

Quantum sampling on a quantum annealer for large volumes in the strong coupling limit for gauge group $U(3)$

Jangho Kim^{1,2,3,*}, Thomas Luu^{1,†} and Wolfgang Unger^{2,‡}

¹*Institute for Advanced Simulation (IAS-4) and JARA-HPC, Forschungszentrum Jülich, 52425 Jülich, Germany*

²*Fakultät für Physik, Bielefeld University, D-33615 Bielefeld, Germany*

³*Lattice Gauge Theory Research Center, Department of Physics and Astronomy, Seoul National University, Seoul 08826, Republic of Korea*



(Received 19 December 2024; accepted 31 March 2025; published 6 May 2025)

In our previous studies [Phys. Rev. D **108**, 074501 (2023)., *Proc. Sci.* LATTICE2023 (2024) 219 [arXiv:2311.07209].], we confirmed that a quantum annealer can be used for importance sampling of gauge theories. In this paper, we extend the previous results to larger 2-dimensional and 4-dimensional lattices to generate ensembles for $U(3)$ gauge theory in the strong coupling limit. We make use of the D-Wave quantum annealer to generate histograms for sublattices, and use the Metropolis-Hastings algorithm to determine thermodynamic observables and their dependence on the physical parameters on large volumes. We benchmark our results to those obtained from classical Monte Carlo simulations.

DOI: 10.1103/PhysRevD.111.094504

I. INTRODUCTION

Quantum annealers are promising devices for ensemble generation in statistical physics and field theory. The examples of quantum sampling range from boson sampling [1] to sampling proteins [2]. Also, many different methods are on the market for quantum annealers: the noisy Gibbs sampler [3], methods to investigate phase transition [4], hybrid approach [5], and the annealer as a quantum thermal sampler [6]. Quantum sampling of statistical ensembles requires however an additional step to obtain ensembles according to the target equilibrium distribution that depends on the physical parameters of the investigated theory [7–9]. In contrast to gate-based quantum computing, quantum annealers are restricted to optimization problems, and a crucial step is to map the physical system on the qubits of the annealer. This embedding poses a scalability challenge [10,11], which limits the usefulness for realistic applications further. Despite these challenges, some progress in high energy physics such as scalar field theory [12], lattice gauge theory with non-Abelian gauge groups $SU(2)$ [13], dihedral group D_3 [14], and for building blocks of standard model

physics [15] has been made. We address in this paper also a non-Abelian gauge group $U(3)$, which at strong coupling has similar properties as $SU(3)$ at vanishing quark density, as explained in Sec. II.

On a classical computer, via a Monte Carlo, importance sampling is realized e.g. by the Metropolis algorithm. The effective theory which we study is based on the dual formulation of lattice QCD at strong coupling, and is governed by a constraint on the dual variables, which restricts admissible configurations. Depending on the physical parameters, ensemble generation is a difficult task, specifically at low temperature where the entropy of the system grows rapidly. In contrast, on a quantum annealer, the admissible configurations can be generated easily with the so-called quadratic unconstrained binary optimization (QUBO) formalism which scales independently of the physical parameters such as the temperature, as explained in Sec. II.

The scope of this paper is to address larger volumes compared to our previous studies [16,17], which requires new strategies to take advantage of the quantum annealer. We use the D-Wave advantage system with the Pegasus topology [18] that has 5760 physical qubits and is the most advanced annealer up to date. In Fig. 1(a) we show its connectivity. This system can however only accommodate small lattice volumes. Large volumes, on the other hand, require an alternative scheme that samples smaller sublattices. The central idea of our work is to make use of an iterative scheme that embeds sublattices in parallel with specific boundaries that are fixed by the previous quantum sampling step. This scheme ultimately allows us to generate larger volume configurations.

* Contact author: jangho@physik.uni-bielefeld.de

† Contact author: t.luu@fz-juelich.de

‡ Contact author: wunger@physik.uni-bielefeld.de

Published by the American Physical Society under the terms of the [Creative Commons Attribution 4.0 International license](#). Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI. Funded by SCOAP³.

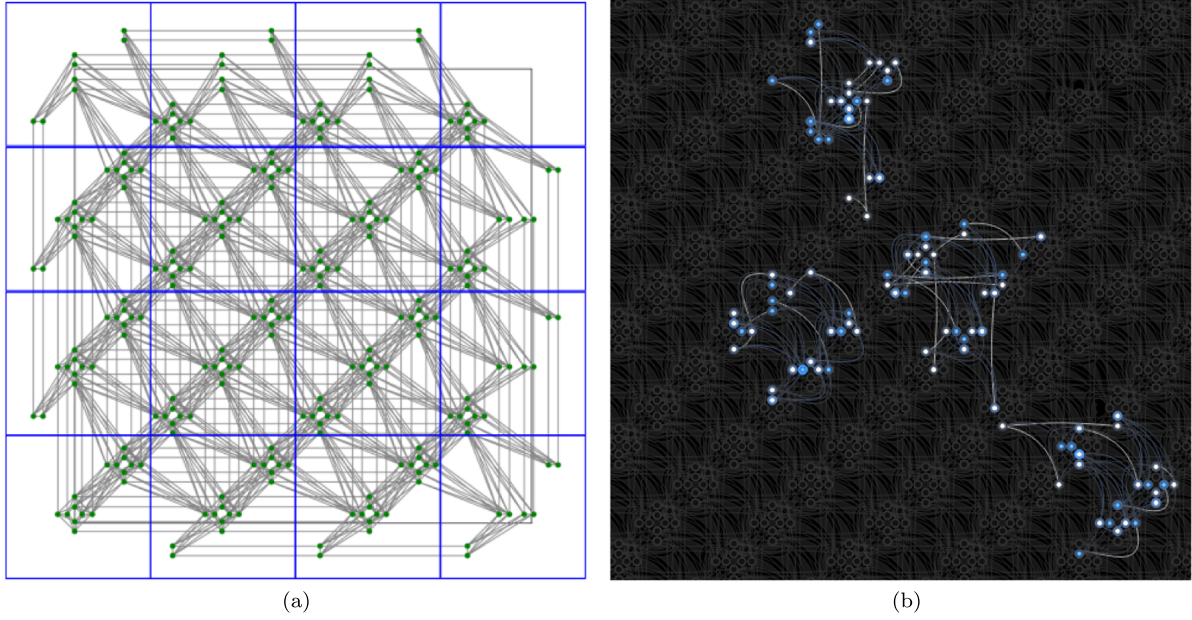


FIG. 1. (a) The connectivity between qubits of the Pegasus topology, showing a part of the full 5760 qubits. Qubits in this topology are considered to have a nominal length of 12 (each qubit is connected to 12 orthogonal qubits through internal couplers) and degree of 15 (each qubit is coupled to 15 different qubits) [19]. (b) Embedding of sublattices in parallel, here shown for a subset of 4, using automatic embedding, illustrating that each sublattice is disconnected. Each sublattice uses 24 physical qubits. Each monomer and dimer needs 2 logical qubits to represent integers between 0 and 3. (a) Pegasus topology. (b) Embedding of 4 sub-lattices.

After short introduction about the effective theory in Sec. II, we explain the technical aspects to address large volumes in Sec. III. We present our results on the volume and parameter dependence of observables in Sec. IV.

II. EFFECTIVE THEORY OF LATTICE QCD

Our effective theory is derived from staggered fermions in the strong coupling limit, which is well known to have a dual representation in terms of integer-valued dual variables:

$$S_F = \sum_{x,\nu} \frac{\gamma^{\delta_{\nu,0}}}{2} \eta_\nu(x) (e^{\mu\delta_{\nu,0}} \bar{\chi}(x) U_\nu(x) \chi(x + \hat{\nu}) - e^{-\mu\delta_{\nu,0}} \bar{\chi}(x + \hat{\nu}) U_\nu^\dagger(x) \chi(x)) + am_q \sum_x \bar{\chi}_x \chi_x \quad (1)$$

$$Z = \sum_{\{d,m\}} \prod_{b=(n,\hat{\nu})} \frac{(N_c - d_\nu(n))!}{N_c! d_\nu(n)!} \gamma^{2d_\nu(n)\delta_{0,\hat{\nu}}} \prod_n \frac{N_c!}{m(n)!} (2am_q)^{m(n)}, \quad (2)$$

where Eq. (2) is obtained from Eq. (1) by integrating out the gauge variables first, and afterwards the Grassmann variables [20]. As we restrict to U(3) gauge group here, baryons are absent in Eq. (2); they would form self-avoiding worldlines for SU(3). We have studied SU(3) in [17] but in this work we restrict ourselves to U(3) for simplicity. The dual variables in Eq. (2) consist of so-called monomers

$m(n)$ and dimers $d_\nu(n)$ (with n a lattice coordinate). Dimers correspond to pion hops, for $N_c = 3$ there can be up to 3 pions on the same link (due to the quark substructure); the monomers contribute to the chiral condensate. On each site, the number of monomers and dimers touching this site has to add up to N_c due to the Grassmann constraint.

This system describes pions that are strongly interacting at low temperatures, resulting in a large entropy, and noninteracting at high temperatures. The chiral transition is located where the pion correlation is maximal (and diverges in the chiral limit). The temperature (in lattice units) is given by $aT = \frac{\gamma^2}{N_t}$ where γ is an anisotropy that favors temporal pion hops over spatial hops for $\gamma > 1$, and vice versa for $\gamma < 1$.

As explained in detail in [16], the integer-valued dual variables can be mapped to a binary vector x which can be adapted for the QUBO formalism:

$$\chi^2 = x^T Q x, \quad Q = W + p(A^T A + \text{diag}(2b^T \cdot A)). \quad (3)$$

Here Q is the quadratic QUBO matrix generated by the weight matrix W that is determined from the effective action, as well as A that incorporates the constraints that stem from the Grassmann integration of the quarks. The penalty factor p favors ($p \gg 1$) or disfavors ($p \ll 1$) the constraint over the weight matrix. As we explain in more detail in the next section, we embed in parallel

2×2 sublattices such that for each sublattice different boundaries b are applied. More precisely, the weight matrix of Q is determined for U(3) gauge group via

$$M = \begin{pmatrix} -2 \log(2am_q) + \log(2) & \log(3) \\ 0 & -\log(2am_q) \end{pmatrix} \quad (4)$$

$$D_s = \begin{pmatrix} \log(12) & 0 \\ 0 & \log(3) \end{pmatrix} \quad (5)$$

$$D_t = \begin{pmatrix} \log(12) - 4 \log(\gamma) & 0 \\ 0 & \log(3) - 2 \log(\gamma) \end{pmatrix} \quad (6)$$

$$W = \text{diag}(M, M, M, M, D_s, D_s, D_t, D_t) \quad (7)$$

and the constraint and boundary are given by

$$A = \begin{pmatrix} 2 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 2 & 1 & 0 & 0 & 2 & 1 & 0 & 0 \\ 0 & 0 & 2 & 1 & 0 & 0 & 0 & 0 & 2 & 1 & 0 & 0 & 0 & 0 & 2 & 1 \\ 0 & 0 & 0 & 0 & 2 & 1 & 0 & 0 & 0 & 0 & 2 & 1 & 2 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 2 & 1 & 0 & 0 & 2 & 1 & 0 & 0 & 2 & 1 \end{pmatrix} \quad (8)$$

$$b = (3 - d_{\text{ext}}^{(1)}, 3 - d_{\text{ext}}^{(2)}, 3 - d_{\text{ext}}^{(3)}, 3 - d_{\text{ext}}^{(4)}). \quad (9)$$

The boundary b is no longer constant as in our previous work, but now depends on the number of external dimers $d_{\text{ext}}^{(x)} \in \{0, 1, 2, 3\}$. This quantity changes after each update as explained in Sec. III A.

III. TECHNICAL DEVELOPMENT

In order to sample our effective theory on D-Wave, we have to map logical qubit x as in Eq. (3) to the physical qubits with connectivities given by the Pegasus topology. This can be achieved by synchronized combination of physical qubits (“chain”) to form a logical qubit, as discussed in detail in [16]. With 5760 physical qubits available, only small volumes can be sampled at once. For larger volumes, we use an iterative scheme that only samples disconnected sublattices, keeping dimers that are not part of the sublattice fixed; see Eq. (9). An example of the decomposition to sublattices is given in Fig. 2. Only the degrees of freedom inside the red boxes are sampled, and this is done in parallel to maximize the use of qubits.

A. Parallel sampling scheme

We now explain the details of the parallel sampling scheme, leaving the details on the Monte Carlo updates for Sec. III C. In order to sample all degrees of freedom, we have to alternate which dimers are frozen and which are updated. This we do by iterating through the base points, as highlighted in Fig. 2, which define what array of sublattices

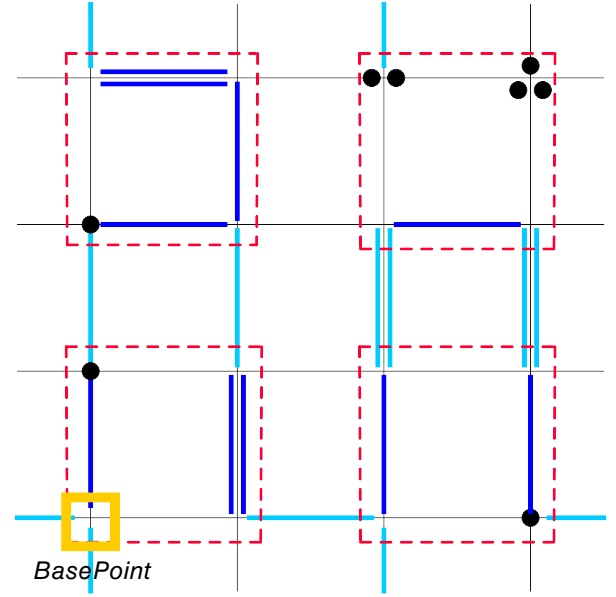


FIG. 2. Selecting 2×2 sub volume with fixed boundary condition. Black dots are monomers, blue lines are dimers. Monomers and dimers inside of red boxes are updated in parallel. The Grassmann constraint is required to be satisfied at each site. The dimers d_{ext} outside of red boxes are external dimers (light blue) and give boundary condition for sampling the red sublattices.

have to be considered for the next update. The number of base points depend on the dimension and the sublattice structure, as shown in Table I. We choose the sublattices of size 2×2 (square sampling), and of size $2 \times 2 \times 2$ (cube sampling). Clearly, the sublattice dimension d_s has to be smaller or equal to the dimension d of the lattice volume. For $d_s < d$, we have to loop through $\binom{d}{d_s}$ sets of directions to sample the full volume. For 2×2 sublattices, 16 logical qubits (~ 24 physical qubits) are required per sublattice and for $2 \times 2 \times 2$ sublattice, 40 logical qubits (~ 88 physical qubits). Since the number of sublattices is V/V_{sub} , the number of logical qubits for lattice size $V = N_s^{(d-1)} \times N_t$ is $N_{\text{logical}} = 2V(1 + d/2)$, with d the dimension of the volume and N_s and N_t are spatial and temporal size of lattice, respectively. Due to the block-diagonal structure of the QUBO matrix Q , the embedding of N_{logical} logical qubits to the physical qubits results in disconnected sublattice embeddings as shown in Fig. 1(b). The quantum annealer then solves for x in a fixed amount of time regardless of how many qubits are used in the parallelization. Hence, as long as the lattice volume V fits on the annealer at once (for 2×2 : 170 sublattices, $2 \times 2 \times 2$: 45 sublattices) the scaling is constant. If the lattice is too large, we sequentially update a block of maximal subset of sublattice at a time. In that case, sampling the sublattices scales linear with the volume.

It is sufficient to sample every degree of freedom by sequentially going through all base points that define the

TABLE I. Parallelization by sublattice sampling. Number of independent sublattices corresponds to the number of sublattices sampled in parallel. Number of base points is the number of sites in sublattice. For one volume sweep, it is required to sample sublattices in parallel for every base point.

V	V_{sub}	Independent sublattice	Base points
4×4	2×2	4	4
8×8	2×2	16	4
16×16	2×2	64	4
32×32	2×2	256	4
$4 \times 4 \times 4$	2×2	16	12
	$2 \times 2 \times 2$	8	8
$4 \times 4 \times 4 \times 4$	2×2	64	24
	$2 \times 2 \times 2$	32	32
$8 \times 8 \times 8 \times 4$	2×2	512	24
	$2 \times 2 \times 2$	256	32

location of sublattice. The number of base points on a d -dimensional lattice is obtained by the binomial coefficient

$$\text{The number of base points} = V_{\text{sub}} \times \binom{d}{d_s}. \quad (10)$$

The solution vector for a new base point depends on the previous solution vector because the fixed boundary $b(d_{\text{ext}}^{(x)})$ in Eq. (9) changes after each update. For the square sampling, we have $(N_c + 1)^4$ (256 for $N_c = 3$) and for cube sampling we have $(N_c + 1)^8$ (65536 for $N_c = 3$) distinct sets of b . While sampling, we store all generated configurations into histograms. This is feasible for the square update (see Fig. 3), but for the cube update the larger number of boundaries precludes us.

The quantum annealer produces a distribution of valid configurations that depend on the penalty factor p of the QUBO matrix defined in Eq. (3), as well as hyper parameters that we discuss in detail in the next section. We classify the valid solution vectors per sublattice into histograms for each set of boundary conditions b , and sort them by monomer number M and number of temporal dimers D_t , which encode the quark mass and temperature dependence. As the set of generated histograms $h_p(b, M, D_t)$ will depend on both the physical parameters and in particular the penalty factor p , the histogram is an estimate of the true distribution $h_{\text{true}}(b, M, D_t)$ that we obtain from exact enumeration (see Fig. 3 and the Appendix). We will discuss in Secs. III C and III D two methods to correct the histograms to obtain valid results.

To be more specific, we measure the N independent solution vectors $x^{(j)}$ ($j = 1, \dots, N$), blocked in n separate sublattices $x_i^{(j)}$ ($i = 1, \dots, n$), as shown in Fig. 4. The boundary condition is the same for all j at fixed i . Since not all configurations are valid in the sense that they fulfill the

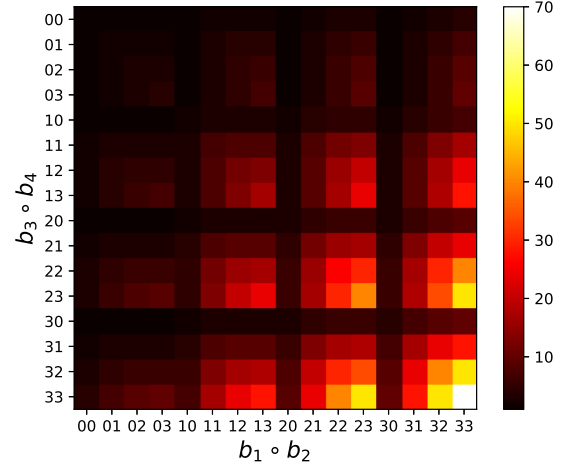


FIG. 3. Number of configurations for each boundary condition for the square sampling (b_1, b_2, b_3, b_4). The color coding reflects the multiplicity of configurations per boundary. For details see the Appendix.

$$\underbrace{\begin{pmatrix} x_1^{(1)} \\ x_2^{(1)} \\ \vdots \\ x_{n-1}^{(1)} \\ x_n^{(1)} \end{pmatrix}, \begin{pmatrix} x_1^{(2)} \\ x_2^{(2)} \\ \vdots \\ x_{n-1}^{(2)} \\ x_n^{(2)} \end{pmatrix}, \begin{pmatrix} x_1^{(3)} \\ x_2^{(3)} \\ \vdots \\ x_{n-1}^{(3)} \\ x_n^{(3)} \end{pmatrix}, \dots, \begin{pmatrix} x_1^{(N)} \\ x_2^{(N)} \\ \vdots \\ x_{n-1}^{(N)} \\ x_n^{(N)} \end{pmatrix}}_{N \text{ solution vectors measured}}$$

FIG. 4. Each column corresponds to a solution vector of a block QUBO matrix. The $x_i^{(j)}$ is a solution vector for each sublattice. We collect $x_i^{(j)}$ for fixed i from $j = 1, \dots, N$ with N solution vectors, and check the validity (blue: valid, red: invalid), see Eq. (12).

constraint equation (8) (red entries), it may happen that solution vectors have invalid blocks. N is typically large enough to ensure that for every j there exists several valid sublattices i from which we construct the histograms.

In Fig. 5, a comparison of the true histogram

$$h_{\text{true}}(b) = \frac{e^{-S[C_{2 \times 2}(b)]}}{\sum_{C_{2 \times 2}(b)} e^{-S[C_{2 \times 2}(b)]}} \quad (11)$$

by exact enumeration [with $C_{2 \times 2}(b)$ the 2×2 configurations at fixed boundary b] and the histogram h_p^b obtained by the D-Wave quantum annealer for penalty factors $p = 1, 2, \infty$ is shown. Here, b labels the 256 distinct boundary conditions $b = \{b_1, \dots, b_4\}$ in lexicographical order (left). The distribution for a specific boundary condition $b = (3, 3, 3, 3)$ (right) is shown for quark mass $am_q = 1.0$ and $\gamma = 0.1$, now sorted by the Boltzmann weight e^{-S} . Both histograms are in good agreement for $p = 1$, but this agreement deteriorates for larger p . This indicates that we have good overlap between the true and the approximate distributions of the 2×2 sublattices, which is

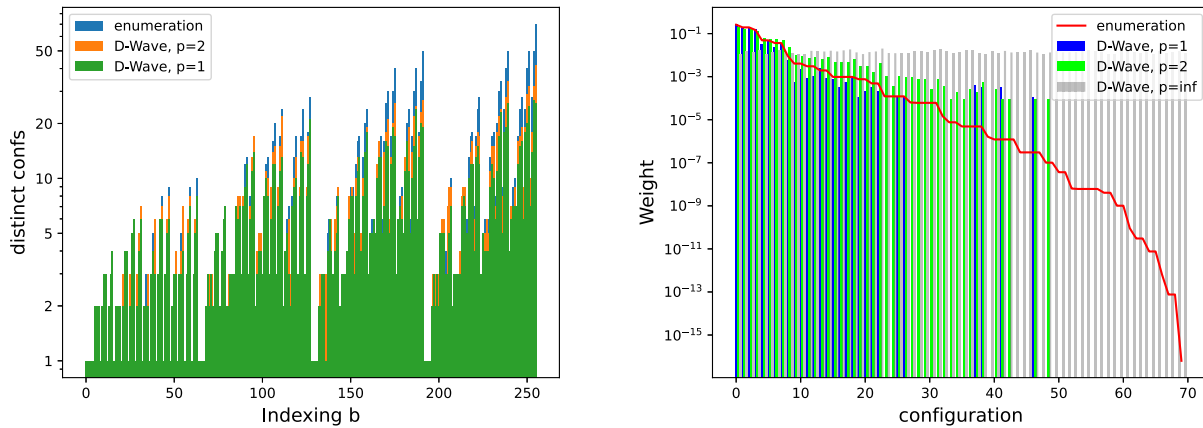


FIG. 5. Comparison for 2×2 sublattices between exact histograms from enumeration with those generated by D-Wave for $p = 1$ and $p = 2$. Left: comparison of the multiplicities for 256 different boundaries. Right: comparison of the weights for a specific boundary $b = (3, 3, 3, 3)$ for $am_q = 1.0$ and $\gamma = 0.1$, with exact Boltzmann distribution $h_{\text{true}}(b)$ compared to approximate distributions from D-Wave $h_p(b)$.

important to implement our strategy for large lattice volumes. How to correct the distribution for the small deviations is discussed in Secs. III C and III D.

B. Optimizing hyper parameters

To maximize the efficiency of the simulation and obtain correct results, the hyper parameters `chain_strength`, `num_reads`, `annealing_time` should be optimized. As described earlier, we must embed logical qubits into the physical layout of qubits so that the constellation of our logical qubits can simulate our QUBO problem (we use D-Wave’s automatized embedding to do this mapping). The physical qubits comprising a logical qubit must act as one; either all take values of 0 or 1. And their connections with other logical qubits must also act accordingly. However, during the annealing process it may (and does) happen that physical qubits within a logical qubit do not act uniformly, and this results in broken “chains.” These broken chains result in invalid solutions.

For submitting a problem to the D-Wave system, one has to provide a value for the `chain_strength`, which acts as a constraint that enforces connectivity of the chains. Increasing this value can reduce chain breaks. However, we cannot arbitrarily increase the value of `chain_strength` such that no broken chains occur, since too large a value makes our QUBO matrix less relevant and the system approaches the state of n independent logical qubits that do not interact with each other. This in turn adversely affects the number of obtained valid solutions. Thus some optimization is required.

We optimize the `chain_strength` to maximize the validity rate v , which is the number of valid solution vectors over all solution vectors:

$$v = \frac{1}{Nn} \sum_{j=1}^N \sum_{i=1}^n x_{i,\text{valid}}^{(j)}. \quad (12)$$

We rescale the QUBO matrix manually by Q_{max} , which is the maximum absolute value of any element of the QUBO matrix, before submitting to the D-Wave solver. This results in the rescaling of the `chain_strength` to order one. For small values of `chain_strength`, the number of broken chains is large. Since broken chains produce almost always invalid configurations, as a result the validity rate is low, as shown in Fig. 6. With increasing `chain_strength`, the validity rate grows with maximum around when `chain_strength` is between 1.0 to 1.4, depending on the penalty factor p and physical parameters. Note that the weight matrix W in the QUBO matrix has a block-diagonal structure, whereas the constraint matrix A has long-range connectivity, requiring a larger `chain_strength`. For larger `chain_strength` the validity rate drops again.

The number N of solution vectors x we request to be generated by the annealer is given by the parameter `num_reads`. The duration of an annealing process is set by `annealing_time`. A longer anneal time can increase the number of generated valid solutions. However this is not guaranteed, and the longer anneal times comes at the expense of overall computational cost. As shown in Fig. 6, we optimize the annealing time per sample. The validity rate is saturated after the annealing time 100 μs . As the annealing time is increased, `num_reads` should be decreased accordingly to keep the computational cost constant. We determine the optimal `chain_strength`/annealing time from the maximum/plateau of the validity rate. Since we have limited amount of quantum computing time, to balance the quantum computing time consumption

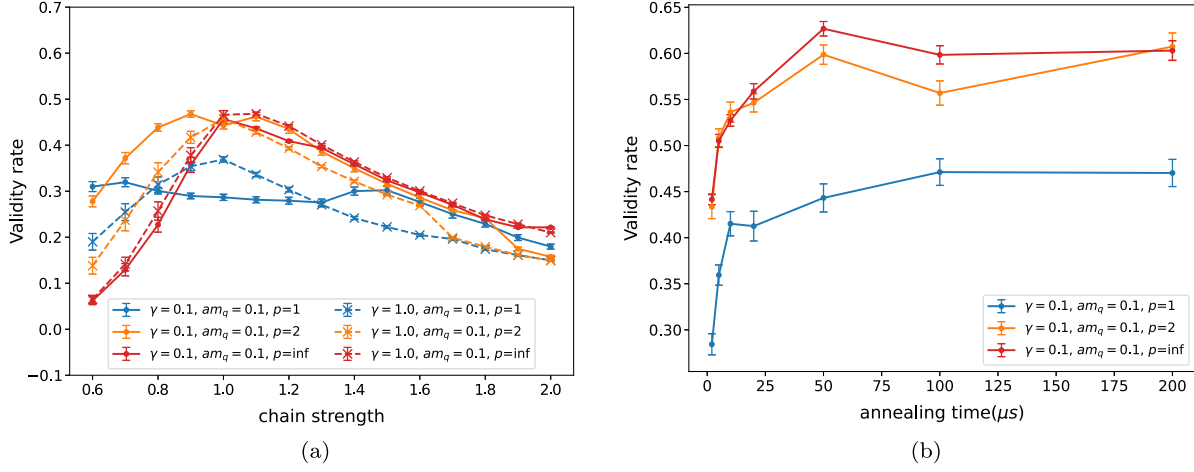


FIG. 6. (a) The validity rate v as a function of chain_strength on 2×2 sublattice with “3 3 3 3” boundary condition. For penalty factor $p = 1, 2$ and infinite. $\gamma = 0.1, 1.0, am_q = 0.1$. (b) The validity rate as function of the annealing time for $p = 1, 2$ and infinite and $\gamma = 0.1, am_q = 0.1$. A plateau is reached after about 100 ms. (a) Annealing time = 100, num_reads = 800. (b) Annealing time dependence.

and quality of the solution, we choose the chain_strength 1.2, annealing.time = 100 μs and num_reads = 200. This allows us to generate histograms for more physical parameters of the interest.

C. Metropolis and Metropolis-Hastings

Although we know the exact distributions of the 2×2 sublattices from enumeration in this simplest case (see the Appendix), this is not the case in the more realistic cases in general [SU(3) including strong coupling expansion]. But even with this exact distribution, it is not feasible to extend to finite periodic lattices: First, analytically extending to larger sublattices such as 4×4 cannot be achieved due the exponential growth of computation involved. Second, a standard Metropolis algorithm attempting to glue the sublattices together would produce extremely small acceptance rates of about 0.1% independent of the physical parameters. Alternative classical algorithms such as the Metropolis algorithm using the parallelization scheme as in Fig. 1 and sweeping through the 4 base points gives acceptance rates of about 20% for low temperatures $\gamma = 0.1$, with a quark mass dependence as shown in Fig. 7(c).

In contrast, with the approximate distribution for 2×2 sublattices measured by D-Wave for penalty factor $p = 1$, we do not rely at all on the exact distribution obtained from classical computation at all. However, we cannot incorporate the standard Metropolis algorithm, but have to use the Metropolis-Hastings algorithm:

$$P_{\text{accept}} = e^{-S_{\text{new}} + S_{\text{old}}} \frac{h_{\text{old}}}{h_{\text{new}}} \quad (13)$$

with $e^{-S_{\text{new}}}$ and $e^{-S_{\text{old}}}$ the Boltzmann factors for the new/old configuration, and h_{new} and h_{old} the new/old histograms of

the configuration. They are nontrivial: although they have the same boundary condition b , they differ in the specific distribution of monomers and dimers within the 2×2 sublattice. The proposal probability $P_{\text{proposal}} = \frac{h_{\text{old}}}{h_{\text{new}}}$ for a set of 2×2 updates is not uniform as it is drawn from the measured histogram. The details of the procedure are described by the following steps:

- (1) On the classical computer, generate a large lattice filled with 3 monomers at every sites.
- (2) Select independent sublattices and use the set of boundaries $\{b\}$ to construct the QUBO matrix for the physical parameters.
- (3) Generate a distribution for the sublattices using the quantum annealer and store it on the memory of classical computer. If the distribution for a specific boundary condition already exists, we can skip this step.
- (4) Select for each sublattice h_{new} from the distribution and make an accept/reject step by comparing with the previous configuration h_{old} on the classical computer according to Eq. (13).
- (5) Store the configuration in terms of monomers and dimers and measure observables.
- (6) Change the base point. Repeat from step 2.

As pointed out in Sec. III A, these histograms have good overlap with the true distribution. Hence we expect that the Metropolis-Hastings algorithm will have a much larger acceptance rate, in particular also for low temperatures. The dependence of the acceptance probability on the penalty factors p is shown in Fig. 7(a). We can clearly see that $p = 1$ has the largest acceptance rate, and that large p will decrease the acceptance rate. In the limit $p \rightarrow \infty$, the weight matrix will not contribute to the QUBO matrix and the 2×2 histograms become flat. Hence, due to $h_{\text{old}} = h_{\text{new}}$, the proposal probability drops out in

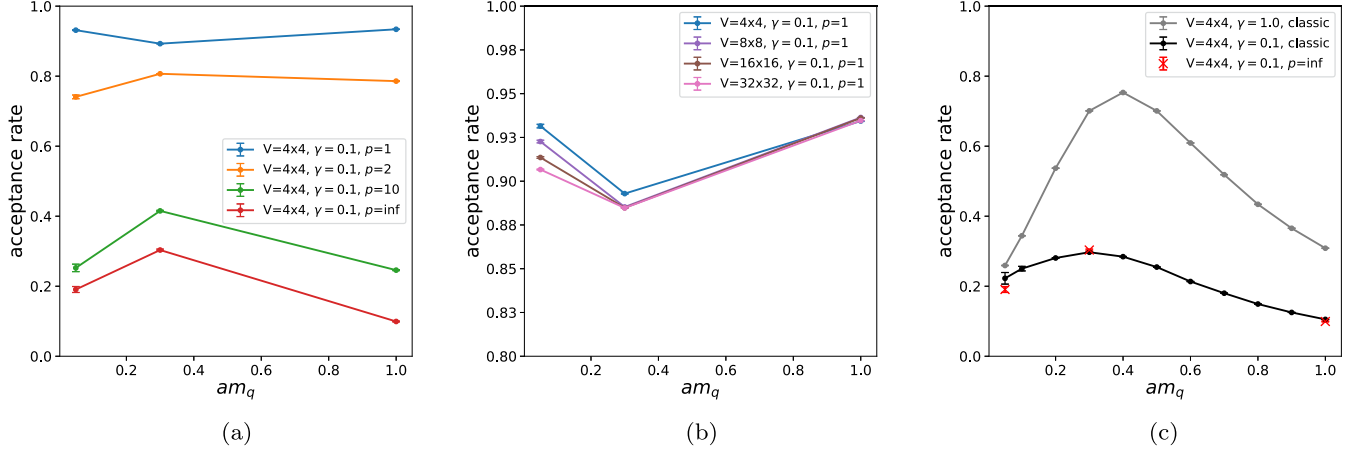


FIG. 7. Comparison of the acceptance rate versus am_q . (a) For a 4×4 lattice from Metropolis-Hastings, with 2×2 sublattice histograms from D-Wave, for various penalty factors p . (b) Illustrating independence of the acceptance rate of the volume. (c) Metropolis from classical computation, for $\gamma = 0.1, 1.0$. Metropolis-Hastings for $p = \infty$ reproduces the classical Metropolis acceptance rate. (a) Metropolis-Hastings (D-Wave) p -dependence. (b) Metropolis-Hastings (D-Wave) volume-dependence. (c) Metropolis (classical).

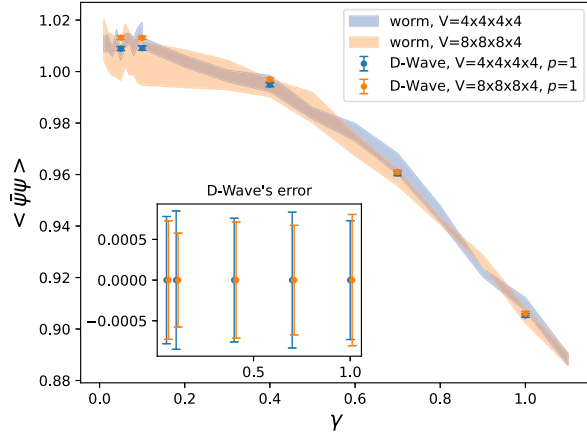


FIG. 8. The chiral condensate as a function of γ which is related to the temperature $aT = \frac{\gamma^2}{N_f}$. While the size of error of classical simulations via worm algorithm depends on the physical parameter, the method that we demonstrate in this paper is independent of γ , as illustrated in the inset. Hence, we confirm that this method has a significant advantage over the classical method at low temperatures.

the Metropolis-Hastings, which simplifies to standard Metropolis. The acceptance rate will hence be minimal as there is poor overlap with the important configurations given by the physical parameters. As shown in Fig. 7(c), the $p = \infty$ acceptance rate agrees with the classical Metropolis algorithm. Recall that classical computations are in particular expensive for low temperatures (small γ), whereas the computational costs for the histograms measured by D-Wave do not depend on γ . This is particularly evident by comparing the statistical uncertainties of the chiral condensate as measured via the worm algorithm versus the Metropolis-Hastings results based on the D-Wave

histograms, as shown in Fig. 8. In the case of worm simulation, the statistical uncertainty grows with decreasing temperature, while it remains constant for the results based on the hybrid approach of D-Wave histograms and Metropolis-Hastings. Interestingly, the error for $8^3 \times 4$ is of comparable size as for $4^3 \times 4$, both having the same number of Metropolis-Hastings updates. This might be related to the fact that for each base point, the solution vector x scales with the volume, as it contains 2×2 sublattices proportional to the volume.

Also, the acceptance rate does not depend much on the volume as shown in Fig. 7(b), as expected.

For the thermodynamic observables such as the chiral condensate and energy density, we show in Fig. 9 the p dependence for $p = 1, 2, 10, \infty$. We clearly see in the inset of Fig. 9(a) that $p = 1$ results in the smallest error, as it has the largest acceptance rate. For the large volume results in the next section, we hence always use $p = 1$. We show in Fig. 10 for $p = 1$ the volume dependence of the chiral condensate and the energy density, for $V = 4 \times 4, 8 \times 8, 16 \times 16, 32 \times 32$ and compare the data with results from a classical computer using the worm algorithm.

D. Branching strategy

In the previous section we studied in detail how to obtain the correct equilibrium distribution from the histograms that D-Wave provides via the Metropolis-Hastings algorithm. Note that if the histograms were already a perfect Boltzmann distribution, we could instead use a heatbath algorithm. Another promising method to obtain valid results that we briefly outline is to repeatedly branch into subbranches, and statistically evaluate each branch to obtain expectation values for thermal observables, such as the chiral condensate and the energy. The essential idea of the branching strategy is to maximize quantum parallelization. While in the

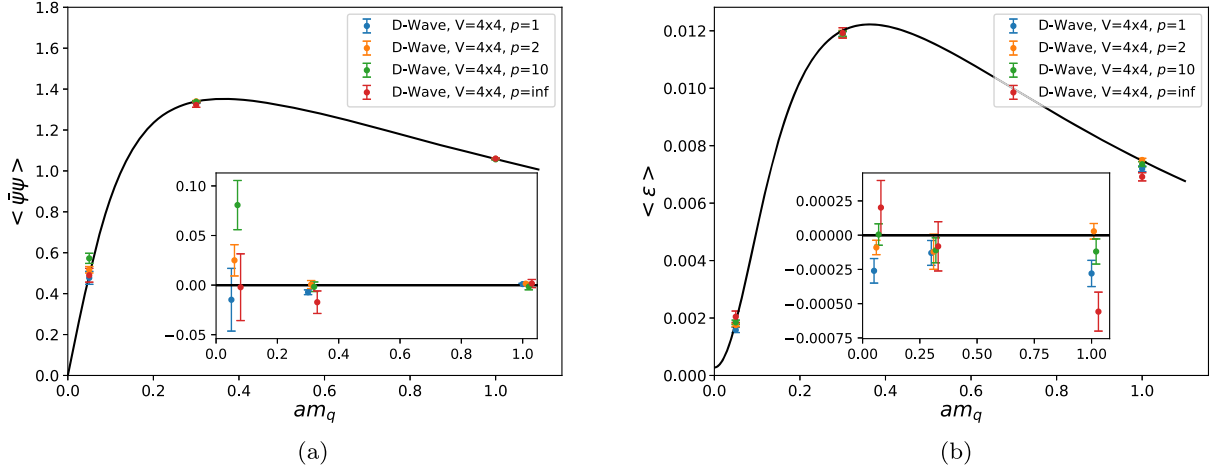


FIG. 9. Results from the Metropolis-Hastings algorithm, all for $\gamma = 0.1$: chiral condensate (a) and energy density (b) for different p values. The solid lines are analytic results from an exact enumeration. The inset figures show the difference between analytic and the numerical results. For any quark mass, $p = 1$ has the smallest error. (a) Chiral condensate at $\gamma = 0.1$. (b) Energy density at $am_q = 0.3$.

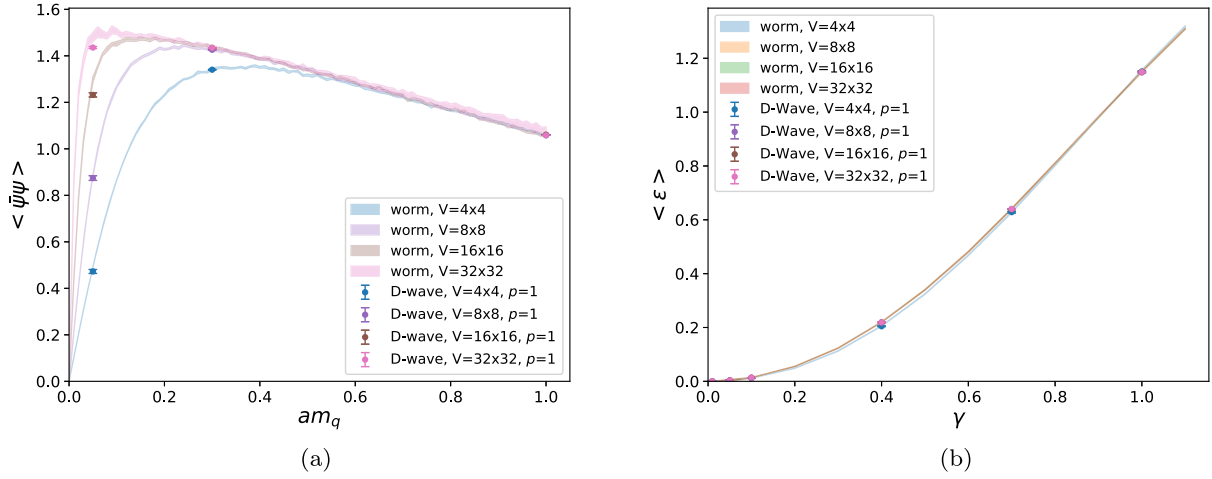


FIG. 10. Results of Metropolis-Hastings method for the chiral condensate for $p = 1$ at $\gamma = 0.1$ (a) and the energy density (b) for $p = 1$ at $am_q = 0.3$ for various 2-dimensional volumes. (a) Chiral condensate at $\gamma = 0.1$. (b) Energy density at $am_q = 0.3$.

Metropolis-Hastings the same accept-reject step is required; the branching strategy performs accept-reject at each branch.

We start with a specific boundary condition (square sampling) with $b = (3, 3, 3, 3)$, i.e. we choose 2×2 sublattices with weights depending on the physical parameters γ and am_q , as shown in Fig. 5(b). The number of samples N_{branch} defines the branches. The same sample can be chosen multiple times in extreme cases (e.g. am_q very large), although this is rare for physical parameters of interest (low temperatures, intermediate and small quark masses).

If we restrict to 2 dimensions, as we have 4 base points for 2×2 sublattices, we have to repeat the above procedure 4 times and obtain further subbranches, in total N_{branch}^4 branches, each representing a valid configuration. While branching into i subbranches, the weight factor

$w_i = e^{-S[i]} / h_{p,i}^b$ for each sample with index of depth i multiplies the weight factors for all existing branches, where h_p^b is the entry of the measured histogram, with b the specific boundary condition. The final weight for each of the N_{branch}^4 branches is

$$W_{\text{branch}} = w_1 \times w_2 \times w_3 \times w_4. \quad (14)$$

Ultimately, to correct for bias, we must reweight our observable $\mathcal{O}$,

$$\langle \mathcal{O} W_{\text{branch}} \rangle / \langle W_{\text{branch}} \rangle.$$

Unfortunately a limitation of this method so far is that the memory requirement for branching grows exponentially

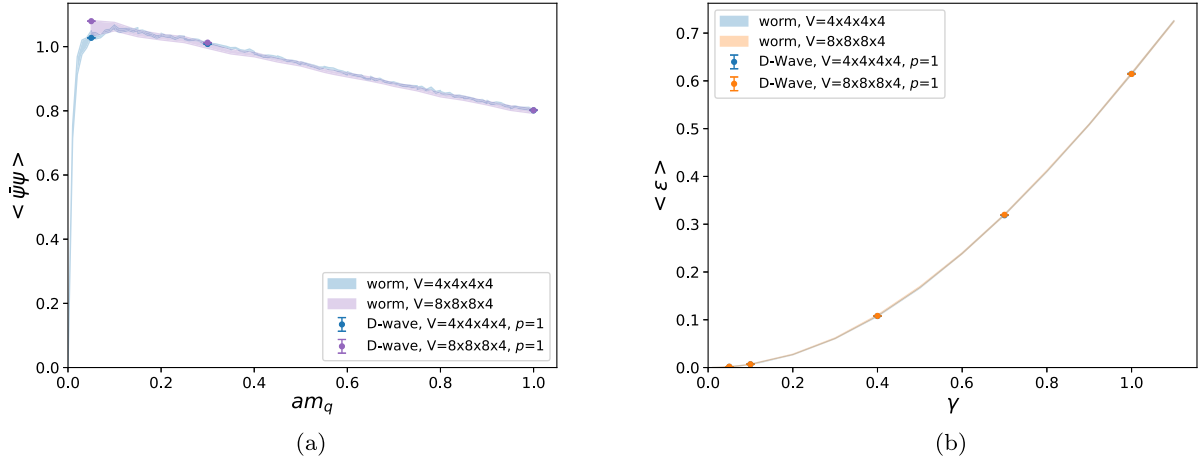


FIG. 11. 4-dimensional results on $V = 4 \times 4 \times 4 \times 4$ and $8 \times 8 \times 8 \times 4$. (a) Chiral condensate at $\gamma = 0.1$ for $am_q = 0.05, 0.3, 1.0$. (b) Energy density at $am_q = 0.3$ for $\gamma = 0.05, 0.1, 0.4, 0.7, 1.0$.

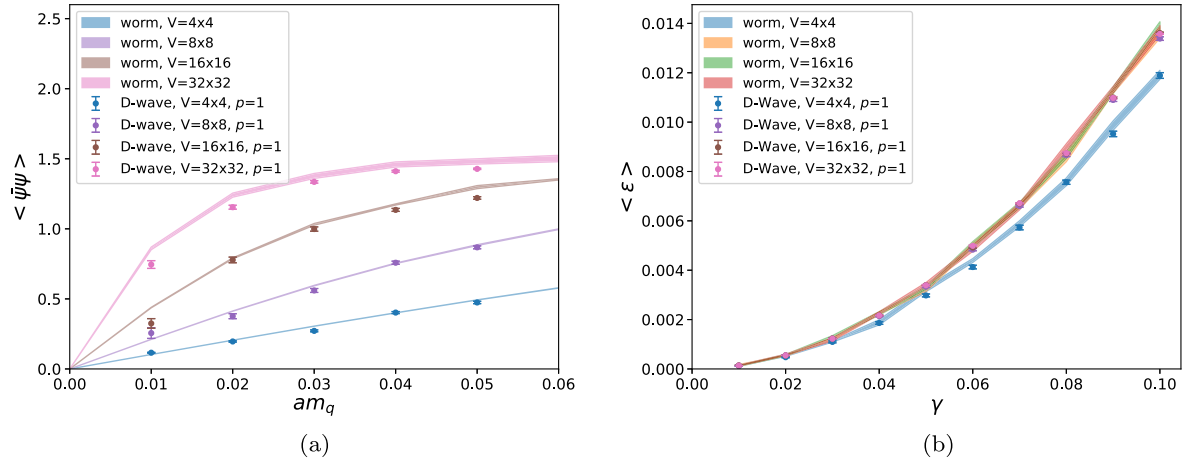


FIG. 12. Chiral extrapolation (a) from histograms at $am_q = 0.05, \gamma = 0.1, p = 1$, obtained via Metropolis-Hastings for several quark masses $am_q \leq 0.05$ and various volumes. (b) Low temperature extrapolation from $am_q = 0.3, \gamma = 0.1, p = 1$ for several $\gamma \leq 0.1$. The bands are obtained by worm simulation on the classical computer. (a) Chiral extrapolation at $\gamma = 0.1$. (b) Low temperature extrapolation at $am_q = 0.3$.

with N_{branch} , preventing us from presenting results at this point in time.

IV. RESULTS

We have compared different 2-dimensional lattices in Sec. III C and extended to 4-dimensional volumes for the thermodynamic observables, as shown in Fig. 11, which includes a $4^3 \times 4$ and $8^3 \times 4$ lattice. We are particularly interested in the chiral limit $am_q \rightarrow 0$ for the chiral condensate, and the low temperature limit $\gamma \rightarrow 0$ for the energy density, as these limits are particularly interesting from the physics point of view, and expensive with classical computers. We also present “extrapolated results,” for which we use a histogram at fixed quark mass and temperature γ and run the Metropolis-Hastings for smaller

quark masses and/or temperatures, given that there is still sufficient overlap with the corresponding target distributions. We demonstrated that these extrapolations are under control for all 2-dimensional volumes under investigation: the lattices $4 \times 4, 8 \times 8, 16 \times 16, 32 \times 32$, as shown in Fig. 12. As this works remarkably well, we can reduce drastically the number of histograms that have to be generated on the annealer for the set of physical parameters.

V. CONCLUSION

We developed an algorithm to simulate U(3) gauge theory in the strong coupling limit for large 2-dimensional and 4-dimensional volumes on a quantum annealer. In particular, we made use of a hybrid quantum/classical approach, whereby we determined the histograms in the

dual variables relevant to the QUBO formalism via the D-Wave quantum annealer, and then ran the Metropolis-Hastings algorithm based on these histograms on a classical computer. We described the hyper parameters we used to reduce the computation time on the annealer, detailing how we numerically determined optimal values for `annealing_time` and `chain_strength` by maximizing the validity rate given our finite computation time resources. Also, the penalty factor p that balances the weight matrix and the constraint in the QUBO matrix, for which the quality of the histograms is the best was determined to be $p = 1$, resulting in large acceptance rates.

Further improvements can be obtained by using the quantum parallelization more extensively and efficiently while branching our sublattices. This would require a better understanding of how the number of branches can be optimized with respect to computation time and memory requirement, in particular in four dimensions. We are actively investigating this option.

Extending to $SU(3)$ and including gauge corrections are important next steps towards a more realistic effective theory of lattice QCD. The first aspect includes baryons while the second aspect includes gluon propagation. The dual representation is well established for both extensions, and again can be mapped on binary vectors. However, these formulations will require more logical qubits. Still, it remains feasible to map the logical qubits required for the QUBO matrix to the physical qubits on the annealer. However, while the histograms required for this work, given square sampling, had 4^4 different boundaries to classify the histograms, these generalizations will require up to 16^4 different boundaries, for which the histograms can no longer be precomputed on D-Wave. Instead they need to be computed during Metropolis-Hastings in between the updates, whenever required. Studying more realistic effective theories will introduce two more physical parameters [21]: the quark chemical potential μ_q and the inverse gauge coupling β that is related to the lattice spacing a . We will report on our findings using the same approach via histograms determined on the annealer in a subsequent publication.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the Jülich Supercomputing Centre [22] for funding this project by providing computing time on the D-Wave Advantage System JUPSI through the Jülich UNified Infrastructure for Quantum computing (JUNIQ). J. K. and T. L. were supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) through the funds provided

to the Sino-German Collaborative Research Center TRR110 “Symmetries and the Emergence of Structure in QCD” (DFG Project-ID 196253076—TRR 110) and as part of the CRC 1639 NuMeriQS—Project No. 511713970, respectively. J. K. and W. U. are supported by the Deutsche Forschungsgemeinschaft (DFG) through the CRC-TR 211 “Strong-interaction matter under extreme conditions,” Project No. 315477589—TRR 211.

DATA AVAILABILITY

The data that support the findings of this article are not publicly available. The data are available from the authors upon reasonable request.

APPENDIX: EXACT ENUMERATION

From exact enumeration, we obtain for 2×2 sublattices with fixed boundary conditions (the building blocks of square sampling) in total 2350 distinct configurations, that are distributed over 256 distinct boundaries (b_1, b_2, b_3, b_4), with $b_i \in \{0, 1, 2, 3\}$.

If we further distinguish the sublattice configurations by its monomer number M and number of temporal dimers D_t , we obtain refined multiplicities. In Table II we give the multiplicities in these sectors M, D_t . Alternatively, the multiplicities can also be considered in the sectors M and B , where $B = b_0 + b_1 + b_2 + b_3$ is the sum of the boundaries, as given in Table III.

TABLE II. Multiplicities from exact enumeration on 2×2 sublattices with fixed boundary conditions sorted by monomer number M and temporal dimer D_t .

M	D_t							Sum
	0	1	2	3	4	5	6	
0	16	18	17	10	6	2	1	70
1	48	60	48	28	12	4	0	200
2	92	110	82	40	14	2	0	340
3	132	148	96	40	8	0	0	424
4	153	154	88	26	3	0	0	424
5	148	132	60	12	0	0	0	352
6	124	92	32	4	0	0	0	252
7	88	52	12	0	0	0	0	152
8	54	24	3	0	0	0	0	81
9	28	8	0	0	0	0	0	36
10	12	2	0	0	0	0	0	14
11	4	0	0	0	0	0	0	4
12	1	0	0	0	0	0	0	1
Sum	900	800	438	160	43	8	1	2350

TABLE III. Multiplicities from exact enumeration on 2×2 sublattices with fixed boundary conditions sorted by monomer number M and sum of boundary conditions $B = b_0 + b_1 + b_2 + b_3$.

M	B													Sum
	0	1	2	3	4	5	6	7	8	9	10	11	12	
0	1	0	4	0	10	0	20	0	19	0	12	0	4	70
1	0	4	0	16	0	40	0	64	0	52	0	24	0	200
2	0	0	10	0	40	0	88	0	116	0	74	0	12	340
3	0	0	0	20	0	72	0	136	0	144	0	52	0	424
4	0	0	0	0	31	0	100	0	158	0	116	0	19	424
5	0	0	0	0	0	40	0	112	0	136	0	64	0	352
6	0	0	0	0	0	0	44	0	100	0	88	0	20	252
7	0	0	0	0	0	0	0	40	0	72	0	40	0	152
8	0	0	0	0	0	0	0	0	31	0	40	0	10	81
9	0	0	0	0	0	0	0	0	0	20	0	16	0	36
10	0	0	0	0	0	0	0	0	0	0	10	0	4	14
11	0	0	0	0	0	0	0	0	0	0	0	4	0	4
12	0	0	0	0	0	0	0	0	0	0	0	0	1	1
Sum	1	4	14	36	81	152	252	352	424	424	340	200	70	2350

- [1] A. P. Lund, M. J. Bremner, and T. C. Ralph, Quantum sampling problems, BosonSampling and quantum supremacy, *npj Quantum Inf.* **3**, 15 (2017).
- [2] D. Ghamari, R. Covino, and P. Faccioli, Sampling a rare protein transition using quantum annealing, *J. Chem. Theory Comput.* **20**, 3322 (2024).
- [3] M. Vuffray, C. Coffrin, Y. A. Kharkov, and A. Y. Lokhov, Programmable quantum annealers as noisy Gibbs samplers, *PRX Quantum* **3**, 020317 (2022).
- [4] D. S. Wild, D. Sels, H. Pichler, C. Zanoci, and M. D. Lukin, Quantum sampling algorithms, phase transitions, and computational complexity, *Phys. Rev. A* **104**, 032602 (2021).
- [5] M. S. Jattana, Quantum annealer accelerates the variational quantum eigensolver in a triple-hybrid algorithm, *Phys. Scr.* **99**, 095117 (2024).
- [6] Z. G. Izquierdo, T. Albash, and I. Hen, Testing a quantum annealer as a quantum thermal sampler, *ACM Trans. Quant. Comput.* **2**, 7 (2021).
- [7] R. Sandt and R. Spatschek, Efficient low temperature Monte Carlo sampling using quantum annealing, *Sci. Rep.* **13**, 6754 (2023).
- [8] D. Ghamari, P. Hauke, R. Covino, and P. Faccioli, Sampling rare conformational transitions with a quantum computer, *Sci. Rep.* **12**, 16336 (2022).
- [9] R. Shibukawa, R. Tamura, and K. Tsuda, Boltzmann sampling with quantum annealers via fast Stein correction, *Phys. Rev. Res.* **6**, 043050 (2024).
- [10] F. A. Quinton, P. A. S. Myhr, M. Barani, P. C. del Granado, and H. Zhang, Quantum annealing applications, challenges and limitations for optimisation problems compared to classical solvers, *Sci. Rep.* **15**, 12733 (2025).
- [11] P. Weinberg, M. Tylutki, J. M. Rönkkö, J. Westerholm, J. A. Åström, P. Manninen, P. Törmä, and A. W. Sandvik, Scaling and diabatic effects in quantum annealing with a D-wave device, *Phys. Rev. Lett.* **124**, 090502 (2020).
- [12] N. Klco and M. J. Savage, Digitization of scalar fields for quantum computing, *Phys. Rev. A* **99**, 052335 (2019).
- [13] S. A Rahman, R. Lewis, E. Mendicelli, and S. Powell, SU(2) lattice gauge theory on a quantum annealer, *Phys. Rev. D* **104**, 034501 (2021).
- [14] M. Fromm, O. Philipsen, and C. Winterowd, Dihedral lattice gauge theories on a quantum annealer, *Eur. Phys. J. Quantum Technol.* **10**, 31 (2023).
- [15] M. Illa and M. J. Savage, Basic elements for simulations of standard-model physics with quantum annealers: Multigrid and clock states, *Phys. Rev. A* **106**, 052605 (2022).
- [16] J. Kim, T. Luu, and W. Unger, U(N) gauge theory in the strong coupling limit on a quantum annealer, *Phys. Rev. D* **108**, 074501 (2023).
- [17] J. Kim, T. Luu, and W. Unger, Testing importance sampling on a quantum annealer for strong coupling SU(3) gauge theory, *Proc. Sci. LATTICE2023* (2024) 219 [arXiv:2311.07209].
- [18] D.-W. Systems, Pegasus: The topology of D-wave quantum processors, Technical Report, D-wave systems Inc., 2019, https://docs.dwavesys.com/docs/latest/c_gs_4.html#topology-intro-pegasus.
- [19] https://docs.dwavesys.com/docs/latest/c_gs_4.html.
- [20] P. Rossi and U. Wolff, Lattice QCD with fermions at strong coupling: A dimer system, *Nucl. Phys.* **B248**, 105 (1984).
- [21] J. Kim, P. Pattanaik, and W. Unger, Nuclear liquid-gas transition in the strong coupling regime of lattice QCD, *Phys. Rev. D* **107**, 094514 (2023).
- [22] <https://www.fz-juelich.de/ias/jsc>.