High-Quality Thermal Gibbs Sampling with Quantum Annealing Hardware

Jon Nelson,[1,2] Marc Vuffray,[3] Andrey Y. Lokhov,[3] Tameem Albash,[4,5] and Carleton Coffrin[2,*]

[1] *Center for Nonlinear Studies, Los Alamos National Laboratory, Los Alamos, New Mexico 87545, USA*
[2] *Advanced Network Science Initiative, Los Alamos National Laboratory, Los Alamos, New Mexico 87545, USA*
[3] *Theoretical Division, Los Alamos National Laboratory, Los Alamos, New Mexico 87545, USA*
[4] *Department of Electrical and Computer Engineering, University of New Mexico, Albuquerque, New Mexico 87131, USA*
[5] *Department of Physics and Astronomy and Center for Quantum Information and Control, University of New Mexico, Albuquerque, New Mexico 87131, USA*

(Received 8 November 2021; revised 24 January 2022; accepted 25 January 2022; published 25 April 2022)

Quantum annealing (QA) was originally intended for accelerating the solution of combinatorial optimization tasks that have natural encodings as Ising models. However, recent experiments on QA-hardware platforms have demonstrated that, in the operating regime corresponding to weak interactions, the QA hardware behaves like a noisy Gibbs sampler at a hardware-specific effective temperature. This work builds on those insights and identifies a class of small hardware-native Ising models that are robust to noise effects and proposes a procedure for executing these models on QA hardware to maximize Gibbs-sampling performance. The experimental results indicate that the proposed protocol results in high-quality Gibbs samples from a hardware-specific effective temperature. Furthermore, we show that this effective temperature can be adjusted by modulating the annealing time and energy scale. The procedure proposed in this work provides an approach to using QA hardware for Ising-model sampling presenting potential opportunities for applications in machine learning and physics simulation.

DOI: 10.1103/PhysRevApplied.17.044046

I. INTRODUCTION

The computational task of sampling—that is, producing independent configurations of random variables from a given distribution—is believed to be among the most challenging computational problems. In particular, many sampling tasks cannot be performed in polynomial time, unless strong and widely accepted conjectures in approximation theory are refuted [1–3]. Consequently, state-of-the-art algorithms for general-purpose sampling are based on Monte Carlo methods that require significant computing resources to sample from distributions of practical interest.

Gibbs distributions, i.e., distributions with the form $P(\sigma) \propto \exp[-\beta H(\sigma)]$, have close ties to the modeling of many physical systems as they capture the equilibrium configurations that matter takes at a given inverse temperature β [4]. It has also been observed that Gibbs distributions provide a powerful foundation for building machine learning models and optimization algorithms [5,6]. However, the computational challenge of sampling from these distributions is often prohibitive for practical applications of machine learning and optimization. It has been observed that a variety of the emerging analog computing devices, including those based on optical [7] and quantum [8] technologies, can be used as fast Gibbs samplers because of the natural connection between probabilistic physical systems and state sampling. The emergence of these unconventional computing devices has the potential for a dramatic impact on state-of-the-art approaches for physics simulation, machine learning, and optimization.

From its inception, quantum annealing (QA) [9–12] has been designed for solving optimization tasks and not sampling tasks. However, in practice, various nonideal properties of QA-hardware platforms lead to outputs that are reminiscent of thermal Gibbs distributions [13–16]. Recent experiments have demonstrated that, in the operating regime corresponding to weak input interactions, QA hardware behaves like a *noisy Gibbs sampler* [17] at a hardware-specific effective temperature, that is, a sampler from a mixture of Gibbs distributions with fluctuating parameters caused by noise. Furthermore, it has been shown in Ref. [17] that samples from this noisy mixture of distributions are indistinguishable from a single

*cjc@lanl.gov

Gibbs distribution with spurious additions to interaction structure of the programmed model. In applications where sampling from a specific Gibbs distribution is required, this distortion of the model parameters may present an undesirable feature because of this mismatch in model structure.

Building on the insights of Ref. [17], this work demonstrates that it is possible to leverage QA hardware to perform high-quality Gibbs sampling from some types of input models. In particular, we identify a class of hardware-native Ising models where the D-Wave 2000Q platform consistently produces samples with a total variation (D_{TV}) distance less than 5% from a desired target Gibbs distribution. To this end, this work explores the full range of operational input parameters, in contrast with Ref. [17], which focused on quantifying the effect of noise on the structure of the output distribution and hence considered a narrow range of weak input couplings. We also show that changing the annealing time enables the QA hardware to sample from a range of inverse temperatures between 1.9 and 5.1 in the hardware's input units, which is an essential feature in a variety of applications. This work leverages the combination of three insights to achieve high-quality Gibbs sampling on the QA hardware: (i) it focuses on the hardware-native Ising models that are resilient to the noise in the hardware; (ii) the input models are rescaled to avoid distortions caused by the transverse field in the computational model; and (iii) both the annealing time and the energy scale of the model are leveraged to control the effective temperature of the target distribution. This results in a prescriptive procedure for conducting high-quality Ising-model Gibbs sampling on QA hardware without the need to tune the input model or perform postprocessing to correct for distorted output distributions. If required, both of these procedures can be leveraged to further improve the results presented in this work, as done in Refs. [13,18,19]. While this work shows that, under specific conditions, Gibbs sampling is possible using quantum annealing hardware, it does not illustrate how it can be used in specific sampling applications.

This work begins by reviewing the computational model of quantum annealing and previous works exploring sampling applications in Sec. II. Building on these previous works, a procedure for conducting high-quality thermal Gibbs sampling is proposed in Sec. III and evaluated on quantum annealing hardware. Section IV explores theoretical models to provide insights into the mechanisms that underpin the proposed sampling approach and Sec. V concludes with a discussion of future directions.

II. PROBLEM FORMULATION AND RELATED WORK

The foundational physical system of interest to this work is the Ising model, a class of graphical models where the nodes, V, represent classical spins (i.e., $\sigma_i \in \{-1, 1\} \forall i \in V$) and the edges, $E \subseteq N \times N$, represent pairwise spin interactions. A local field $h_i \forall i \in V$ is specified for each spin and an interaction strength $J_{ij} \forall i, j \in E$ is specified for each edge. The energy of a given spin configuration is given by the Hamiltonian:

$$H_{\mathrm{Ising}}(\sigma) = -\sum_{i,j \in E} J_{ij}\sigma_i\sigma_j - \sum_{i \in V} h_i\sigma_i. \tag{1}$$

This Ising model is widely used to encode challenging computational problems arising in the study of magnetic materials, machine learning, and optimization [20–23]. The particular computational task of interest to this work is to produce independent identically distributed (IID) samples σ from the Gibbs distribution associated with the Ising model at the inverse temperature [24] α, that is,

$$\mu(\sigma) \propto \exp\left[-\alpha H_{\mathrm{Ising}}(\sigma)\right]. \tag{2}$$

Throughout this work, we assume that the parameters J, h are in the range of -1 to 1, to provide a consistent scaling for α. Given the computational challenge of Ising-model sampling at finite temperature, our objective is to leverage QA hardware to conduct this task.

A. A brief review of quantum annealing

The central idea of quantum annealing is to use the transverse-field Ising model combined with an annealing process to find the low-energy configurations of a classical Ising model. The elementary unit of this model is a qubit $i \in V$ described by the standard vector of Pauli matrices $\{\widehat{\sigma}^x, \widehat{\sigma}^y, \widehat{\sigma}^z\}$ along the three spatial directions $\{x, y, z\}$. The hardware platform provides a programmable Ising-model Hamiltonian on the z axis [9],

$$\widehat{H}_{\mathrm{Ising}} = -\sum_{ij \in E} J_{ij}\widehat{\sigma}_i^z\widehat{\sigma}_j^z - \sum_{i \in V} h_i\widehat{\sigma}_i^z, \tag{3}$$

which encodes the Ising energy function given in Eq. (1) with a one-to-one mapping of qubits to Ising spins. Note that the eigenvalues of the Ising Hamiltonian operator are in bijection with the 2^N possible assignments of the classical Ising model from Eq. (1). The quantum annealing protocol strives to find the low-energy assignments to a user-specified $\widehat{H}_{\mathrm{Ising}}$ model by conducting an analog interpolation process of the following transverse-field Ising-model Hamiltonian:

$$\widehat{H}(s) = -A(s)\sum_{i \in V}\widehat{\sigma}_i^x + B(s)\widehat{H}_{\mathrm{Ising}}. \tag{4}$$

The interpolation process starts with $s = 0$ and ends with $s = 1$. The two interpolation functions $A(s)$ and $B(s)$ are designed such that $A(0) \gg B(0)$ and $A(1) \ll B(1)$, that is, starting with a Hamiltonian dominated by $-\sum_{i\in V}\widehat{\sigma}_i^x$ and slowly transitioning to a Hamiltonian dominated by $\widehat{H}_{\text{Ising}}$. In the closed system setting and when this transition process is sufficiently slow, the quantum annealing is considered to be adiabatic quantum computing. It is known that under these conditions, quantum annealing will find the ground states of $\widehat{H}_{\text{Ising}}$, therefore minimizing the configurations of H_{Ising}, with high probability [25–29]. To that end, the length of the annealing process t (in microseconds) is a user controllable parameter. The outcome of the quantum annealing process is specified by the binary string σ, where each element σ_i takes a value of $+1$ or -1 and corresponds to the observation of the spin projection of qubit i in the computational basis denoted by z.

B. Quantum annealing and sampling

Given the energy-minimizing nature of quantum annealing, it is not immediately clear how it may be applied to the sampling task posed in Eq. (2). It is reasonable to postulate that an adiabatic system would behave similarly to Gibbs sampling at the zero-temperature limit (i.e., $\alpha = \infty$), that is, drawing IID samples from the ground states of H_{Ising}. Interestingly, this is not usually the case, as the transverse-field component of the Hamiltonian, $\widehat{\sigma}_i^x$, induces biases in the annealing process to prefer some ground states over others [30–33]. Nevertheless, protocols have been proposed to help increase the fairness of ground-state sampling using quantum annealing [34,35]. In contrast to previous works, this work is concerned with sampling from thermal distributions (i.e., $\alpha < \infty$), which requires accurate sampling at all of the energy levels of H_{Ising}, presenting a formidable challenge in model accuracy.

Real-world QA hardware is an open quantum system that is subject to a wide variety of nonideal properties, including thermal excitations and relaxations that impact the results of the computation [36–38]. In this regard, the relevant adiabatic theorem is the open-system adiabatic theorem [39–41]: if the dynamics is governed by a master equation of Lindblad form [42] and if the annealing is done sufficiently slowly, the system will converge with high probability to the steady state of the dynamics. This result is promising because, for certain master equations, the thermal Gibbs state is the steady-state solution [43,44].

In addition to the fundamental impacts of open quantum systems, the D-Wave hardware documentation highlights five other sources of deviations from ideal system operations called *integrated control errors* (ICE) [45], which include background susceptibility, flux noise, digital-to-analog conversion quantization, input-output system effects, and variable scale across qubits. Consequently, it has long been observed that output distributions of the QA hardware produced by D-Wave Systems are reminiscent of a Gibbs distribution of the input Hamiltonian H_{Ising} [13–16,18] with a hardware-specific effective temperature of $\beta \approx 10$ [17,46]. The prevailing interpretation of the output distribution of the hardware is the *freeze-out* model, which proposes that the output reflects a quantum Gibbs distribution occurring at an input-dependent point toward the end of the annealing process, where some small amount of $\widehat{\sigma}_i^x$ remains [47]. This model has been notably successful for training quantum machine learning models [48,49] and generating statistics for quantum Gibbs distributions [50–52]. However, these quantum Gibbs distributions are inaccurate when targeting a desired classical Gibbs distribution for sampling applications [6,18,53,54]. The recent insight from Ref. [17] is that when this hardware is operated at a low-energy scale (i.e., $|J|, |h| \leq 0.050$), it behaves as a thermalized classical Gibbs sampler from H_{Ising} but suffers from a notable amount of distortion from instantaneous noise in the local field parameters, h, on the order of 0.036 [46,55], behaving as a so-called noisy Gibbs sampler.

Inspired by the classical Gibbs-sampling insights of Ref. [17], this work demonstrates that there exists a class of Ising Hamiltonians where D-Wave's 2000Q hardware produces high-quality Gibbs samples, with minimal distortions from noise or transverse fields. In particular, we consider the class of H_{Ising} with $J, h \in \{-1, 0, 1\}$, $|V| \leq 16$ that are representable on the D-Wave 2000Q hardware graph. The primary insight of this work is to operate the QA hardware at an energy scale, α_{in}, that is large enough to be noise resilient but small enough to avoid the degeneracy-breaking properties of the transverse field. It is not guaranteed that such a "sweet spot" exists but this work demonstrates that the range of $0.2 \leq \alpha_{\text{in}} \leq 0.4$ on the platform considered achieves the desired properties. Section IV provides a qualitative study postulating why a sweet spot occurs at this particular energy scale. Additionally, this work shows that in the proposed energy scale the annealing time has a consistent effect of tuning the effective temperature (i.e., α) of the Gibbs samples generated by the hardware. When combined, these observations present opportunities for leveraging QA hardware for conducting the Ising-model sampling task described by Eq. (2).

III. EXPLORING GIBBS SAMPLING WITH QUANTUM ANNEALING HARDWARE

This work is concerned with the following four parameters: H_{Ising}, the Ising model that one would like to sample from (restricted to $J, h \in \{-1, 0, 1\}$); $\alpha_{\text{in}} \in [0, 1]$, a scaling factor that will be used when programming the QA hardware; t, the annealing time of the QA protocol; and $\alpha_{\text{out}} \in [0, \alpha_{\max}]$, a scaling factor of the distribution output by the QA hardware, where $\alpha_{\max}$ is an estimated upper bound and is further described in Sec. III A 3. Given H_{Ising}, the QA hardware is programmed with the rescaled model

$H_{\text{Ising}}^{\text{in}} = \alpha_{\text{in}} H_{\text{Ising}}$ and executed with an annealing time t. Note that rescaling is equivalent to replacing J with $\alpha_{\text{in}} J$ and likewise with h. The empirical distribution output by the hardware ν is then compared to Gibbs distributions of H_{Ising} at different effective temperatures, i.e., $\mu(\alpha, \sigma) \propto \exp\left[-\alpha H_{\text{Ising}}(\sigma)\right]$. We define α_{out} as

$$\alpha_{\text{out}} = \text{argmin}_{\alpha} \; D_{\text{TV}}[\nu, \mu(\alpha)], \tag{5}$$

that is, the effective temperature of the closest Gibbs distribution of H_{Ising} to the results output by the hardware, using the D_{TV} distance as the measure of closeness. Details of the D_{TV} metric are provided in Appendix A. The remainder of this section explores how changes in the α_{in} and t parameters impact the output distribution of the D-Wave 2000Q Quantum Annealer located at Los Alamos National Laboratory, known as `DW_2000Q_LANL`. The unexpected finding is that there exists a range of α_{in} values where the QA hardware is a high-quality Gibbs sampler.

A. Experiment design

1. Ising-model selection

Given the challenges of sampling from embedded Ising models [53], this work focuses exclusively on H_{ising} Hamiltonians that have native representation in the QA hardware. In particular, the D-Wave 2000Q system implements a $\mathcal{C}_{16}$ Chimera graph [56], which consists of a 16×16 grid of unit cells each containing eight qubits (four horizontal and four vertical). In each unit cell, every horizontal qubit is connected to every vertical qubit through couplers and various qubits in adjacent unit cells are also connected through couplers. The coupling strength J_{ij} between qubits i and j can be programmed to values in the continuous range $[-1, 1]$ and the local fields h_i can be programmed to values in the range $[-2, 2]$. As part of the execution process, these unitless quantities are transformed into the energy scales of the hardware, of 0.0–6.36 GHz for $A(s)/h$ and 0.07–14.56 GHz for $B(s)/h$, where h is Planck's constant.

In selecting the Ising models for testing Gibbs sampling with QA, we strive to design models that exhibit a variety of sampling difficulties. Previous works have highlighted that QA will prefer some degenerate ground states over others as a result of residual effects from the annealing process [30–33]. As ground states are the most heavily weighted states in the low-temperature Gibbs distribution, these asymmetries introduce significant sampling errors. In this work, we elicit this effect by constructing a set of 13 models with ground-state degeneracies ranging from 1 to 38, to capture both easy and challenging instances to sample from with a quantum annealer.

In particular, this work considers seven *ground-state degeneracy* (GSD) cases with degeneracy values of 2, 4, 6, 8, 10, 24, and 38, without local fields, and six additional *GSD-F* models with one, two, three, four, five, and six ground-state degeneracies including local fields. Note that the GSD-F models tend to have lower degeneracy because the inclusion of the fields breaks symmetries within energy levels. The specific instances are referred to as GSD-*N* and GSD-F-*N*, where the "*N*" indicates the amount of ground-state degeneracy. The Hamiltonians of each instance and a discussion of how they are designed is provided in Appendix B.

2. Quantum annealing data collection

For each of the GSD models, H_{Ising}, a family of hardware inputs is considered by sweeping α_{in} between 0.0 and 1.0 with a step size of 0.0125 below 0.1, 0.025 between 0.1 and 0.5, and 0.1 above 0.5. The increased step size for low scales is helpful, as the output statistics tend to be more sensitive to α_{in} in this regime. For each of these models and scales, 10^6 samples are collected from the QA hardware. After every 100 samples, a gauge transform is applied in order to mitigate the effects of bias. For each gauge transformation, a randomly generated vector $\vec{a} \in \{-1, 1\}^V$ is used to redefine the Ising parameters via $h_i \rightarrow a_i h_i, J_{ij} \rightarrow a_i a_j J_{ij}$. This transformation preserves the energy eigenvalues of $H(s)$ but it relabels the spin configurations σ by $\sigma_i \rightarrow a_i \sigma_i$. The use of gauge transformations is a well-known technique for eliminating these field biases, which can add additional unwanted asymmetries [17,57,58]. After the samples have been collected, the empirical probability of each state is recorded. This data-collection process is repeated for annealing times t of 1 μs, 5 μs, 25 μs, and 125 μs.

3. Estimation of the effective output temperature

As discussed at the beginning of this section, we would like to identify the closest Gibbs distribution of H_{Ising} to the empirical distribution output by the QA hardware, i.e., Eq. (5). However, the exact solution of the optimization task posed by Eq. (5) presents a significant computational challenge because of difficulties in computing candidate Gibbs distributions. The primary reason that this work focuses on the small-scale Ising models with only 16 spins is to make this optimization task computationally tractable. In this work, the optimization problem given in Eq. (5) is solved by a brute-force calculation of the D_{TV} distance for a discrete set of α values. The α values are determined by calculating an optimistic estimate for α_{max} and then scaling this value by the same step sizes that are used for α_{in}. The α_{max} value is calculated by executing the single-qubit protocol proposed in Ref. [46] on each of the relevant spins and taking the average of the spin-by-spin effective temperatures recovered by that protocol. Therefore, the minimum of the range is 0.0 and the maximum is α_{max}. The α value from this discrete set that achieves the minimum D_{TV} distance between the quantum annealing data and the Gibbs distribution is selected as α_{out}. In the results, we observe that $\alpha_{\mathrm{out}} \ll \alpha_{\mathrm{max}}$, indicating that α_{max} is a sufficiently optimistic value for this brute-force optimization procedure in practice.

4. Bounding finite sampling performance

Because the empirical distribution produced by the QA hardware is calculated from finite samples, there is an unavoidable error caused by finite sampling (details are discussed in Appendix A). To understand the impact of this finite sampling error, a lower bound is calculated by a simulation procedure that generates 10^6 samples from the selected α_{out} Gibbs distribution and calculates the D_{TV} distance between this empirical distribution and the exact distribution. Because the samples are generated from a known source distribution, the only source of error is a result of finite sampling. Therefore, this value represents the best D_{TV} distance that is achievable given the number of samples that are used in these experiments: 10^6. This lower bound is included in the result figures to provide a measure of the portion of the D_{TV} distance that can be attributed solely to the limitations of finite sampling. In this work, we find that this sampling issue is only significant when the input scale is low ($\alpha_{\mathrm{in}} < 0.1$), as this regime reflects a distribution that is close to uniform and the probability mass is spread out among exponentially many states. The presence of errors caused by finite sampling motivates our use of a large number of samples (i.e., 10^6 per data point) and provides an explanation for some of the consistent features in the results that are observed in the low-α_{in} regime.

B. Experiment results

1. A typical GSD model

We begin by presenting detailed results on a characteristic GSD model and then provide a summary of the results across the complete collection of models. The D_{TV}-distance results for the GSD-6 model are presented in Fig. 1. As the value of α_{in} is increased, the results show an up-down-up shape in the D_{TV} metric. At very low $\alpha_{\mathrm{in}} < 0.01$, the D_{TV} value is limited by finite sampling effects. As α_{in} is increased, it quickly deviates from the family of H_{Ising} Gibbs distributions and gradually becomes more Gibbs-like. At some point, while increasing α_{in}, a minimum D_{TV} value is reached and it begins to increase monotonically until the maximum α_{in} value. This up-down-up trend is replicated across a variety of the models considered in this work and plausible causes for it are discussed in Sec. IV. The central observation of this analysis is that there is a band of α_{in} between approximately 0.2 and 0.4, where the output distribution of the QA hardware deviates from the target Gibbs distribution by less than 5% in D_{TV} distance, which we designate as *high quality*.

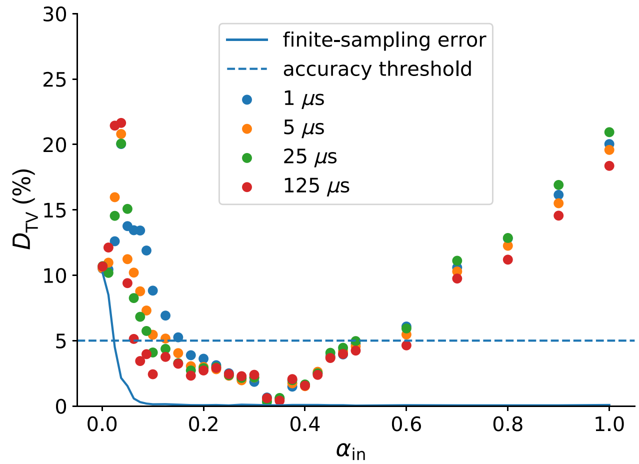

FIG. 1. The D_{TV} distance between QA-hardware output distribution for various scalings α_{in} and the target Gibbs distribution with estimated α_{out} on the GSD-6 Hamiltonian. This model contains 16 spins and six ground states. The experiment is repeated for annealing times of 1 μs (blue), 5 μs (orange), 25 μs (green), and 125 μs (red). The finite-sampling lower bound is indicated by blue curve. The *accuracy threshold* indicates the statistical requirement to be considered a high-quality Gibbs sampler in this work. The surprising finding is the band of α_{in} between approximately 0.2 and 0.4, where the hardware output does conform to a Gibbs distribution of the input Hamiltonian.

Outside of this particular range of α_{in}, the D_{TV} distance increases considerably, showing how the distribution of the QA hardware dramatically shifts away from the family of H_{Ising} Gibbs distributions. Interestingly, there does not seem to be any significant dependence of the D_{TV} distance on annealing time (i.e., each of the colored points in Fig. 1 follow similar trends). These results suggest that this middle regime of α_{in} is favorable for using QA hardware as a Gibbs sampler at each of the annealing times considered. Similar results on all of the GSD instances are provided in Appendix C.

The results from Fig. 1 suggest there are operating regimes where the QA hardware is an accurate Gibbs sampler; however, the effective temperature (i.e., α_{out}) of that distribution is not presented. Figure 2 explores this point by presenting the α_{out} values that are found for each of the high-quality input configurations considered (i.e., $0.2 \leq \alpha_{\mathrm{in}} \leq 0.4$). Given the prevailing theories of how QA hardware operates [46,47], one expects that increasing the energy scale of the input α_{in} or the annealing time will increase the effective temperature of the output distribution, α_{out}. Indeed, that trend is observed in Fig. 2, where the recovered α_{out} values increase monotonically (or nearly so) with both α_{in} and the annealing time. This monotonicity is a particularly useful result, as it suggest a simple procedure for tuning the effective temperature of the distribution output by the QA-hardware platform, which is an essential feature for most practical sampling applications.

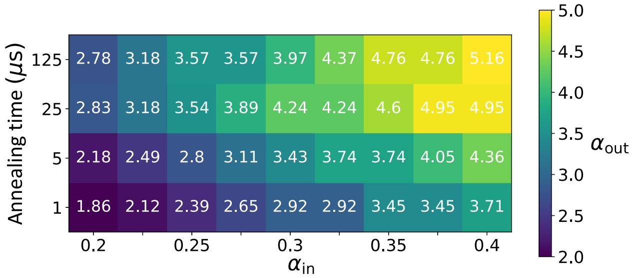

FIG. 2. The values of α_{out} recovered by operating the QA hardware at different values of α_{in} and annealing times in the high-quality regime identified in Fig. 1 for the GSD-6 Hamiltonian. Encouragingly, α_{out} increases monotonically (or nearly so) with both α_{in} and annealing time, allowing the hardware generate distributions at different effective temperatures, which is an essential feature for practical applications.

The generation of Gibbs samples in the α_{out} range of from 1.85 to 5.16 as shown in Fig. 2 is an encouraging result, as the critical points separating high- and low-temperature regimes of Ising spin glasses on lattices and random graphs typically occur for values of $\alpha_{out} < 1.0$ [22]. Sampling from these low-temperature regimes is known to present significant computational challenges for classical algorithms, giving hope that QA hardware may be able to provide a performance enhancement on these tasks. However, additional study is needed to determine if the results presented here can generalize to larger system sizes with different connectivity layouts and to temperatures below the critical point of natively representable Ising spin glasses [59].

2. Summary of the GSD models

To test if the observations on the GSD-6 instance generalize to a broader collection of models, we repeat the previous experiment on all 13 of the GSD models proposed in this work, seven GRD and six GSD-F cases. Figure 3 displays the minimum D_{TV} distance with respect to the ideal Gibbs distribution for each scale of α_{in} and for each model. Only the 1 μs data is presented but results with other annealing times are available in Appendix C. Although one can observe a variety of distinct features across these instances, the encouraging finding is that the range of α_{in} within 0.2 and 0.4 consistently yields statistics with a low D_{TV} distance, suggesting that the observations on the GSD-6 model do generalize to a broader class of models. In particular, $\alpha_{in} \approx 0.3$ achieves a low D_{TV} distance for nearly all of the models considered.

It is important to briefly discuss the GSD-2 and GSD-F-1 models, as these are the only ones that do not follow the up-down-up trajectory shown in Fig. 1. In these cases, the D_{TV} distance is high for low α_{in} and then drops below 5% for α_{in} greater than 0.2, as expected. However, the D_{TV} distance remains low even for α_{in} greater than 0.4.

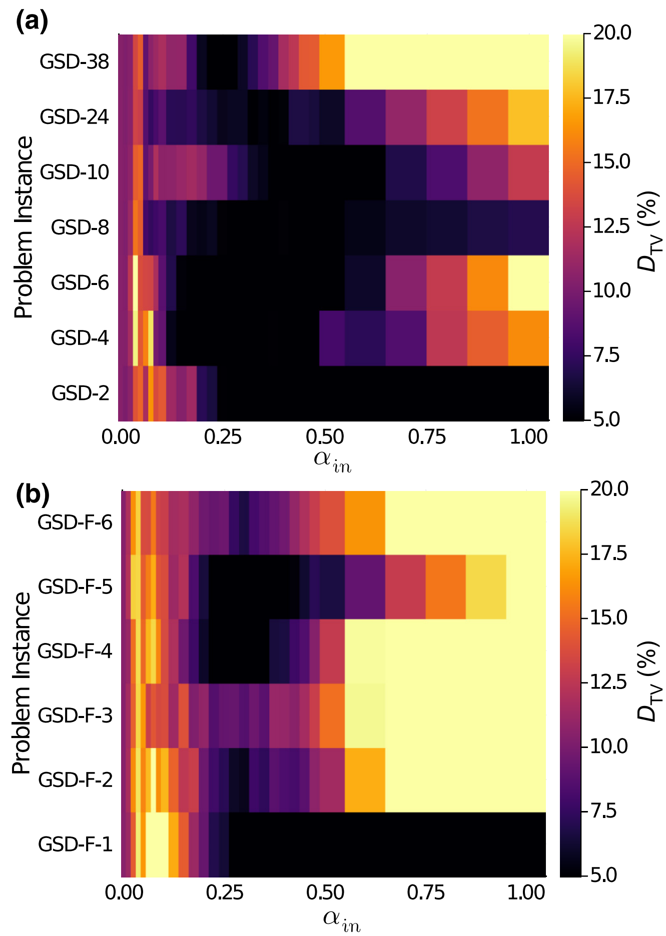

FIG. 3. The D_{TV} distance between the QA-hardware output distributions for various scalings α_{in} and the target Gibbs distribution with estimated α_{out} on the all (a) GSD instances and (b) GSD-F instances. A consistent high-quality Gibbs-sampling band is observed for $0.2 \leq \alpha_{in} \leq 0.4$, with $\alpha_{in} \approx 0.3$ achieving a low D_{TV} distance for nearly all of the GSD models.

This result is because when α_{in} is large, the probability distribution is highly concentrated in the ground states and these instances do not exhibit ground-state degeneracy breaking. As there is only one ground state in the GSD-F case and two symmetrical ground states in the GSD case, the sampling task is relatively easy and effectively reduces to a ground-state identification task. The QA hardware achieves low D_{TV} in these cases by simply outputting the ground states with high probability. Although one can consider these two cases to represent *easy* sampling tasks, we include them to highlight the increased challenge that instances with more ground-state degeneracy pose to using QA for sampling applications.

The results from Fig. 3 indicate that the proposed high-quality Gibbs-sampling regime generalizes to a variety of Ising models; however, the stability of the effective temperatures (i.e., α_{out}) of those distributions is an important

TABLE I. A summary of the smallest and largest α_{out} values recovered by operating the QA hardware at different values of α_{in} and annealing times in the high-quality regime across all GSD models. Although there is some variation in the largest α_{out} values, the QA hardware has a fairly consistent behavior across the GSD models and suggests that samples can be generated effectively in the range of effective temperatures from at least 1.85 to 3.97 for a variety of Hamiltonians.

Instance	min α_{out}	max α_{out}
GSD-2	1.32	3.97
GSD-4	1.85	4.95
GSD-6	1.85	5.16
GSD-8	1.59	4.37
GSD-10	1.32	4.76
GSD-24	1.59	4.58
GSD-38	1.59	4.76
GSD-F-1	1.32	3.97
GSD-F-2	1.32	4.37
GSD-F-3	1.59	4.37
GSD-F-4	1.59	4.58
GSD-F-5	1.59	4.76
GSD-F-6	1.59	4.37

question. Table I provides a summary of this information by presenting the minimum and maximum α_{out} values that can be achieved on each model by varying both α_{in} and the annealing time (for more detailed information, see Appendix D). Although there is some variability in the largest α_{out} values, the results indicate that the range of effective temperatures output by the hardware remains relatively stable and the hardware is suitable for sampling from all models between 1.85 and 3.97. These results also suggest that the observations on the GSD-6 model generalize to a broader class of models.

IV. DISCUSSION

The use of quantum annealing for Gibbs sampling has been proposed as early as 2010 [13]. Since then, several

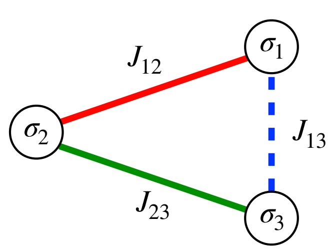

FIG. 4. A diagram of three-spin Ising chain experiment. The solid lines for J_{12} and J_{23} indicate couplings that are programmed in the input Hamiltonian. The dashed line for J_{13} indicates that this coupling is not programmed in the input Hamiltonian even though it is reconstructed from the output statistics and is thus a "spurious coupling."

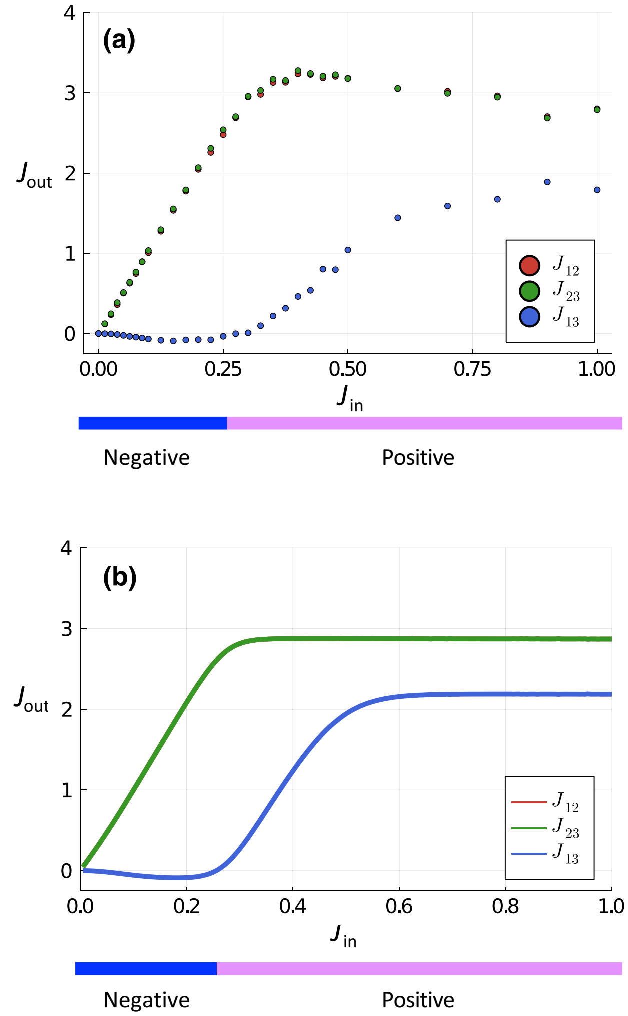

FIG. 5. (a) Reconstructed coupling values for a three-spin Ising chain sampled at various coupling strengths. Only J_{12} (coupling between σ_1 and σ_2, represented by green in the figure) and J_{23} (coupling between σ_2 and σ_3, represented by red in the figure) are programmed in the input Hamiltonian. J_{13} (blue) is not included in the input Hamiltonian and represents a spurious coupling between σ_1 and σ_3. The blue and purple bars represent the intervals of the input coupling strength J_{in} that produce negative and positive spurious couplings, respectively. (b) Results from the model simulation that replicates the three-spin Ising chain experiment. J_{12} and J_{23} are identical and only the green line is visible. J_{13} is spurious coupling and is represented in blue. The blue and purple bars represent the intervals of the input coupling strength J_{in} that produce negative and positive spurious couplings, respectively. Note how in both of these cases J_{13} is reconstructed as negative for low-input couplings and then transitions to positive as input coupling increases beyond 0.275. As this coupling is not programmed in the input Hamiltonian, this nonzero reconstruction marks a disruption to the desired distribution. The negative and positive regions indicate the scaling regimes where noise and quantum effects distort the output distribution, respectively. We observe that the optimal scaling regime to use for Gibbs sampling is when J_{13} is small relative to J_{12} and J_{23}.

studies have indicated that the output distribution of a quantum annealer does not sample from the input Hamiltonian's Gibbs distribution [6,18,32,33]. However, many of these works use an input scaling that is outside of the range identified by this work and are therefore subject to the distortions that we see in Figs. 1 and 3. These figures show that if α_{in} is too small (< 0.2) or too large (> 0.4), then the output distribution of the QA hardware will differ greatly from that of a corresponding Gibbs distribution. Additionally, the observation that the QA operating protocol proposed by this work yields high-quality Gibbs samples in a variety of Ising models suggests a systematic phenomenon that produces this behavior.

To investigate what might account for the distortions in these Gibbs distributions and the unique features that occur in the $0.2 \le \alpha_{\text{in}} \le 0.4$ range, we conduct a detailed study of the output statistics of a very small Ising Hamiltonian, i.e., a three-spin Ising chain. With such a small system, we are able to create a theoretical model that reproduces the experimental data. This theoretical model suggests local field noise and residual transverse-field effects as potential explanations for the observed distortions. The theoretical model finds that the three-spin system sampled from an Ising Hamiltonian with extraneous couplings when α_{in} is in the low- or high-scaling regime. Only when α_{in} remains in the range identified by this work does the system produce the desired Gibbs distribution. This result provides additional evidence that this restricted scaling regime is optimal for conducting Gibbs sampling with QA hardware.

A. Three-spin Ising chain statistics

In this experiment, three spins denoted by σ_1, σ_2, and σ_3 are linked together in a ferromagnetic chain, with σ_1 coupled to σ_2 and σ_2 coupled to σ_3 as shown in Fig. 4. Note that σ_3 is not coupled to σ_1. Formally, the input Hamiltonian is defined as follows:

$$H_{\text{Ising}} = -J_{\text{in}}(\widehat{\sigma}_1^z\widehat{\sigma}_2^z + \widehat{\sigma}_2^z\widehat{\sigma}_3^z), \tag{6}$$

where J_{in} is the value of the coupling. Note that J_{in} is equivalent to α_{in} from the 16-spin experiments. J_{in} is swept between 0.0 and 1.0 with step sizes matching the discretization of α_{in} [60]. For each value of J_{in}, 5×10^6 samples are collected from the QA hardware and the output statistics are recorded. Once again, in order to mitigate bias, a spin-reversal transform is applied after every 100 samples.

The output distribution is analyzed by solving the inverse-Ising problem using the Interaction screening estimation algorithm [22,61]. This algorithm takes the empirical distribution produced by the hardware and estimates the Ising Hamiltonian that most likely produces the given output statistics. The estimated values for each of the couplings are denoted as J_{out} and, more specifically, as J_{12}, J_{23}, and J_{13} to indicate the specific edge in the three-spin chain. These estimated values are compared to the corresponding input values J_{in} in Fig. 5(a). This procedure provides an understanding of the effective Hamiltonian output by the QA hardware, which may differ from the Ising Hamiltonian that has been programmed. The entire protocol is repeated for annealing times of 1 μs, 5 μs, 25 μs, and 125 μs, with similar results. Only the data for a 1-μs annealing time are presented in Fig. 5(a) (for results for the other annealing times, see Appendix E).

Because σ_1 and σ_3 are not coupled together in the input Hamiltonian, one expects J_{13} from the reconstruction to be close to zero. However, as shown in Fig. 5(a), J_{13} appears to be negative and, hence, antiferromagnetic below $J_{\text{in}} = 0.275$ and positive above that point. In Ref. [17], this effect is referred to as a "spurious coupling," because it appears in the output distribution despite not being programmed in the input Hamiltonian. It is striking that this spurious coupling exactly cancels out when $J_{\text{in}} = 0.275$, which coincides with the optimal input scaling for α_{in}, for which almost all GSD models in Fig. 3 are shown to have low D_{TV} distance. Furthermore, the regime where the spurious coupling strength is low relative to the intended coupling strengths can be expanded to approximately include J_{in} between 0.2 and 0.35, which again closely matches the optimal range for α_{in} that is found for the 16-spin experiments. This provides additional support for the observation that the QA hardware considered here is an effective Gibbs sampler in this specific scaling range.

B. A three-qubit Ising chain effective model

To further explore a connection between these experimental observations and the spurious couplings observed in the output of the hardware [i.e., Fig. 5(a)], we propose an extension of the effective single-qubit quantum model developed in Ref. [17] to the three-qubit context. The single-qubit model considered in Refs. [17,46] includes a transverse field with an intensity proportional to the input local field parameter, h. This transverse-field component is able to reproduce an observed saturation of the output field for large input values. Another important feature proposed in Ref. [17] is qubit noise perturbing the input parameters of the model, which explains the spurious couplings in the regime of low input parameters. Building on these characteristics, we consider the following toy model on three qubits, controlled by a single parameter J_{in} in the absence of local fields. We assume that the output distribution is a noise-averaged thermal distribution,

$$\rho = \frac{1}{8}\sum_{s_1,s_2,s_3=\pm 1} \frac{\exp(-\beta H)}{\text{Tr}\exp(-\beta H)}, \tag{7}$$

TABLE II. Input coupling values for GSD models before scaling by α_{in}. The D-Wave programming convention is used, where negative couplings indicate ferromagnetic and positive couplings indicate antiferromagnetic.

	GSD-2	GSD-4	GSD-6	GSD-8	GSD-10	GSD-24	GSD-38
$J_{296,300}$	1.0	−1.0	1.0	1.0	−1.0	1.0	1.0
$J_{296,301}$	1.0	1.0	−1.0	−1.0	1.0	1.0	1.0
$J_{296,302}$	−1.0	1.0	−1.0	−1.0	−1.0	−1.0	−1.0
$J_{296,303}$	−1.0	−1.0	1.0	−1.0	1.0	−1.0	1.0
$J_{297,300}$	1.0	1.0	1.0	−1.0	−1.0	−1.0	1.0
$J_{297,301}$	1.0	1.0	1.0	1.0	−1.0	1.0	1.0
$J_{297,302}$	−1.0	−1.0	1.0	−1.0	1.0	−1.0	1.0
$J_{297,303}$	1.0	−1.0	−1.0	1.0	−1.0	−1.0	−1.0
$J_{298,300}$	−1.0	1.0	−1.0	1.0	−1.0	−1.0	−1.0
$J_{298,301}$	1.0	−1.0	−1.0	−1.0	−1.0	1.0	1.0
$J_{298,302}$	1.0	1.0	−1.0	1.0	1.0	−1.0	−1.0
$J_{298,303}$	1.0	1.0	1.0	−1.0	−1.0	1.0	−1.0
$J_{299,300}$	1.0	1.0	1.0	−1.0	−1.0	1.0	−1.0
$J_{299,301}$	1.0	−1.0	−1.0	−1.0	1.0	1.0	−1.0
$J_{299,302}$	−1.0	1.0	−1.0	−1.0	1.0	−1.0	1.0
$J_{299,303}$	−1.0	1.0	1.0	−1.0	1.0	−1.0	−1.0
$J_{300,308}$	1.0	−1.0	−1.0	−1.0	−1.0	1.0	1.0
$J_{301,309}$	1.0	1.0	−1.0	−1.0	−1.0	1.0	1.0
$J_{302,310}$	1.0	1.0	1.0	1.0	1.0	−1.0	−1.0
$J_{303,311}$	−1.0	1.0	1.0	−1.0	−1.0	−1.0	1.0
$J_{304,308}$	1.0	−1.0	1.0	−1.0	1.0	1.0	1.0
$J_{304,309}$	1.0	−1.0	1.0	−1.0	−1.0	−1.0	−1.0
$J_{304,310}$	1.0	1.0	−1.0	1.0	−1.0	1.0	−1.0
$J_{304,311}$	−1.0	1.0	1.0	−1.0	−1.0	1.0	1.0
$J_{305,308}$	−1.0	1.0	1.0	1.0	−1.0	−1.0	−1.0
$J_{305,309}$	−1.0	−1.0	1.0	1.0	−1.0	−1.0	−1.0
$J_{305,310}$	−1.0	1.0	−1.0	−1.0	1.0	1.0	−1.0
$J_{305,311}$	−1.0	−1.0	1.0	1.0	1.0	−1.0	−1.0
$J_{306,308}$	1.0	1.0	−1.0	−1.0	−1.0	−1.0	1.0
$J_{306,309}$	−1.0	−1.0	1.0	1.0	−1.0	1.0	1.0
$J_{306,310}$	1.0	1.0	−1.0	1.0	−1.0	−1.0	1.0
$J_{306,311}$	1.0	1.0	1.0	1.0	1.0	1.0	1.0
$J_{307,308}$	1.0	1.0	−1.0	−1.0	−1.0	−1.0	1.0
$J_{307,309}$	1.0	1.0	1.0	1.0	−1.0	1.0	−1.0
$J_{307,310}$	−1.0	−1.0	−1.0	1.0	−1.0	1.0	−1.0
$J_{307,311}$	1.0	−1.0	−1.0	−1.0	1.0	−1.0	1.0

where H is a three-qubit Hamiltonian with independent binary qubit noise realized through binary random variables s_i:

$$H = -J_{\text{in}}(\widehat{\sigma}_1^z\widehat{\sigma}_2^z + \widehat{\sigma}_2^z\widehat{\sigma}_3^z) - \sum_{i=1}^{3}\gamma_i J_{\text{in}}\widehat{\sigma}_i^x - \sum_{i=1}^{3}\eta_i s_i\widehat{\sigma}_i^z. \quad (8)$$

Equation (8) describes a three-qubit Ising chain, where J_{in} is a single-input parameter controlling the strength of interactions, while γ_i and η_i are constants rescaling the strength of the transverse field and qubit noise. Leveraging the parameters estimated in Ref. [17], we select $\beta = 11$, $\gamma_i = 0.013$ and $\eta_i = 0.04$ as typical values for these parameters. We would like to highlight that the model from Eq. (8) is very different quantitatively and qualitatively from what is known as the background-susceptibility model, as discussed in Appendix F.

It has been shown in Ref. [17] that for small values of the input parameter J_{in}, the negative spurious coupling can be explained by field noise on the involved qubits. In fact, it is information-theoretically impossible to distinguish between a model with field noise and a model with the corresponding antiferromagnetic coupling. As seen in Fig. 5(b), the toy model in Eq. (7) indeed predicts that when J_{in} is small and thus field noise is significant relative to J_{in}, a negative spurious coupling is preset. We therefore designate this low-scaling regime where $J_{\text{in}} < 0.2$ as the noisy regime. When J_{in} increases, the field noise becomes less pronounced relative to J_{in}. In this regime of higher input values, the model given in Eq. (7) predicts an emergence of a positive spurious coupling—an effect

TABLE III. Input coupling and field values for GSD-F models before scaling by α_{in}. The D-Wave programming convention is used, where negative couplings indicate ferromagnetic and positive couplings indicate antiferromagnetic.

	GSD-F-1	GSD-F-2	GSD-F-3	GSD-F-4	GSD-F-5	GSD-F-6
h_{296}	1.0	1.0	1.0	−1.0	1.0	−1.0
h_{297}	1.0	−1.0	1.0	1.0	−1.0	−1.0
h_{298}	1.0	1.0	−1.0	1.0	−1.0	1.0
h_{299}	1.0	−1.0	−1.0	1.0	−1.0	1.0
h_{300}	−1.0	−1.0	−1.0	−1.0	−1.0	1.0
h_{301}	−1.0	1.0	−1.0	1.0	1.0	1.0
h_{302}	1.0	1.0	1.0	1.0	−1.0	1.0
h_{303}	−1.0	1.0	1.0	−1.0	1.0	−1.0
h_{304}	−1.0	−1.0	1.0	−1.0	−1.0	1.0
h_{305}	−1.0	1.0	−1.0	1.0	1.0	1.0
h_{306}	1.0	−1.0	−1.0	−1.0	−1.0	1.0
h_{307}	1.0	1.0	−1.0	1.0	1.0	−1.0
h_{308}	−1.0	1.0	−1.0	−1.0	1.0	1.0
h_{309}	1.0	−1.0	1.0	−1.0	1.0	1.0
h_{310}	−1.0	1.0	−1.0	−1.0	−1.0	−1.0
h_{311}	−1.0	−1.0	1.0	−1.0	−1.0	−1.0
$J_{296,301}$	−1.0	1.0	1.0	−1.0	−1.0	1.0
$J_{296,302}$	1.0	−1.0	−1.0	1.0	1.0	1.0
$J_{296,303}$	−1.0	−1.0	−1.0	1.0	−1.0	−1.0
$J_{297,300}$	1.0	1.0	1.0	−1.0	1.0	−1.0
$J_{297,301}$	−1.0	−1.0	−1.0	1.0	−1.0	1.0
$J_{297,302}$	1.0	1.0	1.0	1.0	1.0	−1.0
$J_{297,303}$	−1.0	1.0	−1.0	−1.0	1.0	−1.0
$J_{298,300}$	1.0	1.0	1.0	−1.0	−1.0	−1.0
$J_{298,301}$	−1.0	1.0	−1.0	−1.0	−1.0	1.0
$J_{298,302}$	−1.0	−1.0	−1.0	1.0	−1.0	−1.0
$J_{298,303}$	−1.0	−1.0	1.0	−1.0	−1.0	−1.0
$J_{299,300}$	−1.0	1.0	1.0	1.0	1.0	−1.0
$J_{299,301}$	−1.0	1.0	−1.0	−1.0	1.0	1.0
$J_{299,302}$	1.0	−1.0	−1.0	−1.0	−1.0	1.0
$J_{299,303}$	1.0	1.0	1.0	1.0	1.0	1.0
$J_{300,308}$	−1.0	−1.0	1.0	−1.0	1.0	1.0
$J_{301,309}$	−1.0	−1.0	−1.0	−1.0	1.0	1.0
$J_{302,310}$	1.0	1.0	1.0	1.0	−1.0	−1.0
$J_{303,311}$	−1.0	−1.0	−1.0	−1.0	−1.0	1.0
$J_{304,308}$	−1.0	1.0	1.0	1.0	−1.0	−1.0
$J_{304,309}$	−1.0	1.0	1.0	1.0	1.0	−1.0
$J_{304,310}$	1.0	1.0	1.0	−1.0	−1.0	1.0
$J_{304,311}$	1.0	1.0	1.0	−1.0	−1.0	1.0
$J_{305,308}$	−1.0	−1.0	1.0	−1.0	1.0	−1.0
$J_{305,309}$	1.0	−1.0	−1.0	1.0	−1.0	−1.0
$J_{305,310}$	−1.0	−1.0	−1.0	−1.0	1.0	1.0
$J_{305,311}$	1.0	−1.0	1.0	1.0	−1.0	−1.0
$J_{306,308}$	−1.0	1.0	1.0	−1.0	1.0	−1.0
$J_{306,309}$	1.0	1.0	−1.0	−1.0	−1.0	1.0
$J_{306,310}$	1.0	1.0	1.0	1.0	1.0	−1.0
$J_{306,311}$	1.0	−1.0	−1.0	1.0	1.0	1.0
$J_{307,308}$	1.0	1.0	1.0	−1.0	−1.0	1.0
$J_{307,309}$	−1.0	−1.0	−1.0	−1.0	−1.0	1.0
$J_{307,310}$	−1.0	1.0	1.0	−1.0	1.0	1.0
$J_{307,311}$	−1.0	−1.0	−1.0	−1.0	1.0	−1.0

that is also observed in the experimental data. This suggests that a J-dependent residual transverse field on each of the qubits can account for the positive response in the spurious link and the saturation of that link for large inputs. Similarly to how we designate the low-scaling regime as noise-dominated, we can designate the high-scaling

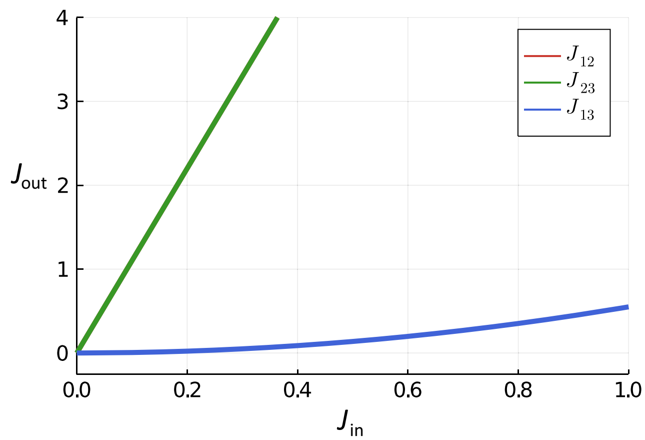

FIG. 6. Input and spurious coupling values as a function of the input coupling intensity according to the background-susceptibility model.

regime as dominated by the effect of residual transverse fields.

An interesting intermediate regime predicted by the toy model as shown in Fig. 5 happens around $J_{\text{in}} = 0.275$, where the noise and transverse-field influences cancel each other, leading to an effective absence of spurious couplings, which also occurs experimentally. This observation suggests that the optimal sampling parameter range for α_{in}, referred to as the *sweet spot* previously in this work, may be explained by the interacting effects of noise and residual fields in the system.

V. CONCLUSION

Inspired by recent work indicating that quantum annealing hardware behaves as a noisy Gibbs sampler in very low energy scales [17], this work identifies an approach for mitigating the impacts of noise and conducting high-quality Gibbs sampling in a range of effective temperatures for a class of small hardware-native Ising models. This approach to using quantum annealing hardware for Gibbs sampling opens opportunities for applications in machine learning and exploring the physics of condensed matter systems.

More broadly, the computational task of Gibbs sampling and the protocol developed in this work could provide an avenue for exploring the potential for quantum advantage in quantum annealing hardware. To that end, two follow-on works would be required. First, a class of Ising models needs to be identified that are challenging to sample from with classical algorithms (e.g., spin glasses) that also adhere to the criteria proposed in this work, i.e., naively representable on the hardware with $J, h \in \{-1, 0, 1\}$. Such examples seem unlikely in the Chimera hardware architecture [59,62]; however, the recent Pegasus hardware architecture [63] will likely provide opportunities for identifying such models. Second, significant additional research is required to verify that the protocol proposed by this work will scale to larger systems, ideally with hundreds to thousands of qubits, so that more of the quantum annealing hardware can be used. This type of verification requires an amazing amount of computation and tuning of sophisticated Monte Carlo methods, as there are no known efficient algorithms for generating Gibbs samples of the target distributions that would be required as a baseline for comparison. Another practical challenge is that scaling to larger systems may also yield unexpected side effects in practice, such as amplifying the role of hardware programming errors [64,65], putting an implicit limit on sampling accuracy at larger scales. Although significant follow-on investigation is required to more deeply understand the potential for performing thermal Gibbs sampling with quantum annealing hardware, this work provides a foundation for maximizing the performance of available hardware platforms when conducting these computational tasks.

ACKNOWLEDGMENTS

This work was supported by the U.S. Department of Energy (DOE) through the Laboratory Directed Research and Development (LDRD) program of the Los Alamos National Laboratory (LANL) under Project No. 20210114ER and the Center for NonLinear Studies (CNLS). This research used computing resources provided by the LANL Institutional Computing Program, which is supported by the DOE National Nuclear Security Administration under Contract No. 89233218CNA000001. This material is also based upon work supported by the National Science Foundation Quantum Leap Big Idea under Grant No. OMA-1936388.

APPENDIX A: COMPARING DISTRIBUTIONS WITH D_{TV} DISTANCE

Given two distributions μ and ν over a binary string of size N, i.e., $\sigma \in \{-1, 1\}^N$, the D_{TV} distance between them is defined as

$$D_{\text{TV}}(\mu, \nu) = \frac{1}{2} \sum_{\sigma \in \{-1,1\}^N} |\mu(\sigma) - \nu(\sigma)|, \tag{A1}$$

that is, the absolute difference between the probability of each state is added together and divided by 2, to normalize the metric to the range of 0 to 1. Given that the metric is in the range of 0 to 1, it is often presented as a percentage between 0 and 100.

The D_{TV} distance can be interpreted as the maximum discrepancy between the probability of an event computed

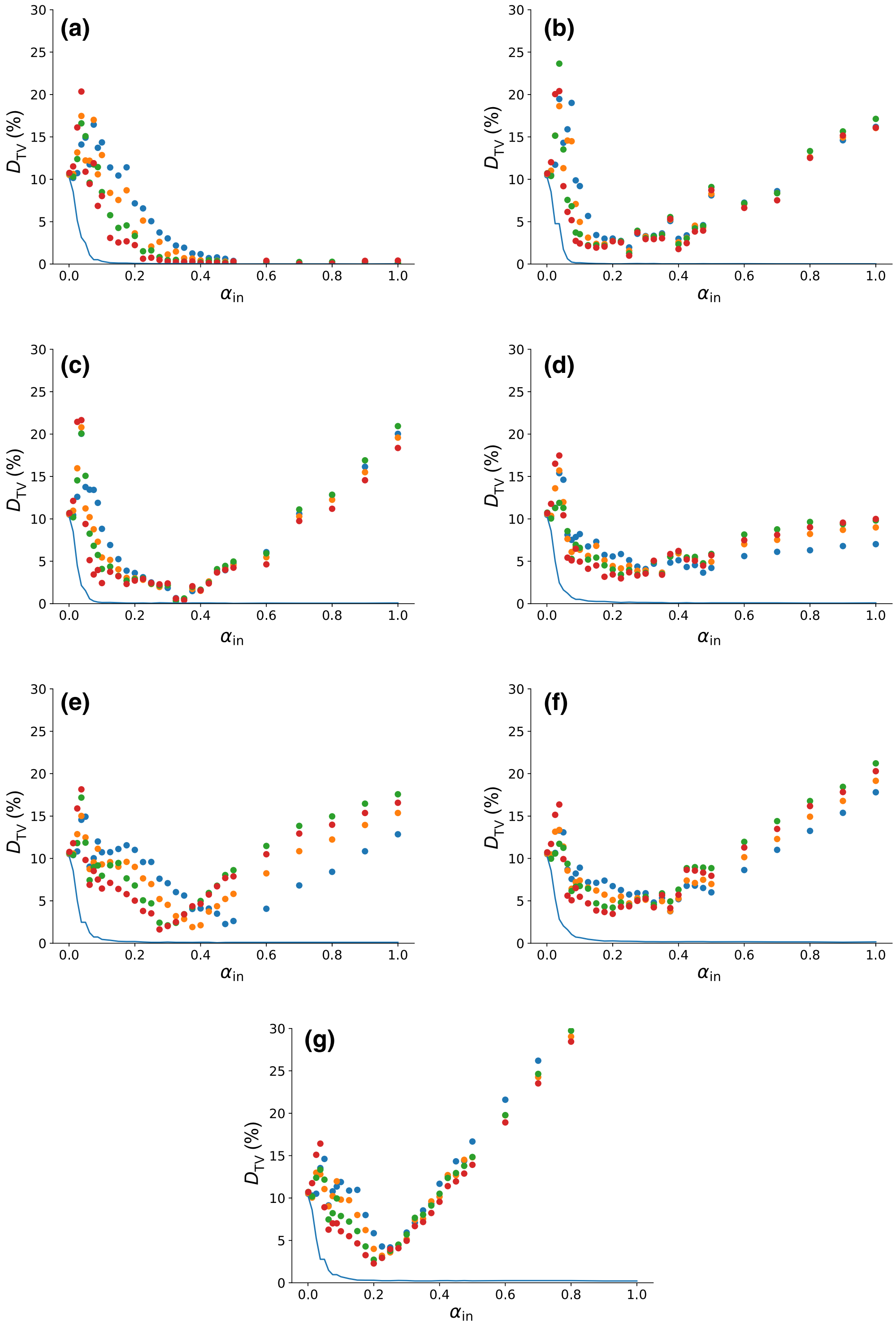

FIG. 7. The dependence of the D_{TV} distance on α_{in} for GSD models. Recall from Fig. 1 the color notation for various anneal times (1 μs, blue; 5 μs, orange; 25 μs, green; 125 μs, red). The D_{TV} lower bound is a result of finite sampling is indicated by the blue line. (a) GSD-2, (b) GSD-4, (c) GSD-6, (d) GSD-8, (e) GSD-10, (f) GSD-24, and (g) GSD-38.

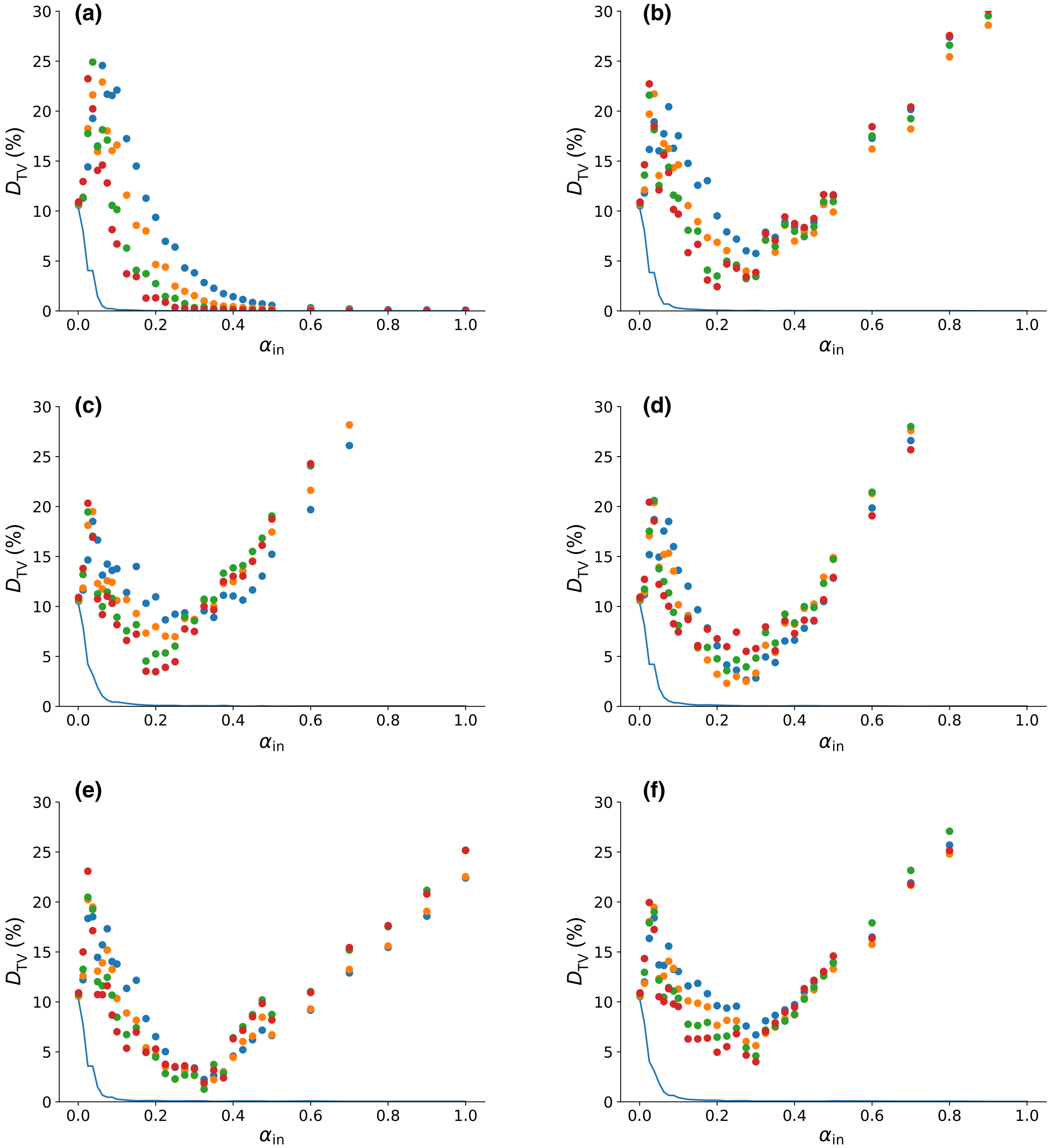

FIG. 8. Dependence of the D_{TV} distance on α_{in} for GSD-F models. Recall from Figs. 1 and 7 the color notation for various anneal times (1 μs, blue; 5 μs, orange; 25 μs, green; 125 μs, red). The D_{TV} lower bound resulting from finite sampling is indicated by the blue line. (a) GSD-F-1, (b) GSD-F-2, (c) GSD-F-3, (d) GSD-F-4, (e) GSD-F-5, and (f) GSD-F-6.

with the distribution μ instead of ν, that is,

$$\sup_A |\mathbb{P}_\mu(A) - \mathbb{P}_\nu(A)| = D_{\mathrm{TV}}(\mu, \nu), \tag{A2}$$

where $A \in \mathcal{P}\left(\{-1,1\}^N\right)$ is an element of the power set of $\{-1,1\}^N$. Therefore, it becomes the distance of choice to measure the error that can be made by estimating probabilities using a surrogate distribution ν of a target distribution μ. However, note that this sets a very strong quality criteria for estimating the reliability of a surrogate distribution, as the D_{TV} is equal to the worse-case estimation error. It is important to note that if μ or ν are only accessible through a set of finite samples, there is an unavoidable error caused

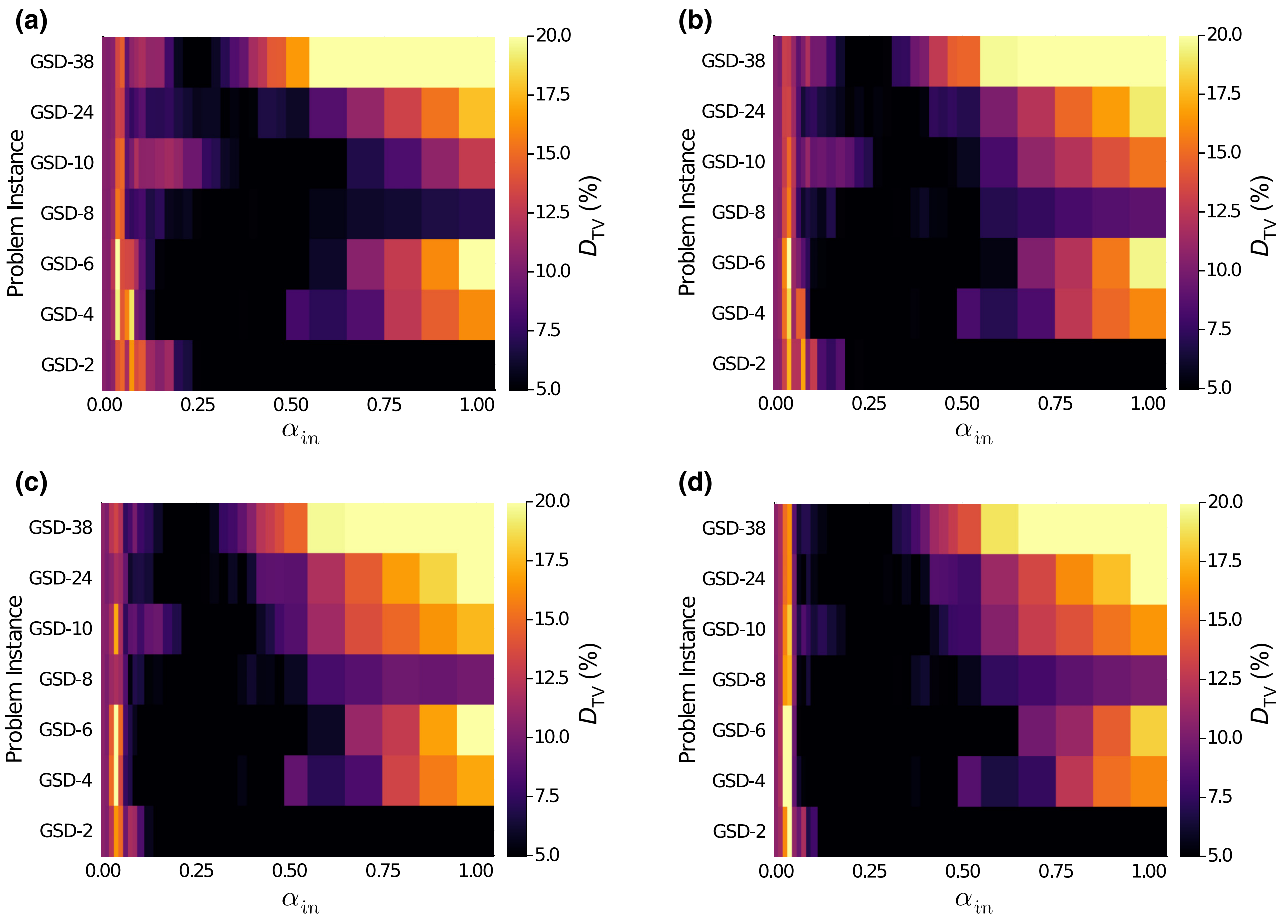

FIG. 9. A heat-map summary of the dependence of the D_{TV} distance on α_{in} and the annealing time for all GSD models. (a) 1 μs, (b) 5 μs, (c) 25 μs, and (d) 125 μs.

by sampling that will be captured by the D_{TV} distance. To avoid finite sampling error in the D_{TV} estimate, one needs to use a number of samples that is on the order of the typical support of μ and ν. This is required to estimate the absolute difference in Eq. (A1) accurately for each state σ. Typically, the number of samples required grows exponentially in N and is maximal for distribution μ and ν that are close to a uniform distribution, where it is in the order of 2^N.

It is fairly common to consider Kullback-Leibler (D_{KL}) divergence as an alternative to the D_{TV} distance, especially when the nonconvexity of the latter becomes a computational bottleneck. However, the D_{KL} divergence does not share the same meaning as the D_{TV} distance: the D_{KL} divergence estimates the inefficiency in compressing a source μ using a code optimized for ν. In general, it is not suitable for measuring the quality of probability estimates between a target distribution and its surrogate. A simple situation that illustrates this last point is when the distribution ν is almost identical to μ except that it has an infinitesimally small probability mass that lies outside of the support of μ. The D_{TV} distance between these distribution is infinitesimally small, while the D_{KL} divergence becomes infinite.

APPENDIX B: THE GROUND-STATE-DEGENERACY MODELS

The Ising-model instances used in this work are generated by the following procedure. First, 16 spins (two unit cells) on the D-Wave 2000Q Quantum Annealer `DW_2000Q_LANL` hardware graph are selected, with qubits 296–303 making up one unit cell and qubits 304–311 making up an adjacent unit cell in the Chimera graph architecture. Then the values for J_{ij} incident to these qubits are assigned randomly from the set $\{-1, 1\}$ for each coupler in the hardware graph. Second, a brute-force calculation is used to compute the ground-state degeneracy for a given instance [66]. This procedure is repeated many times and a subset of cases are selected to highlight a smooth range of ground-state degeneracies. This yields a total of seven GSD cases with degeneracy values of 2, 4, 6, 8, 10, 24, and 38, which are presented in Table II. The same procedure is repeated while also assigning random

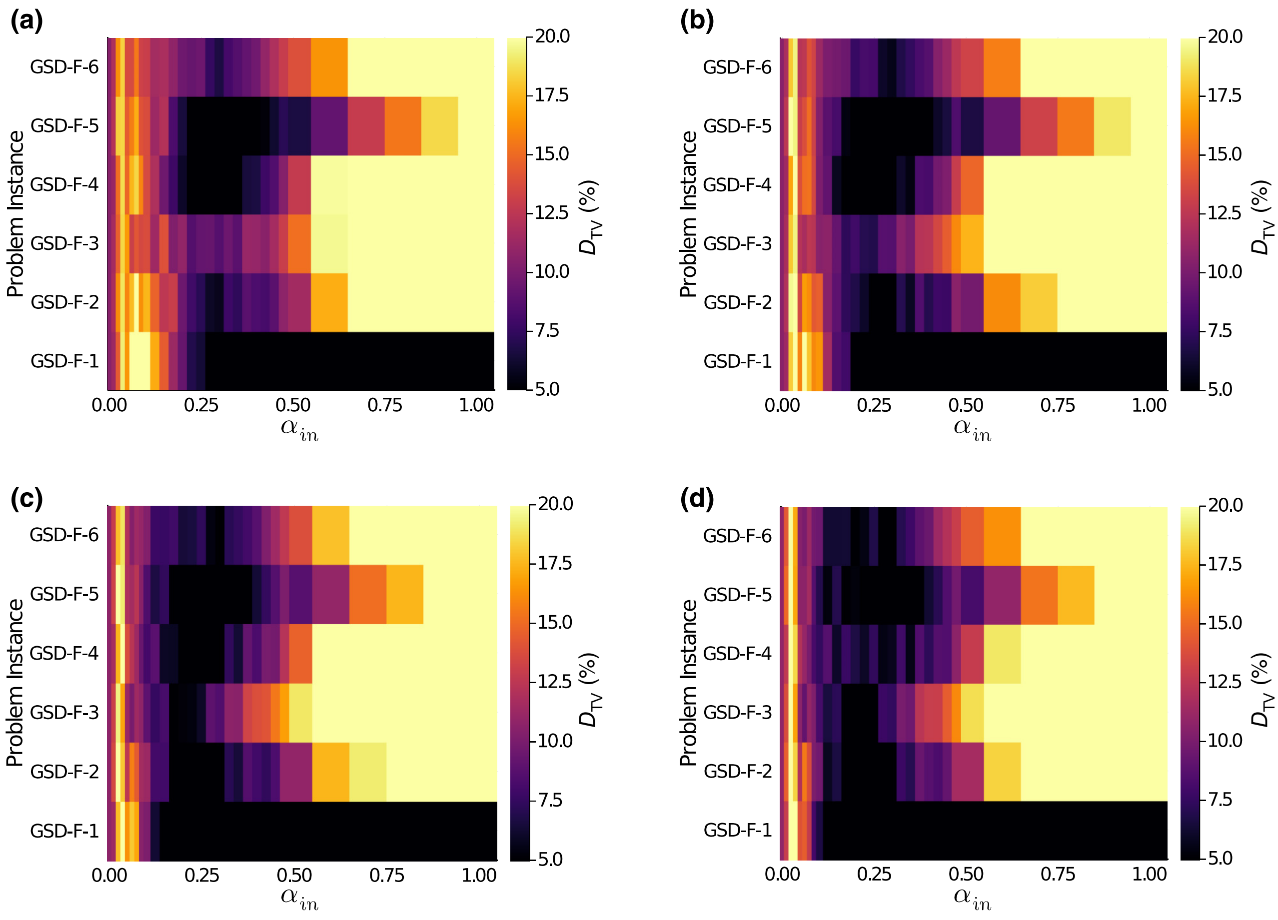

FIG. 10. A heat-map summary of the dependence of the D_{TV} distance on α_{in} and the annealing time for all GSD-F models. (a) 1 μs, (b) 5 μs, (c) 25 μs, and 125 μs.

values to the local fields h_i to $\{-1, 1\}$, yielding six additional independently generated GSD-F models with one, two, three, four, five, and six ground-state degeneracies, which are presented in Table III. The specific instances are referred to as GSD-*N* and GSD-F-*N*, where the "*N*" indicates the amount of ground-state degeneracy. Note that the GSD-F models tend to have lower degeneracy because the inclusion of the fields breaks symmetries within energy levels.

It is worth noting that these models are reminiscent of the random (RAN) and random with fields (RANF) models that have been used to benchmark earlier generations of QA processors [67,68]. The RAN-model class on Chimera graphs has been shown to exhibit a zero-temperature phase transition [59,62], indicating that sampling from this model class at finite temperature is not expected to be computationally hard. Here, however, instances are hand-picked from this class to exhibit varying degrees of ground-state degeneracy. In this sense, we do not expect that the instances of this work to be representative samples from the RAN class, and instead expect the high-degeneracy instances to be challenging for a quantum annealer to sample from due to the degeneracy-breaking effect of the transverse field.

APPENDIX C: D_{TV} DISTANCE AT DIFFERENT SCALES AND ANNEALING TIMES

In this appendix, we present additional figures for each of the experiments in the main text. Figures 7 and 8 display the D_{TV}-distance results for every GSD model as a function of α_{in}. As in Fig. 1, where only GSD-6 is shown, the results for each of the annealing times are overlaid with different colors. Note how GSD-2 and GSD-F-1 are the only instances where the D_{TV} distance decreases as α_{in} increases. This is because models with low degeneracy are easier to sample from in the high-scale regime, which has been previously explained in Sec. III B 2. For the rest of the instances, the consistent U-shaped trend supports our claim that the most effective α_{in} for achieving high accuracy is in the medium-scale regime. This optimal band between α_{in} of 0.2 and 0.4 remains consistent across all GSD models. In addition, these figures show that varying the annealing time does not shift the optimal region by a significant

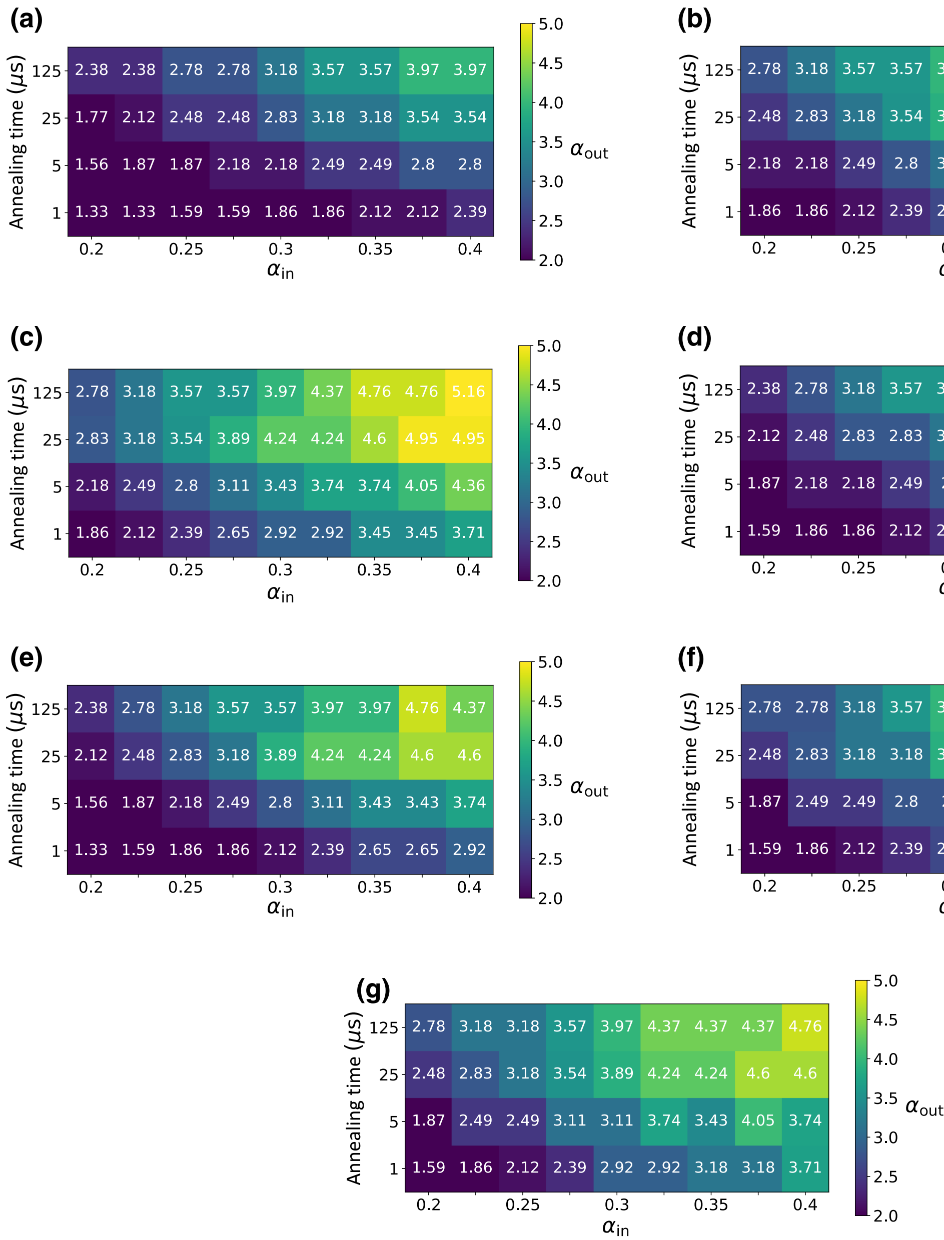

FIG. 11. The dependence of the effective temperature α_{out} on α_{in} and the annealing time for all GSD models. (a) GSD-2, (b) GSD-4, (c) GSD-6, (d) GSD-8, (e) GSD-10, (f) GSD-24, and (g) GSD-38.

amount, with the exception of GSD-10. However, there is a slight yet noticeable shift to the left in almost all instances as the anneal times increase.

Figures 9 and 10 communicate the same information as Figs. 7 and 8, but more directly compares various GSD models by stacking their D_{TV}-distance values in a heat map. Only the 1 μs plot is presented in Fig. 3 in the main text but here the results for all anneal times are shown. Once again, it is evident that the optimal region for α_{in} remains the same for various anneal times.

APPENDIX D: EFFECTIVE TEMPERATURES AT DIFFERENT SCALES AND ANNEALING TIMES

Figures 11 and 12 present the α_{out} values that correspond to the optimal α_{in} range between 0.2 and 0.4. The minimum and maximum value in each heat map is used to construct Table I. These figures show how α_{in} and the anneal time can be used to smoothly tune α_{out} and that the effect is consistent across all GSD models. Although the exact α_{out} values vary slightly from model to model,

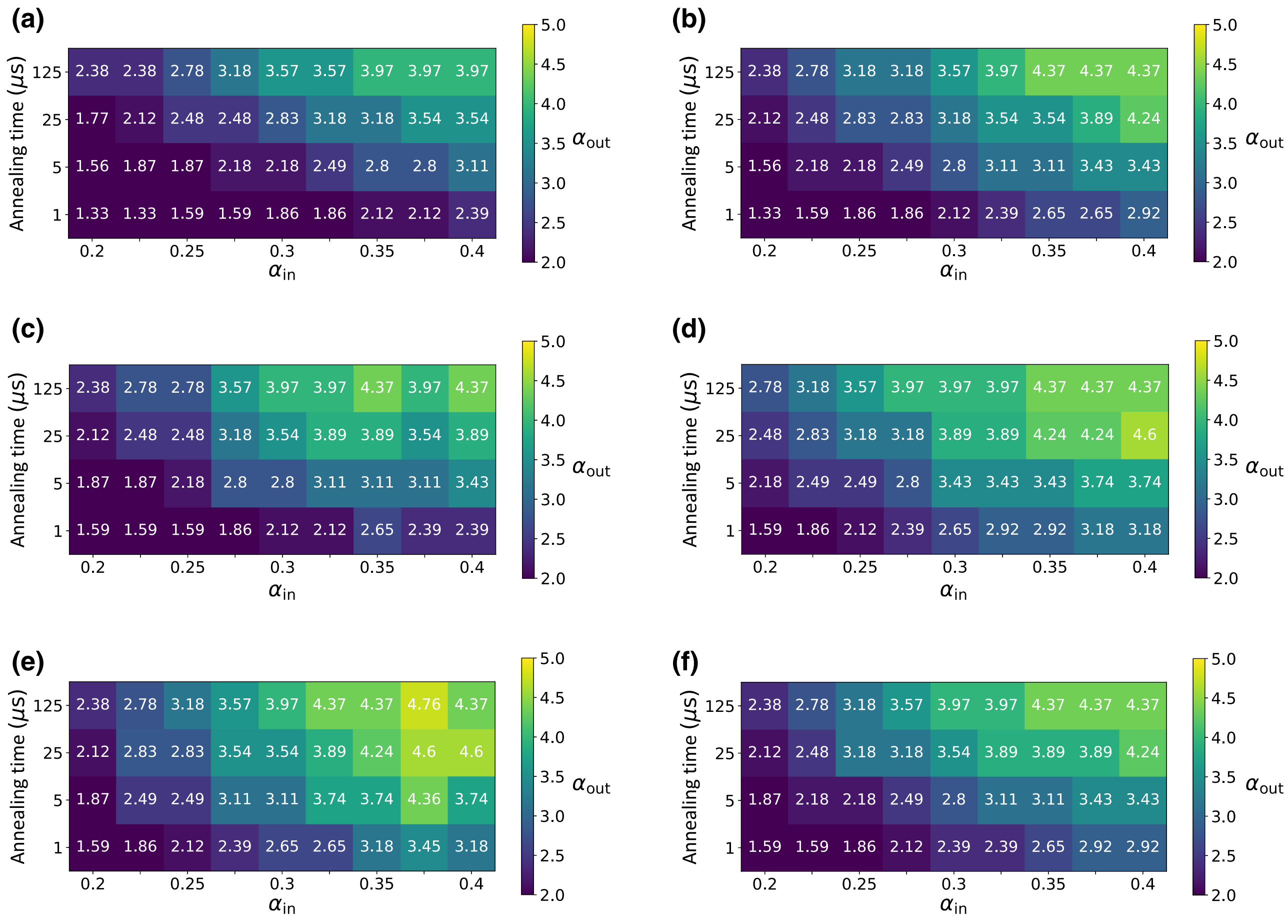

FIG. 12. The dependence of the effective temperature α_{out} on α_{in} and the annealing time for all GSD-F models. (a) GSD-F-1, (b) GSD-F-2, (c) GSD-F-3, (d) GSD-F-4, (e) GSD-F-5, and (f) GSD-F-6.

the trend is predictable and so these figures suggest a straightforward approach for producing Gibbs samples at desired α_{out} value.

APPENDIX E: THREE-SPIN ISING CHAIN ANNEALING TIMES

Finally, Fig. 13 shows results for the three-spin Ising chain experiment for various anneal times. The most important feature of these plots is the point at which the spurious coupling reduces to zero, as we hypothesize that this is the optimal input scale for sampling. From the plots, this optimal J_{in} value slightly decreases from 0.275 to a low of 0.2 as the anneal time increases. However, this value still remains within the optimal region seen in the 16-spin scaling experiments. As the anneal time increases, both programmed and spurious couplings appear to increase more rapidly with the increase of J_{in}. This increase results in the spurious coupling almost saturating by the time J_{in} reaches 0.4 for anneal times of 25 μs and 125 μs. However, the majority of our claimed optimal region between 0.2 and 0.4 remains in the regime where the spurious coupling is small compared to the programmed couplings.

APPENDIX F: BACKGROUND-SUSCEPTIBILITY MODEL FOR THE THREE-SPIN ISING CHAIN

It is believed that physically neighboring qubits are not perfectly isolated from each other and give rise to uncontrolled interactions. These spurious couplings are described by the background-susceptibility model [45] and take the form of an interaction that is proportional to the coupling intensities emanating from neighboring qubits. For the three-spin Ising chain discussed in Sec. IV, it results in the following Hamiltonian with background susceptibility:

$$H_{\text{BS}} = -J_{\text{in}}\widehat{\sigma}_1^z\widehat{\sigma}_2^z - J_{\text{in}}\widehat{\sigma}_2^z\widehat{\sigma}_3^z - \chi J_{\text{in}}^2\widehat{\sigma}_1^z\widehat{\sigma}_3^z, \tag{F1}$$

where $\chi > 0$ is the background-susceptibility constant between qubits one and three. (Note that in Ref. [45], χ is a negative quantity, following the D-Wave programming convention that ferromagnetic couplings are negative.) Note that the intensity of the spurious link between σ_1 and σ_3 is quadratic in the input coupling intensity J_{in} and is of ferromagnetic nature when J_{in} is positive.

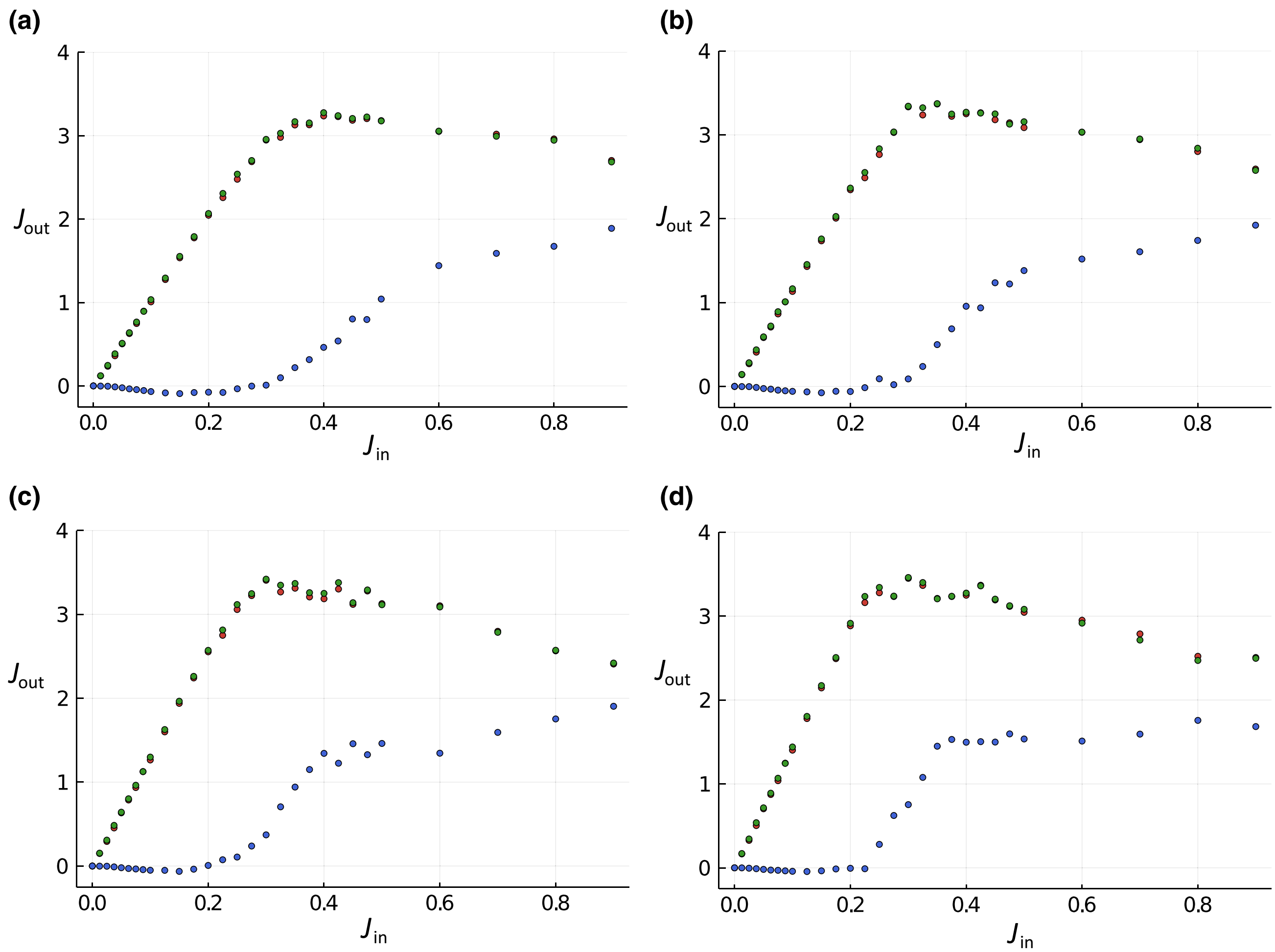

FIG. 13. Three-spin experiments for various annealing times. Green and red denote the programmed edges J_{12} and J_{23}, respectively. Blue represents the spurious coupling J_{13}. (a) 1 μs, (b) 5 μs, (c) 25 μs, and (d) 125 μs.

Because the Hamiltonian in Eq. (F1) is diagonal, we immediately see that this model behaves very differently than the effective Hamiltonian found experimentally (see Fig. 5). The background-susceptibility model predicts no saturation in the input coupling or spurious coupling intensity, as the former grows linearly and the latter grows quadratically. Moreover, the spurious coupling of the background-susceptibility model remains ferromagnetic, unlike what is measured experimentally for small values of J_{in}. This behavior is depicted in Fig. 6 for the typically encountered values of $\beta = 11$ and $\chi = 0.05$.

[1] A. Sinclair and M. Jerrum, Approximate counting, uniform generation and rapidly mixing Markov chains, Inf. Comput. **82**, 93 (1989).
[2] M. R. Jerrum, L. G. Valiant, and V. V. Vazirani, Random generation of combinatorial structures from a uniform distribution, Theor. Comput. Sci. **43**, 169 (1986).
[3] M. Jerrum and A. Sinclair, Polynomial-time approximation algorithms for the Ising model, SIAM J. Comput. **22**, 1087 (1993).
[4] G. Gallavotti, *Statistical Mechanics* (Springer, Berlin, Germany, 1999).
[5] J. Job and D. Lidar, Test-driving 1000 qubits, Quantum Sci. Technol. **3**, 030501 (2018).
[6] A. Perdomo-Ortiz, M, Benedetti, J. Realpe-Gómez, and R. Biswas, Opportunities and challenges for quantum-assisted machine learning in near-term quantum computers, Quantum Sci. Technol. **3**, 030502 (2018).
[7] T. Inagaki, Y. Haribara, K. Igarashi, T. Sonobe, S. Tamate, T. Honjo, A. Marandi, P. L. McMahon, T. Umeki, and K. Enbutsu *et al.*, A coherent Ising machine for 2000-node optimization problems, Science **354**, 603 (2016).
[8] F. Arute, K. Arya, R. Babbush, D. Bacon, J. C. Bardin, R. Barends, R. Biswas, S. Boixo, F. G. S. L. Brandao, D. A. Buell *et al.*, Quantum supremacy using a programmable superconducting processor, Nature **574**, 505 (2019).
[9] T. Kadowaki and H. Nishimori, Quantum annealing in the transverse Ising model, Phys. Rev. E **58**, 5355 (1998).

[10] J. Brooke, D. Bitko, T. F. Rosenbaum, and G. Aeppli, Quantum annealing of a disordered magnet, Science **284**, 779 (1999).

[11] J. Brooke, T. F. Rosenbaum, and G. Aeppli, Tunable quantum tunnelling of magnetic domain walls, Nature **413**, 610 (2001).

[12] E. Farhi, J. Goldstone, S. Gutmann, J. Lapan, A. Lundgren, and D. Preda, A quantum adiabatic evolution algorithm applied to random instances of an NP-complete problem, Science **292**, 472 (2001).

[13] Z. Bian, F. Chudak, W. G. Macready, and G. Rose, The Ising model: Teaching an old problem new tricks, Published online at https://www.dwavesys.com/sites/default/files/weightedmaxsat˙v2.pdf (2010), accessed: 04/28/2017.

[14] J. Raymond, S. Yarkoni, and E. Andriyash, Global warming: Temperature estimation in annealers, Frontiers in ICT **3**, 23 (2016).

[15] J. Marshall, E. G. Rieffel, and I. Hen, Thermalization, Freeze-Out, and Noise: Deciphering Experimental Quantum Annealers, Phys. Rev. Appl. **8**, 064025 (2017).

[16] J. Marshall, D. Venturelli, I. Hen, and E. G. Rieffel, Power of Pausing: Advancing Understanding of Thermalization in Experimental Quantum Annealers, Phys. Rev. Appl. **11**, 044083 (2019).

[17] M. Vuffray, C. Coffrin, Y. A. Kharkov, and A. Y. Lokhov, Programmable Quantum Annealers as Noisy Gibbs Samplers, PRX Quantum **3**, 020317 (2022).

[18] M. Benedetti, J. Realpe-Gómez, R. Biswas, and A. Perdomo-Ortiz, Estimation of effective temperatures in quantum annealers for sampling applications: A case study with possible applications in deep learning, Phys. Rev. A **94**, 022308 (2016).

[19] M. Benedetti, J. Realpe-Gómez, R. Biswas, and A. Perdomo-Ortiz, Quantum-Assisted Learning of Hardware-Embedded Probabilistic Graphical Models, Phys. Rev. X **7**, 041052 (2017).

[20] J. J. Hopfield, Neural networks and physical systems with emergent collective computational abilities, Proc. Natl. Acad. sci. **79**, 2554 (1982).

[21] D. K. Panjwani and G. Healey, Markov random field models for unsupervised segmentation of textured color images, IEEE Trans. Pattern Anal. Mach. Intell. **17**, 939 (1995).

[22] A. Y. Lokhov, M. Vuffray, S. Misra, and M. Chertkov, Optimal structure and parameter learning of Ising models, Sci. Adv. **4**, e1700791 (2018).

[23] G. Kochenberger, J.-K. Hao, F. Glover, M. Lewis, Z. Lü, H. Wang, and Y. Wang, The unconstrained binary quadratic programming problem: A survey, J. Comb. Optim. **28**, 58 (2014).

[24] Throughout this work, α is used to denote the energy scale of a Gibbs distribution and β is used to indicate the effective inverse temperature of the quantum annealer.

[25] A. Einstein, Beiträge zur Quantentheorie, Verh. d. D. Phys. Ges. **16**, 826 (1914).

[26] M. Born and V. Fock, Beweis des Adiabatensatzes, Zeit. f. Physik **51**, 165 (1928).

[27] P. Ehrenfest, Over adiabatische veranderingen van een stelsel in verband met de theorie der quanta, Verslagen Kon. Akad. Amserdam **25**, 412 (1916).

[28] T. Kato, On the adiabatic theorem of quantum mechanics, J. Phys. Soc. Jap. **5**, 435 (1950).

[29] S. Jansen, M.-B. Ruskai, and R. Seiler, Bounds for the adiabatic approximation with applications to quantum computation, J. Math. Phys. **48**, 102111 (2007).

[30] Y. Matsuda, H. Nishimori, and H. G. Katzgraber, Quantum annealing for problems with ground-state degeneracy, J. Phys.: Conference Series **143**, 012003 (2009).

[31] S. Boixo, T. Albash, F. M. Spedalieri, N. Chancellor, and D. A. Lidar, Experimental signature of programmable quantum annealing, Nat. Commun. **4**, 2067 (2013).

[32] B. H. Zhang, G. Wagenbreth, V. Martin-Mayor, and I. Hen, Advantages of unfair quantum ground-state sampling, Sci. Rep. **7**, 1044 (2017).

[33] M. S. Könz, G. Mazzola, A. J. Ochoa, H. G. Katzgraber, and M. Troyer, Uncertain fate of fair sampling in quantum annealing, Phys. Rev. A **100**, 030303 (2019).

[34] V. Kumar, C. Tomlin, C. Nehrkorn, D. O'Malley, and J. Dulny III, Achieving fair sampling in quantum annealing, arXiv:2007.08487 (2020).

[35] E. Pelofske, J. Golden, A. Bärtschi, D. O'Malley, and S. Eidenbenz, Sampling on NISQ devices: "Who's the fairest one of all?" arXiv:2107.06468 (2021).

[36] T. Albash and D. A. Lidar, Decoherence in adiabatic quantum computation, Phys. Rev. A **91**, 062320 (2015).

[37] S. Boixo, V. N. Smelyanskiy, A. Shabani, S. V. Isakov, M. Dykman, V. S. Denchev, M. H. Amin, A. Y. Smirnov, M. Mohseni, and H. Neven, Computational multiqubit tunnelling in programmable quantum annealers, Nat. Commun. **7**, 10327 (2016).

[38] A. Y. Smirnov and M. H. Amin, Theory of open quantum dynamics with hybrid noise, New J. Phys. **20**, 103037 (2018).

[39] W. K. Abou Salem, On the quasi-static evolution of nonequilibrium steady states, Annales Henri Poincaré **8**, 569 (2007).

[40] J. E. Avron, M. Fraas, G. M. Graf, and P. Grech, Adiabatic theorems for generators of contracting evolutions, Commun. Math. Phys. **314**, 163 (2012).

[41] L. C. Venuti, T. Albash, D. A. Lidar, and P. Zanardi, Adiabaticity in open quantum systems, Phys. Rev. A **93**, 032118 (2016).

[42] G. Lindblad, On the generators of quantum dynamical semigroups, Commun. Math. Phys. **48**, 119 (1976).

[43] E. B. Davies, Markovian master equations, Commun. Math. Phys. **39**, 91 (1974).

[44] T. Albash, S. Boixo, D. A. Lidar, and P. Zanardi, Quantum adiabatic Markovian master equations, New J. Phys. **14**, 123016 (2012).

[45] D-Wave Systems Inc., D-wave system documentation, Published online at https://docs.dwavesys.com/docs/latest/ (2020), accessed: 03/17/2021.

[46] J. Nelson, M. Vuffray, A. Y. Lokhov, and C. Coffrin, Single-qubit fidelity assessment of quantum annealing hardware, IEEE Trans. Quantum Eng. **2**, 1 (2021).

[47] M. H. Amin, Searching for quantum speedup in quasistatic quantum annealers, Phys. Rev. A **92**, 052323 (2015).

[48] M. H. Amin, E. Andriyash, J. Rolfe, B. Kulchytskyy, and R. Melko, Quantum Boltzmann Machine, Phys. Rev. X **8**, 021050 (2018).
[49] W. Winci, L. Buffoni, H. Sadeghi, A. Khoshaman, E. Andriyash, and M. H. Amin, A path towards quantum advantage in training deep generative models with quantum annealers, Machine Learning: Sci. Technol. **1**, 045028 (2020).
[50] R. Harris *et al.*, Phase transitions in a programmable quantum spin glass simulator, Science **361**, 162 (2018).
[51] A. D. King *et al.*, Observation of topological phenomena in a programmable lattice of 1,800 qubits, Nature **560**, 456 (2018).
[52] A. D. King, C. Nisoli, and E. D. Dahl, G. Poulin-Lamarre, and A. Lopez-Bezanilla, Qubit spin ice, Science **373**, 576 (2021).
[53] J. Marshall, A. Di Gioacchino, and E. G. Rieffel, Perils of embedding for sampling problems, Phys. Rev. Res. **2**, 023020 (2020).
[54] R. Y. Li, T. Albash, and D. A. Lidar, Limitations of error corrected quantum annealing in improving the performance of Boltzmann machines, Quantum Sci. Technol. **5**, 045010 (2020).
[55] T. Zaborniak and R, de Sousa, Benchmarking Hamiltonian noise in the D-Wave quantum annealer, IEEE Trans. Quantum Eng. **2**, 1 (2021).
[56] P. I. Bunyk, E. M. Hoskinson, M. W. Johnson, E. Tolkacheva, F. Altomare, A. J. Berkley, R. Harris, J. P. Hilton, T. Lanting, A. J. Przybysz, and J. Whittaker, Architectural considerations in the design of a superconducting quantum annealing processor, IEEE Trans. Appl. Supercond. **24**, 1 (2014).
[57] T. F. Rønnow, Z. Wang, J. Job, S. Boixo, S. V. Isakov, D. Wecker, J. M. Martinis, D. A. Lidar, and M. Troyer, Defining and detecting quantum speedup, Science **345**, 420 (2014).
[58] T. Albash, W. Vinci, A. Mishra, P. A. Warburton, and D. A. Lidar, Consistency tests of classical and quantum models for a quantum annealer, Phys. Rev. A **91**, 042314 (2015).
[59] H. G. Katzgraber, F. Hamze, and R. S. Andrist, Glassy Chimeras Could be Blind to Quantum Speedup: Designing Better Benchmarks for Quantum Annealing Machines, Phys. Rev. X **4**, 021008 (2014).
[60] Note when programming a D-Wave system, negative values of J_{in} are used to follow the ferromagnetic sign convention of that platform.
[61] M. Vuffray, S. Misra, A. Lokhov, and M. Chertkov, in *Advances in Neural Information Processing Systems*, Vol. 29, edited by D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett (Curran Associates, Inc., 2016).
[62] M. Weigel, H. G. Katzgraber, J. Machta, F. Hamze, and R. S. Andrist, (Octomore Collaboration), Erratum: Glassy Chimeras Could Be Blind to Quantum Speedup: Designing Better Benchmarks for Quantum Annealing Machines [Phys. Rev. x 4, 021008 (2014)], Phys. Rev. X **5**, 019901 (2015).
[63] K. Boothby, P. Bunyk, J. Raymond, and A. Roy, Next-Generation Topology of D-Wave Quantum Processors, arXiv e-prints, arXiv:2003.00133 (2020).
[64] Z. Zhu, A. J. Ochoa, S. Schnabel, F. Hamze, and H. G. Katzgraber, Best-case performance of quantum annealers on native spin-glass benchmarks: How chaos can affect success probabilities, Phys. Rev. A **93**, 012317 (2016).
[65] T. Albash, V. Martin-Mayor, and I. Hen, Analog errors in Ising machines, Quantum Science and Technology **4**, 02LT03 (2019).
[66] This approach is only computationally feasible for small systems spins, such the ones considered here.
[67] S. Boixo, T. F. Rønnow, S. V. Isakov, Z. Wang, D. Wecker, D. A. Lidar, J. M. Martinis, and M. Troyer, Evidence for quantum annealing with more than one hundred qubits, Nat. Phys. **10**, 218 (2014).
[68] C. Coffrin, H. Nagarajan, and R. Bent, in *Integration of Constraint Programming, Artificial Intelligence, and Operations Research*, edited by Louis-Martin Rousseau and Kostas Stergiou (Springer International Publishing, Cham, 2019), p. 163.