Qubit-efficient encoding schemes for binary optimisation problems

Benjamin Tan[1], Marc-Antoine Lemonde[1], Supanut Thanasilp[1], Jirawat Tangpanitanon[1], and Dimitris G. Angelakis[1,2]

[1]Centre for Quantum Technologies, National University of Singapore, 3 Science Drive 2, Singapore 117543

[2]School of Electrical and Computer Engineering, Technical University of Crete, Chania, Greece 73100

arXiv:2007.01774v2 [quant-ph] 21 Apr 2021

We propose and analyze a set of variational quantum algorithms for solving quadratic unconstrained binary optimization problems where a problem consisting of n_c classical variables can be implemented on $\mathcal{O}(\log n_c)$ number of qubits. The underlying encoding scheme allows for a systematic increase in correlations among the classical variables captured by a variational quantum state by progressively increasing the number of qubits involved. We first examine the simplest limit where all correlations are neglected, i.e. when the quantum state can only describe statistically independent classical variables. We apply this minimal encoding to find approximate solutions of a general problem instance comprised of 64 classical variables using 7 qubits. Next, we show how two-body correlations between the classical variables can be incorporated in the variational quantum state and how it can improve the quality of the approximate solutions. We give an example by solving a 42-variable Max-Cut problem using only 8 qubits where we exploit the specific topology of the problem. We analyze whether these cases can be optimized efficiently given the limited resources available in state-of-the-art quantum platforms. Lastly, we present the general framework for extending the expressibility of the probability distribution to any multi-body correlations. Our encoding scheme allows current generations of quantum hardware to explore the boundaries of classical intractability involving real world problem sizes, and can be implemented on currently available quantum hardware across various platforms, requiring only few dozens of qubits.

Benjamin Tan: b.tan@.nus.edu
Marc-Antoine Lemonde: cqtml@nus.edu.sg
Dimitris G. Angelakis: dimitris.angelakis@qubit.org

1 Introduction

In recent years, important experimental breakthroughs have propelled quantum computing as one of the most thriving fields of research [3, 8, 44, 48], with the long-term goal of building universal quantum computers capable of running algorithms with provable quantum speed-up [19, 45]. As the first generations of quantum hardware, referred to as noisy intermediate-scale quantum (NISQ) devices [40], do not yet fulfill the technical requirements to implement error-corrected universal quantum computing, increasing efforts are dedicated to design near-term algorithms capable of performing computational tasks with imperfect and limited quantum resources [4, 31]. Amongst the most promising paradigms are the variational quantum algorithms (VQA) [13, 22, 33, 35, 38]. In these algorithms, a parameterized quantum circuit is optimized using classical computing resources to generate a quantum state that represents an accurate approximate solution of the problem at hand. While a formal proof of any quantum advantages these algorithms might bring has yet to be found [43], applications of NISQ devices to real-world problems are already being explored in chemistry [32] and unconstrained binary optimisation (QUBO) problems [18].

The QUBO model is an NP-hard combinatorial problem that consists of minimizing a cost function of the form $C_x = \vec{x}^\intercal \mathcal{A} \vec{x}$, where $\vec{x} \in \{0,1\}^{n_c}$ is a vector of n_c classical binary variables and $\mathcal{A}$ is a real and symmetric matrix. VQAs such as the quantum approximate optimization algorithm (QAOA) [1, 5, 13–15, 21, 36, 37, 41, 49] and hardware efficient [6, 42] approaches have been applied to find approximate solutions to QUBO problems. QAOA in particular is able to ensure a lower bound on the quality of its solutions for sparse instances of QUBO problems using shallow circuits. This quality is then able to monotonically converge towards the exact solution for infinitely deep circuits recovering quantum adiabatic computing [13, 51]. Recent experiments have however highlighted the challenges in implementing the QAOA on problem graphs that differ from the native connectivity of the quantum hardware for increasing system sizes [21]. Hardware efficient approaches, on the other hand, are motivated by the simplicity of their imple-

mentations but do not guarantee a lower bound on the quality of the solutions. This is usually accomplished using series of gates native to the quantum platform and is unconstrained by the topology of the QUBO problem. However, depending on the implementation, these hardware efficient approaches can be plagued by exponentially large barren plateaus in their optimization landscapes as the number of qubits increases [34]. In addition to increasing algorithmic difficulties, the engineering overhead of scaling up the quantum hardware also currently limits the size of computational tasks to toy models. Previously proposed schemes to implement quantum algorithms to solve optimization problems have used a number of classical variables equal to the number of qubits available and were therefore limited to problem sizes involving only a few tens of them [1, 5, 6, 21, 36, 37, 41, 42, 49]. This is not representative of real-world optimization problems, where the number of classical variables n_c involved can be on the order of 10^4.

In this work, we tackle this problem by proposing an encoding scheme for QUBO models with n_c variables that can be implemented on $\mathcal{O}(\log n_c)$ number of qubits. We devise a strategy using n_a ancilla qubits and n_r register qubits to divide the QUBO problem into 2^{n_r} subsystems of n_a classical variables, requiring a total of $n_q = n_a + n_r$ qubits. This approach allows for a simultaneous search through each subsystem by exploiting the intrinsic parallelism offered by quantum devices. In this context, the resulting variational quantum state that encodes a probability distribution over all classical solutions is capable of capturing any n_a-body correlations between a number of QUBO variables that scale exponentially with n_r. This heuristic approach allows for the systematic increase in the correlations that can be captured in the probability distribution by progressively increasing the number of qubits. As an example, in the limit where each subsystem is composed of only a single classical variable, i.e. all correlations between classical variables are neglected, $n_r = \log_2(n_c)$ and optimization problems of size $n_c \sim 10^4$ could be tackled on quantum hardware with no more than 15 qubits. We emphasize that this limiting case of $n_a = 1$ can be efficiently classically simulated and therefore should not provide any quantum speed-up. At the other end of the spectrum, the most resource intensive limit of our encoding scheme is reached when $n_q = n_a = n_c$ and recovers the traditional approaches that are classically intractable and possibly offer quantum advantages. This encoding scheme provides a systematic way to traverse between these limits, thus allowing one to balance between capturing selected amounts of correlations whilst respecting the hardware capabilities of modern day devices. With the expected capabilities of upcoming NISQ devices, this scheme paves the way to explore the boundaries of classical intractability for real-world problem sizes.

In what follows, we introduce the general idea of our systematic encoding scheme and numerically demonstrate how the limit of $n_a = 1$ is able to solve QUBO problems while significantly reduce the number of qubits required. From there, we make the first step towards more expressive encoding by considering protocols to capture different subsets of two-body correlations and explore whether they can be optimized efficiently. We demonstrate numerically how a selective encoding scheme can be applied to the Max-Cut problem and show that exploiting the topology of a specific problem to select an efficient subset of correlations leads to better solutions. All protocols proposed in this manuscript are in line with the limitations of the current state-of-the-art quantum platforms.

2 QUBO model and the complete encoding scheme

The QUBO model is an NP-hard combinatorial problem that consists of minimizing a cost function of the form $C_x = \vec{x}^\intercal \mathcal{A} \vec{x}$, where $\vec{x} \in \{0,1\}^{n_c}$ is a vector of n_c classical binary variables and $\mathcal{A}$ is a real and symmetric matrix. This model is of particular interest due to its relationship with other optimization problems such as the Max-Cut, portfolio optimization and facility allocation problems [23, 25, 30]. Existing metaheuristic approaches such as the TABU search, genetic algorithms, and simulated annealing are capable of finding suitable solutions to problems consisting of $n_c \sim 10^4$ classical variables [16, 17].

In recent implementations of VQA applied to solving QUBO problems, each binary variable in $\vec{x}$ is represented by a single qubit, i.e. $n_q = n_c$; a mapping which we will refer to as the *complete encoding*. The resulting quantum state is parameterized by a set of angles $\vec{\theta}$ with the general form

$$|\psi_{\text{cp}}(\vec{\theta})\rangle = \hat{U}_{\text{cp}}(\vec{\theta})|\psi_0\rangle = \sum_{\vec{x}\in\{0,1\}^{n_c}} \alpha_{\vec{x}}(\vec{\theta})|\vec{x}\rangle, \qquad (1)$$

where $\hat{U}_{\text{cp}}(\vec{\theta})$ is the unitary evolution implemented on the quantum platform, $\{|\vec{x}\rangle = \otimes_i^{n_q}|x_i\rangle\}$ with $x_i \in \{0,1\}$ is the complete computational basis spawn by the n_q qubits and $|\psi_0\rangle$ is a given input state. By associating a classical solution $\vec{x}$ with a basis state $|\vec{x}\rangle$, the state $|\psi_{\text{cp}}(\vec{\theta})\rangle$ is able to encode all possible classical solutions in a linear superposition. This unique property of quantum mechanics opens the possibility for multiple classical solutions to be tested simultaneously and this intrinsic parallelism is a strong motivator in developing quantum algorithms for classical problems.

In the case where $\hat{U}_{\text{cp}}(\vec{\theta})$ is a universal quantum circuit, all $\alpha_{\vec{x}}$ in Eq. (1) can in principle be independent (up to the normalization condition). Consequently, this quantum state is able to capture all possible correlations between the classical variables and exhibits

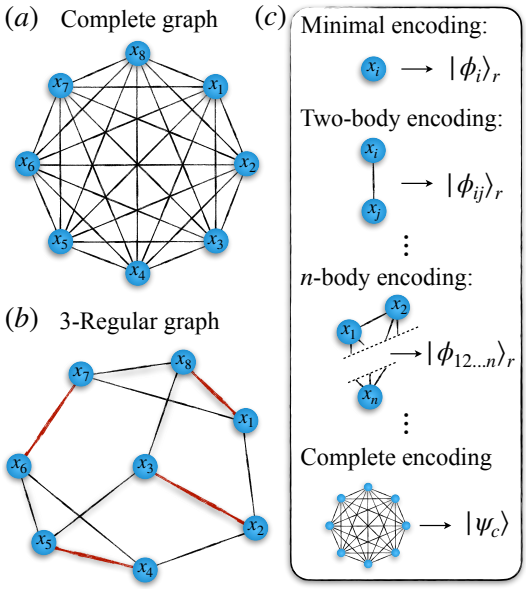

Figure 1: Schematic representation of the encoding schemes. (a) Complete graph representing a dense QUBO matrix $\mathcal{A}$. (b) A 3-Regular graph Max-Cut problem. The set of red edges is an example of perfect matching where each vertex is connected to only a single edge. (c) Encoding schemes where n_a-body correlation are encoded with n_a increasing from top to bottom. In the minimal encoding, each of the 2^{n_r} basis states $|\phi_i\rangle$ is used to represent a single classical variable x_i (vertex). In the n-body (two-body) encoding scheme, groups of n (two) classical variables are formed and each basis state represents a unique encoded group. In the complete encoding, each basis state represents an entire graph.

expressive power that is beyond classical computation [10, 11, 24]. The goal from here would be to efficiently navigate the exponentially large Hilbert space and reach the basis state(s) which represent the exact or approximate solution(s) to the QUBO problem.

In this complete encoding scheme, the QUBO model can be mapped onto an Ising Hamiltonian

$$\hat{H}_{\text{Ising}} = \frac{1}{4}\sum_{i,j}^{n_c} \mathcal{A}_{ij}(1-\hat{\sigma}_z^{(i)})(1-\hat{\sigma}_z^{(j)}), \qquad (2)$$

where $\hat{\sigma}_z^{(i)}$ is the z Pauli matrix acting on qubit i and $\mathcal{A}_{ij}$ are the elements of the matrix $\mathcal{A}$. The ground state of $\hat{H}_{\text{Ising}}$ is a basis state $|\vec{x}\rangle$ that corresponds to an exact solution $\vec{x}$ of the QUBO problem defined by $\mathcal{A}$. For general instances, $\hat{H}_{\text{Ising}}$ represents a system of interacting spins where all two-body interactions may be present.

A variational algorithm can then be implemented to find a suitable solution by using the ansatz $\hat{U}_{\text{cp}}(\vec{\theta})$ to produce trial states and finding the set of parameters $\vec{\theta}$ to minimize the cost function

$$C_{\text{cp}}(\vec{\theta}) = \langle\psi_{\text{cp}}(\vec{\theta})|\hat{H}_{\text{Ising}}|\psi_{\text{cp}}(\vec{\theta})\rangle. \qquad (3)$$

Here, Eq. (3) is a linear function of expectation values with a number of terms polynomial in the number of qubits.

Existing variational ansatzes for optimization problems can be divided in two distinct groups — approaches which require the Hamiltonian $\hat{H}_{\text{Ising}}$ to be implemented on the quantum hardware and those which utilize only native gates unconstrained by the specific problem. Approaches such as the QAOA, as implemented in Refs. [1, 5, 21, 36, 37, 41, 49], fall into the first category and benefit from being able to exploit some extent of adiabatic computing to search the Hilbert space [2]. In principle, the produced variational state is guaranteed to converge towards the exact solution for infinitely long quantum evolution $\hat{U}_{\text{cp}}(\vec{\theta})$. These approaches however, can be difficult to implement for generic QUBO problems. Approaches that fall into the second category have been implemented in Refs. [6, 42] and are designed to circumvent the technical challenges of implementing $\hat{H}_{\text{Ising}}$. However, these approaches do not guarantee the existence of an efficient path to the optimal solution and can become exponentially hard to optimize as the system size increases. While the ansatz proposed in this work belongs to the latter category, there should be no fundamental restrictions on devising circuit structures tailored toward a specific QUBO problem within the proposed encoding schemes.

3 Minimal encoding

While complete encoding schemes allow for all many-body correlations to be captured between classical variables, the number of qubits required limits their application to small system sizes with unfavorable scaling up perspectives. In what follows, we propose an encoding scheme which sacrifices this ability to capture correlations but allows for problem sizes to be scaled exponentially with the number of qubits. We refer to this mapping as the *minimal encoding*.

3.1 Expressibility of the minimal encoding

The minimal encoding scheme considered here requires one ancilla $n_a = 1$ and $n_r = \log_2 n_c$ register qubits for a total number of $n_q = \log_2 n_c + 1$ qubits. The parametrized quantum state can be expressed as

$$|\psi_1(\vec{\theta})\rangle = \sum_{i=1}^{n_c} \beta_i(\vec{\theta})[a_i(\vec{\theta})|0\rangle_a + b_i(\vec{\theta})|1\rangle_a] \otimes |\phi_i\rangle_r, \quad (4)$$

where the states $\{|\phi_i\rangle_r\}$ ($\{|0\rangle_a, |1\rangle_a\}$) are computational basis states of the register (ancilla) qubits. The premise is to define a one-to-one correspondence between each of the n_c classical variables x_i in $\vec{x}$ and a unique basis state $|\phi_i\rangle_r$, as depicted in Fig. 1 (c). The probability of the i^{th} classical variable to have the value 1 or 0 is given by $\Pr(x_i = 1) = |b_i|^2$ and $\Pr(x_i = 0) = 1 - |b_i|^2 = |a_i|^2$ respectively. The coefficients $\beta_i(\vec{\theta})$ capture the likelihood of measuring each register state $|\phi_i\rangle$ and thus the corre-

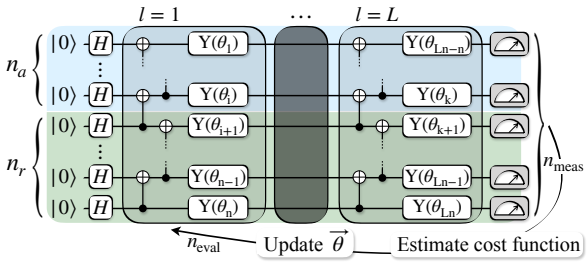

Figure 2: Hardware-efficient variational ansatz. The initial quantum state $|\psi_0\rangle = \frac{1}{\sqrt{2^{n_q}}} \otimes_{i=1}^{n_q} (|0\rangle + |1\rangle)$ is produced by the first layer of Hadamard gates. Each subsequent layer $1 \leq l \leq L$ is composed of a series of CNOT gates and single rotations $R_y(\theta_i)$ (denoted Y) where the $L \times n_q$ variational parameters are grouped in $\vec{\theta}$ ($n = n_q$ for readability). A single evaluation of the cost function requires n_{meas} measurements in the computational basis and one optimization process requires n_{eval} of these evaluations.

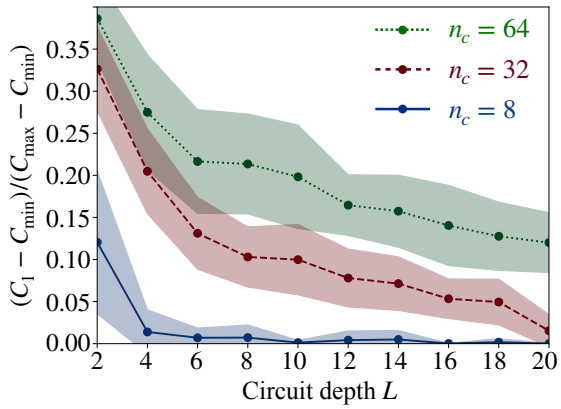

Figure 3: Optimized cost function value in the minimal encoding scheme as a function of circuit depth L for randomly generated problems of different sizes n_c with $n_{\text{meas}} \to \infty$. The lines represent the mean final value of 20 runs from 20 different starting points $\vec{\theta}_{\text{ini}}$. The shaded regions show one standard deviation from the mean. The optimization processes have been terminated at $n_{\text{eval}} = 5000$. The minimum and maximum values of the cost function, $C_{\min}$ and $C_{\max}$ respectively, were found using the classical optimization software Gurobi [20] and were used to normalize the cost function values from 0 to 1.

sponding state of the ancilla qubit. As an example, encoding the probability distribution over all solutions $\vec{x}$ of dimensions $n_c = 4$ requires $n_r = 2$. One can then define the mapping as $|\phi_1\rangle_r \equiv |00\rangle_r$, $|\phi_2\rangle_r = |01\rangle_r$, $|\phi_3\rangle_r \equiv |10\rangle_r$ and $|\phi_4\rangle_r = |11\rangle_r$. In doing so, the quantum state representing the unit probability of sampling $\vec{x} = (1,0,0,1)$ would read $|\psi_1\rangle = (|1\rangle_a|00\rangle_r + |0\rangle_a|01\rangle_r + |0\rangle_a|10\rangle_r + |1\rangle_a|11\rangle_r)/2$. A similar encoding strategy has been utilized in the context of image compression [27].

The limitation of this compact mapping is its ability to only encode distribution functions of statistically independent classical variables, i.e. where the probability of obtaining a particular classical solution $\vec{x}$ from the state is given by $\Pr(\vec{x}) = \prod_{i=1}^{n_c} \Pr(x_i)$. This comes as no surprise as the quantum state uses only n_c coefficients to encode a probability distribution over 2^{n_c} solutions. As a consequence, it is always possible to efficiently capture the resulting distribution functions using classical approaches, as we will discuss below in more detail. Despite these limitations, we examine this limiting case closely as it captures the core elements of the general encoding strategy.

3.2 Cost function to minimize

As with standard VQAs, we defined a cost function to be minimized using a set of parameters $\vec{\theta}$. Given that $|\psi_1(\vec{\theta})\rangle$ represents a distribution function over statistically independent classical variables, it adopts the form

$$C_1(\vec{\theta}) = \sum_{i,j=1}^{n_c} \mathcal{A}_{ij} \frac{\langle \hat{P}_i^1 \rangle_{\vec{\theta}} \langle \hat{P}_j^1 \rangle_{\vec{\theta}}}{\langle \hat{P}_i \rangle_{\vec{\theta}} \langle \hat{P}_j \rangle_{\vec{\theta}}} (1 - \delta_{ij}) + \sum_{i=1}^{n_c} \mathcal{A}_{ii} \frac{\langle \hat{P}_i^1 \rangle_{\vec{\theta}}}{\langle \hat{P}_i \rangle_{\vec{\theta}}}, \tag{5}$$

where $\hat{P}_i = |\phi_i\rangle\langle\phi_i|_r$ ($\hat{P}_i^1 = |1\rangle\langle 1|_a \otimes \hat{P}_i$) are the projectors over the register basis state $|\phi_i\rangle_r$ independent of the ancilla state (with the ancilla being in $|1\rangle_a$). The expectation value can be expressed as $\langle \hat{P}_i \rangle_{\vec{\theta}} = \langle \psi_1(\vec{\theta}) | \hat{P}_i | \psi_1(\vec{\theta}) \rangle$, giving $b_i(\vec{\theta}) = \langle \hat{P}_i^1 \rangle_{\vec{\theta}} / \langle \hat{P}_i \rangle_{\vec{\theta}}$.

The highly entangled quantum state that minimizes Eq. (5) adopts the form $|\psi\rangle = \sum_i \beta_i |\sigma_i\rangle_a \otimes |\phi_i\rangle$ with $\sigma_i = \{0,1\}$ and corresponds unambiguously to the exact solution $\vec{x} = [\sigma_1, \ldots, \sigma_{n_c}]$ that minimizes the QUBO problem defined by the matrix $\mathcal{A}$. This point is crucial as it ensures that finding the global minimum of Eq. (5) leads to the exact classical solution that minimizes the QUBO problem. Another important aspect of $C_1(\vec{\theta})$ is that it only depends on the set of norms $\{|b_i|^2\}$. As a consequence, partial tomography performed by a series of measurements solely in the computational basis is sufficient for its estimation. Finally, the cost function $C_1(\vec{\theta})$ in Eq. (5) cannot be reduced to a linear function of expectation values and therefore the QUBO model in the minimal encoding scheme cannot be described with a suitable Hamiltonian. A detailed derivation of Eq. (5) is presented in appendix A.

3.3 Variational protocol to solve randomly generated QUBO models

The quantum state $|\psi_1(\vec{\theta})\rangle = \hat{U}_1(\vec{\theta})|\psi_0\rangle$ is produced by a parameterized unitary evolution $\hat{U}_1(\vec{\theta})$ applied to an initial product state $|\psi_0\rangle \sim (|0\rangle_a + |1\rangle_a) \otimes \sum_{i=1}^{n_c} |\phi_i\rangle_r$. We consider a hardware efficient circuit as our ansatz $\hat{U}_1(\vec{\theta})$ in the form depicted in Fig. 2. This circuit starts with a layer of Hadamard gates applied to all the qubits initially in $|00\ldots00\rangle$ to produce $|\psi_0\rangle$. It then follows with an alternating sequence of nearest-neighbor CNOT gates and single qubit $R_y(\theta_i)$

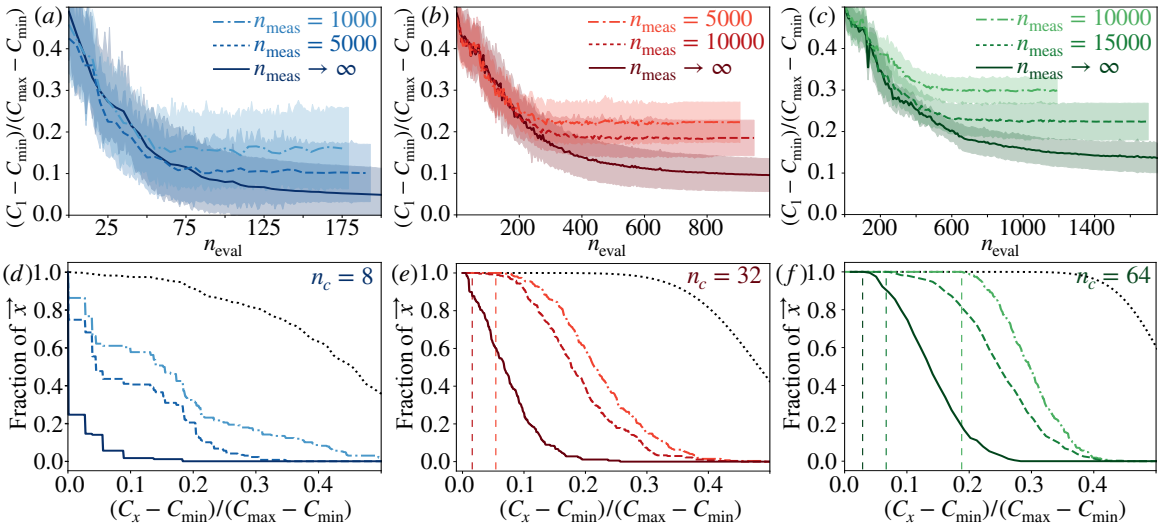

Figure 4: (a)–(c): Evolution of the (renormalized) cost function $\tilde{C}_1 = (C_1 - C_{\min})/(C_{\max} - C_{\min})$ during the optimization process in the minimal encoding scheme for different sizes of $\mathcal{A}$. Shaded regions represent one standard deviation from the average over 30 different starting $\vec{\theta}_{\text{ini}}$. The final value of the cost function decreases as more n_{meas} were used in evaluating $\tilde{C}_1$. (a) $n_q = 4$ qubits with circuit depth of $L = 4$ and matrix size of $n_c = 8$. (b) $n_q = 6$ qubits, circuit depth of $L = 12$, and matrix size of $n_c = 32$. (c) $n_q = 7$ qubits, circuit depth of $L = 18$, and matrix size of $n_c = 64$. (d)–(f): Cumulative distribution of solutions drawn from the final quantum state $|\psi_1(\vec{\theta}_{\text{opt}})\rangle$ using the parameters outlined in (a)–(c). 10 classical solutions were drawn from each starting $\vec{\theta}_{\text{ini}}$ after optimization. The y-axis represents the fraction of solutions with value above the corresponding $\tilde{C}_x$ (x-axis). Vertical lines show the minimum $\tilde{C}_x$ found from the solutions. The black dotted line shows the distribution of (d) all $2^{n_c} = 256$ solutions and (e)–(f) 4×10^8 randomly generated solutions.

rotations. Each successive application of CNOT gates and $R_y(\theta_i)$ rotations make up a single layer. This choice of ansatz represents the simplest case where qubits are arranged in a linear topology with nearest-neighbor couplings. It also produces states with only real-valued coefficients which efficiently restricts the Hilbert space since the cost function in Eq. (5) does not depend on any phases.

The optimization procedure is standard and first consists of randomly choosing a starting point for the variational parameters $\vec{\theta}_{\text{ini}}$ from a uniform distribution and measuring the output quantum state $|\psi(\vec{\theta}_{\text{ini}})\rangle$ in the computational basis. This quantum evolution is repeated n_{meas} times to estimate $C_1(\vec{\theta}_{\text{ini}})$. The results are fed to a classical optimizer which updates the parameters $\theta_{\text{old}} \to \theta_{\text{new}}$. The parameters are updated n_{eval} times until convergence or if a set of termination criteria is met. The resulting parameters are denoted $\vec{\theta}_{\text{opt}}$. From the final quantum state $|\psi(\vec{\theta}_{\text{opt}})\rangle$, a set of solutions $\{\vec{x}\}$ with value $C_x = \vec{x}^\intercal A \vec{x}$ are obtained by sampling each variable independently following $\Pr(x_i = 1) = |b_i|^2$ [cf. Eq. (4)].

In Fig. 3, we show the average optimized cost function as a function of circuit depth for 3 QUBO instances of different sizes, $n_c = 8, 32$ and 64, using $n_q = 4, 6$ and 7 qubits respectively[1]. In each instance, the elements of $\mathcal{A}$ were randomly drawn from a uniform distribution ranging from -1 to 1. COBYLA was chosen as the classical minimizer to update the variational parameters as it was found to give the best results for the least number of cost function evaluations [39]. The effects of a noisy circuit is shown in Appendix C where we compare the performance of a noise-free optimization for a $n_c = 32$ matrix to one with a simplified noise model applied. In Fig. 4 (a)–(c), we compare the infinite-measurement limit to simulated values obtained using $n_{\text{meas}} \sim 1 - 15 \times 10^3$, at specific circuit depths for each of the different problem sizes. Our findings show that increasing the number of measurements reduces the likelihood of the optimizer terminating in a local minima caused by fluctuations in the cost function. It also allows for finer tuning of the optimal parameters due to the increased precision when estimating $C_1(\vec{\theta})$, resulting in an increase in n_{eval}.

From each of the optimized states of Fig. 4 (a)–(c), 10 classical solutions $\vec{x}$ were drawn and distributed according to their normalized cost function value $\tilde{C}_x = (C_x - C_{\min})/(C_{\max} - C_{\min})$; their normalized cumulative sum is shown in Fig. 4 (d)–(f). The resulting histogram $y(\tilde{C}_x)$ (y-axis) represents the fraction of the solutions drawn that have a cost function greater than $\tilde{C}_x$ (x-axis). As an example, a value $y(\tilde{C}_x = 0.2) = 0.3$ means that 30% of solutions drawn have a cost function value of $\tilde{C}_x > 0.2$. The better the

[1] We note that the expressive power of $|\psi_1(\vec{\theta})\rangle$ can be fully captured within the complete encoding scheme by using only a single layer of $R_y(\theta_i)$ rotations applied to each qubit. Studying the minimal encoding scheme is therefore akin to examining the amount of resources required to map the simplest quantum circuit of n_c qubits onto an exponentially narrower circuit.

solutions $\{\vec{x}\}$ obtained, the sharper the histogram will peak at $\tilde{C}_x = 0$. The different coloured lines stand for different number of measurements n_{meas} and are compared to randomly drawn solutions as represented by the dotted black curves. The results show that the minimal encoding scheme was able to produce a significant portion of its solutions to be within 20% of the optimal cost function value for $n_c = 8, 32$ and a majority of solutions produced for the $n_c = 64$ case were found to be within 30% of the optimal cost function value. The numerical results also suggest that an increase in resources such as n_{meas}, n_{eval} and depth L are required to maintain comparable accuracy as the problem sizes increase.

In Appendix C, we investigate the robustness of the minimal encoding to experimental imperfections such as finite gate fidelities, coupling to the environment and readout errors. We further compare its performance to the more standard QAOA protocol in Appendix D. Following the recent state-of-the-art QAOA experiments for a fully connected problem in Ref. [21], implementing the QAOA for $n_c = 8$ variables would require 8 qubits and 612 gates for $p = 3$. In comparison, the minimal encoding with a hardware efficient ansatz only requires 4 qubits and 42 gates for $L = 6$ and was able to achieve an improvement in performance over the QAOA. While current noise levels encountered in state-of-the-art experiments affect all quantum optimization algorithms proposed so far, we show that compared to the QAOA, our efficient encoding is a step in the right direction by drastically reducing the resources required to solve larger-scale problems.

3.4 Classical Simulatability

As previously mentioned, the exponential decrease in the number of qubits offered by the minimal encoding also limits its advantage over classical methods. The probability distribution over statistically independent variables captured by the minimal encoding can be efficiently captured using only n_c continuous variables $\{w_i\}$. Each variable w_i replaces $\langle \hat{P}_i^1 \rangle_{\vec{\theta}} / \langle \hat{P}_i \rangle_{\vec{\theta}}$ in Eq. (5), resulting in a non-convex quadratic optimization problem with continuous variables which can be solved using quadratic programming techniques [50]. This is in contrast to the number of parameters, $N_p = L \times n_q$, required in variational quantum circuits. From our numerical experiments shown in Fig. 3, satisfactory results were obtained using $N_p = (L = 4) \times (n_q = 4) > n_c = 8$ and $N_p = (L = 16) \times (n_q = 6) > n_c = 32$ therefore suggesting classical approaches as more efficient routes. In the following section, we describe the methods for going beyond the limiting case of the minimal encoding where more sophisticated probability distributions can be captured by the quantum state through the use of additional ancilla qubits, therefore providing an opening for possible quantum advantages.

4 Two-body correlations

In this section, we show how two-body correlations between the classical variables of the QUBO problem can be introduced into the probability distribution captured by the quantum state. These correlations refer to the conditional probability of one of the variables taking on a certain value given the value of another variable when sampling the classical solution from the probability distribution. We then describe how the particular topology of the different QUBO instances can influence the subset of correlated pairs to be encoded. Specifically, when applied to a Max-Cut problem, we find that encoding only the correlations between pairs of variables that are connected within the graph leads to an improvement in the classical solutions obtained when compared to the minimal encoding approach.

4.1 General encoding scheme

We propose a general form of the quantum state that allows for the encoding of two-body correlations:

$$|\psi_2(\vec{\theta})\rangle = \sum_{i,j}^{n_{\text{pair}}} \beta_{ij}(\vec{\theta})[a_{ij}(\vec{\theta})|00\rangle_a + b_{ij}(\vec{\theta})|10\rangle_a + c_{ij}(\vec{\theta})|01\rangle_a + d_{ij}(\vec{\theta})|11\rangle_a] \otimes |\phi_{ij}\rangle_r, \tag{6}$$

where the register (ancilla) subspace now comprises $n_r = \log_2(n_{\text{pair}})$ ($n_a = 2$) qubits with n_{pair} being the number of two-body correlations encoded. Similar to the minimal encoding scheme, each basis state $|\phi_{ij}\rangle_r$ of the register space acts as a pointer. However, this pointer now points to the index of a pair of classical variables (x_i, x_j), as depicted in Fig. 1(c). The associated two-qubit ancilla state encodes the bare probability for all pair values, e.g. $\Pr(x_i = 0, x_j = 0) = |a_{ij}|^2$, $\Pr(x_i = 1, x_j = 0) = |b_{ij}|^2$ and so on. This encoding allows one to produce probability distributions that is able to capture correlations beyond statistically independent variables. A similar encoding strategy has been considered to address the issue of limited connectivity in quantum annealing platforms, allowing to simulate all-to-all connectivity from only local interactions [29].

The form of Eq. (6) is general enough to allow correlations to be captured between either all pairs of variables or only a subset of these pairs. In certain cases, one might be able to infer a preferred subset of pairs to encode based on the specific topology of the problem, allowing for an important reduction in the number of qubits required. In what follows, we highlight this point by comparing two general cases of frequently encountered QUBO models.

4.1.1 Selective subsets for sparse matrices

In QUBO instances where $\mathcal{A}$ is sparse, one might naturally expect that the most important correlations are those between the pairs of non-zero elements in $\mathcal{A}$. One seminal instance of sparse QUBO models is the d-regular Max-Cut problem where $d \ll n_c$. Each vertex on the corresponding graph is represented by a classical variable in $\vec{x}$ as depicted in Fig. 1 (b), and each edge by a non-zero off-diagonal element in $\mathcal{A}$. The resulting matrix $\mathcal{A}$ has d unit entries per row and column, and diagonal elements $\mathcal{A}_{i,i} = -d$. By selectively encoding only the $n_{\text{pair}} = n_c \times d/2$ pairs between non-zero elements in $\mathcal{A}$ (i.e. the edges), $n_q = \log_2(n_c \times d)+1$ are required, which is only $\log_2(d)$ qubits more than the minimal encoding scheme.

Illustrating with an example, encoding the 12 edges of the 3-regular graph with $n_c = 8$ shown in Fig. 1 (b) would require $n_r = 4$ register qubits. The pair (x_1, x_2) could be mapped onto the basis state $|\phi_{12}\rangle_r \equiv |0000\rangle_r$, the pair (x_1, x_7) on $|\phi_{17}\rangle_r \equiv |0001\rangle_r$ and so on until each edge is associated with a unique basis state. In the later sections, we apply this selective encoding method to solve a 3-regular Max-Cut problem with $n_c = 42$ number of variables using $n_q = 8$ qubits, allowing us to surpass the performance of the minimal encoding scheme.

4.1.2 Encoding all possible pairings for dense matrices

For more extreme instances where $\mathcal{A}$ is dense, such as the randomly generated QUBO models used in the previous section, selecting a specific subset of two-body correlations becomes completely arbitrary. The only unbiased approach then involves encoding all possible $n_{\text{pair}} = n_c(n_c-1)/2$ pairs of classical variables, requiring the maximal number of qubits $n_q = \log_2[n_c(n_c-1)] + 1$. Using this method to encode the 28 edges in the fully connected graph shown Fig. 1 (a) would require $n_r = 5$ register qubits. The mapping would proceed in a similar fashion as before, where the pair (x_1, x_2) can be associated to $|\psi_{12}\rangle_r \equiv |00000\rangle_r$, (x_1, x_3) to $|\psi_{13}\rangle_r \equiv |00001\rangle_r$ and so on. Despite the "unbiased" choice of pairing the variables, capturing all possible two-body correlations for general dense QUBO problems is typically not an efficient use of quantum resource as we shall observe later.

4.2 Averaging the probabilities and defining the cost function

Interpreting the quantum state $|\psi_2\rangle$ in Eq. (6) as a distribution function $\Pr(\vec{x})$ over the ensemble of classical solutions $\vec{x}$ is not as straightforward as the minimal encoding case. To better understand this statement, let us first consider the limit where the ensemble of pairs $\{(i,j)\}$ encoded would correspond to the set of edges in a 1-regular graph, also known as a perfect matching in graph theory and highlighted in Fig. 1 (b). In this case, each variable x_i is paired with a single other variable x_j and the probability to sample a solution $\vec{x}$ is uniquely defined as $\Pr(\vec{x}) = \prod_{(i,j)} \Pr(x_i, x_j)$. However, in the more general scenarios where at least one variable is included in more than one pair, the probability of sampling a solution $\vec{x}$ is not uniquely defined anymore. For example, in the limit where all pairs are encoded, there are $N_{\text{pm}}(n_c) = (n_c-1)!!$ ways of calculating $\Pr(\vec{x})$ with the possibility of vastly different results, where $N_{\text{pm}}(n_c)$ is the number of perfect matchings in a fully connected graph.

In order to be able to define a cost function in the form of Eq. (5) that is well-behaved despite the non-uniqueness of $\Pr(\vec{x})$, we need to define averaged probabilities $\bar{P}^{i,j}_{\sigma_i,\sigma_j}$ of sampling $x_i = \sigma_i$ and $x_j = \sigma_j$ where $\sigma = \{0,1\}$ that takes into account the multiple ways of calculating $\Pr(\vec{x})$. Doing so, we obtain the averaged probability of sampling $(x_i, x_j) = (1,1)$ from

$$\bar{P}^{i,j}_{1,1} = \sum_{l\neq i,j}^{n_c} \sum_{k\neq i,j,l}^{n_c} R_{ijkl}(\mathcal{G})(|c_{ki}|^2 + |d_{ki}|^2)(|c_{lj}|^2 + |d_{lj}|^2) + R_{ij}(\mathcal{G})|d_{ij}|^2, \tag{7}$$

where c_{ij} and d_{ij} are the amplitudes of the ancilla states given in Eq. (6) ($\vec{\theta}$ is kept implicit). Here, $R_{ij}(\mathcal{G}) = N_{\text{pm}}(\mathcal{G} - v_i - v_j)/N_{\text{pm}}(\mathcal{G})$ is the ratio between the number of perfect matchings after subtracting the vertices v_i and v_j from the graph $\mathcal{G}$ to the total number of perfect matchings in $\mathcal{G}$. Similarly, R_{ijkl} describes the same ratio but with 4 vertices removed instead. The graph $\mathcal{G}$ is built by mapping each classical variable to a vertex and each pair encoded in $|\psi_2\rangle$ to an edge. Expressions similar to Eq. (7) for $\bar{P}^{i,j}_{0,0}$, $\bar{P}^{i,j}_{0,1}$ and $\bar{P}^{i,j}_{1,0}$ are derived in appendix A. Using the same approach, one can also derive the averaged probability of sampling $x_i = 1$, leading to

$$\bar{P}^{i}_{1} = \sum_{k\neq i}^{n_c} R_{ik}(\mathcal{G})(|b_{ik}|^2 + |d_{ik}|^2), \tag{8}$$

where b_{ij} is also defined in Eq. (6).

In the limit where all possible pairs are encoded, these ratios are $R_{ij}(\mathcal{G}) = (n_c-3)!!/(n_c-1)!! = 1/(n_c-1)$ and $R_{ijkl}(\mathcal{G}) = R_{ij}(\mathcal{G})/(n_c-3)$. However, in the case where only a subset of pairs are encoded, $R_{ijkl}(\mathcal{G})$ depends on the vertices $\{i,j,k,l\}$ and is NP-hard to evaluate. One thus needs to resort to approximated ratios and our numerical experiments suggest that estimating $R_{ij}(\mathcal{G}) = 1/d$ and $R_{ijkl}(\mathcal{G}) = R_{ij}(\mathcal{G})/(d-2)$ for a d-regular graph leads to adequate behaviour of the probabilities.

Having the averaged probabilities defined, one can propose a cost function of the form

$$C_2(\vec{\theta}) = \sum_{i,j=1}^{n_c} A_{ij}\bar{P}^{i,j}_{1,1}(\vec{\theta})(1-\delta_{ij}) + \sum_{i=1}^{n_c} A_{ii}\bar{P}^{i}_{1}(\vec{\theta}). \tag{9}$$

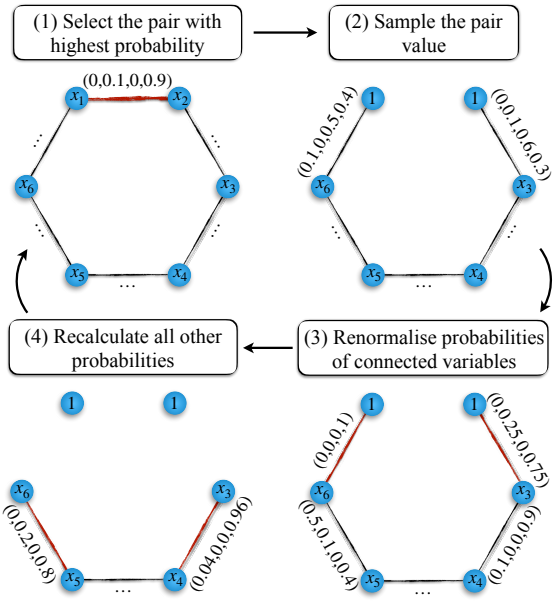

Figure 5: Schematic of the protocol to sample a solution from the ensemble of two-body correlations. The set of numbers displayed next to the edges between the vertices labeled x_i and x_j represents the probabilities $\bar{P}^{i,j} = (\bar{P}^{i,j}_{0,0}, \bar{P}^{i,j}_{1,0}, \bar{P}^{i,j}_{0,1}, \bar{P}^{i,j}_{1,1})$ with the convention $i < j$. In this example, assuming that $\bar{P}^{1,2}_{1,1} = 0.9$ has the largest probability in the entire graph, (x_1, x_2) would be the first pair to be selected in step (1). The remaining steps are outlined in the main text.

The key properties of Eq. (9) are similar to that of Eq. (5) in that (i) its global minimum corresponds unambiguously to the solution $\vec{x}$ that minimizes the QUBO problem, (ii) it can be estimated by a series of measurements solely in the computational basis, and (iii) it cannot be cast as a linear function of expectation values.

4.3 Sampling the classical solution from the quantum state

The form of $|\psi_2(\vec{\theta}_{\text{opt}})\rangle$ provides some flexibility in how solutions can be sampled from it. In the following, we describe a sampling protocol that fully exploits the encoded correlations and we show a simple example in Fig. 5.

The procedure is as follows.

1. Select the pair (i, j) with the most definite mean probabilities, i.e. the pair where $\bar{P}^{i,j}_{\sigma_i,\sigma_j}$ of sampling $x_i = \sigma_i$ and $x_j = \sigma_j$ is the closest to unity. As an example, let us consider that the probability to sample $(x_1, x_2) = (1, 1)$, $\bar{P}^{1,2}_{1,1} = 0.9$, is the largest of all mean probabilities, we select the pair $(1, 2)$.

2. Sample the value of the variables from the set of probabilities $\bar{P}^{i,j} = (\bar{P}^{i,j}_{0,0}, \bar{P}^{i,j}_{1,0}, \bar{P}^{i,j}_{0,1}, \bar{P}^{i,j}_{1,1})$.

3. Renormalize the probabilities of all variables connected to the pair evaluated in (2). Following the example, assume that $x_2 = 1$ has been sampled and is connected to the variable x_3, with probabilities $\bar{P}^{2,3} = (0, 0.1, 0.6, 0.3)$. The probability of sampling $(x_2, x_3) = (0, 1)$, $P^{2,3}_{0,1} = 0.6$, is now irrelevant given that $x_2 = 1$, leading to an updated probability of sampling $x_3 = 0$ of 0.25 and $\Pr(x_3 = 1) = 0.75$.

4. Adjust all remaining probabilities $\bar{P}^{k,l}$ consequently. Going back to the example, x_3 is now connected to x_4 where the probabilities were initially $\bar{P}^{3,4} = (0.1, 0, 0, 0.9)$. Since the probability to sample $x_3 = \{0, 1\}$ have been modified as exemplified in (3), $\bar{P}^{3,4}$ is re-evaluated to $(0.04, 0, 0, 0.96)$.

5. Repeat the steps from (1) until all variables have been assigned a value.

Conceptually, this method allows for a finite propagation of correlations along the graph $\mathcal{G}$ during the sampling. As an example, let us consider the case where correlations in the pairs (x_i, x_k) and (x_k, x_l) are explicitly encoded in $|\psi_2\rangle$ but not for the pair of variables (x_i, x_l). Using this sampling technique makes the probability of sampling $x_l = \{0, 1\}$ change conditionally for the sampled value of x_i, therefore inducing correlations. We stress that these induced correlations are not captured in the optimization process, but only during sampling.

4.4 Application to randomly generated QUBO instances versus a d-regular Max-Cut

In this section, we present the results obtained after optimizing quantum states of the form of Eq. (6) using the cost function $C_2(\vec{\theta})$ for two different instances of the QUBO model — a 3-regular Max-Cut problem of $n_c = 42$ and a randomly generated matrix $\mathcal{A}$ of $n_c = 8$.

4.4.1 Selective encoding for a 3-regular Max-Cut problem

To demonstrate the effectiveness of capturing correlations, we apply our encoding scheme for $n_a = 2$ to a randomly generated 3-regular Max-Cut problem with $n_c = 42$ vertices and 63 edges. In this example, selective encoding was used to only encode correlations between classical variables that are connected by an edge, requiring $n_q = \log_2(63) + 2 \lesssim 8$ qubits. By contrast, encoding all of the 861 possible pairs would require 12 qubits.

Using the same hardware-efficient circuit shown in Fig. 2, we apply the optimization protocol described in Sec. 3.3. In Fig. 6 (a), we show the final cost function $C_2(\vec{\theta}_{\text{opt}})$ as a function of the circuit depth L in the limit of $n_{\text{meas}} \to \infty$. We compare to optimization

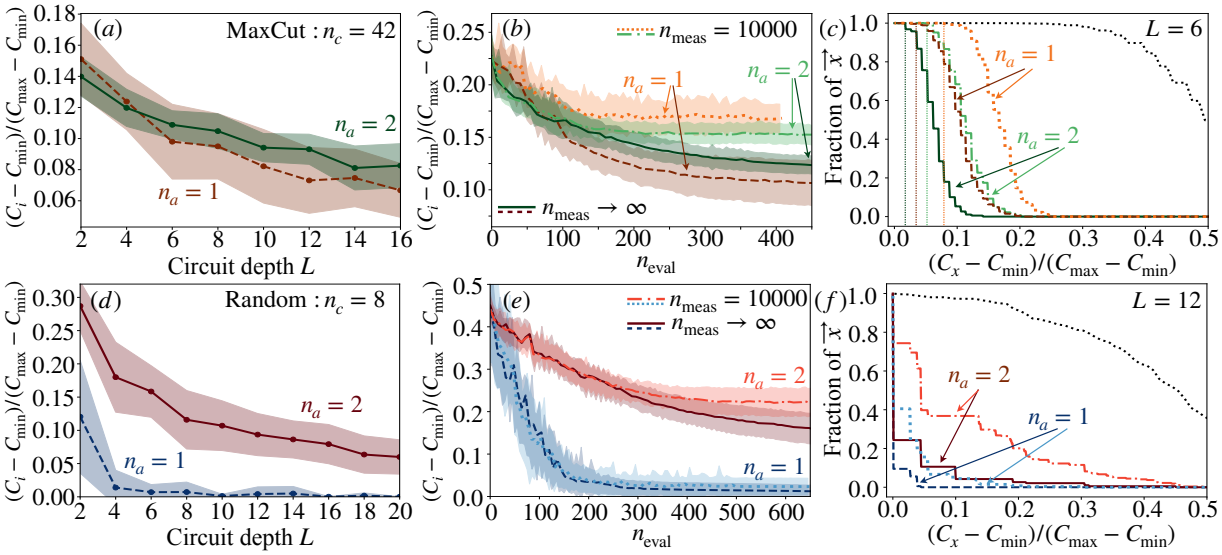

Figure 6: Comparison of results between $n_a = 1$ and $n_a = 2$ encoding schemes for a randomly generated matrix $\mathcal{A}$ of $n_c = 8$ classical variables with all 28 possible pairings being encoded (top row) and a randomly generated 3-regular Max-Cut problem of $n_c = 42$ using selective encoding to encode only vertices joined by an edge (bottom row). The lighter and darker shaded lines show $n_{\text{meas}} = 15000$ and $n_{\text{meas}} \to \infty$ respectively. $n_q = 8$ and $n_q = 6$ ($n_q = 7$ and $n_q = 4$) qubits were used for the Max-Cut and randomly generated instances respectively for the $n_a = 2$ ($n_a = 1$) encoding scheme. (a),(d) Comparison of final cost function values $C_1(\vec{\theta}_{\text{opt}})$ and $C_2(\vec{\theta}_{\text{opt}})$ as a function of the circuit depth L. Points show the mean value over 30 starting points $\vec{\theta}_{\text{ini}}$. Shaded regions represent one standard deviation from the mean. (b), (e) Comparison of $C_1(\vec{\theta}_{\text{opt}})$ and $C_2(\vec{\theta}_{\text{opt}})$ during the course of optimization. Mean number of evaluations for $n_{\text{meas}} \to \infty$ extend beyond what is shown. (c), (f) Cumulative distribution of solutions drawn from the optimized quantum states $|\psi_1(\vec{\theta}_{\text{opt}})\rangle$ of (b) and (e) respectively, sorted according to their energy. 10 classical solutions, $\vec{x}$, were drawn from each optimized quantum state. The black dotted line shows the distribution of (c) all $2^{n_c} = 256$ solutions and (f) 4×10^8 randomly generated solutions.

results for the same problem using the minimal encoding scheme $n_a = 1$. Panel (b) shows the differences in the optimization process between the $n_a = 1$ and $n_a = 2$ encoding schemes for $L = 6$ for $n_{\text{meas}} \to \infty$ and $n_{\text{meas}} = 10^4$. While C_1 and C_2 are both depicted in the same figure to demonstrate their respective performance, we stress that they are different quantities and might lead to substantial differences in the quality of the solutions sampled despite their comparable values. This discrepancy is further accentuated given the fundamentally different sampling protocols.

The distribution of solutions drawn from $|\psi_2\rangle$ show a substantial improvement in quality over the solutions obtained from $|\psi_1\rangle$, as depicted in Fig. 6 (c). Importantly, as we show in Appendix C, this improvement over the minimal encoding is preserved and even amplified in presence of experimental imperfections. Intuitively, this enhanced robustness to noise could be the results of the information redundancy encoded when capturing two-body correlations, a characteristic reminiscent of the general idea of error-correction. The use of selective encoding has thus allowed us to produce better quality solutions through a combination of encoding only the subset of two-body correlations that are expected to be the most relevant and reducing the complexity of the cost function $C_2(\vec{\theta}_{\text{opt}})$.

4.4.2 Encoding all pairs for randomly generated QUBO instances

We conclude the results by revisiting the matrix $\mathcal{A}$ with $n_c = 8$ consisting of elements drawn from a continuous uniform probability distribution. In this instance, all 28 possible pairings between the 8 classical variables are encoded, requiring a total of $n_q = \log_2(28) + 2 \lesssim 7$ qubits.

The results are shown in Fig. 6 (d)–(f) and compared to the results previously obtained in the minimal encoding scheme. Most importantly, panel (c) shows that solutions sampled from the statistically independent distribution function encoded in $|\psi_1\rangle$ are of better quality than those sampled from $|\psi_2\rangle$. These results strongly suggest that encoding all pairs is not an efficient use of quantum resources and can lead to a poorer performance during optimization as well as poorer quality solutions obtained from the final state.

Intuitively, this efficiency loss can be attributed to the use of a much larger Hilbert space to encode highly redundant, and possibly contradictory, information about classical correlations. This suggests that there is a balance to reach regarding the subset of pairs to be encoded and the resources required to do so; a more detailed discussion of the general case for any choice of n_a is presented in the following section.

5 Generalization to multi-body correlations

Now that we have described in detail a framework to make the first step beyond statistically independent classical variables and encode two-body correlations, generalizing the idea to encoding any set of n_a-body correlations is straightforward. Consider a variational quantum state of the form:

$$|\psi_a(\vec{\theta})\rangle = \sum_i \beta_i(\vec{\theta})|\varphi_i(\vec{\theta})\rangle_a|\Phi_i\rangle_r, \tag{10}$$

where the ancilla state $|\varphi_i(\vec{\theta})\rangle_a$ is composed of n_a qubits and is associated with a register state $|\Phi_i\rangle_r$ that points to a specified group i of n_a classical variables. In light of the previous section, whether $|\psi_a(\vec{\theta})\rangle$ can be efficiently optimized to solve a QUBO problem strongly depends on the choice of the encoded groups of n_a classical variables.

One of the simplest mapping strategies consists of encoding a selected set of n_c/n_a independent groups of n_a variables, i.e. where no variable is part of more than one group. The number of qubits needed for this,

$$N_{\text{ind}}(n_a) = \log_2(n_c/n_a) + n_a, \tag{11}$$

increases monotonically until the complete encoding threshold where $n_a = n_c$. In this strategy, there is a one-to-one correspondence between each of the n_c/n_a subgroup of n_a classical variables and a unique basis state of the $n_r = \log_2(n_c/n_a)$ register qubits. The quantum state $|\varphi_i(\vec{\theta})\rangle_a$ of the n_a ancilla qubits associated with the i^{th} subgroup encodes a distribution function that can capture all correlations among the variables of this subgroup. The optimization protocol can be interpreted as partitioning the QUBO problem into subgroups and simultaneously solving each of them using the complete encoding. This choice of mapping is one that is arbitrary as there is no fixed structure as to how the variables should be grouped. However, the minimal use of quantum resources might make this a desirable choice in certain situations.

Another strategy would be to encode all $\frac{n_c!}{n_a!(n_c-n_a)!}$ groups of n_a variables, which is the generalization of encoding all possible pairs for $n_a = 2$. This requires

$$N_{\text{all}}(n_a) = \log_2\left(\frac{n_c!}{n_a!(n_c-n_a)!}\right) + n_a \tag{12}$$

qubits, which is a non-monotonic function of n_a and can substantially exceed the total number of qubits required for the complete encoding, showing an inefficient use of quantum resource.

In between these two extremes are multiple mapping options and whether any of these encoding schemes can efficiently exploit the dominant correlations within a specific family of QUBO models is of great interest. For example, one could imagine encoding an ensemble of $(d+1)$-body correlations that follows the specific topology of a d-regular Max-Cut problem. In this case, each classical variable within the d-regular graph forms a group of $d+1$ elements. Encoding all of those n_c groups into a quantum state would require

$$N_{\text{reg}}(n_a) = \log_2(n_c) + n_a + 1, \tag{13}$$

qubits, where $d = n_a$. For $n_a \to (n_c - 1)$, i.e. a fully connected graph, the number of qubits exceeds the threshold $n_q = n_c$ by $\log_2(n_c)$.

To investigate the resources required to reproduce these probability distributions classically, we consider the simplest encoding strategy described in Eq. (11) where the classical variables are grouped into distinct subgroups. In this scenario, the cost function to minimize is a direct generalization of Eq. (5), i.e. it can still be cast as a quadratic optimization problem over continuous variables. This time however, capturing all correlations encoded in the quantum state would require 2^{n_a} classical variables for each of the 2^{n_r} subsystems, leading to a total of 2^{n_q} total variables, as expected. In contrast, by using a variational quantum circuit, the number of variational parameters involved during optimization remains $N_p = L \times n_q$. While it is expected for L to scale with the number of qubits involved, the exact nature of this scaling is still an open question. Anything sub exponential, which can be expected from previous analysis in the context of random quantum circuits [7], could lead to quantum advantages. One indication favouring this sub exponential depth can be seen in the context of random circuits where an ensemble of random unitaries with approximate t-design properties can be produced with polynomial circuit depth. Because a logarithmic compression in the number of qubits is unlikely to bring about any computational advantages (cf. Section 3.4), the crossover between the minimal encoding and the complete encoding, where the compression in number of qubits is polynomial, is of great interest and needs to be further studied.

6 Conclusion

In this work, we have proposed and analysed a systematic encoding scheme for variational quantum algorithms that allows one to capture increasing amount of correlations between classical variables in optimization problems. We first detailed the implementation of the minimal encoding scheme, using only $n_q = \log_2(n_c) + 1$ qubits to solve a QUBO model of size n_c. This significant reduction in qubits allowed us to tackle randomly generated problem instances of size $n_c = 8, 32$ and 64 using only $n_q = 4, 6$ and 7 qubits respectively. Our numerical solutions was able to find suitable high quality solutions using resources compatible with NISQ devices despite the inability to capture any correlations between the

classical variables. The use of a hardware efficient parameterized circuit allowed us to reduce the number of gates required during implementation. Implementing QAOA according to recent state-of-the-art experiments in performed in Ref. [21] would require 612 gates for $p = 3$ when applied to an $n_c = 8$ variable problem. However, superior solutions can be obtained using our minimal encoding scheme with as little as 42 gates and $n_q = 4$ qubits.

We also detailed encoding protocols that allow for two-body correlations to be captured between the classical variables. The number of qubits required scales logarithmically with the number of pairs encoded and we showed that exploiting the topology of the QUBO instance is essential for efficient optimization of the quantum state. By applying the two-body correlation encoding to a Max-Cut problem of 42 vertices, we were able to obtain better performance compared to the minimal encoding scheme.

The focus of this work was primarily on the encoding schemes outlined in the main text and was not intended as a thorough investigation of the most efficient optimization protocols. We believe that the results presented can still be improved upon substantially. One possible area for exploration could be finding an ansatz that would result in a smoother cost function landscape with shallower circuits. More adapted classical optimization methods may also bring significant improvements in the optimization process as it was found that a considerable fraction of optimization runs got stuck in local minimas [28, 46]. Improvement on that front may also substantially decrease the number of measurements required to reach comparable quality of solutions. Further avenues to explore would be whether generalizations to larger n_a-body correlations can be efficiently optimized and whether alternative ways of capturing correlations for dense problem instances can be found. More importantly, we wish to investigate the intermediate encoding schemes beyond the limit of classical intractability where quantum algorithms may outperform classical approaches.

7 Acknowledgments

This research is supported by the National Research Foundation, Prime Minister's Office, Singapore and the Ministry of Education, Singapore under the Research Centres of Excellence programme. It was also partially funded by Polisimulator project co-financed by Greece and the EU Regional Development Fund, the European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013).

A Derivation of the cost functions

In all of the encoding schemes outlined in the main text, the quantum state $|\psi(\vec{\theta})\rangle$ captures a probability distribution over all 2^{n_c} classical solutions. In this context, we generalize the QUBO cost function, $C_x = \vec{x}^\intercal A \vec{x}$, as a sum over all possible solutions weighted by their respective probability to be sampled, i.e.

$$\begin{aligned}\mathcal{C} &= \sum_{\{\vec{x}\}} \vec{x}^\intercal \mathcal{A}\, \vec{x} \Pr(\vec{x}),\\ &= \sum_{\{\vec{x}\}} \sum_{i,j=1}^{n_c} x_i \mathcal{A}_{ij} x_j \Pr(\vec{x}),\\ &= \sum_{i,j=1}^{n_c} \mathcal{A}_{ij} \sum_{\{\vec{x}|x_i=x_j=1\}} \Pr(\vec{x}). \end{aligned} \tag{14}$$

Here, $\{\vec{x}\}$ represents the ensemble of all the 2^{n_c} possible solutions $\vec{x}$ while $\{\vec{x}|x_i = x_j = 1\}$ represents only the subset of the 2^{n_c-2} solutions where the i^{th} and j^{th} variables in $\vec{x}$ are $x_i = x_j = 1$. To obtain the third line, we have explicitly used the fact that only variables with values equal to 1 contributes to the cost function. In what follows, we present in more details the following steps that lead to Eqs. (5) and (9) of the main text and provide further discussions about their properties.

A.1 Minimal encoding

In the minimal encoding, the state $|\psi_1(\vec{\theta})\rangle$ describes statistically independent classical variables where the probability of sampling $\vec{x}$ is $\Pr(\vec{x}) = \prod_{i=1}^{n_c} \Pr(x_i)$. In this case,

$$\sum_{\{\vec{x}|x_i=x_j=1\}} \Pr(\vec{x}) = \begin{cases} \Pr(x_i=1)\Pr(x_j=1) & \text{if} \quad i \neq j \\ \Pr(x_i=1) & \text{if} \quad i = j \end{cases}, \tag{15}$$

which, in terms of the quantum state amplitudes, reads

$$\sum_{\{\vec{x}|x_i=x_j=1\}} \Pr(\vec{x}) = \begin{cases} |b_i(\vec{\theta})|^2 |b_j(\vec{\theta})|^2 & \text{if} \quad i \neq j \\ |b_i(\vec{\theta})|^2 & \text{if} \quad i = j \end{cases}. \tag{16}$$

By substituting these results into Eq. (14), one gets

$$C_1(\vec{\theta}) = \sum_{i,j=1}^{n_c} \mathcal{A}_{ij} |b_i(\vec{\theta})|^2 |b_j(\vec{\theta})|^2 (1-\delta_{ij}) + \sum_{i=1}^{n_c} \mathcal{A}_{ii} |b_i(\vec{\theta})|^2. \tag{17}$$

The final form presented in Eq. (5) of the main text is obtained by expressing the probabilities $|b_i(\vec{\theta})|^2 = \langle \hat{P}_i^1 \rangle_{\vec{\theta}} / \langle \hat{P}_i \rangle_{\vec{\theta}}$ in terms of the projectors $\hat{P}_i$ and $\hat{P}_i^1$ (defined in the main text).

A.2 Two-body correlations

In the case where the variational quantum state $|\psi_2(\vec{\theta})\rangle$ encodes a given set of two-body correlations,

evaluating Eq. (14) is not as straightforward as in the minimal encoding. This is due to the multiple ways of evaluating the probability $\Pr(\vec{x})$ of sampling a solution $\vec{x}$, each of which capable of producing very different results. More precisely, for $\vec{x} = (\sigma_1, \sigma_2, \ldots, \sigma_{n_c})$, $\Pr(\vec{x}) = \prod_{\{(i,j)\}} \Pr(x_i = \sigma_i | x_j = \sigma_j)$ where $\Pr(x_i = \sigma_i | x_j = \sigma_j)$ represents the conditional probability to sample $x_i = \sigma_i$ given $x_j = \sigma_j$. Here the ensemble $\{(i,j)\}$ represents a set of independent encoded pairs where no variables are repeated, i.e. a perfect matching. Consequently, there are as many ways to evaluate $\Pr(\vec{x})$ as there are perfect matchings $N_{\rm pm}(\mathcal{G})$ in the graph $\mathcal{G}$, corresponding to the encoded pairs in $|\psi_2(\vec{\theta})\rangle$.

To evaluate Eq. (14), we average over all possible ways of evaluating $\Pr(\vec{x})$, denoted by $\{\Pr(\vec{x})\}$, and define the mean probabilities

$$\bar{P}^{i,j}_{1,1} \equiv \frac{1}{N_{\rm pm}(\mathcal{G})} \sum_{\{\Pr(\vec{x})\}} \sum_{\{\vec{x}|x_i=x_j=1\}} \Pr(\vec{x}), \tag{18}$$

for $i \neq j$. The mean probability to sample a single variable $x_i = 1$, $\bar{P}^{i}_{1}$, is given by the same above definition with $i = j$. There are two distinct scenarios that one can encounter while averaging over all possible perfect matchings corresponding to $x_i = x_j = 1$ in $\mathcal{G}$. The first is when the perfect matching contains an edge connecting x_i and x_j. There are $N_{\rm pm}(\mathcal{G}_{ij})$ of such instances, where $\mathcal{G}_{ij}$ is the graph obtained by subtracting the two vertices i and j. For each of these instances, the conditional probability $\Pr(x_i = 1 | x_j = 1) = |d_{ij}(\vec{\theta})|^2$ is directly encoded in the quantum state (see Eq. (6) of the main text). The second scenario occurs when the perfect matching does not include an edge connecting the vertices i and j to each other but instead to other vertices k and l. These cases appear within a subset of $N_{\rm pm}(\mathcal{G}_{ijkl})$ perfect matching instances, where $\mathcal{G}_{ijkl}$ is the graph obtained by subtracting the vertices i, j, k and l. In these scenarios, the conditional probability $\Pr(x_i = 1 | x_j = 1)$ is not directly encoded in the quantum state and has to be inferred from $\Pr(x_i = 1 | x_j = 1) = \Pr(x_k = 0, 1 | x_i = 1) \Pr(x_l = 0, 1 | x_j = 1) = (|c_{ki}|^2 + |d_{ki}|^2)$, where $\Pr(x_k = 0, 1 | x_i = 1)$ is the conditional probability of having $x_k = 0$ or $x_k = 1$ given $x_i = 1$.

Considering these contributions, we obtain the following mean conditional probabilities:

$$\begin{aligned}
\bar{P}^{i,j}_{1,1} &= \frac{N_{\rm pm}(\mathcal{G}_{ijkl})}{N_{\rm pm}(\mathcal{G})} \sum_{l\neq i,j}^{n_c} \sum_{k\neq i,j,l}^{n_c} \Pr(x_k = 0, 1 | x_i = 1) \Pr(x_l = 0, 1 | x_j = 1) + \frac{N_{\rm pm}(\mathcal{G}_{ij})}{N_{\rm pm}(\mathcal{G})} \Pr(x_i = 1 | x_j = 1), \\
&= \frac{N_{\rm pm}(\mathcal{G}_{ijkl})}{N_{\rm pm}(\mathcal{G})} \sum_{l\neq i,j}^{n_c} \sum_{k\neq i,j,l}^{n_c} \left[\left(|c_{ki}|^2 + |d_{ki}|^2\right) \left(|c_{lj}|^2 + |d_{lj}|^2\right) \right] + \frac{N_{\rm pm}(\mathcal{G}_{ij})}{N_{\rm pm}(\mathcal{G})} |d_{ij}|^2,
\end{aligned} \tag{19}$$

for $(i \neq j)$, and

$$\begin{aligned}
\bar{P}^{i}_{1,1} &= \frac{N_{\rm pm}(\mathcal{G}_{ik})}{N_{\rm pm}(\mathcal{G})} \sum_{i\neq k}^{n_c} \Pr(x_k = 0, 1 | x_i = 1), \\
&= \frac{N_{\rm pm}(\mathcal{G}_{ik})}{N_{\rm pm}(\mathcal{G})} \sum_{i\neq k}^{n_c} \left(|c_{ki}|^2 + |d_{ki}|^2\right),
\end{aligned} \tag{20}$$

for $i = j$.

The cost function in Eq. (14) thus adopts the final form

$$C_2 = \sum_{i,j=1}^{n_c} \mathcal{A}_{ij} \bar{P}^{i,j}_{1,1} (1 - \delta_{ij}) + \sum_{i=1}^{n_c} \mathcal{A}_{ii} \bar{P}^{i}_{1,1}, \tag{21}$$

as in Eq. (9) of the main text.

This averaging ensures a well behaved cost function where the quantum state which minimizes this cost function gives the unit probability of sampling the exact solution which minimizes the QUBO problem. The drawback of this method is the partial "washing out" of the encoded correlations as it can be seen by the first term (second scenario) in Eq. (19) which adopts the form of two statistically independent variables.

Following the steps outlined above, the following averaged probabilities can also be derived:

$$\bar{P}^{i,j}_{0,1} = \sum_{l\neq i,j}^{n_c}\sum_{k\neq i,j,l}^{n_c} R_{ijkl}(\mathcal{G})(|a_{ki}|^2+|b_{ki}|^2)(|c_{lj}|^2+|d_{lj}|^2) + R_{ij}(\mathcal{G})|c_{ij}|^2, \tag{22}$$

$$\bar{P}^{i,j}_{1,0} = \sum_{l\neq i,j}^{n_c}\sum_{k\neq i,j,l}^{n_c} R_{ijkl}(\mathcal{G})(|c_{ki}|^2+|d_{ki}|^2)(|a_{lj}|^2+|b_{lj}|^2) + R_{ij}(\mathcal{G})|b_{ij}|^2, \tag{23}$$

$$\bar{P}^{i,j}_{0,0} = 1 - \left[\bar{P}^{i,j}_{0,1} - \bar{P}^{i,j}_{1,0} - \bar{P}^{i,j}_{1,1}\right]. \tag{24}$$

B The cost function landscape

The cost functions $C_1(\vec{\theta})$ and $C_2(\vec{\theta})$, described by Eq. (5) and Eq. (9) of the main text, are nonlinear combinations of expectation values. This form leads to very different behaviours as a function of $\vec{\theta}$ when compared to the linear cost function $C_{\rm cp}(\vec{\theta})$ derived in the complete encoding limit [cf. Eq. (3) of the main text].

These differences are depicted in Fig. 7, where $C_{\rm cp}(\vec{\theta})$, $C_1(\vec{\theta})$ and $C_2(\vec{\theta})$ are plotted as a function of a single parameter θ_i with all other rotation angles being fixed at random values. The $\mathcal{A}$ matrix used in Fig. 7 (a)–(c) is the same randomly generated $n_c = 8$ matrix used in Section 3.3 of the main text, while the same $\mathcal{A}$ matrix describing the $n_c = 42$ 3-regular Max-Cut in Section 4.1.1 was used in panel (d). The circuit used to obtain the landscape of $C_{\rm cp}(\theta_i)$ consists of a single layer of $R_Y(\theta)$ applied in parallel to all qubits. This circuit was chosen as it consists of only single-qubit rotations with no entangling gates. The resulting quantum state can therefore only describe probability distributions of statistically independent classical variables in the complete encoding, and is equally expressible as $|\psi_1(\vec{\theta})\rangle$ in the minimal encoding.

For deep circuits and linear cost functions, Ref. [34] predicts the existence of barren plateaus for 2-design quantum circuits $\hat{U}(\vec{\theta})$. Interestingly, the non linear forms of $C_1(\vec{\theta})$ and $C_2(\vec{\theta})$ do not fulfil the necessary conditions underlying the proof derived in Ref. [34]. Consequently, we expect that a more constrained condition of a t-design quantum circuit, where $t > 2$, would be necessary to demonstrate the existence of these barren plateaus. In addition, for cost functions comprising of a linear combination of a $Poly(n_q)$ number of global observables, Ref. [9] predicts the existence of barren plateaus even for shallow circuits. Despite the fact that each observable considered in this work is a projector, i.e. global operator, the nonlinearity of $C_1(\vec{\theta})$ and $C_2(\vec{\theta})$ combined with the $\mathcal{O}(2^{n_q})$ number of terms involved also do not fulfil the necessary conditions for the proof in Ref. [9]. A more thorough investigation of the barren plateaus for nonlinear cost function is left for future work.

C Effects of noise

In this section, we investigate the performance of our encoding scheme under the effects of a noise model consisting of thermal relaxation errors, imperfect gate fidelities, and readout errors. Thermal relaxation and decoherence can be characterized by the relaxation constants T_1 and T_2 (distinct from T_2^*) respectively. Given a single-qubit density matrix ρ, the effects of thermal processes can be simulated by transforming ρ after a time evolution t as

$$\rho(t) \to \frac{1}{\rho_{00}+\rho_{11}e^{-t/T_1}}\begin{bmatrix}\rho_{00} & \rho_{01}e^{-t/T_2}\\ \rho_{10}e^{-t/T_2} & \rho_{11}e^{-t/T_1}\end{bmatrix}. \tag{25}$$

Gate errors are implemented via a depolarization channel that affects each qubit as it undergoes a gate operation. On top of its intended operation, a gate with error λ has an additional effect on ρ according to

$$\rho \to (1-\lambda)\rho + \frac{\lambda}{2^{n_q}}I \tag{26}$$

where I is the identity matrix representing the maximally mixed state. Readout error is the probability of obtaining an incorrect value of the qubit during measurement, i.e. reading a $|0\rangle$ when the qubit is in the $|1\rangle$ state and vice versa. In experimental quantum platforms, the magnitude for the errors above can differ between qubits, and we implement this model by assigning the qubits values drawn from a normal distribution characterized by a mean and standard deviation for each type of error.

Figure 8 shows a comparison in the performance of the minimal encoding as a function of circuit depth L for different levels of noise. Each data point is obtained by performing the entire optimization protocol in the presence of all the noise sources described above. The lightest orange dashed line (circle data points) shows the results using noise levels characteristic of existing state-of-the-art hardware [47]. For comparison, we also consider more optimistic values that can be expected in the upcoming generations of NISQ devices, shown by the darker lines. The simulation for the triangle data points was achieved using a $2\times$ increase in $\bar{T}$ and an increase in 2-qubit gate fidelity from $F_2 = 99\%$ to $F_2 = 99.9\%$ compared to the circle data points. The diamond points were obtained using a $5\times$ increase in $\bar{T}$ from the circle data points, and an increase in single and 2-qubit gate fidelities to $F_1 = F_2 = 99.99\%$. Mean readout errors were kept unchanged at 1% for all the noisy simulations For comparison, the darkest line (square data points) corresponds to noise-free simulations, i.e. $T = \infty$, $F_1 = F_2 = 1$ and perfect readout

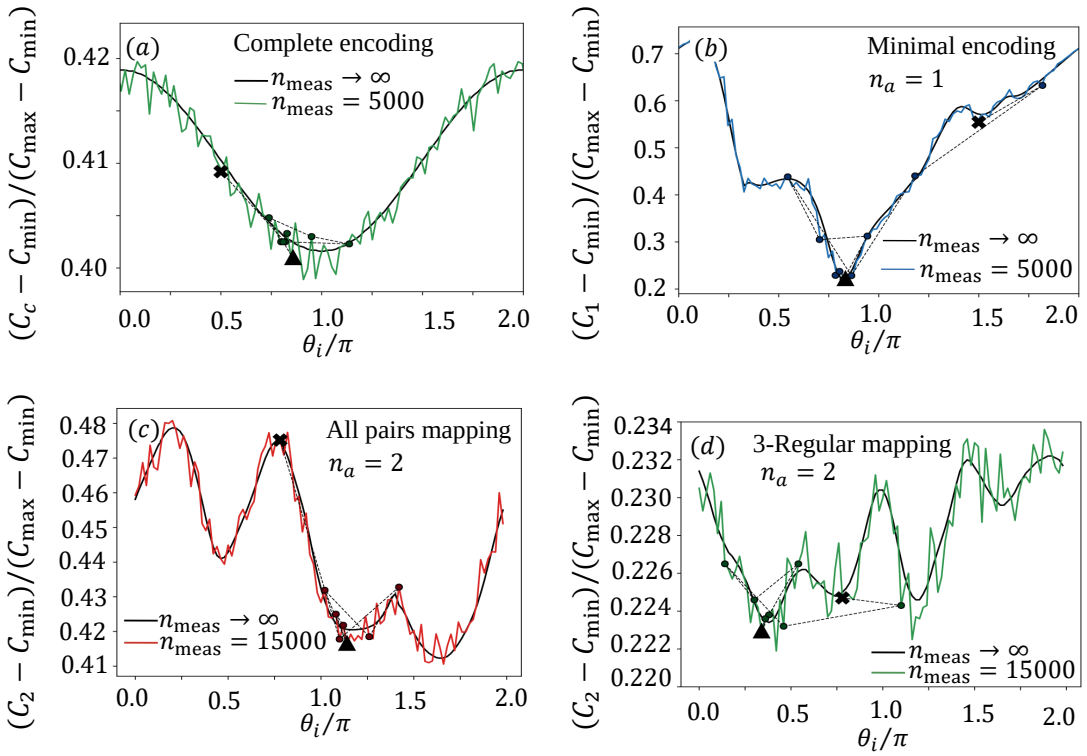

Figure 7: Cost function landscape as a function of a randomly chosen θ_i. Solid line shows the expectation value in the limit $n_{\text{meas}} \to \infty$, compared with the simulated value obtained from finite n_{meas}. The dotted lines from the cross to the triangle show the path taken by the COBYLA optimizer to find the optimal θ_i. (a)–(b) Complete and minimal encoding cost functions, $C_{\text{cp}}(\vec{\theta})$ and $C_1(\vec{\theta})$, for a randomly generated $\mathcal{A}$ with $n_c = 8$. Simulated values obtained using $n_{\text{meas}} = 5000$ and circuit depth $L = 4$. $n_q = 8$ qubits were used for the complete encoding compared to $n_q = 4$ in the minimal encoding. (c) $C_2(\vec{\theta})$ for a randomly generated $\mathcal{A}$ with $n_c = 8$ using $n_a = 2$ encoding scheme. All possible two-body correlations were encoded using $n_q = 7$ qubits. Simulated values obtained using $n_{\text{meas}} = 15000$ and circuit depth $L = 12$. (d) $C_2(\vec{\theta})$ for a 3-regular Max-Cut problem with $n_c = 42$ classical variables using $n_a = 2$ encoding scheme. Selective encoding with $n_q = 8$ qubits was used. Simulated values obtained using $n_{\text{meas}} = 15000$ and circuit depth $L = 6$.

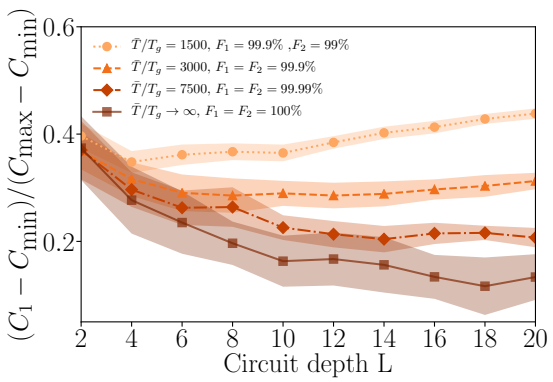

Figure 8: Optimized cost function value using the minimal encoding scheme as a function of circuit depth L under the effects of various noise levels for $n_c = 32$ classical variables and $n_{\text{meas}} = 5 \times 10^4$. T_1/T_g and T_2/T_g values for each qubit in the circuit were drawn from a normal distribution with a mean of $\bar{T}_1 = \bar{T}_2 \equiv \bar{T}$ and standard deviation $\sigma = \bar{T}/20$ obtained from surveying available quantum devices on the IBMQ cloud-based quantum computing platform [47]. Here T_g is the average single-qubit gate time and we used $T_{\text{CNOT}}/T_g = 6$ and $T_{\text{meas}}/T_g = 30$ where T_{CNOT} and T_{meas} are the average time for performing a CNOT gate and a measurement respectively. F_1 and F_2 are the gate fidelities for single-qubit and two-qubit operations respectively. We used a readout error of 1% for all curves except the darkest plain line for which we used 0%. All other parameters are identical to Fig. 3.

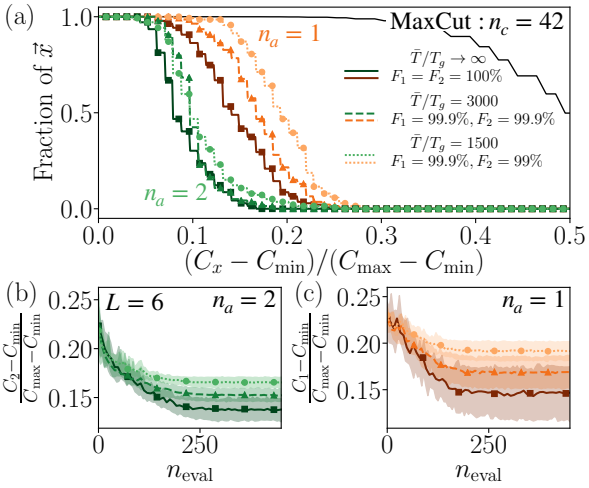

Figure 9: Solving a MaxCut problem with $n_c = 42$ in presence of noise. Reproduction of Fig. 6 (b) and (c) in presence of the noise models described in Appendix C and $n_{\text{meas}} = 5 \times 10^4$. The protocol, QUBO matrix and optimization parameters are identical as for Fig. 6.

but with finite n_{meas}. The result of these optimistic noise levels can be expected from a direct improvement in hardware implementation or from applying additional error-mitigation techniques [12].

In Fig. 9, we reproduce the results for the MaxCut problem of $n_c = 42$ variables as shown in Fig. 6, this time including the noise model introduced above and using $n_{\text{meas}} = 5 \times 10^4$ measurements. The comparison with the minimal encoding scheme shows an enhanced resilience to noise for the two-qubit ancilla encoding. This increased robustness could be attributed to the presence of redundancy in the encoding of correlations which can be thought of as reminiscent of the general ideas behind error encoding schemes. Such results therefore serve as additional motivation to further investigate higher-ancilla encoding schemes as they might procure additional protection against experimental imperfections.

D Comparison with QAOA

In this section, we compared the minimal encoding approach to the Quantum Approximate Optimization Algorithm (QAOA) under the effects of noise. QAOA is a commonly employed technique used to solve binary optimization problems on NISQ devices [1, 5, 13, 21, 36, 37, 41, 49, 51], where each classical variable is represented by a single qubit (complete encoding).

Using QAOA to solve a QUBO problem involves finding the state.

$$|\psi_{\text{QAOA}}(\vec{\gamma}, \vec{\beta})\rangle = \prod_p \hat{U}_x(\beta_p)\hat{U}_H(\gamma_p)|+\rangle^{\otimes n_q}, \qquad (27)$$

that minimizes Eq. (3) by finding the optimal variational parameters $\vec{\gamma}$ and $\vec{\beta}$. The unitaries $\hat{U}_H(\gamma)$ and $\hat{U}_x(\beta)$ take the form

$$\hat{U}_H(\gamma) = e^{-i\gamma \hat{H}_{\text{Ising}}},\ \hat{U}_x(\beta) = e^{-i\beta \sum_i^{n_q} \sigma_x^{(i)}}, \qquad (28)$$

where $\hat{H}_{\text{Ising}}$ is the Ising Hamiltonian described in Eq. (2). One of the main advantages of QAOA is its guaranteed monotonic convergence to the optimal solutions as $p \to \infty$. However, the current capabilities of NISQ devices limits p to small values and its performance has so far been drastically compromised when the interactions in $\hat{H}_{\text{Ising}}$ do not match the connectivity of the physical device.

In what follows, we simulate the QAOA protocol in which we consider a linear topology where two-qubit operations can only be applied on qubits adjacent to each other (similar throughout the manuscript). Solving a general QUBO problem where all-to-all interactions can be encountered therefore requires a network of SWAP gates for all two-qubit $\hat{\sigma}_z^{(i)} \otimes \hat{\sigma}_z^{(j)}$ interactions in $\hat{H}_{\text{Ising}}$ to be implemented. One of the most successful experimental implementation of the QAOA protocol to date relied on an efficient decomposition of the $e^{-i\gamma \hat{\sigma}_z^{(i)} \otimes \hat{\sigma}_z^{(j)}} \cdot \text{SWAP}$ operations into native gates [21] and we simulate the same decomposition with the same gate fidelities and gate times. We note that different platforms require different gate decompositions due to the different native gate sets and

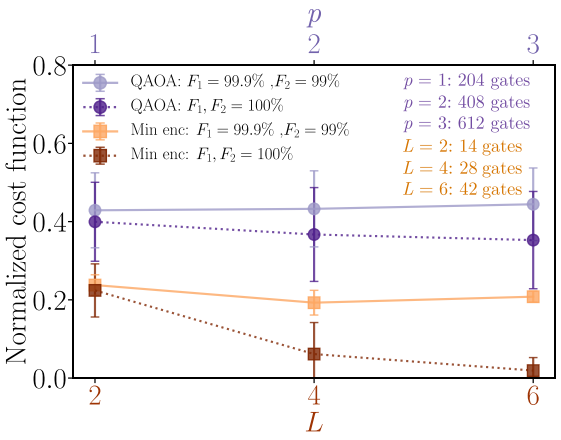

Figure 10: Noisy and noise-free simulations of QAOA and the minimal encoding scheme with hardware efficient ansatz applied to randomly generated A matrices with $n_c = 8$ classical variables. Data points and error bars show the mean and standard deviation of the normalized cost function over multiple matrices (and initial parameters $\vec{\theta}_{\text{ini}}$) after optimization for $p = 1, 2$ and 3 ($L = 2, 4$ and 6) for QAOA (minimal encoding). 8 qubits were used in QAOA while only 4 qubits were required in the minimal encoding scheme. For QAOA, the total fidelity of the $e^{-i\gamma\hat{\sigma}_z^{(i)}\otimes\hat{\sigma}_z^{(j)}}$ · SWAP operations is $F = 96.3\%$ with gate time $\bar{T}/T_g \approx 640$ as reported in Ref. [21]. All other parameters are identical to Fig. 8.

efforts have been devoted to reduce the number of gates required [26].

To find the best parameters $\vec{\gamma}$ and $\vec{\beta}$, we adopt a commonly used optimization strategy that consists of (1) scanning the two-dimensional parameter space spanned by (γ_1,β_1) for $p = 1$, (2) fixing (γ_1,β_1) to their optimal values (3) adding one additional layer $p \to p + 1$ and repeating steps (1)-(3) until reaching the desired final depth. We note that techniques to reduce the required size of the search grid for the parameters associated with $p > 1$ have been proposed [51]. During our simulation, the parameter scan was done over a 50×50 points grid ($\beta \in [0, \pi[$ and $\gamma \in [0, 2\pi[$) with 50 000 measurements per point, well within the capabilities of existing hardware [21]. It is also noteworthy that for general instances of QUBO problems, γ_1 might not be bounded to the domain above. To ensure that the finite grid resolution was not a limiting factor, the optimal parameters found on the initial 50×50 points grid were further improved by performing an additional refined local search.

In Fig. 10, we show the comparison in the performance of our minimal encoding scheme and the QAOA protocol for multiple randomly generated $\mathcal{A}$ matrices of size $n_c = 8$. This is in contrast to problems artificially curated to match the topology of the quantum device commonly used in experimental implementations [21, 26, 36, 37, 49]. Similar to the simulations shown in Appendix C, we use a noise model that, in addition to the finite gate fidelity, also includes thermal relaxations and readout errors. We emphasize that the search protocol in both the QAOA and minimal encoding scheme have been performed in presence of the simulated noise. This is also in contrast to some recent experimental QAOA demonstrations where the optimization is first performed with an ideal simulation and only the optimized circuit is executed on the quantum hardware [21, 26, 49]. For the minimal encoding, we used 15 starting points of randomly chosen parameters. With each optimization run resulting in $n_{\text{eval}} \approx 200$, this leads to a similar amount of circuit evaluations equivalent to $p = 1$ over a 50×50 points search grid.

We see from Fig. 10 that despite the provable monotonic converge of the QAOA for increasing p, practical limitations drastically limit its application to (small) generic QUBO problems. It is therefore not surprising that our minimal encoding considerably outperforms the QAOA given the important difference in the required resources. For a single layer of $p = 1$ for $n_c = 8$, implementing $\hat{U}_H$ via a SWAP network over 8 qubits requires 28 $\hat{\sigma}_z^{(i)} \otimes \hat{\sigma}_z^{(j)}$-SWAP interactions arranged in 8 subsequent layers. Following the gate decomposition used in Ref. [21], our implementation of QAOA required 84 2-qubit gates and 112 single qubit gates for each application of $\hat{U}_H$. In contrast, using a hardware efficient ansatz of the form shown in Fig. 2, our minimal encoding requires 4 qubits, 3 CNOT gates arranged in 2 subsequent layers and 4 R_y parametrized rotations for $L = 1$.

References

[1] Deanna M. Abrams, Nicolas Didier, Blake R. Johnson, Marcus P. da Silva, and Colm A. Ryan. Implementation of xy entangling gates with a single calibrated pulse. *Nature Electronics*, 3 (12):744–750, Nov 2020. ISSN 2520-1131. DOI: 10.1038/s41928-020-00498-1.

[2] Tameem Albash and Daniel A. Lidar. Adiabatic quantum computation. *Rev. Mod. Phys.*, 90:015002, Jan 2018. DOI: 10.1103/RevModPhys.90.015002.

[3] Arya K. Babbush R. et al. Arute, F. Quantum supremacy using a programmable superconducting processor. *Nature*, 574(7779):505–510, 2019. DOI: 10.1038/s41586-019-1666-5.

[4] Bela Bauer, Sergey Bravyi, Mario Motta, and Garnet Kin-Lic Chan. Quantum algorithms for quantum chemistry and quantum materials science. *Chemical Reviews*, 120(22): 12685–12717, Oct 2020. ISSN 1520-6890. DOI: 10.1021/acs.chemrev.9b00829.

[5] Andreas Bengtsson, Pontus Vikstål, Christopher Warren, Marika Svensson, Xiu Gu, Anton Frisk Kockum, Philip Krantz, Christian Križan, Daryoush Shiri, Ida-Maria Svensson, and et al. Improved success probability with greater circuit depth for the quantum approximate optimization

algorithm. *Physical Review Applied*, 14(3), Sep 2020. ISSN 2331-7019. DOI: 10.1103/physrevapplied.14.034010.

[6] Lee Braine, Daniel J Egger, Jennifer Glick, and Stefan Woerner. Quantum algorithms for mixed binary optimization applied to transaction settlement. *arXiv preprint arXiv:1910.05788*, 2019.

[7] Fernando G.S.L. Brandão, Aram W. Harrow, and Michał Horodecki. Local Random Quantum Circuits are Approximate Polynomial-Designs. *Communications in Mathematical Physics*, 346 (2):397–434, 2016. ISSN 14320916. DOI: 10.1007/s00220-016-2706-8.

[8] Colin D. Bruzewicz, John Chiaverini, Robert McConnell, and Jeremy M. Sage. Trapped-ion quantum computing: Progress and challenges. *Applied Physics Reviews*, 6(2):021314, 2019. DOI: 10.1063/1.5088164.

[9] M. Cerezo, Akira Sone, Tyler Volkoff, Lukasz Cincio, and Patrick J. Coles. Cost function dependent barren plateaus in shallow parametrized quantum circuits. *Nature Communications*, 12(1), Mar 2021. ISSN 2041-1723. DOI: 10.1038/s41467-021-21728-w.

[10] Brian Coyle, Daniel Mills, Vincent Danos, and Elham Kashefi. The born supremacy: quantum advantage and training of an ising born machine. *npj Quantum Information*, 6(1), Jul 2020. ISSN 2056-6387. DOI: 10.1038/s41534-020-00288-9.

[11] Yuxuan Du, Min-Hsiu Hsieh, Tongliang Liu, and Dacheng Tao. Expressive power of parametrized quantum circuits. *Physical Review Research*, 2(3), Jul 2020. ISSN 2643-1564. DOI: 10.1103/physrevresearch.2.033125.

[12] Suguru Endo, Zhenyu Cai, Simon C. Benjamin, and Xiao Yuan. Hybrid quantum-classical algorithms and quantum error mitigation. *Journal of the Physical Society of Japan*, 90(3): 032001, Mar 2021. ISSN 1347-4073. DOI: 10.7566/jpsj.90.032001.

[13] Edward Farhi, Jeffrey Goldstone, and Sam Gutmann. A quantum approximate optimization algorithm. *arXiv preprint arXiv:1411.4028*, 2014.

[14] Edward Farhi, Jeffrey Goldstone, and Sam Gutmann. A quantum approximate optimization algorithm applied to a bounded occurrence constraint problem. *arXiv preprint arXiv:1412.6062*, 2014.

[15] Edward Farhi, Jeffrey Goldstone, Sam Gutmann, and Leo Zhou. The quantum approximate optimization algorithm and the sherrington-kirkpatrick model at infinite size. *arXiv preprint arXiv:1910.08187*, 2019.

[16] Dimitris Fouskakis and David Draper. Stochastic optimization: a review. *International Statistical Review*, 70(3):315–349, 2002. DOI: 10.1111/j.1751-5823.2002.tb00174.x.

[17] F. Glover M. Lewis Z.P Lü H.B Wang Y. Wang G. Kochenberger, J.K. Hao. The unconstrained binary quadratic programming problem: a survey. *Journal of Combinatorial Optimization*, 28 (1):58–81, 2014. DOI: 10.1007/s10878-014-9734-0.

[18] Michael R Garey and David S Johnson. *Computers and intractability*, volume 174. freeman San Francisco, 1979.

[19] Lov K. Grover. Quantum mechanics helps in searching for a needle in a haystack. *Phys. Rev. Lett.*, 79:325–328, Jul 1997. DOI: 10.1103/PhysRevLett.79.325.

[20] LLC Gurobi Optimization. Gurobi optimizer reference manual, 2020. URL `http://www.gurobi.com`.

[21] Matthew P. Harrigan, Kevin J. Sung, Matthew Neeley, Kevin J. Satzinger, Frank Arute, Kunal Arya, Juan Atalaya, Joseph C. Bardin, Rami Barends, Sergio Boixo, and et al. Quantum approximate optimization of non-planar graph problems on a planar superconducting processor. *Nature Physics*, 17(3):332–336, Feb 2021. ISSN 1745-2481. DOI: 10.1038/s41567-020-01105-y.

[22] Abhinav Kandala, Antonio Mezzacapo, Kristan Temme, Maika Takita, Markus Brink, Jerry M. Chow, and Jay M. Gambetta. Hardware-efficient variational quantum eigensolver for small molecules and quantum magnets. *Nature*, 549(7671):242–246, 2017. DOI: 10.1038/nature23879.

[23] Richard M. Karp. *Reducibility among Combinatorial Problems*, pages 85–103. Springer US, Boston, MA, 1972. ISBN 978-1-4684-2001-2. DOI: 10.1007/978-1-4684-2001-2˙9.

[24] Nathan Killoran, Thomas R. Bromley, Juan Miguel Arrazola, Maria Schuld, Nicolás Quesada, and Seth Lloyd. Continuous-variable quantum neural networks. *Phys. Rev. Research*, 1:033063, Oct 2019. DOI: 10.1103/PhysRevResearch.1.033063.

[25] Tjalling C. Koopmans and Martin Beckmann. Assignment problems and the location of economic activities. *Econometrica*, 25(1):53–76, 1957. ISSN 00129682, 14680262. DOI: 10.2307/1907742.

[26] Nathan Lacroix, Christoph Hellings, Christian Kraglund Andersen, Agustin Di Paolo, Ants Remm, Stefania Lazar, Sebastian Krinner, Graham J. Norris, Mihai Gabureac, Johannes Heinsoo, and et al. Improving the performance of deep quantum optimization algorithms with continuous gate sets. *PRX Quantum*, 1(2), Oct 2020. ISSN 2691-3399. DOI: 10.1103/prxquantum.1.020304.

[27] Jose I. Latorre. Image compression and entanglement. *arXiv preprint arXiv:0510031*, 2005.

[28] Wim Lavrijsen, Ana Tudor, Juliane Müller, Costin Iancu, and Wibe de Jong. Classical opti-

mizers for noisy intermediate-scale quantum devices. *arXiv preprint arXiv:2004.03004*, 2020.

[29] Wolfgang Lechner, Philipp Hauke, and Peter Zoller. A quantum annealing architecture with all-to-all connectivity from local interactions. *Science Advances*, 1(9):1–6, 2015. ISSN 23752548. DOI: 10.1126/sciadv.1500838.

[30] Harry Markowitz. Portfolio selection. *The Journal of Finance*, 7(1):77–91, 1952. DOI: 10.2307/2975974.

[31] Sam McArdle, Suguru Endo, Alán Aspuru-Guzik, Simon C. Benjamin, and Xiao Yuan. Quantum computational chemistry. *Rev. Mod. Phys.*, 92:015003, Mar 2020. DOI: 10.1103/RevModPhys.92.015003.

[32] Sam McArdle, Suguru Endo, Alán Aspuru-Guzik, Simon C. Benjamin, and Xiao Yuan. Quantum computational chemistry. *Reviews of Modern Physics*, 92(1), Mar 2020. ISSN 1539-0756. DOI: 10.1103/revmodphys.92.015003.

[33] Jarrod R McClean, Jonathan Romero, Ryan Babbush, and Alán Aspuru-Guzik. The theory of variational hybrid quantum-classical algorithms. *New Journal of Physics*, 18(2):023023, feb 2016. DOI: 10.1088/1367-2630/18/2/023023.

[34] Jarrod R. McClean, Sergio Boixo, Vadim N. Smelyanskiy, Ryan Babbush, and Hartmut Neven. Barren plateaus in quantum neural network training landscapes. *Nature Communications*, 9(1):4812, 2018. DOI: 10.1038/s41467-018-07090-4.

[35] Nikolaj Moll, Panagiotis Barkoutsos, Lev S Bishop, Jerry M Chow, Andrew Cross, Daniel J Egger, Stefan Filipp, Andreas Fuhrer, Jay M Gambetta, Marc Ganzhorn, Abhinav Kandala, Antonio Mezzacapo, Peter Müller, Walter Riess, Gian Salis, John Smolin, Ivano Tavernelli, and Kristan Temme. Quantum optimization using variational algorithms on near-term quantum devices. *Quantum Science and Technology*, 3(3):030503, jun 2018. DOI: 10.1088/2058-9565/aab822.

[36] J. S. Otterbach, R. Manenti, N. Alidoust, A. Bestwick, M. Block, B. Bloom, S. Caldwell, N. Didier, E. Schuyler Fried, S. Hong, P. Karalekas, C. B. Osborn, A. Papageorge, E. C. Peterson, G. Prawiroatmodjo, N. Rubin, Colm A. Ryan, D. Scarabelli, M. Scheer, E. A. Sete, P. Sivarajah, Robert S. Smith, A. Staley, N. Tezak, W. J. Zeng, A. Hudson, Blake R. Johnson, M. Reagor, M. P. da Silva, and C. Rigetti. Unsupervised machine learning on a hybrid quantum computer. *arXiv preprint arXiv:1712.05771*, 2017.

[37] Guido Pagano, Aniruddha Bapat, Patrick Becker, Katherine S. Collins, Arinjoy De, Paul W. Hess, Harvey B. Kaplan, Antonis Kyprianidis, Wen Lin Tan, Christopher Baldwin, and et al. Quantum approximate optimization of the long-range ising model with a trapped-ion quantum simulator. *Proceedings of the National Academy of Sciences*, 117(41): 25396–25401, Oct 2020. ISSN 1091-6490. DOI: 10.1073/pnas.2006373117.

[38] Alberto Peruzzo, Jarrod McClean, Peter Shadbolt, Man-Hong Yung, Xiao-Qi Zhou, Peter J. Love, Alán Aspuru-Guzik, and Jeremy L. O'Brien. A variational eigenvalue solver on a photonic quantum processor. *Nature Communications*, 5(1):4213, 2014. DOI: 10.1038/ncomms5213.

[39] M. Powell. A view of algorithms for optimization without derivatives. *Mathematics TODAY*, 43, 01 2007.

[40] John Preskill. Quantum Computing in the NISQ era and beyond. *Quantum*, 2:79, August 2018. ISSN 2521-327X. DOI: 10.22331/q-2018-08-06-79.

[41] Xiaogang Qiang, Xiaoqi Zhou, Jianwei Wang, Callum M. Wilkes, Thomas Loke, Sean O'Gara, Laurent Kling, Graham D. Marshall, Raffaele Santagati, Timothy C. Ralph, Jingbo B. Wang, Jeremy L. O'Brien, Mark G. Thompson, and Jonathan C. F. Matthews. Large-scale silicon quantum photonics implementing arbitrary two-qubit processing. *Nature Photonics*, 12(9):534–539, 2018. DOI: 10.1038/s41566-018-0236-y.

[42] Arthur G. Rattew, Shaohan Hu, Marco Pistoia, Richard Chen, and Steve Wood. A domain-agnostic, noise-resistant, hardware-efficient evolutionary variational quantum eigensolver. *arXiv preprint arXiv:1910.09694*, 2019.

[43] Troels F. Rønnow, Zhihui Wang, Joshua Job, Sergio Boixo, Sergei V. Isakov, David Wecker, John M. Martinis, Daniel A. Lidar, and Matthias Troyer. Defining and detecting quantum speedup. *Science*, 345(6195):420–424, 2014. ISSN 0036-8075. DOI: 10.1126/science.1252319.

[44] M Saffman. Quantum computing with atomic qubits and rydberg interactions: progress and challenges. *Journal of Physics B: Atomic, Molecular and Optical Physics*, 49(20):202001, oct 2016. DOI: 10.1088/0953-4075/49/20/202001.

[45] Peter W. Shor. Polynomial-time algorithms for prime factorization and discrete logarithms on a quantum computer. *SIAM Journal on Computing*, 26(5):1484–1509, 1997. DOI: 10.1137/S0097539795293172.

[46] Kevin J Sung, Jiahao Yao, Matthew P Harrigan, Nicholas C Rubin, Zhang Jiang, Lin Lin, Ryan Babbush, and Jarrod R McClean. Using models to improve optimizers for variational quantum algorithms. *Quantum Science and Technology*, 5 (4):044008, Oct 2020. ISSN 2058-9565. DOI: 10.1088/2058-9565/abb6d9.

[47] IBM Quantum team. Retrieved from

https://quantum-computing.ibm.com. Ibmq, 2020.

[48] G Wendin. Quantum information processing with superconducting circuits: a review. *Reports on Progress in Physics*, 80(10):106001, sep 2017. DOI: 10.1088/1361-6633/aa7e1a.

[49] Madita Willsch, Dennis Willsch, Fengping Jin, Hans De Raedt, and Kristel Michielsen. Benchmarking the quantum approximate optimization algorithm. *Quantum Information Processing*, 19(7), Jun 2020. ISSN 1573-1332. DOI: 10.1007/s11128-020-02692-8.

[50] Stephen J. Wright. Continuous optimization (nonlinear and linear programming). In Nicholas J. Higham, Mark R. Dennis, Paul Glendinning, Paul A. Martin, Fadil Santosa, and Jared Tanner, editors, *The Princeton Companion to Applied Mathematics*, chapter 4, page 281–293. Princeton University Press, Princeton, 2015.

[51] Leo Zhou, Sheng-Tao Wang, Soonwon Choi, Hannes Pichler, and Mikhail D. Lukin. Quantum approximate optimization algorithm: Performance, mechanism, and implementation on near-term devices. *Physical Review X*, 10(2), Jun 2020. ISSN 2160-3308. DOI: 10.1103/physrevx.10.021067.