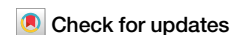


<https://doi.org/10.1038/s42005-025-02072-7>

Speedup of high-order unconstrained binary optimization using quantum \mathbb{Z}_2 lattice gauge theory



Bi-Ying Wang^{1,2}, Xiaopeng Cui³✉, Qingguo Zeng⁴, Yemin Zhan³, Man-Hong Yung^{4,5,6,7}✉ & Yu Shi^{1,3}✉

An important and difficult problem in optimization is the high-order unconstrained binary optimization, which can represent many optimization problems more efficiently than quadratic unconstrained binary optimization, but how to quickly solve it has remained difficult. Here, we present an approach by mapping the high-order unconstrained binary optimization to quantum \mathbb{Z}_2 lattice gauge theory and propose the gauged local quantum annealing, which is the local quantum annealing protected by the gauge symmetry. We present the quantum algorithm and its corresponding quantum-inspired classical algorithm for this problem and achieve algorithmic speedup by using gauge symmetry. By running the quantum-inspired classical algorithm, we demonstrate that the gauged local quantum annealing reduces the computational time by one order of magnitude from that of the local quantum annealing.

High-order unconstrained binary optimization (HUBO) and quadratic unconstrained binary optimization (QUBO) are two high-performance fundamental binary models for the important and wide-range problems of combinatorial optimization. With its simple quadratic-interaction formulation, QUBO has been extensively studied over the past few decades. Numerous classical combinatorial optimization problems, such as Traveling Salesman Problem, Boolean Satisfiability Problem^{1–3}, Maximum Likelihood Detection Problem in the communication technology⁴, error correction based on Low-Density Parity Check, reconfigurable intelligent surfaces beamforming⁵, molecular unfolding⁶ and protein folding problems⁷, as well as the optimal path problem for routes⁸, have been successfully converted to QUBO problems. However, the transformation of these problems to QUBO problems necessitates lots of additional variables^{9–11}, as well as Rosenberg quadratization penalty terms^{12–15}, which increase computational costs and make it challenging for the standard optimizations.

In fact, these problems can be naturally expressed in terms of HUBO problems, where the cost functions are polynomials of orders higher than two. Therefore, instead of transforming them to QUBO problems, solving them in HUBO formulation can reduce both the number of binary variables

and the difficulty in model development, thus saving computational costs^{6,8,16–19}.

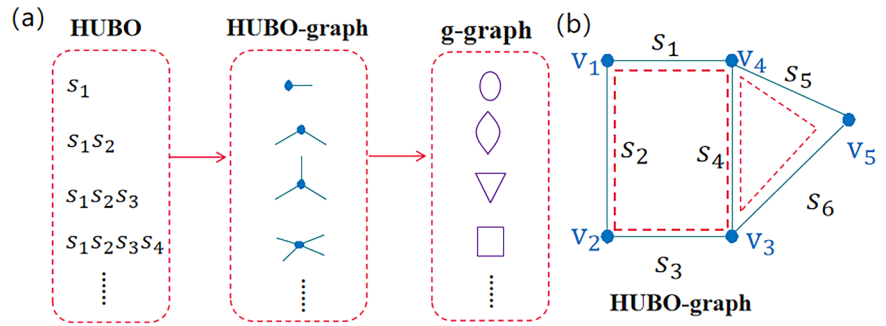
Nevertheless, QUBO still has been used more widely than HUBO. The reason is that in QUBO, the quadratic-interaction formulation for binary variables maps to Ising model²⁰, hence, the tremendous amount of knowledge on the Ising model has been useful in the design of quantum algorithms^{21,22} and quantum-inspired algorithms for QUBO^{23–26}. Therefore, if Ising-like approaches are also established for HUBO, it is hopeful for HUBO to outperform QUBO.

It has been noted that quantum \mathbb{Z}_2 lattice gauge theory (QZ2LGT)^{27–30}, with high-order interactions, can be studied by using the method of quantum simulation^{31,32} and maps to the Hamiltonian Cycle problem³³. On the other hand, gauge symmetry has already been used in quantum error detection, by measuring the conserved quantity with gauge operators without disrupting the quantum evolution^{34–36}.

In this paper, we propose a method to map HUBO to QZ2LGT, which is regarded as a formulation of variable interaction in HUBO, and subsequently leverage the gauge symmetry to improve HUBO solvers. A problem graph of HUBO is constructed³³, and QZ2LGT is defined on the dual graph. Afterwards, based on the gauge symmetry, a speedup scheme similar to

¹Wilczek Quantum Center, Shanghai Institute for Advanced Studies, Shanghai, China. ²Yangtze River Delta Industrial Innovation Center of Quantum Science and Technology, Suzhou, China. ³Department of Physics & State Key Laboratory of Surface Physics, Fudan University, Shanghai, China. ⁴Shenzhen Institute for Quantum Science and Engineering, Southern University of Science and Technology, Shenzhen, China. ⁵International Quantum Academy, Shenzhen, China. ⁶Guangdong Provincial Key Laboratory of Quantum Science and Engineering, Southern University of Science and Technology, Shenzhen, China. ⁷Shenzhen Key Laboratory of Quantum Science and Engineering, Southern University of Science and Technology, Shenzhen, China. ✉e-mail: xpclove@126.com; yung@sustech.edu.cn; yu_shi@ustc.edu.cn

Fig. 1 | Procedure of the mapping process. **a** Four different terms with interaction order varying from 1 to 4 in the objective function of HUBO map to four kinds of vertices in the *HUBO-graph* and four kinds of plaquettes in the *G-graph*. **b** An example of an inefficient cycle in the *HUBO-graph*. In this example, the cycle $s_1s_2s_3s_4s_5$ is inefficient, as it can be broken down into cycles $s_1s_2s_3s_4$ and $s_4s_5s_6$.



quantum Zeno dynamics^{37–39} is introduced for computational speedup. The scheme is also adapted to the corresponding quantum-inspired classical algorithms.

The gauge operators commute with the Hamiltonian. For the quantum adiabatic evolution, the time-dependent state is close to the instantaneous ground state, hence, the measurements of the gauge operators enforce the reduction of the state to instantaneous ground state with high probabilities, as the so-called quantum Zeno effect. This feature is used in our quantum algorithm and the corresponding quantum-inspired classical algorithm. We apply our method to upgrade the local quantum annealing (LQA) to the gauged local quantum annealing (gLQA). In the quantum-inspired classical algorithm, gLQA is a classical feedback process. For comparison, we calculate the ground state energies of the QZ2LGT on a 2D lattice and on a four-regular graph by using LQA, gLQA, and simulated annealing (SA) by leveraging the capabilities of the advanced Python package OpenJij^{40,41}. It is shown that gLQA outperforms LQA, which in turn outperforms SA.

Methods

Mapping HUBO to \mathbb{Z}_2 gauge theory

We first discuss how to map HUBO to QZ2LGT. Then the method to find out the gauge operators of the corresponding QZ2LGT is presented.

The objective of a HUBO is to minimize the classical Hamiltonian

$$H(\mathbf{s}) = \sum_{i_1} J_{i_1} s_{i_1} + \sum_{i_1 < i_2} J_{i_1 i_2} s_{i_1} s_{i_2} + \dots + \sum_{i_1 < i_2 < \dots < i_N} J_{i_1 i_2 \dots i_N} \prod_{j=1}^N s_{i_j}, \quad (1)$$

for $N \geq 3$ with real-number coefficients J 's and $\mathbf{s} \in \{-1, 1\}^N$.

The procedure to map HUBO to QZ2LGT involves two steps. First, we use a graph to describe the HUBO problem, which we refer to as the *HUBO-graph*, in which each edge is occupied by a binary spin s_i of the HUBO, while one vertex represents a term of the HUBO. Second, we map the *HUBO-graph* to its dual, which we refer to as *G-graph*. In this transformation, each edge in the *HUBO-graph* is crossed by one link in the *G-graph*, thus each vertex in the *HUBO-graph* maps to a plaquette in the *G-graph*, while two adjacent vertices in the *HUBO-graph* map to two adjacent plaquettes in the *G-graph*.

To illustrate the procedure, we present the mapping from four different terms with interaction order varying from 1 to 4 in the objective function of HUBO to four kinds of vertices of *HUBO-graph* and four kinds of plaquettes of *G-graph* in Fig. 1(a). As seen from the figure, each term in HUBO problem is represented as a vertex in the *HUBO-graph* and is subsequently mapped to a plaquette in the *G-graph*. Finally, by placing spins at links of this *G-graph*, we build QZ2LGT on the *G-graph* and thus map the original HUBO to QZ2LGT, with the quantum Hamiltonian

$$\hat{H} = \hat{Z} + g\hat{X}, \quad (2)$$

where g is the coupling parameter and $\hat{X} = -\sum_i \hat{\sigma}_x^i$ denotes minus the sum of $\hat{\sigma}_x$ operations on all spins. $\hat{Z} = \sum_p J_p \prod_{i \in p} \hat{\sigma}_z^i$ represents the $\hat{\sigma}_z$ products

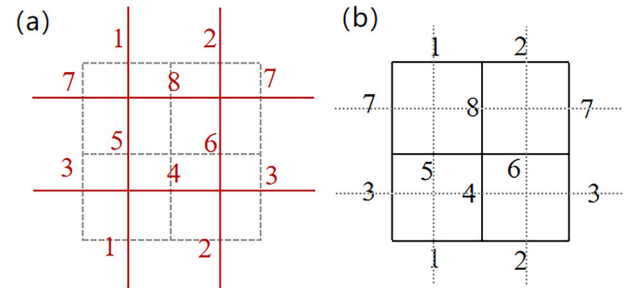


Fig. 2 | An example to illustrate the mapping process. **a** *HUBO-graph* and **(b)** *G-graph* for HUBO with objective in Eq. (4). Both graphs are obtained through the method introduced in the Methods. The solid lines in each figure represent the edges of the graph, while the dashed lines indicate its dual graph. The *G-graph* and the *HUBO-graph* are the dual graphs for each other. The numbers denote the labels of binary spins in the HUBO problem.

operating on the spins of all plaquettes. The gauge operator that commutes with Hamiltonian \hat{H} is defined as the product of $\hat{\sigma}_x$ on all links of each site v

$$\hat{G}_v = \prod_{i \in v} \hat{\sigma}_x^i. \quad (3)$$

The ground state of the Hamiltonian of QZ2LGT is a superposition of product states of definite σ_z of each spin. Each product state is an eigenstate of \hat{Z} , and corresponds to an optimal solution to the classical HUBO problem.

The *G-graph* is complicated, so it is not easy to obtain the gauge operators by directly counting the sites. Since each site on the *G-graph* comes from an efficient cycle on the *HUBO-graph*, we apply the closed-loop search algorithm to identify these cycles directly in the *HUBO-graph*. An efficient cycle is the smallest cycle that cannot be further decomposed into smaller cycles. For instance, in Fig. 1b, the cycle $s_1s_2s_3s_4s_5$ is inefficient, as it can be broken down into cycles $s_1s_2s_3s_4$ and $s_4s_5s_6$, which are efficient. Our algorithm is detailed in the Supplementary Methods.

For the mapping process, an example is given in the following. For a HUBO problem with an objective

$$H = J_1 s_1 s_3 s_5 s_4 + J_2 s_2 s_4 s_6 s_3 + J_3 s_1 s_8 s_5 s_7 + J_4 s_2 s_7 s_6 s_8, \quad (4)$$

with eight variables s_i 's and four real-number coefficients J 's. As introduced above, after mapping each item of the objective function into a vertex, the *HUBO-graph* is obtained as in Fig. 2a. Then, each edge of *HUBO-graph* is crossed by a link, and the links around the same vertex should be connected to define the surrounded region as a plaquette in the *G-graph*, as shown in Fig. 2b. As introduced in Eq. (3), the gauge operators for QZ2LGT defined in the *G-graph* can be obtained by counting the sites. As seen in Fig. 2b, with

periodic boundary condition, there are gauge operators $\hat{G}_1 = \hat{\sigma}_x^4 \hat{\sigma}_x^6 \hat{\sigma}_x^8 \hat{\sigma}_x^5$, $\hat{G}_2 = \hat{\sigma}_x^1 \hat{\sigma}_x^8 \hat{\sigma}_x^2 \hat{\sigma}_x^4$, $\hat{G}_3 = \hat{\sigma}_x^3 \hat{\sigma}_x^5 \hat{\sigma}_x^7 \hat{\sigma}_x^6$ and $\hat{G}_4 = \hat{\sigma}_x^1 \hat{\sigma}_x^2 \hat{\sigma}_x^3 \hat{\sigma}_x^7$.

In this paper, we will first propose a quantum speedup scheme for quantum adiabatic evolution based on the quantum Hamiltonian and then downgrade the quantum Hamiltonian to classical Hamiltonian, and the speedup scheme is extended to a classical one, as the quantum-inspired classical algorithm.

Before presenting our speedup scheme, we review the previous scheme of LQA. In quantum annealing, or adiabatic quantum computing, one considers a system evolving under the time-dependent Hamiltonian

$$\hat{H}(t) = t\gamma\hat{H}_t - (1-t)\hat{H}_x, \quad (5)$$

with γ controlling the fraction of the energy of target Hamiltonian \hat{H}_t in the total Hamiltonian. The system is initially prepared in the state $|+\rangle^{\otimes n}$, which is the ground state of the Hamiltonian $-\hat{H}_x = -\sum_{i=1}^n \hat{\sigma}_x^i$, where n is the number of spins in the system and $\hat{\sigma}_x^i$ is a Pauli operator on the i th spin. The Hamiltonian varies from the initial Hamiltonian $-\hat{H}_x$ at $t=0$ to the target Hamiltonian \hat{H}_t at time $t=1$. If the variation speed of the Hamiltonian is slow enough to meet the adiabatic condition, the state of the system stays at the instantaneous ground state during the evolution, finally reaching the ground state of the target Hamiltonian finally.

In LQA, which is inspired by quantum annealing, one only considers the states of the local form¹⁸

$$|\theta\rangle = |\theta_1\rangle \otimes |\theta_2\rangle \otimes \dots \otimes |\theta_n\rangle. \quad (6)$$

where θ_i denotes the angle between the state of the i th-spin with the z -axis, and is written in the form

$$|\theta_i\rangle = \cos\frac{\theta}{2}|+\rangle + \sin\frac{\theta}{2}|-\rangle. \quad (7)$$

The cost function of the LQA is defined as

$$C(t, \theta) = \langle \theta | \hat{H}(t) | \theta \rangle, \quad (8)$$

which can be written as a function of variable θ . A variable $w_i \in \mathbb{R}$ is used to parameterize θ_i as $\theta_i = \frac{\pi}{2} \tanh w_i$, in order to limit the range of θ_i to be $\theta_i \in [-\frac{\pi}{2}, \frac{\pi}{2}]$. As a result, the cost function $C(t, \mathbf{w})$ is expressed as a function of variable $w_i (i=1, \dots, n)$ and time t .

In the corresponding quantum-inspired classical scheme, with the time discretized as $t_j = j/N_{\text{iter}} (j=1, \dots, N_{\text{iter}})$, an iteration of variables w_i , depending on the gradient of cost function is performed, $\mathbf{v} \leftarrow \mu \mathbf{v} - \eta \nabla_{\mathbf{w}} C(\mathbf{w}, t)$, $\mathbf{w} \leftarrow \mathbf{w} + \mathbf{v}$, where \mathbf{v} is an additional vector representing the updating speed of \mathbf{w} , $\mu \in [0, 1]$ and η are parameters¹⁸. After the iteration, the final spin configuration can be obtained as $s_i = \text{sign}(w_i)$.

Our speedup scheme and the corresponding classical scheme

During an adiabatic evolution, as the evolution is not infinitesimally slow, the state $|\psi(t)\rangle$ deviates from the instantaneous ground state $|\psi_g(t)\rangle$. A gauge operator \hat{G} commutes the Hamiltonian $\hat{H}(t)$ at any time t , and it also commutes both $\hat{H}(0) = \hat{H}_x$ and $\hat{H}(1) = \hat{H}_t$. Therefore $|\psi_g(t)\rangle$ is also an eigenstate of \hat{G} ,

$$\hat{G}|\psi_g(t)\rangle = a|\psi_g(t)\rangle, \quad (9)$$

with constant a .

Our speedup scheme for quantum adiabatic evolution is the following. A measurement of the gauge operator \hat{G} is made on the state $|\psi(t)\rangle$, and a correction is subsequently made according to the measurement result G , to force the system to its ground state. This process is repeated until the state $|\psi(t)\rangle$ becomes the ground state $|\psi_g(t)\rangle$.

In quantum simulation with a finite step size, the adiabatic condition is not strictly met. Gauge symmetric measurements can be employed to safeguard the adiabatic process. For a model with gauge symmetry, which should be preserved throughout the entire process, a measurement of the gauge operator does not interfere with the evolution. Thus, under the gauge symmetric measurements, the adiabaticity can be maintained within fewer steps, leading to speedup.

The above gauge-protected scheme for quantum adiabatic evolution can be extended to a classical speedup scheme with similar protection. After transforming the problem to the classical binary optimization, the optimization Hamiltonian $H(\mathbf{s})$ and symmetry operator $G(\mathbf{s})$ can be described in terms of classical spin configuration $\mathbf{s} \equiv \{s_i\}$. In our speedup scheme, a symmetry-forced operation based on gradient is made, on the spin configuration \mathbf{s} in every step of the algorithm,

$$s_i \leftarrow s_i - B[G(\mathbf{s}) - a] \frac{\partial G(\mathbf{s})}{\partial s_i}, \quad (10)$$

where B is a parameter controlling the evolution speed. This is a classical feedback process inspired by quantum measurement and the quantum Zeno effect.

As an example of the above general scheme, we now present the formulation of gLQA by introducing gauge symmetry into LQA. During the quantum annealing process, the state is always the instantaneous ground state. Since each gauge operator commutes with the system Hamiltonian, the state in the adiabatic evolution is also an eigenstate of each gauge operator. Thus, in LQA,

$$\hat{G}_{v_i}|\theta\rangle = |\theta\rangle, \quad (11)$$

where v_i represents a vertex of link i . As an example of the above classical feedback (10), after obtaining the localized classical formula of the gauge operator G_{v_i} , an additional gradient-based iteration generated from the gauge operator is applied to force the state to respect the gauge symmetry,

$$w_i \leftarrow w_i + B \sum_{v_i} (G_{v_i} - 1) \frac{\partial G_{v_i}}{\partial w_i}, \quad (12)$$

where v_i represents the sites on link i , and B is a constant. By replacing G_{v_i} with $\prod_{l \in v_i} x_l$, where $x_l \in [-1, 1]$, the iteration for the gLQA becomes

$$\begin{aligned} \nu_i &\leftarrow \mu \nu_i - \eta \nabla_{w_i} C(\mathbf{w}, t) \\ w_i &\leftarrow w_i + \nu_i \\ w_i &\leftarrow w_i - B \sum_{v_i} \left(\prod_{l \in v_i} x_l - 1 \right) \frac{\partial \prod_{l \in v_i} x_l}{\partial w_i}. \end{aligned} \quad (13)$$

Results

As introduced in the Methods, a HUBO task can be mapped to the calculation of the ground state and its energy of QZ2LGT on a G -graph. In order to benchmark our algorithm, we apply LQA and gLQA to calculate the ground energies of QZ2LGT on two kinds G -graph, namely, 2D square lattice and the dual graph of a random four-regular graph. In the first subsection, the structures of the two graphs are introduced. Then, the calculation results of each graph are presented, which are performed on an Intel® CPU i7-8700 operating at a frequency of 3.2GHz. To benchmark the speed of algorithms, we introduce a standard merit – time to solution (TTS), which is defined as the computation time of finding an optimal value or solution with 99% probability^{42,43}. TTS can be calculated through the real

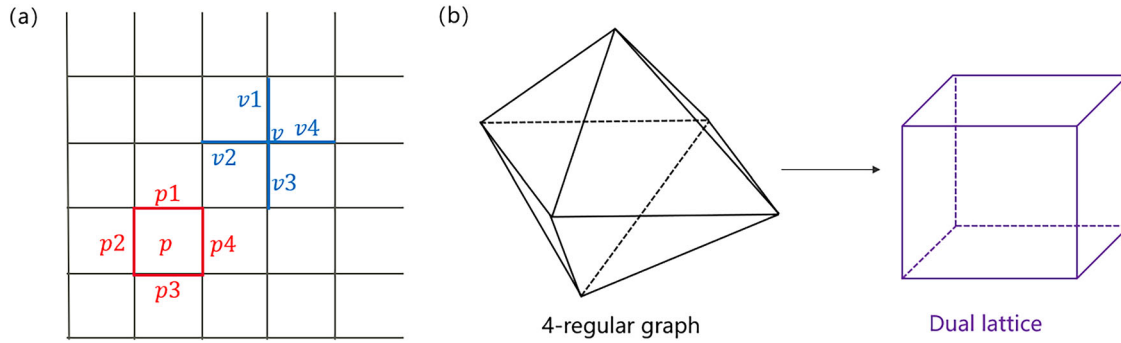


Fig. 3 | Examples to illustrate the dual graphs. a The structure of a square lattice. The spins $p_i (i = 1, 2, 3, 4)$ represent the spins located on the edges of a plaquette, while the spins $v_i (i = 1, 2, 3, 4)$ correspond to the spins connected to a vertex. **b** The G -graph obtained as the dual lattice of a four-regular graph.

computation time t_p as

$$\text{TTS} = t_p \frac{\log(1 - 0.99)}{\log(1 - p)}, \quad (14)$$

where success probability

$$p \equiv \frac{n_{\text{sol}}}{n_{\text{sam}}} \quad (15)$$

denotes the ratio between the number of times n_{sol} achieving the ground state energy and the total number of sampling times n_{sam} with the corresponding computation time t_p .

2D lattice and four-regular graph

On a square lattice of size $L \times L$, as a dual graph, there are $2L^2$ links and L^2 plaquettes, and one spin locates at each link, so altogether there are $N = 2L^2$ spins, $N_p = L^2$ plaquettes and $N_v = L^2$ vertices. The lattice is depicted in Fig. 3(a). The Hamiltonian of \mathbb{Z}_2 lattice gauge theory is as in Eq. (2). As seen from the figure, each magnetic term contains $\hat{\sigma}_z$ operating on the four spins ($p1, p2, p3, p4$) of one plaquette with $\hat{Z} = -\sum_p \hat{\sigma}_z^{p1} \hat{\sigma}_z^{p2} \hat{\sigma}_z^{p3} \hat{\sigma}_z^{p4}$, while the gauge operator \hat{G}_v contains four links of site v and can be written as $\hat{G}_v = \hat{\sigma}_x^{v1} \hat{\sigma}_x^{v2} \hat{\sigma}_x^{v3} \hat{\sigma}_x^{v4}$.

We generate a random G -graph by finding the dual lattice of a random four-regular graph as presented in Fig. 3b. As seen from the figure, every edge in the four-regular graph maps to one link in the G -graph which crosses the edge, and every vertex or site in the four-regular graph maps to a plaquette in the G -graph, an efficient cycle in the four-regular graph maps to a site in the G -graph. As explained in the Methods, by counting all the vertices, the operator \hat{Z} and gauge operator \hat{G}_v of the G -graph can be obtained. In the four-regular graph, each vertex has four connected edges, which consequently lead to four links $\{p1, p2, p3, p4\}$ in every plaquette p of the obtained G -graph. Thus, the \hat{Z} operator can be written as $\hat{Z} = -\sum_p \hat{\sigma}_z^{p1} \hat{\sigma}_z^{p2} \hat{\sigma}_z^{p3} \hat{\sigma}_z^{p4}$. Due to the uncertain length of the cycle in a randomly generated four-regular graph, the number of links in one site s of G -graph is uncertain. For convenience, in the following, we only collect the gauge operators for which the number of spins included is smaller than a threshold k_m for the G -graphs generated by four-regular graphs.

Numerical results for 2D lattices

In this subsection, we present our gLQA results of optimization for a square lattice, as the dual graph. We first discuss the results for the lattice size $L = 10$, then present the results for a range of lattice sizes from 10 to 40. For comparison, we also include the results of LQA optimization. For $L = 10$, optimization from SA is also presented.

A lattice contains $N_p = L^2$ plaquettes, so the ground state energy is $E_s = -N_p$. For $L = 10$, the minimum energy E_{min} and median energy E_{med} , which is the median value of the energy over all the samples, as functions of the number of iteration steps n_{iter} for sampling number $n_{\text{sam}} = 200$, are

depicted in Fig. 4a and b, respectively. As can be seen from Fig. 4a and b, though E_{min} and E_{med} obtained from SA (green curve in the figures) exhibit rapid decreases and subsequent saturations with the increase of iteration steps n_{iter} , they fail to converge to the ground state values within 200 times of sampling. As seen from Fig. 4a, as n_{iter} increases, E_{min} obtained from LQA and that from gLQA rapidly decrease and saturate towards the ground state energy $E_s = -100$ within 60 steps of iteration. Moreover, E_{min} obtained from gLQA reaches the ground state energy quicker than from LQA. As seen from Fig. 4b, similar behavior can be observed for E_{med} calculated from LQA and that from gLQA, which decreases rapidly before gradually slows down until saturation. Besides, E_{med} obtained from gLQA is smaller than that from LQA. The probability p and TTS as functions of iteration step n_{iter} are displayed in Fig. 4c and d, respectively. As seen from Fig. 4c, the probability p calculated from gLQA and that from LQA gradually increase with n_{iter} , leading to the decrease of TTS in the Fig. 4d. For $n_{\text{iter}} \geq 500$, the probability p calculated in gLQA saturates to a high value ~ 0.81 , inducing a slow increase of TTS with n_{iter} . The probability p calculated in LQA saturates to $p \sim 0.35$ for $n_{\text{iter}} \geq 1000$. It is clearly seen in Fig. 4d that gLQA achieves shorter TTS than LQA. Notably, the shortest TTS of approximately 0.15 s in gLQA for this lattice occurs at $n_{\text{iter}} = 500$, which is less than a quarter of the shortest TTS of approximately 0.60 s for LQA, demonstrating a fourfold increase of the speed by our gLQA.

The results with $n_{\text{sam}} = 1000$ for lattices with various values of a number of spins N (L ranging from 10 to 40) are illustrated in Fig. 5. Since the results obtained from SA cannot reach the ground state for a small lattice with $L = 10$, we only compare LQA and gLQA. With the increase of the number of spins N , for the iteration step fixed at 1000, the success probability p in Fig. 5a decreases, which consequently leads to the increased TTS in Fig. 5c. To make a clear comparison, the N dependence of ratio $r_p (r_{\text{TTS}})$ between the success probability (TTS) calculated in our gLQA and that in LQA is presented in Fig. 5b (Fig. 5d). The figures clearly indicate that our gLQA method achieves at least a twofold increase in success probability p and a reduction of TTS by 60%.

Results for G -graphs dual to four-regular graphs

In this subsection, we present our results for G -graphs generated as the dual from four-regular graphs. We first discuss the results for a G -graph with $N = 400$. Then, the outcomes for N ranging from 400 to 1600 are presented.

The success probability and TTS are calculated as functions of the number of iteration steps n_{iter} for $n_{\text{sam}} = 2000$ samplings of a random G -graph with 400 links, as shown in Fig. 6a and c. As seen from Fig. 6a, with the increase of n_{iter} , the success probability p in LQA and that in gLQA both start to increase gradually from 0 for $n_{\text{iter}} > 100$. Then the success probability p in gLQA saturates around 0.25 for $n_{\text{iter}} \geq 500$, while p in LQA only reaches ~ 0.05 for $n_{\text{iter}} = 800$. As $p = 0$ for $n_{\text{iter}} < 100$, TTS in this regime is meaningless. As seen from Fig. 6c, with the increase of n_{iter} , TTS' in both LQA and gLQA first decrease, because of the increase of p . It is notable that the shortest TTS in LQA is around 11s, which is more than fivefold of the one in

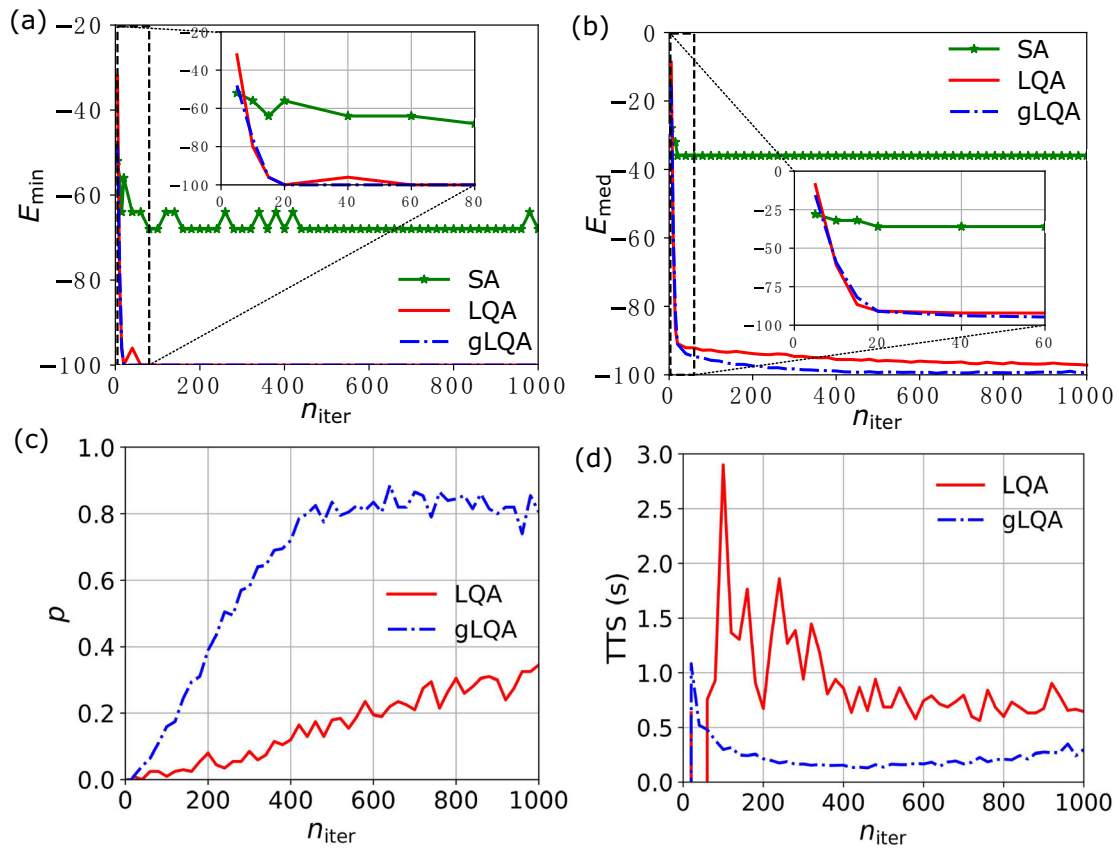


Fig. 4 | Calculation results for sampling number $n_{\text{sam}} = 200$ of a square lattice with size $L = 10$, as functions of iteration steps n_{iter} , obtained from three algorithms (LQA, gLQA and SA). **a Minimum energy E_{min} . **b** Medium energy E_{med} .**

c Success probability p . **d** Time to solution (TTS). The insets in (a) and (b) provide an enlarged view of the data for small iteration steps in the corresponding graph.

Fig. 5 | Results for square lattices as functions of number of spins N . **a Success probability p . **b** Success possibility ratio r_p of algorithm gLQA to algorithm LQA. **c** Time to solution TTS. **d** Time-to-solution ratio r_{TTS} of algorithm gLQA to algorithm LQA. The red bold curves and blue dotted-dashed curves in (a) and (c) denote the results from the LQA and gLQA algorithms, respectively. The green bold curves in (b) and (d) represent the ratio of results between gLQA and LQA.**

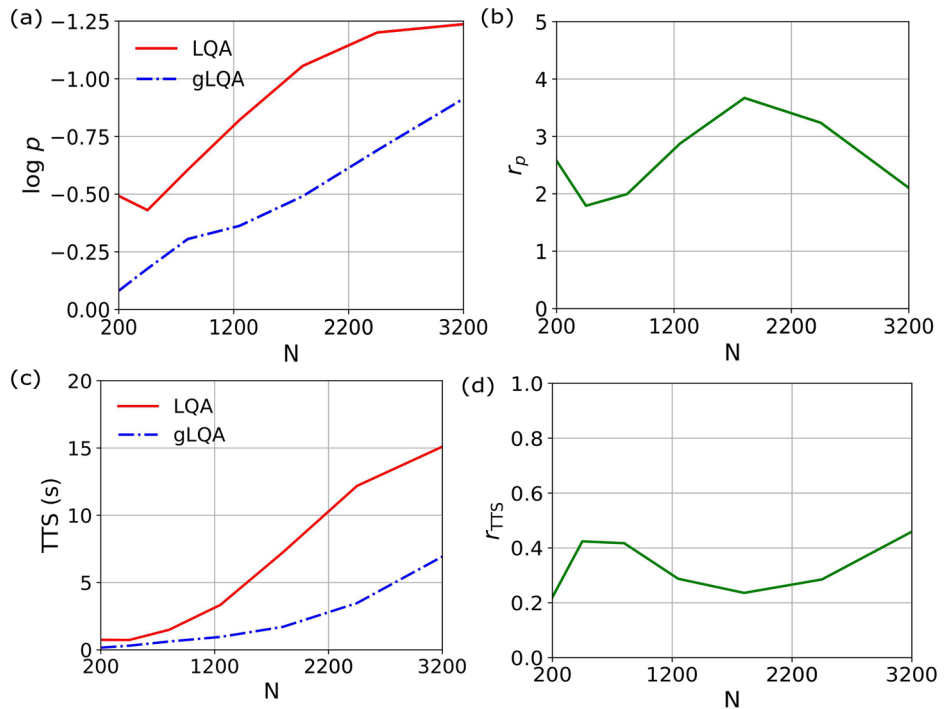


Fig. 6 | Results for the G -graph as the dual of a random four-regular graph with number of spins $N = 400$, as functions of iteration steps n_{iter} . **a Success probability p . **b** Success possibility ratio r_p of algorithm gLQA to algorithm LQA. **c** Time to solution TTS. **d** Time-to-solution ratio r_{TTS} of algorithm gLQA to algorithm LQA. The red bold curves and blue dotted-dashed curves in (a) and (c) denote the results obtained from the LQA and gLQA algorithms, respectively. The green bold curves in (b) and (d) represent the ratio of results between gLQA and LQA.**

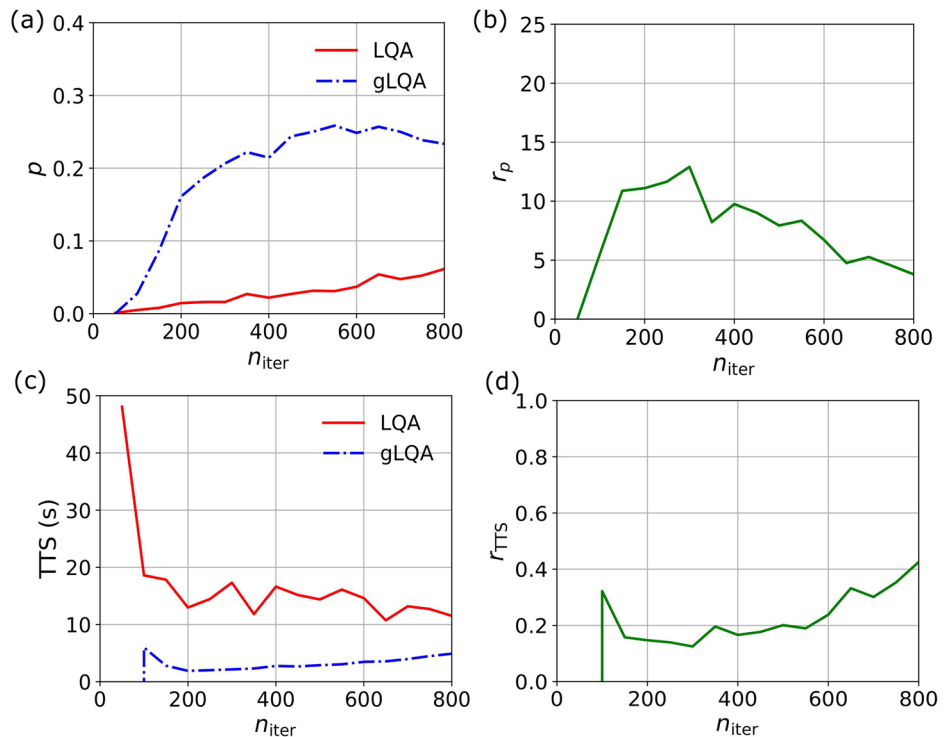
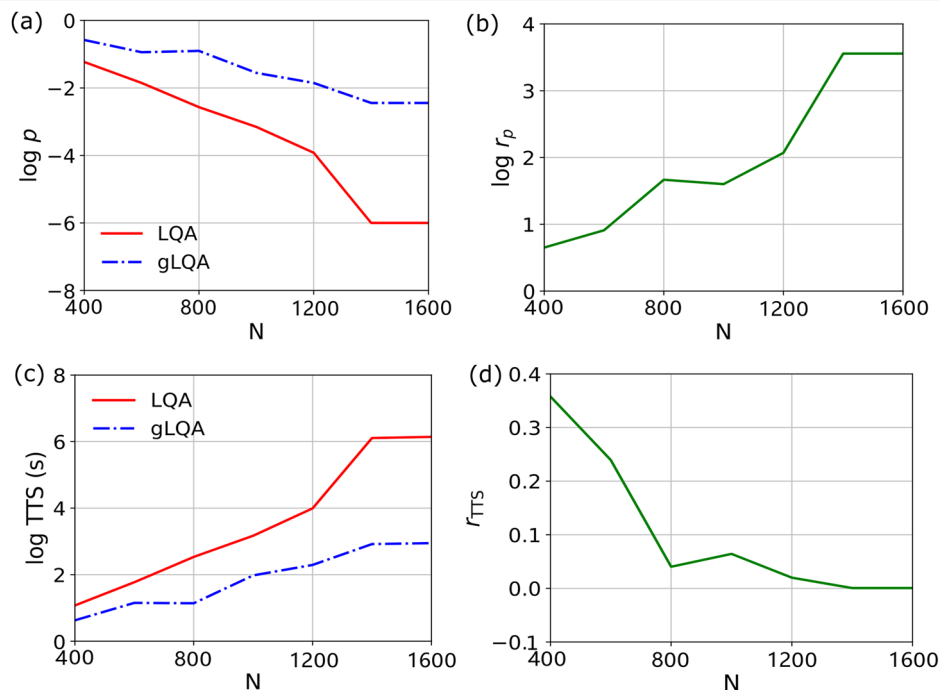


Fig. 7 | Result for the G -graphs as the duals of random four-regular graphs, as functions of the number of spins N . **a Success probability p . **b** Success possibility ratio r_p of algorithms gLQA to LQA. **c** Time to solution (TTS). **d** Time-to-solution ratio r_{TTS} of algorithms gLQA to LQA. The red bold curves and blue dotted-dashed curves in (a) and (c) denote the results obtained from the LQA and gLQA algorithms, respectively. The green bold curves in (b) and (d) represent the ratio of results between gLQA and LQA.**



our gLQA, which is ~ 2 s. The ratios r_p and r_{TTS} as functions of n_{iter} are shown in the Fig. 6b and d, respectively. Fig. 6b displays at least a fourfold increase of the success probability by our gLQA, which reaches thirteenfold at $n_{\text{iter}} = 300$. As shown in Fig. 6d, TTS in our gLQA has a $\sim 60\%$ reduction compared with the one from LQA, which reaches $\sim 85\%$ at $n_{\text{iter}} = 300$.

The results with $n_{\text{sam}} = 100000$ for G -graph generated from four-regular graphs with the number of spins N ranging from 400 to 1600 are presented in Fig. 7. With the increase of N , for the iteration step kept constant at 500, the success probability p decreases, as shown in Fig. 7a,

and consequently leads to the increase of TTS, as shown in Fig. 7c. Besides, a sudden increase of TTS in LQA appears for $N > 1200$ due to the sudden decrease of near-zero success probability p . To make a clear comparison, N dependences of ratio r_p and r_{TTS} are presented in Fig. 7b and d. The figures clearly indicate that our gLQA method achieves an obvious increase in p , while displays a reduction of TTS larger than 75%. For large G -graph with $N > 1200$, the increase of success probability for our gLQA can be larger than 3000 and the reduction of TTS can reach three orders of magnitude.

Conclusions

In this paper, we first presented a quantum algorithm for HUBO by mapping it to a graph problem, which is subsequently transformed to QZ2LGT defined on the dual graph, referred to as the *G-graph*. The HUBO problem is thus transformed into the problem of finding the ground state and its energy in QZ2LGT. The gauge operators commute with the Hamiltonian, hence, their measurements enforce the state to be in the ground state during its evolution, leading to the speedup of the adiabatic algorithm.

Then we presented the corresponding quantum-inspired classical algorithm facilitated by the gauge symmetry and have introduced a gauge-forced iteration step to further speed up the computation.

We have demonstrated the advantage of our method by introducing gLQA in the quantum algorithm and the corresponding quantum-inspired algorithm.

The benchmarking analyses have been made on the quantum-inspired algorithms using LQA and gLQA, on two types of *G-graphs* with spin numbers ranging from 200 to 3200. Our benchmarking analyses of TTS based on LQA and gLQA demonstrated that gLQA significantly outperforms LQA. For comparison, we have also applied SA to solve these problems, which fails even for a small 2D lattice of size $L = 10$.

We have successfully identified \mathbb{Z}_2 gauge theory as suitable for HUBO and leveraged gauge symmetry to expedite the solution process.

There are several potential avenues for future developments. First, it would be beneficial to test our speedup scheme on a wider range of instances and develop an automated technique for tuning the parameters in our scheme, such as constant B and the time step, which are determined through preliminary searches now.

Moreover, the proposed gauge-symmetry-protected scheme has the potential to accelerate computation for all algorithms based on quantum adiabatic theory. It is also interesting to explore whether other features of QZ2LGT are useful in this scheme.

Data availability

The data generated and/or analyzed in this work are available from the corresponding author upon reasonable request.

Code availability

The codes generated in this work are available from the corresponding author upon reasonable request.

Received: 29 January 2024; Accepted: 27 March 2025;

Published online: 11 April 2025

References

- Lucas, A. Ising formulations of many np problems. *Front. Phys.* **2**, 4 (2014).
- Kalinin, K. P. & Berloff, N. G. Computational complexity continuum within Ising formulation of NP problems. *Commun. Phys.* **5**, 20 (2022).
- Mohseni, N., McMahon, P. L. & Byrnes, T. Ising machines as hardware solvers of combinatorial optimization problems. *Nat. Rev. Phys.* **4**, 363–379 (2022).
- Kim, M., Venturelli, D. & Jamieson, K. Leveraging quantum annealing for large mimo processing in centralized radio access networks. In *Proceedings of the ACM Special Interest Group on Data Communication*, 241–255 (2019).
- Ross, C., Gradoni, G., Lim, Q. J. & Peng, Z. Engineering reflective metasurfaces with ising hamiltonian and quantum annealing. *IEEE Trans. Antennas Propag.* **70**, 2841–2854 (2022).
- Mato, K., Mengoni, R., Ottaviani, D. & Palermo, G. Quantum molecular unfolding. *Quantum Sci. Technol.* **7**, 035020 (2022).
- Robert, A., Barkoutsos, P. K., Woerner, S. & Tavernelli, I. Resource-efficient quantum algorithm for protein folding. *npj Quantum Inf.* **7**, 38 (2021).
- Glos, A., Krawiec, A. & Zimborás, Z. Space-efficient binary optimization for variational quantum computing. *npj Quantum Inf.* **8**, 39 (2022).
- Xia, R., Bian, T. & Kais, S. Electronic structure calculations and the Ising Hamiltonian. *J. Phys. Chem. B* **122**, 3384–3395 (2017).
- Boykov, Y. & Kolmogorov, V. An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. *IEEE Trans. Pattern Anal. Mach. Intell.* **26**, 1124–1137 (2004).
- Fujisaki, J., Oshima, H., Sato, S. & Fujii, K. Practical and scalable decoder for topological quantum error correction with an ising machine. *Phys. Rev. Res.* **4**, 043086 (2022).
- Rosenberg, I. G. Reduction of bivalent maximization to the quadratic case (1975).
- Ayodele, M. Penalty weights in QUBO formulations: Permutation problems. 159–174 (Springer International Publishing).
- García, M. D., Ayodele, M. & Moraglio, A. Exact and sequential penalty weights in quadratic unconstrained binary optimisation with a digital annealer. GECCO '22, 184–187 (Association for Computing Machinery, New York, NY, USA, 2022). <https://doi.org/10.1145/3520304.3528925>.
- Verma, A. & Lewis, M. Penalty and partitioning techniques to improve performance of qubo solvers. *Discret. Optim.* **44**, 100594 (2022).
- Blondel, M., Fujino, A., Ueda, N. & Ishihata, M. Higher-order factorization machines. NIPS'16, 3359–3367 (Curran Associates Inc., Red Hook, NY, USA, 2016).
- Ide, N., Asayama, T., Ueno, H. & Ohzeki, M. Maximum likelihood channel decoding with quantum annealing machine. In *2020 International Symposium on Information Theory and Its Applications (ISITA)*, 91–95 (IEEE, 2020).
- Bowles, J., Dauphin, A., Huembeli, P., Martinez, J. & Acín, A. Quadratic unconstrained binary optimization via quantum-inspired annealing. *Phys. Rev. A* **18**, 034016 (2022).
- Norimoto, M., Mori, R. & Ishikawa, N. Quantum algorithm for higher-order unconstrained binary optimization and mimo maximum likelihood detection. *IEEE Trans. Commun.* **71**, 1926–1939 (2023).
- Barahona, F. On the computational complexity of Ising spin glass models. *J. Phys. A: Math. Gen.* **15**, 3241 (1982).
- Ebadi, S. et al. Quantum optimization of maximum independent set using Rydberg atom arrays. *Science* **376**, 1209–1215 (2022).
- Graham, T. et al. Multi-qubit entanglement and algorithms on a neutral-atom quantum computer. *Nature* **604**, 457–462 (2022).
- Goto, H., Tatsumura, K. & Dixon, A. R. Combinatorial optimization by simulating adiabatic bifurcations in nonlinear hamiltonian systems. *Sci. Adv.* **5**, eaav2372 (2019).
- Reifenstein, S., Kako, S., Khoiratee, F., Leleu, T. & Yamamoto, Y. Coherent ising machines with optical error correction circuits. *Adv. Quantum Technol.* **4**, 2100077 (2021).
- Leleu, T. et al. Scaling advantage of chaotic amplitude control for high-performance combinatorial optimization. *Commun. Phys.* **4**, 266 (2021).
- McGeoch, F. P., C. The d-wave advantage system: an overview. Tech. Rep., D-Wave Systems Inc, Burnaby, BC, Canada (2020).
- Sachdev, S. Topological order, emergent gauge fields, and Fermi surface reconstruction. *Rep. Prog. Phys.* **82**, 014001 (2019).
- Wen, X. G. Mean-field theory of spin-liquid states with finite energy gap and topological orders. *Phys. Rev. B* **44**, 2664–2672 (1991).
- Wen, X.-G. Quantum orders and symmetric spin liquids. *Phys. Rev. B* **65**, 165113 (2002).
- Hamma, A. & Lidar, D. A. Adiabatic preparation of topological order. *Phys. Rev. Lett.* **100**, 030502 (2008).
- Cui, X., Shi, Y. & Yang, J.-C. Circuit-based digital adiabatic quantum simulation and pseudoquantum simulation as new approaches to lattice gauge theory. *J. High. Energy Phys.* **2020**, 160 (2020).

32. Cui, X. & Shi, Y. Trotter errors in digital adiabatic quantum simulation of quantum \mathbb{Z}_2 lattice gauge theory. *Int. J. Mod. Phys. B* **34**, 2050292 (2020).
33. Cui, X. & Shi, Y. Correspondence between Hamiltonian Cycle Problem and the Quantum \mathbb{Z}_2 Lattice Gauge Theory. *EPL* **144**, 48001 (2023).
34. Kitaev, A. Fault-tolerant quantum computation by anyons. *Ann. Phys.* **303**, 2–30 (2003).
35. Zhang, H. et al. Encoding error correction in an integrated photonic chip. *PRX Quantum* **4**, 030340 (2023).
36. Sivak, V. V. et al. Real-time quantum error correction beyond break-even. *Nature* **616**, 50–55 (2023).
37. Herman, D. et al. Constrained optimization via quantum zeno dynamics. *Commun. Phys.* **6**, 219 (2023).
38. Schäfer, F. et al. Experimental realization of quantum zeno dynamics. *Nat. Commun.* **5**, 3194 (2014).
39. Facchi, P., Marmo, G. & Pascazio, S. Quantum zeno dynamics and quantum zeno subspaces. *J. Phys.: Conf. Ser.* **196**, 012017 (2009).
40. S. K. et al. Optimization by simulated annealing. *Science* **220**, 671 (1983).
41. The Comapny Jij. OpenJij. Software <https://github.com/OpenJij/OpenJij> (2024).
42. Hamerly, R. et al. Experimental investigation of performance differences between coherent ising machines and a quantum annealer. *Sci. Adv.* **5**, eaau0823 (2019).
43. Aramon, M. et al. Physics-inspired optimization for quadratic unconstrained problems using a digital annealer. *Front. Phys.* **7**, 48 (2019).

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grant no. 12075059).

Author contributions

B.Y.W., X.C., and Y.S. conceived the work and formulated the theoretical framework. B.Y.W. performed the numerical calculations and analyzed the data with useful discussions with X.C., Y.S., and M.H.Y. B.Y.W., X.C., Q.Z., Y.Z., and Y.S. prepared the manuscript. All authors contributed to analyzing the data, discussing the results and commented on the writing.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s42005-025-02072-7>.

Correspondence and requests for materials should be addressed to Xiaopeng Cui, Man-Hong Yung or Yu Shi.

Peer review information *Communications Physics* thanks the anonymous reviewers for their contribution to the peer review of this work. A peer review file is available.

Reprints and permissions information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025