# Variational quantum optimization with multibasis encodings

Taylor L. Patti,[1,2,*] Jean Kossaifi,[2] Anima Anandkumar,[3,2] and Susanne F. Yelin[1]
[1]*Department of Physics, Harvard University, Cambridge, Massachusetts 02138, USA*
[2]*NVIDIA, Santa Clara, California 95051, USA*
[3]*Department of Computing + Mathematical Sciences (CMS), California Institute of Technology (Caltech), Pasadena, California 91125, USA*

(Received 6 October 2021; revised 26 January 2022; accepted 4 August 2022; published 22 August 2022)

Despite extensive research efforts, few quantum algorithms for classical optimization demonstrate a realizable quantum advantage. The utility of many quantum algorithms is limited by high requisite circuit depth and nonconvex optimization landscapes. We tackle these challenges by introducing a variational quantum algorithm that benefits from two innovations: multibasis graph encodings using single-qubit expectation values and nonlinear activation functions. Our technique results in increased observed optimization performance and a factor-of-two reduction in requisite qubits. While the classical simulation of many qubits with traditional quantum formalism is impossible due to its exponential scaling, we mitigate this limitation with exact circuit representations using factorized tensor rings. In particular, the shallow circuits permitted by our technique, combined with efficient factorized tensor-based simulation, enable us to successfully optimize the MaxCut of the 512-vertex DIMACS library graphs on a single GPU. By improving the performance of quantum optimization algorithms while requiring fewer quantum resources and utilizing shallower, more error-resistant circuits, we offer tangible progress for variational quantum optimization.

DOI: 10.1103/PhysRevResearch.4.033142

## I. INTRODUCTION

NP-hard optimization problems, such as Traveling Salesman and MaxCut, are central to a wide array of fields, such as operational research, engineering, and network design [1]. Despite the classical nature of these problems, there is immense interest in identifying variational quantum algorithms (VQAs) that can solve them faster or more precisely than any classical method, a concept known as quantum advantage [2–5].

One popular VQA is the variational quantum eigensolver (VQE), where energy minimization yields the ground state of a problem-encoded Hamiltonian through variational update (e.g., gradient descent minimization) of the quantum circuit parameters [6–8]. The quantum approximate optimization algorithm (QAOA) is a related protocol in which unitary evolutions using both an initial and a problem-encoded Hamiltonian are alternated in order to find a solution-encoded ground state [9–13]. In addition to VQE and QAOA, numerous other VQAs with distinct encoding strategies have been considered in [14–16]. The optimization landscape of VQAs can be quite favorable. For instance, the optimization landscape of VQE for NP-hard optimization problems can be made convex [8], such that it is absent of local minima and the algorithm is likely to obtain the ground truth. While this deterministic finding of solutions is superior to the performance of polynomial complexity classical algorithms (e.g., Goemans-Williamson [17–19]), we emphasize that this guarantee requires between polynomially and exponentially many gates in the number of qubits $n$, as well as a specific family of *Ansätze* [8]. Such circuit depths limit the algorithms' potential to demonstrate quantum advantage, rendering them not only computationally inefficient, but also highly susceptible to quantum noise [10,11,20] and barren plateaus [21–26]. The effectiveness of local VQAs, or algorithms where the quantum state update is limited to only explicitly connected degrees of freedom (e.g., graph vertices connected by an edge), is even more limited. Local VQAs have demonstrably poorer performance than classical methods, e.g., the Goemans-Williamson algorithm, on particularly challenging and large graph instances [27–29]. However, there is evidence that these limitations may resolve with substantial circuit depths [30,31].

The difficulty of classically simulating large-scale quantum circuits is a central challenge to algorithm development. This is because the traditional mathematical formalism of quantum mechanics automatically represents the full Hilbert space and thus scales exponentially in the number of qubits, $n$, with matrix operators of size $2^{2n}$ operating on state vectors of size $2^n$. When a quantum system does not occupy the full Hilbert space, these intractable dimensions for quantum network simulation can be remediated by employing a factorized tensor formalism [32]. While many varieties of decomposed tensors exist, tensor rings have proven particularly popular in the quantum sciences due to their modularity and rank structure, which have close parallels to quantum

*taylorpatti@g.harvard.edu

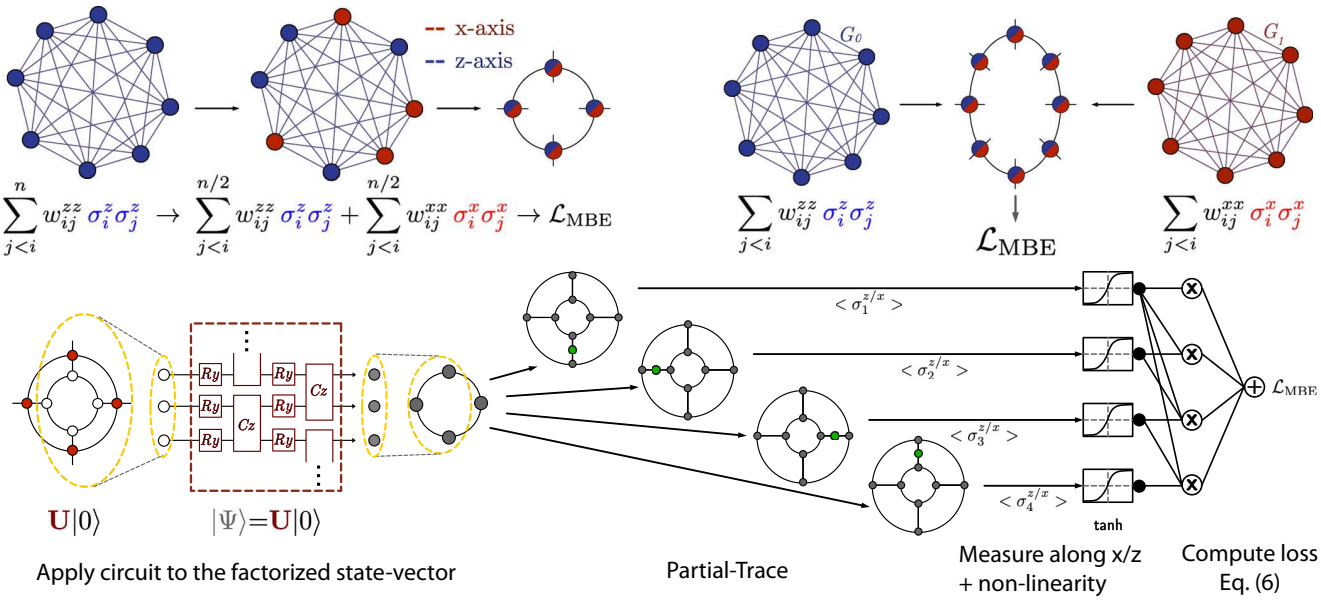


FIG. 1. Left: Multibasis encoding (MBE) of a graph. An $n$-vertex graph (blue) is represented as an Ising model. We reassign $n/2$ vertices from $\sigma^z$ (blue) to $\sigma^x$ (red) operators, allowing us to map the graph to just $n/2$ qubits (here, a nearest-neighbor connected, blue/red tensor ring). The MaxCut is obtained by optimizing this state via single-qubit measurements. Although tensor cores in the tensor ring formalism only share bonds with neighboring cores, they effectively solve MaxCut graphs with diverse edge distributions. Right: Multibasis encoding (MBE) with two distinct $n$-qubit graphs. Each graph is mapped to the classical Ising model, with $G_0$ (blue) encoded along the $z$ basis (as is traditional) and $G_1$ (red) utilizing the $x$ basis, resulting in an $n$-qubit quantum state (blue/red). This encoding is similar to MBE with a single graph, except that the $x$ and $z$ bases *independently* encode two separate graphs and thus no cross terms between the $z$ and $x$ bases are required. Bottom: Overview of our multibasis encoding approach for an $n$-vertex graph. To find MaxCut($G$) variationally, the ceil($n/2$)-qubit (here, 4-qubit) null input state $|\mathbf{0}\rangle$ (an MPS) is evolved under a parameterized quantum circuit $U$ (an MPO), producing an output state $|\psi\rangle$. $U$ encodes a circuit of depth $L$ (here, $L = 4$; red box). After partial tracing, each qubit is measured independently in the $x$ and $z$ Pauli bases and a nonlinear activation function (here, tanh) is applied. The MBE loss $\mathcal{L}_{\mathrm{MBE}}$ [Eq. (6)] is minimized via gradient descent. The $x$ and $z$ Pauli spin values of the resulting wave function $|\psi\rangle$ are then rounded to $\pm 1$ [see Eq. (8)]. This rounding makes states near the the global minimum of $\mathcal{L}_{\mathrm{MBE}}$ correspond to $|\psi\rangle = |\psi_g\rangle$, the MaxCut ground truth solution.

entanglement. In the tensor ring formalism, both quantum states and quantum operators are represented in factorized form by matrix product states (MPSs) and matrix product operators (MPOs), respectively [33–35]. However, tensor formalism is often unsuitable for high-depth and connectivity regimes, which are most commonly used in quantum optimization, since tensor rings quickly become prohibitively large (high-rank/bond-dimension) when simulating deep or complicated circuits [36]. Moreover, they are limited to only nearest-neighbor interactions.

Due in part to these limitations, no simulation of more than ∼100 qubits has demonstrated quantum optimization rivaling that of classical methods, except on graphs with simple edge distributions, e.g., regular graphs with edges only between a small subset of nearest-neighbor vertices, such as toroid graphs. (Here, "nearest-neighbor" vertices are defined as those vertices closest in vertex numbers. Although the vertex number assignment can, in general, be arbitrary, the limitation of edges between such vertices in certain graph types, e.g., regular graphs, can impose some graph structure.) This is also the case in [37], where exact representations of general tensor architectures with optimal contraction schemes are used. Other large-scale implementations have focused on more restrictive problems. For instance, QAOA MaxCut optimization with up to 210 qubits has been achieved for 3-regular graphs with randomly distributed edges [38]. QAOA MaxCut optimization has also been implemented with several-thousand qubits when optimizing over only a subset of a graphs total edges, a method which did not yield high average performance [39]. Moreover, large-scale optimization on NP-hard problems (e.g., MaxCut) has not been explored using VQE.

*Quantum computing contribution.* This manuscript introduces a different family of quantum algorithms. We benchmark our techniques against VQE implementations using the same "circuit *Ansatz*," a term that we use to reference the structure of a quantum circuit, including its gate types, gate order, and circuit depth. The gate type and gate order used for all circuits in this work is provided in Figs. 1 (bottom) and 2, and the circuit depth is discussed in reference to each numerical simulation. When the same circuit *Ansatz* is used, our algorithm outperforms VQE on a variety of optimization tasks while requiring fewer qubits. Although aspects of our algorithm may make it easier to simulate classically, we emphasize that the algorithm is a quantum protocol that is classically intractable for large-scale problems. In particular:

(i) We propose multibasis encodings (MBEs), a quantum optimization algorithm that introduces additional constraints (regularization), reducing its susceptibility to local minima in the training landscape. We demonstrate our algorithm on the MaxCut problem and empirically find that MBE outperforms VQE by 5–7% when using the same circuit *Ansatz*.

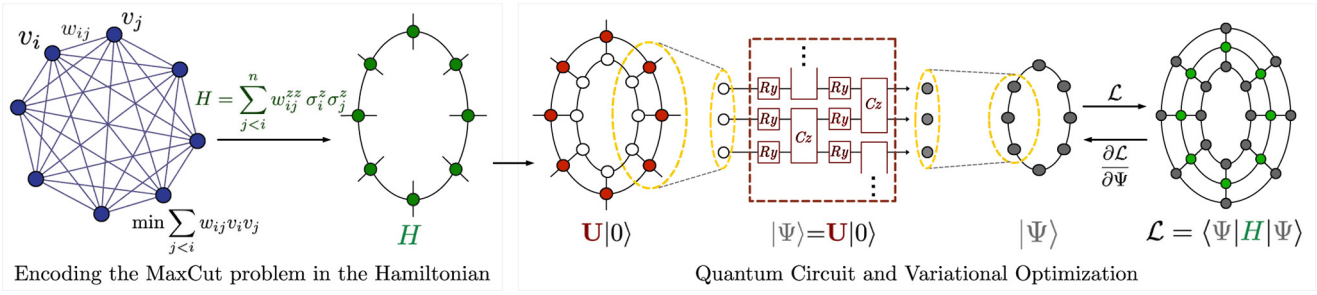


FIG. 2. Overview of traditional MaxCut encoding and VQE using tensor ring factorizations, which are tensor train networks with periodic boundary conditions. Left: A graph $G$ with $n$ vertices $v_i$, $v_j$ and weights $w_{ij}$ is mapped into an $n$-qubit Hamiltonian $H$ in MPO form. The MPS ground state $|\psi_g\rangle$ of $H$ encodes the solution to MaxCut($G$). Right: To find MaxCut($G$) variationally, the null input state $|\mathbf{0}\rangle$ (an MPS) is evolved under a parameterized quantum circuit $U$ (an MPO), producing an output state $|\psi\rangle$. $U$ encodes a circuit of depth $L$ (here, $L = 4$; red box) in this manuscript's layer (block) pattern: one layer (block) of single-qubit $y$-axis rotations $R_y$ followed by a layer of CONTROL-Z gates which alternate between even and odd qubits. The energy expectation value $\mathcal{L} = E$ is minimized via gradient descent. The global minimum of $\mathcal{L}$ corresponds to $|\psi\rangle = |\psi_g\rangle$.

(ii) By doubling the amount of optimization features encoded into a single qubit, MBEs halve the number of qubits required for a given optimization task. This is a valuable asset for a developing field which has invested millions of dollars and spent multiple decades to achieve ∼50-qubit registers and where additional coherence limitations emerge at scale [40].

(iii) By combining our MBEs with nonlinear activation functions, we use relatively shallow quantum circuits to solve MaxCut optimization problems for various graphs, including graphs with random edge distributions. As variational quantum heuristics tend to perform better with increasing circuit expressivity (e.g., circuit depth), the relatively low requisite circuit depths of MBE may indicate that their use of classical nonlinearities (which are used to increase the expressivity of classical networks) is a trade-off of quantum for classical expressivity. Furthermore, sampling ∼5 initializations of our MBE experiments on shallow circuits (depth $L = 7$ for 100-vertex graphs, such that $L$ is approximately logarithmic in the number of vertices) leads to optimal cut convergence with near unit probability. This shallow-circuit, multishot procedure is both more coherent and time efficient than deterministic convergence with deep circuits, which requires up to an exponential number of parameters.

*Large-scale simulation contribution.* This work utilizes tensor networks, developing software in order to simulate practical quantum algorithms at an unprecedented scale. Specifically:

(i) MBE's ability to operate with relatively shallow circuits enables us to use tensor networks with lower rank (bond dimension). As the rank of a tensor structure determines the time and memory complexity of its contraction, we can simulate high-accuracy implementations of MBE at considerable scales.

(ii) We develop TENSORLY-QUANTUM [41,42], a software package for simulating efficient quantum circuits with decomposed tensors on CPU and GPU. TENSORLY-QUANTUM is based on the TENSORLYsoftware family [43].

(iii) Using TENSORLY-QUANTUM on a single NVIDIA A100 GPU, we simulate solving a 512-vertex MaxCut problem using MBE, which in our experiments demonstrates superior performance over VQE with an identical circuit *Ansatz*. This sets a record for the large-scale simulation of a successful quantum optimization algorithm.

By introducing a variety of algorithms that improve the observed optimization performance, require fewer qubits, and operate on shallower, more error-resistant circuits, we offer tools to increase the utility of variational quantum algorithms.

### A. MaxCut optimization problems

The Maximum Cut problem, most commonly referred to as MaxCut, is a partitioning problem on unidirected graphs $G = (V, A)$, where $V$ is a set of vertices (blue orbs in Fig. 2, left) connected by edges $A$ (black lines connecting orbs) [44]. The objective is to optimally assign all vertices $v_i$, $v_j \in \{-1, 1\}$, so as to maximize the edge weights $w_{ij} \in A$, where any such assignment is referred to as a "cut." In this work, we will consider a generalized form of the problem known as *weighted* MaxCut, in which $w_{ij}$ take arbitrary real values.

Two formulations of MaxCut exist: the NP-complete decision problem and the NP-hard optimization problem [45]. The former seeks to determine if a cut of size $c$ or greater exists for a given graph $G$, whereas the latter attempts to identify the largest cut of $G$ possible. Here we focus on the more general optimization problem formulation, the ground truth, which we denote as MaxCut($G$). It is common practice to express the objective function in its binary quadratic form [44]:

$$\text{maximize} \quad \frac{1}{2} \sum_{j<i} w_{ij}(1 - v_i v_j). \tag{1}$$

### B. VQE framework and tensor network formalism

To find the MaxCut of a given graph on a quantum computer, it is convenient to minimize the equivalent summation, $\sum_{j<i} w_{ij} v_i v_j$. For a graph with $n$ vertices $v_i$, this reduces the problem to finding the $n$-qubit wave function $|\psi\rangle$ that minimizes the energy expectation value $E = \langle\psi|H|\psi\rangle$ of the classical Ising model Hamiltonian,

$$H = \sum_{j<i}^{n} w_{ij}^{zz} \sigma_i^z \sigma_j^z. \tag{2}$$

$H$ is obtained by substituting vertices $v_i$ for the Pauli-$Z$ spin operators $\sigma_i^z$, as depicted in Fig. 2, and $w_{ij}^{zz} = w_{ij}$ is a relabeling to specify the $zz$-spin interactions. As $H$ contains only terms in the $z$ basis, its eigenvectors are classical (zero-entanglement product states), such that $|\psi_i\rangle = \bigotimes_s |s\rangle$, where $|s\rangle \in \{|0\rangle, |1\rangle\}$. Here we denote the lowest eigenvalue or "ground-state" solution as $|\psi_g\rangle$, the qubits of which form a bijection with the optimal $v_i$ of MaxCut($G$). As Eq. (2) has $\mathbb{Z}_2$ symmetry, $|\psi_g\rangle$ is degenerate with the state $X^{\otimes n}|\psi_g\rangle$.

Figure 2 (right) depicts the VQE framework [6–8]. Equation (1) is optimized by defining the loss function $\mathcal{L} = E$ and varying the parameters $\hat{\theta}$ of a quantum circuit with unitary $U(\hat{\theta})$, which acts on the input quantum state (Fig. 2, right). Without loss of generality, we define the input state as the $n$-qubit zero state $|\mathbf{0}\rangle = \bigotimes_n |0\rangle$, such that

$$|\psi\rangle = U(\hat{\theta})|\mathbf{0}\rangle. \tag{3}$$

We decompose this unitary matrix $U$ as $\Lambda$ subunitaries $U(\hat{\theta}) = \prod_k^{\Lambda} U_k(\hat{\theta}_k)$, where $\hat{\theta}_k$ is the corresponding subset of $\hat{\theta}$ and $U_k(\hat{\theta}_k) = \prod_{j=1}^{n} \exp(-i\hat{\theta}_j W_j) M_k$ for generic Hermitian operators $W_j$ and unitary matrices $M_k$. Thus, the gradient $g_l(\hat{O}) = \frac{\partial\langle\hat{O}\rangle}{\partial\theta_l}$ of operator $\hat{O}$ with respect to any parameter $\theta_l \in \hat{\theta}$ is

$$g_l(\hat{O}) = i\langle\mathbf{0}|U_R^\dagger[W_l, U_L^\dagger \hat{O} U_L]U_R|\mathbf{0}\rangle, \tag{4}$$

where $U_L$ and $U_R$ are the compositions of unitaries $U_k$ with $k \geqslant l$ and $k < l$, respectively. Rather than using circuits with extensive connectivity, we instead focus on one-dimensional (1D) tensor ring circuits of $n$ qubits. In particular, tensor rings have periodic boundary conditions such that qubit $n-1$ is connected to qubit 0. Such nearest-neighbor connectivity makes the circuit amenable to both near-term quantum hardware [10,12] and simulation via decomposed tensors. We accomplish this simulation with TENSORLY-QUANTUM [41,42]. A nascent and expanding public software package, TENSORLY-QUANTUM strives to leverage the structure of decomposed tensors in order to simulate quantum machine learning in the most efficient, nonapproximate manner possible. While tensor-ring-based tensor networks are typically used for approximate inference and obtained by applying tensor decomposition to dense state vectors and operators, we build a low-rank but exact factorized representation of the simulated quantum circuits. When judiciously constructed, tensor simulations yield a low-rank quantum formalism that permits enormous compression of state and operator spaces. Although in the quantum sciences tensor methods are most frequently associated with state approximations and truncations, such as the density matrix renormalization group [46], here we advocate for their use in exact quantum simulation. Similarly, due to their nearest-neighbor connectivity, tensor ring factorizations in quantum computing have traditionally been employed for locally connected optimization problems, such as 3-regular MaxCut [47]; however, here we emphasize their utility for general purpose optimization tasks.

To analyze VQE with tensor formalism, the Hamiltonian of Eq. (2) is represented as an MPO $H^{\{\beta,\gamma\}}$, with physical indices $\beta$ and $\gamma$. The energy $\mathcal{L} = E$ is then calculated with a single large contraction (Fig. 2, right),

$$E = \sum_{\{\beta,\gamma,\delta,\epsilon\}} \Psi^{\{\beta\}} U^{\{\beta,\gamma\}} H^{\{\gamma,\delta\}} U^{\{\delta,\epsilon\}} \Psi^{\{\epsilon\}}, \tag{5}$$

where

$$\Psi^{\{\beta\}} = \Psi^{\beta_0,\ldots,\beta_{m-1}} = \sum_{\{\alpha\}} \psi^{\beta_0}_{\alpha_0\alpha_1}, \ldots, \psi^{\beta_{m-1}}_{\alpha_{m-1}\alpha_0}$$

is an $n$-qubit MPS of $m$ cores and

$$U^{\{\beta,\gamma\}} = U^{\beta_0,\gamma_0,\ldots,\beta_{m-1},\gamma_{m-1}} = \sum_{\{\alpha\}} u^{\beta_0,\gamma_0}_{\alpha_0,\alpha_1}, \ldots, u^{\beta_{m-1},\gamma_{m-1}}_{\alpha_{m-1},\alpha_0}$$

is the corresponding MPO unitary.

As we work in the absence of quantum noise, states $|\psi\rangle$ display time-reversal symmetry and can be fully expressed with real numbers [48]. We thus restrict our rotations to those of the Pauli-$Y$ generator $\sigma^y$ and implement a simple, repeating subunitary pattern of two layers, also known as blocks. The pattern is illustrated in Fig. 2 (right): a row of parameterized single-qubit rotations $R_y(\theta)$ ($W = \sigma^y$) is followed by a row of CONTROL-Z (CZ) gates, with the latter alternating control between even and odd qubits. As each single-qubit rotation is a $2 \times 2$ dense matrix and each two-qubit CONTROL-Z gate is a rank-2 MPO of two eight-element cores, the memory requirements of the uncontracted circuit representation scale only linearly in both $n$ and $L$. This is an up-to-exponential reduction in classical memory resources compared to circuits described in traditional quantum formalism. Likewise, a factorized representation of the input state $|\mathbf{0}\rangle$ in tensor ring form requires exponentially fewer terms, as it is represented by a rank-$\prod_{i=0}^{n} 1$ MPS with just $n$ two-element cores.

## II. MULTIBASIS ENCODING (MBE)

*Algorithm.* MBE encodes variables into two (or more) qubit axes and evaluates the multivariable interactions in the loss function as the product of single-qubit terms. MBE for weighted graphs is depicted in Fig. 1 (left). An $n$-vertex graph $G$ is expressed similarly to the Ising model Hamiltonian in Eq. (2), save that only the first ceil($n/2$) vertices are mapped to the $z$ axis (blue), while the second floor($n/2$) vertices are mapped to the $x$ axis (red), thus enabling $n$ vertices to be encoded into only ceil($n/2$) qubits [Fig. 1 (bottom)]. If $n$ is odd, then the $x$ axis of the $n$th qubit is unneeded, as it is absent from the loss function and can go unmeasured. In this work, the vertex order is taken to be that of the publicly available BiqMac and DIMACS graph data files [49,50], which is arbitrary due to the random or symmetric distribution of graph edges. In future work, more sophisticated vertex partitionings can be explored, such as mappings that reflect graph topology. MBE halves the number of qubits required for a given optimization, providing a meaningful decrease in quantum hardware overhead.

In order to optimize both axes as independent vertices, we must make several alterations to standard VQE. To begin, the Hamiltonian formed by the product of the MBE encoding operators would be an unsuitable loss function, as the quantum ground state it would encode would not correspond to the classical MaxCut of $G$. We instead focus on the products of single-qubit measurements $\langle\sigma_i^x\rangle$ and $\langle\sigma_i^z\rangle$, such that $\sigma_i^x$ and $\sigma_i^z$

operators are simultaneously optimized. This yields the MBE loss function

$$\mathcal{L}_{\mathrm{MBE}} = \sum_{j<i}^{n/2} w_{ij}^{zz} \tanh\left(\langle\sigma_i^z\rangle\right) \tanh\left(\langle\sigma_j^z\rangle\right) + \sum_{j<i}^{n/2} w_{ij}^{xx} \tanh\left(\langle\sigma_i^x\rangle\right) \tanh\left(\langle\sigma_j^x\rangle\right) + \sum_{i,j}^{n/2} w_{ij}^{zx} \tanh\left(\langle\sigma_i^z\rangle\right) \tanh\left(\langle\sigma_j^x\rangle\right), \quad (6)$$

where $\tanh(x)$ is trivially implemented on the classical computer controlling gradient descent. This procedure is graphically depicted in Fig. 1 (bottom) for the four-qubit encoding of an $n = 8$ vertex graph. For a detailed example, let us consider an $n = 4$ vertex graph with three edges: one joining vertices 1 and 2 with weight $\omega_{12}$, one joining vertices 3 and 4 with weight $\omega_{34}$, and one joining vertices 1 and 3 with weight $\omega_{13}$. The MaxCut of this graph can be optimized using VQE by encoding it into the four-qubit Ising model Hamiltonian,

$$H = \omega_{12}\sigma_1^z\sigma_2^z + \omega_{34}\sigma_3^z\sigma_4^z + \omega_{13}\sigma_1^z\sigma_3^z,$$

and extremizing for the ground state. We can instead use MBE with only two qubits to optimize for the MaxCut of this graph. Specifically, we minimize the two-qubit MBE loss function,

$$\mathcal{L}_{\mathrm{MBE}} = w_{12}^{zz} \tanh\left(\langle\sigma_1^z\rangle\right) \tanh\left(\langle\sigma_2^z\rangle\right) + w_{12}^{xx} \tanh\left(\langle\sigma_1^x\rangle\right) \tanh\left(\langle\sigma_2^x\rangle\right) + w_{11}^{zx} \tanh\left(\langle\sigma_1^z\rangle\right) \tanh\left(\langle\sigma_1^x\rangle\right),$$

and round our results, as detailed in Eq. (8). We again emphasize that as Eq. (6) is comprised of distinct Pauli strings that are independently measured on separate circuit preparations (i.e., one set of preparations/measurements to estimate the $z$-axis expectation values and one for the $x$ axis), the uncertainty principle is not violated for $w_{ij}^{zx}$ with $j = i$. The projection of high-dimensional quantum data into a lower-dimensional representation has also been explored in [51,52], and the utilization of two, rather than a single, quantum bases has proven useful in other quantum machine learning algorithms [53].

The inclusion of the nonlinear activation function $\tanh(x)$ disincentives the extremization of one basis at the expense of another, which could otherwise occur because the optimal values of both $\sigma_i^x$ and $\sigma_i^z$ cannot be linearly encoded by a single quantum state due to the normalization condition of the Bloch sphere of each qubit $i$:

$$\langle\sigma_i^z\rangle^2 + \langle\sigma_i^x\rangle^2 \leqslant 1, \quad (7)$$

where equality holds for real-valued pure states. As the gradient of $\tanh(x)$ reduces near the $\pm 1$ poles (inset Fig. 3, top), full optimization of one axis at the expense of the other is discouraged. The optimal cuts are deduced by a rounding procedure (detailed below), which assigns integer vertex values, but does not affect the optimization process [loss function, parameter update, or the normalization condition of Eq. (7)]. In this manner, MBE is a dual-axis quantum analog to linear programming relaxations [54]. Furthermore, the normalization constraint of Eq. (7) means that $\mathcal{L}_{\mathrm{MBE}}$ can only ever partially descend into local minima and is better equipped to escape their regions of attraction. The robustness of MBE against local minima is likely partially attributable to its use of global optimization [27,28], such as the global optimization of single-qubit states and the interdependence between the generally unrelated $z$- and $x$-axis encoded vertices. We note that we have, for simplicity, neglected both external fields and $y$-basis interactions in Eq. (6); however, the addition of $y$-basis terms could immediately be used to both improve the

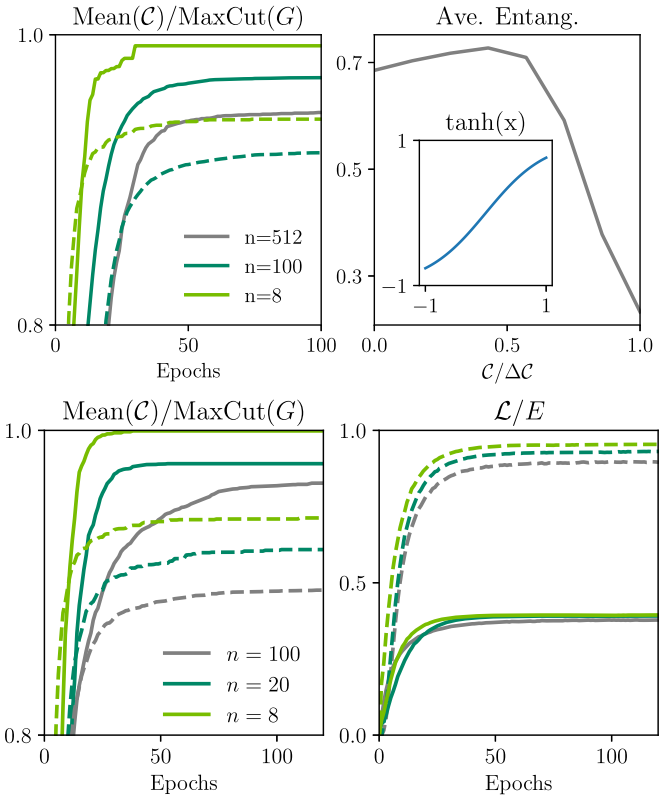


FIG. 3. Top right: Average cut (ratio of cut obtained with largest known solution) $\mathcal{C}$ convergence (left) for both MBE (solid lines) and VQE using the same circuit *Ansatz* (i.e., gate types, gate order, and circuit depth; dashed lines) with $L = 7$ ($n = 8$, 100) and $L = 13$ ($n = 512$). We note that in these experiments with $n = 8$, 100, MBE significantly outperforms VQE using the same gate *Ansatz*. While VQE with $n = 512$ was prohibitively memory inefficient to simulate for comparison, MBE with $n = 512$ outperforms VQE with $n = 8$, a system 1/64th of its size, as well as the leading single-shot classical algorithm (Table II). Top left: Average entanglement entropy for two-qubit subpartitions (maximum value per qubit is 1) vs fraction of calculated MaxCut convergence for nonlinear loss functions. Product state formation occurs because minimizing $\mathcal{L}_{\mathrm{MBE}}$ maximizes $\langle\sigma_i^z\rangle^2 + \langle\sigma_i^x\rangle^2$. Inset: $\tanh(x)$ nonlinear activation function further disincentivizes the maximization of one axis at the expense of the other. Bottom: Average cut (ratio of cut obtained with largest known solution) $\mathcal{C}$ convergence (left) and raw loss function $\mathcal{L}$ (right) with both two-graph MBE (solid lines) and VQE using the same circuit *Ansatz* (dashed lines) for $n = 8$, 20, and 100. MBE improves the *calculated* MaxCut convergence $\mathcal{C}$, although its ability to satisfy by the two encoded Ising models is limited by the normalization condition of Eq. (7). This is remedied by the rounding procedure of Eq. (12).

algorithm's performance as well as simultaneously optimize three (rather than two) graph vertices. Finally, we indicate that the use of nonlinear loss functions, which are known to increase the expressivity of classical neural networks, may contribute to the success of MBE with relatively shallow quantum circuits (compared, e.g., in this manuscript to VQE of the same circuit *Ansatz*). That is, the nonlinear loss functions also increase the expressivity classically, rather than with higher quantum entanglement and correlation alone.

As minimizing Eq. (6) under the constraints of Eq. (7) cannot yield classical solutions to Eq. (1), we define a rounding procedure for the classification and scoring of a cut $\mathcal{C}$ for a graph $G$:

$$\begin{aligned}\mathcal{C}_{\mathrm{MBE}}(\hat{\theta};G) = &\sum_{j<i}^{n/2} \frac{w_{ij}^{zz}}{2}\left[1 - R\left(\langle\sigma_i^z\rangle\right)R\left(\langle\sigma_j^z\rangle\right)\right] \\ &+ \sum_{j<i}^{n/2} \frac{w_{ij}^{xx}}{2}\left[1 - R\left(\langle\sigma_i^x\rangle\right)R\left(\langle\sigma_j^x\rangle\right)\right] \\ &+ \sum_{i,j}^{n/2} \frac{w_{ij}^{zx}}{2}\left[1 - R\left(\langle\sigma_i^z\rangle\right)R\left(\langle\sigma_j^x\rangle\right)\right], \qquad (8)\end{aligned}$$

where the classically implemented function $R$ rounds the measured expectation values to $\pm 1$. We note that this scoring is our true, or *computational*, MaxCut estimate, as it is the MaxCut assignment which results from projecting the qubit measurements of our quantum state from the $[-0.76, 0.76]$ codomain of our linear programming relaxation [$\tanh(x)$ activation function] back into the $\pm 1$ codomain of MaxCut nodes.

## III. RESULTS

In this section, we empirically validate our approach's performance by solving the MaxCut problem on a diverse set of graphs with up to 512 vertices. We first introduce the experimental settings and implementation details before presenting the results for two scenarios: (i) using MBE to solve $n$-vertex MaxCut problems with only $n/2$ qubits, and (ii) using MBE to encode two separate MaxCut graph instances in a single circuit. In addition to having an inherently lower quantum hardware overhead, both implementations of MBE demonstrate superior optimization performance.

Figure 3 (top) illustrates the average performance (ratio of cut obtained with largest known solution) of both MBE and VQE circuits for graphs of $n = 8$, 100 vertices and the MBE circuit alone for $n = 512$. In this manuscript, we use VQE circuits with the same gate *Ansätze* as the MBE circuits, as detailed in Fig. 2 and the referencing text. The $n = 512$ graph with VQE using the same circuit *Ansatz* was too memory inefficient for evaluation on a single NVIDIA A100 GPU. The simulations were completed using TENSORLY-QUANTUM, which runs on a PYTORCH [56] backend and implements tensor contractions with OPT-EINSUM [57]. The $n = 8$ instances are complete [all-to-all, $n(n-1)/2$-edge] graphs for which we calculated the exact ground truth through brute force computation, the $n = 100$ graphs are the first three 0.9 density weighted (4455-edge) MaxCut graphs (cataloged as the w09-100 instances) from the extensively studied Biq Mac library [49], and the $n = 512$ graph is the pm3-8-50 instance of the DIMACS library [50]. While the pm3-8-50 graph is relatively sparse (1536 edges), its edges are not limited to nearest-neighbor vertices (where nearest-neighbor vertices are defined by the closest vertex numbers and the limitation of edges between such vertices imposes some graph structure). Like other recent works [22,58], we implement simple entanglement-based pretraining prior to the MBE algorithm. Shallow circuits of depth $L = 7$ ($n = 8$ and $n = 100$ graphs) and $L = 13$ ($n = 512$ graph) are selected in order to adopt a protocol that is suitable for near-term quantum devices; however, the performance of the larger graphs ($n = 100, 512$) increases with moderately deeper circuits.

In our experiments, MBE consistently demonstrates a 5%–7% average performance increase (ratio of average cut obtained with largest known solution) across all $n$, as seen in Fig. 3 (top). We emphasize that not only does the MBE algorithm achieve larger cuts with higher frequency than VQE using the same circuit *Ansatz*, it simultaneously solves MaxCut($G$) with *half* the required qubits and parameters, as summarized in Table I. As quantum state space scales exponentially in $n$, this factor-of-two reduction in required qubits remains significant for quantum computing at scale. Even with shallow circuit-depth growth, such that $L$ increases only moderately from the 100-vertex BiqMac graphs ($L = 7$) compared to the 512-vertex DIMACS graph ($L = 13$), MBE outperforms the leading single-shot classical algorithm (Table II). Specifically, this observed requisite depth is only a sublogarithmic increase in $L$ compared to $n$, i.e., $\log_2(512)/13 < \log_2(100)/7$. MBE achieves an average cut of ∼95% of the largest known solution [59]. MBE also outperforms the classical algorithm in terms of the largest cut obtained for any given run, with ∼98% accuracy from just 30 total runs compared to ∼97% accuracy from 100 total runs. These performance increases may be greater for deeper circuits; however, our current contraction algorithm yields a maximum MBE circuit depth of $L = 13$ for 512-vertex graphs on a single GPU. As the simulation of these networks is ultimately memory bound, with memory requirements growing exponentially with circuit depth (as bond dimension grows with quantum correlations/entanglement), implementations of the algorithm are generally not classically tractable at scale. For instance, causal/lightcone-based methods for simplifying the simulation of a circuit become intractable for sufficiently complex circuits, and the implementation of MBE for larger system sizes requires distributed programming with considerable classical resources [60]. The simulation of deeper circuits could be provided by tensor contraction backends with improved memory management, such as the cuTensor library, while implementations of this scale on quantum hardware are consistent with the projections for moderate-term quantum devices. Although computational benchmarking for optimization problems has been demonstrated for thousands of qubits [39], to our knowledge, MBE with $n = 512$ appears to be the largest simulation of successful quantum optimization algorithms on graphs that are not limited to edges between nearest-neighbor vertices (where nearest-neighbor vertices are defined by the closest vertex numbers and the limitation of edges between such vertices imposes some graph structure).

TABLE I. Comparison of single-graph MBE and VQE using the same circuit *Ansatz* (i.e., gate types, gate order, and circuit depth) for $n = 100$ vertex graphs for circuits of depth $L = 7$. MBE requires half the number of qubits and parameters as VQE, yet produces significantly better solutions (higher cut $\mathcal{C}$), both on average and with higher probability.

| Method | Depth | No. Vertices | No. Qubits | No. Param | Mean($\mathcal{C}$)/MaxCut($G$) | $P(\mathcal{C} > T)$ |
|---|---|---|---|---|---|---|
| VQE | $L = 7$ | 100 | 100 | 400 | 0.921 | 12.5% |
| MBE (Ours) | $L = 7$ | 100 | 50 | 200 | *0.971* | *50.0%* |

MBE's improved performance on optimization problems is due to the two-axis constraint on each qubit, which only permits convergence to local minima that are *bistable* points for both the $z$ and $x$ axes. This is in contrast with the monostable (e.g., only stable along the $z$ axis) condition of VQE. Convergence to a local minima with bistability requires the concurrence of a zero gradient for both independently parametrized axes at a single, nonoptimal point in parameter space. As $\mathcal{L}_{\text{MBE}}$ is best extremized by larger $\langle\sigma^{\zeta}\rangle$, the circuit will tend towards satisfying the equality in Eq. (7). As this corresponds to entanglement-free qubits, there is a systematic disentanglement of the circuit into product states throughout training (Fig. 3, top right). To understand this process, note that for the general wave function

$$|\phi\rangle = \alpha|0_i 0_r\rangle + \beta|0_i 1_r\rangle + \gamma|1_i 0_r\rangle + \delta|1_i 1_r\rangle$$

describing any two qubits $i$ and $r$, the left-hand side of Eq. (7) for qubit $i$ can be written as

$$\langle\sigma_i^z\rangle^2 + \langle\sigma_i^x\rangle^2 = [(\beta+\gamma)^2 + (\alpha-\delta)^2][(\beta-\gamma)^2 + (\alpha+\delta)^2]. \tag{9}$$

In this form, we note that Eq. (7) is maximized when the concurrence (entanglement [61,62]) is minimized, and vice versa, driving the wave function towards product states as training progresses. Once disentanglement nears completion, the equality in Eq. (7) begins to hold and for any $\theta_t$ and qubit $i$, such that

$$\langle\sigma_i^z\rangle g_t(\sigma_i^z) = -\langle\sigma_i^x\rangle g_t(\sigma_i^x), \tag{10}$$

where $g_t$ are the gradients as given by Eq. (4). As $\langle\sigma_q^{\zeta}\rangle = 0$ is unfavorable for the optimization of $\mathcal{L}_{\text{MBE}}$, both axes of each qubit $i$ must be bistable with respect to each angle $\theta_t$ in order for the update of that parameter to halt.

In this manner, MBE is a sort of quantum analog to alternating minimization in classical algorithms [63], but which uses both quantum superposition and classical nonlinearity to minimize two cost functions simultaneously, rather than one sequentially. Alternating minimization has also proven useful in QAOA protocols [15,64–66], as have other perturbations such as filtered measurements [67]. Because $\mathcal{L}_{\text{MBE}}$ is calculated from single-qubit measurements, it is a form of measurement-based quantum computation [68–70]. MBE, like many of the most useful optimization algorithms, is a heuristic. Unlike polynomial-time approximation algorithms such as Goemans-Williamson [17], heuristic algorithms such as MBE are not guaranteed to obtain any specific approximation ratio for generic problems. Moreover, many existing approximation algorithms are conjectured to have approximation ratios that are already optimal [71]. For instance, the Unique Games Conjecture asserts that the 0.878 approximation ratio of Goemans-Williamson for MaxCut is optimal [19], meaning that efficient algorithms could not exceed this ratio [72]. Despite this lack of guarantees, heuristic algorithms often outperform approximation algorithms in practice.

TABLE II. Comparison of single-graph MBE with circuits of depth $L = 13$ and the leading single-shot classical relaxation heuristic [55], with comparable parameters, for $n = 512$ vertex graphs. MBE produces improved solutions (higher cut $\mathcal{C}$), both on average and in the most successful run.

| Method | Mean($\mathcal{C}$)/ MaxCut($G$) | Max($\mathcal{C}$)/ MaxCut($G$) |
|---|---|---|
| Classical relaxation | 0.939 | 0.969 |
| MBE (Ours) ($L = 13$) | *0.948* | *0.978* |

MBE can also encode two *distinct* $n$-vertex graphs into a single register of $n$ qubits and solve their two MaxCuts in parallel. This is equivalent to the simplified case of $w_{ij}^{zx} = 0\ \forall i, j \leqslant n$ in Eqs. (6) and (8) using $n$ qubits, yielding

$$\mathcal{L}_{\text{MBE}} = \sum_{j<i}^{n} w_{ij}^{zz} \tanh\left(\langle\sigma_i^z\rangle\right) \tanh\left(\langle\sigma_j^z\rangle\right) + \sum_{j<i}^{n} w_{ij}^{xx} \tanh\left(\langle\sigma_i^x\rangle\right) \tanh\left(\langle\sigma_j^x\rangle\right) \tag{11}$$

and

$$\mathcal{C}_{\text{MBE}}(\hat{\theta}; G) = \sum_{j<i}^{n} \frac{w_{ij}^{zz}}{2}\left[1 - R\left(\langle\sigma_i^z\rangle\right)R\left(\langle\sigma_j^z\rangle\right)\right] + \sum_{j<i}^{n} \frac{w_{ij}^{xx}}{2}\left[1 - R\left(\langle\sigma_i^x\rangle\right)R\left(\langle\sigma_j^x\rangle\right)\right]. \tag{12}$$

The average performance of MBE for solving two $n$-vertex graphs in parallel vs that of VQE with the same circuit *Ansatz* solving a single graph is displayed in Fig. 3 (bottom) for graphs of $n = 8$, 20, and 100 vertices with $L = 7$. For $n = 8$ and $n = 20$, we generate exact solutions to complete (all-to-all) graphs through brute force computation, whereas the $n = 100$ graphs are again the first three 0.9 density weighted MaxCut graphs from the Biq Mac library [49]. While for this fixed $L$, both VQE and two-graph MBE suffer decreasing performance with increasing $n$, two-graph MBE consistently demonstrates a 5%–7% average performance increase across $n$. We again note that the performance for large-$n$ graphs increases with greater $L$.

We emphasize that not only does the two-graph MBE algorithm find larger cuts more often than VQE when using

the same circuit *Ansatz*, it simultaneously solves MaxCut($G$) for *two* graphs $G$, rather than only one as with VQE. Applications that require MaxCut optimization typically require the evaluation of numerous graph instances, making parallel programming via MBE a useful technique. To provide an example, MaxCut is used to process graphics for image segmentation. As image data sets typically require hundreds of thousands, if not millions, of images, the factor-of-two speedup furnished by MBE offers a substantial and practical advantage. The parallel computation of multiple MaxCut values is also a special case of the qubit-efficient single-graph version, which enjoys the same robustness to local minima and further demonstrates the generality of the technique.

While in an ideal world optimization algorithms would produce optimal solutions (ground truths) with 100% probability, NP-hard problems such as MaxCut cannot be solved deterministically on a classical device in polynomial time unless $\mathbf{P} = \mathbf{NP}$ [73], which is often considered unlikely. As a result, approximate and heuristic algorithms are intensely researched. Although they lack performance guarantees, many heuristic methods can exceed the performance of polynomial-time algorithms when they are implemented using greater than a polynomial amount of computational resources. We note that these findings are not in conflict with complexity conjectures such as $\mathbf{P} \neq \mathbf{NP}$ [73] and the Unique Games Conjecture [71], as they specifically refer to implementations that exceed polynomial time. For instance, studies have indicated that it requires up to an exponential number of parameters and specific *Ansätze* to formulate VQE with a convex loss function, such that it is robust against local minima and likely to converge to the global minimum during gradient descent [8]. This is an impractical quantity, reaching $\sim 2^{99}$ ($\sim 2^{511}$) parameters for the $n = 100$ ($n = 512$) graphs considered here. Similarly, it has long been known that as the number of parameterized evolutions $p \to \infty$ in QAOA, ideal adiabatic computing can be obtained [9]. In contrast, the cumulative effects of probabilistic sampling (that is, running the randomly initialized circuit multiple times) lead to high-confidence convergence with markedly few repetitions $r$. In what follows, we reason that a probabilistic sampling of various shallow MBE circuit initializations is a more efficient alternative. As larger values of $\mathcal{C}$ are a direct certificate of superior optimization, there should be no preference for less efficient single-shot techniques. Furthermore, shallow implementations are particularly important for near-term quantum devices, which are prohibitively susceptible to noise at even moderate circuit depth.

Figure 4 (top) displays the probability that an optimal cut, which we define as $\mathcal{C} > T = 0.97 \times \text{MaxCut}(G)$, will be found for $n = 100$ graphs with both MBE and VQE. For depth $L = 7$, MBE produces an optimal cut with upwards of 50% probability for both the single-graph ($n$ vertices in $n/2$ qubits; Fig. 4, top left) and double-graph (two $n$ vertex graphs in $n$ qubits; Fig. 4, top right) protocols. In contrast, VQE using the same circuit *Ansatz* produces optimal cuts with just 12.5% probability. Furthermore, the likelihood of obtaining an optimal cut with MBE increases considerably with moderate circuit depth, rising to approximately 80% for $L = 13$ (left). We note that $L = 1$ circuits (right) obtain

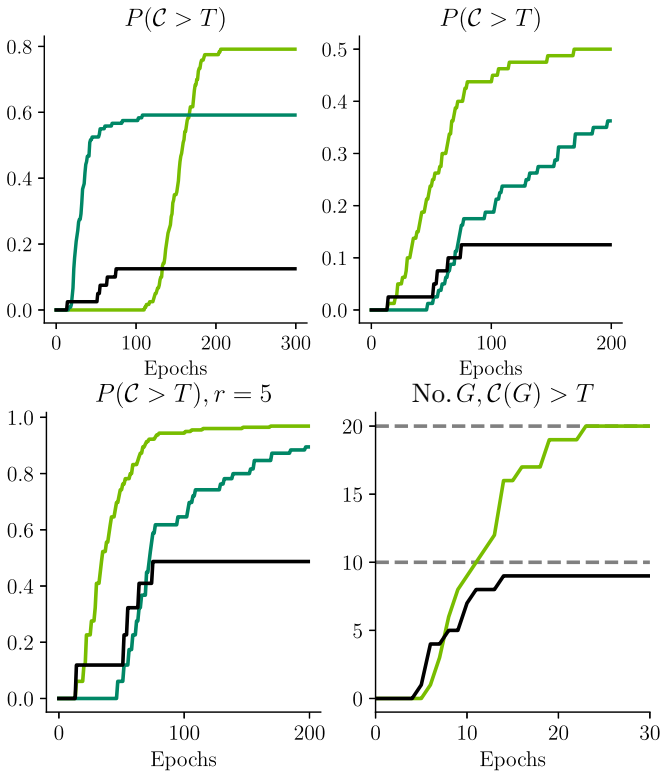


FIG. 4. Top left: The probability $P(\mathcal{C} > T)$ that cut $\mathcal{C}$ of an $n = 100$ graph is optimal using MBE with $L = 13$ (light green), MBE with $L = 7$ (dark green), and VQE with $L = 7$ (black). Increasing the depth from $L = 7$ to $L = 13$, while still shallow for $n = 100$, markedly improves the performance of these experiments. Top right: $P(\mathcal{C} > T)$ of $n = 100$ graphs using two-graph MBE with $L = 7$ (light green), two-graph MBE with $L = 1$ (dark green), and VQE with $L = 7$ (black). While the $L = 1$ case is entanglement free, it benefits from MBE's two-axis constraints. Bottom left: The probability of achieving an optimal cut ($\mathcal{C} > T$) of an $n = 100$ graph with $r = 5$ repeats using two-graph MBE with $L = 7$ (light green), two-graph MBE with $L = 1$ (dark green), and VQE with $L = 7$ (black). For the shallow $L = 7$ MBE circuit, five repetitions produce nearly deterministic results with less than 200 epochs. Bottom right: Number of $n = 20$ graphs with identified optimal cuts from the set of 10 instances and $r = 10$ repeats using two-graph MBE (green), and VQE (black). MBE not only successfully optimizes all (vs 90%) of $G$, it solves twice as many graphs in the same number of epochs.

optimal cuts with probability 0.36, tripling the convergence rate of standard VQE under the same *Ansatz* structure with 1/7th the circuit depth. As circuits with $L = 1$ are comprised of only local rotations without control gates, the totality of the performance is due to mutual constraints on multibasis *superpositions*, and not due to quantum entanglement. Like other entanglement-free formulations [74–76], this renders the circuit efficient for classical simulation and indicates that algorithms for a simulated superposition with multibasis constraints may hold promise as "quantum-inspired" classical algorithms. However, we note that quantum implementations are still of interest because other entanglement-free relaxations are known to suffer decreased performance with increasing circuit width $n$ [8]. Furthermore, MBE with even

modest entanglement and circuit depth markedly increases the probability of optimal convergence.

Figure 4 (bottom, left) shows the probability of obtaining at least one optimal cut for $n = 100$ graphs with $L = 7$ and $r = 5$, which nears 97% in fewer than 100 training steps for two-graph MBE circuits. For $r = 10$, convergence is greater than 99.9% and the $4nr = 4000$ parameters utilized for 10 repetitions still pale in comparison to the exponentially many required by deep-circuit techniques. As VQE with the same circuit *Ansatz* but $n = 100$ produces optimal cuts only 12.5% of the time, MBE is four times more effective than VQE for probabilistic optimization.

MBE also demonstrates superior performance over VQE with the same circuit *Ansatz* in terms of the diversity of tenable graphs (Fig. 4, bottom right). For $r = 10$, not only does two-graph MBE find optimal solutions for *all* of the complete $n = 20$ graphs tested (compared to 90% for VQE), its parallel implementation doubles the number of MaxCut instances optimized.

*Simulation considerations.* Numerically, $\mathcal{L}_{\mathrm{MBE}}$ is more compact for large or dense graphs, where the MPO $H$ quickly becomes cumbersome. However, for the single-qubit measurements required for $\mathcal{L}_{\mathrm{MBE}}$, contraction with a simple, single-qubit operator needs to occur $n$ times. In order to efficiently compute $n$ single-qubit measurements on large, exact tensor networks without either reconstructing an exponentially large ($2^{n/2}$) space or contracting over the full network $\sim n$ times, we use an efficient partial trace-based contraction scheme in which we construct $k$ distinct reduced density matrix operators,

$$\rho_k = \sum_{\{\beta,\gamma,\delta\notin K\}} \Psi^{\{\beta\}} U^{\{\beta,\gamma\}} U^{\{\gamma,\delta\}} \Psi^{\{\delta\}}, \tag{13}$$

where $K$ is the $k$th set of kept indices. $K$ should be sufficiently small so that the $2^{|K|}$ elements of $\rho_k$ remain numerically tractable. For each $\rho_k$, $|K|$ smaller partial traces are done to isolate single-qubit density matrices $\rho_q$, with which we take the single-qubit expectation values of Eq. (11),

$$\langle \sigma_q^{\zeta} \rangle = \mathrm{Tr}\big[\sigma_q^{\zeta} \rho_q\big], \tag{14}$$

where $\zeta = z, x$.

## IV. MBE FOR OTHER OPTIMIZATION PROBLEMS

MBE can be directly adapted to numerous quantum algorithms for optimization tasks by (1) distributing the problem variables into multiple bases of the available qubits (e.g., distributing the variables from $n$ qubits in the $z$ basis alone to $n/2$ qubits in the $z$ and $x$ bases, respectively), (2) replacing the multiqubit expectation values in the loss function with the products of single-qubit expectation values and nonlinear activation functions, and (3) classifying the final variable assignment by rounding the postoptimization spin values to the nearest allowable integer.

As an example, we detail the adaptation of MBE to the NP-hard $m$-node Traveling Salesman problems. The preexisting variational quantum implementation is detailed in [77]. The encoding requires $m^2$ qubits, one for each node $i$ and each step $p$. The objective function,

$$\begin{aligned} C = \sum_{i,j} \frac{w_{i,j}}{4} \sum_p \langle [1 - \sigma_{i,p}^z][1 - \sigma_{j,p+1}^z] \rangle \\ + K \sum_p \left( 1 - \sum_i \left[ 1 - \langle \sigma_{i,p}^z \rangle \right]/2 \right)^2 \\ + K \sum_i \left( 1 - \sum_p \left[ 1 - \langle \sigma_{i,p}^z \rangle \right]/2 \right)^2, \end{aligned} \tag{15}$$

minimizes the distances $w_{i,j}$ between nodes $i$ and $j$ as weighted two-qubit interactions between adjacent Traveling Salesman steps $p$. In addition, the regularizing terms (proportional to $K$) constrain each step to include exactly one node, and constrain each node to be visited at exactly one step.

MBE can encode this problem by making minor modifications to the typical implementation above. The only changes are applying nonlinear activation functions (i.e., $\tanh[\langle \sigma_{i,p}^z \rangle]$ instead of $\langle \sigma_{i,p}^z \rangle$) and substituting the product of two single-qubit observables for two-qubit observables (i.e., $\tanh[\langle \sigma_{i,p}^z \rangle] \tanh[\langle \sigma_{j,p+1}^z \rangle]$ instead of $\langle \sigma_{i,p}^z \sigma_{j,p+1}^z \rangle$). Once the optimization of the circuit is completed, the best path found by the algorithm can be read out by selecting the highest valued spin for each Traveling Salesman step $p$.

## V. DISCUSSION

In this manuscript, we introduced multibasis encoding (MBE), a technique for quantum optimization algorithms. In our experiments, MBE's performance on a diverse set of graphs exceeds that of VQE using the same circuit *Ansatz*. MBE also provides meaningful efficiency improvements over other VQAs that encode one variable per qubit for classical optimization problems, potentially closing the gap between near-term implementations and quantum advantage by reducing the overhead of quantum algorithms, such as a factor-of-two decrease in required qubits, which can readily be extended to a factor of three with the inclusion of the $y$ basis. While simulated using classically tractable *Ansätze*, the performance of our algorithm benefits from increased circuit depth. As the classical simulation complexity increases exponentially in circuit depth, this indicates that MBEs may enjoy meaningful quantum advantages at scale. Furthermore, when we extend our definition of accuracy to encompass probabilistic sampling of various circuit initializations, the effectiveness of MBE in our experiments was likewise strong.

MBE can be expanded to a broad framework of multi-axis qubit encodings, which would include any nonlinear quantum loss function that permits the optimization of multiple, mutually regularizing observables on a single qubit. These findings are likely to spur additional research in efficient qubit encodings and the application of our techniques to related algorithms. These include algorithms with high circuit depth or high circuit connectivity, which have broad enough light cones such that they are fundamentally intractable on classical hardware, and thus require quantum hardware. Since deeper circuits are attainable with more efficient tensor contraction methods or distributed computing efforts, this work encourages further development of large-scale quantum

simulation with tensor methods. Most critically, as these simulations are ultimately memory bound, the implementation of MBE at scale constitutes a strong candidate for quantum advantage.

We also leverage the powerful tensor techniques packaged in TENSORLY-QUANTUM to complete large-scale simulations of effective optimization algorithms on a single, consumer-grade GPU. Here we have produced what appears to be the largest to-date simulation of a quantum optimization algorithm with randomly distributed graphs that rivals classical performance. Such a successful and large-scale implementation demonstrates that simple and low-rank tensor representations are sufficient to model various techniques in quantum machine learning, and to do so without truncation or approximation. Finally, through the use of large-scale graphs, we demonstrate that the global qubit connectivity and high entanglement capacity lacked by both the MPS formalism and linearly connected near-term quantum devices do not preclude quantum optimization routines.

## ACKNOWLEDGMENTS

This work was done during T.L.P.'s internship at NVIDIA. At CalTech, A.A. is supported in part by the Bren endowed chair, and Microsoft, Google, Adobe faculty fellowships. S.F.Y. thanks the AFOSR and the NSF for funding. The authors would like to thank Brucek Khailany, Johnnie Gray, Garnet Chan, Andreas Hehn, and Adam Jedrych for conversations.

## APPENDIX A: INTUITION FOR MBE

Our MBE protocol uses a loss function which is inspired by, but not equivalent to, the long-range, *ZX* Hamiltonian,

$$H_{zx} = \sum_{j<i} w_{ij}^{zz} \sigma_i^z \sigma_j^z + \sum_{j<i} w_{ij}^{xx} \sigma_i^x \sigma_j^x + \sum_{j\neq i} w_{ij}^{zx} \sigma_i^z \sigma_j^x. \quad \text{(A1)}$$

The key difference between Eq. (A1) and MBE is that MBE utilizes the product of *single-qubit* measurements and nonlinear activation functions to encode separate vertices into the $z$ and $x$ bases [as explained in Eqs. (6) and (8)].

## APPENDIX B: INITIALIZATION

As an initialization procedure for the MBE, we restrict the amount of entanglement between groups of qubits and the untrained weights. We heuristically note that this leads to mildly improved performance and hypothesize that it may be due to the deleterious effects of random entanglement on barren plateaus and problem convexity, as explained in the main text. In particular, for the $n = 100, 512$ ($n = 8, 20$) cases, we pretrain up to five blocks of two qubits each by minimizing their entanglement with the remainder of the circuit until it is less than 15% (0.05%) of its maximal value, which can be done experimentally through state tomography or balanced single-qubit measurements, or until some maximal pretrain step limit is exceeded. Judiciously setting $\hat{\theta}$ is an alternative method that also satisfies our constraint. Despite this initial entanglement minimization, the qubits become rapidly entangled during the learning process, as illustrated in Fig. 3(a).

[1] W. Li, Y. Ding, Y. Yang, R. S. Sherratt, J. H. Park, and J. Wang, Parameterized algorithms of fundamental NP-hard problems: A survey, Hum. Cent. Comput. Inf. Sci. **10**, 29 (2020).

[2] A. Lucas, Ising formulations of many NP problems, Front. Phys. **2**, 5 (2014).

[3] D. Wecker, M. B. Hastings, and M. Troyer, Progress towards practical quantum variational algorithms, Phys. Rev. A **92**, 042303 (2015).

[4] J. R McClean, J. Romero, R. Babbush, and A. Aspuru-Guzik, The theory of variational hybrid quantum-classical algorithms, New J. Phys. **18**, 023023 (2016).

[5] M. Cerezo, A. Arrasmith, R. Babbush, S. C. Benjamin, S. Endo, K. Fujii, J. R. McClean, K. Mitarai, X. Yuan, L. Cincio, and P. J. Coles, Variational quantum algorithms, Nat. Rev. Phys. **3**, 625 (2021).

[6] A. Peruzzo, J. McClean, P. Shadbolt, M.-H. Yung, X.-Q. Zhou, P. J. Love, A. Aspuru-Guzik, and J. L. O'Brien, A variational eigenvalue solver on a photonic quantum processor, Nat. Commun. **5**, 4213 (2014).

[7] A. Kandala, A. Mezzacapo, K. Temme, M. Takita, M. Brink, J. M. Chow, and J. M. Gambetta, Hardware-efficient quantum optimizer for small molecules and quantum magnets, Nature (London) **549**, 242 (2017).

[8] J. Lee, A. B. Magann, H. A. Rabitz, and C. Arenz, Progress toward favorable landscapes in quantum combinatorial optimization, Phys. Rev. A **104**, 032401 (2021).

[9] E. Farhi, J. Goldstone, and S. Gutmann, A quantum approximate optimization algorithm, arXiv:1411.4028.

[10] M. P. Harrigan, K. J. Sung, M. Neeley, K. J. Satzinger, F. Arute, K. Arya *et al.*, Quantum approximate optimization of non-planar graph problems on a planar superconducting processor, Nat. Phys. **17**, 332 (2021).

[11] G. G. Guerreschi and A. Y. Matsuura, Qaoa for Max-Cut requires hundreds of qubits for quantum speed-up, Sci. Rep. **9**, 6903 (2019).

[12] G. Pagano, A. Bapat, P. Becker, K. S. Collins, A. De, P. W. Hess, H. B. Kaplan, A. Kyprianidis, W. L. Tan, C. Baldwin, L. T. Brady, A. Deshpande, F. Liu, S. Jordan, A. V. Gorshkov, and C. Monroe, Quantum approximate optimization of the long-range Ising model with a trapped-ion quantum simulator, Proc. Natl. Acad. Sci. USA **117**, 25396 (2020).

[13] L. Zhou, S.-T. Wang, S. Choi, H. Pichler, and M. D. Lukin, Quantum Approximate Optimization Algorithm: Performance, Mechanism, and Implementation on Near-Term Devices, Phys. Rev. X **10**, 021067 (2020).

[14] I. H. Kim and B. Swingle, Robust entanglement renormalization on a noisy quantum computer, arXiv:1711.07500.

[15] Z. Wang, N. C. Rubin, J. M. Dominy, and E. G. Rieffel, *xy* mixers: Analytical and numerical results for the quantum alternating operator ansatz, Phys. Rev. A **101**, 012320 (2020).

[16] F. G. Fuchs, H. Kolden, A. Øie, N. Henrik, and G. Sartor, Efficient encoding of the weighted Max $k$-Cut on a quantum computer using QAOA, SN Comput. Sci. **2**, 89 (2021).

[17] M. X. Goemans and D. P. Williamson, Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming, J. ACM **42**, 1115 (1995).

[18] J. Håstad, Some optimal inapproximability results, J. ACM **48**, 798 (2001).

[19] S. Khot, G. Kindler, E. Mossel, and R. O'Donnell, Optimal inapproximability results for Max-Cut and other 2-variable csps? SIAM **37**, 319 (2007).

[20] J. Preskill, Quantum Computing in the NISQ era and beyond, Quantum **2**, 79 (2018).

[21] J. R. McClean, S. Boixo, V. N. Smelyanskiy, R. Babbush, and H. Neven, Barren plateaus in quantum neural network training landscapes, Nat. Commun. **9**, 4812 (2018).

[22] T. L. Patti, K. Najafi, X. Gao, and S. F. Yelin, Entanglement devised barren plateau mitigation, Phys. Rev. Research **3**, 033090 (2021).

[23] C. O. Marrero, M. Kieferová, and N. Wiebe, Entanglement induced barren plateaus, arXiv:2010.15968.

[24] R. Wiersema, C. Zhou, Y. de Sereville, J. F. Carrasquilla, Y. B. Kim, and H. Yuen, Exploring entanglement and optimization within the Hamiltonian variational ansatz, PRX Quantum **1**, 020319 (2020).

[25] Z. Holmes, K. Sharma, M. Cerezo, and P. J. Coles, Connecting ansatz expressibility to gradient magnitudes and barren plateaus, PRX Quantum **3**, 010313 (2022).

[26] M. Cerezo, A. Sone, T. Volkoff, L. Cincio, and P. J. Coles, Cost-function-dependent barren plateaus in shallow quantum neural networks, Nat. Commun. **12**, 1791 (2021).

[27] M. B. Hastings, Classical and quantum bounded depth approximation algorithms, arXiv:1905.07047.

[28] S. Bravyi, A. Kliesch, R. Koenig, and E. Tang, Obstacles to Variational Quantum Optimization from Symmetry Protection, Phys. Rev. Lett. **125**, 260505 (2020).

[29] K. Marwaha and S. Hadfield, Bounds on approximating Max *k*xor with quantum and classical local algorithms, Quantum **6**, 757 (2022).

[30] K. Marwaha, Local classical MAX-CUT algorithm outperforms $p = 2$ QAOA on high-girth regular graphs, Quantum **5**, 437 (2021).

[31] J. Basso, E. Farhi, K. Marwaha, B. Villalonga, and L. Zhou, The quantum approximate optimization algorithm at high depth for MaxCut on large-girth regular graphs and the Sherrington-Kirkpatrick model, *17th Conference on the Theory of Quantum Computation, Communication and Cryptography (TQC 2022)*, Leibniz International Proceedings in Informatics (LIPIcs) Vol. 232, edited by F. Le Gall and T. Morimae (Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl, Germany, 2022), pp. 7:1–7:21.

[32] M. Fishman, S. R. White, and E. M. Stoudenmire, The iTensor software library for tensor network calculations, arXiv:2007.14822.

[33] R. Orús, A practical introduction to tensor networks: Matrix product states and projected entangled pair states, Ann. Phys. **349**, 117 (2014).

[34] 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).

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

[36] Y. Zhou, E. M. Stoudenmire, and X. Waintal, What Limits the Simulation of Quantum Computers? Phys. Rev. X **10**, 041038 (2020).

[37] E. S. Fried, N. P. D. Sawaya, Y. Cao, I. D. Kivlichan, J. Romero, and A. Aspuru-Guzik, qtorch: The quantum tensor contraction handler, PLoS One **13**, e0208510 (2018).

[38] D. Lykov, R. Schutski, A. Galda, V. Vinokur, and Y. Alexeev, Tensor network quantum simulator with step-dependent parallelization, arXiv:2012.02430.

[39] C. Huang, M. Szegedy, F. Zhang, X. Gao, J. Chen, and Y. Shi, Alibaba cloud quantum development platform: Applications to quantum algorithm design, arXiv:1909.02559.

[40] N. P. de Leon, K. M. Itoh, D. Kim, K. K. Mehta, T. E. Northup, H. Paik, B. S. Palmer, N. Samarth, S. Sangtawesin, and D. W. Steuerman, Materials challenges and opportunities for quantum computing hardware, Science **372**, eabb2823 (2021).

[41] T. L. Patti, J. Kossaifi, S. F. Yelin, and A. Anandkumar, Tensorly-quantum: Quantum machine learning with tensor methods, arXiv:2112.10239.

[42] T. L. Patti, J. Kossaifi, and A. Anandkumar, Tensorly-quantum: Tensor-based quantum machine learning, arXiv:2112.10239.

[43] J. Kossaifi, Y. Panagakis, A. Anandkumar, and M. Pantic, Tensorly: Tensor learning in python, J. Mach. Learn. Res. **20**, 925 (2019).

[44] C. W. Commander, Maximum cut problem, max-cutmaximum cut problem, max-cut, in *Encyclopedia of Optimization*, edited by Christodoulos A. Floudas and Panos M. Pardalos (Springer, Boston, 2009), pp. 1991–1999.

[45] R. M. Karp, *Reducibility Among Combinatorial Problems*, edited by J. D. Bohlinger, R. E. Miller, and J. W. Thatcher (Springer, Boston, 1972).

[46] S. R. White, Density Matrix Formulation for Quantum Renormalization Groups, Phys. Rev. Lett. **69**, 2863 (1992).

[47] J. Wurtz and P. Love, Maxcut quantum approximate optimization algorithm performance guarantees for $p > 1$, Phys. Rev. A **103**, 042612 (2021).

[48] H. Shen, P. Zhang, Y.-Z. You, and H. Zhai, Information Scrambling in Quantum Neural Networks, Phys. Rev. Lett. **124**, 200504 (2020).

[49] A. Wiegele, Biq Mac library - a collection of Max-Cut and quadratic 0-1 programming instances of medium size, Tech. Rep. (2007).

[50] The Dimacs library of mixed semidefinite-quadratic-linear programs (unpublished).

[51] A. Eddins, M. Motta, T. P. Gujarati, S. Bravyi, A. Mezzacapo, C. Hadfield, and S. Sheldon, Doubling the size of quantum simulators by entanglement forging, PRX Quantum **3**, 010309 (2022).

[52] M. Matty, Y. Zhang, T. Senthil, and E.-A. Kim, Entanglement clustering for ground-stateable quantum many-body states, Phys. Rev. Research **3**, 023212 (2021).

[53] X. Gao, E. R. Anschuetz, S.-T. Wang, J. I. Cirac, and M. D. Lukin, Enhancing generative models via quantum correlations, arXiv:2101.08354.

[54] K. I. Aardal and R. Weismantel, *Polyhedral Combinatorics: An Annotated Bibliography*, Universiteit Utrecht, UU-CS, Department of Computer Science (Utrecht University, Netherlands, 1996).

[55] S. Burer, R. D. C. Monteiro, and Y. Zhang, Rank-two relaxation heuristics for Max-Cut and other binary quadratic programs, SIAM J. Optim. (2001).

[56] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai *et al.*, Pytorch: An imperative style, high-performance deep learning library, in *Advances in Neural Information Processing Systems*, edited by H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Curran Associates, Inc., 2019), Vol. 32.

[57] D. G. a. Smith and J. Gray, Opt_einsum - A PYTHON package for optimizing contraction order for Einsum-like expressions, J. Open Source Software **3**, 753 (2018).

[58] J. Dborin, F. Barratt, V. Wimalaweera, L. Wright, and A. G. Green, Matrix product state pre-training for quantum machine learning, arXiv:2106.05742.

[59] P. Festa, P. M. Pardalos, M. G. C. Resende, and C. C. Ribeiro, Randomized heuristics for the Max-Cut problem, Opt. Method. Softw. **17**, 1033 (2002).

[60] S. Stanwyck, Nvidia sets world record for quantum computing simulation with cuquantum running on dgx superpod (unpublished).

[61] W. K. Wootters, Entanglement of formation and concurrence, Quantum Inf. Comput. **1**, 27 (2001).

[62] X. Gao, A. Sergio, K. Chen, S. Fei, and X. Li-Jost, Entanglement of formation and concurrence for mixed states, Front. Comput. Sci. China **2**, 114 (2008).

[63] P. Jain and P. Kar, Non-convex optimization for machine learning, FNT Mach. Learn. **10**, 142 (2017).

[64] S. Hadfield, Z. Wang, B. O'Gorman, E. G. Rieffel, D. Venturelli, and R. Biswas, From the quantum approximate optimization algorithm to a quantum alternating operator ansatz, Algorithms **12**, 34 (2019).

[65] L. Zhu, H. L. Tang, G. S. Barron, F. A. Calderon-Vargas, N. J. Mayhall, E. Barnes, and S. E. Economou, An adaptive quantum approximate optimization algorithm for solving combinatorial problems on a quantum computer, arXiv:2005.10258.

[66] J. Cook, S. Eidenbenz, and A. Bärtschi, The quantum alternating operator ansatz on maximum $k$-vertex cover, *2020 IEEE International Conference on Quantum Computing and Engineering (QCE)*, 2020, pp. 83–92.

[67] D. Amaro, C. Modica, M. Rosenkranz, M. Fiorentini, M. Benedetti, and M. Lubasch, Filtering variational quantum algorithms for combinatorial optimization, Quantum Sci. Technol. **7**, 015021 (2022).

[68] R. Raussendorf and H. J. Briegel, A One-Way Quantum Computer, Phys. Rev. Lett. **86**, 5188 (2001).

[69] R. Raussendorf, D. E. Browne, and H. J. Briegel, Measurement-based quantum computation on cluster states, Phys. Rev. A **68**, 022312 (2003).

[70] R. R. Ferguson, L. Dellantonio, A. A. Balushi, K. Jansen, W. Dür, and C. A. Muschik, Measurement-Based Variational Quantum Eigensolver, Phys. Rev. Lett. **126**, 220501 (2021).

[71] S. Khot, *On the Power of Unique 2-Prover 1-Round Games* (Association for Computing Machinery, New York, NY, 2002), pp. 767–775.

[72] R. O'Donnell, Ugc-hardness of Max Cut (unpublished).

[73] C. Wrathall, Complete sets and the polynomial-time hierarchy, Theor. Comput. Sci. **3**, 23 (1976).

[74] T. Wang and J. Roychowdhury, Oim: Oscillator-based Ising machines for solving combinatorial optimisation problems, arXiv:1903.07163.

[75] H. Goto, K. Tatsumura, and A. R. Dixon, Combinatorial optimization by simulating adiabatic bifurcations in nonlinear Hamiltonian systems, Sci. Adv. **5**, eaav2372 (2019).

[76] E. Crosson and A. W. Harrow, Simulated quantum annealing can be exponentially faster than classical simulated annealing, *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)* (2016), pp. 714–723.

[77] Qiskit Optimization Development Team, Max-Cut and Traveling Salesman problem (unpublished).