In classical data contexts, generative models play a role analogous to tensor networks for quantum many-body states. These models aim to replicate the probability distribution of datasets, such as images and text. Given a dataset $\mathcal{D}$ with M samples $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_M\}$ sampled from the true distribution $\mathbb{P}_{\mathrm{data}}(\boldsymbol{x})$, the task of generative modeling is to construct a model distribution $\mathbb{P}_{\mathrm{model}}(\boldsymbol{x})$ capable of generating new samples resembling the original data.

Generative models fall into two categories: explicit density models and implicit density models. Explicit density models directly learn $\mathbb{P}_{\mathrm{model}}(\boldsymbol{x})$ and allow computation of probabilities for any input $\boldsymbol{x}$. Examples include probabilistic graphical models [85], autoregressive neural networks [5,7,86,87], and sequence models in natural language processing [88]. Some tensor networks, known in machine learning as tensor trains [89] and tensor trees, also belong to this category and benefit from efficient contraction properties. Implicit density models, such as Boltzmann machines [90] and generative adversarial networks (GANs) [6], generate data samples without explicitly computing probabilities. Although PEPSs are theoretically explicit density models, the computational

[2]Strictly speaking, the EE of TTN scales logarithmically only when averaged over translations or over different possible placements of the subsystem. See Ref. [79].

complexity of contracting them exactly (which is #P-Complete [91]) necessitates approximate contraction methods, categorizing them as implicit models in practice [92]. Both model types have demonstrated success across various tasks. Explicit models are particularly valuable when the log-likelihood needs to be computed, whereas implicit models excel in high-quality sample generation.

Drawing from experiences in quantum many-body physics, for a model (whether a neural network or tensor network) to effectively learn a dataset, it must be capable of generating probability distributions with MI scaling that is at least as rapid as that of the actual data. Thus, we investigate the MI scaling behavior in classical data to assess the scalability of tensor networks as generative models for large-scale machine-learning tasks.

As illustrated in Fig. 1, we partition the data into two subsystems A and B to study their mutual information $I(A{:}B)$. For one-dimensional data such as text, we use a left-right partition. For two-dimensional data, such as images, we examine two types of separations. The first is a horizontal cut, dividing the system into top-bottom (T : B) regions, with L representing the length of the top region. The second is a center-surroundings partition (C : S), where L is the side length of the central square.

If the area law holds, the mutual information between two regions should be proportional to the interface area between them. For one-dimensional text data, this implies a constant MI, $I(\mathrm{L}:\mathrm{R}) = \text{constant}$, provided A and B are sufficiently large. For two-dimensional image data, $I(\mathrm{T}:\mathrm{B})$ should remain constant for top-bottom partitions, while for center-surroundings partitions, the interface area increases linearly with L, suggesting $I(\mathrm{C}:\mathrm{S}) \sim L$. In finite systems, MI is expected to grow nonlinearly before stabilizing at the area-law plateau, as MI is zero at the boundaries. In the case of volume laws, the MI between the two regions grows with the volume of the smaller region. This occurs, e.g., when each part of the system is correlated with every other part. Initially, for one region very small and the other large, this yields $I(\mathrm{L}:\mathrm{R}) \sim L$, and for image data, $I(\mathrm{T}:\mathrm{B}) \sim L$ and $I(\mathrm{C}:\mathrm{S}) \sim L^2$. Intermediate MI scaling behaviors, such as logarithmic $I(A:B) \sim \ln(L)$ or power-law $I(A:B) \sim L^\alpha$ for $0 < \alpha < 1$, are also possible.

III. ESTIMATORS OF MUTUAL INFORMATION

Estimating mutual information (MI) from empirical data is crucial for uncovering complex relationships within datasets. This problem can be formally stated as

Problem (mutual information estimation from samples). Given N independent and identically distributed (i.i.d.) samples (a_i, b_i), $i = 1, \ldots, N$, from the joint probability density $\mathbb{P}_{AB}$, estimate the mutual information $I(A:B)$ defined as

$$I(A:B) = \int_{\mathcal{A}\times\mathcal{B}} \mathbb{P}_{AB}(a,b) \ln \frac{\mathbb{P}_{AB}(a,b)}{\mathbb{P}_A(a)\mathbb{P}_B(b)} \,\mathrm{d}a\,\mathrm{d}b, \tag{4}$$

where $\mathbb{P}_A(a) = \int_{\mathcal{B}} \mathbb{P}_{AB}(a,b)\,\mathrm{d}b$ and $\mathbb{P}_B(b) = \int_{\mathcal{A}} \mathbb{P}_{AB}(a,b)\,\mathrm{d}a$ are the marginal probability densities of A and B, respectively.

This task is central to various fields, including deep learning and information theory [9,93]. However, accurately determining MI from high-dimensional data remains challenging due to the curse of dimensionality.

Estimation methods are broadly classified into parametric and nonparametric approaches [94]. Nonparametric methods do not assume a specific data distribution model but often struggle with high-dimensional data. Parametric methods, conversely, use model-based approaches with adjustable parameters to approximate the underlying distributions.

In this work, we employ both parametric and nonparametric methods to estimate MI, leveraging their respective strengths and cross-validating their results. We introduce an MI estimator using density estimates provided by advanced autoregressive neural networks [86,95,96]. We implement the mutual information neural estimator (MINE) [65] enhanced by convolutional neural networks. To complement these parametric methods, we use the standard k-nearest neighbor (kNN) estimator as a nonparametric benchmark. This section provides an overview of these estimators, with detailed implementations discussed in Appendixes A and B.

A. Estimation from trained autoregressive networks

We propose using autoregressive neural networks, a tractable explicit density model, to estimate entropy and MI. Autoregressive models [5,7,86,87] decompose the joint probability distribution into a product of conditional probabilities:

$$\mathbb{P}(\boldsymbol{x}) = \prod_i \mathbb{P}(x_i|\boldsymbol{x}_{<i}), \tag{5}$$

where $\boldsymbol{x}_{<i} = [x_1, x_2, \ldots, x_{i-1}]$ represents the vector of variables preceding x_i. These conditional probabilities are defined as parametrized functions with a fixed number of parameters. We consider the conditional distributions $\mathbb{P}(x_i|\boldsymbol{x}_{<i})$ as Bernoulli random variables, defined by a function (to be learned) that maps $\boldsymbol{x}_{<i}$ to the mean of the Bernoulli distribution. Popular architectures for autoregressive models include WaveNet [97], PixelRNN [5], PixelCNN [7], and PixelCNN++ [87], which have demonstrated excellent performance in various tasks. Training an autoregressive neural network involves maximizing the likelihood of the observed data by optimizing the parameters $\boldsymbol{\theta} = \{\theta_1, \theta_2, \ldots\}$:

$$\arg\max_{\boldsymbol{\theta}\in\mathcal{M}} \frac{1}{|\mathcal{D}|} \sum_{\boldsymbol{x}\in\mathcal{D}} \ln \mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{x}) = \frac{1}{|\mathcal{D}|} \sum_{\boldsymbol{x}\in\mathcal{D}} \sum_{i=1}^{n} \ln \mathbb{P}_{\theta_i}(x_i|\boldsymbol{x}_{<i}), \tag{6}$$

where we have substituted the factorized joint distribution of an autoregressive model [Eq. (5)].

After training, we can compute entropies using the learned conditional probabilities. Given that all conditional probabilities are normalized, we can calculate entropies of subregions respecting the sequential ordering through Monte Carlo sampling. However, estimating MI remains challenging as the sequential structure prevents obtaining arbitrary marginal density functions. With a trained network, we can estimate $\ln \mathbb{P}^{\mathrm{AN1}}(x)$ and $\ln \mathbb{P}^{\mathrm{AN1}}(x,y)$. We then train another network with reverse ordering to access $\ln \mathbb{P}^{\mathrm{AN2}}(y)$ and $\ln \mathbb{P}^{\mathrm{AN2}}(y,x)$ for points x and y in A and B, respectively (Fig. 2). Applying to our partition, AN1 is optimized for computing marginals of left or top regions, while AN2 is optimized for right or bottom

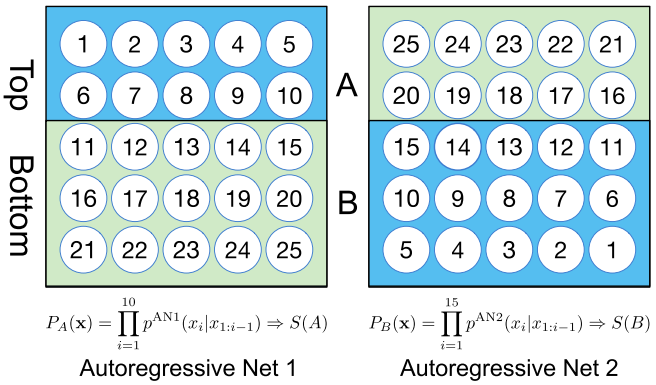


FIG. 2. Autoregressive neural networks and two distinct orderings for computing marginal probabilities. (left) Autoregressive Network 1 processes a 5×5 image in raster scan order (1 to 25), computing conditional probabilities $\mathbb{P}_A(\boldsymbol{x}) = \prod_{i=1}^{10} p^{\mathrm{AN1}}(x_i|x_{1:i-1})$ to estimate entropy $S(A)$ of the top region. (right) Network 2 uses reverse ordering (25 to 1) to compute $\mathbb{P}_B(\boldsymbol{x}) = \prod_{i=1}^{15} p^{\mathrm{AN2}}(x_i|x_{1:i-1})$ for entropy $S(B)$ of the bottom region. These two networks with complementary orderings are trained separately, allowing estimation of marginal entropies for both top and bottom regions independently and thus mutual information via Eq. (7).

regions. The MI is then estimated as

$$I(A:B) = S^{\mathrm{AN1}}(A) + S^{\mathrm{AN2}}(B) - \frac{S^{\mathrm{AN1}}(A,B) + S^{\mathrm{AN2}}(A,B)}{2}. \tag{7}$$

Meanwhile, using a single forward-indexed network to compute marginals for right and bottom regions would require summing over exponentially many configurations of the complementary left and top regions, leading to intractable computations and high variance. By employing both forward and reverse orderings, we ensure that marginal probabilities are always computed via efficient analytical summations over the complement region, resulting in stable and accurate mutual information estimates.

For this estimator to be reliable, consistency between the two model distributions must be ensured. As shown in Fig. 3(a), the difference in estimated entropies at $L = L_{\max}$ is negligible for MNIST, Fashion-MNIST, and CIFAR-10 datasets, indicating close distributions.

We employ PixelCNN [7,98] and PixelCNN++ [87] architectures [99] for estimating conditional probabilities. These models process images in a top-to-bottom sequence but are not compatible with a spiral processing path needed for center-surrounding partitions. PixelCNN assumes a discrete data distribution, requiring discretization of pixel values into 256 bins, processed using a logistic mixture likelihood model. This discretization introduces a scaling factor in the MI estimates, absent in models like MINE and kNN that assume continuous distributions. We adjust the results from PixelCNN and PixelCNN++ in Fig. 4 by a common scaling factor for consistency with MINE and kNN models.

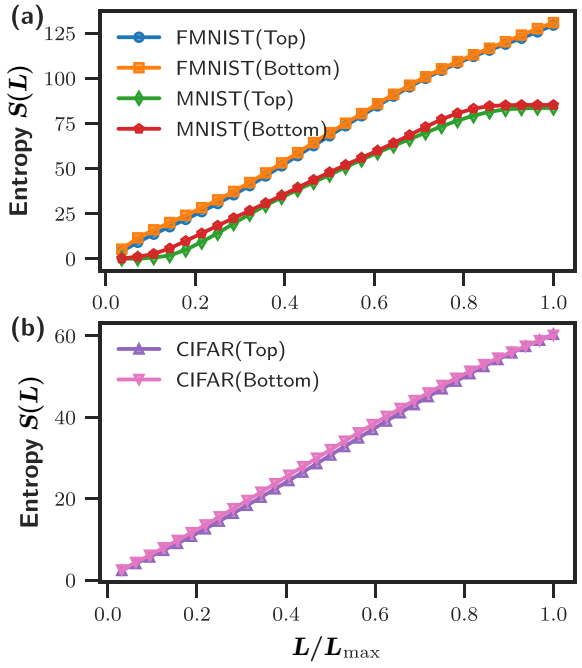


FIG. 3. Shannon entropy scaling in image datasets estimated using PixelCNN [5] and PixelCNN++ architectures [87]. (a) Entropy curves for MNIST and Fashion-MNIST (28×28 pixels) showing volume-law scaling $S(L) \propto L$ for both top and bottom regions. The close agreement between the top and bottom curves indicates consistent probability distributions captured by the trained models. (b) Corresponding analysis for CIFAR-10 ($32 \times 32 \times 3$ pixels) demonstrating similar volume-law scaling. The x axis shows normalized region length $L/L_{\max}$, while the y axis displays entropy $S(L)$ in bits.

B. Estimation from samples: mutual information neural estimation (MINE)

MINE is a parametric estimator that employs variational neural networks to estimate MI [65]. It is particularly effective for complex partitions like center-surroundings in images and text, and serves as a benchmark for autoregressive models in simpler partitions like top-bottom. The idea behind MINE is interpreting mutual information as the Kullback-Leibler (KL) divergence between joint and marginal distributions, transformed into a dual representation. Mutual information is represented as the KL divergence between the joint, $\mathbb{P}_{AB}$, and the product of the marginals, $\mathbb{P}_A \otimes \mathbb{P}_B$: $I(A:B) = D_{\mathrm{KL}}(\mathbb{P}_{AB} \,||\, \mathbb{P}_A \otimes \mathbb{P}_B)$, where D_{KL} is defined as $D_{\mathrm{KL}}(\mathbb{P} \,||\, \mathbb{Q}) := \mathbb{E}_{\mathbb{P}}[\ln \frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}]$. Applying the Donsker-Varadhan dual representation of the KL divergence [100]:

$$D_{\mathrm{KL}}(\mathbb{P}\|\mathbb{Q}) = \sup_{T:\Omega\to\mathbb{R}} (\mathbb{E}_{\mathbb{P}}[T] - \ln(\mathbb{E}_{\mathbb{Q}}[e^T])), \tag{8}$$

where the supremum is taken over all functions $T : \Omega \to \mathbb{R}$ such that the two expectations are finite. Hence,

$$I_{\mathrm{MINE}}(A,B) := \sup_{\theta\in\Theta} \left(\mathbb{E}_{\mathbb{P}_{AB}}[T_\theta] - \ln\left(\mathbb{E}_{\mathbb{P}_A\otimes\mathbb{P}_B}\left[e^{T_\theta}\right]\right)\right). \tag{9}$$

By limiting the class of score functions T_θ to those represented by deep neural networks θ, Eq. (9) provides a lower bound on mutual information [65]. This bound is tight for the optimal

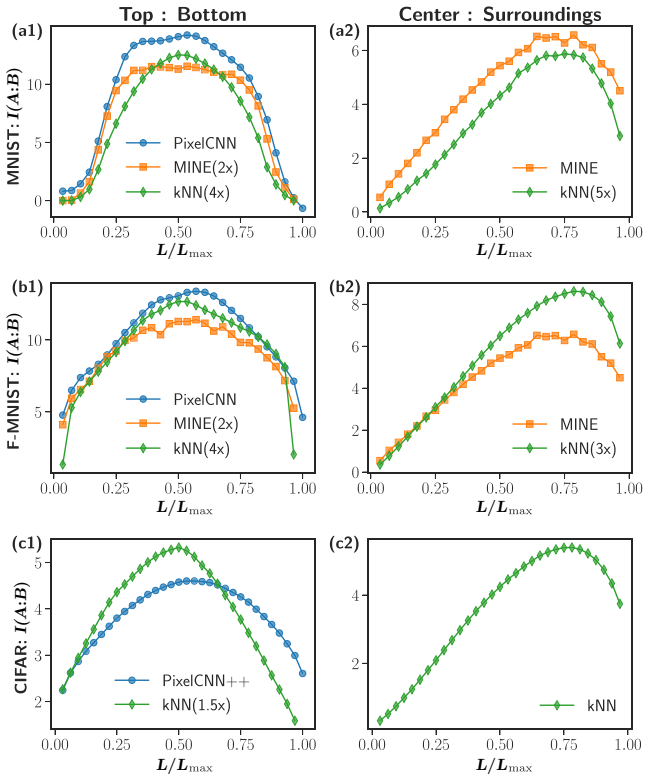


FIG. 4. Mutual information scaling in image datasets analyzed using three complementary estimation methods: (i) PixelCNN and PixelCNN++ autoregressive networks computing exact conditional probabilities, (ii) mutual information neural estimation (MINE) using convolutional neural networks as variational functions, and (iii) k-nearest-neighbor (kNN) density estimation. Left panels show top-bottom (T : B) partitioning; right panels show center-surroundings (C : S) partitioning for (a1), (a2) MNIST, (b1), (b2) Fashion-MNIST (28×28 pixels), and (c1), (c2) CIFAR-10 ($32 \times 32 \times 3$ pixels). For MNIST, the simplest dataset, we observe evidence of area-law scaling through T : B saturation and linear C : S growth. For more complex datasets (Fashion-MNIST and CIFAR-10), $I(\mathrm{C}:\mathrm{S})$ still grows linearly but $I(\mathrm{T}:\mathrm{B})$ does not reach a plateau, suggesting faster than area-law scaling. The x axis represents the normalized partition length $L/L_{\max}$, and the y axis shows the mutual information $I(A:B)$ in bits, with kNN results globally adjusted for consistent comparison.

score function T^*, and enhancing the expressive power of neural networks ensures that MI can be approximated to the desired accuracy. As a lower-bound estimator, a higher MI value using a different network provides a better estimation, allowing for systematic improvement of results.

In practice, this lower bound is estimated on the entire dataset $\mathcal{D}$, with optimization solved by stochastic minibatch gradient descent. The optimized $I_{\mathrm{MINE}}(A, B)$ is considered a lower bound of true MI.

MINE naturally aligns with two-category classification tasks, such as image or text classification, where the neural network outputs a real number indicating the classification result. In the original work of Ref. [65], a fully connected feedforward neural network was used to represent score functions. We incorporate convolutional neural networks as our score functions T_θ, which have proven highly effective for image [3] and text [4] classification. This allows us to achieve improved variational estimates (see Appendix B for comparison) and scale up our calculations. The mutual information neural estimator is versatile and applicable to both top-bottom and center-surroundings partitions.

C. Estimation from samples: k-nearest neighbor

The k-nearest-neighbor (kNN) estimator is a nonparametric method that approximates the data distribution by assuming it is constant within high-dimensional simplices defined by the k-nearest neighbors. The density $\hat{\mathbb{P}}(x)$ at a point x is estimated as

$$\hat{\mathbb{P}}(x) \approx \frac{k}{M \mathrm{Vol}_{\mathrm{kNN}}(x)}, \tag{10}$$

where $\mathrm{Vol}_{\mathrm{kNN}}(x)$ is the volume covering the k nearest data points to x out of M samples. This estimated distribution is used to compute Shannon entropy and, consequently, the MI of the datasets. We use a refined kNN estimator [64,101]. Due to implementation specifics, MI is determined up to an additive constant that depends on k and the sample number n_{data} (see Appendix A). We adjust our results globally to account for this when comparing with MINE and autoregressive estimators (see Fig. 4).

The kNN estimator performs well on low-dimensional data but its accuracy diminishes with increasing data dimensionality. Nevertheless, it can provide MI estimates for text data and all image partitions considered in our study.

In the following sections, we utilize these estimators to analyze mutual information scaling in text and image datasets, assessing the viability of tensor network representations based on the observed scaling behaviors.

IV. MUTUAL INFORMATION SCALING IN IMAGES

Images inherently possess spatial correlations reflecting real-world relationships. For example, in a face image, an eye on one side implies the presence of another on the opposite side. These correlations occur at various scales, with short-range correlations being more prevalent. Such short-range correlations are typically captured by the initial layers of convolutional neural networks [3]. Consequently, we hypothesize that the mutual information in images will scale close to an area law, with possible additional contributions from longer-range correlations.

To test this hypothesis, we employ autoregressive network modeling, MINE, and kNN MI estimators to analyze low-resolution real-world image datasets. These include the MNIST handwritten digit dataset [47], the Fashion-MNIST clothing images [48], and the CIFAR-10 dataset [49], which comprises natural images of animals and vehicles. These datasets are widely used to benchmark machine-learning approaches.

A. Entropy scaling

We begin by examining the entropy of subregions within image data using trained autoregressive neural networks (elaborated in Sec. III A). The results are presented in Fig. 3 for the top-bottom (T : B) partitioning of MNIST, Fashion-MNIST,

and CIFAR-10 datasets. We include the results obtained from autoregressive networks of both orderings.

From Fig. 3, we observe that entropy scales similarly to thermal entropy in physical systems, adhering to a volume law. The volume law in entropy supports our hypothesis that MI is a more suitable metric for studying the information structure in classical data. The closeness of the entropy curves for the two orderings also supports the consistency of the probability distributions captured by the trained autoregressive models.

Image datasets typically focus on a central object with fewer distinctive features towards the edges. This effect is particularly pronounced in MNIST, where edges are often blank. This is evident in Fig. 3(a), which shows a minimal slope near the boundaries ($L \approx 0$ or $L \approx L_{\max}$). As discussed in Appendix A, the MI of MNIST data also decreases towards the edges due to the lack of features. To mitigate this edge effect, we analyzed the data after making the images translationally invariant by randomly displacing the images, as depicted in Fig. 1. We note that this effect is less pronounced in datasets with natural images.

B. Mutual information scaling

Figure 4 presents the mutual information (MI) curves for the top-bottom (T : B) and center-surroundings (C : S) partitions on the translationally invariant MNIST dataset. The $I(\mathrm{T}:\mathrm{B})$ curves display a noticeable plateau in the central region, indicative of area-law scaling. Concurrently, the $I(\mathrm{C}:\mathrm{S})$ curves grow linearly at smaller L values, further supporting the area-law hypothesis.

We extend our investigation to more complex image datasets, including Fashion-MNIST and CIFAR-10. Similar to MNIST, we randomly displaced the images to ensure translation invariance. We employed MINE and PixelCNN++ autoregressive networks along with kNN methods for Fashion-MNIST. The CIFAR-10 dataset, with its three color channels, increases the data dimensionality, presenting an additional challenge for MI estimation. Although MINE results for CIFAR-10 are not available due to computational constraints, our analysis indicates that while $I(\mathrm{C}:\mathrm{S})$ continues to grow linearly, $I(\mathrm{T}:\mathrm{B})$ does not reach a plateau (Fig. 4). This observation suggests that in more generic images, MI scaling could exceed the area law.

These findings have important implications for tensor network representations of image data. The area-law scaling observed in MNIST suggests that two-dimensional tensor networks like PEPS could efficiently represent this dataset, aligning with previous successful applications of tensor network algorithms to MNIST classification tasks [14,33,38,102] and attempts in modeling MNIST images [31,103]. However, the faster-than-area-law scaling in more complex datasets like Fashion-MNIST and CIFAR-10 indicates that traditional tensor network approaches may face scalability challenges for larger, more complex images. This finding motivates exploring more generalized tensor network states or hybrid models that combine tensor networks with neural networks [104], potentially offering a solution to these scalability issues while maintaining computational efficiency.

V. MUTUAL INFORMATION SCALING IN TEXT

Natural text is inherently complex due to factors like grammatical structure, semantics, style, and cross-references, leading to correlations at various scales—from sentence-level grammar to paragraph or document-level coherence. These multiscale correlations suggest high mutual information between different text segments. Supporting this, recent studies have observed algebraically decaying correlations at both the character and word levels [20,43,44]. Advanced transformer models [105] capable of generating long, coherent text also exhibit power-law mutual information scaling. In contrast, models that struggle with long coherence, such as recurrent neural networks [44] and long short-term memory networks (LSTM) [106], show exponentially decaying correlations.

To accommodate the distinct structure of text data, we utilize two MI estimation methods introduced in Sec. III: the mutual information neural estimator (MINE) [65] and the k-nearest-neighbor (kNN) estimator [64]. We apply these methods to a dataset consisting of Wikipedia articles and analyze the mutual information scaling. Subsequently, we introduce a random pair model and a dependency tree model as toy models of text data to understand the observed scaling.

A. Power law scaling in text

We analyze the WikiText-2 dataset, which consists of 600 training articles and 2 million tokens [107]. Words are converted into a computer-readable format using pretrained word-level embeddings from the 50-dimensional GloVe model [108]. This model generates dense vectors for each word, ensuring that words appearing in similar contexts, and thus sharing similar meanings, are proximate in the feature space. Consequently, our mutual information estimation incorporates word meanings. We have verified the robustness of our results by testing them against changes in the dimensionality of the embedding space to 200.

We employ the kNN and MINE estimators using both fully connected and convolutional neural networks as variational functions in the estimator, as detailed in Sec. III. Both methods yield nearly identical results. Initially, we utilized a fully connected feedforward neural network as the score function, in line with the original version of Ref. [65]. To stabilize the optimization, we applied the moving average gradient trick mentioned in Refs. [65,98]. Subsequently, we improved the mutual information neural estimator by using a text convolutional neural network [4] as the score function. While this introduces some biases into the estimation, the results estimated with CNNs are larger and thus more accurate than those estimated with feedforward neural networks. Additionally, CNNs have fewer tunable parameters than feedforward networks, enabling us to scale up to deeper and wider networks.

As anticipated, the strong correlations between different segments of the text result in an MI scaling that is significantly steeper than the logarithmic scaling observed in critical systems, as shown in Figs. 5(b1) and 5(b2). Specifically, for the WikiText-2 dataset, we observe approximate power-law correlations for intermediate lengths L, with an exponent $\nu = 0.82(2)$. This scaling is nearly equivalent to a volume law,

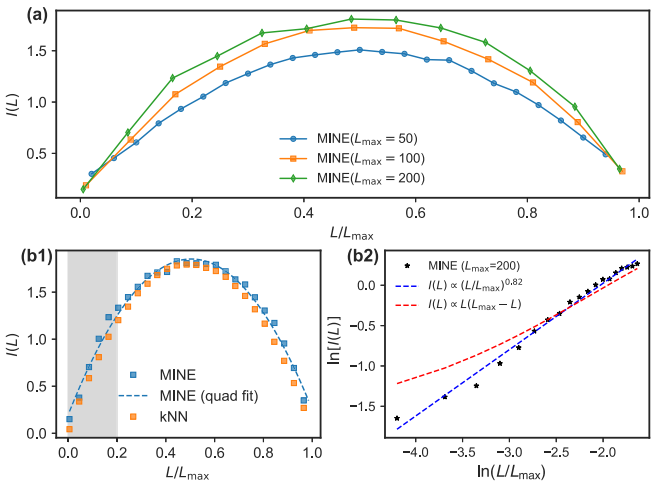


FIG. 5. Mutual information analysis of the WikiText-2 dataset using 50-dimensional GloVe word embeddings. (a) MINE estimates for varying sequence lengths $L_{max} = 50, 100, 200$ words, showing consistent scaling behavior after normalization. A text convolutional neural network [4] is used as the score function. (b1) Comparison between MINE and kNN ($k = 20$, MaxNText = 10 000) estimators for $L_{max} = 200$, with power-law fitting region highlighted in gray. (b2) Log-log analysis of the initial part of the $I(L)$ curve revealing power-law scaling $I(L) \propto L^{0.82(2)}$ (blue dashed line) for small L, compared with the theoretical upper bound $I(L) \propto L(L_{max} - L)$ (red dashed line) derived for maximally correlated elements in Sec. V B. The x axis represents the normalized length of the left region, L/L_{max}, and the y axis shows the mutual information $I(\mathrm{L} : \mathrm{R})$ in bits.

where MI grows linearly with the system size. For context, an area-law scaling would result in a constant MI, independent of system size. Furthermore, as illustrated by the dashed line in Fig. 5(c), the scaling closely aligns with a model where all words are equally correlated, regardless of their distance. This would result in a scaling of $I(\mathrm{L} : \mathrm{R}) \propto L$, $(L_{max} - L)$. However, deviations from a perfect power law are observed, especially at smaller system sizes, indicating a more complex scaling behavior that may involve multiple length scales or crossover phenomena.

These findings have significant implications for representing text using tensor networks. The observed power-law scaling, approaching a volume law, suggests that traditional one-dimensional tensor network approaches, such as matrix product states (MPSs) or tree tensor networks (TTNs), may not scale efficiently to long texts. This is because these tensor network structures are designed to capture area-law or logarithmic scaling of entanglement, which is much slower than the observed near-volume-law scaling in text data.

Moreover, our results challenge the assertion made by Ref. [43] that languages exhibit critical distributions, implying that the structures of natural languages significantly differ from local critical systems. In quantum critical systems, power-law-decaying correlations typically lead to logarithmic scaling of entanglement entropy [76–78]. However, our observations show that in classical text data, power-law correlations coexist with near-volume-law scaling of mutual information. This discrepancy highlights the fundamental differences between quantum and classical systems in terms of information structure.

Another significant observation from our findings is the universal scaling of the entire distribution, as shown in Fig. 5(b). The consistency of the MI curve as we increase L_{max}, with only a constant rescaling of the overall height, suggests that high levels of mutual information persist even in longer texts. For instance, doubling the text length from $L_{max} = 100$ to $L_{max} = 200$ results in a similar MI curve shape, merely scaled by a constant factor. This scaling behavior indicates that the information structure of text remains consistent across different length scales.

However, we must also consider the potential existence of an intermediate characteristic scale L_{max}, beyond the scope of our numerical analysis, where correlations might exhibit a different decay pattern. Indeed, previous research by Shen [44] identified a length scale at which correlations between individual words disappear. Therefore, it is conceivable that significantly larger values of L_{max} could reveal a different scaling behavior. This possibility underscores the need for further investigation into the information structure of text at various scales.

B. Scaling of correlation functions and a random pair model for text

The observed power-law scaling of MI in text data, combined with the known algebraic decay of correlations between characters or words from the literature [43,44], presents an intriguing puzzle. This behavior differs significantly from what is typically observed in quantum critical systems, where power-law correlations usually lead to logarithmic scaling of entanglement entropy. To better understand this phenomenon and its implications for modeling natural language, we first review the role of correlation functions in quantum and classical systems, then introduce a simplified toy model that captures these key features of text data.

In principle, any probability distribution can be decomposed into the form of Eq. (5) by tabulating every conditional probability $\mathbb{P}(x_i|\boldsymbol{x}_{<i})$. However, this approach becomes inefficient for large systems due to the exponential growth in parameters. Early language models, such as n-grams, addressed this by limiting connectivity:

$$\mathbb{P}(x_{t+1} \mid x_t, \ldots, x_1) = \mathbb{P}(x_{t+1} \mid x_t, \ldots, x_{t-n+2}). \tag{11}$$

This Markovian property allows for matrix product state representations [33,109,110], resulting in area-law scaling of mutual information and exponential decay of correlations.

Matrix product states (MPSs) naturally reproduce the typical decay of correlations characteristic of gapped systems, which further explains why MPSs effectively represent ground states of gapped models. Specifically, the correlations between two sites i and j are primarily created through the tensors in the shortest path connecting them. Mathematically, via transfer operators, the two-point correlation function $C_{\mathrm{MPS}}(A_i, B_j) = \langle A_i B_j \rangle - \langle A_i \rangle \langle B_j \rangle$ in a constant bond dimension MPS decays exponentially in the asymptotic limit:

$$C_{\mathrm{MPS}}(A_i, B_j) \propto e^{-|i-j|/\xi}, \tag{12}$$

for some correlation length $\xi > 0$.

Previous research [18,20,43,44] has identified algebraic decay in correlations between characters or individual words in natural language, a feature reminiscent of critical systems in physics [111]. In these critical physical systems, gapless excitations with infinite range lead to power-law-decaying correlations

$$C(A_i, B_j) \propto |i-j|^{-\alpha}, \tag{13}$$

where $\alpha \geqslant 0$ is some exponent. In quantum critical models, this implies logarithmic scaling of the entanglement entropy [76–78]. For a finite system with N sites, it is possible to increase the bond dimension D of the matrix product state polynomially with N to reproduce algebraic correlations at long distances. In contrast, in tree tensor networks and multiscale entanglement renormalization ansatz (MERA), correlations decay algebraically, as required in gapless models.

This observation led Lin and Tegmark to construct a character-level statistical language model that exhibits similar long-range correlations [20,43]. In their model, long-range correlations emerge from hidden variables representing linguistic structures and meanings. This binary Markov tree-based model can then be represented by tree tensor networks (TTNs), and is expected to result in power-law correlations and logarithmic mutual information scaling. Based on this analogy, it has been argued that natural languages exhibit critical behavior [18,43], and therefore one could expect languages to behave like a critical quantum system. Therefore, languages should possess critical properties of mutual information, which can be considered analogous to entanglement in classical systems [42].

However, our observations indicate a much steeper growth of mutual information, suggesting that long-range correlations play a more significant role than assumed by Lin and Tegmark. This might seem surprising at first, given that critical physical systems typically exhibit logarithmic scaling of entanglement entropy [76–78]. We argue that the algebraic scaling of mutual information in classical data does not contradict the presence of algebraic correlations. This difference arises because natural language data does not exhibit the notion of locality typically imposed in quantum systems, leading to a broader range of possible mutual information scaling behaviors.

To address this puzzle, we introduce a random pair model, a classical analog of the random singlet model introduced in Ref. [22]. This model demonstrates how algebraic correlations in classical probability distributions can coexist with power-law mutual information scaling, in contrast with the logarithmic scaling typically observed in quantum critical systems. This is an initial minimal model reproducing our observations but does not take linguistic structure into account. We capture those aspects in a more realistic model in Sec. V C.

Consider a probability distribution describing a one-dimensional lattice of classical degrees of freedom, such as words. Each word is assigned a coordinate x. The probability distribution is characterized by each site x being correlated with only one other site y, with the pair (x, y) sharing the maximum possible mutual information (Fig. 6 provides an illustration). The correlated pairs are randomly distributed across the lattice, following a probability distribution

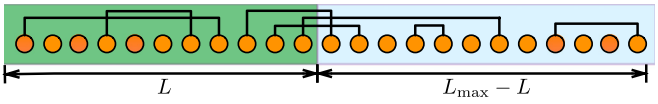


FIG. 6. A schematic presentation of the random pair model. Lattice sites form maximally correlated pairs with one randomly chosen other site following a probability distribution $\rho(x-y)$. The mutual information $I(A:B)$ between two regions A and B is given by the number of links connecting them.

$\rho(x-y) = C|x-y|^{-\alpha}$, where $\alpha > 1$ is a model parameter, and C is a normalization constant. This model exhibits algebraically decaying correlation functions $\propto |x-y|^{-\alpha}$.

In this model, the mutual information is determined by the number of pairs where one element is in region L and the other in R. Hence, we can express the mutual information as

$$I(\mathrm{L}:\mathrm{R}) = \sum_{x=1}^{L} \sum_{y=L+1}^{L_{\max}} \rho(x-y), \tag{14}$$

where $L_{\max}$ is the total length of the system. To extract the leading asymptotic behavior, we can approximate the sum over y by an integral:

$$\rho_L(z) \equiv \sum_{w=z}^{\infty} \rho(w) \approx \int_z^{\infty} C w^{-\alpha} dw = \frac{C}{\alpha - 1} z^{1-\alpha}, \tag{15}$$

for $z > 0$. This allows us to simplify the expression for mutual information:

$$I(\mathrm{L}:\mathrm{R}) \approx \sum_{x=1}^{L} \rho_L(L-x+1) = \frac{C}{\alpha-1} \sum_{x=1}^{L} (L-x+1)^{1-\alpha}. \tag{16}$$

For large L, we can approximate this sum by an integral:

$$I(\mathrm{L}:\mathrm{R}) \approx \frac{C}{\alpha-1} \int_0^L (L-x)^{1-\alpha}\, \mathrm{d}x = \frac{C}{(\alpha-1)(2-\alpha)} L^{2-\alpha}. \tag{17}$$

This result shows that the model exhibits both algebraic correlations and algebraic scaling of mutual information. The exponent of the mutual information scaling, $2-\alpha$, is directly related to the exponent of the correlation decay, α. This relationship provides insight into how different correlation structures in the data can lead to various mutual information scaling behaviors: (i) For $1 < \alpha < 2$, we observe a power-law scaling of mutual information with an exponent between 0 and 1. This regime corresponds to our observations in the WikiText-2 dataset, where we found $\nu \approx 0.82$, implying $\alpha \approx 1.18$. (ii) For $\alpha = 2$, the model would result in logarithmic mutual information scaling, reminiscent of critical quantum systems. (iii) For $\alpha > 2$, the mutual information would saturate to a constant value for large L, corresponding to an area law.

The flexibility of this model in producing different scaling behaviors highlights the rich structure possible in classical data, which can differ significantly from quantum systems. Note that while this random pair model captures the observed mutual information scaling, it may not have an efficient matrix product state representation. However, it can be efficiently

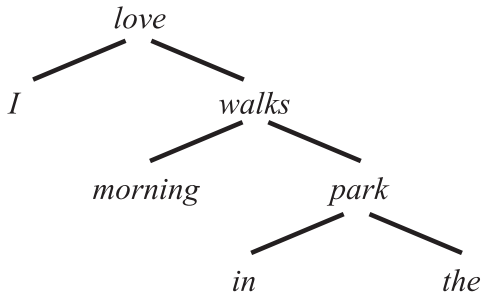


FIG. 7. Example of a dependency parsing tree for the sentence "I love morning walks in the park." The directed edges represent grammatical dependencies between words, forming a hierarchical structure. This tree representation motivates our dependency tree model (Sec. V C), where both tree structure and word choices are generated probabilistically to capture linguistic patterns.

represented by restricted Boltzmann machines [112], suggesting that alternative network architectures might be more suitable for capturing the information structure of natural language.

An interesting limiting case of the random pair model occurs when pairs are uniformly distributed, i.e., $\alpha = 0$. In this scenario, the number of correlated pairs that a fraction of the system can form is proportional to the volume (length in one dimension) of that fraction. Consequently, the number of correlated pairs between two fractions of the system is given by the product of their volumes. This is achieved in the random pair model by setting $\rho(x - y) = C$. As a result, the mutual information scales as

$$I(\mathrm{L} : \mathrm{R}) = |\mathrm{L}||\mathrm{R}| \propto L(L_{\max} - L), \tag{18}$$

where $L_{\max}$ is the total number of lattice sites. This curve serves as a benchmark for assessing the deviation of the data's probability distribution from a scenario where all elements are correlated.

C. Dependency tree model

While the random pair model provides a minimal framework for understanding MI scaling in text, it falls short in capturing the complex linguistic structures inherent in natural language. To provide a more realistic description, we introduce the *dependency tree model*, a generative model that better reflects the grammatical and semantic relationships between words in a sentence and allows us to infer the length distribution of word-to-word mutual information (MI).

In natural text, meaning and grammatical structure lead to complex correlations among words. Building on Chomsky's foundational work on formal grammars [45,113], a paradigmatic tool to capture these grammatical relations is *dependency parsing* [46,114]. In dependency parsing, the syntactic structure of a sentence is modeled as a tree, where nodes represent words and edges denote directed grammatical relations between them. An example is presented in Fig. 7.

Based on this grammatical structure, we introduce the *dependency tree model*, a generative model for grammatical relations between words during text generation (see Fig. 8). In this model, the dependency tree structure is sampled from a random distribution (discussed below), and the words of the sentence are generated using a Markov model.

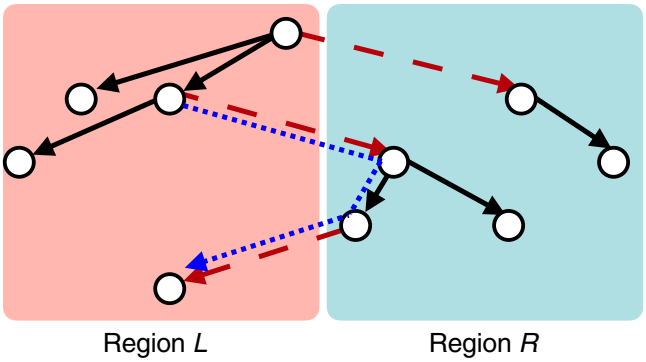


FIG. 8. Schematic of the dependency tree model for mutual information calculation in text. Words (white circles) form nodes in a dependency parsing tree, with black arrows indicating grammatical dependencies sampled from a random distribution. The tree structure is sampled from a random distribution, while words are generated via a Markov process. For calculating mutual information between regions L and R via Eq. (20), dashed red arrows indicate boundary crossings contributing positive MI terms I(crossing), while dotted blue lines represent return paths requiring subtraction of I(return) terms to avoid overcounting. Under the Markov assumption, these contributions fully determine the mutual information between regions.

While the random pair model introduces correlations between pairs of words independently, the dependency tree model captures more complex linguistic dependencies that more closely resemble correlations in natural language. Within this model, MI scaling provides additional information about linguistic patterns. The discussion in this section up to Eq. (20) is general. By making some further assumptions, we analytically find that the length distribution of edges in the dependency parsing tree also follows a power law with an exponent $\nu - 2$, where ν is the exponent of the MI scaling. These simplifying assumptions are not mandatory; the model can be made more complex to capture additional aspects of natural text. In such cases, the MI scaling can be modeled numerically. The main results of this calculation are presented here, while detailed derivations can be found in Appendix C.

While earlier work by Lin and Tegmark [20,43] proposed a generative model relying on a fixed, binary tree-shaped graphical model, our dependency tree model offers more flexibility. In their model, the bottom row of the graph represents words in a sentence, while higher nodes encode meaning and grammatical relations between the words. While this structure correctly predicts power-law correlations between individual words, it produces logarithmic mutual information scaling between regions, as the distributions can be expressed as tree tensor networks (TTNs), which contradicts our observations.

1. Model description and assumptions

The dependency tree model offers more flexibility than the binary tree-based structure of Refs. [20,43], as it allows both the words and the tree structure to be sampled randomly. Each node in the tree corresponds to a word, and the structure of the

tree represents the grammatical dependencies between words. We assume the Markov property within the graph; thus, each word is generated by its parent irrespective of nodes higher in the graph:

$$\mathbb{P}(\text{child} \mid \text{all ancestor words}) = \mathbb{P}(\text{child} \mid \text{parent}). \quad (19)$$

This simplification allows us to focus on how the randomly generated graph structure affects mutual information scaling. While this is a strong simplifying assumption, the model can be extended by allowing the nodes of the tree to carry additional information besides the word itself, such as meaning, similar to how recurrent neural networks use hidden state vectors to carry the meaning of a sentence.

2. Mutual information contributions

The Markov property significantly simplifies the calculation of the mutual information between two regions of text, L (left) and R (right). Specifically, there are only two types of contributions to the mutual information, which we call *crossings* and *returns*:

$$I(L:R) = \sum_{\text{crossings}} I(\text{crossing}) - \sum_{\text{returns}} I(\text{return}), \quad (20)$$

as we derive in Appendix C. These contributions are illustrated in Fig. 8 as dashed and dotted arrows, respectively.

More formally, $I(\text{crossing}) = I(W_p : W_c)$ represents the mutual information between the parent word W_p and the child word W_c of an arrow crossing the boundary between L and R. The correlation between these words introduces mutual information between the two regions, contributing positively to $I(L:R)$. However, these contributions are not independent, and simply summing them would lead to overcounting. To counter this, contributions from *returns* must be subtracted. *Returns* correspond to the shortest path connecting a parent word W_p at the boundary, before the path crosses to the other region, and its earliest descendant W_d in the same region as the parent when the path returns. $I(\text{return})$ indicates the mutual information between the parent and the descendant word, $I(W_p : W_d)$. These MI contributions need to be subtracted to avoid overcounting, as this is the amount of information that is returned to the region where the path started. The derivation of Eq. (20) can be found in Appendix C.

3. Mutual information scaling between regions

To gain analytical insights into the mutual information scaling, we now introduce some simplifying assumptions. These assumptions, while not necessary for numerical simulations, allow us to derive analytical expressions that provide intuition about the model's behavior. The scaling in our model is determined by the random distribution generating the dependency tree. By comparing these analytical results with the power-law scaling of the mutual information observed in Sec. V A, we can infer properties of the dependency structure in real text.

To preserve the tree structure, each word in the generated graph can have at most one incoming and arbitrarily many outgoing edges. Let $q(L)$ be the probability distribution of the length L of these edges, where L is the length along the text direction (positive or negative depending on the direction of the edge). We neglect boundary effects and assume that $q(L)$ is uniform across the text.

Based on our discussion in Appendix C above Eq. (C11), we assume that all MI contributions from *crossings* are identical. Furthermore, Eq. (C15) in Appendix C shows that under certain simplifying assumptions, the *return* terms decay exponentially. Motivated by this, we neglect contributions from *returns*. This approximation also serves as a strict upper bound, as these terms only decrease the mutual information in Eq. (20).

Under these assumptions, the mutual information between two regions becomes proportional to the number of *crossings* across the boundary. While these simplifications do not capture word-word correlation quantified by MI in real text with perfect accuracy, they provide a reasonable approximation for our analysis.

Let $\mathrm{Cr}(L)$ be the average number of crossings between regions L and R of lengths L and $L_{\max} - L$, where $L_{\max}$ is the total text length. Then,

$$\begin{aligned}\mathrm{Cr}(L) = &\sum_{j=1}^{L_{\max}-1} \min(L, j, L_{\max} - j) q(j) \\ &+ \sum_{j=1}^{L_{\max}-1} \min(L_{\max} - L, j, L_{\max} - j) q(-j), \end{aligned} \quad (21)$$

where the first line corresponds to the expected number of crossings from L to R and the second line vice versa. Although this is not a simple expression to analyze, its first and second discrete derivatives are significantly more tractable. Assuming without loss of generality that the left region is shorter, $L < L_{\max}/2$, then the first discrete derivative is

$$\mathrm{Cr}(L+1) - \mathrm{Cr}(L) = \sum_{j=L+1}^{L_{\max}/2} q(j) - \sum_{j=N/2+1}^{L_{\max}-L-1} q(-j). \quad (22)$$

The second discrete derivative is particularly informative:

$$\begin{aligned}&\mathrm{Cr}(L+2) + \mathrm{Cr}(L) - 2\,\mathrm{Cr}(L+1) \\ &\quad = q(L+1) + q(L_{\max} - L). \end{aligned} \quad (23)$$

This is precisely the symmetrized edge length distribution. In our simple model, this must be proportional to the second derivative of the mutual information curve in Fig. 5(c). Based on the power-law scaling observed in Sec. V A for short and intermediate scales, we expect

$$q(L+1) + q(L_{\max} - L) \propto L^{\nu-2}$$

for small L, where $\nu \approx 0.82$ as observed in the WikiText-2 dataset. We note, however, that our approximation of vanishing *return* contributions is expected to be more accurate on longer scales.

Finally, we mention that the simplifying assumptions about the *crossing* and *return* terms were only necessary for the analytical calculation of the scaling. To model language more accurately, these assumptions can be relaxed. In such cases, numerical simulations based on Eq. (20) can provide more precise estimations of the scaling across all length scales, albeit with fewer analytical insights.

VI. SUMMARY AND OUTLOOK

In this paper, we have explored the application of mutual information (MI) as a tool for analyzing natural datasets, leveraging the strong interplay between machine-learning and tensor network representations from quantum many-body theory. Our investigation uncovers several key insights into the structure of text and image data, with important implications for their efficient representation and processing.

For text data, we observed a power-law scaling of MI, suggesting that traditional one-dimensional tensor network approaches like matrix product states (MPS) and tree tensor networks (TTN) are not optimal for representing long texts. This contrasts with quantum systems, where power-law-decaying correlations typically lead to logarithmic entanglement entropy scaling; hence MPS states provide efficient numerical descriptions. Our results indicate that classical text data exhibit a fundamentally different information structure, with power-law correlations coexisting with near-volume-law MI scaling. To better understand this phenomenon, we introduced a random pair model and an enhanced Markov generative model based on dependency parsing trees, which capture linguistic dependencies more accurately than the former. Both models successfully reproduce the observed mutual information scaling and the scaling of the correlation functions by carefully choosing the power-law distribution for the lengths of dependencies between words. This suggests that the hierarchical and statistical properties of natural language play a significant role in shaping its information structure.

For image data, our findings were more nuanced. For simpler datasets like MNIST [47], we observed a clear area-law scaling when the data were made translationally invariant. This result aligns well with previous successes in applying tensor-network-based machine-learning algorithms to MNIST classification tasks [14,33,38,102]. However, for more complex image datasets such as Fashion-MNIST [48] and CIFAR-10 [49], the results were less definitive. While the MI for center-surrounding partitions adhered to an area law, the top-bottom partitions scaled more rapidly, indicating a deviation from area-law scaling. This suggests that more sophisticated tensor network architectures [33,115] or hybrid models combining tensor networks with neural networks [104] might be necessary for effectively representing and processing these more complex image datasets. Further refinement and exploration of these hybrid models could pave the way for scalable tensor network applications in diverse machine-learning tasks.

Our study underscores the potential of MI as a theoretical tool to guide the selection, improvement, and evaluation of machine-learning models. Just as entanglement entropy and MI have played a crucial role in developing algorithms for quantum many-body physics, MI could serve a similar function in machine learning, aiding in capturing the necessary representation power to efficiently characterize complex datasets.

Several exciting avenues for future research emerge from our work. One promising direction is to characterize the MI scaling of distributions learned by state-of-the-art neural networks, such as gated recurrent units [116] and transformers [105]. Such an analysis could provide insights into why contemporary language models, like BERT [117] and GPT [118], outperform earlier models like recurrent neural networks [88] and long short-term memory networks [106] in generating coherent text over long sequences. Additionally, studying the dynamics of MI across the layers of deep learning networks, which has already offered new perspectives on their learning and information processing capabilities [9,119,120], could be a fruitful area of future research. This could involve exploring how MI between data subregions evolves as neural networks process images or text sequentially. Another intriguing line of inquiry involves identifying other information-theoretic measures that could further our understanding of how datasets occupy only a small fraction of the entire parameter space and the traits that enable their efficient compression by tensor networks or analysis via neural networks. Advancements in these areas could enhance machine-learning architectures and contribute to demystifying the inner workings of neural networks.

ACKNOWLEDGMENTS

We thank Álvaro M. Alhambra, Georgios Styliaris, and Samuel Scalet for fruitful discussions. This research is part of the Munich Quantum Valley, which is supported by the Bavarian state government with funds from the Hightech Agenda Bayern Plus. The work is partially supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy—EXC-2111—390814868. I.K. is supported by the Max-Planck-Harvard Research Center for Quantum Optics (MPHQ).

APPENDIX A: MORE DETAILS ON THE k-NEAREST-NEIGHBOR ESTIMATOR AND FURTHER ANALYSIS

Throughout this paper, we have utilized the well-known k-nearest-neighbor (kNN) estimator [64] as a benchmark for other estimators. The kNN estimator is particularly effective for low-dimensional data. A key feature of this estimator is the additive constant in its mutual information (MI) estimates, dependent on the number n_{data} of input data points and the parameter k, representing the number of nearest neighbors used in the density estimation step (refer to Sec. III for more details).

In Fig. 9, we present the MI estimates derived from the kNN method for the MNIST dataset without applying translational invariance to the images. Our empirical findings indicate that the scaling depends solely on the ratio k/n_{data}. Due to this dependency, we adjust the kNN results by a global additive constant when comparing with other methods unaffected by this issue.

Figure 9 also illustrates the impact of the blank areas located at the edges of MNIST images, as the MI diminishes towards these edges. Moreover, the curves for the top-bottom partition do not exhibit the flat plateau observed in Fig. 4(a), indicative of area-law behavior in translationally invariant MNIST images. This deviation arises from the edges suppressing MI values, even near the center of the images.

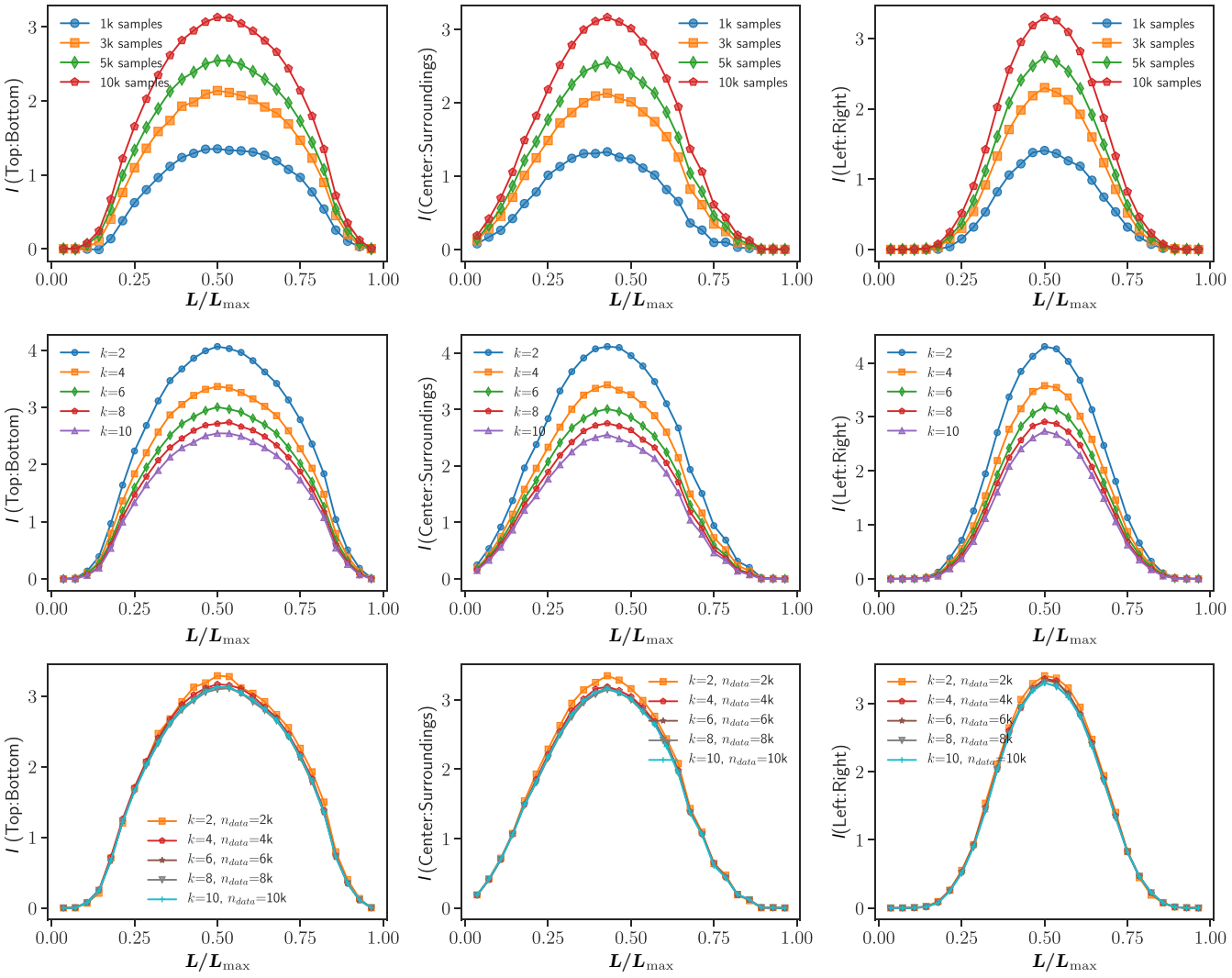


FIG. 9. Detailed analysis of the kNN estimation of MI in the MNIST dataset. (top row, from left to right) Estimated $I(\mathrm{L}:\mathrm{R})$, $I(\mathrm{C}:\mathrm{S})$, $I(\mathrm{T}:\mathrm{B})$ with $k = 10$ for 1k, 3k, 5k, 10k samples. (second row, from left to right) Estimated $I(\mathrm{L}:\mathrm{R})$, $I(\mathrm{C}:\mathrm{S})$, $I(\mathrm{T}:\mathrm{B})$ with 5k samples and $k = 2,\ 4,\ 6,\ 8,\ 10$. The kNN estimation is influenced by the parameters k and the number of samples used. (bottom row) Estimated $I(\mathrm{L}:\mathrm{R})$, $I(\mathrm{C}:\mathrm{S})$, $I(\mathrm{T}:\mathrm{B})$ with $k \propto n_{\mathrm{data}}$. For $k \propto n_{\mathrm{data}}$, all curves collapse, yielding a universal estimate. The x axis represents the normalized length of the left, central, or top region, L/L_{max}, and the y axis shows the mutual information $I(A:B)$. These results demonstrate the consistency and scalability of the kNN estimator, as well as its sensitivity to dataset characteristics such as edge effects.

APPENDIX B: MORE DETAILS ON THE MUTUAL INFORMATION NEURAL ESTIMATOR AND FURTHER ANALYSIS

In this Appendix, we provide additional technical details on our implementation of the mutual information neural estimator (MINE) [65]. This approach allows for the flexible use of various neural networks as score functions. As a lower-bound estimator, the accuracy of the MI estimation can be enhanced by achieving a higher MI estimate using a more expressive score function.

Initially, as per Belghazi *et al.* [65], we used a fully connected feedforward neural network (FC-FFNN) as the score function T_θ. To leverage a network with greater expressive power and suitability for images and text, we switched to using convolutional neural networks (CNNs) as T_θ. The CNN architecture consists of multiple layers: a three-dimensional (3D) convolution layer, a two-dimensional (2D) max-pooling layer, and a dropout layer [121] with a rate of 0.15. This is followed by a fully connected flattening layer with 0.55 dropout regularization. The last layer is then connected to the final fully connected layer with a single output. This output serves as the score function T_θ in Eq. (9). We employed the Adam optimizer [122,123] with a batch size of 128 and a learning rate of 10^{-4} during training. The activation function used was ReLU [124].

We compared the above-described CNN with a FC-FFNN, composed of an input layer matching the data dimensionality (e.g., 28×28 for MNIST), a hidden layer with 500 neurons using ReLU activation, and a subsequent output layer with a single linear neuron. Both network types were trained under identical conditions regarding optimizer settings (learning rate, batch size) and training epochs. The mutual information neural estimation procedure was implemented using PyTorch [125].

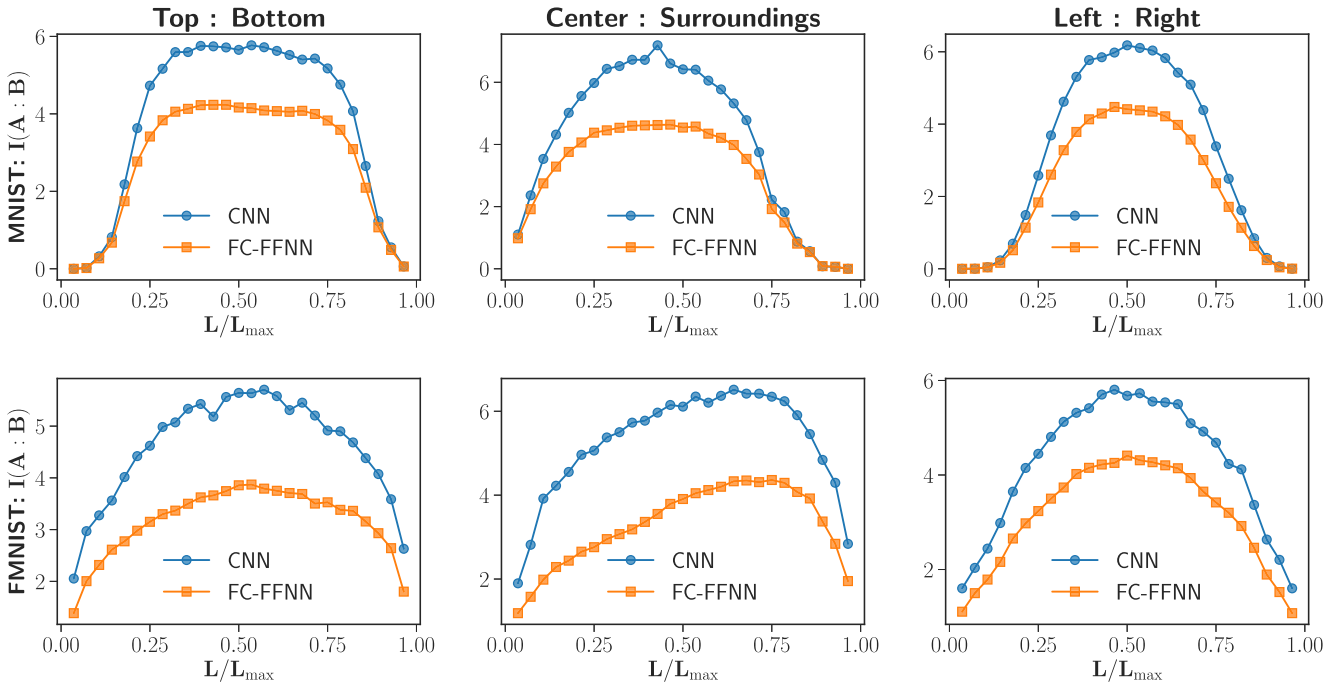


FIG. 10. Detailed comparison of MINE estimation of MI in MNIST and Fashion-MNIST datasets using different score functions. (top row) $I(\mathrm{L}:\mathrm{R})$, $I(\mathrm{C}:\mathrm{S})$, $I(\mathrm{T}:\mathrm{B})$ estimates when T_θ is a convolutional neural network (CNN). (bottom row) $I(\mathrm{L}:\mathrm{R})$, $I(\mathrm{C}:\mathrm{S})$, $I(\mathrm{T}:\mathrm{B})$ estimates when T_θ is a fully connected feedforward neural network (FC-FFNN). The CNN consistently outperforms the FC-FFNN by providing higher MI estimates, indicating greater accuracy and stability. The x axis represents the normalized length of the left, central, or top region, $L/L_{\max}$, and the y axis shows the mutual information $I(A:B)$ in bits. These comparisons demonstrate the importance of choosing appropriate neural network architectures for accurate MI estimation in complex datasets.

In Fig. 10, we compare the performance of MINE using FC-FFNN and CNN score functions on MNIST and Fashion-MNIST data. We found that CNNs consistently outperform FC-FFNNs by providing higher MI values. Consequently, we utilized CNN-based MINE results throughout the main text.

APPENDIX C: DEPENDENCY TREE MODEL

In this Appendix, we provide detailed calculations for the dependency tree model introduced in Sec. V C. First, we derive Eq. (20) in the simplified case of a linear dependency tree, as depicted in Fig. 11(a). This simplification is for notational convenience, and the derivations are similar in the general case.

In this linear tree of length n, the sequence starts with words in the left region, $L_1 = (W_1, W_2, \ldots, W_i)$, crosses into the right region, $R = (W_{i+1}, W_{i+2}, \ldots, W_j)$, and then returns to the left region with words $L_2 = (W_{j+1}, W_{j+2}, \ldots, W_n)$. The mutual information between the left $L = L_1 \cup L_2$ and right R regions is given by

$$\begin{aligned} I(L:R) &= S(L) + S(R) - S(L,R) \\ &= S(L_1,L_2) + S(R) - S(L_1,L_2,R) \\ &= [S(L_1) + S(L_2) - I(L_1:L_2)] \\ &\quad + S(R) - S(L_1,L_2,R), \end{aligned} \tag{C1}$$

where $I(L_1:L_2)$ is the mutual information between L_1 and L_2.

Assuming the Markov property—that each word depends only on its parent—we can express the entropy of a sequence of words from W_l to W_m as

$$S(W_l,\ldots,W_m) = S(W_l) + \sum_{k=l}^{m-1} S(W_{k+1} \mid W_k). \tag{C2}$$

Applying this to the terms in Eq. (C1) and simplifying, many terms cancel telescopically, resulting in:

$$I(L:R) = I(W_{i+1}:W_i) + I(W_{j+1}:W_j) - I(L_1:L_2). \tag{C3}$$

Here, $I(W_{i+1}:W_i)$ and $I(W_{j+1}:W_j)$ represent the MI contributions from the *crossings* at the boundaries between L and R

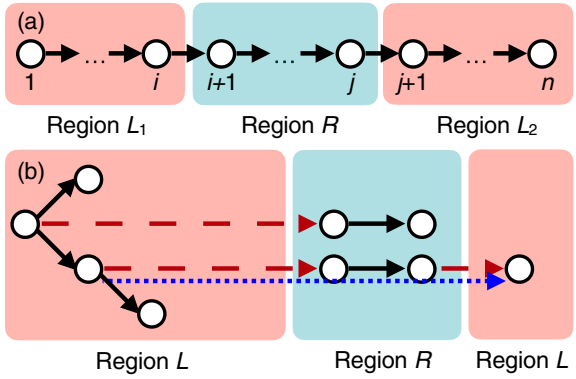


FIG. 11. Markov chain model of text. (a) The tree in Fig. 8 expanded horizontally such that boundary crossings only occur from left to right. (b) A simplified case of (a) when the tree consists of a single path.

[sites $(i, i+1)$ and $(j, j+1)$]. The term $I(L_1 : L_2)$ accounts for the MI within the left region due to the *returns*—paths that start and end in L after passing through R, as we show below. The joint probability distribution of the words in L_1 and L_2 can be written as

$$\mathbb{P}(L_1, L_2) = \mathbb{P}(L_1)\mathbb{P}(L_2)\frac{\mathbb{P}(W_{j+1}|W_i)}{\mathbb{P}(W_{j+1})}, \tag{C4}$$

due to the Markov property. Consequently, the entropy of the two regions is

$$S(L_1, L_2) = S(L_1) + S(L_2) - I(W_{j+1} : W_i). \tag{C5}$$

Hence, the mutual information between the left and right regions is given by the mutual information between the words at the boundary,

$$I(L_1 : L_2) = S(L_1) + S(L_2) - S(L_1, L_2) \tag{C6}$$

$$= I(W_{j+1} : W_i), \tag{C7}$$

which is a *return* term.

1. General case: Branching dependency tree

In the general case, where the dependency tree has multiple branches, as depicted in Fig. 11(b), the same reasoning applies independently to each branch connecting a parent word to one of its descendants. An analogous derivation leads to Eq. (20). Assuming Markovian language generation, this formula is exact. This provides an opportunity to numerically benchmark various language models against empirical observations on real text, as in Sec. V A. This will be addressed in future work. In the remainder of this section, we aim to gain analytical understanding of the MI scaling based on the simplifying assumptions mentioned in Sec. V C.

2. Mutual information estimate between individual words

Let us denote the Markov matrix for word generation $\mathbf{M}$. We make the standard assumption that $\mathbf{M}$ is both irreducible and aperiodic, thus it has a unique stationary distribution $\boldsymbol{\pi} = \mathbf{M}\boldsymbol{\pi}$. According to the Perron–Frobenius theorem, all other eigenvalues λ_i satisfy $1 > |\lambda_2| \geqslant |\lambda_3| \geqslant \cdots \geqslant 0$. With the left and right eigenvectors $\mathbf{l}_i$ and $\mathbf{r}_i$, $\mathbf{M}$ has the eigendecomposition:

$$\mathbf{M} = \boldsymbol{\pi}\mathbf{1}^{\mathrm{T}} + \sum_{i=2}^{m} \lambda_i \mathbf{r}_i \mathbf{l}_i^{\mathrm{T}}, \tag{C8}$$

where m is the dimensionality of the word space and $\mathbf{1} = (1, 1, \ldots, 1)^{\mathrm{T}}$ is a constant column vector. Due to the biorthogonality between left and right eigenvectors $\mathbf{l}_i^{\mathrm{T}}\mathbf{r}_j = \delta_{ij}$, the nth power of the Markov matrix is

$$\mathbf{M}^n = \boldsymbol{\pi}\mathbf{1}^{\mathrm{T}} + \sum_{i=2}^{m} \lambda_i^n \mathbf{r}_i \mathbf{l}_i^{\mathrm{T}}, \tag{C9}$$

which decays exponentially with n, with the decay rate dominated by the second largest eigenvalue λ_2.

Next, let us estimate the mutual information between two words within the graph. Using Eq. (20), we relied on this result to estimate MI scaling between text regions in Sec. V C.

The crossing terms in Eq. (20) are between a parent word W_p and its child word W_c in the tree structure where the distribution of W_p is given by $\boldsymbol{\mu}$. The joint distribution of W_p and W_c is given by $\mathbb{P}(W_p = b, W_c = a) = \mu_b M_{ba}$, where $M_{ba} = \mathbb{P}(W_c = a|W_p = b)$. The marginal distribution of W_c is $\mathbb{P}(W_c = a) = (\mathbf{M}^{\mathrm{T}}\boldsymbol{\mu})_a$. Thus, the mutual information is

$$I(W_p : W_c) = \sum_{a,b} \mu_b M_{ba} \ln\left(\frac{M_{ba}}{(\mathbf{M}^{\mathrm{T}}\boldsymbol{\mu})_a}\right). \tag{C10}$$

In our approximations, we assume that $\boldsymbol{\mu}$ is close to the stationary distribution $\boldsymbol{\mu} \approx \boldsymbol{\pi}$. Thus, the mutual information is approximately constant for *crossings* in the tree graph:

$$I(\text{crossing}) \approx \mathcal{C}_c. \tag{C11}$$

Assume that words U and V, with word distribution vectors $\boldsymbol{\mu}_U$ and $\boldsymbol{\mu}_V$, share an earliest common ancestor X in the tree, with distances Δ_U and Δ_V to U and V, respectively. Similarly to the previous linear case, we assume that the distribution of X is close to the stationary one $\boldsymbol{\mu}_X \approx \boldsymbol{\pi}$. To derive a similar formula for *returns*, we use much of the notation from Ref. [43] and their assumption that the distance between words is large enough so that the words are approximately independent from each other,

$$\mathbb{P}(U = u, V = v) \approx \mathbb{P}(U = u)\mathbb{P}(V = v). \tag{C12}$$

The mutual information can thus be approximated as [43]

$$\begin{aligned} I(U : V) &= \sum_{u,v} \mathbb{P}(u, v) \ln\left(\frac{\mathbb{P}(u, v)}{\mathbb{P}(u)\mathbb{P}(v)}\right) \\ &= \sum_{u,v} \mathbb{P}(u, v) \ln\left(1 + \frac{\mathbb{P}(u, v)}{\mathbb{P}(u)\mathbb{P}(v)} - 1\right) \\ &\lesssim \sum_{u,v} \mathbb{P}(u, v)\left(\frac{\mathbb{P}(u, v)}{\mathbb{P}(u)\mathbb{P}(v)} - 1\right) \\ &= \sum_{u,v} \frac{\mathbb{P}(u, v)^2}{\mathbb{P}(u)\mathbb{P}(v)} - 1. \end{aligned} \tag{C13}$$

Note that this is a strict upper bound due to Jensen's inequality [$\ln(1 + x) \leqslant x$ for $x \geqslant 0$], which is a good approximation if Eq. (C12) holds. The joint probability distribution can be further approximated up to leading order in the eigenvalues of $\mathbf{M}$ as

$$\begin{aligned} \mathbb{P}(u, v) &\approx \sum_x \pi_x\left(\pi_u + \lambda_2^{\Delta_U} A_{xu}\right)\left(\pi_v + \lambda_2^{\Delta_V} A_{xv}\right) \\ &= \pi_u\pi_v + \lambda_2^{\Delta_U + \Delta_V} \sum_x \pi_x A_{xu} A_{xv}, \end{aligned}$$

where the matrix $\mathbf{A} = \mathbf{r}_2\mathbf{l}_2^T$ is the projector to the subspace with the second largest eigenvalue. A_{xu} captures the influence of ancestor $X = x$ on word $U = u$. In the formula above, the cross terms vanish for this reason, since π_x belongs to the subspace of $\lambda_1 = 1$. Similarly, in Eq. (C13), the cross terms in $\mathbb{P}^2(u, v)$ vanish since $\sum_u A_{xu}\pi_u = 0$. Hence, up to the same order, the mutual information becomes

$$I(U : V) \approx \left(\sum_{u,v} \frac{\left(\sum_x \pi_x A_{xu} A_{xv}\right)^2}{\pi_u\pi_v}\right)\lambda_2^{2(\Delta_U + \Delta_V)}. \tag{C14}$$

This expression indicates that the MI between two words U and V decays exponentially with their combined distances from their earliest common ancestor in the dependency tree. Therefore, the *return* contributions in Eq. (20) scale with dist, the number of edges between them in the graph, as

$$I(\text{return}) \approx \mathcal{C}_r \lambda_2^{2\ \text{dist}} + O\left(\left|\frac{\lambda_3}{\lambda_2}\right|^{2\ \text{dist}}\right). \tag{C15}$$

[1] Y. LeCun, Y. Bengio, and G. Hinton, Deep learning, Nature (London) **521**, 436 (2015).

[2] M. I. Jordan and T. M. Mitchell, Machine learning: Trends, perspectives, and prospects, Science **349**, 255 (2015).

[3] A. Krizhevsky, I. Sutskever, and G. E. Hinton, Imagenet classification with deep convolutional neural networks, in *Advances in Neural Information Processing Systems* (Curran Associates, Inc., Red Hook, NY, 2012), Vol. 25, pp. 1097–1105.

[4] Y. Kim, Convolutional neural networks for sentence classification, in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (Association for Computational Linguistics, Stroudsburg, PA, 2014), pp. 1746–1751.

[5] A. Van Den Oord, N. Kalchbrenner, and K. Kavukcuoglu, Pixel recurrent neural networks, in *33rd International Conference on Machine Learning, ICML 2016* (IMLS, Corvallis, OR, 2016), Vol. 4, pp. 2611–2620.

[6] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, Generative adversarial networks, in *Advances in Neural Information Processing Systems* (Curran Associates, Inc., Red Hook, NY, 2014), Vol. 3, pp. 2672–2680.

[7] A. Van Den Oord, N. Kalchbrenner, O. Vinyals, L. Espeholt, A. Graves, and K. Kavukcuoglu, Conditional image generation with PixelCNN decoders, in *Advances in Neural Information Processing Systems* (Curran Associates, Inc., Red Hook, NY, 2016), pp. 4797–4805.

[8] Y. Bengio, A. Courville, and P. Vincent, Representation learning: A review and new perspectives, IEEE Trans. Pattern Anal. Mach. Intell. **35**, 1798 (2013).

[9] N. Tishby and N. Zaslavsky, Deep learning and the information bottleneck principle, in *2015 IEEE Information Theory Workshop, ITW 2015* (Institute of Electrical and Electronics Engineers Inc., Piscataway, NJ, 2015).

[10] N. Cohen, O. Sharir, and A. Shashua, On the expressive power of deep learning: A tensor analysis, J. Mach. Learn. Res. **49**, 698 (2015).

[11] Y. Levine, D. Yakira, N. Cohen, and A. Shashua, Deep learning and quantum entanglement: Fundamental connections with implications to network design, in *6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings* (OpenReview.net, 2018).

[12] Y. Levine, O. Sharir, N. Cohen, and A. Shashua, Quantum entanglement in deep learning architectures, Phys. Rev. Lett. **122**, 065301 (2019).

[13] Y.-H. Zhang, Entanglement entropy of target functions for image classification and convolutional neural network, arXiv:1710.05520.

[14] Y. Huang, Provably efficient neural network representation for image classification, arXiv:1711.04606.

[15] Y. Bahri, J. Kadmon, J. Pennington, S. S. Schoenholz, J. Sohl-Dickstein, and S. Ganguli, Statistical mechanics of deep learning, Annu. Rev. Condens. Matter Phys. **11**, 501 (2020).

[16] C. Bény, Deep learning and the renormalization group, arXiv:1301.3124.

[17] P. Mehta and D. J. Schwab, An exact mapping between the variational renormalization group and deep learning, arXiv:1410.3831.

[18] Á. J. Gallego and R. Orús, Language design as information renormalization, SN Comput. Sci. **3**, 140 (2022).

[19] M. Koch-Janusz and Z. Ringel, Mutual information, neural networks and the renormalization group, Nat. Phys. **14**, 578 (2018).

[20] H. W. Lin, M. Tegmark, and D. Rolnick, Why does deep and cheap learning work so well? J. Stat. Phys. **168**, 1223 (2017).

[21] J. Eisert, M. Cramer, and M. B. Plenio, *Colloquium*: Area laws for the entanglement entropy, Rev. Mod. Phys. **82**, 277 (2010).

[22] M. M. Wolf, F. Verstraete, M. B. Hastings, and J. I. Cirac, Area laws in quantum systems: Mutual information and correlations, Phys. Rev. Lett. **100**, 070502 (2008).

[23] J. I. Cirac, D. Perez-Garcia, N. Schuch, and F. Verstraete, Matrix product states and projected entangled pair states: Concepts, symmetries, and theorems, Rev. Mod. Phys. **93**, 045003 (2021).

[24] U. Schollwöck, The density-matrix renormalization group in the age of matrix product states, Ann. Phys. (NY) **326**, 96 (2011).

[25] N. Schuch, M. M. Wolf, F. Verstraete, and J. I. Cirac, Entropy scaling and simulability by matrix product states, Phys. Rev. Lett. **100**, 030504 (2008).

[26] F. Verstraete and J. I. Cirac, Matrix product states represent ground states faithfully, Phys. Rev. B **73**, 094423 (2006).

[27] G. Vidal, Class of quantum many-body states that can be efficiently simulated, Phys. Rev. Lett. **101**, 110501 (2008).

[28] Y.-Y. Shi, L.-M. Duan, and G. Vidal, Classical simulation of quantum many-body systems with a tree tensor network, Phys. Rev. A **74**, 022320 (2006).

[29] E. Robeva and A. Seigal, Duality of graphical models and tensor networks, Inf. Inference J. IMA **8**, 273 (2019).

[30] I. Glasser, N. Pancotti, and J. I. Cirac, From probabilistic graphical models to generalized tensor networks for supervised learning, IEEE Access **8**, 68169 (2020).

[31] Z.-Y. Han, J. Wang, H. Fan, L. Wang, and P. Zhang, Unsupervised generative modeling using matrix product states, Phys. Rev. X **8**, 031012 (2018).

[32] E. M. Stoudenmire, Learning relevant features of data with multi-scale tensor networks, Quantum Sci. Technol. **3**, 034003 (2018).

[33] I. Glasser, R. Sweke, N. Pancotti, J. Eisert, and J. I. Cirac, Expressive power of tensor-network factorizations for probabilistic modeling, in *Advances in Neural Information*

Processing Systems (Curran Associates, Inc., Red Hook, NY, 2019), pp. 1498–1510.

[34] S. Cheng, L. Wang, T. Xiang, and P. Zhang, Tree tensor networks for generative modeling, Phys. Rev. B **99**, 155131 (2019).

[35] S. Efthymiou, J. Hidary, and S. Leichenauer, TensorNetwork for machine learning, arXiv:1906.06329.

[36] W. Huggins, P. Patel, K. B. Whaley, E. M. Stoudenmire, P. Patil, B. Mitchell, K. B. Whaley, and E. M. Stoudenmire, Towards quantum machine learning with tensor networks, Quantum Sci. Technol. **4**, 024001 (2019).

[37] J. Miller, G. Rabusseau, and J. Terilla, Tensor networks for probabilistic sequence modeling, in *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics*, Proceedings of Machine Learning Research, edited by A. Banerjee and K. Fukumizu (PMLR, 2021), Vol. 130, pp. 3079–3087.

[38] S. Cheng, L. Wang, and P. Zhang, Supervised learning with projected entangled pair states, Phys. Rev. B **103**, 125117 (2021).

[39] G. Carleo, I. Cirac, K. Cranmer, L. Daudet, M. Schuld, N. Tishby, L. Vogt-Maranto, and L. Zdeborová, Machine learning and the physical sciences, Rev. Mod. Phys. **91**, 045002 (2019).

[40] A. Novikov, D. Podoprikhin, A. Osokin, and D. Vetrov, Tensorizing neural networks, in *Advances in Neural Information Processing Systems* (2015), Vol. 28, pp. 442–450.

[41] Z.-F. Gao, S. Cheng, R.-Q. He, Z. Y. Xie, H.-H. Zhao, Z.-Y. Lu, and T. Xiang, Compressing deep neural networks by matrix product operators, Phys. Rev. Res. **2**, 023300 (2020).

[42] J. Wilms, M. Troyer, and F. Verstraete, Mutual information in classical spin models, J. Stat. Mech. (2011) P10011.

[43] H. W. Lin and M. Tegmark, Critical behavior in physics and probabilistic formal languages, Entropy **19**, 299 (2017).

[44] H. Shen, Mutual information scaling and expressive power of sequence models, arXiv:1905.04271.

[45] N. Chomsky, *Syntactic Structures* (Mouton, The Hague, 1957).

[46] S. Kübler, R. McDonald, and J. Nivre, *Dependency Parsing*, Synthesis Lectures on Human Language Technologies (Morgan & Claypool Publishers, San Rafael, CA, 2009).

[47] Y. LeCun, C. Cortes, and C. J. C. Burges, MNIST handwritten digit database (1998), http://yann.lecun.com/exdb/mnist/.

[48] H. Xiao, K. Rasul, and R. Vollgraf, Fashion-MNIST: A novel image dataset for benchmarking machine learning algorithms, arXiv:1708.07747.

[49] A. Krizhevsky, V. Nair, and G. Hinton, CIFAR-10 (Canadian Institute for Advanced Research) (2009), http://www.cs.toronto.edu/~kriz/cifar.html.

[50] F. Verstraete and J. I. Cirac, Renormalization algorithms for quantum-many body systems in two and higher dimensions, arXiv:cond-mat/0407066.

[51] V. Murg, F. Verstraete, and J. I. Cirac, Variational study of hard-core bosons in a two-dimensional optical lattice using projected entangled pair states, Phys. Rev. A **75**, 033605 (2007).

[52] P. Corboz, S. R. White, G. Vidal, and M. Troyer, Stripes in the two-dimensional t-J model with infinite projected entangled-pair states, Phys. Rev. B **84**, 041108(R) (2011).

[53] B. Bauer, P. Corboz, A. M. Läuchli, L. Messio, K. Penc, M. Troyer, and F. Mila, Three-sublattice order in the SU(3) Heisenberg model on the square and triangular lattice, Phys. Rev. B **85**, 125116 (2012).

[54] P. Corboz, Variational optimization with infinite projected entangled-pair states, Phys. Rev. B **94**, 035133 (2016).

[55] M. Rader and A. M. Läuchli, Finite correlation length scaling in Lorentz-invariant gapless iPEPS wave functions, Phys. Rev. X **8**, 031030 (2018).

[56] B. Ponsioen, S. S. Chung, and P. Corboz, Period 4 stripe in the extended two-dimensional Hubbard model, Phys. Rev. B **100**, 195141 (2019).

[57] B. Vanhecke, J. Hasik, F. Verstraete, and L. Vanderstraeten, A scaling hypothesis for projected entangled-pair states, Phys. Rev. Lett. **129**, 200601 (2022).

[58] P. C. G. Vlaar and P. Corboz, Simulation of three-dimensional quantum systems with projected entangled-pair states, Phys. Rev. B **103**, 205137 (2021).

[59] S. Cheng, J. Chen, and L. Wang, Information perspective to probabilistic modeling: Boltzmann machines versus Born machines, Entropy **20**, 583 (2017).

[60] J. Martyn, G. Vidal, C. Roberts, and S. Leichenauer, Entanglement and tensor networks for supervised image classification, arXiv:2007.06082.

[61] I. Convy, W. Huggins, H. Liao, and K. B. Whaley, Mutual information scaling for tensor network machine learning, Mach. Learn.: Sci. Technol. **3**, 015017 (2022).

[62] A. Torralba, R. Fergus, and W. T. Freeman, 80 million tiny images: A large data set for nonparametric object and scene recognition, IEEE Trans. Pattern Anal. Mach. Intell. **30**, 1958 (2008).

[63] P. Grassberger, Entropy estimates from insufficient samplings, arXiv:physics/0307138.

[64] A. Kraskov, H. Stögbauer, and P. Grassberger, Estimating mutual information, Phys. Rev. E **69**, 066138 (2004).

[65] M. I. Belghazi, A. Baratin, S. Rajeswar, S. Ozair, Y. Bengio, A. Courville, and R. D. Hjelm, MINE: Mutual Information Neural Estimation, in *35th International Conference on Machine Learning, ICML 2018* (International Machine Learning Society, Corvallis, OR, 2018), Vol. 2, pp. 864–873.

[66] J. Song and S. Ermon, Understanding the limitations of variational mutual information estimators, in *Proceedings of ICLR* (OpenReview.net, 2020).

[67] N. Carrara and J. Ernst, On the estimation of mutual information, *39th International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering* (MaxEnt, 2019).

[68] D. McAllester and K. Stratos, Formal limitations on the measurement of mutual information, arXiv:1811.04251.

[69] M. B. Hastings, An area law for one-dimensional quantum systems, J. Stat. Mech. (2007) P08024.

[70] A. Anshu, I. Arad, and D. Gosset, Entanglement subvolume law for 2D frustration-free spin systems, in *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing* (ACM, New York, 2020), pp. 868–874.

[71] K. Van Acoleyen, M. Mariën, and F. Verstraete, Entanglement rates and area laws, Phys. Rev. Lett. **111**, 170501 (2013).

[72] M. Mariën, K. M. R. Audenaert, K. Van Acoleyen, and F. Verstraete, Entanglement Rates and the stability of the area law for the entanglement entropy, Commun. Math. Phys. **346**, 35 (2016).

[73] M. Srednicki, Entropy and area, Phys. Rev. Lett. **71**, 666 (1993).

[74] C. Holzhey, F. Larsen, and F. Wilczek, Geometric and renormalized entropy in conformal field theory, Nucl. Phys. B **424**, 443 (1994).

[75] G. Vidal, J. I. Latorre, E. Rico, and A. Kitaev, Entanglement in quantum critical phenomena, Phys. Rev. Lett. **90**, 227902 (2003).

[76] P. Calabrese and J. Cardy, Entanglement entropy and quantum field theory, J. Stat. Mech. (2004) P06002.

[77] P. Calabrese and J. Cardy, Entanglement entropy and conformal field theory, J. Phys. A: Math. Theor. **42**, 504005 (2009).

[78] H. Casini and M. Huerta, Entanglement entropy in free quantum field theory, J. Phys. A: Math. Theor. **42**, 504007 (2009).

[79] P. Silvi, V. Giovannetti, S. Montangero, M. Rizzi, J. I. Cirac, and R. Fazio, Homogeneous binary trees as ground states of quantum critical Hamiltonians, Phys. Rev. A **81**, 062335 (2010).

[80] Z. Landau, U. Vazirani, and T. Vidick, A polynomial time algorithm for the ground state of one-dimensional gapped local Hamiltonians, Nat. Phys. **11**, 566 (2015).

[81] N. Schuch, M. M. Wolf, K. G. H. Vollbrecht, and J. I. Cirac, On entropy growth and the hardness of simulating time evolution, New J. Phys. **10**, 033032 (2008).

[82] T. Cover and J. Thomas, *Elements of Information Theory* (Wiley-Interscience, Bristol, UK, 2006).

[83] M. B. Hastings, Solving gapped Hamiltonians locally, Phys. Rev. B **73**, 085115 (2006).

[84] A. Molnar, N. Schuch, F. Verstraete, and J. I. Cirac, Approximating Gibbs states of local Hamiltonians efficiently with projected entangled pair states, Phys. Rev. B **91**, 045138 (2015).

[85] C. M. Bishop, *Pattern Recognition and Machine Learning*, Information Science and Statistics (Springer, New York, 2006).

[86] B. Uria, M. A. Cote, K. Gregor, I. Murray, and H. Larochelle, Neural autoregressive distribution estimation, J. Mach. Learn. Res. **17**, 1 (2016).

[87] T. Salimans, A. Karpathy, X. Chen, and D. P. Kingma, PixelCNN++: Improving the PixelCnn with discretized logistic mixture likelihood and other modifications, in *5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings* (OpenReview.net, 2017).

[88] I. Sutskever, J. Martens, and G. E. Hinton, Generating text with recurrent neural networks, in *28th International Conference on Machine Learning, ICML 2017, Bellevue, Washington, June 28 - July 2, 2011*, edited by L. Getoor and T. Scheffer (Omnipress, Madison, WI, 2011), pp. 1017–1024.

[89] I. V. Oseledets, Tensor-train decomposition, SIAM J. Sci. Comput. **33**, 2295 (2011).

[90] R. Salakhutdinov and G. Hinton, An efficient learning procedure for deep Boltzmann machines, Neural Comput. **24**, 1967 (2012).

[91] N. Schuch, M. M. Wolf, F. Verstraete, and J. I. Cirac, Computational complexity of projected entangled pair states, Phys. Rev. Lett. **98**, 140506 (2007).

[92] M. Lubasch, J. I. Cirac, and M.-C. Bañuls, Unifying projected entangled pair state contractions, New J. Phys. **16**, 033014 (2014).

[93] R. Devon Hjelm, K. Grewal, P. Bachman, A. Fedorov, A. Trischler, S. Lavoie-Marchildon, and Y. Bengio, Learning deep representations by mutual information estimation and maximization, in *7th International Conference on Learning Representations, ICLR 2019* (ICLR, 2019), pp. 1–24.

[94] L. Paninski, Estimation of entropy and mutual information, Neural Comput. **15**, 1191 (2003).

[95] K. Gregor, I. Danihelka, A. Mnih, C. Blundell, and D. Wierstra, Deep autoregressive networks, in *31st International Conference on Machine Learning, ICML 2014* (IMLS, Corvallis, OR, 2014), Vol. 4, pp. 2991–3000.

[96] S. Reed, A. Van denOord, N. Kalchbrenner, S. G. Colmenarejo, Z. Wang, Y. Chen, D. Belov, and N. De Freitas, Parallel multiscale autoregressive density estimation, in *34th International Conference on Machine Learning, ICML 2017* (International Machine Learning Society, 2017), Vol. 6, pp. 4447–4456.

[97] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, WaveNet: A generative model for raw audio, arXiv:1609.03499.

[98] B. Poolel, S. Ozair, A. Van Den Oord, A. A. Alemi, and G. Tucker, On variational bounds of mutual information, in *36th International Conference on Machine Learning, ICML 2019* (IMLS, Corvallis, OR, 2019), Vol. 97, pp. 9036–9049.

[99] Pixel_models (2019), https://github.com/kamenbliznashki/pixel_models.

[100] M. D. Donsker and S. R. S. Varadhan, Asymptotic evaluation of certain Markov process expectations for large time. IV, Commun. Pure Appl. Math. **36**, 183 (1983).

[101] NPEET (2019), https://github.com/gregversteeg/NPEET.

[102] E. M. Stoudenmire and D. J. Schwab, Supervised learning with tensor networks, in *Advances in Neural Information Processing Systems* (Curran Associates, Inc., Red Hook, NY, 2016), pp. 4806–4814.

[103] T. Vieijra, L. Vanderstraeten, and F. Verstraete, Generative modeling with projected entangled-pair states, arXiv:2202.08177.

[104] J. Liu, S. Li, J. Zhang, and P. Zhang, Tensor networks for unsupervised machine learning, Phys. Rev. E **107**, L012103 (2023).

[105] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, Attention is all you need, in *Advances in Neural Information Processing Systems* (Curran Associates, Inc., Red Hook, NY, 2017), Vol. 30, pp. 5999–6009.

[106] S. Hochreiter and J. Schmidhuber, Long short-term memory, Neural Comput. **9**, 1735 (1997).

[107] S. Merity, C. Xiong, J. Bradbury, and R. Socher, Pointer sentinel mixture models, arXiv:1609.07843.

[108] J. Pennington, R. Socher, and C. Manning, Glove: Global vectors for word representation, in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (Association for Computational Linguistics, Stroudsburg, 2014), Vol. 19, pp. 1532–1543.

[109] K. Kato and F. G. S. L. Brandão, Quantum approximate Markov chains are thermal, Commun. Math. Phys. **370**, 117 (2019).

[110] X. Gao, E. R. Anschuetz, S.-T. Wang, J. I. Cirac, and M. D. Lukin, Enhancing generative models via quantum correlations, Phys. Rev. X **12**, 021037 (2022).

[111] J. Cardy, *Scaling and Renormalization in Statistical Physics*, Cambridge Lecture Notes in Physics (Cambridge University Press, Cambridge, 1996).

[112] X. Gao and L.-M. Duan, Efficient representation of quantum many-body states with deep neural networks, Nat. Commun. **8**, 662 (2017).

[113] N. Chomsky, *Aspects of the Theory of Syntax* (MIT Press, Cambridge, 1965).

[114] J. Nivre, Dependency grammar and dependency parsing, in *Technical Report* (Research Laboratory of Electronics at the Massachusetts Institute of Technology, Sweden, 2005).

[115] I. Glasser, N. Pancotti, M. August, I. D. Rodriguez, and J. I. Cirac, Neural-network quantum states, string-bond states, and chiral topological states, Phys. Rev. X **8**, 011006 (2018).

[116] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, Empirical evaluation of gated recurrent neural networks on sequence modeling, arXiv:1412.3555.

[117] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Long and Short Papers)*, edited by J. Burstein, C. Doran, and T. Solorio (Association for Computational Linguistics, Minneapolis, 2019), Vol. 1, pp. 4171–4186.

[118] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse *et al.*, Language models are few-shot learners, in *Advances in Neural Information Processing Systems* (Curran Associates, Inc., Red Hook, NY, 2020), Vol. 33, pp. 1877–1901.

[119] N. Tishby, F. C. Pereira, and W. Bialek, The information bottleneck method, arXiv:physics/0004057.

[120] Z. Goldfeld, E. van den Berg, K. Greenewald, I. Melnyk, N. Nguyen, B. Kingsbury, and Y. Polyanskiy, Estimating information flow in deep neural networks, in *36th International Conference on Machine Learning, ICML 2019* (IMLS, Corvallis, OR, 2018), Vol. 97, pp. 4153–4162.

[121] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, Dropout: A simple way to prevent neural networks from overfitting, J. Mach. Learn. Res. **15**, 1929 (2014).

[122] D. P. Kingma and J. L. Ba, Adam: A method for stochastic optimization, in *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings* (OpenReview.net, 2015).

[123] S. J. Reddi, S. Kale, and S. Kumar, On the convergence of adam and beyond, in *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings* (ICLR, 2018).

[124] V. Nair and G. E. Hinton, Rectified linear units improve restricted Boltzmann machines, in *Proceedings of the 27th International Conference on Machine Learning (ICML-10)* (Omnipress, Madison, WI, 2010), pp. 807–814.

[125] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai *et al.*, PyTorch: An imperative style, high-performance deep learning library, in *Advances in Neural Information Processing Systems 32*, edited by H. Wallach, H. Larochelle, A. Beygelzimer, F. d Alche-Buc, E. Fox, and R. Garnett (Curran Associates, Inc., Red Hook, NY, 2019), pp. 8024–8035.