Mapping of attention mechanisms to a generalized Potts model

Riccardo Rende, Federica Gerace, Alessandro Laio, and Sebastian Goldt*
Scuola Internazionale Superiore di Studi Avanzati (SISSA), Via Bonomea 265, 34136 Trieste, Italy

(Received 27 April 2023; revised 14 December 2023; accepted 4 March 2024; published 16 April 2024)

Transformers are neural networks that revolutionized natural language processing and machine learning. They process sequences of inputs, like words, using a mechanism called self-attention, which is trained via masked language modeling (MLM). In MLM, a word is randomly masked in an input sequence, and the network is trained to predict the missing word. Despite the practical success of transformers, it remains unclear what type of data distribution self-attention can learn efficiently. Here, we show analytically that if one decouples the treatment of word positions and embeddings, a single layer of self-attention learns the conditionals of a generalized Potts model with interactions between sites and Potts colors. Moreover, we show that training this neural network is exactly equivalent to solving the inverse Potts problem by the so-called pseudolikelihood method, well known in statistical physics. Using this mapping, we compute the generalization error of self-attention in a model scenario analytically using the replica method.

DOI: 10.1103/PhysRevResearch.6.023057

I. INTRODUCTION

Transformers [1] are a powerful type of neural network that have achieved state-of-the art results in natural language processing (NLP) [2–6], image classification [7], and even protein structure prediction [8]. While standard neural networks can be thought of as functions of a single input, transformers act on sets of "tokens," like words in a sentence. The key to the success of transformers is a technique called masked language modeling (MLM), where transformers are trained to predict missing words in a sentence [2–6]; cf. Fig. 1(a). This technique has the advantage that it can leverage large amounts of raw text (or images, or protein sequences) without any annotation. By learning the conditional distribution of having a word in a specific position of the sentence, given the other words, transformers ostensibly learn the relationships between words in a robust way.

The basic building block of transformers is the *self-attention* (SA) mechanism [9,10], which transforms a sequence of tokens $\mathbf{x}_j$ into another sequence $\mathbf{h}_j$. We illustrate self-attention on a masked language modeling task in Fig. 1. The sentence is first transformed into a set of representations $\mathbf{x}_j = \mathbf{e}_j + \mathbf{p}_j$, where $\mathbf{e}_j$ is a vector representing the jth word, and the vector $\mathbf{p}_j$ encodes its position. SA then computes a linear transformation of the representations to yield the values $\mathbf{v}_j$. The kth output vector $\mathbf{h}_k$ is then a linear combination of the values $\mathbf{v}_j$ weighted by an attention matrix A, whose elements A_{kj} quantify the relative importance of the jth input token for the kth output vector, for example based on their semantic similarity. The functions to compute values $\mathbf{v}_j$ and the attention matrix A both have trainable parameters; see Eq. (2) for a precise definition. The flexibility of self-attention comes from the attention weights A_{kj}, which are not fixed, but computed given the context, i.e., the surrounding tokens.

The practical success of transformers raises several fundamental questions: what are the statistical structures that self-attention learns with MLM? More precisely, since the MLM objective is to learn the conditional probability distribution of words given a set of surrounding words, which family of conditional probabilities can self-attention learn? And how many samples are required to achieve good performance? Here, we take a step towards answering these questions by exploiting tools from the statistical physics of learning [11–14].

The first challenge is to design a data model that mimics the structure of real sentences. While classical works modeled inputs as vectors of i.i.d. random variables, recent work has introduced more sophisticated data models for neural networks [15–21], which allowed the study of unsupervised learning [22,23]. To analyze the self-supervised learning of MLM, we model sequences of words as system of spins, interacting via a generalized Potts Hamiltonian [24,25] with couplings between colors (=words) and positions. We sample a synthetic data set from the Potts model using Monte Carlo, and we perform masked language modeling by training a transformer to predict masked spins in spin sequences. While an off-the-shelf transformer requires several layers of self-attention to learn this simple probability distribution, we show analytically that a single layer of *factored* self-attention, where we separate the treatment of positions and inputs, can reconstruct the couplings of the Potts model *exactly* in the limit of a large training set. In particular, we derive an exact mapping between the output of the self-attention mechanism and the conditional distribution of a Potts spin given the others. We finally use this mapping to compute the generalization

*sgoldt@sissa.it

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.

loss of a single layer of self-attention analytically using the replica method.

II. A GENERALIZED POTTS MODEL TO SAMPLE SEQUENCES

We model sentences as sequences of spins $\mathbf{s} = (\mathbf{s}_1, \ldots, \mathbf{s}_L)$, with $\mathbf{s}_i \in \mathbb{R}^C$ taking values from a vocabulary of C colors, which we encode as one-hot vectors. Each color can be thought of as a word in natural text, an amino acid in a protein, etc. In a standard Potts Hamiltonian, only spins of the same color interact with each other via an interaction matrix J. This is an unrealistic model for real data: it treats all colors as orthogonal, even though words and amino acids have varying degrees of similarity. We therefore generalize the Potts Hamiltonian to

$$\mathcal{H}(\mathbf{s}) = -\frac{1}{2}\sum_{i,j=1}^{L} J_{ij}\mathbf{s}_i^T U \mathbf{s}_j, \tag{1}$$

where $J \in \mathbb{R}^{L\times L}$ governs the interactions between spins at different positions, and $U \in \mathbb{R}^{C\times C}$ encodes the similarities between colors (we denote matrices by capital letters and vectors in boldface). Without loss of generality, we set $J_{ii} = 0$ and sample sequences from the Boltzmann distribution $P(\mathbf{s}) \propto \exp[-\beta\mathcal{H}(\mathbf{s})]$. We recover the standard Potts model by choosing U as the identity matrix.

III. MASKED LANGUAGE MODELING WITH TRANSFORMERS

Given the generative model (1), the MLM objective amounts to predicting the ith spin given the sequence $\mathbf{s}_{\setminus i}$ where that spin is "masked," i.e., $\mathbf{s}_i = \mathbf{t}$, the masking token. To apply self-attention to a sequence $\mathbf{s}_{\setminus i}$, we first compute the values $\mathbf{v}_j = V(E\mathbf{s}_j + a\mathbf{p}_j)$, where the embedding matrix $E \in \mathbb{R}^{d\times C}$ maps Potts colors into d-dimensional representation vectors, and $V \in \mathbb{R}^{d\times d}$ is a weight matrix; both E and V are trainable parameters. The scalar parameter a controls the relative importance between the embedding and positional encoding vectors. The output vector $\mathbf{h}_i$ corresponding to the masked token is a linear combination of the values, weighted by an exponential attention function [1]:

$$\mathbf{h}_i(\mathbf{s}_{\setminus i}) = \sum_{j=1}^{L} \frac{\exp[(E\mathbf{t} + \mathbf{p}_i)^\top Q^\top K(aE\mathbf{s}_j + \mathbf{p}_j)]}{\sum_k \exp[(E\mathbf{t} + \mathbf{p}_i)^\top Q^\top K(aE\mathbf{s}_k + \mathbf{p}_k)]}\mathbf{v}_j. \tag{2}$$

Crucially, the ith spin $\mathbf{s}_i$ in this expression has to be replaced with the masking token $\mathbf{t}$, since it is the masked input. The matrices $Q, K \in \mathbb{R}^{d\times d}$ are also trainable parameters of the model. In the following, we take the embedding dimension equal to the number of colors, $d = C$, in order to be able to map the output vector $\mathbf{h}_i$ into a probability distribution $\tilde{\mathbf{p}}_i$ over the colors through the *softmax* nonlinearity [26].

IV. TRAINING A VANILLA TRANSFORMER ON THE GENERALIZED POTTS MODEL

For our first experiment, we emulate the setting of protein structure prediction, so we choose a vocabulary of size $C = 20$ and sample a symmetric interaction matrix $J_{ij} = \{0, 1\}$, which we show in Fig. 2(b). We draw the entries of the symmetric interaction matrix U i.i.d. from the standard Gaussian distribution. Given these parameters, we use Gibbs sampling to generate a data set with $M = 3000$ sequences of length $L = 20$. We tune the inverse temperature β to ensure an average Hamming distance of 0.3 between sampled sequences, typical for protein families [27].

We then train off-the-shelf transformers consisting of one and three layers on this data set by minimizing the cross-entropy loss between the output distribution and the missing spin using stochastic gradient descent (see the Appendix A for the numerical details) on the loss $\mathcal{L}(\mathbf{s}) = -L^{-1}\sum_{i=1}^{L}\sum_{\alpha=1}^{C} s_{i\alpha}\log\tilde{p}_{i\alpha}(\mathbf{s})$, for a sequence $\mathbf{s}$. In Fig. 2, we show the test loss

$$\epsilon_g = \mathbb{E}_{\mathbf{s}\sim P}[\mathcal{L}(\mathbf{s})] \tag{3}$$

during training, where $\mathbb{E}_{\mathbf{s}\sim P}$ denotes an average over the generative model (1). A transformer with a single layer of self-attention does not attain the optimal generalization error (black dashed line). By plotting the attention matrix of the single layer, we see that the transformer recovers the original interaction matrix to some degree, albeit not perfectly. Training transformers with three layers on the same data set improves the accuracy at the cost of losing interpretability: there is no straightforward way to collapse several layers of nonlinear transformations of the input sequence into a single attention map; we show the average of the final two attention layers in Fig. 2(b).

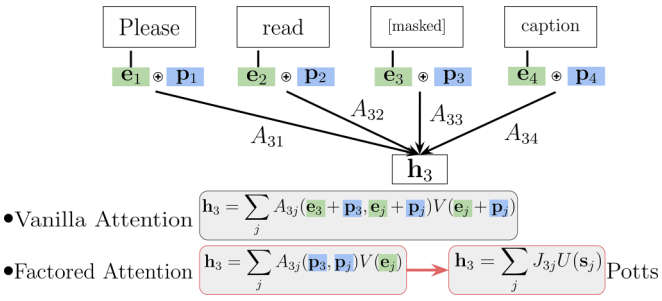

FIG. 1. Masked language modeling (MLM) with a single layer of self-attention. The goal of MLM is to predict the masked word in a given sentence. Self-attention first maps words into representations $\mathbf{e}_j + \mathbf{p}_j$, where $\mathbf{e}_j$ are embedding vectors representing words, and $\mathbf{p}_j$ encode their positions. For a given masked word, the associated attention vector $\mathbf{h}_k$ is computed as a linear combination of the values $\mathbf{v}_j = V(\mathbf{e}_j + \mathbf{p}_j)$ of all other tokens, weighted by the attention weights A_{kj}. In vanilla self-attention, values and attention weights depend on embeddings *and* positional vectors, while in *factored* attention, attention weights depend only on positions, and values only on the embeddings. By identifying the attention weights A with the interaction matrix J of a Potts model (1), the value matrix V with the color similarity matrix U, and the embedding vectors with the one-hot spins, we get a learning model identical to a Potts model.

V. FACTORED SELF-ATTENTION LEARNS THE GENERALIZED POTTS MODEL

We now consider a variant of self-attention in which the treatment of positions and values is decoupled. We set $a = 0$

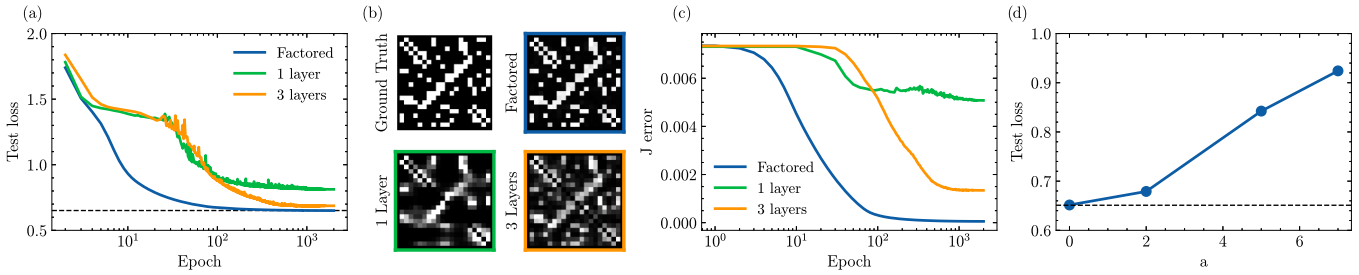

FIG. 2. A single layer of factored self-attention learns the generalized Potts model efficiently. (a) Test loss (3) for factored self-attention and for vanilla transformers with one and three layers during training with stochastic gradient descent. The optimal generalization loss is shown as a black dashed line. (b) Interaction matrix J of the generative Potts model (1) compared to the attention maps learned by transformers with vanilla and factored self-attention. For the three-layer transformer, the attention map was obtained by averaging the maps of the last two layers. (c) Reconstruction error of the interaction $(J - A)^2$ as a function of the number of epochs for all considered architectures. (d) Test loss as a function of perturbation level a. Decoupling the treatment between positions and colors by decreasing a decreases the test loss. *Parameters:* sequence length $L = 20$, vocabulary size $C = 20$, embedding dimension $d = 20$, $M = 3000$ data points.

in Eq. (2), set the masking token $\mathbf{t} = \mathbf{0}$, choose one-hot encodings for the positions, and fix the embedding matrix at $E = I_C$, so that

$$h_{i\alpha}(\mathbf{s}_{\setminus i}) = \sum_{j=1}^{L} A_{ij}(V\mathbf{s}_j)_\alpha, \tag{4}$$

where $A_{ij} \equiv e^{(Q^\top K)_{ij}} / \sum_{k=1}^{L} e^{(Q^\top K)_{ik}}$. This modified self-attention has exactly the same form as the conditional distribution of the generalized Potts model if one sets $U = V$ and $\beta J = A$, which is

$$p(s_{i\alpha} = 1|\mathbf{s}_{\setminus i}) = \frac{\exp\left(\beta \sum_{j=1}^{L} J_{ij}(U\mathbf{s}_j)_\alpha\right)}{\sum_{\gamma=1}^{C} \exp\left(\beta \sum_{j=1}^{L} J_{ij}(U\mathbf{s}_j)_\gamma\right)}. \tag{5}$$

This equivalence between factored self-attention and the Potts model is our first main result; we now discuss its ramifications.

Decoupling positions and colors leads to a significant improvement in the performance of a single layer, which reaches the optimal generalization error and converges faster; cf. Fig. 2(a). Factored self-attention recovers the interaction matrix J perfectly, cf. Fig. 2(b) for the attention map and Fig. 2(c) for the reconstruction error of the interaction matrix. In Fig. 2(d), we show that decoupling the treatment of positions and colors completely performs better than any intermediate solution with $a > 0$.

Factored-attention layers, and thus input-independent attention weights, have been already used as a building block for deep transformers, outperforming standard attention in different applications [28] used it to analyze protein sequences and found that a single layer of factored self-attention performed as well as a deep transformer, and significantly better than a single layer of vanilla self-attention, without explicitly explaining this observation. Moreover, using factored attention is key to obtaining state-of-the-art results in approximating ground states of many-body quantum systems [29–31].

Intriguingly, the form of the loss for masked language modeling with factored self-attention as described above exactly matches the loss of the pseudolikelihood method, which has been used for solving the inverse Ising problem [32–36]. The pseudolikelihood method is statistically consistent [37–39], i.e., its parameter estimates converge to the true parameters as the number of samples goes to infinity. A direct consequence of the mapping in Eq. (4) is thus that MLM with factored self-attention enjoys the same asymptotic optimality.

VI. THE SAMPLE COMPLEXITY OF SELF-ATTENTION

A key quantity in machine learning problems is the sample complexity, namely how many samples are required to achieve a small generalization loss ϵ_g with a given model. The mapping introduced in this work allows us to address this question precisely for a single layer of self-attention by means of the replica method from statistical physics. The main difficulty in the calculation lies in handling the nontrivial data distribution (1). This difficulty can be mitigated thanks to recent advances in statistical physics, which allow us to extend the replica method of disordered systems to structured data [18,20,40]. To perform the replica calculation, we first relax the discrete nature of Potts spins by rewriting the generalized Potts Hamiltonian [Eq. (1)] in terms of spin magnetization $\mathbf{m} = \langle \mathbf{s} \rangle_{P(\mathbf{s})}$ following mean-field theory. The associated Boltzmann measure then turns into a multivariate Gaussian distribution, whose covariance matrix is the negative inverse of the interaction matrix, i.e., $\Sigma = -J^{-1}$ [41,42]. We then draw sequences $\{\mathbf{m}^\mu\}_{\mu=1}^{M}$ of length L from the multivariate Gaussian with zero mean and covariance matrix $\Sigma = (\Omega L^{-1/2} + \nu \mathbb{I})^{-1}$, where Ω is a symmetric full-rank random matrix sampled from the Gaussian orthogonal ensemble, while $\nu\mathbb{I}$ is a diagonal matrix centering the spectrum of Ω in ν. To ensure Σ is positive-definite, we set $\nu > 2$ due to the semicircle law [43]. By fixing the location i of the masked spin across all input sequences, solving the MLM task is equivalent to inferring the ith row of the interaction matrix $\mathbf{J}_i$.

To accomplish this task, we train a single layer of factored self-attention by empirical risk minimization of a square loss

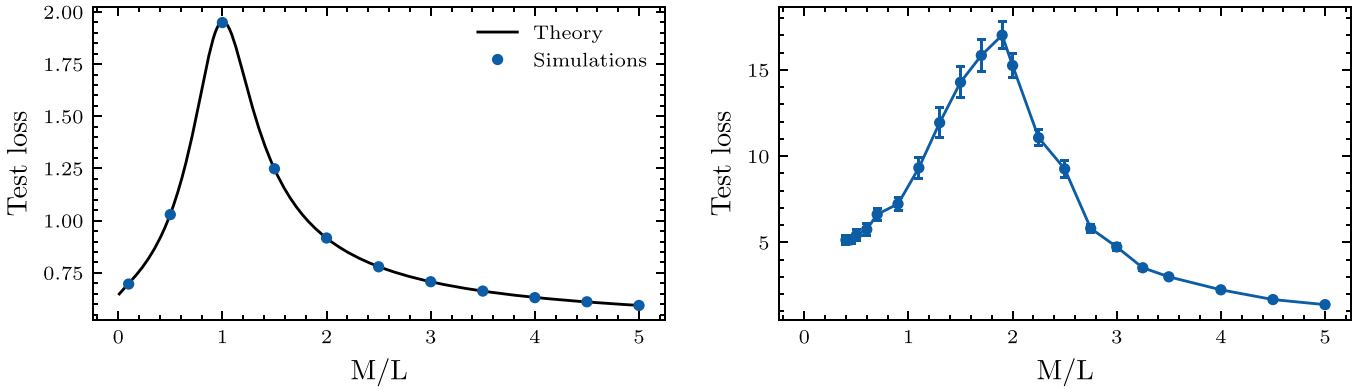

FIG. 3. The interpolation peak of factored attention in theory and practice. Left: A replica analysis predicts the test loss exactly. Test loss of a single layer of factored self-attention as a function of the number of samples per input dimension, as computed using replica theory (solid line). The blue points represent the outcome of numerical minimization of the square loss (6), averaged over 30 realizations, and show perfect agreement with the theory. Error bars are smaller than point size. Right: Same plot for a single layer of factored self-attention in the setting of Fig. 2 ($L = C = 20$), showing the same qualitative behavior. The simulations are averaged over $n = 30$ different realizations.

with ℓ_2-regularization:

$$\hat{\mathbf{A}}_i = \underset{\mathbf{A}_i}{\operatorname{argmin}} \left[\frac{1}{2} \sum_{\mu=1}^{M} \left(m_i^\mu - \mathbf{A}_i \cdot \mathbf{m}_{\setminus i}^\mu \right)^2 + \frac{\lambda}{2} ||\mathbf{A}_i||_2^2 \right]. \quad (6)$$

Our goal is to characterize the generalization loss ϵ_g (3) of factored attention with the parameters obtained from minimizing the loss (6). In the high-dimensional limit, where the number of samples and the sequence length M, L tend to infinity while their ratio $\alpha \equiv M/L \sim O(1)$, we can express ϵ_g using replica theory as a function of four scalar quantities:

$$\epsilon_g = \rho + q^\star - 2r^\star + 1/\nu. \quad (7)$$

Here, ν is the center of the spectrum of Ω, $\rho = \operatorname{tr} \Sigma_{\setminus i}/(\nu^2 L)$ is a function of the covariance matrix $\Sigma_{\setminus i}$, where we have removed the ith row and column, while $q^\star$ and $r^\star$ are the so-called overlap parameters. They correspond to practically measurable quantities over different realization of the training set, involving the estimator of the ith row of the interaction matrix:

$$q^\star = \frac{\hat{\mathbf{A}}_i^t \Sigma_{\setminus i} \hat{\mathbf{A}}_i}{L}, \quad r^\star = -\frac{(\mathbf{A}_i^\star)^t \Sigma_{\setminus i} \hat{\mathbf{A}}_i}{\nu L}. \quad (8)$$

The parameter r^* can thus be interpreted as an overlap between the estimated attention $\hat{\mathbf{A}}_i$ and the ground-truth value $\mathbf{A}_i$ while q^* is the overlap of the estimated attention, both mediated by the modified covariance matrix $\Sigma_{\setminus i}$.

As we show in the Appendix B, the values of these order parameters for a given training set of size α can be obtained by solving the optimization problem,

$$f_\beta = \underset{q,r,\delta q,\hat{q},\hat{r},\delta\hat{q}}{\operatorname{extr}} \left[-\frac{1}{2}(\hat{q}\delta q - q\delta\hat{q}) + \nu r\hat{r} + \lim_{L\to\infty} \frac{1}{L} \Psi_s + \alpha \Psi_e \right], \quad (9)$$

which yields the *typical* value (over the data) of the free energy density associated with a Gibbs measure at inverse temperature β whose Hamiltonian corresponds to the loss function in (6). Note that the optimization only involves the scalar parameters q, r, δq and their conjugates $\hat{q}$, $\hat{r}$, and $\delta\hat{q}$, with $\delta q = q - q_s$, $\delta\hat{q} = \hat{q} + \hat{q}_s$, and q_s being the self-overlap among replicas. The so-called entropic potential Ψ_s is a function of the input covariance $\Sigma_{\setminus i}$:

$$\Psi_s = \lim_{L\to\infty} \frac{\hat{r}^2}{2L} \operatorname{tr}((\Sigma_{\setminus i})^t \Sigma_{\setminus i} (\nu\mathbb{I} + \delta\hat{q}\Sigma_{\setminus i})^{-1}) + \lim_{L\to\infty} \frac{\hat{q}}{2L} \operatorname{tr}(\Sigma_{\setminus i} (\nu\mathbb{I} + \delta\hat{q}\Sigma_{\setminus i})^{-1}), \quad (10)$$

while the energetic potential Ψ_e only depends on the specific choice of the loss function. As shown in the Appendix B, for the optimization problem in Eq. (6), Ψ_e is given by

$$\Psi_e = -\frac{1}{2(1+\delta q)} \left(\frac{1}{\nu} + \rho + q - 2r \right). \quad (11)$$

Using this approach, we estimated analytically the generalization loss in Eq. (7) as a function of the rescaled number of samples $\alpha = M/L$ (see the Appendixes for details). The result is shown in the right panel of Fig. 3. As can be noticed, the test loss *increases* in the small data regime, before peaking at $\alpha = 1$. This value corresponds to the interpolation threshold, which is the largest number of samples that the neural network can perfectly fit, which in fact happens at $M = L$. Below this threshold, the model overfits to its training data; beyond this threshold, the generalization error decreases monotonically with the training set size; for large α, we found $\epsilon_g \sim \alpha^{-1/2}$. A similar peak in the generalization loss has been observed in supervised learning [44] and it is connected to the well-known "double descent" curve observed in deep neural networks in the presence of label noise [45]. There, the peak is a consequence of overfitting induced by the noise in the labels, and it appears after an initial decay of the test loss at small α. In the self-supervised learning regime explored in this work, we find instead that the peak appears naturally as a consequence of the intrinsic stochasticity of the inputs. Indeed, in masked language modeling, the labels are a part of the input itself. The noise affecting the labels is thus highly correlated to that affecting the input. If the noise in the input is too high, the model starts immediately overfitting and the initial descent is not observed. The absence of the initial descent can therefore be ascribed to the high level of the noise in the input.

We verify the predictions of the replica theory by plotting the generalization loss of a single layer of factored self-attention trained on the generalized Potts model in the setting of Fig. 2 at small regularisation (right side of Fig. 3). We see the same qualitative behavior as predicted by replica theory, even though in this case we did not apply any of the assumptions required for the replica analysis (the mean-field limit, and the usage of a full-rank J matrix and of a U matrix fixed to the identity). In particular, the test loss increases when adding more data for small α. The only difference between the plots is the location of the peak. For the square loss that we analyzed with replicas, as we already commented, the peak is at the interpolation threshold $M = L$. For the simulations with logistic loss, the peak appears at the linearly separability threshold, which is the largest number of points a linear classifier can classify correctly, and which can be larger than 1 [11,18].

VII. CONCLUDING PERSPECTIVES

In this work, we have characterized the probability distributions that a single layer of self-attention can learn when trained on a masked language modeling task, considered as a simple prototype of self-supervised learning. In particular, we have shown analytically and numerically that with a single factored-attention layer, it is possible to exactly reconstruct the couplings of a generalized Potts model with two-body interactions between both sites and colors. More precisely, we showed that training factored self-attention on the MLM objective is equivalent to solving the inverse Potts problem using the pseudolikelihood method [32–36], and therefore it yields consistent estimators of the parameters. These findings make factored attention a powerful, theoretically driven building block for deep transformers. Our replica analysis of self-attention enabled us to compute the generalization loss of the model exactly and yielded a nontrivial generalization behavior.

Learning higher-order interactions will require additional layers: a detailed study of how this can be achieved is an interesting direction for future research. It will be interesting also to study the learning *dynamics* of self-attention using methods from statistical physics [46–49], both on MLM and on supervised tasks [50–53]. In short, our work clarifies the limits of standard self-attention trained on data where two-body interactions dominate, and it highlights the potential of factored attention as a component of transformer models.

ACKNOWLEDGMENT

We thank Marc Mézard, Manfred Opper, and Lenka Zdeborová for stimulating discussions, and Santiago Acevedo for critically reading the manuscript.

APPENDIX A: NUMERICAL DETAILS

The numerical simulations were performed using JAX [54]. Both the *factored* attention layer and the vanilla transformer architecture were optimized using SGD with a minibatch size of 100 and a cosine annealer as the learning rate decay scheduler, both standard choices in the literature. The initial learning rate was adjusted to the specific simulations, choosing between 0.1 and 0.01. The vanilla transformer code has been taken from Ref. [55], with no modifications made. In particular, as already pointed out in the main text, each element $\mathbf{s}_i$ of a sequence $\mathbf{s} = (\mathbf{s}_1, \ldots, \mathbf{s}_L)$ is first transformed into a token $\mathbf{x}_i = \mathbf{e}_i + \mathbf{p}_i$, with $\mathbf{e}_i$ being the embedding of $\mathbf{s}_i$, and $\mathbf{p}_i$ being the positional encoding. The tokenized sequences are then fed to a layer made of two distinct sublayers. The first sublayer is composed of a single-head attention, while the second sublayer contains a two-layer fully connected neural network. The inputs of both sublayers are connected to their outputs through skip-connections, and layer normalization is then applied.

Finally, there is an output layer consisting of a linear transformation from d- to C-dimensional space in order to obtain a probability distribution over the colors through the *softmax* nonlinearity. For a graphic visualization of the transformer encoder architecture, reference can be made to the original paper of Vaswani *et al.* [1]. The list of transformer hyperparameters used for the simulations of Fig. 2 is as follows: embedding dimension, 20; number of heads, 1; number of layers, 1–3; dropout probability, 0.0; number of classes, 20.

The data set was generated using Gibbs sampling, starting from a random sequence of $L = 20$ sites and $C = 20$ Potts colors and cyclically sampling the spins by exploiting the knowledge of the exact conditional probabilities, Eq. (5). To decorrelate the samples, 10 000 Gibbs sweeps were made between each of the two saved configurations.

The simulations in the left panel of Fig. 3 have been performed by sampling the input data points from a multivariate Gaussian distribution and the masked token from the same distribution, conditioned on the other elements in the sequence. The optimization problem in Eq. (6) is then solved in closed form thanks to the Moore-Penrose inverse, as in Ref. [18].

APPENDIX B: REPLICA ANALYSIS OF SELF-ATTENTION

In this Appendix, we discuss in detail the replica analysis of factored single-layer self-attention. The computation builds upon the recent advances in the statistical physics of learning with regard to the extension of replica theory to structured data [18,20,40]. In the following, we will show how to slightly modify this new approach in order to deal with masked language modeling tasks, under the following simplified assumptions: mean-field limit, full-rank J matrix, and U matrix fixed to the identity and thus not learned.

1. Statistical physics formulation of machine learning problems

In statistical physics, learning is considered as a dynamical and exploratory process across the space of the learnable parameters. At equilibrium, these parameters are assumed to follow a Boltzmann-Gibbs distribution, where the role of the Hamiltonian is actually played by the loss function:

$$\pi_\beta(\mathbf{A}_i, \mathcal{D}) = \frac{P(\mathbf{A}_i)}{\mathcal{Z}_\beta} e^{-\beta \sum_{\mu=1}^{M} \ell\left(m_i^\mu, \frac{\mathbf{A}_i \cdot \mathbf{m}_{\setminus i}^\mu}{\sqrt{L}}\right)} = \frac{P_A(\mathbf{A}_i)}{\mathcal{Z}_\beta} \prod_{\mu=1}^{M} P_G\left(m_i^\mu | \frac{\mathbf{A}_i \cdot \mathbf{m}_{\setminus i}^\mu}{\sqrt{L}}\right), \tag{B1}$$

with β being the inverse temperature and $\mathcal{D}$ the training set. In the zero-temperature limit (i.e., $\beta \to \infty$), the Boltzmann-Gibbs distribution concentrates around the minima of the loss function, which are merely the solutions of the optimization problem in Eq. (6):

$$\hat{\mathbf{A}}_i \underset{\beta\to\infty}{=} \mathbb{E}_{\pi_\beta}[\mathbf{A}_i]. \tag{B2}$$

Up to this point, reframing a machine learning problem in terms of statistical physics did not seem to be very advantageous since sampling from a high-dimensional Boltzmann-Gibbs distribution is known to be impracticable.

This is where the replica theory comes into play. In particular, it states that, in the high-dimensional limit [i.e., $M, L \to \infty$ with $\alpha \equiv M/L \sim O(1)$], the free energy of a learning system concentrates around its typical value over the input data distribution:

$$f_\beta = -\lim_{L\to\infty} \frac{1}{L} \mathbb{E}_{\{\mathbf{m}^\mu_{\setminus i}, m^\mu_i\}} [\log \mathcal{Z}_\beta]. \tag{B3}$$

As we will see in the next section, this expectation can be tackled by means of the replica trick. From this quantity, all the high-dimensional metrics of interest can be computed as a function of simple scalar quantities. This is the case, for instance, of the generalization loss in Eq. (7). In particular, the overlap parameters $m_\star$ and $q_\star$ correspond to practically measurable quantities over different realization of the training set, involving the estimator of the ith row of the interaction matrix:

$$q^\star = \frac{\hat{\mathbf{A}}_i^t \Sigma_{\setminus i} \hat{\mathbf{A}}_i}{L}, \quad r^\star = -\frac{\mathbf{J}_i^t \Sigma_{\setminus i} \hat{\mathbf{A}}_i}{\nu L}. \tag{B4}$$

In the next section, we will outline the main steps of the replica trick leading to the generalization loss formula in Eq. (7).

2. Replica calculation

As anticipated in the previous section, the replica trick allows us to compute the typical value of the free-energy density in Eq. (B3) by expressing this quantity as a function of the solely replicated partition function $\mathcal{Z}^n_\beta$, obtained by constructing $n > 0$ different and independent copies of the same learning system:

$$f_\beta = -\lim_{n\to 0^+} \frac{d}{dn} \lim_{L\to\infty} \left[\frac{1}{L} \mathbb{E}_{\{\mathbf{m}^\mu_{\setminus i}, m^\mu_i\}} \mathcal{Z}^n_\beta \right]. \tag{B5}$$

a. Average over the training set

As a first step, the replica calculation focuses on the expectation of the replicated partition function over the training set, which, written in a more explicit form, looks like

$$\mathbb{E}_{\{\mathbf{m}^\mu_{\setminus i}, m^\mu_i\}} \mathcal{Z}^n_\beta = \int \prod_{a=1}^n d\mathbf{A}^a_i \prod_{a=1}^n P_A(\mathbf{A}^a_i) \times \prod_{\mu=1}^M \mathbb{E}_{m^\mu_i|\mathbf{m}^\mu_{\setminus i}} \mathbb{E}_{\mathbf{m}^\mu_{\setminus i}} \left[P_G\left(m^\mu_i \mid \frac{\mathbf{A}^a_i \cdot \mathbf{m}^\mu_{\setminus i}}{\sqrt{L}} \right) \right], \tag{B6}$$

with P_G and P_A being, respectively, the Gibbs and the Gaussian measure associated with the ith row of the attention matrix as in Eq. (B1). Indeed, as already pointed out in the main manuscript, the interaction matrix is drawn from the Gaussian orthogonal ensemble, therefore its rows will correspond to Gaussian random vectors. At this point, we can notice an important aspect of MLM tasks. In this case, the labels are not provided by a teacher vector as in standard teacher-student settings. On the contrary, the masked tokens are directly sampled from the input distribution by conditioning over all the other elements composing the sequence:

$$m^\mu_i \sim \frac{P_J(\mathbf{J}_i)}{\sqrt{2\pi\nu^{-1}}} e^{-\frac{1}{2\nu^{-1}}\left(m^\mu_i + \frac{\mathbf{J}_i\cdot\mathbf{m}^\mu_{\setminus i}}{\nu\sqrt{L}}\right)^2} = P_J(\mathbf{J}_i) P_0\left(m^\mu_i \mid \frac{\mathbf{J}_i \cdot \mathbf{m}^\mu_{\setminus i}}{\nu\sqrt{L}} \right). \tag{B7}$$

Note that the noise in the labels arises as a consequence of the one already affecting the input data points, meaning that its intensity cannot be chosen independently from the intrinsic stochasticity of the input. Due to these considerations, by explicitly expressing the outer expectation in Eq. (B6), we then obtain

$$\mathbb{E}_{\{\mathbf{m}^\mu_{\setminus i}, m^\mu_i\}} \mathcal{Z}^n_\beta = \int d\mathbf{J}_i\, P_J(\mathbf{J}_i) \prod_{a=1}^n \left(\int d\mathbf{A}^a_i\, P_A(\mathbf{A}^a_i) \right) \times \prod_{\mu=1}^M \left\{ \int dm^\mu_i \mathbb{E}_{\mathbf{m}^\mu_{\setminus i}} \left[P_0\left(m^\mu_i \mid \frac{\mathbf{J}_i\cdot\mathbf{m}^\mu_{\setminus i}}{\nu\sqrt{L}} \right) \times P_G\left(m^\mu_i \mid \frac{\mathbf{A}^a_i \cdot \mathbf{m}^\mu_{\setminus i}}{\sqrt{L}} \right) \right] \right\}. \tag{B8}$$

To compute the expectation over the input sequences with a masked element at position i, we first define the preactivations as

$$h^\mu_a = \frac{\mathbf{A}^a_i \cdot \mathbf{m}^\mu_{\setminus i}}{\sqrt{L}}, \quad z^\mu = -\frac{\mathbf{J}_i \cdot \mathbf{m}^\mu_{\setminus i}}{\nu\sqrt{L}}, \tag{B9}$$

and then we express these definitions in terms of Dirac deltas and their corresponding integral representation:

$$1 \propto \int \prod_{\mu=1}^M \prod_{a=1}^n \frac{dh^\mu_a d\hat{h}^\mu_a}{2\pi} \prod_{\mu=1}^M e^{i\hat{h}^\mu_a \left(h^\mu_a - \frac{\mathbf{A}^a_i\cdot\mathbf{m}^\mu_{\setminus i}}{\sqrt{L}}\right)},$$
$$1 \propto \int \prod_{\mu=1}^M \frac{dz^\mu d\hat{z}^\mu}{2\pi} \prod_{\mu=1}^M e^{i\hat{z}^\mu \left(z^\mu + \frac{\mathbf{J}_i\cdot\mathbf{m}^\mu_{\setminus i}}{\nu\sqrt{L}}\right)}. \tag{B10}$$

By plugging these factors into Eq. (B8), we then obtain

$$\begin{aligned} &\mathbb{E}_{\{\mathbf{m}^\mu_{\setminus i}, m^\mu_i\}} \mathcal{Z}^n_\beta \\ &= \int d\mathbf{J}_i\, P_J(\mathbf{J}_i) \int \prod_{a=1}^n d\mathbf{A}^a_i \prod_{a=1}^n P_A(\mathbf{A}^a_i) \\ &\times \int \prod_{\mu=1}^M dm^\mu_i \int \prod_{\mu=1}^M \frac{dz^\mu d\hat{z}^\mu}{2\pi} \prod_{\mu=1}^M e^{i\hat{z}^\mu z^\mu} \prod_{\mu=1}^M P_0(m^\mu_i \mid z^\mu) \\ &\times \int \prod_{\mu=1}^M \prod_{a=1}^n \frac{dh^\mu_a d\hat{h}^\mu_a}{2\pi} \prod_{\mu=1}^M \prod_{a=1}^n e^{i\hat{h}^\mu_a h^\mu_a} \prod_{\mu=1}^M \prod_{a=1}^n P_G(m^\mu_i \mid h^\mu_a) \\ &\times \prod_{\mu=1}^n \mathbb{E}_{\mathbf{m}^\mu_{\setminus i}} \left[e^{i\hat{z}^\mu \frac{\mathbf{J}_i\cdot\mathbf{m}^\mu_{\setminus i}}{\nu\sqrt{L}}} \prod_{a=1}^n e^{-i\hat{h}^\mu_a \frac{\mathbf{A}^a_i\cdot\mathbf{m}^\mu_{\setminus i}}{\sqrt{L}}} \right]. \end{aligned} \tag{B11}$$

The expectation over the masked input sequences is a simple multivariate Gaussian integral, whose solution is given by

$$\mathbb{E}_{\mathbf{m}^\mu_{\setminus i}}\left[e^{i\hat{z}^\mu \frac{\mathbf{J}_i\cdot\mathbf{m}^\mu_{\setminus i}}{\nu\sqrt{L}}}\prod_{a=1}^{n} e^{-i\hat{h}^\mu_a \frac{\mathbf{A}^a_i\cdot\mathbf{m}^\mu_{\setminus i}}{\sqrt{L}}}\right] = e^{-\frac{1}{2}\frac{\mathbf{J}_i^t\Sigma_{\setminus i}\mathbf{J}_i}{\nu^2 L}(z^\mu)^2} e^{\sum_{a=1}^n \frac{\mathbf{J}_i^t\Sigma_{\setminus i}\mathbf{A}^a_i}{\nu L}\hat{h}^\mu_a\hat{z}^\mu - \frac{1}{2}\sum_{a,b=1}^n \frac{(\mathbf{A}^a_i)^t\Sigma_{\setminus i}\mathbf{A}^a_i}{L}\hat{h}^\mu_a\hat{h}^\mu_b}, \tag{B12}$$

where, as already pointed out in the main text, $\Sigma_{\setminus i}$ corresponds to the covariance matrix of the masked input sequences, that is, the input covariance matrix without the contribution of the row and column associated with the ith masked token. By replacing the solution of the expectation in Eq. (B11), we then obtain

$$\begin{aligned}
&\mathbb{E}_{\{\mathbf{m}^\mu_{\setminus i}, m^\mu_i\}}\mathcal{Z}^n_\beta \\
&= \int d\mathbf{J}_i\, P_J(\mathbf{J}_i)\int\prod_{a=1}^n d\mathbf{A}^a_i \prod_{a=1}^n P_A(\mathbf{A}^a_i) \\
&\times \int\prod_{\mu=1}^M dm^\mu_i \int\prod_{\mu=1}^M \frac{dz^\mu d\hat{z}^\mu}{2\pi}\prod_{\mu=1}^M e^{i\hat{z}^\mu z^\mu}\prod_{\mu=1}^M P_0(m^\mu_i \mid z^\mu) \\
&\times \int\prod_{\mu=1}^M\prod_{a=1}^n \frac{dh^\mu_a d\hat{h}^\mu_a}{2\pi}\prod_{\mu=1}^M\prod_{a=1}^n e^{i\hat{h}^\mu_a h^\mu_a}\prod_{\mu=1}^M\prod_{a=1}^n P_G(m^\mu_i \mid h^\mu_a) \\
&\times \prod_{\mu=1}^n e^{-\frac{1}{2}\frac{\mathbf{J}_i^t\Sigma_{\setminus i}\mathbf{J}_i}{\nu^2 L}(z^\mu)^2 + \sum_{a=1}^n \frac{\mathbf{J}_i^t\Sigma_{\setminus i}\mathbf{A}^a_i}{\nu L}\hat{h}^\mu_a\hat{z}^\mu - \frac{1}{2}\sum_{a,b=1}^n \frac{(\mathbf{A}^a_i)^t\Sigma_{\setminus i}\mathbf{A}^a_i}{L}\hat{h}^\mu_a\hat{h}^\mu_b}.
\end{aligned} \tag{B13}$$

b. Rewriting the averaged replicated partition function in terms of saddle-point integrals

As a consequence of the average over the training set, the different replicas are now interacting among each other through the following set of overlap parameters:

$$\rho = \frac{\mathbf{J}_i^t\Sigma_{\setminus i}\mathbf{J}_i}{\nu^2 L}, \quad r_a = -\frac{(\mathbf{A}^a_i)^t\Sigma_{\setminus i}\mathbf{J}_i}{\nu L}, \quad q_{ab} = -\frac{(\mathbf{A}^a_i)^t\Sigma_{\setminus i}\mathbf{A}^b_i}{L}. \tag{B14}$$

Once again, to proceed further in the calculation, we can insert their definition by means of Dirac δ's and their integral representation:

$$\begin{aligned}
1 &\propto \int \frac{d\rho d\hat{\rho}}{2\pi}\, e^{i\hat{\rho}(\nu^2 L\rho - \mathbf{J}_i^t\Sigma_{\setminus i}\mathbf{J}_i)}, \\
1 &\propto \int\prod_{a=1}^n \frac{dr_a d\hat{r}_a}{2\pi}\prod_{a=1}^n e^{i\hat{r}_a(-\nu L r_a - (\mathbf{A}^a_i)^t\Sigma_{\setminus i}\mathbf{J}_i)}, \\
1 &\propto \int\prod_{a\leqslant b}\frac{dq_{ab}d\hat{q}_{ab}}{2\pi}\prod_{a\leqslant b} e^{i\hat{q}_{ab}(q_{ab} - (\mathbf{A}^a_i)^t\Sigma_{\setminus i}\mathbf{A}^b_i)}.
\end{aligned} \tag{B15}$$

By substituting the overlap definition in Eq. (B13), plugging in the corresponding factors, and performing the change of variables—$i\hat{\rho}\to-\hat{\rho}$, $i\hat{r}_a\to\hat{r}_a$, and $i\hat{q}_{ab}\to\hat{q}_{ab}$—we can rewrite the averaged replicated partition function in terms of saddle-point integrals over the overlap parameters:

$$\begin{aligned}
\mathbb{E}_{\{\mathbf{m}^\mu_{\setminus i}, m^\mu_i\}}\mathcal{Z}^n_\beta &= \int\frac{d\rho d\hat{\rho}}{2\pi}\int\prod_{a=1}^n\frac{dr_a d\hat{r}_a}{2\pi} \\
&\times\int\prod_{a\leqslant b}\frac{dq_{ab}d\hat{q}_{ab}}{2\pi}\, e^{L\Psi^{(n)}},
\end{aligned} \tag{B16}$$

where the action $\Psi^{(n)}$ is a nontrivial function of the overlap parameters:

$$\Psi^{(n)} = -\nu^2\rho\hat{\rho} + \nu\sum_{a=1}^n r_a\hat{r}_a - \sum_{a\leqslant b} q_{ab}\hat{q}_{ab} + \frac{1}{L}\Psi_s + \frac{M}{L}\Psi_e, \tag{B17}$$

where Ψ_s and Ψ_e are the so-called entropic and energetic potential, and, in the specific case of a single-layer factored attention, they are given by

$$\begin{aligned}
\Psi_s &= \log\Bigg[\int d\mathbf{J}_i P_J(\mathbf{J}_i)\int\prod_{a=1}^n\left[d\mathbf{A}^a_i P_A(\mathbf{A}^a_i)\right] \\
&\quad\times e^{\hat{\rho}\mathbf{J}_i^t\Sigma_{\setminus i}\mathbf{J}_i + \sum_{a=1}^n \hat{r}_a(\mathbf{A}^a_i)^t\Sigma_{\setminus i}\mathbf{J}_i + \sum_{a\leqslant b}\hat{q}_{ab}(\mathbf{A}^a_i)^t\Sigma_{\setminus i}\mathbf{A}^b_i}\Bigg], \\
\Psi_e &= \log\Bigg[\int dm_i\int\frac{dzd\hat{z}}{2\pi}\, e^{-\frac{\rho}{2}\hat{z}^2 + i\hat{z}z}P_0(m_i|z) \\
&\quad\times\int\prod_{a=1}^n\frac{dh_a d\hat{h}_a}{2\pi}\, e^{-\frac{1}{2}\sum_{a,b=1}^n q_{ab}\hat{h}_a\hat{h}_b - \hat{z}\sum_{a=1}^n r_a\hat{h}_a} \\
&\quad\times e^{i\sum_{a=1}^n\hat{h}_a h_a}P_G(m_i|h_a)\Bigg].
\end{aligned} \tag{B18}$$

Note that we have dropped the dependency of Ψ_e on the μ index since all the μ-dependent terms decouple with respect to μ.

c. Replica symmetric assumption

To proceed further in the calculation, we need to assume a specific replica structure. Since all replicas have been introduced independently from each other with no specific differences among them, it seems natural to assume that replicas should all play the same role and that, therefore, the overlap parameters should not depend on the specific replica index. In particular, under the replica symmetric ansatz, we assume

$$\begin{aligned}
q_{ab} &= \begin{cases} g & \text{if } a=b, \\ q & \text{otherwise,}\end{cases} \qquad -i\hat{q}_{ab} = \begin{cases} -\frac{\hat{g}}{2} & \text{if } a=b, \\ \hat{q} & \text{otherwise,}\end{cases} \\
r_a &= r, \quad -i\hat{r}_a = \hat{r} \quad \forall a.
\end{aligned} \tag{B19}$$

By plugging the replica symmetric assumption in Eqs. (B17) and (B18) and applying the following Hubbard Stratonovich transformations:

$$\begin{aligned}
e^{\frac{\hat{q}}{2}\sum_{a\leqslant b}(\mathbf{A}^a)^t\Sigma_{\setminus i}\mathbf{A}^b} &= \int\mathcal{D}\boldsymbol{\xi}\exp\left(\sum_{a=1}^n(\mathbf{A}^a_i)^t(\hat{q}\Sigma_{\setminus i})^{1/2}\boldsymbol{\xi}\right), \\
e^{-\frac{q}{2}\sum_{a,b=1}^n\hat{h}_a\hat{h}_b} &= \int\mathcal{D}\xi\exp\left(i\sqrt{q}\sum_{a=1}^n\hat{h}^2_a\,\xi\right),
\end{aligned} \tag{B20}$$

with $\xi \sim \mathcal{N}(0,1)$, we then get the expression for the replica symmetric action:

$$\Psi^{(n)} = -\nu^2\rho\hat{\rho} + \nu r\hat{r} + \frac{n}{2}(\delta q + q)(\delta\hat{q} - \hat{q}) - \frac{n(n-1)}{2}q\hat{q} + \frac{1}{L}\Psi_s + \frac{M}{L}\Psi_e, \tag{B21}$$

where we have defined $\delta\hat{q} = \hat{g} + \hat{q}$ and $\delta q = g - q$. The replica symmetric potentials Ψ_s and Ψ_e are are given by

$$\Psi_s = \log\left[\int d\mathbf{J}_i P_J(\mathbf{J}_i) e^{\hat{\rho}\mathbf{J}_i^t\Sigma_{\setminus i}\mathbf{J}_i}\int \mathcal{D}\boldsymbol{\xi} \times\left(\int d\mathbf{A}_i P_A(\mathbf{A}_i) e^{\hat{r}\mathbf{A}_i^t\Sigma_{\setminus i}\mathbf{J}_i - \frac{\hat{l}}{2}\mathbf{A}_i^t\Sigma_{\setminus i}\mathbf{A}_i + \mathbf{A}_i^t(\hat{q}\Sigma_{\setminus i})^{1/2}\boldsymbol{\xi}}\right)^n\right],$$

$$\Psi_e = \log\left[\int dm_i \int \frac{dz d\hat{z}}{2\pi} e^{-\frac{\rho}{2}\hat{z}^2 + i\hat{z}z} P_0(m_i|z)\int \mathcal{D}\xi \times\left(\int \frac{dh d\hat{h}}{2\pi} e^{-\frac{l}{2}\hat{h}^2 + i\sqrt{q}\xi\hat{h} - m\hat{z}\hat{h} + i\hat{h}h} P_G(m_i|h)\right)^n\right]. \tag{B22}$$

d. Zero replica limit

By taking the limit of $n \to 0$ in Eqs. (B21) and (B22) and solving the integrals with respect to the $\hat{z}$ and $\hat{h}$ variables, we then get the following expression for the action potential:

$$\Psi^{(n\to 0)} = \nu r\hat{r} + \frac{1}{2}(l+q)(\hat{l} - \hat{q}) + \frac{1}{2}q\hat{q} + \frac{1}{L}\Psi_s^{(n\to 0)} + \frac{M}{L}\Psi_e^{(n\to 0)} \tag{B23}$$

with the entropic and energetic potentials in the zero replicas limit given by

$$\Psi_s^{(n\to 0)} = \int \mathcal{D}\boldsymbol{\xi}\int d\mathbf{J}_i P_J(\mathbf{J}_i) \times \log\left[\int d\mathbf{A}_i P_A(\mathbf{A}_i) e^{\hat{r}\mathbf{A}_i^t\Sigma_{\setminus i}\mathbf{J}_i - \frac{\hat{l}}{2}\mathbf{A}_i^t\Sigma_{\setminus i}\mathbf{A}_i + \mathbf{A}_i^t(\hat{q}\Sigma_{\setminus i})^{1/2}\boldsymbol{\xi}}\right],$$

$$\Psi_e^{(n\to 0)} = \int \mathcal{D}\xi \int dm_i \int \frac{dz}{\sqrt{2\pi\left(\rho - \frac{r^2}{q}\right)}} e^{-\frac{\left(z + \frac{r}{\sqrt{q}}\xi\right)^2}{2\left(\rho - \frac{r^2}{q}\right)}} \times P_0(m_i|z)\log\left[\int \frac{dh}{\sqrt{2\pi l}} e^{-\frac{(h+\sqrt{q}\xi)^2}{2l}} P_G(m_i|h)\right]. \tag{B24}$$

Note that, as in standard teacher-student settings [18], in order to avoid divergent terms in this limit, the overlap ρ and its conjugate $\hat{\rho}$ need to be constrained to $\mathbb{E}_{\mathbf{J}_i}[\mathbf{J}_i^t\Sigma_{\setminus i}\mathbf{J}_i]/\nu^2$ and 0, respectively.

e. Typical free-energy density

Having determined the expression for the replicated partition function in the zero-temperature limit, we can actually compute the typical free-energy density as

$$f_\beta = -\lim_{n\to 0^+}\frac{d}{dn}\lim_{L\to\infty}\left[\frac{1}{L}\mathbb{E}_{\{\mathbf{m}_{\setminus i}^\mu, m_i^\mu\}}\mathcal{Z}_\beta^n\right] = \lim_{L\to\infty}\frac{1}{L}\int\frac{d\rho d\hat{\rho}}{2\pi}\int\frac{dr d\hat{r}}{2\pi}\int\frac{dq d\hat{q}}{2\pi}\, e^{L\Psi^{(n\to 0)}}. \tag{B25}$$

In the high-dimensional limit, we can solve the integrals over the overlap parameters by saddle-point, thus obtaining

$$f_\beta = \underset{q,r,\delta q,\hat{q},\hat{r},\delta\hat{q}}{\text{extr}}\left[\nu r\hat{r} + \frac{1}{2}(\delta q + q)(\delta\hat{q} - \hat{q}) + \frac{1}{2}q\hat{q} + \lim_{L\to\infty}\frac{1}{L}\Psi_s^{(n\to 0)} + \alpha\Psi_e^{(n\to 0)}\right], \tag{B26}$$

where the values of the overlap parameters extremizing the typical free-energy density can therefore be determined by solving the following system of coupled saddle-point equations:

$$\begin{aligned} q &= -2\frac{\partial\Psi_s^{(n\to 0)}}{\partial\delta\hat{q}}, \quad \hat{q} = 2\frac{\partial\Psi_e^{(n\to 0)}}{\partial\delta q}, \\ \delta\hat{q} &= -2\frac{\partial\Psi_e^{(n\to 0)}}{\partial q}, \quad \delta q = 2\frac{\partial\Psi_s^{(n\to 0)}}{\partial\hat{q}}. \\ r &= -\frac{1}{\nu}\frac{\partial\Psi_s^{(n\to 0)}}{\partial\hat{r}}, \quad \hat{r} = -\frac{1}{\nu}\frac{\partial\Psi_e^{(n\to 0)}}{\partial r}, \end{aligned} \tag{B27}$$

Up to this point, we have performed the replica calculation in full generality, without specifying either the interaction matrix or the loss function. In the next section, we will evaluate the typical free-energy density for the specific MLM task of Eq. (6) under the simplified assumptions of Sec. VI of the main text.

f. Zero-temperature limit and Gaussian priors

As already pointed out in the main text, the interaction matrix J is sampled from the GOE. It is then natural to assume a Gaussian prior on the ith row of the attention matrix:

$$P_A(\mathbf{A}) = \frac{1}{\sqrt{2\pi}} e^{\beta\lambda\mathbf{A}_i^t\mathbf{A}_i}, \tag{B28}$$

with β being the inverse temperature parameter, while λ is the L_2 regularization strength. Moreover, the optimization problem in Eq. (6) optimizes a square loss to solve the corresponding MLM task. This means that the Gibbs measure of Eq. (B1) associated with this task is

$$P_G(m_i|h) \propto e^{-\frac{\beta}{2}(m_i - h)^2}. \tag{B29}$$

By plugging these two specific forms of both the prior and the Gibbs measure in Eq. (B6) and taking the zero-temperature limit as exemplified in [18,40], we get the following expression for the typical free-energy density in the zero-temperature limit:

$$f \underset{\beta\to\infty}{=} \underset{q,r,\delta q,\hat{q},\hat{r},\delta\hat{q}}{\text{extr}}\left[\nu r\hat{r} - \frac{1}{2}(\delta q\hat{q} - q\delta\hat{q}) + \lim_{L\to\infty}\frac{1}{L}\Psi_s^{(n\to 0)} + \alpha\Psi_e^{(n\to 0)}\right], \tag{B30}$$

with the entropic and energetic potential given by

$$\Psi_s^{n\to 0} \underset{\beta\to\infty}{=} \frac{1}{2}[\hat{r}^2 \mathrm{tr}(\Sigma_{\setminus i})^t \Sigma_{\setminus i}(\beta\lambda\mathbb{I} + \delta\hat{q}\Sigma_{\setminus i})^{-1} + \hat{q}\mathrm{tr}\Sigma_{\setminus i}(\beta\lambda\mathbb{I} + \delta\hat{q}\Sigma_{\setminus i})^{-1}],$$
$$\Psi_e^{n\to 0} \underset{\beta\to\infty}{=} -\frac{1}{2}\frac{\nu^{-1} + \rho + q - 2r}{1 + \delta q}, \tag{B31}$$

and $\rho = \mathrm{tr}\Sigma_{\setminus i}/(\nu^2 L)$. Note that this functional form of the typical free-energy density corresponds exactly to that of supervised learning with the noisy label and the Gaussian structured data [40]. However, we should point out again that the variance of the noise in the labels, controlled by the shift factor ν, is a direct consequence of the intrinsic noise already affecting the input. Therefore, it cannot be tuned independently from it. This is also reflected in the slightly different functional forms of the saddle-point equations in (B27), which, in the zero-temperature limit with Gaussian priors and square loss, are given by

$$q = \mathrm{tr}[(\hat{q}\Sigma_{\setminus i} + \hat{r}^2\Sigma_{\setminus i}(\Sigma_{\setminus i})^t)\Sigma_{\setminus i}(\lambda\mathbb{I} + \delta\hat{q}\Sigma_{\setminus i})^{-2}],$$
$$\delta q = \mathrm{tr}[(\lambda\mathbb{I} + \delta\hat{q}\Sigma_{\setminus i})^{-1}\Sigma_{\setminus i}],$$
$$r = -\frac{\hat{r}}{\nu}\mathrm{tr}[\Sigma_{\setminus i}(\Sigma_{\setminus i})^t(\lambda\mathbb{I} + \delta\hat{q}\Sigma_{\setminus i})^{-1}],$$
$$\hat{q} = \frac{\nu^{-1} + \rho + q - 2r}{(1 + \delta q)^2},$$
$$\delta\hat{q} = \frac{1}{1 + \delta q},$$
$$\hat{r} = -\frac{1}{\nu(1 + \delta q)}. \tag{B32}$$

As in the case of supervised learning settings [18,40], the solution of this system of coupled saddle-point equations in the zero-temperature limit allows us to express the generalization loss as shown in the main text [Eq. (7)], with the exception that the noise in the labels is the direct consequence of the intrinsic noise of the inputs.

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, Attention is all you need, *Advances in Neural Information Processing Systems*, edited by I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Curran Associates, 2017), Vol. 30.

[2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv:1810.04805.

[3] J. Howard and S. Ruder, Universal language model fine-tuning for text classification, arXiv:1801.06146 [cs.CL].

[4] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, Improving language understanding with unsupervised learning, Technical Report (OpenAI, 2018), https://openai.com/research/language-unsupervised.

[5] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. 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. M. Ziegler, J. Wu, C. Winter, C. Hesse *et al.*, Language models are few-shot learners, Adv. Neural Info. Proc. Sys. **33**, 1877 (2020).

[6] OpenAI, Gpt-4 technical report, arXiv:2303.08774 [cs.CL].

[7] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, An image is worth 16x16 words: Transformers for image recognition at scale, arXiv:2010.11929 [cs.CV].

[8] J. Jumper, R. Evans *et al.*, Highly accurate protein structure prediction with AlphaFold, Nature (London) **596**, 583 (2021).

[9] D. Bahdanau, K. Cho, and Y. Bengio, Neural machine translation by jointly learning to align and translate, arXiv:1409.0473 [cs.CL].

[10] Y. Kim, C. Denton, L. Hoang, and A. M. Rush, Structured attention networks, arXiv:1702.00887 [cs.CL].

[11] E. Gardner and B. Derrida, Three unfinished works on the optimal storage capacity of networks, J. Phys. A **22**, 1983 (1989).

[12] H. S. Seung, H. Sompolinsky, and N. Tishby, Statistical mechanics of learning from examples, Phys. Rev. A **45**, 6056 (1992).

[13] A. Engel and C. Van den Broeck, *Statistical Mechanics of Learning* (Cambridge University Press, Cambridge, 2001).

[14] 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).

[15] S. Y. Chung, D. D. Lee, and H. Sompolinsky, Classification and geometry of general perceptual manifolds, Phys. Rev. X **8**, 031003 (2018).

[16] S. Goldt, M. Mézard, F. Krzakala, and L. Zdeborová, Modeling the influence of data structure on learning in neural networks: The hidden manifold model, Phys. Rev. X **10**, 041044 (2020).

[17] S. Spigler, M. Geiger, and M. Wyart, Asymptotic learning curves of kernel methods: empirical data versus teacher-student paradigm, J. Stat. Mech.: Theor. Expt. (2020) 124001.

[18] F. Gerace, B. Loureiro, F. Krzakala, M. Mézard, and L. Zdeborová, Generalisation error in learning with random features and the hidden manifold model, J. Stat. Mech.: Theor. Expt. (2021) 124013.

[19] A. Favero, F. Cagnetta, and M. Wyart, Locality defeats the curse of dimensionality in convolutional teacher-student scenarios, J. Stat. Mech.: Theor. Expt. (2022) 114012.

[20] B. Loureiro, C. Gerbelot, H. Cui, S. Goldt, F. Krzakala, M. Mézard, and L. Zdeborová, Learning curves of generic features maps for realistic datasets with a teacher-student model, J. Stat. Mech.: Theor. Expt. (2022) 114001.

[21] A. Ingrosso and S. Goldt, Data-driven emergence of convolutional structure in neural networks, Proc. Natl. Acad. Sci. USA **119**, e2201854119 (2022).

[22] M. Refinetti and S. Goldt, The dynamics of representation learning in shallow, non-linear autoencoders, in *International Conference on Machine Learning* (PMLR, New Orleans, LA, 2022), pp. 18499–18519.

[23] H. Cui and L. Zdeborová, High-dimensional asymptotics of denoising autoencoders, *37th Conference on Neural Information Processing Systems (NeurIPS 2023)* (Curran Associates, 2024), Vol. 36.

[24] R. B. Potts, Some generalized order-disorder transformations, Math. Proc. Cambridge Philos. Soc. **48**, 106 (1952).

[25] F. Y. Wu, The Potts model, Rev. Mod. Phys. **54**, 235 (1982).

[26] For a vector $\mathbf{x} = (x_i)$, the softmax nonlinearity yields $x'_i = \exp(x_i)/\sum_j \exp(x_j)$. The elements of $\mathbf{x}'$ can be interpreted as a probability distribution, since they are positive and sum to 1.

[27] R. D. Finn, A. Bateman *et al.*, Pfam: the protein families database, Nucl. Acids Res. **42**, D222 (2014).

[28] N. Bhattacharya, N. Thomas, R. Rao, J. Daupras, P. K. Koo, D. Baker, Y. S. Song, and S. Ovchinnikov, Interpreting Potts and transformer protein models through the lens of simplified attention, *Biocomputing 2022* (World Scientific, Singapore, 2022), pp. 34–45.

[29] L. L. Viteritti, R. Rende, and F. Becca, Transformer variational wave functions for frustrated quantum spin systems, Phys. Rev. Lett. **130**, 236401 (2023).

[30] R. Rende, L. L. Viteritti, L. Bardone, F. Becca, and S. Goldt, A simple linear algebra identity to optimize large-scale neural network quantum states, arXiv:2310.05715 [cond-mat.str-el].

[31] L. L. Viteritti, R. Rende, A. Parola, S. Goldt, and F. Becca, Transformer wave function for the Shastry-Sutherland model: emergence of a spin-liquid phase, arXiv:2311.16889 [cond-mat.str-el].

[32] J. Besag, Statistical analysis of non-lattice data, J. R. Stat. Soc. Ser. D **24**, 179 (1975).

[33] S. Cocco and R. Monasson, Adaptive cluster expansion for inferring Boltzmann machines with noisy data, Phys. Rev. Lett. **106**, 090601 (2011).

[34] F. Ricci-Tersenghi, The Bethe approximation for solving the inverse Ising problem: a comparison with other inference methods, J. Stat. Mech.: Theor. Expt. (2012) P08015.

[35] M. Ekeberg, C. Lövkvist, Y. Lan, M. Weigt, and E. Aurell, Improved contact prediction in proteins: Using pseudolikelihoods to infer Potts models, Phys. Rev. E **87**, 012707 (2013).

[36] S. Cocco, C. Feinauer, M. Figliuzzi, R. Monasson, and M. Weigt, Inverse statistical physics of protein sequences: a key issues review, Rep. Prog. Phys. **81**, 032601 (2018).

[37] A. Hyvärinen, Consistency of pseudolikelihood estimation of fully visible Boltzmann machines, Neural Comput. **18**, 2283 (2006).

[38] P. Ravikumar, M. J. Wainwright, and J. D. Lafferty, High-dimensional Ising model selection using ℓ_1-regularized logistic regression, Ann. Statist. **38**, 1287 (2010).

[39] E. Aurell and M. Ekeberg, Inverse Ising inference using all the data, Phys. Rev. Lett. **108**, 090201 (2012).

[40] S. Goldt, B. Loureiro, G. Reeves, F. Krzakala, M. Mézard, and L. Zdeborová, The Gaussian equivalence of generative models for learning with shallow neural networks, in *Proceedings of the 2nd Mathematical and Scientific Machine Learning Conference* (PMLR, 2022), pp. 426–471.

[41] F. Morcos, A. Pagnani, B. Lunt, A. Bertolino, D. S. Marks, C. Sander, R. Zecchina, J. N. Onuchic, T. Hwa, and M. Weigt, Direct-coupling analysis of residue coevolution captures native contacts across many protein families, Proc. Natl. Acad. Sci. USA **108**, E1293 (2011).

[42] H. C. Nguyen, R. Zecchina, and J. Berg, Inverse Statistical problems: from the inverse Ising problem to data science, Adv. Phys. **66**, 197 (2017).

[43] G. Livan, M. Novaes, and P. Vivo, *Introduction to Random Matrices* (Springer, Cham, 2018).

[44] M. Opper and W. Kinzel, Statistical mechanics of generalization, *Models of Neural Networks III: Association, Generalization, and Representation* (Springer, New York, NY, 1996).

[45] P. Nakkiran, G. Kaplun, Y. Bansal, T. Yang, B. Barak, and I. Sutskever, Deep double descent: where bigger models and more data hurt, J. Stat. Mech.: Theor. Expt. (2021) 124003.

[46] D. Saad and S. A. Solla, Exact solution for on-line learning in multilayer neural networks, Phys. Rev. Lett. **74**, 4337 (1995).

[47] M. Biehl and H. Schwarze, Learning by on-line gradient descent, J. Phys. A **28**, 643 (1995).

[48] S. Goldt, M. S. Advani, A. M. Saxe, F. Krzakala, and L. Zdeborová, Dynamics of stochastic gradient descent for two-layer neural networks in the teacher-student setup, J. Stat. Mech.: Theor. Expt. (2020) 124010.

[49] R. Veiga, L. Stephan, B. Loureiro, F. Krzakala, and L. Zdeborová, Phase diagram of stochastic gradient descent in high-dimensional two-layer neural networks, Adv. Neur. Inf. Proc. Syst. **35**, 23244 (2023).

[50] C. Zhang, M. Raghu, J. Kleinberg, and S. Bengio, Pointer value retrieval: A new benchmark for understanding the limits of neural network generalization, arXiv:2107.12580 [cs.LG].

[51] Y. Zhang, A. Backurs, S. Bubeck, R. Eldan, S. Gunasekar, and T. Wagner, Unveiling transformers with LEGO: a synthetic reasoning task, arXiv:2206.04301 [cs.LG].

[52] E. Abbe, S. Bengio, E. Cornacchia, J. Kleinberg, A. Lotfi, M. Raghu, and C. Zhang, Learning to reason with neural networks: Generalization, unseen data and Boolean measures, Adv. Neural Info. Proc. Sys. **35**, 2709 (2022).

[53] A. Seif, S. A. M. Loos, G. Tucci, É Roldán, and S. Goldt, The impact of memory on learning sequence-to-sequence tasks, Mach. Learn.: Sci. Technol. **5**, 015053 (2024).

[54] J. Bradbury, R. Frostig, P. Hawkins, M. J. Johnson, C. Leary, D. Maclaurin, G. Necula, A. Paszke, J. Vander Plas, S. Wanderman-Milne, and Q. Zhang, JAX: composable transformations of Python+NumPy programs, version 0.3.13 (2018), http://github.com/google/jax.

[55] P. Lippe, Welcome to the UvA deep learning tutorials, https://uvadlc-notebooks.readthedocs.io/en/latest/ (2022).