

Loop series expansions for tensor networks

Glen Evenbly ^{*}, Nicola Pancotti , and Ashley Milsted 

AWS Center for Quantum Computing, Pasadena, California 91125, USA
and California Institute of Technology, Pasadena, California 91125, USA

Johnnie Gray  and Garnet Kin-Lic Chan

California Institute of Technology, Pasadena, California 91125, USA



(Received 26 March 2025; accepted 13 January 2026; published 5 March 2026)

Belief propagation (BP) can be a useful tool to approximately contract a tensor network, provided that the contributions from any closed loops in the network are sufficiently weak. In this article, we describe how a loop series expansion can be applied to systematically improve the accuracy of a BP approximation to a tensor network contraction, in principle converging arbitrarily close to the exact result. More generally, our result provides a framework for expanding a tensor network as a sum of component networks in a hierarchy of increasing complexity. We benchmark this proposal for the contraction of infinite projected entangled pair states, either representing the ground state of an Affleck-Kennedy-Lieb-Tasaki (AKLT) model or with randomly defined tensors, where it is shown to improve in accuracy over standard BP by several orders of magnitude while incurring only a minor increase in computational cost. These results indicate that the proposed series expansions could be a useful tool to accurately evaluate tensor networks in cases that otherwise exceed the limits of established contraction routines.

DOI: [10.1103/vqks-cr6x](https://doi.org/10.1103/vqks-cr6x)

I. INTRODUCTION

Tensor networks (TNs) are powerful tools for the classical simulation of quantum systems [1–10]. Although they were traditionally developed for model systems in the context of condensed matter physics, they have more recently been applied toward the benchmarking of nascent quantum computers [11–21]. Tensor network states were pioneered by the introduction of matrix product states [22,23] (MPSs), which possess tensors connected in a one-dimensional (1D) geometry. For several decades, numeric algorithms based on the MPS, such as the density matrix renormalization group [24–26], have been widely regarded as the gold standard of classical simulation algorithms for 1D quantum systems, producing high-accuracy and unbiased results for a wide range of problems. A large part of this success can be attributed to the computational efficiency with which MPS can be contracted. Subsequent developments have seen the proposal of sophisticated tensor networks, such as projected entangled pair states [27,28] (PEPSs) and multiscale entanglement renormalization ansatz [29,30], designed to efficiently represent quantum states on large two-dimensional (2D) systems. However, a significant challenge of these 2D tensor networks is

the high computational expense required for their contraction [31] even when approximate contraction schemes are used [32–38]. In practice, this expense limits the amount of entanglement that can be encoded in these networks and therefore also limits the range of problems for which they produce accurate results. Thus, the development of strategies for more efficient contraction of TNs remains a vital task.

A promising approach for the contraction of TNs is through the adoption of belief propagation [39–41] (BP), a message passing algorithm with a long history in both computer science [42–45] and in statistical physics [46–48]. In recent times, the application of BP toward TNs has generated much interest [49–58] and produced a notable success [20] of simulating a quantum experiment performed on IBM’s 127-qubit Eagle processor [59], a problem that was seen to be intractable for other 2D TN algorithms [60]. However, a major drawback of BP is that it is an *uncontrolled approximation*; one cannot systematically improve the accuracy of a BP result as is otherwise done in the context of tensor networks by simply growing the bond dimension. Nevertheless, generalized variants of BP [43–45] can offer improved accuracy, and additional schemes based on a preliminary clustering of the network have also been proposed to enhance BP’s performance [55].

In this article, we present a method to improve upon the accuracy of BP for the contraction of a tensor network by incorporating and building upon the loop series expansion proposed by Chertkov and Chernyak [61,62]. Assuming that a BP fixed point of the tensor network has been found, we describe how the dominant loop excitations, otherwise neglected by BP, can be reincorporated into the evaluation of this network. This approach can also be understood as providing a

^{*}Contact author: evenbly@amazon.com

general notion of a *series expansion for tensor networks* that a complicated tensor network may be expanded as a sum of component networks in a hierarchy of increasing complexity, with the BP approximation as the zeroth-order term in the series. We demonstrate numerically that the proposed series expansion can provide both accurate and computationally cheap evaluations of PEPS, such that it could be a useful tool to contract tensor networks in cases that otherwise exceed the limits of established contraction routines.

II. BELIEF PROPAGATION

Here, we provide a brief introduction to BP as a necessary precursor to our main results. Let $\mathcal{T}$ be a closed tensor network composed of N tensors labeled $\{T_1, T_2, T_3, \dots, T_N\}$ and with M indices labeled $\{i_1, i_2, i_3, \dots, i_M\}$. Assume that our goal is to contract $\mathcal{T}$ to a scalar $\mathcal{Z}(\mathcal{T})$,

$$\mathcal{Z}(\mathcal{T}) = \sum_{i_1, i_2, \dots} T_1(\xi_1) T_2(\xi_2) T_3(\xi_3) \dots, \quad (1)$$

where ξ_r describes the subset of indices $\xi_r \subseteq \{i_1, i_2, \dots, i_M\}$ connected to tensor T_r . The starting point of BP is to define a pair of vectors, called messages, for each edge of the network; if (r, s) denotes an edge shared between tensors T_r and T_s , then we define messages $\{\mu_{r \rightarrow s}, \mu_{s \rightarrow r}\}$ on this edge. We refer to a message $\mu_{s \rightarrow r}$ as outgoing from T_s and incoming to T_r . BP prescribes the following algorithm to update each of the messages: The updated outgoing message for T_r toward tensor T_s is given by contracting the incoming messages to T_r on all other indices,

$$\mu_{r \rightarrow s} = \left(\prod_{q \in \xi_r/s} \mu_{q \rightarrow r} \right) T_r. \quad (2)$$

The set of $2M$ messages realizes a *BP fixed point* if all of the equations are simultaneously satisfied up to multiplicative constants. Let us assume that we have found a BP fixed point for our network $\mathcal{T}$ and that each matching pair of messages has been normalized such that their dot product (i.e., without complex conjugation) is unity, $\langle \mu_{r \rightarrow s} \cdot \mu_{s \rightarrow r} \rangle = 1$. Let us define scalar $\tilde{T}_r$ as resulting from the contraction of a tensor T_r with all incoming messages,

$$\tilde{T}_r = \left(\prod_{q \in \xi_r} \mu_{q \rightarrow r} \right) T_r, \quad (3)$$

which we refer to as the (BP) vacuum contribution from tensor T_r . The BP approximation to the network contraction $\mathcal{Z}$ from Eq. (1) is given by multiplication of these scalars,

$$\mathcal{Z}(\mathcal{T}) \approx \prod_r \tilde{T}_r. \quad (4)$$

It follows that the free energy $\mathcal{F} = -\log \mathcal{Z}$ can be approximated as

$$\mathcal{F}(\mathcal{T}) \approx - \sum_r \log(\tilde{T}_r), \quad (5)$$

which is also known as the Bethe free energy [39].

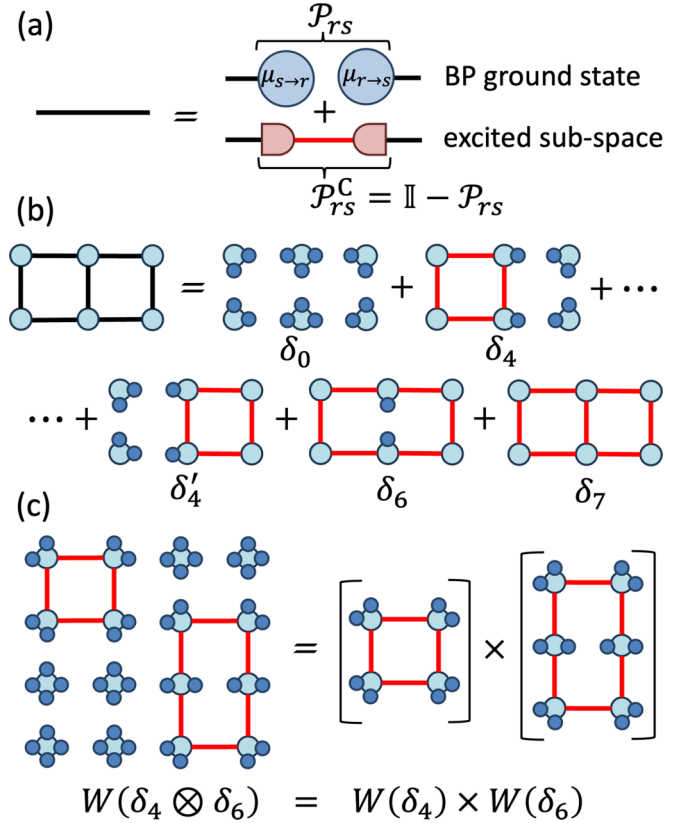


FIG. 1. (a) An edge within a tensor network is resolved into the sum of a projector onto the BP ground state, formed from the outer product of the BP messages, and the complementary projector onto the excited subspace. (b) A network is expanded as a sum of five configurations, where indices in each configuration have been projected onto either the ground or excited subspace. The zero-weight configurations (i.e., those containing dangling excitations) have been omitted. (c) The weight of a configuration containing several distinct excitations is equal to the product of the individual excitation weights [assuming that the vacuum contributions defined in Eq. (3) have been normalized to unity].

III. CONFIGURATIONS IN THE BP BASIS

We now outline a methodology to improve upon the accuracy of a known BP fixed point by introducing additional terms into Eq. (4). For any edge (r, s) of the network, assumed to be of dimension d , we define a rank-1 projector $\mathcal{P}_{rs}$ onto the message subspace and its rank $(d - 1)$ complement $\mathcal{P}_{rs}^C$,

$$\mathcal{P}_{rs} = \mu_{s \rightarrow r} \otimes \mu_{r \rightarrow s}, \quad \mathcal{P}_{rs}^C = \mathbb{I} - \mathcal{P}_{rs}, \quad (6)$$

where $\mathbb{I}$ as the d -dim identity [see also Fig. 1(a)]. We henceforth refer to $\mathcal{P}_{rs}$ as the projector onto the (BP) ground state of edge (r, s) and $\mathcal{P}_{rs}^C$ as the projector onto the excited subspace. It follows that the tensor network $\mathcal{T}$ can be resolved as a sum over 2^M (BP basis) configurations, formed from all combinations of projections of each edge into either the ground or excited subspace. We define the *degree* of a configuration as the number of excited edges that it contains, using δ_x to denote a degree- x configuration and $W(\delta_x)$ to denote its *weight* (i.e., the scalar resulting from its contraction). Thus, we can recognize δ_0 as the *BP vacuum*, the configuration with all edges

projected into their BP ground state. It follows that $W(\delta_0)$ evaluates to the standard BP result of Eq. (4). Our goal is now to characterize the weights of the remaining configurations, which may be thought of as excitations on top of the BP vacuum, such that the high-weight contributions can then be reintroduced into the approximation.

A significant step toward this goal comes from the understanding derived in Ref. [61] that any configuration with a *dangling excitation*, i.e., contains a tensor with a single excited index, has weight zero. This result follows directly from the definition of the BP fixed point; let us assume that edge (r, s) of tensor T_r is projected via $\mathcal{P}_{rs}^C$ into the excited subspace and all other edges are in the BP ground space. It follows that

$$\left(\prod_{q \in \xi_r/s} \mu_{q \rightarrow r} \right) T_r \mathcal{P}_{rs}^C = \mu_{r \rightarrow s} \mathcal{P}_{rs}^C = 0, \quad (7)$$

where Eq. (2) was used to replace T_r with the outgoing message $\mu_{r \rightarrow s}$, which is, by construction in Eq. (6), orthogonal to the projector $\mathcal{P}_{rs}^C$. This result, which implies that any nontrivial excitation must contain one or more closed loops, is useful in reducing the number of basis configurations that need be considered. For instance, when applied to the network of $M = 7$ edges in Fig. 1(b), the number of configurations is reduced from $2^7 = 128$ to only five configurations of nonzero weight labeled $\{\delta_0, \delta_4, \delta'_4, \delta_6, \delta_7\}$. Note that, although Eq. (7) is strictly only valid for an exact BP fixed point, we find that any discrepancies when applied to a (well-converged) numerically optimized BP fixed point are negligibly small.

Another useful observation, as examined in Ref. [63], is that the weight of a configuration containing separate excitations, i.e., excitations that do not share any common tensors, is given by the product of the individual weights

$$W(\delta_x \otimes \delta_y) = W(\delta_x) \times W(\delta_y). \quad (8)$$

See also the example shown in Fig. 1(c). This result is further simplified by rescaling each tensor T_r so that its vacuum state contribution, i.e., $\tilde{T}_r$ from Eq. (3), is unity, which allows the vacuum component of any configuration to be ignored. This observation implies that one only needs to understand the connected excitations in order to fully characterize a tensor network contraction.

For any reasonably large tensor network, the full number of nontrivial excitations will still exceed practical consideration. However, as detailed further in Appendix A, we posit that in most spatially homogeneous tensor networks, the weight of a configuration δ_x will be exponentially suppressed in its degree x ,

$$W(\delta_x) \approx \exp(-kx), \quad (9)$$

with k a positive scalar. While it is easy to construct counterexamples where this scaling will not hold, it is argued in Appendix A that this behavior will hold for networks where the BP fixed point already provides a reasonably accurate approximation to the network contraction. Thus, one can understand that the (BP basis) configurations, when ordered in increasing degree, form a series expansion for a network contraction. While the number of δ_x configurations may still grow (exponentially) quickly with x , the exponential suppression of the contribution of each configuration provides the possibility

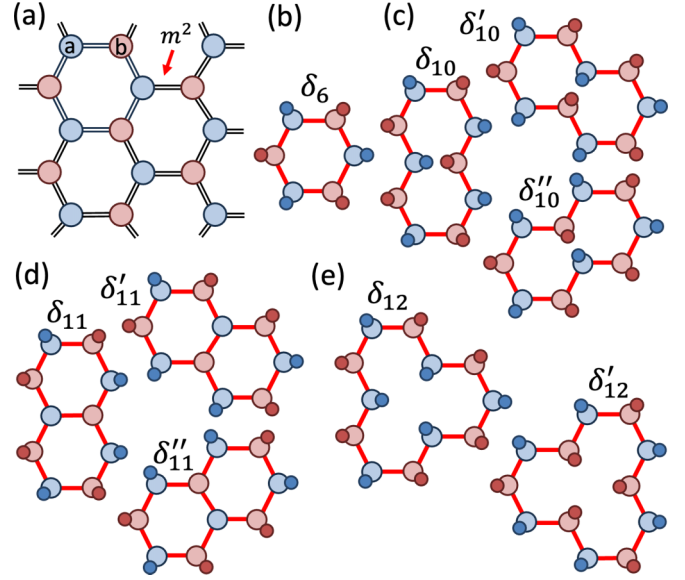


FIG. 2. (a) A closed tensor network representing $\mathcal{T} = \langle \psi | \psi \rangle$ results from contracting the physical indices of a hexagonal lattice PEPS $|\psi\rangle$, bond dimension m , with its conjugate. (b)–(e) Depiction of the excitations δ_x up to 12th degree.

that a partial summation will rapidly converge to the exact result as higher-degree excitations are added.

IV. CONTRACTION OF IPEPS

In this section, we discuss how the proposed series expansion can be applied to contract infinite projected entangled pair states (iPEPSs). For concreteness, we focus on a hexagonal lattice PEPS $|\psi\rangle$ with local dimension d and virtual dimension m , formed from a single pair of unique tensors A and B tiled on alternate sites (such that the state has a two-site unit cell). A closed hexagonal lattice tensor network $\mathcal{T} = \langle \psi | \psi \rangle$ of bond dimension m^2 is formed by taking the product of the PEPS with its conjugate and contracting over the site indices [see also Fig. 2(a)]. The network $\mathcal{T}$ is composed of two unique tensors a and b , which result from contraction of the physical index in $(A \otimes A^*)$ and $(B \otimes B^*)$, respectively. In what follows, we assume that we have found a BP fixed point for $\mathcal{T}$ and that tensors a and b have been rescaled such that their vacuum contributions, as defined in Eq. (3), are unity.

A. Free-energy density

A common task in tensor network algorithms is to evaluate the scalar $\mathcal{Z}$ associated with contracting a closed tensor network, as per Eq. (1). However, in the context of an infinite tensor network this scalar would diverge, so we instead evaluate an intensive equivalent: the free-energy density f ,

$$f = \lim_{N \rightarrow \infty} \left(\frac{-\log(\mathcal{Z}_N)}{N} \right), \quad (10)$$

where $\mathcal{Z}_N$ represents the evaluation of a patch of N tensors from the full network. Note that due to our normalization of

the BP vacuum, the BP contribution to $\mathcal{Z}_N$ is unity for all N (or equivalently, the Bethe free energy is zero [39]).

Our first step is to evaluate the weights $W(\delta_x)$ of the lowest-degree excitations δ_x ; in the calculations performed in this article, we evaluate the excitations up to degree $x = 14$ [see Figs. 2(b)–2(e)]. Each of these weights is given by contracting a small tensor network formed by capping the external indices (i.e., of tensors within the excitation) with their fixed-point messages and projecting the internal indices onto their excited subspace with projectors $\mathcal{P}_{rs}^C$ as defined in Eq. (6). However, even once these weights $W(\delta_x)$ are known, it still remains a nontrivial combinatorial task to compute their impact on the free-energy density f . Here, we shall follow the approximate counting strategy derived in Appendix B 2, which gives the following result for the free-energy correction:

$$f \approx - \sum_{\{\delta_i\}} L(\delta_x) W(\delta_x) e^{S(\delta_x)f}. \quad (11)$$

Here, $L(\delta_x)$ is the number of locations per lattice site that the excitation can be placed and $S(\delta_x)$ as the number of hexagonal plaquettes that the excitation occupies (as detailed further in Appendix B). For instance, the δ_6 excitation shown in Fig. 2(b) has $L(\delta_6) = 1/2$ (since there is one hexagon per two lattice sites) and $S(\delta_6) = 7$. Given that Eq. (11) contains the unknown f on both sides, we cannot solve for f directly. Instead, we solve Eq. (11) iteratively by first setting $f = 0$ on the right-hand side, computing an updated value for f , then feeding updated f back into right-hand side, and repeating until f is converged.

B. Transfer matrix

We now examine how one can compute the transfer matrix T_{AB} associated with cutting open a single edge from the network $\mathcal{T} = \langle \psi | \psi \rangle$, as shown in Fig. 3(a). Such transfer matrices are utilized in determining the optimal truncation of an internal index from the PEPS and thus play an important role in many PEPS optimization algorithms [32–37].

The BP vacuum contribution to T_{AB} is shown in Fig. 3(b), together in Fig. 3(c) with examples of contributions from loop excitations, all of which individually evaluate to a matrix with the same dimensions as T_{AB} . The series approximation to T_{AB} is constructed by evaluating each of the loop-excitation networks (up to some desired degree), then adding the contributions together after weighting each by a suppression factor $\exp(S(\delta_x)f)$ similar to Eq. (11). Note this requires that the free-energy density f be evaluated prior to computing T_{AB} . Finally, we remark that the excitation terms that do not include either of the tensors adjacent to an open index do not need to be considered explicitly; their contributions, which are proportionate to the vacuum term, are already accounted for through the suppression factors.

C. Density matrix

We now describe how the density matrix ρ_{AB} associated with a pair of adjacent lattice sites can be computed. Let us first define tensor a' from taking $(A \otimes A^*)$, while leaving the physical indices uncontracted, and similarly for b' [see Fig. 3(e)]. The network for ρ_{AB} can be realized from a

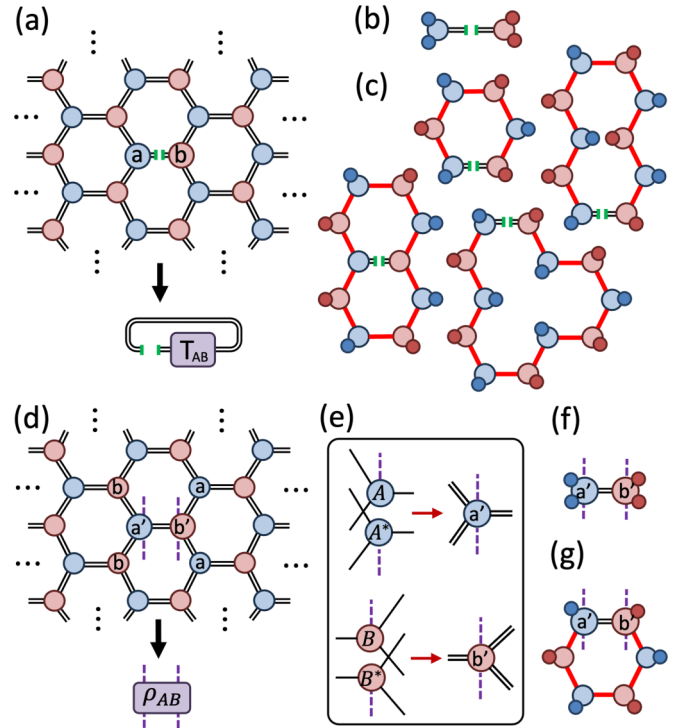


FIG. 3. (a) The transfer matrix T_{AB} is formed by contracting the network after a single edge has been cut. (b) The BP vacuum contribution to T_{AB} . (c) Several examples of loop corrections to T_{AB} . (d) The density matrix ρ_{AB} is obtained from contracting the network with a pair of impurity tensors a' and b' that have their physical indices left open. (e) Definition of the impurity tensors a' and b' . (f) The BP vacuum contribution to ρ_{AB} . (g) A loop correction to ρ_{AB} .

substitution of a' and b' as impurities in the original network $\mathcal{T} = \langle \psi | \psi \rangle$, as shown in Fig. 3(d).

Similarly to the transfer matrix, only the loop excitations that include one (or both) of the tensors a' and b' will give nontrivial contribution to ρ_{AB} . The BP vacuum contribution to ρ_{AB} is shown in Fig. 3(f), together in Fig. 3(g) with an example of a loop excitation. The series approximation to ρ_{AB} is thus constructed in the same way as the transfer matrix; the contributing loop-excitation networks are evaluated up to some desired degree and then added together after weighting each by a suppression factor $\exp(S(\delta_x)f)$.

V. BENCHMARK RESULTS

We now benchmark our approach, beginning with contraction of a PEPS representing the ground state $|\psi\rangle$ of the hexagonal lattice AKLT model [64,65] in the thermodynamic limit. The AKLT model is particularly useful as a test bed, since $|\psi\rangle$ can be represented exactly [27] with an iPEPS of bond dimension $m = 2$. Following the methodology prescribed in Sec. IV, we compute the free-energy density f , transfer matrix T_{AB} , and the density matrix ρ_{AB} for n th-degree loop expansions with $n = \{6, 10, 12, 14\}$. Here, an n th-degree expansion is defined as including all nontrivial generalized loop excitations of degree n or lower.

We compare against evaluations using a boundary MPS contraction [33] whose dimension χ has been chosen large

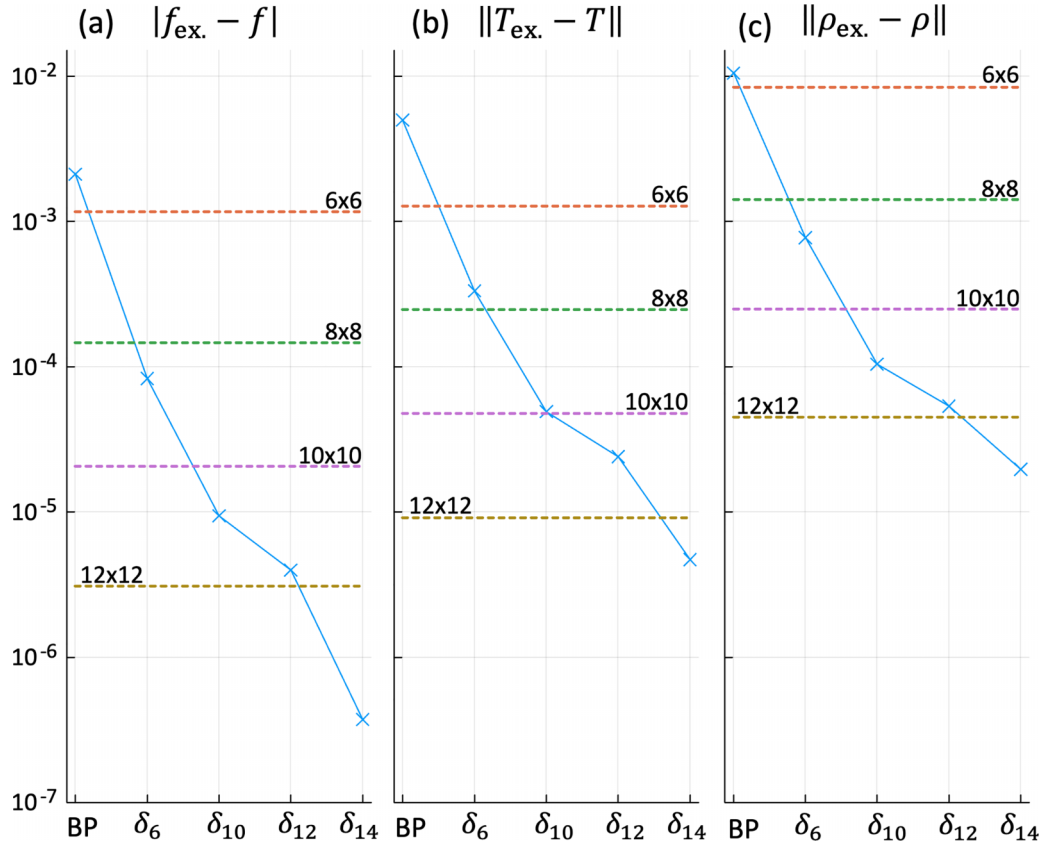


FIG. 4. Results comparing the accuracy of BP and loop expansion up to degree δ_x against exact results for the ground state of the (thermodynamic limit) hexagonal AKLT model. Displayed are errors in (a) the free-energy density f , (b) the transfer matrix T_{AB} , and (c) the two-site density matrix ρ_{AB} . Dashed lines show results from an exact contraction of an $N \times N$ system.

enough, $\chi \approx 30$, such that any truncation errors are negligible and the results can be considered as numerically exact. Figure 4 shows that the addition of the loop corrections improves the accuracy of the BP results dramatically, such that they appear to converge toward the exact results as the degree of the expansion is increased. The expansion to 14th-degree excitations δ_{14} gave the free-energy density f accurate to $\varepsilon = 3 \times 10^{-7}$, which represents an improvement of 4 orders of magnitude over the BP result. That the infinite hexagonal tensor network can be characterized so accurately from only the BP vacuum plus a few of the low-degree loop excitations is remarkable. Note that we also compare against results obtained from the exact contraction of the PEPS on a finite $N \times N$ geometry (with periodic boundaries) for $N = \{6, 8, 10, 12\}$. In all cases, we see that an N th-degree loop expansion roughly approximates the accuracy of an exact evaluation on a $N \times N$ geometry, providing additional confirmation that the loop expansions are properly accounting for all short-range contributions.

In Fig. 5(a), we present results for the accuracy of the free-energy density evaluated from randomly defined iPEPS (with two-site unit cell) of local dimension d and bond dimension m . These were prepared by placing on each edge of the lattice a maximally entangled pair of $\text{dim-}m$ particles and then projecting each vertex onto a $\text{dim-}d$ space via a $(m^3 \times d)$ isometry. The isometries were formed by first populating a $(m^3 \times d)$ matrix with random elements chosen uniformly

from the interval $(0, 1)$, then enforcing rotational symmetry (useful to reduce the number of unique terms in the loop expansion), and finally orthogonalizing the columns of the matrix. The results from random iPEPS, which are broadly similar to those from the AKLT model, demonstrate that the series expansion seems to be valid for typical instances of iPEPS. Notice that increasing the local dimension d (while keeping a fixed bond dimension m) was generally found to reduce the accuracy of our approach; this could be expected as increasing d also increases the per-site entanglement entropy of the random PEPS.

Figures 5(b) and 5(c) present equivalent results for kagome and square lattices. For iPEPS defined on kagome lattices, we perform local manipulations to exactly transform the network into a hexagonal geometry (similar to the form of PEPS described in Ref. [66]) prior to applying the series expansion (see Appendix C for details). The square-lattice results were obtained by applying the expansion directly to the square lattice, evaluated up to the δ_8 excitations as shown in Appendix C. Similar to the hexagonal lattice, the results from the kagome and square lattices also show robust improvements over the BP approximation. However, we see that the convergence on the square lattice is slower with respect to increasing the degree of expansion; we attribute this to the fact that the quantity of distinct loop configurations at any given degree grows much faster than on the hexagonal lattice.

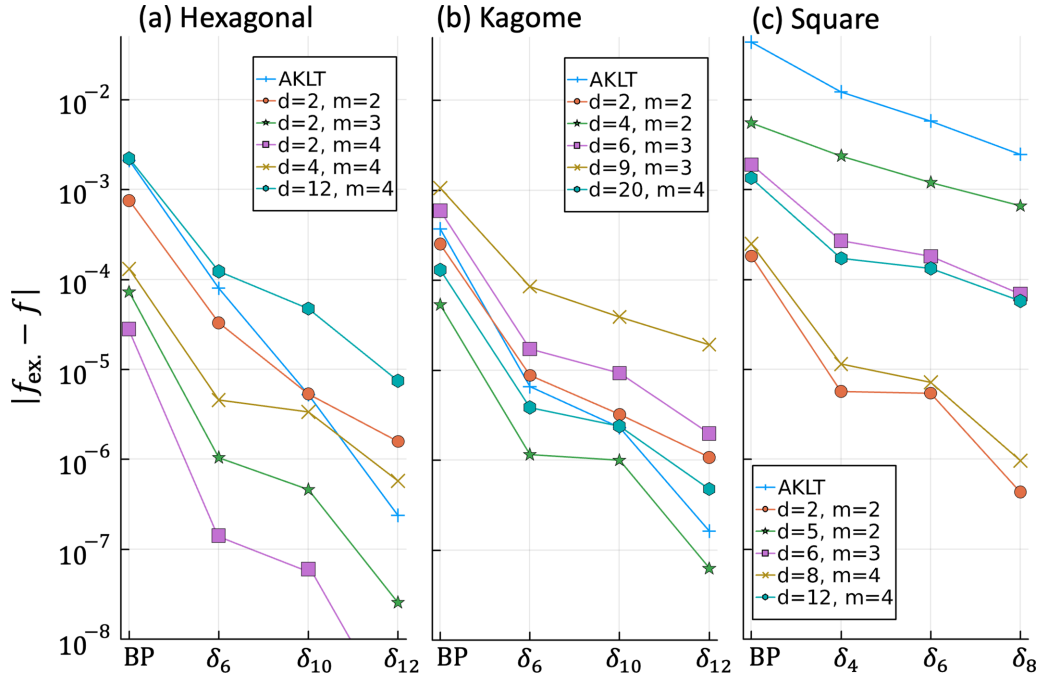


FIG. 5. Results comparing the accuracy of BP and loop expansion up to degree δ_s for evaluation of the free-energy density f from iPEPS on either (a) the hexagonal lattice, (b) the kagome lattice, or (c) the square lattice. The iPEPS either encode the AKLT ground state (on each respective lattice geometry) or are formed from random tensors with local dimension d and bond dimension m .

VI. DISCUSSION

A key result presented in this article is a systematic way to improve on BP results for the contraction of tensor networks, turning BP from an uncontrolled approximation into a well-controlled approximation. This result would be directly useful in contexts where BP has been employed for the contraction of tensor networks, such as the simulation of quantum experiments [20].

Another direct application of these results could be for PEPS optimization algorithms. Given that recent work [53,58] established the equivalence of BP to the “simple update” strategy for PEPS, the series expansion could significantly enhance the accuracy of the “simple update” strategy while still remaining relatively computationally efficient. For instance, the genus-1 loop excitations on the hexagonal lattice, i.e., the δ_6 and δ_{10} terms in Fig. 2, can be evaluated with $O(m^6)$ cost for PEPS of bond dimension m , and this cost could also potentially be reduced via approximation strategies like sparse diagonalization. Thus, the dominant terms in the expansion are still feasible to compute even for PEPS with bond dims in excess of $m = 100$, well beyond the computational limits of boundary MPS or corner transfer matrix methods for PEPS contractions [33–38]. A general comparison of the computational efficiency of loop expansions versus MPS-based methods for iPEPS contractions remains beyond the scope of the present article, given that such a comparison would be heavily dependent on the properties of the state under consideration. Nonetheless, in Appendix D we demonstrate that certain (short-range correlated) PEPS states can indeed be contracted more efficiently using the loop expansion than with MPS-based methods.

We also expect the series expansions to be useful in the context of coarse-graining tensor algorithms for tensor networks, especially in relation to proposals that extend the tensor renormalization group [67] by including mechanisms to identify and remove short-range loop correlations [68–71]. Given that our results provide a general framework precisely for efficiently identifying loop correlations, it is likely that they would find significant utility in this setting.

Although our proposal worked well for the test cases considered, it is clear that there are also several potential ways that it could fail. Most obviously, application of the series expansion first requires finding a BP fixed point. It is unknown how difficult this task is for a PEPS representing the ground state of a local Hamiltonian or other physically relevant tensor networks. Even if a BP fixed point is found, the series expansion would not be expected to converge if the network under consideration possessed multiple degenerate (or almost degenerate) fixed points, as is typical in, for instance, glassy systems. In this scenario, Eq. (9) would fail; relative to any single BP fixed point, there would exist high-degree configurations that still carry large weight (i.e., related to a different BP fixed point). An example of this would be a PEPS encoding of a Greenberger-Horne-Zeilinger (GHZ) state, where a BP fixed point could realize either the $|\uparrow\uparrow\uparrow\cdots\rangle$ or $|\downarrow\downarrow\downarrow\cdots\rangle$ component of the network; relative to one of these fixed points, the other component would represent a maximal-degree excitation with equal weight.

Most existing tensor network algorithms are built on the twin pillars of tensor contractions and decompositions. The results presented in this article, which prescribe a

general procedure for resolving a complicated network into its fundamental components, could represent a third pillar for tensor network algorithms, one that could have interesting synergies with established routines. Exploring the ways that series expansions could complement or enhance existing tensor network algorithms remains an interesting avenue for future research.

ACKNOWLEDGMENTS

We thank AWS for supporting the quantum computing program. G.K.C. acknowledges support from the U.S. DOE, Office of Science, National Quantum Information Science Research Centers, Quantum Systems Accelerator (QSA), and a generous gift from AWS. We acknowledge funding provided by the Institute for Quantum Information and Matter, an NSF Physics Frontiers Center (NSF Grant No. PHY-2317110).

DATA AVAILABILITY

The data supporting this study's findings are available within the article.

APPENDIX A: EXPONENTIAL SUPPRESSION OF HIGH-DEGREE EXCITATIONS

The utility of the proposed loop expansion is built on the premise that the weights of excitations (i.e., on top of the BP vacuum) are exponentially diminishing in their degree; this property allows us to accurately approximate a tensor network contraction using only a few low-degree excitations. In this Appendix, we attempt to justify this premise using an analysis similar to the one used to show that MPSs are finitely correlated [4].

Let $\mathcal{T}$ be a finite tensor network containing at least a single closed loop. We assume that a BP fixed point has been found and that the tensors in network $\mathcal{T}$ have been normalized such that their vacuum contributions are unity (or equivalently that the Bethe free energy is zero). Finally, we also assume that the BP fixed point accurately approximates the network contraction; that the difference between the full network contraction $\mathcal{Z}(\mathcal{T})$ and the BP approximation $\mathcal{Z}_{\text{BP}}$ is small. We now consider isolating a single length- x loop of the network by absorbing the fixed-point messages on all edges outside of the loop into their adjoining tensors, such that the reduced network is realized as the trace of a product of matrices, as shown in the example of Fig. 6(a). For simplicity, we assume that the matrices within the loop are copies of a single matrix A , which would only occur in practice if the original network possessed sufficient spatial symmetries, although the essence of our argument does not rely on this assumption. It follows that the network contraction can be written as $\mathcal{Z}(\mathcal{T}) = \text{tr}\{A^x\}$. The reduced network can be decomposed as the BP vs the non-BP component:

$$\mathcal{Z}(\mathcal{T}) = \text{tr}\{(A\mathcal{P})^x\} + \text{tr}\{(A\mathcal{P}^C)^x\}, \quad (\text{A1})$$

with $\mathcal{P}$ as the rank-1 projector onto the message subspace and $\mathcal{P}^C$ its complement [see also Fig. 6(b)]. Note that $\mathcal{Z}_{\text{BP}} = \text{tr}\{(A\mathcal{P})^x\} = 1$ due to the normalization condition imposed. Given that BP fixed points are known to correspond to

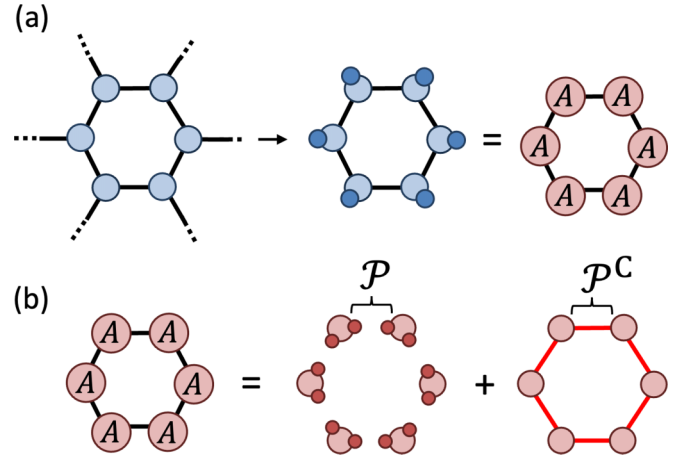


FIG. 6. (a) A loop of six tensors within an extended network is isolated by setting external edges with their fixed-point messages and then absorbing the messages into the adjoining tensors. The resulting network is given as the trace of a product of matrices A , which are assumed to be identical due to symmetry. (b) The loop network is decomposed into the BP component (projecting each edge into the message subspace) plus the loop excitation (projecting each edge into the excited subspace) [see also Eq. (A1)].

stationary points of the Bethe free energy, and that we have assumed our known fixed point accurately approximates the network contraction, it follows that the projector $\mathcal{P}$ onto message subspace should maximize $\mathcal{Z}_{\text{BP}}$. Thus, $\mathcal{P}$ must equal the outer product of the dominant left/right eigenvectors of matrix A (or, equivalently, the fixed-point messages within the loop must be the left/right eigenvectors of A). Let $\{\lambda_0, \lambda_1, \lambda_2, \dots\}$ be the eigenvalues of A ordered with descending magnitude, where the normalization condition implies that $|\lambda_0| = 1$. It follows that the weight $W(\delta_x)$ of the loop correction term can be written as

$$W(\delta_x) = \text{tr}\{(A\mathcal{P}^C)^x\} = \sum_{r=1}^d ((\lambda_r)^x). \quad (\text{A2})$$

Let us now consider how Eq. (A2) would scale in the limit of large length x (still assuming that the matrices within the loop are identical). First, we should realize that the subleading eigenvalues should have magnitude strictly less than unity, i.e., $|\lambda_1| < 1$; otherwise, the loop correction term outweighs the BP term, violating the prior assumption that the BP fixed point accurately approximates the full network contraction $\mathcal{Z}(\mathcal{T})$. In the large- x limit, Eq. (A2) will be dominated by the largest magnitude eigenvalues, so if λ_1 of A is n -fold degenerate, it follows

$$W(\delta_x) \approx n(\lambda_1)^x, \quad (\text{A3})$$

which may be rewritten as

$$W(\delta_x) \approx e^{-kx}, \quad (\text{A4})$$

with $k = -\log(\lambda_1) - (1/x)\log(n)$. Notice that k tends to a positive constant with large x , given that $\lambda_1 < 1$, such that the weight $W(\delta_x)$ tends to purely exponential decay in degree x as was posited by Eq. (9) of the main text.

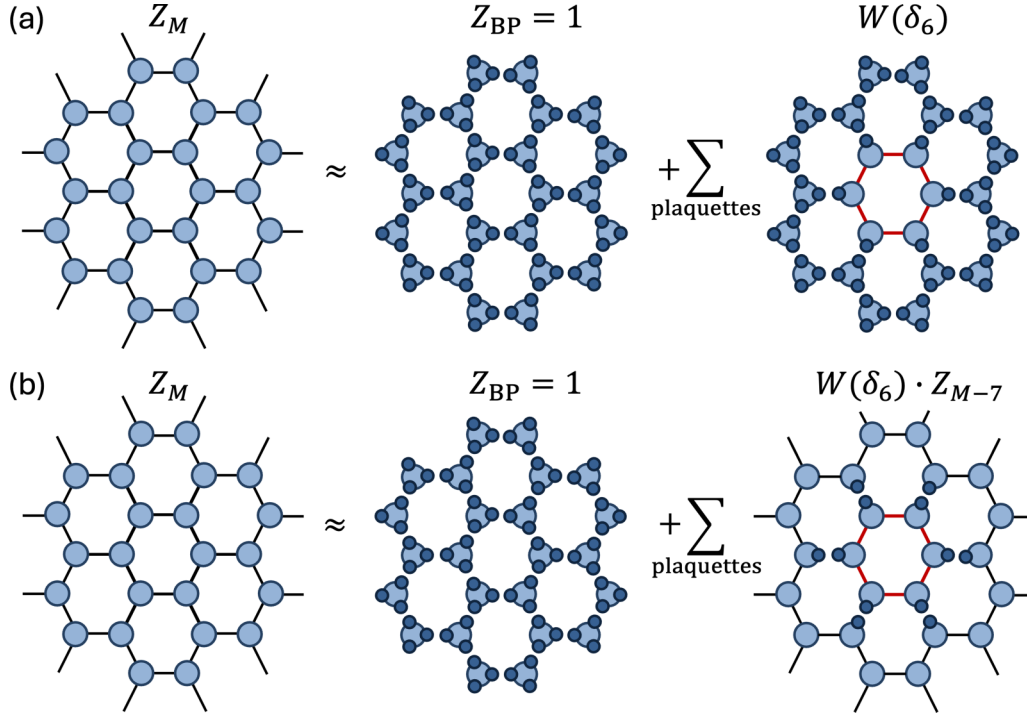


FIG. 7. (a) A hexagonal lattice tensor network is approximated via the *single-excitation* approximation for the δ_6 loop excitations. The tensor indices in each expansion term, except those which are part of the δ_6 excitation, are set in the BP ground state. (b) The *self-consistent completion* is used to expand the tensor network for the δ_6 loop excitations. In each expansion term, only the tensor indices directly connected to the δ_6 excitation are set in the BP ground state.

Although our argument assumed, for simplicity, that the transfer matrices with any loop were identical, this condition is not strictly necessary; it is only required that the transfer matrices have a gap between their dominant and subleading eigenvalues in order to realize (approximate) exponential decay. Finding a general characterization of the networks that are expected to have a gap in the transfer matrix spectrum associated with any closed loop is likely a difficult task (and certainly beyond the scope of this article). However, given that the loop series expansion is only expected to be viable for networks where the loop corrections are small in comparison to the BP approximation, it follows that the gap condition must hold in this circumstance. Finally, we remark that our argument is only valid for excitations that realize genus-1 loops, not, for instance, higher genus loops like the δ_{11} excitations shown in Fig. 2(d). Nonetheless, in practice the scaling of Eq. (A4) does seem to hold for higher genus loop excitations.

APPENDIX B: LOOP EXCITATIONS IN THE THERMODYNAMIC LIMIT

In any finite tensor network, it is possible, at least in principle, to enumerate over the different possible loop configurations, organize them by degree, and then evaluate the terms out to some desired degree. In contrast, in the thermodynamic limit, even counting the density of terms of some fixed degree may be an intractable combinatorial problem. To this end, we now discuss three different methods to approximate the loop corrections to the free-energy density f , as defined in Eq. (10), of a homogeneous tensor network on a lattice in the thermodynamic limit (e.g., an iPEPS).

1. Single-excitation approximation

The first method that we propose to estimate the contribution of loop excitations to the free-energy density involves counting only the configurations that involve only a single excitation on top of the BP vacuum. Let $\mathcal{T}_M$ be a homogeneous tensor network on a hexagonal lattice of M plaquettes that, when contracted, evaluates to some scalar $\mathcal{Z}_M$. We begin by assuming that we have already found the BP fixed point of the network, and that the network has been normalized such that the BP vacuum is unity, $\mathcal{Z}_{BP} = 1$ (or equivalently that the Bethe free energy is zero).

We now consider evaluating the corrections arising from the δ_6 excitations [see Fig. 7(a)], although our derivation is easily extendable to include arbitrary loop excitations. Given that there are M distinct locations that the δ_6 excitation could occupy, with each configuration having a weight of $W(\delta_6)$, it follows that the scalar $\mathcal{Z}_M$ of the network is

$$\mathcal{Z}_M \approx 1 + MW(\delta_6). \quad (B1)$$

The (per-site) free-energy density can thus be approximated as

$$f \approx \frac{-\log(1 + MW(\delta_6))}{2M} \approx -\frac{1}{2}W(\delta_6). \quad (B2)$$

To get to the second line of working in Eq. (B2), we have used $\log(1 + \epsilon) \approx \epsilon$, which requires $1/M \gg W(\delta_6)$ to be valid. The single-excitation approximation is only valid for small M makes sense; for larger M lattices, the contributions from configurations with double excitations (or higher) will dominate as their multiplicity scales at least as M^2 . However, given that the result from Eq. (B2) is independent of M , it may still be

used as an approximation to the free-energy density f in the thermodynamic limit. Another way of interpreting this derivation is that we are using the single-excitation approximation to calculate the free-energy density on only a small patch of the otherwise infinite system. We can easily generalize Eq. (B2) to include arbitrary excitations δ_x ,

$$f \approx \sum_{\{\delta_x\}} -L(\delta_x)W(\delta_x), \quad (\text{B3})$$

where $L(\delta_x)$ is the number of locations per lattice site that the excitation δ_x can be placed.

2. Self-consistent completion

The second method that we propose to estimate the contribution of loop excitations uses a self-consistent completion. For simplicity, we again focus on the δ_6 excitations, noting that the derivation is easily extended to include arbitrary excitations. We now resolve the hexagonal network of M plaquettes as the BP contribution $\mathcal{Z}_{\text{BP}}$ plus M distinct contributions arising from the placement of a δ_6 excitation on the lattice, as shown in Fig. 7(b). However, in each contribution we freeze only edges connecting the excitation to their BP ground while leaving the other edges in the network free. It follows that each of these contributions has total weight $W(\delta_6)\tilde{\mathcal{Z}}_{M-7}$, where $\tilde{\mathcal{Z}}_{M-7}$ represents the scalar from contracting the remaining tensor network, which now consists of $(M-7)$ complete hexagonal plaquettes. Thus, the scalar $\mathcal{Z}_M$ from the network contraction is approximated as

$$\mathcal{Z}_M \approx 1 + MW(\delta_6)\tilde{\mathcal{Z}}_{M-7}. \quad (\text{B4})$$

In order to make headway with Eq. (B4), another approximation is needed; we now assume that $\tilde{\mathcal{Z}}_{M-7}$ and $\mathcal{Z}_M$ have the same free-energy density f , which implies

$$\tilde{\mathcal{Z}}_{M-7} \approx e^{7f} \mathcal{Z}_M. \quad (\text{B5})$$

Note that this approximation for $\tilde{\mathcal{Z}}_{M-7}$ ignores boundary effects (i.e., resulting from fixing some of the network edges in their BP ground). Substitution of Eq. (B5) into Eq. (B4) gives

$$\mathcal{Z}_M \approx 1 + \mathcal{Z}_M(e^{7f}MW(\delta_6)), \quad (\text{B6})$$

which can be rearranged as

$$\mathcal{Z}_M \approx (1 - e^{7f}MW(\delta_6))^{-1}. \quad (\text{B7})$$

Finally, the (per-site) free-energy density f can be expressed as

$$f \approx \frac{\log(1 - e^{7f}MW(\delta_6))}{2M} \approx -\frac{1}{2}W(\delta_6)e^{7f}, \quad (\text{B8})$$

where the approximation $\log(1 + \epsilon) \approx \epsilon$ is used to obtain the final result. This formula is easily generalized to include arbitrary excitations δ_x ,

$$f \approx \sum_{\{\delta_x\}} -L(\delta_x)W(\delta_x)e^{S(\delta_x)f}, \quad (\text{B9})$$

with $L(\delta_x)$ as the number of locations per lattice site that the excitation can be placed and $S(\delta_x)$ as the number of hexagonal plaquettes that each excitation ‘‘occupies.’’ More precisely, we define $S(\delta_x)$ as the reduction in the number of complete hexagonal plaquettes remaining in the network after the excitation

is placed [see Fig. 7(b)]. Notice that Eq. (B9) is identical to Eq. (B3) apart from the addition of the *suppression factors* $\exp(S(\delta_x)f)$, which serve to approximate the contributions arising from configurations containing multiple excitations.

Finally, we remark that, in general, it will not be possible to solve Eq. (B9) directly for (the loop corrections to) the free-energy density f as it appears on both sides of the equation. However, given that these corrections are assumed to be relatively weak (otherwise the series expansion itself would not be valid), we are able to use an iterative strategy to approximately solve for f . This strategy is as follows: First, we set $f = 0$ on the right-hand side of Eq. (B9) to compute an updated value for f , before then feeding updated f back into the right-hand side and repeating until f is converged. In our benchmark calculations, we observed that this strategy only required a few iterations in order to converge to high precision.

3. Multiexcitation counting

A third strategy to estimate the contribution of loop excitations is simply to explicitly count the products of distinct excitations [72]. Again, we consider a network on a hexagonal lattice of M plaquettes that evaluates to some scalar $\mathcal{Z}_M$. Here, the loop corrections, when expanded out to the degree-10 terms, remain identical to those from the single-excitation approximation in Eq. (B3),

$$\mathcal{Z}_M = MW(\delta_6) + 3MW(\delta_{10}). \quad (\text{B10})$$

However, the corrections from degree-12 excitations, $(\Delta\mathcal{Z})_{12}$, now include the δ_{12} excitations in addition to the product of two degree-6 excitations δ_6 ,

$$(\Delta\mathcal{Z})_{12} = 2MW(\delta_{12}) + \frac{M(M-7)}{2}W(\delta_6)^2, \quad (\text{B11})$$

where the factor $\frac{M(M-7)}{2}$ counts the number of distinct locations that the two degree-6 excitations can be placed. The corrections from the degree-14 excitations remain unchanged,

$$(\Delta\mathcal{Z})_{14} = 12MW(\delta_{14}), \quad (\text{B12})$$

but the degree-16 excitations additionally include the product of δ_6 and δ_{10} excitations,

$$(\Delta\mathcal{Z})_{16} = 18MW(\delta_{16}) + 3M(M-10)W(\delta_6)W(\delta_{10}). \quad (\text{B13})$$

This methodology, which explicitly counts product excitations, can easily be extended out to arbitrary degree, e.g., where the degree-18 expansion would also include products of three δ_6 excitations.

Taking the logarithm of $\mathcal{Z}_M$, the (per-site) free-energy density f is given as

$$\begin{aligned} f = & -\frac{1}{2}W(\delta_6) - \frac{3}{2}W(\delta_{10}) - W(\delta_{12}) \\ & + \frac{7}{4}W(\delta_6)^2 - 6W(\delta_{14}) - 9W(\delta_{16}) \\ & + 15W(\delta_6)W(\delta_{10}) + \dots, \end{aligned} \quad (\text{B14})$$

where it is noted that the terms of order M^2 canceled out in the logarithm expansion.

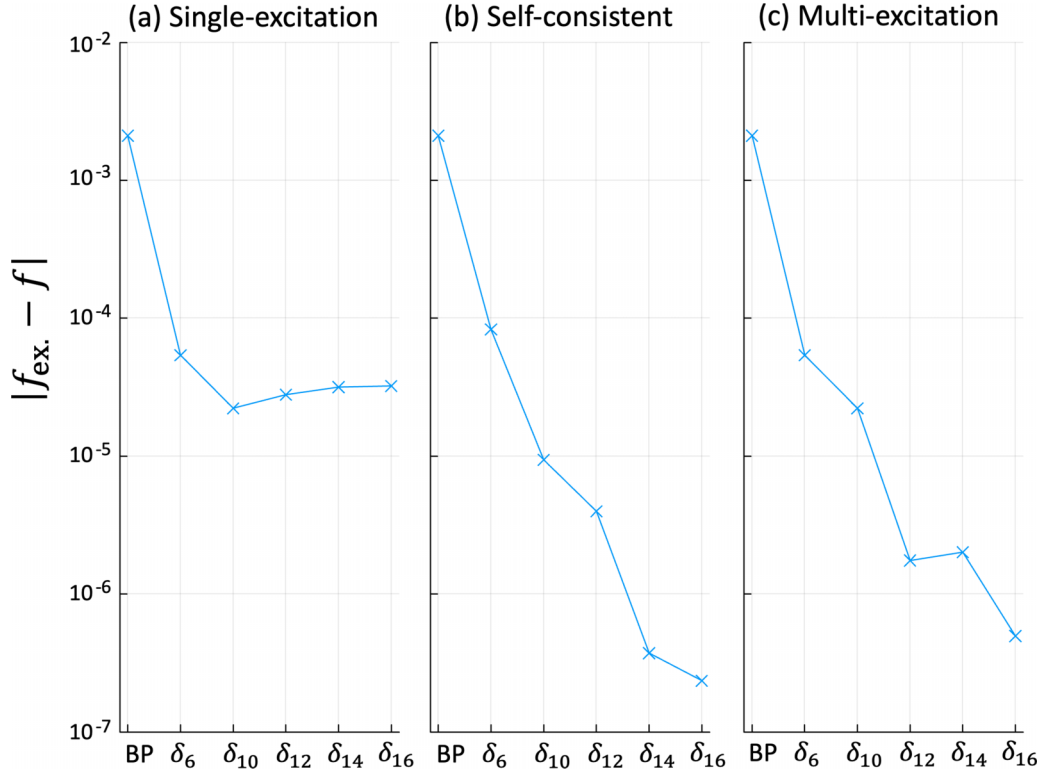


FIG. 8. Results comparing the accuracy of (a) the single-excitation approximation, (b) the self-consistent completion, and (c) the multiexcitation counting strategies for computing the free-energy density f of the hexagonal AKLT model when including all excitations δ_x up to degree x . While the single-excitation approximation saturates at a finite accuracy, the other strategies appear to converge to the exact result in the limit of a large degree expansion.

4. Comparison

We have presented three approaches for using the loop expansion to estimate the free-energy density on an infinite, homogeneous tensor network. A numerical comparison of these approaches is shown in Fig. 8 for loop corrections (expanded out to degree-16) in the hexagonal lattice AKLT model. While the expansion using the single-excitation approximation saturates at a finite accuracy, both the self-consistent completion and the multiexcitation approach appear to converge to the exact result as the degree of the expansion is increased. Similar results were also observed for the AKLT models on square and kagome lattices, and for randomly initialized PEPS tensor networks.

While both the self-consistent completion and the multiexcitation counting are seen to give comparable accuracy, the numerical results of the main text were derived using the former approach. A reason for this is that the self-consistent completion is easier to apply to the case of computing transfer and density matrices, given that the same suppression factors $\exp(S(\delta_x)f)$ can be used as for the free-energy density calculation.

APPENDIX C: KAGOME AND SQUARE LATTICES

In this Appendix, we provide some of the details of the benchmark results provided for iPEPS on kagome and square lattices. An iPEPS for the kagome lattice is shown in Fig. 9(a). Although we could perform the series expansion on the (square norm) of this network directly, we instead found it

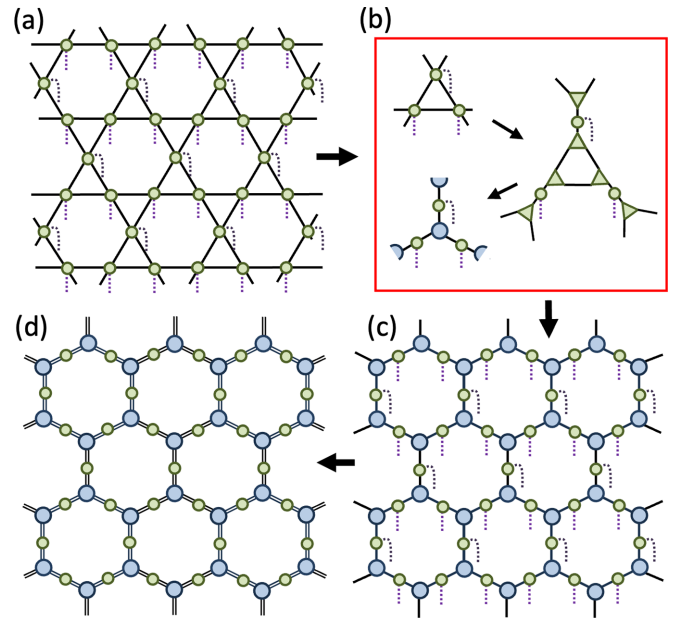


FIG. 9. (a) A kagome lattice iPEPS representing quantum state $|\psi\rangle$. (b) Each PEPS tensor is decomposed into a product of three 3-index tensors via singular value decompositions, and then the tensors surrounding each triangle of the kagome lattice are contracted into a single tensor. (c) A modified PEPS representation of the quantum state $|\psi\rangle$ is obtained. (d) The square norm $\langle\psi|\psi\rangle$ now evaluates to a hexagonal lattice tensor network.

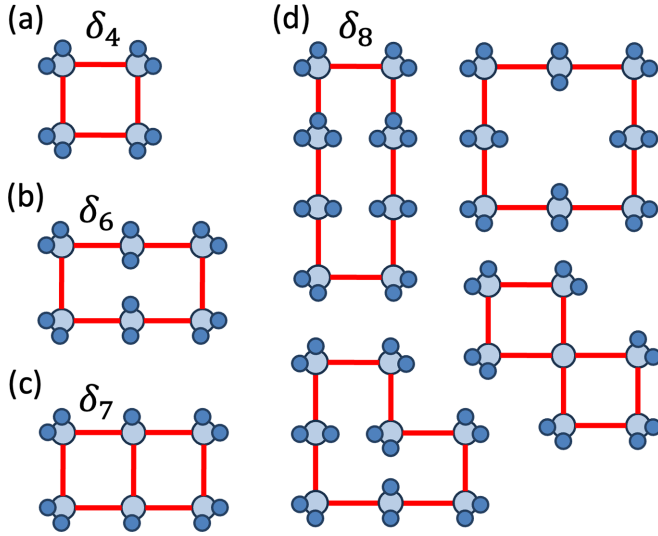


FIG. 10. Loop correction terms for a square-lattice tensor network with a one-site unit cell. (a) The fourth-degree term. (b) One (out of two) of the sixth-degree terms. One (out of two) of the seventh-degree terms. Four (out of nine) of the eighth-degree terms.

preferable to perform some preliminary restructuring. The reasoning for this was to remove the triangular loops from the network, which are otherwise problematic since (1) the weight of the excitations around the triangular loops is likely to be large (since the loops are short) and (2) they generate a multitude of low-degree terms that would need to be included in the expansion. The restructuring is depicted in Fig. 9(b), where each (5-index) PEPS tensor is first decomposed as a product of three 3-index tensors, and then the tensors surrounding each triangle of the kagome lattice are subsequently contracted into a single tensor (located at the center of each kagome lattice triangle). These manipulations result in a PEPS with a hexagonal lattice geometry but where the physical indices are located on tensors set on the edges of the hexagonal lattice, as shown in Fig. 9(c), similar to a so-called projected entangled simplex state [66]. Upon forming the square norm $\langle \psi | \psi \rangle$, a purely hexagonal lattice tensor network is obtained, as shown in Fig. 9(d), which can be treated with the same expansion as described for the hexagonal lattice PEPS. Note that the manipulations performed in Fig. 9(b) may require an increase in PEPS bond dimension from $m \rightarrow m^2$, although in many cases the updated bond dimension could be truncated without incurring significant error. This example of manipulating the kagome lattice into a hexagonal lattice highlights an interesting point: The accuracy of the series expansion can depend on the particulars of the tensor network representation of a quantum state $|\psi\rangle$, rather than solely on the properties of $|\psi\rangle$ itself (e.g., entanglement and correlations).

The benchmark results of Sec. V also explore contraction of PEPS on the square lattice, both for the ground state of the AKLT model and for randomly initialized networks. The square-lattice results were computed similarly to the methods prescribed for the hexagonal lattice, both in terms of evaluating the series expansion and of evaluating the numerically exact results using boundary MPS. Several of the loop correction terms are shown in Fig. 10, where it can be seen that the

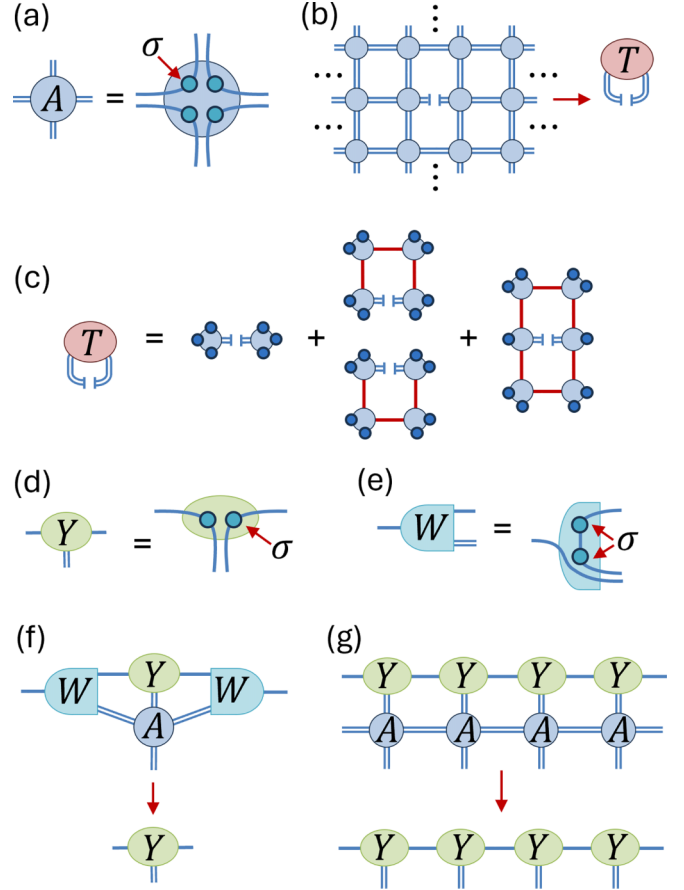


FIG. 11. (a) Definition of a CDL tensor. (b) A transfer matrix T formed by opening a single index from an infinite network of CDL tensors. (c) The transfer matrix T is exactly computed from the BP vacuum plus the single and double loop corrections. (d) Definition tensor Y . (e) Definition of isometry W . (f), (g) A boundary MPS, formed from copies of Y , is seen to be an exact fixed point of the boundary MPS contraction when implemented in conjunction with isometries W .

number of distinct loop excitations of degree- x scales more quickly as a function of x than was seen with the hexagonal lattice.

APPENDIX D: CONTRACTION EFFICIENCY FOR PEPS VERSUS MPS-BASED APPROACHES

In the main text, we demonstrated the use of the loop expansion to approximately contract PEPS tensor networks, although we did not attempt to quantify the efficiency (in terms of accuracy versus computational cost) versus established PEPS contraction methods such as boundary MPS or corner transfer matrix renormalization group. One reason for this was that a proper comparison would require a thorough and dedicated investigation, as the relative accuracy is likely highly variable on the test model under consideration, and also that the cost scaling of MPS-based methods can depend on the finer details of their implementation. Given that our aim was to present the loop expansion as a general method to approximate arbitrary tensor network contractions, with the application to contract PEPS only as a single example, we

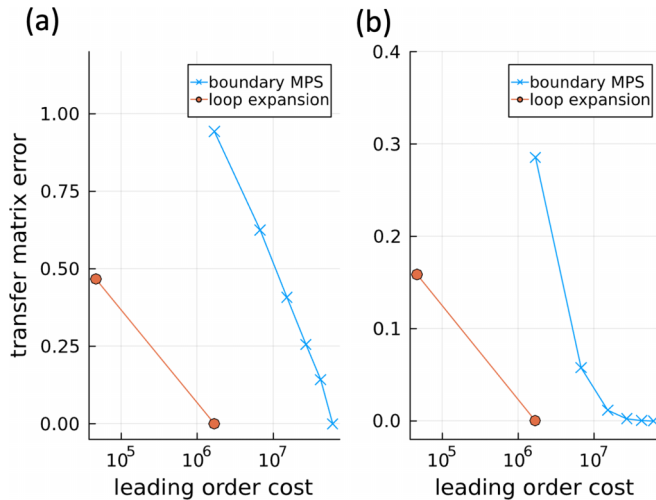


FIG. 12. (a) Leading order computational cost versus accuracy for the error in the transfer matrix, $\|T - T_{\text{exact}}\|$, from a network of CDL tensors of dimension $D = 36$ for $\alpha = 0.1$ and as defined in Eq. (D1). Compared are bond dimensions $m = 1, 2, \dots, 6$ for the boundary MPS versus the loop expansion with either single or double loop corrections. The result demonstrate that the loop expansion is the more efficient approach it produces higher accuracy for lower computational cost. (b) A comparison with $\alpha = 0.4$ shows similar results.

felt such a comparison was beyond the scope of the present work. Nonetheless, in this Appendix we present an example that demonstrates that there exist 2D networks that are demonstrably more efficient to contract with a loop expansion than with previous MPS-based approaches.

Consider a homogeneous square-lattice tensor network of bond dimension D formed from corner double-line (CDL) tensors, themselves composed of corner-sharing entangled pairs mediated by matrices σ of dimension $\sqrt{D}$ as shown in Fig. 11(a). We assume that the σ matrices are diagonal and

contain exponentially decaying weights,

$$\sigma_{ij} = \delta_{ij} \exp(-\alpha i), \quad (\text{D1})$$

with α a positive constant. We focus on the task of computing the transfer matrix T formed by cutting open a single index from the infinite network, as shown in Fig. 11(b). We begin by computing T using the loop series expansion, where computing the series expansion out to the double-plaquette term [see Fig. 11(c)] is found to give the exact result. The cost of evaluating the double-plaquette correction is $O(D^4)$, although this cost reduces to $O(D^3)$ if only the single-plaquette loop corrections are considered, which may still provide a good approximation to T .

We now consider calculating T through a boundary MPS approach [33]. Alternatively, a corner transfer matrix approach [34] could be used, although this is found to be equivalent in efficiency for the network under consideration. It can be seen that an MPS formed from tensors Y as defined in Fig. 11(d), in conjunction with isometries W as defined in Fig. 11(d), produces an exact fixed-point boundary MPS, as shown in Figs. 11(f) and 11(g). The transfer matrix T can then be evaluated from (two copies of) the boundary MPS using standard methodology. If the bond dimension m of the boundary MPS is fixed at $m = \sqrt{D}$, then the contraction for T is exact, although accurate approximations to T can be obtained with smaller m . The leading order cost of contracting for T using the boundary MPS approach is known to scale as $O(D^4 m^2)$.

In Fig. 12, we compare the efficiency of the loop expansion versus the boundary MPS approach (in terms of the error $\|T_{\text{exact}} - T_{\text{approx}}\|$ versus the leading order computational cost) for several values of α in Eq. (D1). It can be seen that the loop expansion is the more efficient approach for contracting the CDL network: It produces more accurate results with less computational effort over the values of α considered. Although the CDL tensor network only has a simple entanglement structure, this example nonetheless demonstrates that the loop expansion may offer a more efficient contraction than traditional MPS-based approaches for contraction of certain 2D tensor networks.

- [1] J. I. Cirac and F. Verstraete, Renormalization and tensor product states in spin chains and lattices, *J. Phys. A: Math. Theor.* **42**, 504004 (2009).
- [2] T. G. Kolda and B. W. Bader, Tensor decompositions and applications, *SIAM Rev.* **51**, 455 (2009).
- [3] G. Evenbly and G. Vidal, Tensor network states and geometry, *J. Stat. Phys.* **145**, 891 (2011).
- [4] R. Orus, A practical introduction to tensor networks: Matrix product states and projected entangled pair states, *Ann. Phys.* **349**, 117 (2014).
- [5] J. C. Bridgeman and C. T. Chubb, Hand-waving and interpretive dance: An introductory course on tensor networks, *J. Phys. A: Math. Theor.* **50**, 223001 (2017).
- [6] S. Montangero, *Introduction to Tensor Network Methods: Numerical Simulations of Low-Dimensional Many-Body Quantum Systems* (Springer, Cham, 2018).
- [7] R. Orus, Tensor networks for complex quantum systems, *Nat. Rev. Phys.* **1**, 538 (2019).
- [8] P. Silvi, F. Tschirsich, M. Gerster, J. Jünemann, D. Jaschke, M. Rizzi, and S. Montangero, The Tensor Networks Anthology Simulation techniques for many-body quantum lattice systems, *SciPost Phys. Lect. Notes*, **8** (2019).
- [9] S.-J. Ran, E. Tirrito, C. Peng, X. Chen, L. Tagliacozzo, G. Su, and M. Lewenstein, *Tensor Network Contractions: Methods and Applications to Quantum Many-Body Systems* (Springer, Cham, 2020).
- [10] J. I. Cirac, D. Pérez-García, N. Schuch, and F. Verstraete, Matrix product states and projected entangled pair states: Concepts, symmetries, theorems, *Rev. Mod. Phys.* **93**, 045003 (2021).
- [11] E. S. Fried, N. P. D. Sawaya, Y. Cao, I. D. Kivlichan, J. Romero, and A. Aspuru-Guzik, qTensor: The quantum tensor contraction handler, *PLoS One* **13**, e0208510 (2018).
- [12] B. Villalonga, S. Boixo, B. Nelson, C. Henze, E. Rieffel, R. Biswas, and S. Mandrà, A flexible high-performance simulator for verifying and benchmarking quantum circuits implemented on real hardware, *npj Quantum Inf.* **5**, 86 (2019).

- [13] R. Schutschi, D. Lykov, and I. Oseledets, Adaptive algorithm for quantum circuit simulation, *Phys. Rev. A* **101**, 042335 (2020).
- [14] Y. Zhou, E. M. Stoudenmire, and X. Waintal, What limits the simulation of quantum computers? *Phys. Rev. X* **10**, 041038 (2020).
- [15] M. Levental, Tensor networks for simulating quantum circuits on FPGAs, [arXiv:2108.06831](https://arxiv.org/abs/2108.06831).
- [16] T. Vincent, L. J. O’Riordan, M. Andrenkov, J. Brown, N. Killoran, H. Qi, and I. Dhand, Jet: Fast quantum circuit simulations with parallel task-based tensor-network contraction, *Quantum* **6**, 709 (2022).
- [17] F. Pan and P. Zhang, Simulation of quantum circuits using the big-batch tensor network method, *Phys. Rev. Lett.* **128**, 030501 (2022).
- [18] F. Pan, K. Chen, and P. Zhang, Solving the sampling problem of the Sycamore quantum circuits, *Phys. Rev. Lett.* **129**, 090502 (2022).
- [19] T. Ayral, T. Louvet, Y. Zhou, C. Lambert, E. M. Stoudenmire, and X. Waintal, Density-matrix renormalization group algorithm for simulating quantum circuits with a finite fidelity, *PRX Quantum* **4**, 020304 (2023).
- [20] J. Tindall, M. Fishman, E. M. Stoudenmire, and D. Sels, Efficient tensor network simulation of IBM’s Eagle kicked Ising experiment, *PRX Quantum* **5**, 010308 (2024).
- [21] T. Begušić, J. Gray, and G. K.-L. Chan, Fast and converged classical simulations of evidence for the utility of quantum computing before fault tolerance, *Sci. Adv.* **10**, eadk4321 (2024).
- [22] M. Fannes, B. Nachtergaele, and R. F. Werner, Finitely correlated states on quantum spin chains, *Commun. Math. Phys.* **144**, 443 (1992).
- [23] S. Ostlund and S. Rommer, Thermodynamic limit of density matrix renormalization, *Phys. Rev. Lett.* **75**, 3537 (1995).
- [24] S. R. White, Density matrix formulation for quantum renormalization groups, *Phys. Rev. Lett.* **69**, 2863 (1992).
- [25] S. R. White, Density-matrix algorithms for quantum renormalization groups, *Phys. Rev. B* **48**, 10345 (1993).
- [26] U. Schollwöck, The density-matrix renormalization group, *Rev. Mod. Phys.* **77**, 259 (2005).
- [27] F. Verstraete and J. I. Cirac, Renormalization algorithms for quantum-many-body systems in two and higher dimensions, [arXiv:cond-mat/0407066](https://arxiv.org/abs/cond-mat/0407066).
- [28] F. Verstraete, J. I. Cirac, and V. Murg, Matrix product states, projected entangled pair states, and variational renormalization group methods for quantum spin systems, *Adv. Phys.* **57**, 143 (2008).
- [29] L. Cincio, J. Dziarmaga, and M. M. Rams, Multiscale entanglement renormalization ansatz in two dimensions: Quantum Ising model, *Phys. Rev. Lett.* **100**, 240603 (2008).
- [30] G. Evenbly and G. Vidal, Entanglement renormalization in two spatial dimensions, *Phys. Rev. Lett.* **102**, 180406 (2009).
- [31] N. Schuch, M. M. Wolf, F. Verstraete, and J. I. Cirac, Computational complexity of projected entangled pair states, *Phys. Rev. Lett.* **98**, 140506 (2007).
- [32] T. Nishino and K. Okunishi, Corner transfer matrix renormalization group method, *J. Phys. Soc. Jpn.* **65**, 891 (1996).
- [33] J. Jordan, R. Orús, G. Vidal, F. Verstraete, and J. I. Cirac, Classical simulation of infinite-size quantum lattice systems in two spatial dimensions, *Phys. Rev. Lett.* **101**, 250602 (2008).
- [34] R. Orús and G. Vidal, Simulation of two-dimensional quantum systems on an infinite lattice revisited: Corner transfer matrix for tensor contraction, *Phys. Rev. B* **80**, 094403 (2009).
- [35] P. Corboz, T. M. Rice, and M. Troyer, Competing states in the t - J model: Uniform d -wave state versus stripe state, *Phys. Rev. Lett.* **113**, 046402 (2014).
- [36] M. Lubasch, J. I. Cirac, and M.-C. Bañuls, Unifying projected entangled pair state contractions, *New J. Phys.* **16**, 033014 (2014).
- [37] M. Lubasch, J. I. Cirac, and M.-C. Bañuls, Algorithms for finite projected entangled pair states, *Phys. Rev. B* **90**, 064425 (2014).
- [38] M. T. Fishman, L. Vanderstraeten, V. Zauner-Stauber, J. Haegeman, and F. Verstraete, Faster methods for contracting infinite two-dimensional tensor networks, *Phys. Rev. B* **98**, 235148 (2018).
- [39] H. A. Bethe, Statistical theory of superlattices, *Proc. R. Soc. London A* **150**, 552 (1935).
- [40] J. Pearl, Reverend bayes on inference engines a distributed hierarchical approach, in *Proceedings of the Second AAAI Conference on Artificial Intelligence*, AAAI’82 (AAAI Press, Pittsburgh, Pennsylvania, 1982), pp. 133–136.
- [41] J. Pearl, Fusion, propagation, and structuring in belief networks, *Artif. Intell.* **29**, 241 (1986).
- [42] M. Mézard, G. Parisi, and R. Zecchina, Analytic and algorithmic solution of random satisfiability problems, *Science* **297**, 812 (2002).
- [43] J. S. Yedidia, W. Freeman, and Y. Weiss, Generalized belief propagation, in *Advances in Neural Information Processing Systems*, edited by T. Leen, T. Dietterich, and V. Tresp (MIT Press, Cambridge, MA, USA, 2000), Vol. 13, pp. 668–674.
- [44] J. S. Yedidia, W. T. Freeman, and Y. Weiss, Understanding belief propagation and its generalizations, in *Exploring Artificial Intelligence in the New Millennium*, edited by G. Lakemeyer and B. Nebel (Morgan Kaufmann, San Francisco, CA, 2003), pp. 239–269.
- [45] J. S. Yedidia, W. T. Freeman, and Y. Weiss, Constructing free-energy approximations and generalized belief propagation algorithms, *IEEE Trans. Inf. Theory* **51**, 2282 (2005).
- [46] M. Mézard, G. Parisi, and M. A. Virasoro, *Spin Glass Theory and Beyond: An Introduction to the Replica Method and its Applications* (World Scientific, Singapore, 1987), Vol. 9.
- [47] S. Yoon, A. V. Goltsev, S. N. Dorogovtsev, and J. F. F. Mendes, Belief-propagation algorithm and the Ising model on networks with arbitrary distributions of motifs, *Phys. Rev. E* **84**, 041144 (2011).
- [48] B. Karrer, M. E. J. Newman, and L. Zdeborová, Percolation on sparse networks, *Phys. Rev. Lett.* **113**, 208702 (2014).
- [49] M. S. Leifer and D. Poulin, Quantum graphical models and belief propagation, *Ann. Phys.* **323**, 1899 (2008).
- [50] D. Poulin and E. Bilgin, Belief propagation algorithm for computing correlation functions in finite-temperature quantum many-body systems on loopy graphs, *Phys. Rev. A* **77**, 052318 (2008).
- [51] A. Wrigley, W. S. Lee, and N. Ye, Tensor belief propagation, in *Proceedings of the 34th International Conference on Machine Learning*, edited by D. Precup and Y. W. Teh (PMLR, 2017), Vol. 70, pp. 3771–3779.
- [52] E. Robeva and A. Seigal, Duality of graphical models and tensor networks, *Inf. Inference* **8**, 273 (2018).

- [53] R. Alkabetz and I. Arad, Tensor networks contraction and the belief propagation algorithm, *Phys. Rev. Res.* **3**, 023073 (2021).
- [54] S. Sahu and B. Swingle, Efficient tensor network simulation of quantum many-body physics on sparse graphs, [arXiv:2206.04701](#).
- [55] C. Guo, D. Poletti, and I. Arad, Block belief propagation algorithm for two-dimensional tensor networks, *Phys. Rev. B* **108**, 125111 (2023).
- [56] N. Pancotti and J. Gray, One-step replica symmetry breaking in the language of tensor networks, [arXiv:2306.15004](#).
- [57] Y. Wang, Y. E. Zhang, F. Pan, and P. Zhang, Tensor network message passing, *Phys. Rev. Lett.* **132**, 117401 (2024).
- [58] J. Tindall and M. Fishman, Gauging tensor networks with belief propagation, *SciPost Phys.* **15**, 222 (2023).
- [59] Y. Kim, A. Eddins, S. Anand, K. X. Wei, E. Van Den Berg, S. Rosenblatt, H. Nayfeh, Y. Wu, M. Zaletel, K. Temme, and A. Kandala, Evidence for the utility of quantum computing before fault tolerance, *Nature (London)* **618**, 500 (2023).
- [60] M. P. Zaletel and F. Pollmann, Isometric tensor network states in two dimensions, *Phys. Rev. Lett.* **124**, 037201 (2020).
- [61] M. Chertkov and V. Y. Chernyak, Loop calculus in statistical physics and information science, *Phys. Rev. E* **73**, 065102(R), (2006).
- [62] M. Chertkov and V. Chernyak, Loop series for discrete statistical models on graphs, *J. Stat. Mech.* (2006) P06009.
- [63] V. Gómez, J. Mooij, and H. Kappen, Truncating the loop series expansion for belief propagation, *J. Mach. Learn. Res.* **8**, 1987 (2007).
- [64] I. Affleck, T. Kennedy, E. H. Lieb, and H. Tasaki, Rigorous results on valence-bond ground states in antiferromagnets, *Phys. Rev. Lett.* **59**, 799 (1987).
- [65] I. Affleck, T. Kennedy, E. H. Lieb, and H. Tasaki, Valence bond ground states in isotropic quantum antiferromagnets, *Commun. Math. Phys.* **115**, 477 (1988).
- [66] Z. Y. Xie, J. Chen, J. F. Yu, X. Kong, B. Normand, and T. Xiang, Tensor renormalization of quantum many-body systems using projected entangled simplex states, *Phys. Rev. X* **4**, 011025 (2014).
- [67] M. Levin and C. P. Nave, Tensor renormalization group approach to two-dimensional classical lattice models, *Phys. Rev. Lett.* **99**, 120601 (2007).
- [68] G. Evenbly and G. Vidal, Tensor network renormalization, *Phys. Rev. Lett.* **115**, 180405 (2015).
- [69] S. Yang, Z.-C. Gu, and X.-G. Wen, Loop optimization for tensor network renormalization, *Phys. Rev. Lett.* **118**, 110504 (2017).
- [70] G. Evenbly, Gauge fixing, canonical forms, and optimal truncations in tensor networks with closed loops, *Phys. Rev. B* **98**, 085155 (2018).
- [71] M. Hauru, C. Delcamp, and S. Mizera, Renormalization of tensor networks using graph-independent local truncations, *Phys. Rev. B* **97**, 045111 (2018).
- [72] I. Lukin and A. Sotnikov (private communication).