Degeneracy Engineering for Classical and Quantum Annealing: A Case Study of Sparse Linear Regression in Collider Physics

Eric R. Anschuetz,[1,*] Lena Funcke,[1,2,3,†] Patrick T. Komiske,[1,3,‡] Serhii Kryhin,[1,3,§] and Jesse Thaler[1,2,3,¶]

[1] *Center for Theoretical Physics, Massachusetts Institute of Technology, Cambridge, MA 02139, USA*
[2] *Co-Design Center for Quantum Advantage*
[3] *The NSF AI Institute for Artificial Intelligence and Fundamental Interactions*

Classical and quantum annealing are computing paradigms that have been proposed to solve a wide range of optimization problems. In this paper, we aim to enhance the performance of annealing algorithms by introducing the technique of degeneracy engineering, through which the relative degeneracy of the ground state is increased by modifying a subset of terms in the objective Hamiltonian. We illustrate this novel approach by applying it to the example of ℓ_0-norm regularization for sparse linear regression, which is, in general, an NP-hard optimization problem. Specifically, we show how to cast ℓ_0-norm regularization as a quadratic unconstrained binary optimization (QUBO) problem, suitable for implementation on annealing platforms. As a case study, we apply this QUBO formulation to energy flow polynomials in high-energy collider physics, finding that degeneracy engineering substantially improves the annealing performance. Our results motivate the application of degeneracy engineering to a variety of regularized optimization problems.

arXiv:2205.10375v2 [quant-ph] 12 Sep 2022

CONTENTS

* eans@mit.edu
† lfuncke@mit.edu
‡ pkomiske@mit.edu
§ serhin@mit.edu
¶ jthaler@mit.edu

I. INTRODUCTION

Quantum annealing [1–3] is a computing paradigm for solving optimization problems, with applications ranging across computer science problems [4], machine learning [5], quantum chemistry [6], protein folding [7], and beyond. Such optimization problems often require minimizing a cost function, which can be reformulated as finding the ground state of a classical Ising Hamiltonian [8]. Many problems of practical importance, however, have cost functions over exponentially many spin configurations, reminiscent of classical spin glasses [9–11]. These characteristics make it extremely difficult for classical algorithms, including classical annealing, to find the ground state of the classical Ising Hamiltonian [3].

Quantum annealing was conceived as an alternative to solve this task, where one elevates the classical Ising Hamiltonian to a quantum spin Hamiltonian to take advantage of tunneling in the optimization landscape [3]. Since the first quantum annealing device became commercially available in 2011 [12], a large number of proof-of-principle demonstrations have been performed (see, e.g., Refs. [13–18]). Quantum annealing still faces several conceptual and hardware challenges, however—in particular the inability to outperform classical annealing algorithms in many applications (see Ref. [19] for a review).

In this paper, we introduce the technique of *degeneracy engineering* in order to enhance the performance of classical and quantum annealing. We show that for some applications, one can bias the spectral landscape toward more optimal solutions, dramatically improving both classical and quantum annealing performance on these problems. We illustrate this novel concept by applying it to ℓ_0-norm regularization for sparse linear regression, which is a non-convex optimization problem that is, in general, NP-hard [20]. Specifically, we first show how to cast ℓ_0-norm regularization as a quadratic unconstrained binary optimization (QUBO) problem, suitable for implementation on (quantum) annealing platforms. The key insight is to use a *redundant* (qu)bit encoding scheme for the linear fit coefficients, which allows the ℓ_0-norm penalty term to be written in quadratic form. The smallest redundant encoding scheme requires only one extra (qu)bit per coefficient. By using a higher degree of redundancy, though, one is able to increase the *relative degeneracy* of the desired ground-state configuration to the first excited state of the regularizer, which, in practice, yields better annealing performance on the full problem.

Sparse linear regression is a topic of general interest, but here we focus on a case study in high-energy collider physics. *Energy flow polynomials* (EFPs) are a linear basis of collider observables [21], which can be used to accomplish a broad range of classification and regression tasks in collider physics. Most EFP studies to date have used standard linear regression with a subset of $O(1000)$ EFPs [22, 23], but it is likely that many collider tasks could be accomplished to the desired accuracy using only a handful of EFPs. This is a natural venue to explore sparse linear regression, but there are known cases where the two most popular sparse linear regression approaches—ridge regression using ℓ_2-norm regularization [24] and lasso regression using ℓ_1-norm regularization [25]—yield unsatisfactory results [21, 26]. While the ℓ_0-norm penalty is expected to yield better performance in such cases, it is computationally daunting to implement. These considerations make this problem an ideal test bed for exploring the performance of degeneracy engineering.

In detailed numerical simulations, we assess the potential gains from quantum annealing by comparing standard simulated annealing [27] to path integral Monte Carlo (PIMC) [28]. While PIMC is a classical annealing strategy, it is a useful proxy for quantum annealing [29], and it is exact in the long equilibration time limit. We compare five different regularization methods, including the standard ℓ_2-, ℓ_1-, and ℓ_0-norm regularizations, as well as two novel heuristics. Focusing on ℓ_0-norm regularization, we then compare two different encoding schemes with different degrees of redundancy, thus examining the potential benefits of degeneracy engineering. Our case study is based on EFP sparse regression tasks with known analytic solutions, so that we have an absolute performance benchmark. Using our QUBO implementation with the smallest redundant encoding scheme, we find relatively poor regression performance. Going to a higher degree of redundancy, though, we achieve significantly better performance. This motivates further studies of degeneracy engineering for other optimization problems beyond sparse linear regression.

The remainder of this paper is organized as follows. In Sec. II, we review ℓ_p-norm-regularized linear regression and its binary encoding on quantum or classical computers, followed by a derivation of ℓ_0-norm regularization in QUBO form. In Sec. III, we introduce the concept of degeneracy engineering, which improves the annealing performance by increasing the relative degeneracy of the ground-state configuration. In Sec. IV, we outline different optimization strategies, including a review of classical annealing and PIMC, a proposal of novel heuristics, and considerations for quantum annealing. In Sec. V, we review EFPs, including a detailed overview of the observables and data sets used in our case study. In Sec. VI, we present the numerical results of our case study, comparing the smallest redundant encoding scheme to the scheme with two redundant qubits, comparing the ℓ_0-norm regularization to its ℓ_1- and ℓ_2-norm counterparts, and comparing simulated annealing to PIMC. We conclude in Sec. VII, including a broader discussion of the role of redundant encoding schemes for classical and quantum annealing.

II. SPARSE LINEAR REGRESSION AS A QUBO PROBLEM

A. Review of ℓ_p-Norm Regularization

For generic regression problems, the goal is to find a function $h : \mathbb{R}^n \to \mathbb{R}$ that approximates the mapping of inputs $\vec{x}$ to outputs y seen in a training data set $\mathcal{S}$. One way to achieve this is by minimizing the mean squared error (MSE) loss function:

$$L_{\text{MSE}} = \sum_{s\in\mathcal{S}} \Big(y_s - h(\vec{x}_s)\Big)^2. \tag{1}$$

For linear regression, one chooses a set of K functions $h_a(\vec{x})$ and real fit coefficients c_a, such that

$$h(\vec{x};\{c_a\}) = \sum_{a=1}^{K} c_a\, h_a(\vec{x}). \tag{2}$$

To avoid overfitting, one is often interested in finding a sparse, approximate minimizer of the MSE. To achieve this in practice, one introduces a regulator R that penalizes non-zero values of c_a:

$$L = L_{\text{MSE}} + \lambda\, R, \tag{3}$$

where λ controls the strength of the regularization. For ℓ_p-norm regularization, the regularization term is

$$R^{(p)} = \sum_{a=1}^{K} R_a^{(p)}, \qquad R_a^{(p)} = |c_a|^p, \tag{4}$$

where $|\cdot|$ is the absolute value. When $p = 0$, we define the limit as

$$\lim_{p\to 0} R_a^{(p)} = \begin{cases} 0 & c_a = 0 \\ 1 & c_a \neq 0 \end{cases}, \tag{5}$$

such that the ℓ_0-norm penalty depends only on whether c_a is non-zero, independent of its magnitude. Since ℓ_0-norm regularized regression is computationally challenging to implement, this problem is well suited for exploring the performance of degeneracy engineering.

B. Redundant Binary Encodings

To formulate a quadratic representation of the ℓ_0-norm regularizer, we first consider binary encodings of the real fit coefficients c_a. For an M-bit representation, we have

$$c_a = \sum_{i=1}^{M} g_i\, b_a^{(i)}, \tag{6}$$

where g_i are fixed real numbers, and the binary coefficients $b_a^{(i)}$ take values of 0 or 1.

For a non-redundant encoding, one typically chooses a standard binary encoding, such as $g_i = 2^i$. More generally, though, g_i can take any desired fixed value, including a negative value, at the expense of having multiple binary representations for the same real number [30]. As a concrete example, consider a four-bit encoding, where

$$\vec{g} = \{-2, -1, 1, 2\}. \tag{7}$$

For a fixed a, there are $2^4 = 16$ possible choices for the values of $b_a^{(i)}$, but only 7 unique values of c_a, namely

$$c_a \in \{-3, -2, -1, 0, 1, 2, 3\}. \tag{8}$$

In the context of annealing, these redundant encodings are irrelevant for the ground state if the corresponding values of the loss function are *the same or higher* than for the default encoding. We will exploit this freedom in implementing ℓ_0-norm regularization.

C. Quadratic Loss for ℓ_0-Norm Penalty

When inserting the binary representation for the fit coefficients c_a in Eq. (6) into the MSE loss function in Eq. (1), we see that the dependence on the binary coefficients $b_a^{(i)}$ is at most quadratic. Thus, standard linear regression can be cast as a QUBO problem.

A QUBO problem consists of finding a vector

$$x^* = \arg\min_{x\in\mathbb{B}^n} Q(x) \tag{9}$$

that is minimal with respect to a quadratic polynomial $Q : \mathbb{B}^n \to \mathbb{R}$ over binary variables $x_i \in \mathbb{B}$ for $\mathbb{B} = \{0, 1\}$ and $i \in [n]$,

$$Q(x) = \sum_{i=1}^{n}\sum_{j=1}^{i} J_{ij}\, x_i\, x_j. \tag{10}$$

Here, the coefficients $J_{ij} \in \mathbb{R}$ satisfy $1 \le j \le i \le n$, and $[n]$ is the set of strictly positive integers less than or equal to n.

When adding the ℓ_p-norm regularization in Eq. (4), we still have a QUBO form for $p = 2$, but not for any other value of p. To understand the role of redundant encodings in this context, it is instructive to first consider the $p = 1$ case. Because of the absolute value signs in Eq. (4), this is not of QUBO form, but it is "almost QUBO" since one could remove the absolute value sign if one knew that a given c_a was either always positive or always negative. Taking inspiration from this observation, consider a redundant encoding of c_a where there are both positive and negative values of g_i, such as in Eq. (7). In that case, we have the following inequality:

$$|c_a| \le \sum_{i=1}^{M} |g_i|\, b_a^{(i)}. \tag{11}$$

This would be an equality if $b_a^{(i)}$ were only non-zero when g_i was positive, or when g_i was negative, but not both. From the perspective of minimizing Eq. (3), though, cases with non-zero values of $b_a^{(i)}$ for mixed signs of g_i are irrelevant, as long as there is another encoding of c_a that only uses all positive or all negative values of g_i (and therefore satisfies Eq. (11) as an equality). This is indeed the case for the example in Eqs. (7) and (8). Therefore, without changing the solution of the sparse regression problem, we can use a modified ℓ_1-norm regulator:

$$R_a^{(1-\mathrm{mod})} = \sum_{i=1}^{M} |g_i|\, b_a^{(i)}, \tag{12}$$

which is now of QUBO form.

We can do something similar for the ℓ_0-norm regulator:

$$R_a^{(0-\mathrm{mod})} = \sum_{i=1}^{M} b_a^{(i)}, \tag{13}$$

which is again of QUBO form. Here, though, for an N-bit binary encoding, there are only $N + 1$ values of c_a that have the correct regulator, namely all of the individual g_i values (which get a penalty of 1) and the value 0 (which gets a penalty of 0). Ideally, we would want a large fraction of achievable c_a values to have at least one $b_a^{(i)}$ configuration with the right penalty. This can be achieved by leveraging a redundant encoding using ancilla bits, as we explain next.

D. Single Ancilla Bit Encoding

The first example of a redundant encoding involves just a single ancilla bit per fit coefficient. This ancilla bit r_a plays no role in determining the value of c_a, but it appears in the ℓ_0-norm regulator as follows:

$$R_a^{(0-\text{single})} = r_a + (1 - r_a) \sum_{i=1}^{M} b_a^{(i)}. \tag{14}$$

This single ancilla bit encoding (ABE) is shown graphically in Fig. 1a, where to match Eq. (17) below, we have separated out $b_a^{(i)}$ into positive ($p_a^{(i)}$) and negative ($n_a^{(i)}$) fit coefficients.

In the context of annealing, we care about the lowest energy configuration. Minimizing Eq. (14) over the ancilla bit r_a, we find that

$$\min_{r_a} R_a^{(0-\text{single})} = \begin{cases} 0 & c_a = 0 \\ 1 & c_a \neq 0 \end{cases}, \tag{15}$$

which is precisely the desired ℓ_0-norm regulator.

We note that an approximate formulation of ℓ_0-norm regularization as an optimization problem has recently been proposed in Ref. [31]. This approach, however, is based on the general expression of k-local problems as QUBO problems, which requires potentially inefficient gadgetization techniques [32–34].

III. DEGENERACY ENGINEERING

A. General Principles

The key idea behind degeneracy engineering is to increase the relative ground-state to excited-state degeneracies of a tractable subset of terms in a given problem Hamiltonian via the addition of ancilla (qu)bits. More specifically, this technique changes the relative degeneracies (but not the values) associated with this subset of Hamiltonian terms, which in our case is the ℓ_0-norm regularizer. Consequently, if one were to optimize the problem Hamiltonian, the success probability of finding the true ground-state energy would be enhanced. Heuristically, the success probability of finding the true ground-state energy of the *full* Hamiltonian is also enhanced. Degeneracy engineering is motivated by similar techniques in variational quantum simulation, where it has been shown that a strong over-parametrization of quantum circuits improves the chance of finding a good approximation of the true solution [35–38].

As we demonstrate in the next subsection, the concept of degeneracy engineering is particularly well suited for Hamiltonians including a penalty term. While ground-state energies of generic Hamiltonians can be negative, penalty terms employ absolute values and thus vanish under minimization. This feature makes penalty terms the ideal candidates for degeneracy engineering. While it is generally hard to engineer multiple *negative* values for generic ground-state energies, one can straightforwardly engineer multiple *zero* values for the ground-state energy of a penalty term. In particular, this can be achieved by exploiting cancellations of positive and negative contributions to the ground-state energy, as we will exemplify in Eq. (18) below. Thus, degeneracy engineering could provide advantages for any optimization problem containing a penalty term, including penalty terms enforcing physical symmetries.

B. Double Ancilla Bit Encoding

To illustrate the concept of degeneracy engineering, we apply it to the example of ℓ_0-norm regularization for sparse linear regression.

The ℓ_0-norm regulator in Eq. (14) has a single minimum, $\min_{r_a} R_a^{(0-\text{single})} = 0$, where $r_a = 0$ and $b_a^{(i)} = 0$. However, the regulator also has an exponentially large degeneracy of the first excited state, $\min_{r_a} R_a^{(0-\text{single})} = 1$, where $r_a = 1$. Thus, in practice, the optimization using the single ABE is expected to perform poorly.

To mitigate this problem, we want to modify the relative degeneracy of the states under consideration. Our goal is to match the degeneracy levels of the minimum and the first excited state, without changing the energy values. To this end, we consider a double ancilla bit encoding (double ABE) of the ℓ_0-norm loss function.

For concreteness, consider the binary encoding:

$$g_i = 2^i. \tag{16}$$

Next, we introduce a redundant encoding where the fit coefficient zero has multiple representations:

$$c_a = \sum_{i=0}^{M} g_i \big(p_a^{(i)} - n_a^{(i)}\big). \tag{17}$$

Here, $p_a^{(i)}$ ($n_a^{(i)}$) are binary coefficients that yield positive (negative) contributions to the fit coefficients.

For the double ABE, we add two ancilla bits (q_a and r_a) per fit coefficient:

$$\begin{aligned} R_a^{(0-\text{double})} = {} & q_a + (1 + 2q_a - r_a) \sum_{i=1}^{M} p_a^{(i)} \\ & + r_a + (1 + 2r_a - q_a) \sum_{i=1}^{M} n_a^{(i)} \\ & - 2 \sum_{i=1}^{M} p_a^{(i)}\, n_a^{(i)}, \end{aligned} \tag{18}$$

as shown graphically in Fig. 1b. Minimizing Eq. (18) over the ancilla bits r_a and q_a, we recover the desired ℓ_0-norm regulator in Eq. (15), but with a higher relative ground-state degeneracy; we now describe in more detail why this is so.

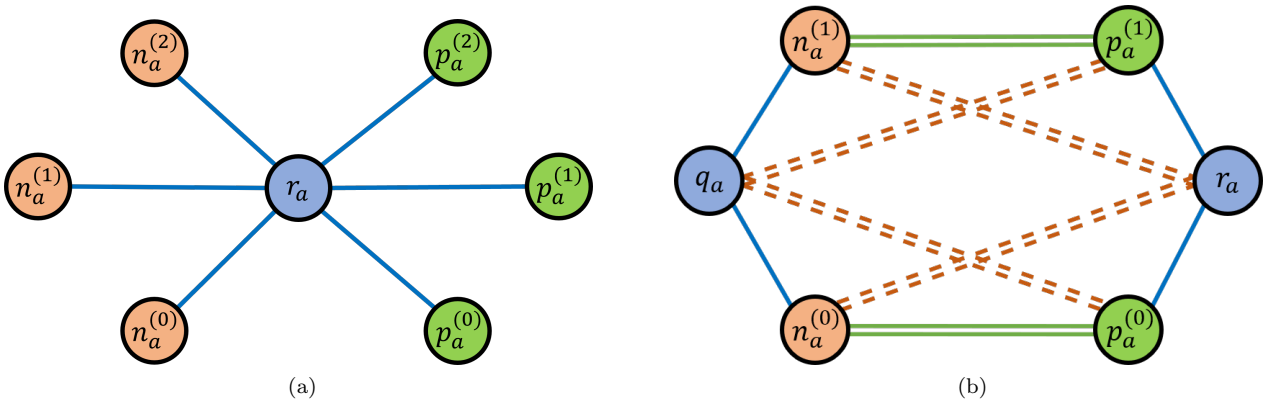


FIG. 1. Graphical representation of the ℓ_0-norm regularizer with (a) single ABE and (b) double ABE. Circles correspond to a penalty of $+1$ for the ancilla bits r_a and q_a (blue) and positive contributions $p_a^{(i)}$ (green) and negative contributions $n_a^{(i)}$ (orange) to the fit coefficients. Lines correspond to penalties of -1 (single solid blue), -2 (double solid green), and $+2$ (double dashed orange).

C. Comparing the Encodings

The graphical illustrations in Fig. 1 can help build intuition about the differing behaviors of the single ABE in Eq. (14) and the double ABE in Eq. (18). Here, the ancilla bits r_a and q_a are depicted as blue nodes, the positive contributions $p_a^{(i)}$ to the fit coefficients are shown as green nodes, and the negative contributions $n_a^{(i)}$ are shown as orange nodes. Turning on any of the nodes is associated with a penalty of $+1$. Solid blue edges correspond to a pairwise penalty of -1, which comes from Eq. (14) and from the first two lines of Eq. (18). Double dashed orange edges correspond to a pairwise penalty of $+2$ from the first two lines of Eq. (18), while double solid green edges correspond to a pairwise penalty of -2 from the third line of Eq. (18).

For the single ABE, the only configuration with zero penalty is the one with all nodes turned off, corresponding to $c_a = 0$. The configurations with penalty $+1$ arise from connected graphs, where the connection is enabled by turning on the ancilla bits r_a. Thus, there is only one ground-state configuration with $c_a = 0$ and a slew of excited-state configurations for $c_a \neq 0$.

For the double ABE, by contrast, there are a large number of configurations with zero penalty and $c_a = 0$, particularly the 2^M configurations associated with turning on pairs of nodes connected by double solid green edges. The configurations with penalty $+1$ and $c_a \neq 0$ arise from connected graphs that do not involve any double solid green edges, of which there are 2^M. Thus, there is a balance between the number of $c_a = 0$ and $c_a \neq 0$ configurations and therefore an improved loss landscape for our ℓ_0-norm regularizer.

It is instructive to compare the single and double ABE in the simplest case of $M = 1$, with two binary fit coefficients p_a and n_a. For the single ABE, we have one ancilla bit r_a. There are four different ways to encode $c_a = 0$, of which the lowest lying state with $R_a^{(0-\text{single})} = 0$ arises from:

(i) turning off all bits.

There are two different ways to encode $c_a = 1$, which are the lowest lying states with $R_a^{(0-\text{single})} = 1$:

(i) turning on just p_a; and

(ii) turning on just p_a and r_a.

Thus, the relative degeneracy of the lowest lying $c_a = 0$ and $c_a = 1$ configurations is 1:2.

For the double ABE, we have two ancilla bits r_a and q_a. There are now eight different ways to encode $c_a = 0$, of which the two lowest lying states with $R_a^{(0-\text{double})} = 0$ are:

(i) turning off all bits, just as for the single ABE; and

(ii) turning on just p_a and n_a.

Similarly, we can encode $c_a = 1$ in four different ways, of which the two lowest lying states with $R_a^{(0-\text{double})} = 1$ arise from:

(i) turning on just p_a; and

(ii) turning on just p_a and r_a.

Thus, the relative degeneracy between the lowest lying $c_a = 0$ and $c_a = 1$ configurations is 1:1.

In this way, we have used the double ABE to successfully engineer a larger ground-state degeneracy without changing the lowest lying energy levels of the system. This general principle of exponentially increasing the ground-state degeneracy of the regularizer can be generalized to $M > 1$ in a straightforward fashion, by turning on various combinations of pairs of $(p_a^{(i)}, n_a^{(i)})$.

There is some freedom in Eq. (18) that could be exploited for practical applications. We chose a penalty of +2 in Eq. (18) (i.e. the dotted orange edges between q_a and $p_a^{(i)}$ and between r_a and $n_a^{(i)}$) to reduce the degeneracy of the first excited states. With a penalty of +1 instead, one could take a connected configuration with total penalty +1 and turn on additional $p_a^{(i)}$ and $n_a^{(i)}$ pairs without additional costs. As long as it is greater than +1, the precise value of this penalty term could be adjusted to optimize the loss landscape.

D. Possible Generalizations

For concreteness, we perform our case studies using just the two example encodings described above. There is, however, a whole family of related redundant encodings that might be relevant for practical applications.

As one extreme example, it is possible to avoid highly connected ancilla bits and instead implement tree graph structures, where each node has penalty +1 and each edge has penalty −1. In this encoding, there are separate graphs for positive and negative coefficients. In each graph, the $g_i = 1$ node is directly connected to the $g_i = 2$ node, instead of being indirectly connected via the ancilla bit. Then, $g_i = 2$ is directly connected to $g_i = 4$, which is directly connected to $g_i = 8$, and so on. Meanwhile, $g_i = 4$ is connected to an additional $g_i = 1$ bit, $g_i = 8$ is connected to additional $g_i = 1$ and $g_i = 2$ bits, and so on. However, such an encoding not only requires a large overhead of additional bits, but the only configuration with $c_a = 0$ and zero penalty is the one with all nodes turned off.[1] Thus, even though such tree graph structures might be advantageous for specific tasks, the double ABE encoding discussed above is, in general, more efficiently implementable. We leave to future work a study combining these redundant tree graphs with partially connected ancilla bits.

IV. OPTIMIZATION STRATEGIES

The results in Sec. VI are based on three different optimization strategies—classical annealing, PIMC, and sparse regularization heuristics—which we describe in this section. While we do not perform quantum annealing on a quantum computer, we review why PIMC is a useful proxy for studying quantum optimization, and we discuss some general considerations when implementing sparse regression on physical quantum devices.

[1] We used powers of 2 for simplicity, but there are ways to optimize the coefficients to reduce the size of the required graph.

A. Review of Classical Annealing

As a representative measure of the performance of traditional classical optimization algorithms, we perform population annealing [27]. For this, we consider a family of canonical distributions parametrized by the inverse temperature β,

$$p_\beta(x) = \frac{1}{\mathcal{Z}_\beta} \mathrm{e}^{-\beta E(x)}, \tag{19}$$

where $E(x)$ is the energy of the state x and $\mathcal{Z}_\beta$ is the partition function. As an alternative to the traditional simulated annealing method of optimization, population annealing considers a population of R_0 replicas of the state x. This population is initialized randomly (i.e. infinite temperature), the first annealing step is performed at temperature $1/\beta_0$, and then the system is cooled to some finite temperature $1/\beta_\ell$ by an annealing schedule of ℓ steps. Unlike simply performing simulated annealing R_0 times, however, with each cooling step, replicas are duplicated or deleted based on an estimate of their relative Boltzmann weights. At each cooling step, the population is reequilibrated according to some Monte Carlo algorithm. As a representative classical method, we equilibrate using Metropolis-Hastings [39].

B. Path Integral Monte Carlo as a Proxy for Quantum Annealing

As a representative measure of the performance of quantum optimization algorithms, we consider a proxy for quantum annealing called the PIMC method. In most stoquastic formulations of quantum annealing, one considers the following parametrized quantum Hamiltonian:

$$H(s) = (1-s)H_i + sH_f = \Gamma(s) \sum_{i=1}^{N} \sigma_i^x + J(s)\, \tilde{L}, \tag{20}$$

where H_i is the initial Hamiltonian and H_f is the final Hamiltonian, called the problem Hamiltonian. The annealing parameter $s = t/t_f \in [0, 1]$ is given by the ratio of the time t and the total annealing time t_f, thus linearly increasing from 0 to 1. Here, $\tilde{L}$ is the operator form of the loss function L from Eq. (3) encoded in the σ^z basis.

To numerically simulate the performance of quantum annealing, we use PIMC. This method employs the Trotter-Suzuki mapping of the quantum annealing Hamiltonian in Eq. (20) to a classical energy function with an extra imaginary time dimension, which is discretized into M imaginary time slices [40]. This well-known mapping from a d-dimensional quantum Ising system to a $(d+1)$-dimensional classical Ising system can be straightforwardly derived using the Trotter breakup formula and spin-1/2 algebra; see App. A for details. We then perform Monte Carlo sampling using the Swendsen-Wang cluster update algorithm [41] with the population

annealing update heuristic [27], forming clusters only in the imaginary time direction on the mapped set of spins [42]. PIMC has been numerically found to accurately simulate quantum annealing in many stoquastic systems [29].

C. Refined Regression as Novel Heuristics

To assess the performance of annealing strategies for ℓ_0-norm regression, we study two novel heuristics: *refined ℓ_1-norm regression* and *refined ℓ_0-norm regression*.

Regression with ℓ_1-norm regularization is often used as a proxy for regression with ℓ_0-norm regularization due to the efficiency of the former. Because the ℓ_1-norm penalty has constant absolute slope everywhere except the origin, it leads to sparse solutions, just like the ℓ_0-norm case. We take this a step further, and consider refined ℓ_1-norm regression. In this strategy, coefficients c_a that are set to zero by the initial ℓ_1-norm regularized regression are clamped to zero. Then, ordinary linear regression is performed on the remaining coefficients to minimize the MSE loss function of Eq. (1). The solution found via this heuristic performs at least as well as the originally found solution in terms of sparsity and MSE loss, though not necessarily in terms of the regularized loss.

We use a similar heuristic to post-process the results of our annealing strategies for ℓ_0-norm regularized linear regression. In refined ℓ_0-norm regression, coefficients set to zero by the annealing process are clamped to zero, and ordinary linear regression is performed on the remaining coefficients. Here, the solution found via this heuristic performs at least as well as the annealed ℓ_0-regularized solution on all performance measures. Given the low computational overhead of unregularized linear regression, we implement this refinement step when presenting our baseline annealing results.

D. Considerations for Quantum Annealing

General adiabatic quantum computation is known to be equivalent to the gate model of quantum computation [2]. Due to experimental considerations, however, most current implementations of quantum annealing platforms use time-dependent *stoquastic* Hamiltonians of the form of Eq. (20), yielding a model of computation that is not believed to be as powerful as general quantum computation. Recently, it was shown that even under a restriction to stoquastic Hamiltonians, there exist oracle separations between quantum annealing and classical algorithms for certain classes of problems [43].

Outside of these specific classes of problems, however, it has been numerically shown that for many QUBO problems, PIMC—a classical algorithm—performs essentially as well as quantum annealing [29]. For this reason, we consider PIMC to serve as a good proxy for quantum annealing in our study.

As we will emphasize in Sec. VII, the concept of degeneracy engineering has important implications for both classical and quantum annealing, beyond just QUBO problems. When solving any optimization problem that employs penalty terms, one can try to engineer multiple zero values for the lowest-energy contribution to the penalty. Extrapolating from the construction in Fig. 1, it appears that degeneracy engineering generally requires ancilla qubit(s) that employ a large degree of connectivity to the other qubits on the platform. Our results therefore stress the importance of good qubit connectivity in quantum annealing platforms.

V. CASE STUDY WITH ENERGY FLOW POLYNOMIALS

The results in Sec. VI are based on a case study in collider physics, where we apply our QUBO formulation of ℓ_0-norm regularization to EFPs. In this section, we briefly review the key properties of EFPs and introduce the observable relations and data sets used in our study.

A. Review of Energy Flow Polynomials

EFPs were introduced in Ref. [21] to accomplish a wide range of jet analysis tasks in high-energy collider physics. EFPs form a discrete linear basis for all infrared- and collinear-safe observables, and many common jet observables are exact linear combinations of EFPs. Many collider tasks can be accomplished using only a handful of EFPs, which makes them an ideal candidate to explore sparse linear regression.

To visualize and calculate the EFPs, Ref. [21] established a one-to-one correspondence between EFPs and loopless multigraphs. For an M-particle jet and a multigraph G with N vertices and d edges $(k,l) \in G$, the corresponding functional expression for the EFP reads

$$\mathrm{EFP}_G = \sum_{i_1=0}^{M} \cdots \sum_{i_N=0}^{M} z_1 \cdots z_{i_N} \prod_{(k,l)\in G} \theta_{i_k i_l}^{\beta}, \tag{21}$$

where β is an angular weighting factor (not to be confused with inverse annealing temperature). For our numerical studies, we take $\beta = 2$. In our case study, the energy fraction z_i carried by particle i and the angular distance θ_{ij} between particles i and j are defined as

$$z_i = \frac{p_{Ti}}{\sum_j p_{Tj}} \quad \text{and} \quad \theta_{ij} = (\Delta y_{ij}^2 + \Delta\phi_{ij}^2)^{1/2}. \tag{22}$$

Here, p_{Ti} is the transverse momentum of particle i, and we use the definitions $\Delta y_{ij} = y_i - y_j$ and $\Delta\phi_{ij} = \phi_i - \phi_j$, where y_i and ϕ_i are the rapidity and the azimuthal angle of particle i.

There are a rich variety of linear relations between different jet observables and EFPs [21, 26], a few of which

we study in this paper. Even for fixed β, the set of all EFPs is an overcomplete basis and therefore needs to be explored using regularized linear regression. This motivates the application of ℓ_0-norm regression to study these linear relations.

For our numerical study, we use the `EnergyFlow` module, which is based on Ref. [21]. This PYTHON package provides all the necessary tools to compute EFPs on collider events, as well as tools to download, read, and manipulate the data sets described in Sec. V C. In our study, we test twelve different linear relations between collider observables and EFPs, which are described in Sec. V B and summarized in Table I.

B. Testing Relations Between Observables

Many common jet observables, including the jet mass, energy correlation functions [44], and angularities [45, 46], are exact finite linear combinations of EFPs. This makes them useful targets for our annealing studies since there is a ground truth definition of successful regularized regression. We consider twelve different linear relations between collider observables and EFPs, which have been extensively studied in Refs. [21, 26]. These twelve relations, summarized in Table I, will serve as benchmarks for testing our QUBO formulation of ℓ_0-norm regression. In Table I, the fourth column represents Eqs. (1) and (2), where y_s corresponds to the observable on the left-hand side, h_a corresponds to the EFPs on the right-hand side, and c_a corresponds to the coefficients to be determined, which optimally match the numbers given in the table.

The first set of observables is given by the infrared- and collinear-safe jet angularities [45, 46] defined as

$$\lambda^{(\alpha)} = \sum_{i=1}^{M} z_i \theta_i^{\alpha}, \tag{23}$$

where $\alpha > 0$ is an angular exponent and θ_i denotes the distance of particle i to the p_T-weighted centroid axis (y_J, ϕ_J) of the jet located at

$$y_J = \sum_{j=1}^{M} z_j y_j, \quad \phi_J = \sum_{j=1}^{M} z_j \phi_j. \tag{24}$$

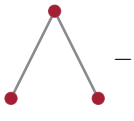
Using Eq. (24), the angularities in Eq. (23) can be expressed in terms of pairwise distances as

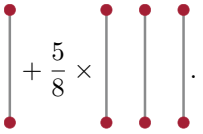
$$\lambda^{(\alpha)} = \sum_{i_1=1}^{M} z_{i_1} \left(\sum_{i_2=1}^{M} z_{i_2} \theta_{i_1 i_2}^2 - \frac{1}{2} \sum_{i_2=1}^{M} \sum_{i_3=1}^{M} z_{i_2} z_{i_3} \theta_{i_2 i_3}^2 \right)^{\alpha/2}, \tag{25}$$

where $\theta_{ij} = (\Delta y_{ij}^2 + \Delta \phi_{ij}^2)^{1/2}$.

For even α, the parenthetical in Eq. (25) can be expanded and identified to be a linear combination of EFPs with $N = \alpha$ and $d = \alpha$ [47]. Focusing on the cases $\alpha \in \{2, 4, 6\}$ and using the multigraph representation of Eq. (21), we can write down the following linear relations for the jet angularities:

$\lambda^{(2)} = \frac{1}{2} \times$  , (26)

$\lambda^{(4)} = \frac{1}{2} \times$  $- \frac{3}{4} \times$  , (27)

$\lambda^{(6)} = \frac{1}{2} \times$  $- \frac{3}{2} \times$  $+ \frac{5}{8} \times$  . (28)

In these three multigraph representations, each edge (k, l) corresponds to a term $\theta_{i_k i_l}$ in Eq. (25), and each vertex j corresponds to a summation $\sum_{i_j=1}^{M} z_{i_j}$.

Next, we consider a jet observable based on the two-dimensional geometric moment tensor of the energy distribution in the (y, ϕ)-plane [47, 48]:

$$C = \sum_{i=1}^{M} z_i \begin{bmatrix} (y_i - y_J)^2 & (\phi_i - \phi_J)(y_i - y_J) \\ (\phi_i - \phi_J)(y_i - y_J) & (\phi_i - \phi_J)^2 \end{bmatrix}, \tag{29}$$

where the distances are measured with respect to the p_T-weighted centroid axis (y_J, ϕ_J) of the jet in Eq. (24). Both the trace and the determinant of this matrix can be expressed as a linear combination of EFPs. The trace is related to the $\alpha = 2$ angularity, while the determinant satisfies [21]

$\det C = \frac{1}{4} \times$  $- \frac{1}{2} \times$  . (30)

In Ref. [26], a variety of relations were derived from cutting the graph nodes. These relations only hold for a limited number of particles, and they can be derived from the fact that anti-symmetrizing L indices of a tensor in M dimensions yields zero for $L > M$. A useful organizational scheme for the EFPs is by the number of edges d in the associated multigraph. We consider two linear relations at $d = 3$, called "Triple Dumbbell" and "Lollipop," which are valid only for events containing $M \leq 2$

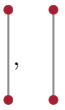
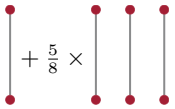
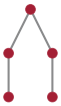
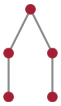
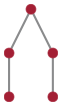
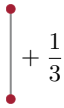
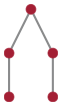
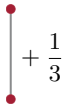
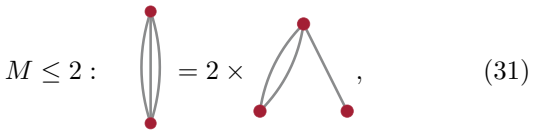
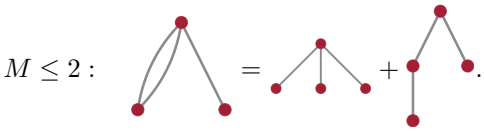
Label	Name of Observable	Restriction	Multigraph Representation of Linear EFP Relation
(a)	Angularity $\alpha = 2$	None	$\lambda^{(2)} = \frac{1}{2} \times$
(b)	Angularity $\alpha = 4$	None	$\lambda^{(4)} = \frac{1}{2} \times$  $- \frac{3}{4} \times$  ,
(c)	Angularity $\alpha = 6$	None	$\lambda^{(6)} = \frac{1}{2} \times$  $- \frac{3}{2} \times$  $+ \frac{5}{8} \times$ 
(d)	Determinant C	None	$\det C = \frac{1}{4} \times$  $- \frac{1}{2} \times$ 
(e)	Triple Dumbbell	$M \leq 2$	 $= 2 \times$ 
(f)	Triple Dumbbell (Approx.)	None	 $\approx 2 \times$ 
(g)	Lollipop	$M \leq 2$	 $=$  $+$ 
(h)	Lollipop (Approx.)	None	 $\approx$  $+$ 
(i)	Five Dots	$M \leq 3$	 $=$  $+ \frac{1}{2} \times$  $- \frac{1}{2} \times$ 
(j)	Five Dots (Approx.)	None	 $\approx$  $+ \frac{1}{2} \times$  $- \frac{1}{2} \times$ 
(k)	Planar Event	$n \leq 2$	 $= \frac{1}{2} \times$  $+ \frac{1}{2} \times$  $+ \frac{1}{3} \times$  $- \frac{1}{6} \times$   $- \frac{1}{4} \times$  
(l)	Planar Event (Approx.)	None	 $\approx \frac{1}{2} \times$  $+ \frac{1}{2} \times$  $+ \frac{1}{3} \times$  $- \frac{1}{6} \times$   $- \frac{1}{4} \times$

TABLE I. Labels and names of the twelve observable relations used in our EFP case study. The third column indicates possible restrictions on their range of applicability, where M is the number of particles in the jet and n is the number of spatial dimensions. The fourth column gives the corresponding multigraph representations of the linear EFP relations and represents Eqs. (1) and (2), where y_s corresponds to the observable on the left-hand side, h_a corresponds to the EFPs on the right-hand side, and c_a corresponds to the coefficients to be determined.

particles [26]:

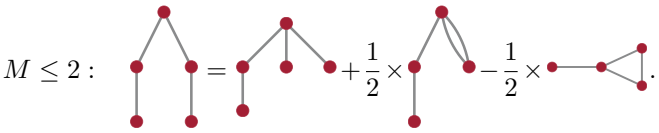


(31)



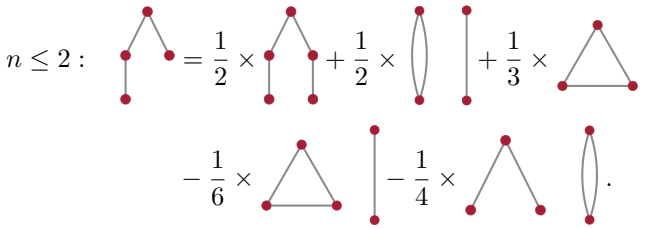
(32)

We consider one example at $d = 4$, called "Five Dots," for events containing $M \leq 3$ particles [26]:



(33)

As the last example, we consider a linear relation called "Planar Event," which is subject to a spatial constraint on the event. In particular, this relation is only applicable to planar events with two (or fewer) spatial degrees of freedom [26]:



(34)

A summary of these linear relations is given in Table I, along with the restrictions that constrain their range of applicability. Additionally, we list "approximate" linear relations, where we consider exact linear relations outside of their range of applicability. This allows us to test the performance of sparse linear regression in regimes where we expect to find good, but not perfect, solutions. In total, we have four exact relations that always hold, four exact relations that hold only with restrictions, and four approximate relations, which yields twelve linear relations that are tested in our numerical study.

C. Data Sets

For our numerical study, we use a data set from the CMS Open Data Portal [49, 50] in the MOD HDF5 Format [51], which was created for jet-based studies. These dijet events are generated in Pythia 6.4.25 [52], and we do not consider any detector simulation effects. In our study, we use 100,000 shower-generated events with $p_T \in [475, 525]$ GeV and absolute values of the rapidity $|y| < 1.9$. Even though the event samples are weighted, for simplicity we treat the events as having equal weights.[2]

For most of the observables in Sec. V B, we can use generic events to test the given functional relations. In specific cases, however, we need to constrain the data to incorporate the specific conditions listed in Table I. For example, some of the linear relations are only applicable to planar events or to events with a specific number of particles. To generate planar events, we constrain the particle motion to two spatial dimensions, which is accomplished by setting the azimuthal angles of all particles to zero, $\phi_i = 0$. To generate events with a fixed particle number, we consider events with larger particle numbers and sequentially delete random particles until reaching the required number. In this process, we preserve the total transverse momentum p_T of the jet by rescaling the transverse momenta of the remaining particles.

As mentioned above, we only apply this preprocessing when testing the "exact" linear relations that are subject to constraints. When testing the "approximate" versions of these linear relations, we leave the data unmodified.

VI. NUMERICAL RESULTS

We now present the results of our numerical study, in which we apply sparse regression to test the twelve linear EFP relations in Table I. Since we have two different annealing encoding schemes, five different optimization strategies, and twelve observable relations to test, we only present selected results to highlight the main points of our study; we present some additional results in App. B.

First, we demonstrate the advantage of degeneracy engineering by comparing the baseline encoding from Sec. II D to the degeneracy-engineered encoding in Sec. III B. Second, we demonstrate the advantage of the refinement approach introduced in Sec. IV C, showing that refined ℓ_0-norm regression performs better than its unrefined version. Third, we demonstrate the advantage of ℓ_0-norm regularization, by showing that it yields a better sparsity/performance trade-off than ℓ_1- or ℓ_2-norm regularization. Finally, we assess the potential gains from quantum computing by comparing classical annealing to PIMC, finding no dramatic difference in performance.

A. Advantage of Degeneracy Engineering

To evaluate the performance of degeneracy engineering from Sec. III, we compare the performance of the single ABE in Eq. (14) versus the double ABE in Eq. (18). For both encodings, we use the same classical annealing

[2] Event weights could be straightforwardly incorporated by generalizing the MSE loss function.

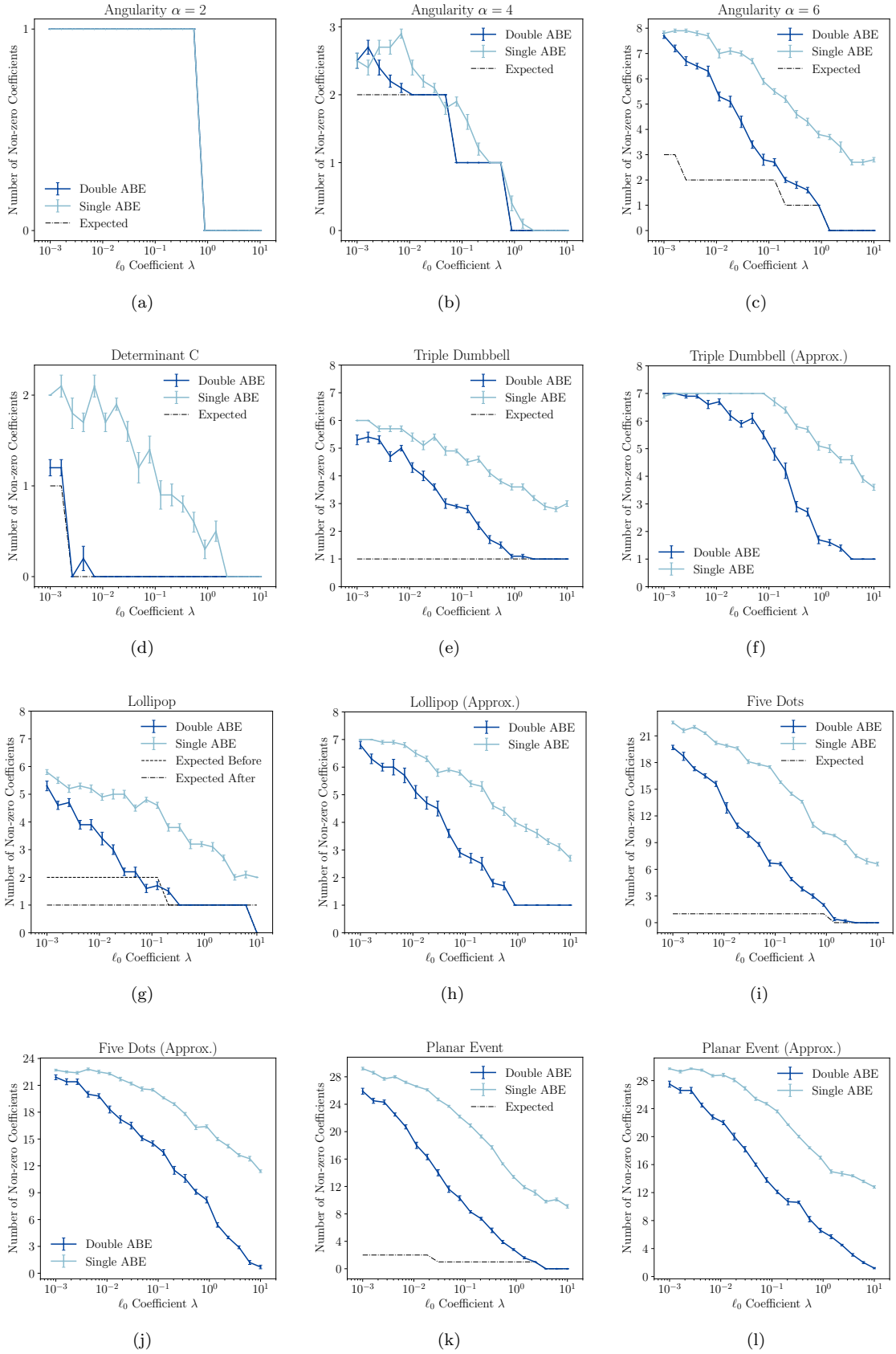


FIG. 2. Number of non-zero fit coefficients as a function of the ℓ_0-norm coefficient λ, comparing single ABE (light blue) with double ABE (dark blue) on classical annealing. The twelve observable relations and their (a)–(l) labels are given in Table I.

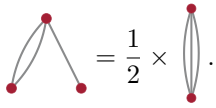
algorithm with the same training parameters for each observable. As described in Sec. IV A, this optimization algorithm is based on classical population annealing with a geometric annealing schedule, with inverse temperature at annealing step i given by

$$\beta_i = \beta_0 \left(\frac{\beta_\ell}{\beta_0}\right)^{\frac{i}{\ell}}. \tag{35}$$

The population is initialized at the temperature $\beta_0 = 1/T_0 = 10$ and then cooled to the temperature $\beta_\ell = 1/T_l = 10^{10}$ by an annealing schedule of $\ell = 2^{14}$ steps. For the ℓ_0-norm coefficient λ, we study a range that spans four orders of magnitude, $\lambda \in [10^{-3}, 10]$.

In Fig. 2, we show the number of non-zero fit coefficients as a function of λ, comparing the single ABE (light blue) to the double ABE (dark blue). The twelve plots in this figure correspond to the twelve different relations in Table I. The results are averaged over ten independent runs, with the standard deviation shown as error bars. For all observables, we find that the degeneracy-engineered version with double ABE performs either equally well or better in terms of the number of identified non-zero fit coefficients. In Fig. 6 of App. B, we plot the loss function versus λ as an alternative way to highlight the improved behavior of the double ABE.

For all non-approximate relations in Table I, we can analytically compute the best-case theoretical expectation by considering all possible combinations of non-zero coefficients given by a particular analytical relation. In Fig. 2, this analytical result is displayed as the black-dashed "expected" line. Interestingly, for the Lollipop observable in Fig. 2g, our original theoretical expectation from Eq. (32) was outperformed by the double ABE algorithm. Indeed, the corresponding black-dashed line ("Expected Before") has more non-zero coefficients than we found numerically. This inspired us to find a different analytic relation for the Lollipop observable, which yielded an improved black-dotted line ("Expected After") with a smaller loss function:

$M \leq 2$:  . (36)

Amusingly, this is just the reversed Triple Dumbbell relation from Eq. (31).

B. Advantage of Refinement

We now evaluate the refinement approach given by the two novel heuristics introduced in Sec. IV C. For this, we use the degeneracy-engineered classical annealing with double ABE, employing the same annealing parameters, annealing schedule, and observables as in Sec. VI A.

When studying the performance of the two novel heuristics, we have to account for the fact that ℓ_0-norm and ℓ_1-norm regression have different loss functions; see Eq. (3). This requires us to choose an alternative presentation compared to Fig. 2 since the meaning of λ differs. We choose to plot the median of the unregularized MSE loss function in Eq. (1) as a function of the mean number of non-zero fit coefficients, since both of these quantities have meaning for any regularization scheme. To compute the error bars for the MSE, we take the 25% and 75% quantiles from ten distinct runs. To compute the mean number and the corresponding error bars of the non-zero fit coefficients, we average over these ten distinct runs.

In Fig. 3, we compare the standard ℓ_0-norm regression (blue) to the two novel heuristics: refined ℓ_0-norm regression (red) and refined ℓ_1-norm regression (orange). As explained in Sec. IV C, we use unregularized regression to refine the non-zero coefficient values while clamping coefficients that were set to zero in the original regularized regression. Refinement improves the MSE performance of standard ℓ_0- and ℓ_1-norm regression with only moderate computational overhead.

The large fluctuations in the median MSE in Fig. 3 are due to the fact that even after refinement, single bit flips in the solution can yield large changes in the model. This makes it somewhat difficult to interpret these plots, but we can draw two general lessons. First, there is a tradeoff between lowering the number of relevant non-zero fit coefficients—implicitly via making λ larger—and increasing the MSE. As the number of non-zero coefficients decreases, the accuracy of the regression solution worsens as expected. Second, the refined ℓ_0-norm regression and the refined ℓ_1-norm regression perform similarly well for all twelve observables we studied. For a fixed number of non-zero coefficients, both refined heuristics yield substantially lower MSE compared to unrefined ℓ_0-norm regression.

C. Advantage of ℓ_0-Norm Regularization

The key premise of our analysis is that ℓ_0-norm regularization should yield sparser solutions to EFP regression problems than ℓ_1- and ℓ_2-norm regularization. To test this, we compare the refined version of ℓ_0-norm regularization to the standard versions of ℓ_1- and ℓ_2-norm regression. As in Sec. VI B, we plot the median MSE loss as a function of the mean number of non-zero fit coefficients.

Results are shown in Fig. 4, for refined ℓ_0-norm regression (red), ℓ_1-norm regression (orange), and ℓ_2-norm regression (green). Because ℓ_2-norm regression does not yield a sparse solution for any value of λ, the green line is vertical on these plots. For specific observables, including the Lollipop observable in Fig. 4g, only refined ℓ_0-norm regression manages to consistently find the exact solution, independently of the number of non-zero coefficients. This can be seen from comparing the very small MSE values for the ℓ_0-norm case to the large MSE values obtained for ℓ_1- and ℓ_2-norm regression. Thus, the

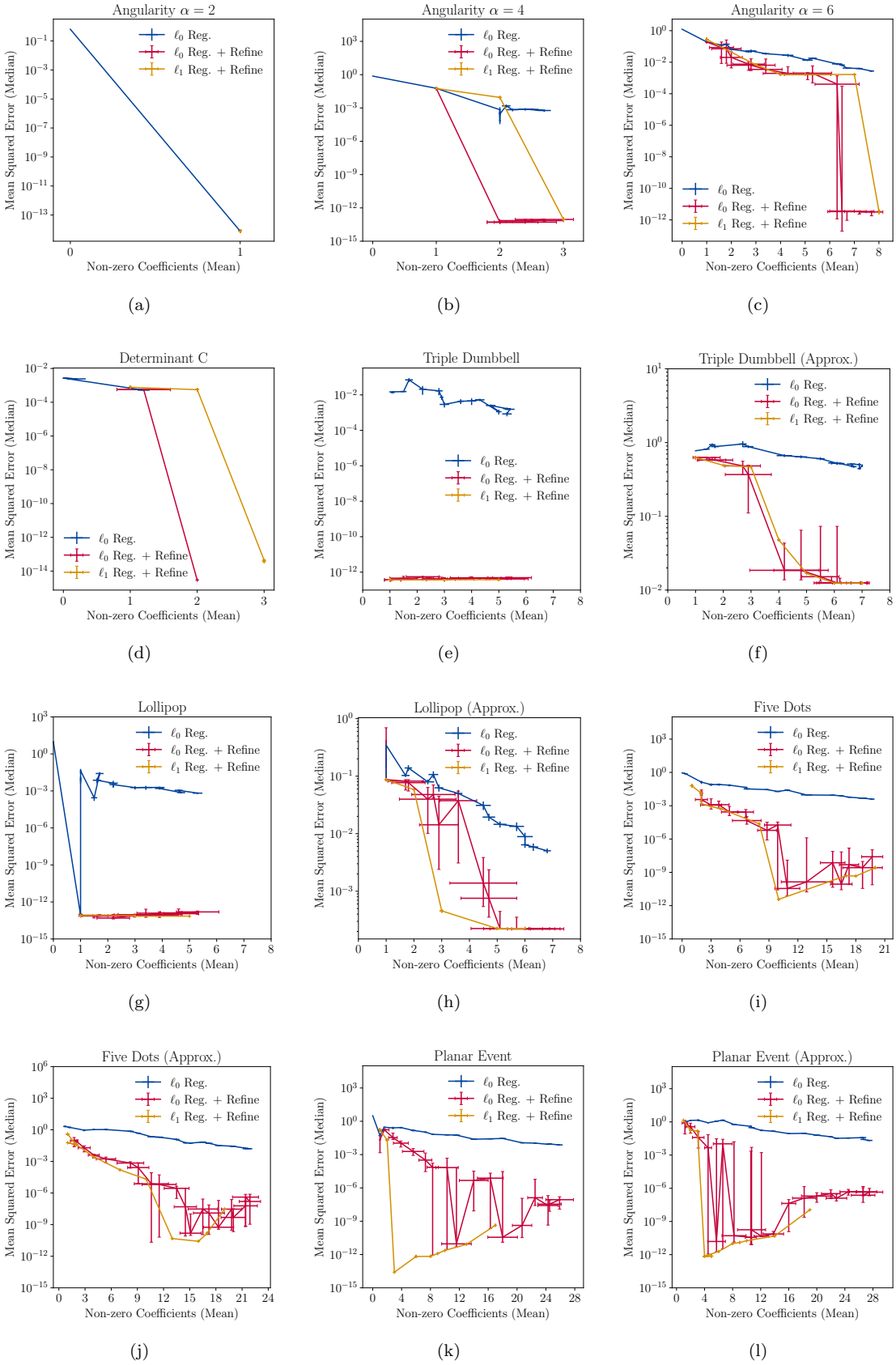


FIG. 3. Median MSE loss function in Eq. (1) as a function of the mean number of non-zero coefficients, comparing ℓ_0-norm regression (blue) with refined ℓ_0-norm (red) and ℓ_1-norm (orange) regression. The same twelve observables from Table I with (a)–(l) labels are shown.

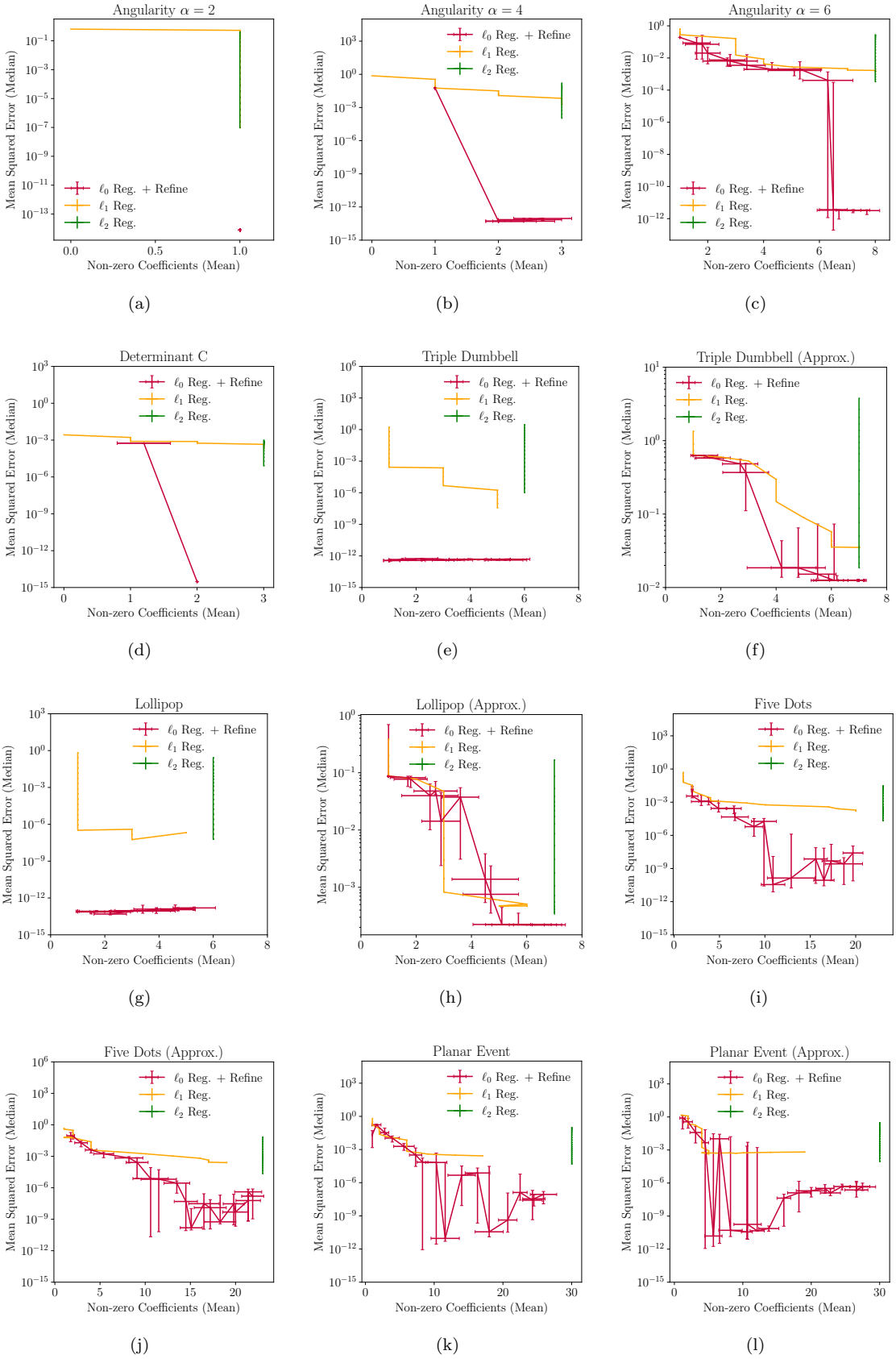


FIG. 4. Same as Fig. 3, but now comparing refined ℓ_0-norm regression (red) to standard ℓ_1-norm (orange) and ℓ_2-norm (green) regression. The (a)–(l) are defined in Table I.

heuristic of refined ℓ_0-norm regression manages to minimize the number of non-zero coefficients as effectively as ℓ_1-norm regression, while also finding an exact solution. For all observables, only refined ℓ_0-norm regression manages to consistently find a nearly exact solution (as measured by MSE), when the number of non-zero coefficients is large.

D. Challenges for Quantum Annealing

As our final numerical study, we assess the potential gains from quantum computing by comparing classical annealing to PIMC. Recall from Sec. IV B that PIMC serves as a proxy for quantum annealing. We use the same observables as Sec. VI A, but a different annealing schedule to attempt an apples-to-apples comparison. For classical annealing, the distributions are initialized at the inverse temperature $\beta_0 = 1/T_0 = 10$ and then cooled to the inverse temperature $\beta_\ell = 1/T_\ell = 10^8$ by a geometric annealing schedule of $\ell = 2048$ steps. For PIMC, $J(s)$ increases geometrically from 10 to 10^8, while $\Gamma(s)$ decreases geometrically from $\Gamma = 1$ to $\Gamma = 0$, once again over $\ell = 2048$ annealing steps. We use double ABE for both methods.

In Fig. 5, we plot the number of non-zero coefficients as a function of the ℓ_0-norm coefficient λ. We compare the performance of classical annealing (solid blue) with PIMC (dashed blue). As in Fig. 2, the error bars are computed by averaging the results over ten distinct runs, and we plot the best-case analytical expectation (black dashed) for all non-approximate relations. In Fig. 7 of App. B, we plot the loss as a function of λ as an alternative way to assess the potential gains from quantum computing.

For the Lollipop observable in Fig. 5g, we again observe that our original theoretical expectation was outperformed by both classical annealing and PIMC. PIMC is actually able to do a better job in the vicinity of $\lambda \simeq 0.1$, though classical annealing does slightly better at smaller λ. For all observables, we find that the performance of classical annealing and PIMC are similar, both with respect to the number of non-zero coefficients and with respect to the loss function; see Fig. 7. The main difference between classical annealing and PIMC is the significantly higher (classical) computation cost of the latter.[3]

These results demonstrate the robustness of the regression performance with respect to changing the annealing method. On the other hand, these findings suggest that true quantum annealing may not yield performance gains for this particular optimization problem.

[3] When trying to perform PIMC on the twelfth observable "Planar Event (Approx.)", we burnt out a laptop power supply, as illustrated in Fig. 5l. We decided against tempting fate to test this observable on a high-performance computer.

VII. CONCLUSIONS

In this paper, we introduced the technique of degeneracy engineering, which is a strategy to improve the performance of both classical and quantum annealing algorithms by increasing the relative degeneracy of the ground state by manipulating a subset of terms in the problem Hamiltonian. We applied this new concept to the NP-hard problem of ℓ_0-norm regularization for sparse linear regression, focusing on a case study in high-energy collider physics.

The key theoretical insights of this paper are twofold. First, we discovered an efficient representation of ℓ_0-norm regularization as a QUBO problem, which opens up the possibility to study this problem with quantum annealing without relying on potentially inefficient gadgetization [32]. Second, we found that the *relative degeneracy* of the ground state of the ℓ_0 regularizer can be increased by increasing the degree of *redundancy* in the qubit encoding scheme for the linear fit coefficients. In practice, our numerical simulations suggest that this changes the spectrum of the total problem Hamiltonian to a spectrum that is more amenable to annealing strategies.

In detailed numerical experiments, we demonstrated the advantages of using ℓ_0-norm regularization for sparse linear regression and of employing degeneracy engineering. In a case study on energy flow polynomials in collider physics, we compared five different regularization methods, including the standard ℓ_2-, ℓ_1-, and ℓ_0-norm regularization, as well as two novel heuristics that refine ℓ_0-norm regularization. We found an advantage of ℓ_0-norm regularization compared to ℓ_1- and ℓ_2-norm regularization, with the best performance obtained using the two refinement heuristics. We also compared standard simulated annealing to path integral Monte Carlo as a proxy for quantum annealing, where we found similar performances for both approaches. Most importantly, we compared different encoding schemes with different degrees of redundancy, finding significantly better performance from the degeneracy-engineered QUBO implementation with a higher degree of redundancy.

What are the prospects, limitations, and requirements of degeneracy engineering? The concept of degeneracy engineering potentially has wide-range applicability, in particular, for Hamiltonians containing a penalty term that is easy to study analytically. Penalty terms are ubiquitous in optimization problems and beyond, ranging from ℓ_0-norm regularization terms to penalty terms enforcing physical constraints or symmetries (see Ref. [53] for an example of penalty terms in particle track reconstruction). While ground-state energies of generic Hamiltonians can be negative, penalty terms employ absolute values and thus vanish under minimization. This feature makes penalty terms the ideal candidates for degeneracy engineering, as one can potentially engineer multiple zero values for the ground-state energy of a penalty term. As we exemplified in Eq. (18), this can be achieved by exploiting cancellations of positive and negative contri-

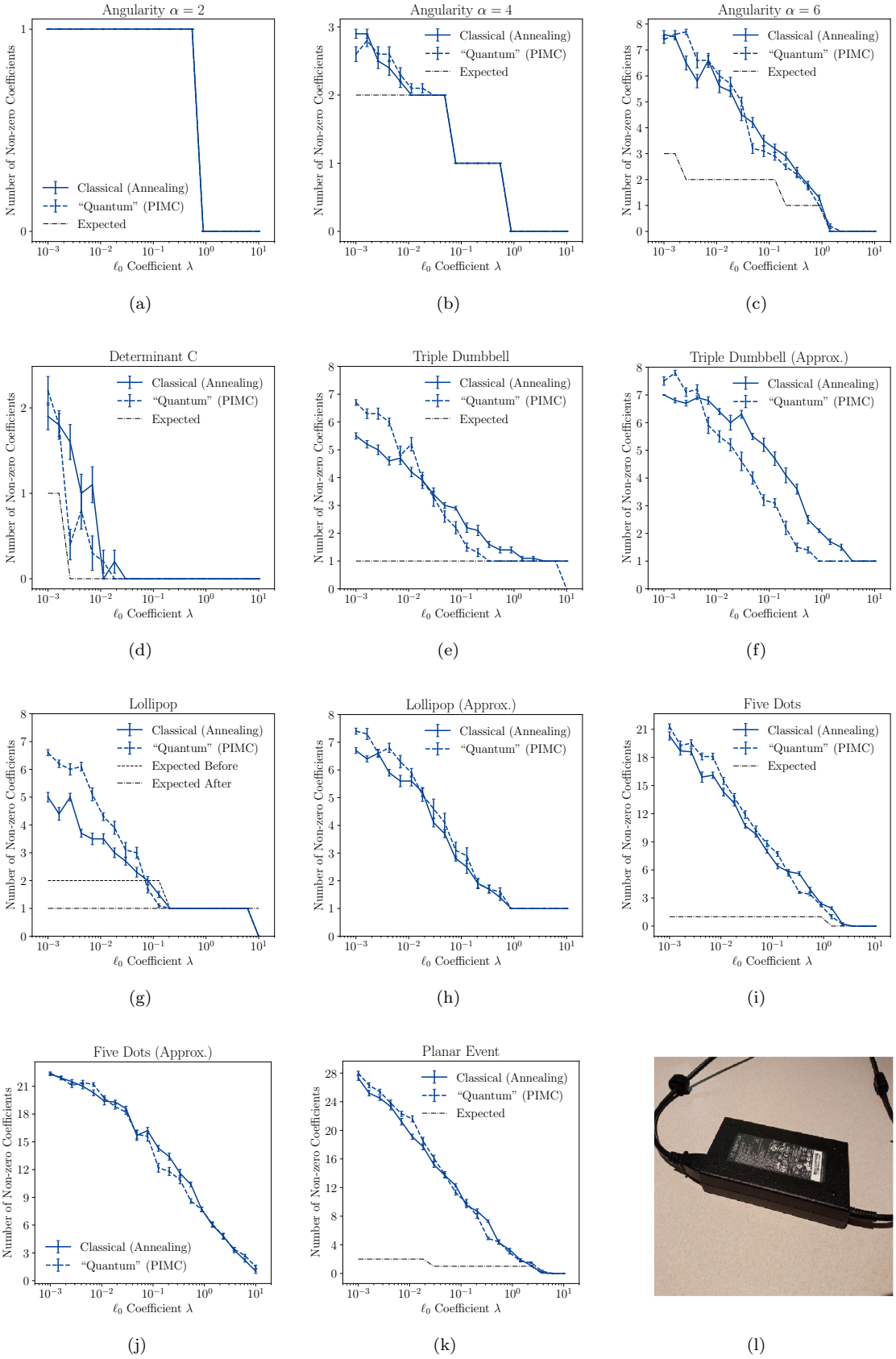


FIG. 5. Same as Fig. 2, but comparing classical annealing (solid blue) to PIMC (dashed blue) as a proxy for quantum annealing, using the double ABE. As discussed in footnote 3, Fig. 5l has been replaced by a burnt charger. The (a)–(l) are defined in Table I.

butions to the ground-state energy. Quantum annealing platforms could substantially benefit from this concept, but would require a large degree of connectivity for the ancilla qubit(s), as illustrated in Fig. 1.

Our results motivate studies of degeneracy engineering for optimization problems beyond sparse linear regression. We also expect degeneracy engineering to be applicable to optimization methods beyond classical and quantum annealing, including variational quantum simulations on digital quantum computers [54, 55] and classical optimization methods like tensor-network methods [56]. If a task is aimed at finding the ground-state energy of a Hamiltonian, then it will likely benefit from engineering more ground-states configurations.

ACKNOWLEDGEMENTS

E.R.A. is supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. 4000063445, and a Lester Wolfe Fellowship and the Henry W. Kendall Fellowship Fund from M.I.T. L.F. and J.T. are supported by the U.S. Department of Energy (DOE), Office of Science, National Quantum Information Science Research Centers, Co-design Center for Quantum Advantage (C^2QA) under Contract No. DE-SC0012704 and by the DOE QuantISED program through the theory consortium "Intersections of QIS and Theoretical Particle Physics" at Fermilab (FNAL 20-17). L.F. is additionally supported by the U.S. DOE Office of Nuclear Physics under Grant Contracts No. DE-SC0011090 and No. DE-SC0021006. This work was supported by the U.S. DOE Office of High Energy Physics under Grant Contract No. DE-SC0012567 and by the National Science Foundation under Cooperative Agreement No. PHY-2019786 (The NSF AI Institute for Artificial Intelligence and Fundamental Interactions, http://iaifi.org/).

Appendix A: Technical Details of Path Integral Monte Carlo

In this appendix, we review some technical details [57] of deriving the path-integral representation of the Ising model used to simulate quantum annealing. We start with the transverse Ising Hamiltonian in Eq. (20),

$$H = \sum_{\langle ij\rangle} J_{ij}\sigma_i^z\sigma_j^z + \Gamma\sum_{i=1}\sigma_i^x, \tag{A1}$$

where J_{ij} are couplings between nearest-neighbor sites and Γ is the transverse field. The latter does not commute with the classical Ising term and therefore turns the Ising model from classical to quantum.

To derive the path-integral representation of the quantum Hamiltonian in Eq. (A1), we first split this Hamiltonian into its kinetic energy term K and its potential energy term U given by

$$K = \Gamma\sum_{i=1}\sigma_i^x, \quad U = \sum_{\langle ij\rangle} J_{ij}\sigma_i^z\sigma_j^z, \tag{A2}$$

such that $H = K + U$ and $[K, U] \neq 0$.

Then, we write down the partition function Z at the temperature $T = 1/\beta$ as

$$\begin{aligned} Z &= \mathrm{Tr}e^{-\beta H} \\ &= \mathrm{Tr}\left(e^{-\beta(K+U)/P}\right)^P \\ &= \sum_{s^1}\cdots\sum_{s^P}\langle s^1|e^{-\beta(K+U)/P}|s^2\rangle \\ &\times\langle s^2|e^{-\beta(K+U)/P}\ldots|s^P\rangle\langle s^P|e^{-\beta(K+U)/P}|s^1\rangle, \end{aligned} \tag{A3}$$

where we inserted the identity operator $\mathbb{1} = \sum_{s^m}|s^m\rangle\langle s^m|$ in the last equality and denoted $s^m = \{s_i^m\}$ as a configuration of all spins in the mth Trotter slice.

Next, we turn the exact expression for the partition function in Eq. (A3) into an approximate expression,

$$\begin{aligned} Z \approx Z_P &= \sum_{s^1}\cdots\sum_{s^P}\langle s^1|e^{-\beta K/P}e^{-\beta U/P}|s^2\rangle \\ &\times\langle s^2|e^{-\beta K/P}e^{-\beta U/P}\ldots|s^P\rangle\langle s^P|e^{-\beta K/P}e^{-\beta U/P}|s^1\rangle, \end{aligned} \tag{A4}$$

by using the Trotter breakup formula,

$$e^{-\beta(K+U)/P} \approx e^{-\beta K/P}e^{-\beta U/P}, \tag{A5}$$

which neglects non-zero commutators of K and U. The expression for Z_P in Eq. (A4) approximates the original partition function Z in Eq. (A3) with an error that is proportional to $(\Delta t)^2$, where $\Delta t = \beta/P$ is the so-called Trotter breakup time.

As a next step, we observe that the potential energy U is diagonal in the chosen spin basis. Thus, the only non-trivial term in Eq. (A4) is the average of the kinetic term K between two Trotter slices,

$$\begin{aligned} &\langle s^m|e^{-\beta K/P}e^{-\beta U/P}|s^{m+1}\rangle \\ &= \langle s^m|e^{-\beta K/P}|s^{m+1}\rangle e^{-\beta U(s^{m+1})/P}. \end{aligned} \tag{A6}$$

The kinetic part of this equation contains a sum over the spin sites in the exponential, which can be expressed as a product of expectation values,

$$\begin{aligned} \langle s^m|e^{-\beta K/P}|s^{m+1}\rangle &= \langle s^m|\exp\left(-\frac{\beta\Gamma}{P}\sum_{i=1}^N\sigma_i^x\right)|s^{m+1}\rangle \\ &= \prod_{i=1}^N\langle s^m|\exp\left(-\frac{\beta\Gamma}{P}\sigma_i^x\right)|s^{m+1}\rangle, \end{aligned} \tag{A7}$$

because spin operators at different sites k and $k+1$ commute. Here, N is the number of lattice sites.

The most crucial step of the derivation, which turns the model from quantum into classical, is the following. In the case of spin-1/2, one can show that

$$\begin{aligned}\langle\uparrow|e^{\alpha\sigma_x}|\uparrow\rangle &= \langle\downarrow|e^{\alpha\sigma_x}|\downarrow\rangle = \cosh(\alpha),\\ \langle\uparrow|e^{\alpha\sigma_x}|\downarrow\rangle &= \langle\downarrow|e^{\alpha\sigma_x}|\uparrow\rangle = \sinh(\alpha),\end{aligned} \tag{A8}$$

which implies that one can rewrite the transversal-field (quantum) term as an Ising-like (classical) interaction between different spins s and s' with $ss' = \pm 1$,

$$\begin{aligned}\langle s|e^{\alpha\sigma_x}|s'\rangle &= \sqrt{(1/2)\sinh(2\alpha)}e^{-(1/2)\ln\tanh(\alpha)ss'}\\ &\equiv Ce^{Bss'}.\end{aligned} \tag{A9}$$

Combining Eqs. (A6), (A7), and (A9), we find

$$\begin{aligned}&\langle s^m|e^{-\beta K/P}e^{-\beta U/P}|s^{m+1}\rangle\\ &= C^N \exp\left(\frac{J_\perp}{PT}\sum_i s_i^m s_i^{m+1}\right)\exp\left(\frac{1}{PT}\sum_{\langle ij\rangle}J_{ij}s_i^m s_j^m\right),\end{aligned} \tag{A10}$$

where we have defined

$$\begin{aligned}J_\perp &= \frac{PT}{2}\ln\tanh\left(\frac{\Gamma}{PT}\right) > 0,\\ C^2 &= \frac{1}{2}\sinh\left(\frac{2\Gamma}{PT}\right).\end{aligned} \tag{A11}$$

Thus, the $J_\perp$ term in Eq. (A10) yields a ferromagnetic Ising-like coupling between the spins s_i^m and s_i^{m+1}, which are nearest neighbors along the Trotter dimension.

Finally, we can express the partition function of the d-dimensional quantum system in Eq. (A4) as a partition function of a $(d+1)$-dimensional classical system,

$$Z \approx Z_P = C^{NP}\sum_{s^1}\cdots\sum_{s^P} e^{-H_{d+1}/PT}, \tag{A12}$$

where the $(d+1)$-dimensional classical Hamiltonian is given by

$$H_{d+1} = -\sum_{m=1}^{P}\left(\sum_{\langle ij\rangle}J(s)s_i^m s_j^m + J^T\sum_i s_i^m s_i^{m+1}\right). \tag{A13}$$

Here, $s^m = \{s_i^m\}$ denotes a configuration of all the spins in the mth Trotter slice, where $M+1$ is identified with m and J^T is the uniform coupling along the extra (imaginary time) direction.

Appendix B: Additional Plots

In this appendix, we present additional plots to complement the discussion in Sec. VI.

In Fig. 6, we give an alternative comparison of double ABE versus single ABE. The advantage of using double ABE was already shown in Fig. 2 in terms of the number of identified non-zero fit coefficients as a function of the ℓ_0-norm coefficient λ. Here, we plot the ℓ_0-norm regularized loss from Eq. (3) as a function of λ, comparing the single ABE (light blue) to the double ABE (dark blue). For all observables, we find that the degeneracy-engineered version with double ABE performs equally well or better in terms of lowering the loss function.

In Fig. 7, we give an alternative comparison of classical annealing and PIMC. Like for Fig. 5, we use the degeneracy-engineered encoding with double ABE, but now plotting the ℓ_0-norm regularized loss as a function of λ. Comparing classical annealing (solid blue) to PIMC (dashed blue), we find similar performance across the twelve relations.

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

[2] E. Farhi, J. Goldstone, S. Gutmann, and M. Sipser, Quantum computation by adiabatic evolution (2000), arXiv:quant-ph/0001106 [quant-ph].

[3] T. Kadowaki, Study of optimization problems by quantum annealing (2002), arXiv:quant-ph/0205020 [quant-ph].

[4] 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**, 472–475 (2001).

[5] S. Lloyd, M. Mohseni, and P. Rebentrost, Quantum algorithms for supervised and unsupervised machine learning (2013), arXiv:1307.0411 [quant-ph].

[6] R. Babbush, P. J. Love, and A. Aspuru-Guzik, Adiabatic quantum simulation of quantum chemistry, Scientific Reports **4**, 10.1038/srep06603 (2014).

[7] A. Perdomo-Ortiz, N. Dickson, M. Drew-Brook, G. Rose, and A. Aspuru-Guzik, Finding low-energy conformations of lattice protein models by quantum annealing (2012), arXiv:1204.5485 [quant-ph].

[8] A. Lucas, Ising formulations of many np problems, Frontiers in Physics **2**, 10.3389/fphy.2014.00005 (2014).

[9] K. Binder and A. P. Young, Spin glasses: Experimental facts, theoretical concepts, and open questions, Rev. Mod. Phys. **58**, 801 (1986).

[10] H. Nishimori, *Statistical physics of spin glasses and information processing: an introduction*, 111 (Clarendon Press, 2001).

[11] M. Mézard, G. Parisi, and M. A. Virasoro, *Spin glass theory and beyond: An Introduction to the Replica Method and Its Applications*, Vol. 9 (World Scientific Publishing Company, 1987).

[12] M. W. Johnson, M. H. Amin, S. Gildert, T. Lanting, F. Hamze, N. Dickson, R. Harris, A. J. Berkley, J. Jo-

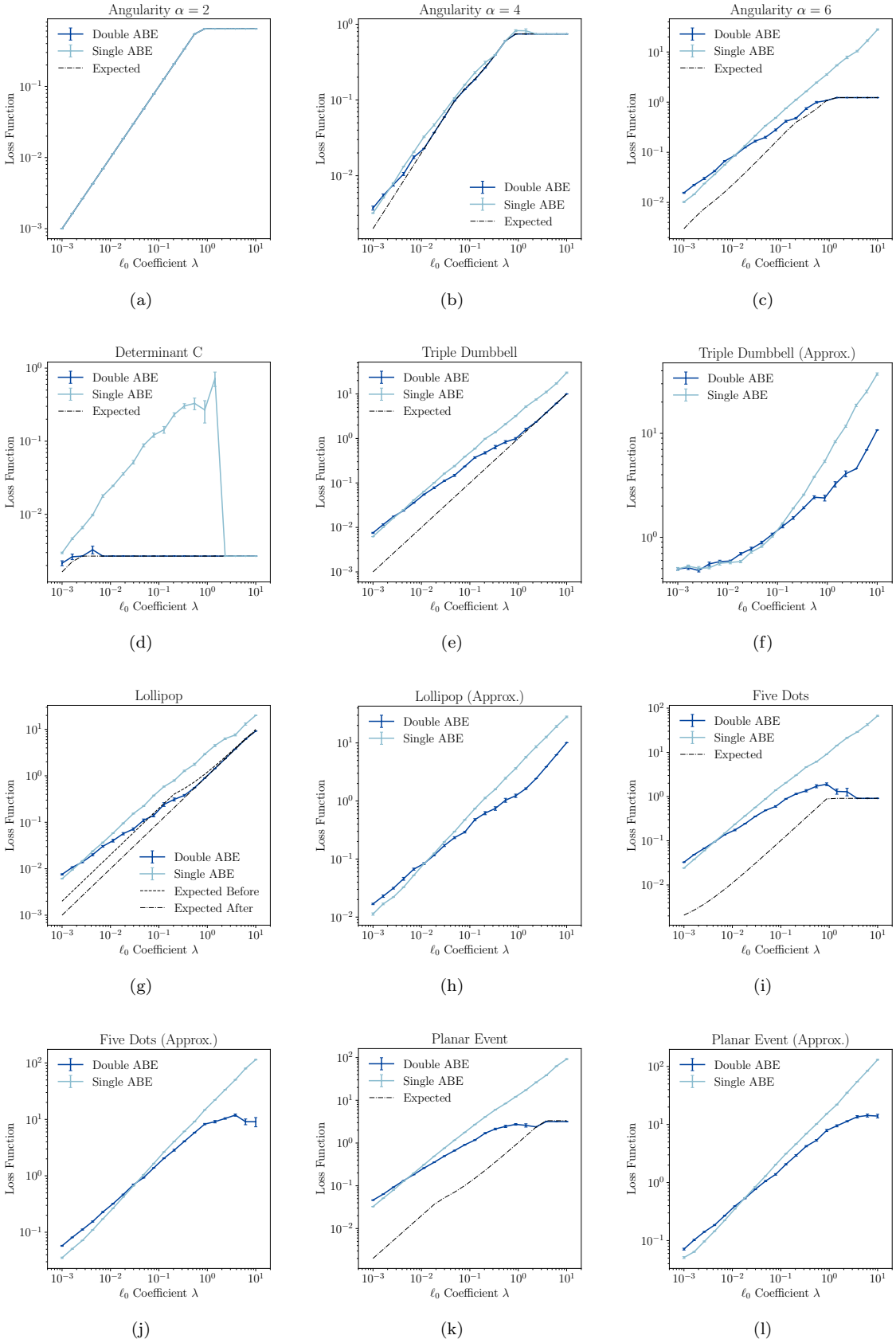


FIG. 6. Same as Fig. 2, but plotting the ℓ_0-norm regularized loss function in Eq. (3) as a function of the ℓ_0-norm coefficient λ. The (a)–(l) are defined in Table I.

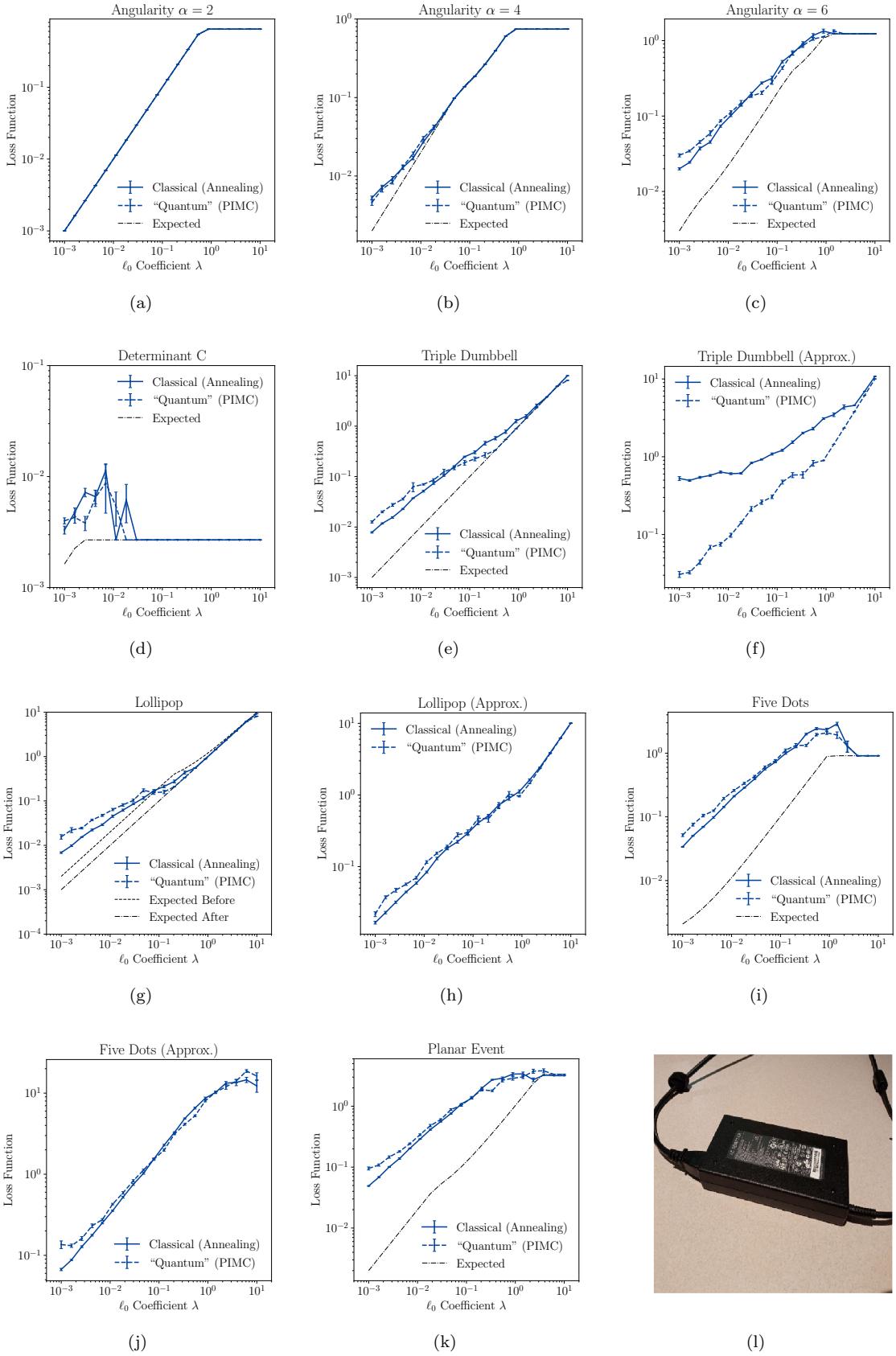


FIG. 7. Same as Fig. 5, but plotting the ℓ_0-norm regularized loss function in Eq. (3) as a function of the ℓ_0-norm coefficient λ. Fig. 7l is identical to Fig. 5l. The (a)–(l) are defined in Table I.

hansson, P. Bunyk, *et al.*, Quantum annealing with manufactured spins, Nature **473**, 194 (2011).
[13] T. Albash and D. A. Lidar, Demonstration of a scaling advantage for a quantum annealer over simulated annealing, Phys. Rev. X **8**, 031016 (2018).
[14] 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).
[15] H. G. Katzgraber, F. Hamze, Z. Zhu, A. J. Ochoa, and H. Munoz-Bauza, Seeking quantum speedup through spin glasses: The good, the bad, and the ugly, Phys. Rev. X **5**, 031026 (2015).
[16] I. Hen, J. Job, T. Albash, T. F. Rønnow, M. Troyer, and D. A. Lidar, Probing for quantum speedup in spin-glass problems with planted solutions, Physical Review A **92**, 042325 (2015).
[17] A. Mott, J. Job, J.-R. Vlimant, D. Lidar, and M. Spiropulu, Solving a higgs optimization problem with quantum annealing for machine learning, Nature **550**, 375 (2017).
[18] A. Zlokapa, A. Mott, J. Job, J.-R. Vlimant, D. Lidar, and M. Spiropulu, Quantum adiabatic machine learning by zooming into a region of the energy surface, Phys. Rev. A **102**, 062405 (2020).
[19] P. Hauke, H. G. Katzgraber, W. Lechner, H. Nishimori, and W. D. Oliver, Perspectives of quantum annealing: methods and implementations, Reports on Progress in Physics **83**, 054401 (2020).
[20] B. Natarajan, Sparse approximate solutions to linear systems, SIAM J. Comput. **24**, 227 (1995).
[21] P. T. Komiske, E. M. Metodiev, and J. Thaler, Energy flow polynomials: A complete linear basis for jet substructure, JHEP **04**, 013, arXiv:1712.07124 [hep-ph].
[22] P. T. Komiske, E. M. Metodiev, and J. Thaler, An operational definition of quark and gluon jets, JHEP **11**, 059, arXiv:1809.01140 [hep-ph].
[23] A. Butter *et al.*, The Machine Learning landscape of top taggers, SciPost Phys. **7**, 014 (2019), arXiv:1902.09914 [hep-ph].
[24] A. E. Hoerl and R. W. Kennard, Ridge regression: Biased estimation for nonorthogonal problems, Technometrics **12**, pp. 55 (1970).
[25] R. Tibshirani, Regression shrinkage and selection via the lasso, Journal of the Royal Statistical Society: Series B (Methodological) **58**, 267 (1996).
[26] P. T. Komiske, E. M. Metodiev, and J. Thaler, Cutting Multiparticle Correlators Down to Size, Phys. Rev. D **101**, 036019 (2020), arXiv:1911.04491 [hep-ph].
[27] K. Hukushima and Y. Iba, Population Annealing and Its Application to a Spin Glass, AIP Conf. Proc. **690**, 200 (2003).
[28] J. A. Barker, A quantum-statistical Monte Carlo method; path integrals with boundary conditions, J. Chem. Phys. **70**, 2914 (1979).
[29] S. V. Isakov, G. Mazzola, V. N. Smelyanskiy, Z. Jiang, S. Boixo, H. Neven, and M. Troyer, Understanding quantum tunneling through quantum monte carlo simulations, Phys. Rev. Lett. **117**, 180402 (2016).
[30] D. S. Phatak and I. Koren, Hybrid signed-digit number systems: A unified framework for redundant number representations with bounded carry propagation chains, IEEE Trans. Computers **43**, 880 (1994).
[31] S. S. T. Desu, P. K. Srijith, M. V. P. Rao, and N. Sivadasan, Adiabatic quantum feature selection for sparse linear regression (2021), arXiv:2106.02357 [cs.LG].
[32] N. Dattani, Quadratization in discrete optimization and quantum mechanics, arXiv preprint arXiv:1901.04405 (2019).
[33] S. Abel, J. C. Criado, and M. Spannowsky, Completely quantum neural networks (2022), arXiv:2202.11727 [quant-ph].
[34] T. Gabor, M. L. Rosenfeld, S. Feld, and C. Linnhoff-Popien, How to approximate any objective function via quadratic unconstrained binary optimization (2022), arXiv:2204.11035 [quant-ph].
[35] E. Fontana, N. Fitzpatrick, D. M. Ramo, R. Duncan, and I. Rungger, Evaluating the noise resilience of variational quantum algorithms, Phys. Rev. A **104**, 10.1103/physreva.104.022403 (2021).
[36] J. Kim, J. Kim, and D. Rosa, Universal effectiveness of high-depth circuits in variational eigenproblems, Phys. Rev. Res. **3**, 10.1103/physrevresearch.3.023203 (2021).
[37] E. R. Anschuetz, Critical points in quantum generative models, in *International Conference on Learning Representations* (2022).
[38] E. R. Anschuetz and B. T. Kiani, Beyond barren plateaus: Quantum variational algorithms are swamped with traps (2022), arXiv:2205.05786 [quant-ph].
[39] W. K. Hastings, Monte Carlo sampling methods using Markov chains and their applications, Biometrika **57**, 97 (1970).
[40] M. Suzuki, Relationship between d-Dimensional Quantal Spin Systems and (d+1)-Dimensional Ising SystemsEquivalence, Critical Exponents and Systematic Approximants of the Partition Function and Spin Correlations, Prog. Theor. Phys. **56**, 1454 (1976).
[41] R. H. Swendsen and J.-S. Wang, Nonuniversal critical dynamics in monte carlo simulations, Phys. Rev. Lett. **58**, 86 (1987).
[42] S. V. Isakov and R. Moessner, Interplay of quantum and thermal fluctuations in a frustrated magnet, Phys. Rev. B **68**, 104409 (2003).
[43] M. B. Hastings, The Power of Adiabatic Quantum Computation with No Sign Problem, Quantum **5**, 597 (2021), arXiv:2005.03791 [quant-ph].
[44] A. J. Larkoski, G. P. Salam, and J. Thaler, Energy Correlation Functions for Jet Substructure, JHEP **06**, 108, arXiv:1305.0007 [hep-ph].
[45] S. D. Ellis, C. K. Vermilion, J. R. Walsh, A. Hornig, and C. Lee, Jet Shapes and Jet Algorithms in SCET, JHEP **11**, 101, arXiv:1001.0014 [hep-ph].
[46] A. J. Larkoski, J. Thaler, and W. J. Waalewijn, Gaining (Mutual) Information about Quark/Gluon Discrimination, JHEP **11**, 129, arXiv:1408.3122 [hep-ph].
[47] G. Gur-Ari, M. Papucci, and G. Perez, Classification of Energy Flow Observables in Narrow Jets, (2011), arXiv:1101.2905 [hep-ph].
[48] J. Gallicchio and M. D. Schwartz, Quark and Gluon Jet Substructure, JHEP **04**, 090, arXiv:1211.7038 [hep-ph].
[49] Cms releases first batch of high-level lhc open data (2014).
[50] Cms releases new batch of research data from lhc (2016).
[51] EnergyFlow Documentation, CMS Open Data and the MOD HDF5 Format, `https://energyflow.network/docs/datasets/#cms-open-data-and-the-mod-hdf5-format`, accessed: 2021-11-18.
[52] T. Sjostrand, S. Mrenna, and P. Z. Skands, PYTHIA

6.4 Physics and Manual, JHEP **05**, 026, arXiv:hep-ph/0603175.

[53] H. M. Gray, Quantum pattern recognition algorithms for charged particle tracking, Phil. Trans. Roy. Soc. Lond. A **380**, 20210103 (2021).

[54] A. Peruzzo, J. McClean, P. Shadbolt, M. Yung, X. Zhou, P. J. Love, A. Aspuru-Guzik, and J. L. O'Brien, A variational eigenvalue solver on a photonic quantum processor, Nat. Comm. **5**, 4213 (2014).

[55] J. R. McClean, J. Romero, R. Babbush, and A. Aspuru-Guzik, The theory of variational hybrid quantum-classical algorithms, New J. Phys. **18**, 023023 (2016).

[56] R. Orús, Tensor networks for complex quantum systems, Nature Reviews Physics **1**, 538 (2019), arXiv:1812.04011 [cond-mat.str-el].

[57] R. Martoňák, G. E. Santoro, and E. Tosatti, Quantum annealing by the path-integral monte carlo method: The two-dimensional random ising model, Phys. Rev. B **66**, 094203 (2002).