Seeking Quantum Speedup through Spin Glasses: The Good, the Bad, and the Ugly

Helmut G. Katzgraber,[1,2,3] Firas Hamze,[4] Zheng Zhu,[1] Andrew J. Ochoa,[1] and H. Munoz-Bauza[1]

[1]*Department of Physics and Astronomy, Texas A&M University, College Station, Texas 77843-4242, USA*
[2]*Materials Science and Engineering Program, Texas A&M University, College Station, Texas 77843, USA*
[3]*Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, New Mexico 87501, USA*
[4]*D-Wave Systems, Inc., 3033 Beta Avenue, Burnaby, British Columbia, V5G 4M9, Canada*

(Dated: September 3, 2015)

There has been considerable progress in the design and construction of quantum annealing devices. However, a conclusive detection of quantum speedup over traditional silicon-based machines remains elusive, despite multiple careful studies. In this work we outline strategies to design hard tunable benchmark instances based on insights from the study of spin glasses—the archetypal random benchmark problem for novel algorithms and optimization devices. We propose to complement head-to-head scaling studies that compare quantum annealing machines to state-of-the-art classical codes with an approach that compares the performance of different algorithms and/or computing architectures on different classes of computationally hard tunable spin-glass instances. The advantage of such an approach lies in having to compare only the performance hit felt by a given algorithm and/or architecture when the instance complexity is increased. Furthermore, we propose a methodology that might not directly translate into the detection of quantum speedup, but might elucidate whether quantum annealing has a "quantum advantage" over corresponding classical algorithms like simulated annealing. Our results on a 496-qubit D-Wave Two quantum annealing device are compared to recently used state-of-the-art thermal simulated annealing codes.

PACS numbers: 75.50.Lk, 75.40.Mg, 05.50.+q, 03.67.Lx

I. INTRODUCTION

Optimization plays an integral role across disciplines. Not only does modern manufacturing and transport heavily depend on efficient optimization methods to reduce cost and emissions, many fields of research depend on a multitude of optimization techniques to solve a wide variety of problems. Similarly, the ever-increasing amount of data available to mankind means an urgent need for more efficient approaches in querying, parsing, and mining data, approaches that often depend on optimization techniques. Within physics-related disciplines alone, optimization is needed to solve many difficult problems ranging from frustrated spin systems [1–3] to novel approaches in material discovery, as well as the efficient parsing of high-energy event data or astrophysical spectra. As such, the search for more efficient optimization approaches is of great importance. Because the speedup of current silicon-based computing technologies is slowly coming to an end mostly due to manufacturing and material constraints [4], interest in developing faster optimization methods has shifted to the development of new state-of-the-art algorithms, as well as novel computing paradigms, e.g., based on quantum architectures.

Quantum computing [5, 6] and, in particular, adiabatic quantum optimization [7–18] has gained increased momentum since D-Wave Systems Inc. introduced the D-Wave Two (DW2) quantum annealing device [19]. Inspired by the work of Santoro *et al.* [12], multiple teams have attempted to demonstrate that quantum adiabatic optimization—or quantum annealing (QA) [20–23]—has advantages over conventional thermal optimization techniques, such as, for example, simulated annealing (SA) [24]. The idea behind QA is to adiabatically quench quantum fluctuations to optimize a cost function (Hamiltonian) of a given complex optimization problem. Potentially, the wave function of the problem might be able to quantum tunnel through barriers in the free-energy landscape, i.e., QA might be able to outperform other approaches like SA where temperature fluctuations are slowly reduced to find the optimum. Towards the end of the annealing schedule in SA, when these temperature fluctuations are small, the system is unable to overcome free-energy barriers and, especially for problems with rough energy landscapes such as in spin glasses [25, 26] and related problems, it might become trapped in metastable states, thus missing the true optimum of the problem.

The fact that a broad range of hallmark optimization problems, such as the satisfiability problem (k-SAT), the number partitioning problem, vertex covers, knapsack problems, coloring problems, the traveling salesman problem, etc. can be mapped onto quadratic unconstrained binary optimization problems [27], means that devices that are tailored to solve these, such as the DW2, could revolutionize today's optimization efforts. Although not a fully programmable universal quantum computer, the D-Wave device represents a sizable advance in (quantum) computing.

The seminal work of Rønnow *et al*. [28] took great care and detail in defining the notion of *quantum speedup*. While at the moment the demonstration of *strong quantum speedup* remains a distant goal, the detection of *limited quantum speedup* [29]—a speedup relative to a given corresponding classical algorithm such as SA—seems more graspable. The number of studies (see, for example, Refs. [28, 30–33, 33]) attempting to detect quantum speedup is growing at a fast pace; however, the definite detection of quantum speedup remains elusive. So why, despite these large efforts, does quantum speedup remain to be demonstrated? Potentially, there are many reasons

why this might be the case. On one hand the complex circuitry, combined with the extreme fragility of quantum states to perturbations might be a source of decoherence and thus loss of any advantage over conventional techniques. On the other hand, the systems currently available (maximally 512 qubits on DW2, soon up to ~ 1000) might be too small for the benchmarks to be in the asymptotic scaling regime. However, a more mundane reason that is relatively easy to fix is the choice of the wrong benchmark problem. In Ref. [34], Katzgraber *et al*. demonstrated that the native benchmark to search for quantum speedup on a device like the DW2—an Ising spin glass with discrete uncorrelated disorder—is likely a problem that not only might be too easy to detect any speedup (think of two world-class skiers on a bunny slope), but the energy landscape of a spin glass on the DW2 Chimera topology [35] might actually favor thermal approaches like SA, simply because the spin-glass state exists only at zero temperature. Furthermore, the use of either bimodal or uniform range-k disorder [28, 31–33, 33] creates an energy landscape that has a *huge* number of configurations that minimize the cost function. As such, any method like SA run with multiple restarts will naturally excel in optimizing such a problem. Attempts to mitigate this issue by *planting* solutions [36] delivers problem instances that might not be challenging enough for both classical algorithms and quantum devices alike.

To overcome the limitations imposed by the small size of current devices, it is imperative to use a *native* benchmark problem that uses as many qubits N as possible on the device. Any *embedding* of a potentially harder problem [37] will further reduce the number of logical qubits, thus pushing the asymptotic regime farther away. Furthermore, it is hard to mitigate the effects of noise on both qubits and couplers without improving manufacturing. However, it is considerably easier to design hard benchmark instances that attempt to work around the flaws and limitations of the DW2 architecture. Reference [38] focuses on designing instance problems that are affected as little as possible by the chip's intrinsic noise. Here, we present a simple road map that uses insights from the study of spin glasses to design hard, as well as tunable, benchmark instances.

In addition, we propose to search for quantum advantages over classical architectures not only by comparing to state-of-the-art classical algorithms [39], but by studying the effects of tuning the instance complexity for a given type of disorder on both classical and quantum approaches. By studying the performance hit felt by the different approaches on carefully tailored problems with a free-energy landscape that is either dominated by large barriers or is reminiscent of a ferromagnetic system, further insights into the nature of quantum annealing devices can be gained. To perform a fair comparison across instances, here we fix the ground-state degeneracy (ideally) to 1 (or as low as possible) and vary the complexity of the free-energy landscape by using the spin-glass order parameter distribution as a proxy to the dominant features of the landscape [40, 41]. We show that, indeed, the spin-glass order parameter distribution produces tunable instances, and that predictions from the study of spin glasses on the complexity of the energy landscape allows us to produce problems on average considerably harder than any previous study.

We emphasize that we are *not* attempting to perform a scaling analysis as done in previous studies, simply because we believe that the currently accessible system sizes of up to 512 qubits are too small to be in the asymptotic limit [42]. We base this statement on previous simulations of two-dimensional Ising spin glasses on a square lattice at zero temperature with discrete disorder [43] where corrections to scaling due to the finite system sizes were very strong for systems with $\sim 10^3$ spins.

Our results show that the DW2 device is outperformed at finding the ground state by classical state-of-the-art optimization algorithms. However, there is a potential signature that the DW2 device might be able to optimize certain classes of carefully designed native spin-glass problems more efficiently than the classical counterpart SA, especially if noise is reduced. This suggests that the DW2 device potentially has a "quantum advantage" over corresponding classical algorithms like SA for certain problems. In addition, there are signs that the DW2 device might in some cases be more effective at generating low-lying states, as opposed to strict ground states than SA. Finally, our results suggest that "classical computational hardness" in spin glasses seems to carry over to quantum annealing devices, therefore facilitating the design of spin-glass-based instances. The day that quantum annealing machines have lower noise levels, higher connectivity to enable the simple embedding of spin-glass problems with, e.g., a finite transition temperature [34, 37], or a larger numbers of qubits, a combination of the approach presented in Ref. [28], with error-correction techniques [31, 44], and designer instances described in this work will likely show if quantum speedup is myth or reality.

The paper is structured as follows. In Sec. II, we introduce the native benchmark problem, followed by a detailed description of the limitations of current approaches as well as how we design hard instance problems in Sec. III. Section IV summarizes results on both the DW2 device, as well as classical simulation codes, followed by a discussion and summary. Appendix A outlines our experimental methodology on the DW2 device housed at D-Wave Systems Inc., followed by simulation details in Appendix B and numerical results in Appendix C. Appendix D summarizes less fruitful efforts experimenting with other instance classes.

II. NATIVE BENCHMARK: SPIN GLASSES

We illustrate our benchmarking ideas using the D-Wave Systems, Inc., D-Wave Two quantum annealing machine [45]. The *native* benchmark problem for the DW2 device is an Ising spin glass [6, 25–27] defined on the Chimera topology of the system [35],

$$\mathcal{H} = -\sum_{\{i,j\}\in\mathcal{V}} J_{ij} S_i^z S_j^z - \sum_{i\in\mathcal{V}} S_i^z h_i \,. \tag{1}$$

The N Ising spins $S_i^z \in \{\pm 1\}$ are defined on the vertices $\mathcal{V}$ of the Chimera lattice (see Fig. 7) and can be coupled to

a (local) field h_i. The sum is over all edges $\mathcal{E}$ connecting vertices $\{i, j\} \in \mathcal{V}$. In this study we set $h_i = 0\ \forall i$.

We emphasize that it is of paramount importance to study *native* problems that use as many qubits as possible to prevent overhead that might yield smaller embedded problems. At the moment, with approximately 500 (soon 1000) qubits at hand, it will be difficult to detect any quantum speedup. As such, our focus does not lie in performing a detailed scaling analysis with the problem size N, but to show how to select tunable hard problems that have the *same* disorder distribution, i.e., have the same strengths or weaknesses with respect to the intrinsic noise found in these devices. Tuning the complexity of the problem instances will then allow for a systematic testing of any potential advantages or disadvantages that the DW2 device might have over other architectures and/or simulation approaches. Note that in this study we disregard the effects of noise on the couplers and qubits and will report on these in a subsequent publication with strategies on how to mitigate the effects of perturbed problem Hamiltonians [38]. However, for the generated problems, the resilience to noise (robustness to perturbations) on the qubits and couplers is roughly similar and mostly agrees within error bars for the different instance subclasses that use interactions based on Sidon sets [46]; see Sec. III B for details. This means that the noise of the DW2 does not affect our results.

III. DESIGNING HARD INSTANCES

We start by describing the shortcomings of previous instances to detect quantum speedup and then outline our approach to produce tunable, hard instances.

In Ref. [34] it was shown that a spin glass on the Chimera topology has a zero-temperature phase transition. Although the worst case complexity of finding a ground state of an Ising spin glass on the Chimera graph falls into the NP hard class, performing any minimization of the energy based on any annealing approach will likely have a rather simple phase space to traverse for small system sizes because dominant barriers will not be as pronounced. Embedding problems that have a finite-temperature spin-glass transition is difficult, mainly due to the large overhead; i.e., only systems with few logical qubits can be studied because many physical qubits are needed to emulate long-range interactions. Because the resulting systems are small, the problems are far from the asymptotic regime to detect any quantum speedup in a scaling analysis.

A more promising route is thus to use insights from the study of spin glasses and carefully design the interactions between the qubits on the native Chimera graph, such that the problems are as hard as possible in order to challenge any optimization approach.

A. Problems with current approaches

In addition to a restrictive geometry, the D-Wave hardware has clear restrictions as to what values the interactions between the spins can have. This is rather limiting and, as such, only discrete and well-separated values of the couplers can be set. The simplest approach used in previous studies [28, 31–33, 33] is to select the disorder from a bimodal distribution, i.e., $J_{ij} \in \{\pm 1\}$ (we shall refer to these as U_1), followed by uniform range-k problems where the interactions J_{ij} are chosen from the integer set $\{\pm 1, \pm 2, \ldots, \pm k\}$. We refer to the latter as U_k. The problem with these choices for systems up to $N = 512$ variables is the huge degeneracy of the ground states that yields again benchmarks too simple to challenge any optimization approach (see Sec. IV). A simple analogy to this problem is a game of golf where the green has, for example, 10^7 holes. Hitting a hole in one is a trivial task! However, having a course with only one hole makes the sport truly challenging. As such, we design herein problems that—within the hardware restrictions of the machine—have a unique configuration that minimizes the Hamiltonian in Eq. (1).

Other approaches [36, 47] using planted solutions suffer from similar problems: While the instances are harder than for the problems in the U_k class, they often still have a large degeneracy and their complexity is not high enough for the current available systems of up to $\sim 10^3$ qubits. In particular, the very careful work presented in Ref. [36] shows a clear easy-hard-easy transition of the planted k-SAT solutions that could be exploited to generate hard instances. However, one problem that these instances have is that the disorder is not drawn from a particular distribution; i.e., two different planted k-SAT instances will likely have a very different (classical) energy spectrum and thus also be differently susceptible to the intrinsic noise found in the DW2 device [48]. Furthermore, we perform experiments with planted k-SAT solutions as presented in Ref. [36] using the benchmark codes in Ref. [39] and find that these instances are at times easier than the ones in the U_1 class. The authors of Ref. [36] do emphasize that harder problems must be designed to allow for the optimization of the annealing time, as well as the need to find problems where the benefits of quantum annealing can be assessed *ahead of time*.

Finally, setting the spin-spin interactions within the $\mathrm{K}_{4,4}$ unit cell of Chimera (see Fig. 7) to be of larger magnitude than those between the cells (often referred to as "cluster problems") has given DW2 an advantage over classical codes in a scaling analysis [49] when cluster Monte Carlo updates are not allowed. However, by design, simulated annealing (and any other Monte Carlo-like simple-sampling variation) will have a large disadvantage. The addition of simple clusterlike moves would again give classical approaches the upper hand and, as such, these approaches are not a viable route to detect any speedup, especially because they are unphysical.

B. Designing tunable hard instances

Our approach to generate hard instances capitalizes on the similarity between classical hardness of spin-glass-like problems and quantum hardness. In Fig. 6 of Ref. [40], it was shown in detail how the "mixing" or "autocorrelation" time strongly correlates to the complexity of the spin-glass order

parameter distribution while performing the simulations with state-of-the-art parallel tempering Monte Carlo methods [50–52]. Autocorrelation times uniquely characterize the time a classical algorithm needs to completely decorrelate the system. As such, the time can be used as an indirect proxy of the time complexity of a particular disorder instance.

In spin glasses, order is measured by comparing two copies of the system with the same disorder [25]. For simplicity, we set $S_i^z \equiv S_i$, because we are studying the system classically. In that case, the overlap between two replicas α and β with the same disorder $\mathcal{J}$ but independent Markov chains is defined via

$$q = \frac{1}{N}\sum_{i=1}^{N} S_i^\alpha S_i^\beta, \tag{2}$$

where the sum is over all spins N. One can then study the distribution of the order parameter $P(q)$ which characterizes a given disorder instance $\mathcal{J}$. After a disorder average $[\cdots]_{\rm av}$ over many instances $\mathcal{P}(q) = [P(q)]_{\rm av}$ displays a single peak around $q \sim 0$ for high temperatures. For $T \to 0$ two peaks at $\pm q_{\rm EA}$ emerge [53, 54], a characteristic signature of a broken symmetry. However, for a given instance the structure of the distribution $P(q)$ can be rather complex and can have multiple peaks at different values of q in addition to the two dominant peaks at $\pm q_{\rm EA}$. Individual peaks can be identified with pairs of dominant valleys in the (free-) energy landscape [26]. When these peaks are close to $q \approx 0$, one can assume that a thick barrier separates these valleys, whereas when the peaks are close the barriers are typically thin.

Reference [40] showed that when the distribution $P(q)$ has large support for an area close to $q = 0$, then the autocorrelation times were typically larger than when the support around $q = 0$ is close to zero. As such, by measuring the distribution function $P(q)$, we can predict approximately the time complexity of a particular disorder instance [41]. This is illustrated in the main panel (bottom left) of Fig. 1. There, three characteristic instances are shown (color coded). An instance with many peaks close to $q = 0$ will typically be computationally harder than one that has only two peaks at $q \sim 1$ (red line). Our experiments (shown herein) on the DW2 device show that, indeed, the complexity of an instance can be tuned by studying the structure of $P(q)$ where the distance between two dominant peaks corresponds roughly to the barrier thickness in phase space and the relative depth between the peaks and maxima can be interpreted approximately as the barrier depth. While we are confident that there is a clear correlation between the distance Δq of two well-defined peaks and the thickness of barriers in the energy landscape, the correlation of the depth between the peaks and the height of the barriers remains to be tested experimentally by a more precise mining of the data. However, if the depth between the peaks is nonzero, then it is safe to assume that there is some relatively trivial path that connects the valleys [55].

In addition to selecting instances according to the complexity of the phase space by studying the behavior of the spin-glass order parameter distribution, we estimate the number of configurations for a given instance that minimize the Hamiltonian in Eq. (1). The goal is to make the problem as difficult as possible by restricting the number of minimizing configurations ideally to one, i.e., a unique ground state. To estimate the number of ground-state configurations a given instance has, we use the method pioneered in Refs. [56, 57] where states at very low temperatures are sampled with parallel tempering Monte Carlo techniques. Once the ground-state energy is found, a histogram with minimizing configurations is created (indexed by translating the binary configuration string to a number) and sampled until every bin has at least 50 hits. We make sure that we find the true ground-state energy by studying every instance with different simulational heuristics. However, we cannot be completely certain that we have found *all* configurations that minimize the Hamiltonian, simply because in some cases this number can be huge (in the worst case 2^N). Having exactly one ground state is not a necessary condition to generate a hard problem. However, if our efficient low-temperature search is unable to find more states that minimize the cost function, it will be unlikely that other methods will.

A large source of degeneracy in an Ising Hamiltonian is due to zero local fields. The Hamiltonian in Eq. (1) can be written as a single-spin expression, namely,

$$\mathcal{H} = \sum_{i\in\mathcal{V}}^{N} \mathcal{F}_i S_i, \tag{3}$$

where the local fields $\mathcal{F}_i$ are given by

$$\mathcal{F}_i = -\sum_{j\neq i} J_{ij} S_j - h_i. \tag{4}$$

Whenever for a given disorder $\mathcal{F}_i = 0$, spin S_i can take any value without influencing the energy of the system. Therefore, if a given disorder instance has k spins where $\mathcal{F}_i = 0$, the degeneracy of the ground state will grow by a factor 2^k. To prevent this from happening, we need to choose the disorder from a distribution that—within the restrictions of the device—minimizes the cases where the local fields are zero. The most convenient choice is thus to select the values of $|J_{ij}|$ from a Sidon set [46]. In a Sidon set, the sum of two members of the set gives a number that is not part of the set. For example, the set $\{2, 5, 10\}$ is a Sidon set because the pairwise sum of members of the set never adds up to a member of the set. This is not the case for $\{2, 5, 7\}$, where $2 + 5 = 7$.

To illustrate our ideas, we choose the interactions between the spins from the Sidon set S_{28}

$$J_{ij} \in \{\pm 8/28, \pm 13/28, \pm 19/28, \pm 28/28\}, \tag{5}$$

where we normalize the interactions to be restricted between ± 1 [58]. To select instances with particular properties, we can therefore generate large numbers of random problems using different disorder distributions and then mine the data. We first fix the number of ground-state configurations to 1, and then we divide the instances into subclasses by studying the (normalized) overlap distribution $P(q)$ for each instance. For example, we define the following classes:

(a) *Hard instances with thick barriers*: These are instances where $P(q) > 5$ for $|q| \leq 0.75$. See Fig. 1, main panel.

We are interested in instances that have dominant peaks in the central (blue/dark) window. Based on classical simulations, we expect these instances to be on average among the hardest. In particular, we expect that both simulated, as well as quantum annealing will have trouble finding the optimum – see Fig. 1(a).

(b) *Hard instances with thin barriers*: These are instances where $P(q) \approx 0$ for $|q| \leq 0.50$ and where $P(q) > 2.5$ for $|q| \geq 0.5$ with at least two peaks in the range $|q| \in [0.5, 1.0]$. See Fig. 1, main panel. We are interested in instances that have dominant peaks that are close to each other in the gray boxes close to $|q| > 0.5$. Based on classical simulations, we expect these instances to be on average hard, however, not as hard as the instances with a thick barrier. We expect that while simulated annealing will have similar problems than with the instances with a thick barrier, quantum annealing might show an enhanced performance *if* the device has some quantum advantage over classical codes – see Fig. 1(b).

(c) *(Hard) instances with small barriers*: These are instances where $P(q) < 0.1$ for $|q| \leq 0.75$. The overlap distribution is reminiscent of a ferromagnet at low temperature. In this case no peaks are allowed in the large central (red/light) box of Fig. 1, main panel. In these instances we expect one dominant energy valley (up to smaller wiggles), i.e., these should be the easiest instances on average for any annealing approach. See Fig. 1(c).

Note that the individual windows we use are tuned such that from 10^5 randomly simulated instances approximately 5000 match the aforementioned criteria. After filtering the instances that have more than one minimizing configuration, we obtain approximately 2500 instances to experiment with. The detailed simulation strategy, as well as simulation parameters, are listed in Appendix B.

Noise on the DW2 device is approximately 5% of a particular external field (qubit noise) h and 3.5% of a spin-spin interaction (coupler) J_{ij}. For the instances in S_{28}, the smallest classical energy gap is $\Delta E = 2/28$, i.e., slightly larger than the noise found on the DW2 device. While this will affect the success probabilities, it will affect *all* instances, either easy or hard, approximately the same way. To verify this, we perform detailed simulations where we compute the ground-state energy and configuration of a given instance with no degeneracy, perturb the couplers and qubits with Gaussian random noise of a typical strength found in the current DW2 device, and recompute the ground-state configuration. We apply 10 noise gauges and compute how stable the different instance subclasses defined below are on average. Our results show that all Sidon-set-based instance subclasses with different barrier thicknesses are affected *similarly* by the intrinsic noise of the device (not shown). As such, when comparing instance classes, on average a fair comparison is performed.

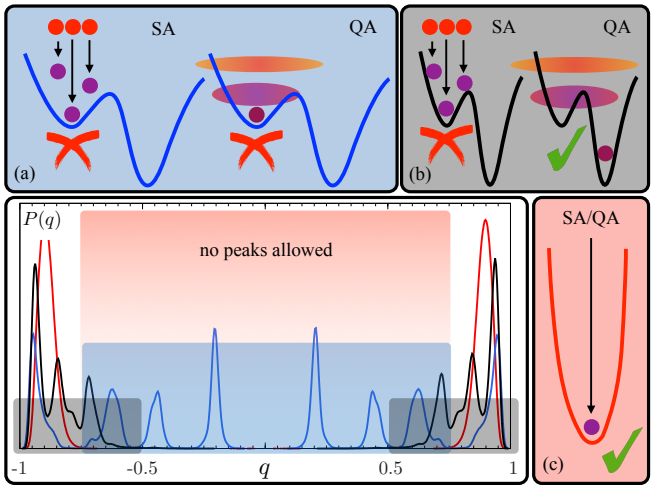

FIG. 1: Main panel: Overlap distribution $P(q)$ for three characteristic disorder instances. For the *hard instances with thick barriers*, we choose instances in the central [blue (dark)] box that have features that extend outside this domain. Based on classical simulations, we expect these instances to be on average among the hardest. In (a), we show the expected outcome of experiments with both simulated annealing (SA) and quantum annealing on the DW2 device (QA). Because the barriers are large and thick, we expect both classical and quantum approaches to have difficulties. In (b), we illustrate the expected behavior when the barriers are thin, i.e., double peaks (or more) that protrude from the dark boxes in the region $|q| > 0.5$. The features in the energy landscape of these *hard instances with thin barriers* are still very pronounced, but we expect the barriers to be thinner than in (a). While SA should show little to no advantage when the barriers remain high but are thinner, if the DW2 device has any quantum advantage, it might be able to overcome these barriers. Finally, we study instances that have no features for $|q| < 0.75$ (large red box in the main panel) and only have a single peak at $\pm q_{\rm EA}$. These *(hard) instances with small barriers* have the simplest energy landscape (c) with mostly only one dominant feature. As such, we expect any annealing approach to efficiently find the optimum of the problem (on average). Note that these are cartoons intended to illustrate the different instance classes and do not represent actual data.

IV. RESULTS

A detailed list of the average success probabilities is given in Appendix C. To make sure that an approximately fair comparison with a known baseline study is performed, we tune the number of sweeps for the SA codes [39] such that the average success probabilities for SA and the DW2 device are approximately the same for bimodal disorder. This is the case for $N_{\rm sw} = 900$ sweeps. Note also that below we quote mainly average success probabilities. The reason is that for the hardest instance classes the DW2 device is often unable to minimize the cost function for the number of runs performed; i.e., a median would be zero and thus deliver no useful information. Because probabilities are restricted to be in the interval $[0, 1]$, an average is well defined.

A. The ugly—D-Wave Two fails often

Figure 2 shows sorted success probabilities p for SA (left) and the DW2 device (right) and different instance classes normalized by the number of samples $N_{\rm sa}$ studied. We compare classes S_{28} with thick, thin, and small barriers with uniform range-4 (U_4) instances and bimodal disorder (U_1) used in previous studies [28]. The data for the DW2 device show a clear progression in complexity and, in particular, that the device is unable to solve many of the harder problems (success probabilities below 10^{-4}). The SA simulations using the codes of Ref. [39] show that bimodal disorder is considerably easier than all other instance classes. Furthermore, for the number of sweeps used, the complexity of U_4 is similar to S_{28} with small ("none") barriers. Interestingly, the SA codes do not distinguish between S_{28} instances with thin and thick barriers. Note that this is not the case for the DW2 device.

Furthermore, SA can solve a much wider range of instances, as can be seen by the distributions dropping to zero only close to $n \to N_{\rm sa}$. This means that while the typical (median) probability to solve a problem is finite for the SA codes, for the hardest instance classes the median is zero for the DW2 device. A double-peaked success behavior of the quantum annealer is consistent with what has been reported in Refs. [28, 32], who present it as evidence of quantum behavior, although the hypothesis has been subsequently challenged by studies of quasiclassical models [59, 60]. Finally, we emphasize that by optimizing the number of sweeps in the SA codes these can be tuned to outperform the DW2 device for all disorder classes studied.

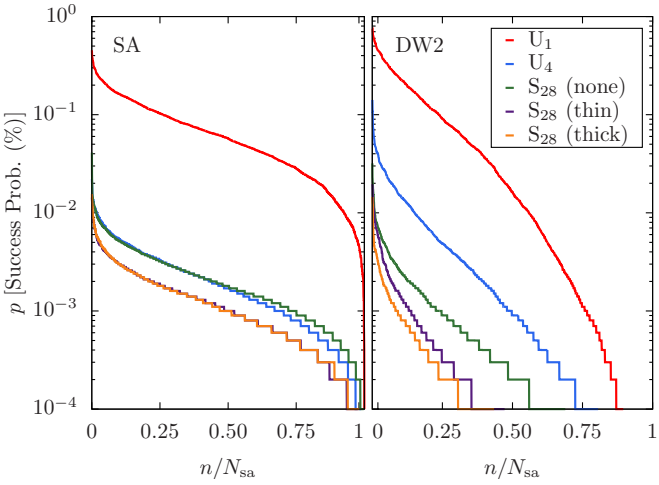

FIG. 2: Sorted success probabilities p (after a gauge average) in percent SA (left) and the DW2 device (right) and different instance classes. The instance index n is normalized by the number of instances $N_{\rm sa}$ per class for better viewing. For both cases, bimodal disorder (U_1) is the easiest problem class to solve. Although the shape of these functions is different, the number of sweeps in SA are chosen such that on average the success probabilities for the U_1 are similar using SA and the DW2. Using SA, uniform range-4 (U_4) instances are comparable to Sidon instances S_{28} with small ("none") barriers. Furthermore, SA does not distinguish between S_{28} instances with thin and thick barriers. There is a clear progression in complexity for the different instance classes on the DW2 device. In particular, while SA can solve almost all instances studied, this is not the case for the DW2. The median success probability for the hardest instance classes (S_{28}) is zero on the DW2 for the number of runs performed; i.e., the machine would need many more runs to be able to find the optimum of hard native problems. Error bars are omitted for better viewing.

B. The bad—Previous instance classes are too easy

Figure 3 shows averaged (and gauge-averaged) success probabilities in logarithmic scale for both DW2 and SA for different instance classes. The data clearly illustrate that the average success probabilities for bimodal disorder are approximately 1 order of magnitude larger than any other type of disorder studied. Note that we choose the number of sweeps for SA such that the average success probability in the bimodal class is comparable to the DW2 device. For the DW2 device, one can clearly see a progression in difficulty between U_1, U_4, as well as the Sidon set S_{28} with small barriers, followed by the Sidon sets with thin and thick barriers. For the choice of sweeps in SA, U_4 is comparable to S_{28} with no dominant barriers, and the S_{28} instances with thick and thin barriers have approximately the same average success probabilities. For all Sidon instance classes studied, the classical SA simulations outperform DW2 based on raw success probabilities. This is seen in more quantitative detail in Fig. 4, which shows the ratio of the average success probability for SA divided by the average success probability for DW2 for each instance class. To establish any quantum speedup, a system-size scaling is needed. However, the fact that the average success probabilities for the bimodal disorder for DW2 and the classical SA codes are much larger than for all other problems suggests that bimodal disorder (or, more generally, highly degenerate random problems) is too easy a problem to detect any quantum speedup. Running any classical SA code in repetition mode with highly degenerate problems potentially represents an advantage over any quantum annealing scheme. Overall, DW2 has far lower average success probabilities on the Sidon sets. This can be explained by the inherent noise present in the device. In the Sidon sets the gap to the first excited state is considerably smaller than for, e.g., bimodal disorder. As such, solving a Hamiltonian that is not the target Hamiltonian due to noise-induced perturbations is likely. Therefore, in an attempt to filter out these effects, we study relative probabilities between instance classes and *not* between optimization techniques. Because the problem instances are randomly generated, one can expect that within a given instance type, e.g., S_{28}, the noise affects all instance classes in a similar fashion [58], as we see in our simulations. This means also that the difference in the performance of DW2 for S_{28} instances with thick and thin barriers is likely not an artifact of the chosen values for the couplers.

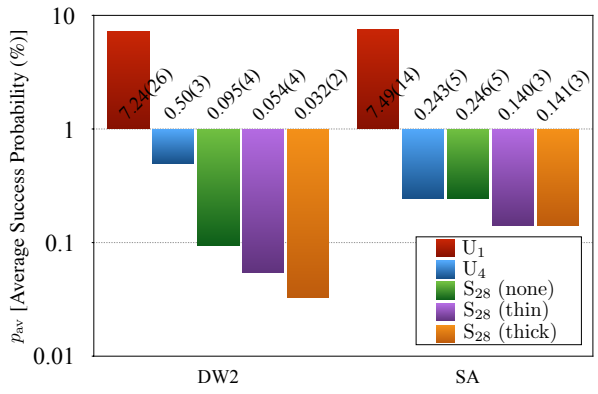

FIG. 3: Average success probabilities $p_{\rm av}$ (after a gauge average) for DW2 and SA using different types of disorder. In all Sidon instance classes (S_{28}) the classical codes outperform DW2. Furthermore, success probabilities for bimodal disorder (U_1) are much larger than for any other instance class, therefore suggesting that the degeneracy produced by bimodal disorder makes this instance class too easy to detect quantum speedup. Note also that the classical codes, on average, do not seem to distinguish between instances with thick and thin barriers. Labels are from left to right.

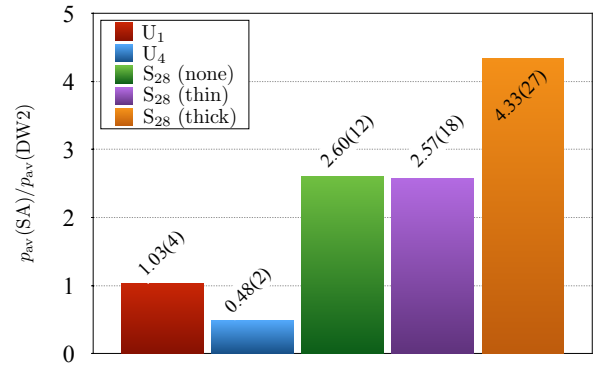

FIG. 4: Ratio between the average success probability for SA and DW2, $p_{\rm av}({\rm SA})/p_{\rm av}({\rm DW2})$, for different disorder classes. In all S_{28} cases, SA outperforms DW2 when the number of SA sweeps is tuned such that $p_{\rm av}({\rm SA})/p_{\rm av}({\rm DW2}) \approx 1$ for the bimodal U_1 class. Labels are from left to right.

C. The good—Evidence of a quantum advantage?

Figure 3 suggests that—at least with the choice of annealing parameters made—in the Sidon instance class the classical codes do not seem to differentiate between thin and thick barriers on average, whereas DW2 does seem to show an improvement in the average success probabilities when the barrier thickness is decreased.

Given the stochastic nature of the classical algorithms, the thickness of a barrier should have a much weaker effect on the algorithmic efficiency than its height. We have selected the instances in such a way that barriers are predominantly tall. Although we have no exact control at the moment as to how tall these barriers are, we can expect them to be on average of similar height for both Sidon sets with thin and thick barriers. However, by selecting instances with peaks in the overlap distribution at a given distance from each other, we have good control over the barrier thickness. Figure 5 shows the ratio of average success probabilities when reducing the barrier thickness (left) and removing dominant barriers (right) for both SA and DW2. While reducing the barrier thickness has no effect on average on the classical algorithms, DW2 experiences a performance increase. To make sure this is not an artifact of our choice of simulation parameters, we run the SA codes with both $N_{\rm sw} = 900$ and 2000 sweeps obtaining qualitatively the same results. Furthermore, we find no correlation between the barrier thickness and the effects noisy couplers and qubits have on the success probabilities for both instance classes. When removing dominant barriers altogether, both classical and quantum algorithms show a noticeable performance increase. One can, therefore, surmise that when the barriers are thin enough (and tall) the DW2 device might experience a quantum advantage over classical approaches. However, a far more careful and systematic study must be performed before strong conclusions can be drawn.

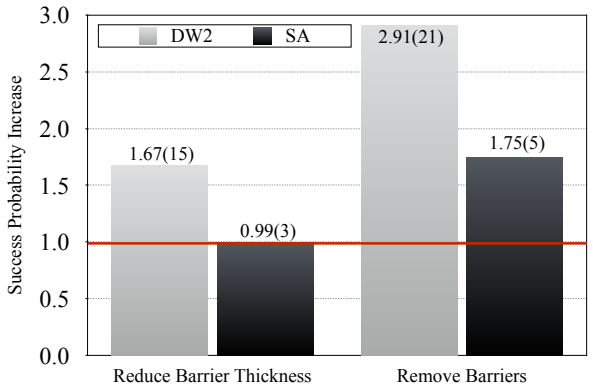

FIG. 5: Average success probability increase when reducing the barrier thickness (ratio between the average success probabilities for S_{28} thick and S_{28} thin) and removing the barriers (ratio between the average success probabilities for S_{28} thick and S_{28} none). While in the latter case both classical algorithms and the quantum annealer show a performance boost on average, in the former only the quantum annealer shows improvement.

To gain a deeper understanding of the noise effects that affect the DW2 device, we relax our criterion for a successful optimization run by allowing the k lowest excited states to count towards a "successful" run in the Sidon sets. In this case, the smallest classical energy gap when flipping a spin is $\Delta E = 2/28 \approx 0.0714$. This should be compared with the disorder-averaged ground state energy of the system, i.e, $[E_0]_{\rm av} \approx= -551$. We compute the success probabilities for energies in the interval $[E_0, E_0 + k\Delta E]$ for different instance

classes using SA and the DW2. Figure 6 shows the average success probabilities as a function of the number of energy levels k. Although we only fix the average success probabilities for the U_1 class to be similar for DW2 (full symbols) and SA (empty symbols) and $k = 0$, it seems this result holds for at least the first 10 excited states. As can be seen, average success probabilities increase with an increased inclusion of low-lying energy levels for all instance classes. The trend is far more pronounced for the DW2 device than for SA in the case of the Sidon sets S_{28}, indicating that noise clearly affects the ability of the machine to detect ground states. Furthermore, note that allowing for the lowest 10 energy levels in the S_{28} class corresponds to an increase in less than 1% in the overall energy of the system. Averaging over gauges (i.e., different instances of noise terms in the Hamiltonian) does help the DW2 device, thus illustrating that an increased performance strongly depends on reducing noise, and also performing multiple quenches.

Is the DW2 device of any use then? For problems affected by noise due to device restrictions, the DW2 thus might efficiently deliver low-lying energy states. This is of particular relevance to problem domains such as machine learning [61] and Bayesian statistical analysis [62].

For optimization, the data suggest that error-correction strategies [31] that enhance robustness to noise should be explored in greater depth. Combined with a hybrid approach that either breaks up the problem into smaller groups that are easier to tackle [63–65], or uses other efficient computing architectures [66] to complement the minimization, the DW2 device (or any other quantum annealing machine) might be an efficient optimization tool one day.

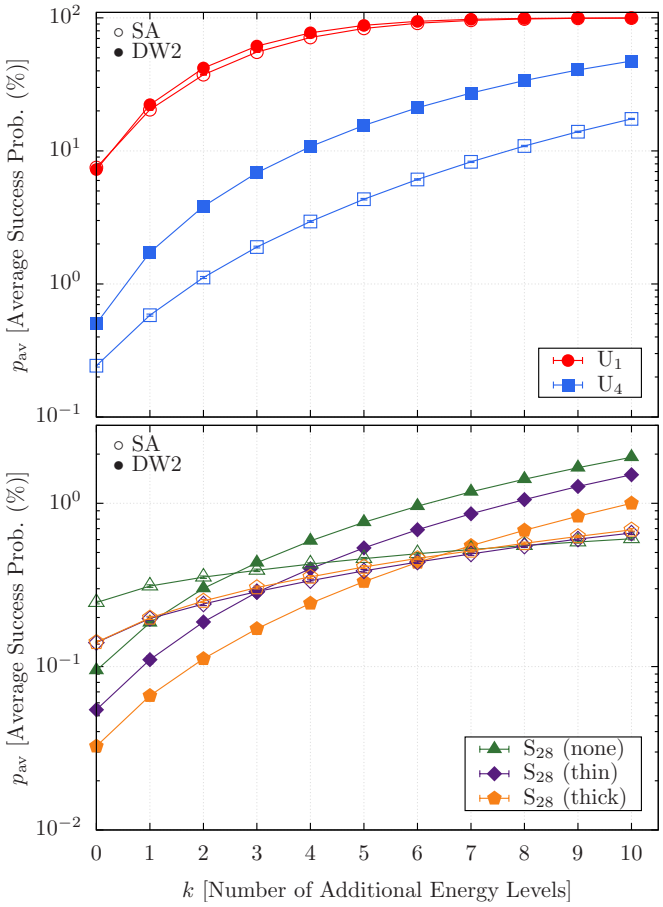

FIG. 6: Average success probabilities p_{av} (after a gauge average) for both DW2 and SA as a function of the number k of low-lying energy levels above the ground state for different instance classes. The top panel shows data for the U_1 and U_4 instance classes, whereas the bottom panel shows data for the S_{28} class. Both panels have the same horizontal axis; we split up the data for better viewing. The trend displayed by DW2 compared to the SA codes for the S_{28} class suggests that noise might be the dominant source of the overall poor performance of the DW2 device.

V. DISCUSSION

We illustrate that a careful design of the benchmark instances is key when attempting to detect quantum speedup. In particular, using insights from the study of spin glasses can help in designing benchmark problems that are considerably harder than previous attempts, and are tunable. Noise levels combined with the small number of qubits on the DW2 device make it difficult to detect any quantum speedup at the moment. Below, we attempt to discuss sources of the poor performance of the device as seen from the spin-glass perspective.

Disordered frustrated binary systems are the *native*, likely hardest, as well as simplest benchmark problems for any new (quantum) computing paradigm. It is important to consider some of the hallmark properties of spin glasses that could make it extremely difficult to detect any (quantum) speedup in the presence of coupler, as well as local-field qubit noise.

A. Effects of coupler noise

The extreme fragility of the spin-glass state was predicted a long time ago [67, 68] and analyzed on the basis of scaling arguments [69, 70]. These scaling arguments predict that the configurations that dominate the partition function change drastically and randomly when temperature, local fields, or the interactions between the spins are modified. There is strong (numerical) evidence of disorder chaos (coupler noise) in spin glasses [71–78]. Therefore, small perturbations of the couplers due to noise might lead to the destruction of the spin-glass state, as well as to a change of the problem to be solved. The latter can be alleviated slightly by performing multiple gauges. However, the weak chaos regime is dominated by rare events that can flip large spin domains that can directly affect experimental results [77]. Increasing the classical energy gap beyond the noise level of the machine can partially reduce these effects, however at the cost of producing considerably easier benchmark instances [38].

One might argue that the minimum classical gap of the Sidon instances ($\Delta E = 2/28$) is too small compared to the machine restrictions when encoding problems. However, we perform tests with a different instance class with a larger clas-

sical energy gap and where the couplers are drawn from the Sidon set $\{\pm 5, \pm 6, \pm 7\}$, finding qualitatively similar results.

B. Effects of local-field noise

In mean-field theory [79], an Ising spin-glass system has a line of transitions in a field [80], known as the de Almeida-Thouless line that separates the paramagnetic phase at high temperatures and fields from the spin-glass phase at lower temperatures and fields [81–86]. Although the existence of a de Almeida-Thouless line for short-range spin glasses is still under some debate (see, for example, Refs. [87–89]), there is vast numerical evidence for a multitude of geometries and, in particular, low-dimensional systems that the spin-glass state is strongly affected by any longitudinal (random) fields [90–93]. As for the case of disorder chaos in spin glasses, the spin-glass state can be easily affected by the intrinsic qubit noise of the DW2 device. Therefore, it might be plausible that, again, the high levels of noise might reduce the success probabilities because the studied system is perturbed and dominant barriers are affected.

VI. SUMMARY AND CONCLUSIONS

We find that for most disorder types studied, DW2 is systematically slower at finding the ground state than the state-of-the-art classical SA codes developed by Isakov *et al.* [39]. Note that, by optimizing the number of sweeps in the SA codes, these can be tuned to outperform the DW2 device for all disorder classes studied. Although this might be discouraging at first, we argue that an improved machine calibration [94], noise reduction [95], and the ability to likewise optimize the quantum annealing schedule combined with larger system sizes and tailored spin-glass problems might help in the quest for quantum speedup. We also show that a "classically computationally hard" problem seems to typically also be a hard problem for the quantum annealing device. However, it could also be that the DW2 device is a thermal annealer [59, 60, 96–99] in disguise.

For the hardest Sidon instances the DW2 device does show a promising trend when the success constraints are relaxed. Furthermore, reducing the thickness between barriers in the free-energy landscapes suggests that for the large Sidon instances studied some quantum advantage might be present. However, this would not be enough to deem the hardware to be efficient, especially because it is unclear if this effect persists for larger problem sizes. We conclude by stressing that a careful design of benchmark instances is key to detecting quantum speedup [28] or any quantum advantage a novel quantum annealing device might have. We thus expect that a combination of the methodologies outlined in this work with the approach outlined in Ref. [28] that defines the notion of "quantum speedup" in detail, combined with better hardware (and maybe quantum error correction [31, 44]), will finally show whether or not quantum annealing has an advantage over classical thermal annealing.

Acknowledgments

We are grateful to M. Amin, R. S. Andrist, J. Job, A. King, D. Lidar, J. Machta, W. Macready, O. Melchert, C. Moore, A. Perdomo-Ortiz, T. F. Rønnow, M. Troyer, D. Venturelli, and I. Zintchenko for many discussions. We especially thank I. Zintchenko for performing double-blind sanity checks for us with his codes and O. Melchert for computing the misfit parameter [100] for our instances on Chimera and for performing detailed studies that link the structure of the spin-overlap distribution directly to the free-energy landscape [101]. H. G. K. acknowledges support from the NSF (Grant No. DMR-1151387) and thanks the City University of New York at Staten Island for an uneventful job interview that culminated with him attending a very interesting Mathematics Department colloquium where he learned about Sidon sets. H. G. K. would also like to thank E. Schmeichel for making this paper happen. We thank the Texas Advanced Computing Center (TACC) at The University of Texas at Austin for providing HPC resources (Stampede Cluster), ETH Zurich for CPU time on the Brutus and Euler clusters, and Texas A&M University for access to their Ada, Eos and Lonestar clusters. We especially thank O. Byrde for beta access to the Euler cluster, as well as S. Vellas and F. Dang for beta access to the Ada cluster. This research is based upon work supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via MIT Lincoln Laboratory Air Force Contract No. FA8721-05-C-0002. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purpose notwithstanding any copyright annotation thereon.

Appendix A: D-Wave Two quantum annealer description

The D-Wave device implements the quantum annealing algorithm via superconducting compound Josephson junction flux qubits [19]. The objective is to find the ground state of the Ising problem Hamiltonian $\mathcal{H}_{\rm P}$ presented in Eq. (1) defined on the D-Wave *Chimera* graph; see Fig. 7. This is attempted by applying and slowly removing a transverse field. The time-dependent Hamiltonian is thus given by

$$\mathcal{H}(s) = A(s)\mathcal{H}_{\rm D} + B(s)\mathcal{H}_{\rm P}, \tag{A1}$$

where the *driver* Hamiltonian $\mathcal{H}_D = \sum_i \sigma_i^x$, $s \in [0,1]$, and $A(s), B(s)$, which control the relative magnitudes of driver and problem Hamiltonians are, respectively, decreasing and increasing in s. Plots of $A(s)$ and $B(s)$ are shown in Fig. 8. The parameter s can be translated to time t via the relation $t = st_f$, where t_f is the annealing time.

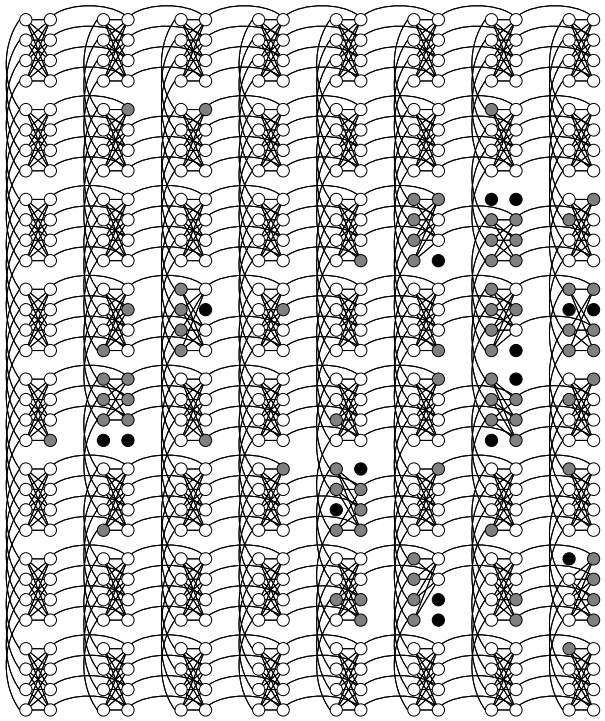

FIG. 7: Adjacency matrix of the D-Wave Two chip used in this study. Circles represent the individual qubits and lines the couplers. White circles represent fully functional qubits, whereas light gray circles represent working qubits with missing couplers. Broken qubits are represented by dark circles (16). This means that the total number of working qubits is 496.

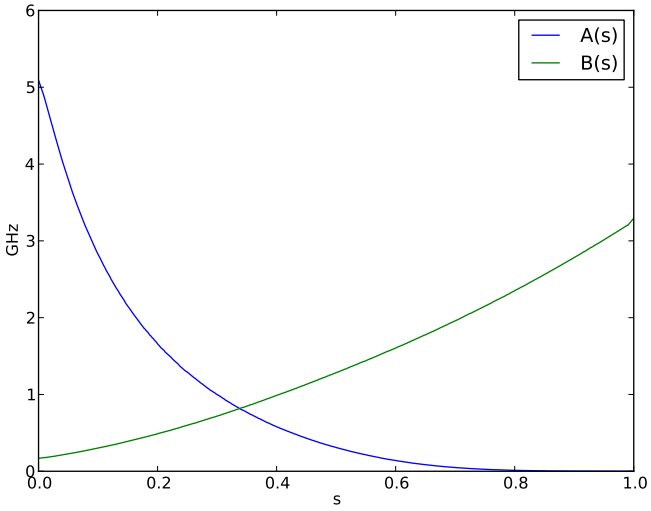

FIG. 8: (Color online) Quantum annealing schedules employed by the D-Wave processor, where $s = t/t_f$.

1. D-Wave Two Methodology

An annealing time of $20\mu s$ is used for all experimental runs on the DW2 processor, which is cooled to a temperature of 18mK. Each problem instance is run $N_{\rm R} = 10^4$ times in $N_{\rm G} = 10$ batches of randomly-chosen gauge transformations in order to provide protection against parameter noise and control errors. To generate a gauge transformation, a set of N random variables $\{t_i\}$, with $t_i \in \{-1, 1\}$, is sampled uniformly, and the transformation

$$h'_i \leftarrow h_i t_i \qquad J'_{ij} \leftarrow J_{ij} t_i t_j \tag{A2}$$

is made. In principle, this procedure does not fundamentally change the problem, but due to parameter noise on the physical device, each gauge transformation of a given instance will, in reality, correspond to a different Hamiltonian.

Following the analysis performed in Ref. [33], an instance's success probability across gauges is derived from the geometric mean of the gauges' failure rates. If p_g is the observed success probability of a gauge g, then

$$\bar{p} = 1 - \prod_{g=1}^{N_{\rm G}} (1 - p_g)^{1/N_{\rm G}}. \tag{A3}$$

A "success" is defined as the occurrence of a state meeting a criterion, for example, of having ground-state energy E_0, or with energy lying in a range $[E_0, E_0 + \Delta]$, $\Delta > 0$, of the minimum.

The DW2 device is run in the so-called "autoscaling" mode for all problems, which adjusts the nominally specified J and h parameters to fully use the range allowed by the device.

2. Simulated annealing methodology

For the software-based simulated annealing experiments, we use the codes developed by Isakov *et al.* [39] to ensure a fair comparison with previous studies. The authors present a variant of SA that exploits the bipartite nature of topologies such as the Chimera graph's in order to halve the number of variables being simulated. This optimization results in considerably improved performance over plain SA. In this study we use the `an_ss_ge_nf_bp_vdeg` routine.

All instances are simulated $N_{\rm R} = 10^4$ times for $N_{\rm sw} = 900$ Monte Carlo sweeps each; clearly, no advantage would be gained from gauge transformations in the software case. The default geometric annealing schedule described in Ref. [39] was adequate for our purposes, but the (inverse) temperature scales were appropriately adjusted for each instance class. The parameters of the simulation are listed in Table I.

Note that we choose $N_{\rm sw} = 900$, such that the average success probabilities for the DW2 device agree with the SA simulations for the commonly studied bimodal (U_1) disorder. We choose this approach to provide a baseline for all other instance classes. Simulations with $N_{\rm sw} = 2000$ sweeps showed qualitatively similar results.

TABLE I: Simulated annealing parameters used for the different instance classes. For each type of disorder class J_{ij}, $N_{\rm sw}$ Monte Carlo sweeps are performed on an annealing schedule from $\beta_{\rm i}$ to $\beta_{\rm f}$.

Class	J_{ij}	$N_{\rm sw}$	$\beta_{\rm i}$	$\beta_{\rm f}$
U_1	$\{\pm 1\}$	900	0.1000	3.000
U_4	$\{\pm 1, \pm 2, \pm 3, \pm 4\}$	900	0.2500	7.500
S_{28}	$\{\pm 8, \pm 13, \pm 19, \pm 28\}$	900	0.0357	1.071

TABLE II: Raw results of the experiments on the DW2 device and the simulated annealing (SA) codes for the different instance classes we study. Listed are average success probabilities ($p_{\rm av}$) in percent, as well as the number of disorder instances $N_{\rm sa}$ studied.

Class	$N_{\rm sa}$	$p_{\rm av}(\%)$ [DW2]	$p_{\rm av}(\%)$ [SA]
S_{28} (thick barriers)	2239	0.032(2)	0.141(3)
S_{28} (thin barriers)	1816	0.054(4)	0.140(3)
S_{28} (small barriers)	2637	0.095(4)	0.246(5)
U_4	2000	0.50(3)	0.243(5)
U_1	2000	7.24(26)	7.49(14)

Appendix B: Parallel tempering Monte Carlo simulation details

To compute the overlap distribution $P(q)$ we perform finite-temperature parallel tempering Monte Carlo simulations [50–52] combined with isoenergetic cluster moves [102] to speed up the simulations. We choose a temperature set with 30 temperatures and the lowest temperature $T_{\rm min} = 0.212$ is chosen such that thermalization can be completed in a meaningful time and features in the overlap distribution are well defined. Two replicas with $N = 496$ spins and the same disorder are thermalized for 2^{23} Monte Carlo sweeps and $P(q)$ is measured over an additional 2^{23} Monte Carlo sweeps to obtain high-resolution data. We compute 10^5 randomly chosen disorder instances for each problem class. The data are then mined according to predefined criteria (see Sec. III B).

Appendix C: Experimental Results

Table II lists the numerical values of the average success probabilities for the different instance classes we study either on the DW2 device or with SA codes. All numbers are averaged via a jackknife procedure over $N_{\rm sa}$ instances of the disorder.

Appendix D: Other Instance Classes Studied

We also perform other experiments with different instance classes. However, these are either too easy or it is extremely difficult to obtain unique ground-state instances. Note that for the J_4 instances [34], where the interactions are bimodally distributed and the bonds in the $K_{4,4}$ cells are a $1/4$, as well as the $S_{1,3,7}$ small Sidon instances, we limit the number of configurations that minimize the Hamiltonian to less than 32 because too few unique ground states could be found. As such, we are merely mentioning here the results to prevent other researchers from attempting to study these systems. Average success probabilities are listed in Table III.

TABLE III: Raw results of the experiments on the DW2 device and the SA codes for additional instance classes we study. Listed are average success probabilities ($p_{\rm av}$) in percent, as well as the number of disorder instances $N_{\rm sa}$ studied.

Class	$N_{\rm sa}$	$p_{\rm av}(\%)$ [DW2]	$p_{\rm av}(\%)$ [SA]
J_4 (thick barriers)	1250	0.50(3)	4.1(1)
J_4 (small barriers)	2035	1.96(6)	13.3(2)
$S_{1,3,7}$ (thick barriers)	1615	0.063(4)	0.59(1)
$S_{1,3,7}$ (small barriers)	1582	0.22(1)	1.14(2)

[1] A. K. Hartmann and H. Rieger, *Optimization Algorithms in Physics* (Wiley-VCH, Berlin, 2001).

[2] *Lecture notes in computer science 2241*, in *Computational Combinatorial Optimization*, edited by M. Jünger and D. Naddef (Springer Verlag, Heidelberg, 2001), vol. 2241.

[3] A. K. Hartmann and H. Rieger, *New Optimization Algorithms in Physics* (Wiley-VCH, Berlin, 2004).

[4] G. Moore, *Cramming more components onto integrated circuits*, Electronics Magazine **38**, 114 (1965).

[5] M. A. Nielsen and I. L. Chuang, *Quantum Computation and Quantum Information* (Cambridge University Press, Cambridge, 2000).

[6] H. Nishimori, *Statistical Physics of Spin Glasses and Information Processing: An Introduction* (Oxford University Press, New York, 2001).

[7] A. B. Finnila, M. A. Gomez, C. Sebenik, C. Stenson, and J. D. Doll, *Quantum annealing: A new method for minimizing multidimensional functions*, Chem. Phys. Lett. **219**, 343 (1994).

[8] T. Kadowaki and H. Nishimori, *Quantum annealing in the transverse Ising model*, Phys. Rev. E **58**, 5355 (1998).

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

[10] E. Farhi, J. Goldstone, S. Gutmann, and M. Sipser, *Quantum Computation by Adiabatic Evolution* (2000), arXiv:quant-ph/0001106.

[11] J. Roland and N. J. Cerf, *Quantum search by local adiabatic evolution*, Phys. Rev. A **65**, 042308 (2002).

[12] G. Santoro, E. Martoňák, R. Tosatti, and R. Car, *Theory of quantum annealing of an Ising spin glass*, Science **295**, 2427 (2002).

[13] A. Das and B. K. Chakrabarti, *Quantum Annealing and Related Optimization Methods* (Edited by A. Das and B.K. Chakrabarti, Lecture Notes in Physics 679, Berlin: Springer, 2005).

[14] G. E. Santoro and E. Tosatti, *TOPICAL REVIEW: Optimization using quantum mechanics: quantum annealing through adiabatic evolution*, J. Phys. A **39**, R393 (2006).

[15] D. A. Lidar, *Towards Fault Tolerant Adiabatic Quantum Computation*, Phys. Rev. Lett. **100**, 160506 (2008).

[16] A. Das and B. K. Chakrabarti, *Quantum Annealing and Ana-*

log Quantum Computation, Rev. Mod. Phys. **80**, 1061 (2008).
[17] S. Morita and H. Nishimori, *Mathematical Foundation of Quantum Annealing*, J. Math. Phys. **49**, 125210 (2008).
[18] S. Mukherjee and B. K. Chakrabarti, *Multivariable optimization: Quantum annealing and computation*, Eur. Phys. J. Special Topics **224**, 17 (2015).
[19] M. W. Johnson, M. H. S. Amin, S. Gildert, T. Lanting, F. Hamze, N. Dickson, R. Harris, A. J. Berkley, J. Johansson, P. Bunyk, et al., *Quantum annealing with manufactured spins*, Nature **473**, 194 (2011).
[20] M. H. S. Amin and V. Choi, *First-order quantum phase transition in adiabatic quantum computation*, Phys. Rev. A **80**, 062326 (2009).
[21] A. P. Young, S. Knysh, and V. N. Smelyanskiy, *First-Order Phase Transition in the Quantum Adiabatic Algorithm*, Phys. Rev. Lett. **104**, 020502 (2010).
[22] I. Hen and A. P. Young, *Exponential complexity of the quantum adiabatic algorithm for certain satisfiability problems*, Phys. Rev. E **84**, 061152 (2011).
[23] Y. Matsuda, H. Nishimori, and H. G. Katzgraber, *Ground-state statistics from annealing algorithms: quantum versus classical approaches*, New J. Phys. **11**, 073021 (2009).
[24] S. Kirkpatrick, C. D. Gelatt, Jr., and M. P. Vecchi, *Optimization by simulated annealing*, Science **220**, 671 (1983).
[25] K. Binder and A. P. Young, *Spin glasses: Experimental facts, theoretical concepts and open questions*, Rev. Mod. Phys. **58**, 801 (1986).
[26] D. L. Stein and C. M. Newman, *Spin Glasses and Complexity*, Primers in Complex Systems (Princeton University Press, 2013).
[27] A. Lucas, *Ising formulations of many NP problems*, Front. Physics **12**, 5 (2014).
[28] 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** (2014).
[29] For the sake of brevity we shall refer to "*limited quantum speedup*" simply as "*quantum speedup*.".
[30] R. D. Somma, D. Nagaj, and M. Kieferová, *Quantum Speedup by Quantum Annealing*, Phys. Rev. Lett. **109**, 050501 (2012).
[31] K. L. Pudenz, T. Albash, and D. A. Lidar, *Error-corrected quantum annealing with hundreds of qubits*, Nat. Commun. **5**, 3243 (2014).
[32] S. Boixo, T. Albash, F. M. Spedalieri, N. Chancellor, and D. A. Lidar, *Experimental signature of programmable quantum annealing*, Nat. Comm. **4**, 2067 (2013).
[33] 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).
[34] 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).
[35] P. Bunyk, *Architectural Considerations in the Design of a Superconducting Quantum Annealing Processor*, IEEE Trans. Appl. Supercond. **24**, 1 (2014).
[36] I. Hen, J. Job, T. Albash, T. F. Rønnow, M. Troyer, and D. Lidar, *Probing for quantum speedup in spin glass problems with planted solutions* (2015), (arXiv:quant-phys/1502.01663).
[37] D. Venturelli, S. Mandrà, S. Knysh, B. O'Gorman, R. Biswas, and V. Smelyanskiy, *Quantum Optimization of Fully-Connected Spin Glasses* (2014), (arXiv:cond-mat/1406.7553).
[38] Z. Zhu, A. J. Ochoa, F. Hamze, S. Schnabel, and H. G. Katzgraber, *Best-case performance of quantum annealers on native spin-glass benchmarks: How chaos can affect success probabilities* (2015), (arXiv:1505.02278).
[39] S. V. Isakov, I. N. Zintchenko, T. F. Rønnow, and M. Troyer, *Optimized simulated annealing for Ising spin glasses*, Comput. Phys. Commun. **192**, 265 (2015), (see also ancillary material to arxiv:cond-mat/1401.1084).
[40] B. Yucesoy, J. Machta, and H. G. Katzgraber, *Correlations between the dynamics of parallel tempering and the free-energy landscape in spin glasses*, Phys. Rev. E **87**, 012104 (2013).
[41] We have performed [101] detailed simulations to verify the conjecture that the spin-overlap distribution faithfully mirrors the free-energy landscape of a disordered system. Using both simple Monte Carlo [103] and improved extremal optimization [104, 105] — two methods that optimize in an inherently different way — we study the algorithm's efficiency to traverse the energy landscape for the different Sidon instance classes studied. For both algorithms, when the spin-overlap distribution has a lot of structure (hard instances with thick and thin barriers), the Hamming distances between subsequent energy maxima and minima grow when approaching the ground-state energy. This means that the energy landscape is highly nontrivial in this case; i.e., deep valleys and tall mountains seem to exist. This is not the case for instances for which the spin overlap has a trivial structure reminiscent of a ferromagnet (easy instances). Here the Hamming distance as a function of the energy valley index decreases when the energies are close to the ground state, suggesting a featureless energy landscape.
[42] Figure 7 in Ref. [43] shows the change in energy when switching boundary conditions in two-dimensional Ising spin glasses on a square lattice. When the disorder is chosen from a discrete distribution (in this case bimodal), corrections to scaling are large and systems with more than approximately 4000 spins are needed to truly probe the thermodynamic limit. This is also highlighted in the erratum to Ref. [34], Ref. [106], where weak, but persistent corrections to scaling affected the determination of the critical exponents in the ferromagnetic sector.
[43] I. A. Campbell, A. K. Hartmann, and H. G. Katzgraber, *Energy size effects of two-dimensional Ising spin glasses*, Phys. Rev. B **70**, 054429 (2004).
[44] K. L. Pudenz, T. Albash, and D. A. Lidar, *Quantum Annealing Correction for Random Ising Problems*, Phys. Rev. A **91**, 042302 (2015).
[45] See http://www.dwavesys.com.
[46] S. Sidon, *Ein Satz über trigonometrische Polynome und seine Anwendung in der Theorie der Fourier-Reihen*, Mathematische Annalen **106**, 536 (1932).
[47] T. Neuhaus, *Monte Carlo Search for Very Hard KSAT Realizations for Use in Quantum Annealing* (2014), (arXiv:cond-mat/1412.5361).
[48] Reference [107] shows that if the instances are generated such that the coupling range is restricted, the results on DW2 are more favorable, as would be expected due to increased noise robustness.
[49] S. Boixo, V. N. Smelyanskiy, A. Shabani, S. V. Isakov, M. Dykman, V. S. Denchev, M. Amin, A. Smirnov, M. Mohseni, and H. Neven, *Computational Role of Collective Tunneling in a Quantum Annealer* (2014), arXiv:1411.4036.
[50] K. Hukushima and K. Nemoto, *Exchange Monte Carlo method and application to spin glass simulations*, J. Phys. Soc. Jpn. **65**, 1604 (1996).
[51] C. Geyer, in *23rd Symposium on the Interface*, edited by E. M. Keramidas (Interface Foundation, Fairfax Station, VA, 1991), p. 156.
[52] H. G. Katzgraber, S. Trebst, D. A. Huse, and M. Troyer,

Feedback-optimized parallel tempering Monte Carlo, J. Stat. Mech. P03018 (2006).

[53] S. F. Edwards and P. W. Anderson, *Theory of spin glasses*, J. Phys. F: Met. Phys. **5**, 965 (1975).

[54] G. Parisi, *Order parameter for spin-glasses*, Phys. Rev. Lett. **50**, 1946 (1983).

[55] Clearly, phase space is a high-dimensional space. As such, it is very likely that there is always some path that will connect two points, thus avoiding a barrier. However, in this case, we refer to relatively straightforward paths, whatever that might mean in the frustrating world of spin glasses.

[56] J. J. Moreno, H. G. Katzgraber, and A. K. Hartmann, *Finding low-temperature states with parallel tempering, simulated annealing and simple Monte Carlo*, Int. J. Mod. Phys. C **14**, 285 (2003).

[57] H. G. Katzgraber and A. P. Young, *Geometry of large-scale low-energy excitations in the one-dimensional Ising spin glass with power-law interactions*, Phys. Rev. B **68**, 224408 (2003).

[58] In the meantime, we have discovered that for the vanilla Chimera graph the optimal Sidon set is $\mathrm{U}_{5,6,7}$ with random interactions drawn from the set $\{\pm 5, \pm 6, \pm 7\}$. This set produces many instances with a low degeneracy and is also rather robust to the intrinsic noise of the DW2 device because the classical energy gap $\Delta E = 2/7$ is considerably larger than the noise of the device. We report on these results in a subsequent publication [38].

[59] J. A. Smolin and G. Smith, *Classical signature of quantum annealing*, arXiv preprint arXiv:1305.4904 (2013).

[60] S. W. Shin, G. Smith, J. A. Smolin, and U. Vazirani, *How "Quantum" is the D-Wave Machine?* (2014), (arXiv:1401.7087).

[61] C. Bishop, *Pattern Recognition and Machine Learning* (Springer-Verlag, New York, 2006).

[62] A. Gelman, J. B. Carlin, H. L. Stern, and D. B. Rubin, *Bayesian Data Analysis* (Chapman and Hall/CRC, London, 2003).

[63] J. Houdayer and O. C. Martin, *Renormalization for discrete optimization*, Phys. Rev. Lett. **83**, 1030 (1999).

[64] C. K. Thomas, O. L. White, and A. A. Middleton, *Persistence and memory in patchwork dynamics for glassy models*, Phys. Rev. B **77**, 092415 (2008).

[65] I. Zintchenko, M. B. Hastings, and M. Troyer, *From local to global ground states in Ising spin glasses*, Phys. Rev. B **91**, 024201 (2015).

[66] F. Belletti, M. Cotallo, A. Cruz, L. A. Fernández, A. Gordillo, A. Maiorano, F. Mantovani, E. Marinari, V. Martín-Mayor, A. Muñoz-Sudupe, et al., *Simulating spin systems on IANUS, an FPGA-based computer*, Comp. Phys. Comm. **178**, 208 (2008).

[67] S. R. McKay, A. N. Berker, and S. Kirkpatrick, *Spin-Glass Behavior in Frustrated Ising Models with Chaotic Renormalization-Group Trajectories*, Phys. Rev. Lett. **48**, 767 (1982).

[68] G. Parisi, *Spin glasses and replicas*, Physica A **124**, 523 (1984).

[69] D. S. Fisher and D. A. Huse, *Ordered phase of short-range Ising spin-glasses*, Phys. Rev. Lett. **56**, 1601 (1986).

[70] A. J. Bray and M. A. Moore, *Chaotic Nature of the Spin-Glass Phase*, Phys. Rev. Lett. **58**, 57 (1987).

[71] I. Kondor, *On chaos in spin glasses*, J. Phys. A **22**, L163 (1989).

[72] M. Ney-Nifle and A. P. Young, *Chaos in a two-dimensional Ising spin glass*, J. Phys. A **30**, 5311 (1997).

[73] M. Ney-Nifle, *Chaos and universality in a four-dimensional spin glass*, Phys. Rev. B **57**, 492 (1998).

[74] A. Billoire and E. Marinari, *Evidence against temperature chaos in mean-field and realistic spin glasses*, J. Phys. A **33**, L265 (2000).

[75] A. Billoire and E. Marinari, *Overlap among states at different temperatures in the SK model*, Europhys. Lett. **60**, 775 (2002).

[76] M. Sasaki, K. Hukushima, H. Yoshino, and H. Takayama, *Temperature Chaos and Bond Chaos in Edwards-Anderson Ising Spin Glasses: Domain-Wall Free-Energy Measurements*, Phys. Rev. Lett. **95**, 267203 (2005).

[77] H. G. Katzgraber and F. Krzakala, *Temperature and Disorder Chaos in Three-Dimensional Ising Spin Glasses*, Phys. Rev. Lett. **98**, 017201 (2007).

[78] In fact, a recent study similar to our preliminary results (see, e.g., https://youtu.be/C8fSpHW9XHk) attempts to design harder benchmark instances by exploiting the chaotic effects in spin-glass systems [108].

[79] G. Parisi, *The order parameter for spin glasses: a function on the interval* 0–1, J. Phys. A **13**, 1101 (1980).

[80] J. R. L. de Almeida and D. J. Thouless, *Stability of the Sherrington-Kirkpatrick solution of a spin glass model*, J. Phys. A **11**, 983 (1978).

[81] R. N. Bhatt and A. P. Young, *Search for a transition in the three-dimensional $\pm J$ Ising spin-glass*, Phys. Rev. Lett. **54**, 924 (1985).

[82] A. Billoire and B. Coluzzi, *Numerical study of the Sherrington-Kirkpatrick model in a magnetic field*, Phys. Rev. E **68**, 026131 (2003).

[83] A. Barrat and L. Berthier, *Real-space application of the mean-field description of spin-glass dynamics*, Phys. Rev. Lett. **87**, 087204 (2001).

[84] H. Takayama and K. Hukushima, *Field-shift aging protocol on the 3D Ising spin-glass model: dynamical crossover between the spin-glass and paramagnetic states*, J. Phys. Soc. Jpn. **73**, 2077 (2004).

[85] J. Houdayer and O. C. Martin, *Ising spin glasses in a magnetic field*, Phys. Rev. Lett. **82**, 4934 (1999).

[86] F. Krzakala, J. Houdayer, E. Marinari, O. C. Martin, and G. Parisi, *Zero-temperature responses of a 3D spin glass in a field*, Phys. Rev. Lett. **87**, 197204 (2001).

[87] L. Leuzzi, G. Parisi, F. Ricci-Tersenghi, and J. J. Ruiz-Lorenzo, *Ising Spin-Glass Transition in a Magnetic Field Outside the Limit of Validity of Mean-Field Theory*, Phys. Rev. Lett. **103**, 267201 (2009).

[88] R. A. Baños, A. Cruz, L. A. Fernandez, J. M. Gil-Narvion, A. Gordillo-Guerrero, M. Guidetti, D. Iñiguez, A. Maiorano, E. Marinari, V. Martin-Mayor, et al., *Thermodynamic glass transition in a spin glass without time-reversal symmetry*, Proc. Natl. Acad. Sci. U.S.A. **109**, 6452 (2012).

[89] M. Baity-Jesi, R. Alvarez Baños, A. Cruz, L. A. Fernandez, J. M. Gil-Narvion, Gordillo-Guerrero, D. Iñiguez, A. Maiorano, F. Mantovani, E. Marinari, et al., *Dynamical transition in the $D = 3$ Edwards-Anderson spin glass in an external magnetic field*, Phys. Rev. E **89**, 032140 (2014).

[90] A. P. Young and H. G. Katzgraber, *Absence of an Almeida-Thouless line in Three-Dimensional Spin Glasses*, Phys. Rev. Lett. **93**, 207203 (2004).

[91] H. G. Katzgraber and A. P. Young, *Probing the Almeida-Thouless line away from the mean-field model*, Phys. Rev. B **72**, 184416 (2005).

[92] H. G. Katzgraber, D. Larson, and A. P. Young, *Study of the de Almeida-Thouless line using power-law diluted one-dimensional Ising spin glasses*, Phys. Rev. Lett. **102**, 177205 (2009).

[93] D. Larson, H. G. Katzgraber, M. A. Moore, and A. P. Young, *Spin glasses in a field: Three and four dimensions as seen from one space dimension*, Phys. Rev. B **87**, 024414 (2013).

[94] A. Perdomo-Ortiz, B. O'Gorman, J. Fluegemann, R. Biswas, and V. N. Smelyanskiy, *Determination and correction of persistent biases in quantum annealers* (2015), (arXiv:quant-phys/1503.05679).

[95] A. Perdomo-Ortiz, J. Fluegemann, R. Biswas, and V. N. Smelyanskiy, *A Performance Estimator for Quantum Annealers: Gauge selection and Parameter Setting* (2015), (arXiv:quant-phys/1503.01083).

[96] G. Smith and J. Smolin, *Putting "Quantumness" to the Test*, Physics **6**, 105 (2013).

[97] T. Lanting, A. J. Przybysz, A. Y. Smirnov, F. M. Spedalieri, M. H. Amin, A. J. Berkley, R. Harris, F. Altomare, S. Boixo, P. Bunyk, et al., *Entanglement in a quantum annealing processor*, Phys. Rev. X **4**, 021041 (2014).

[98] T. Albash, W. Vinci, A. Mishra, P. A. Warburton, and D. A. Lidar, *Consistency Tests of Classical and Quantum Models for a Quantum Device*, Phys. Rev. A **91**, 042314 (2015).

[99] T. Albash, T. F. Rønnow, M. Troyer, and D. A. Lidar, *Reexamining classical and quantum models for the D-Wave One processor*, Eur. Phys. J. Spec. Top. **224**, 111 (2015).

[100] S. Kobe and T. Klotz, *Frustration: How it can be measured*, Phys. Rev. E **52**, 5660 (1995).

[101] O. Melchert *et al.*, in preparation (2015).

[102] Z. Zhu, A. J. Ochoa, and H. G. Katzgraber, *Efficient Cluster Algorithm for Spin Glasses in Any Space Dimension*, Phys. Rev. Lett. **115**, 077201 (2015).

[103] H. G. Katzgraber, *Introduction to Monte Carlo Methods* (2009), (arXiv:0905.1629).

[104] A. A. Middleton, *Improved extremal optimization for the ising spin glass*, Phys. Rev. E **69**, 055701(R) (2004).

[105] S. Boettcher and A. G. Percus, *Optimization with Extremal Dynamics*, Phys. Rev. Lett. **86**, 5211 (2001).

[106] M. Weigel, H. G. Katzgraber, J. Machta, F. Hamze, R. S. Andrist, and 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).

[107] A. D. King, *Performance of a quantum annealer on range-limited constraint satisfaction problems* (2015), arXiv:1502.02098.

[108] V. Martin-Mayor and I. Hen, *Unraveling Quantum Annealers using Classical Hardness* (2015), (arXiv:1502.02494).