FERMILAB-PUB-22-185-V

Reducing the cost of energy estimation in the variational quantum eigensolver algorithm with robust amplitude estimation

Peter D. Johnson,[1] Alexander A. Kunitsa,[1] Jérôme F. Gonthier,[1] Maxwell D. Radin,[1]
Corneliu Buda,[2] Eric J. Doskocil,[2] Clena M. Abuan,[3] and Jhonathan Romero[1]

[1]*Zapata Computing, Inc., 100 Federal St., Boston, MA 02110, USA*
[2]*Applied Chemistry and Physics Centre of Expertise, BP Group Research,*
150 West Warrenville Road, Naperville, IL 60563, USA
[3]*Digital Science and Engineering, BP Innovation and Engineering, 501 Westlake Park Blvd, Houston, TX 77079, USA*

(Dated: March 15, 2022)

Quantum chemistry and materials is one of the most promising applications of quantum computing. Yet much work is still to be done in matching industry-relevant problems in these areas with quantum algorithms that can solve them. Most previous efforts have carried out resource estimations for quantum algorithms run on large-scale fault-tolerant architectures, which include the quantum phase estimation algorithm. In contrast, few have assessed the performance of near-term quantum algorithms, which include the variational quantum eigensolver (VQE) algorithm. Recently, a large-scale benchmark study [1] found evidence that the performance of the variational quantum eigensolver for a set of industry-relevant molecules may be too inefficient to be of practical use. This motivates the need for developing and assessing methods that improve the efficiency of VQE. In this work, we predict the runtime of the energy estimation subroutine of VQE when using robust amplitude estimation (RAE) to estimate Pauli expectation values. Under conservative assumptions, our resource estimation predicts that RAE can reduce the runtime over the standard estimation method in VQE by one to two orders of magnitude. Despite this improvement, we find that the runtimes are still too large to be practical. These findings motivate two complementary efforts towards quantum advantage: 1) the investigation of more efficient near-term methods for ground state energy estimation and 2) the development of problem instances that are of industrial value and classically challenging, but better suited to quantum computation.

I. INTRODUCTION

Identifying a promising candidate for practical quantum advantage lies at the frontier of modern quantum computing research [1–7]. Rapid improvement in quantum hardware [8–12] has given us hope that in the near future we will enter a new technological era marked by the widespread application of quantum algorithms to simulating chemistry and materials, solving differential equations, and modeling financial markets. Is it possible to predict the onset of quantum advantage for a particular industrially relevant problem? What are the quantum resources needed to achieve this using imperfect near-term devices? Our work is motivated by these questions in the context of chemistry simulation, focusing on the prototypical task of ground state energy estimation using the variational quantum eigensolver (VQE) algorithm [13]. The molecular ground state energy is useful for the important molecular modelling task of predicting the enthalpies of hydrocarbon combustion reactions. The starting point of this work is a related study of Gonthier et al. [1], which raised the question: *how do recent advances in quantum estimation [14] improve the pessimistic findings of VQE performance?* Our work aims to answer this question. We approach this problem from a resource estimation perspective, using the same set of molecules as in [1] to facilitate comparison.

Ground state energy estimation is a fundamental problem in molecular quantum chemistry holding the key to multiple industrial applications, such as material design and accelerated drug discovery. It was among the first applications of both fault-tolerant [15] and near-term variational quantum algorithms [13]. Several authors identified it as a promising candidate for industrially relevant quantum advantage, even though a specific problem for which it can be established remains a subject of debate [16–19].

An important step along the path to quantum advantage for quantum chemistry and materials is the development of viable problem instances: identify industrially-relevant quantum systems for which solving the ground state energy problem with sufficient accuracy is classically intractable. Typical examples include so-called multireference systems, where the ground state cannot be even qualitatively described by the Hartree-Fock mean-field approximation. Motivated by this, previous work in resource estimation explored strongly correlated systems such as metalloenzyme cofactors [2, 7, 17], transition metal compounds [18], and the 2-D Fermi-Hubbard model [19]. In the present work, we instead look at systems that are well-described by single-reference methods. However, reaching sufficient accuracy can still be quite costly [1], and these systems have the advantage of readily available, accurate experimental data to gauge the accuracy of classical algorithms. These properties make these molecules a good benchmark set.

Most of these previous resource estimates in quantum chemistry have been devoted to assessing algorithms

This manuscript has been authored by Fermi Research Alliance, LLC under Contract No. DE-AC02-07CH11359 with the U.S. Department of Energy, Office of Science, Office of High Energy Physics.

for large-scale fault-tolerant quantum computers. This leaves open the question: how might near-term quantum computers be used to realize quantum advantage in quantum chemistry?

In the past decade, a host of quantum algorithms have been developed that are suited to the limitations of noisy intermediate-scale quantum (NISQ) devices. The archetype of these methods is the variational quantum eigensolver (VQE) algorithm [13]. VQE is a heuristic algorithm that leverages the variational principle of quantum mechanics to find the best approximation for a ground state of a given molecular Hamiltonian H for a particular choice of a circuit ansatz A [13, 20, 21]. The progress of these near-term quantum algorithms suggests an alternative route to discovering quantum advantage: begin with the capabilities of current quantum devices and determine what minimal improvements are needed for them to solve useful problems. However, carrying out resource estimations for near-term quantum algorithms like VQE is challenging because they are heuristic: unlike traditional quantum algorithms for the ground state energy estimation, such as QPE, VQE does not provide theoretical performance guarantees and needs to be benchmarked on a per case basis, taking into account the target precision and typical problem size.

Recent work [1] carried out a large-scale benchmark study on the resources needed to run VQE. The authors considered a set of molecular Hamiltonians representing industrially-relevant hydrocarbon molecules. They found that the runtime required to reach chemical accuracy (i.e. 1 kcal/mol) for the reaction energies is prohibitive under realistic assumptions for quantum gate times. This large runtime was mostly due to the time needed for the subroutine of energy expectation value estimation with standard sampling; the statistical nature of the energy estimation entails an inverse quadratic scaling of the number of measurements with respect to the target precision.

This statistical phenomenon is responsible for the so-called "measurement problem" of VQE with standard sampling: the number of statistical samples required to obtain sufficiently accurate energy estimates is large, leading to prohibitively large runtimes for the VQE algorithm. From the pessimistic findings regarding the measurement problem in [1], the authors concluded that techniques for speeding up the estimation subroutine used in VQE would be necessary in order to make the algorithm competitive with state-of-the-art classical methods on industry-relevant problem instances. They suggest that techniques like *Robust Amplitude Estimation* (RAE) [14, 22], which increase the rate of information gain in estimation, will be needed to realize quantum advantage for the problem of ground state energy estimation.

RAE offers a new feature among near-term quantum algorithms for estimation: improvements in the quantum computer (as measured by reduction in gate error rates) translate into a proportional improvement in estimation performance (as measured by reduction in runtime). Key to this feature is the robustness of the algorithm: it accommodates a degree of error in the operations by learning a model of the error's effect.

Recent work has investigated the use of similar techniques for the application of Monte Carlo integration in finance [4, 23]. To our knowledge, ours is the first effort to assess these methods for the application of quantum chemistry. The objective of this work is to carry out a resource estimation for robust amplitude estimation applied to VQE energy estimation for the problem instances defined in [1]. Our resource estimates predict that, for the molecules considered, RAE gives between a 13 and 64 fold reduction in runtime over VQE.

The paper is structured as follows. In Section II we review the expectation value estimation techniques of standard sampling (as traditionally used in VQE) and robust amplitude estimation. In Section III we describe the methods used to carry out the resource estimations including algorithm performance modeling, circuit compilation, and the accounting of error correction overhead. In Section IV we describe the results of validating the RAE algorithm performance model and the results comparing the performance of standard sampling to robust amplitude estimation for the benchmark set of molecules. Finally, we conclude in Section V with an outlook on future directions for discovering quantum advantage in quantum chemistry.

II. TECHNICAL BACKGROUND

A. Standard sampling

Before introducing the robust amplitude estimation algorithm we briefly review the standard sampling estimation method. In the simplest setting, standard sampling is used in VQE to estimate expectation values of Pauli strings. For a Hamiltonian decomposed into a linear combination of Pauli strings $H = \sum_j \mu_j P_j$ and "ansatz state" $|A\rangle$, the energy expectation value is estimated as a linear combination of Pauli expectation value estimates

$$\hat{E} = \sum_j \mu_j \hat{\Pi}_j, \tag{1}$$

$$\mathrm{Var}(\hat{E}) \leq \varepsilon^2_{\text{chem. acc.}} \tag{2}$$

where $\hat{\Pi}_j$ is the estimate of $\langle A|P_j|A\rangle$. For a given Pauli operator P, the standard sampling estimation procedure is as follows: prepare $|A\rangle$ and measure operator P receiving outcome $d = 0, 1$; repeat M times, receiving k outcomes labeled 0 and $M - k$ outcomes labeled 1; estimate $\Pi = \langle A|P|A\rangle$ as $\hat{\Pi} = \frac{k-(M-k)}{M}$. In the case that one has some prior information about the value of Π, we can use a Bayesian inference variant of the above estimation process. In this case, expectation values are modeled as binomial distributions with beta priors, such that the measurement process can be assimilated as updating the distribution based on new measurements according to Bayes rule. This approach is referred to as *Bayesian VQE* (BVQE), and is described in [24].

As determined in [1], reaching a high-accuracy energy estimate with standard sampling requires too many independent measurements for VQE to compete with state-of-the-art classical quantum chemistry methods. This is due to the large constant of proportionality K relating the estimation runtime T to the target accuracy ε,

$$T = MC = \frac{CK}{\varepsilon^2}, \tag{3}$$

where M is the number of measurements and C is the time cost of state preparation and measurement. This large proportionality constant of CK is the source of the measurement problem. As described in [1], state-of-the-art methods reduce the value of K but still fall several orders of magnitude short. We refer to this issue as *the measurement problem*. This finding illuminates an obstacle for using VQE as a practical problem solving method and motivates the need for methods which reduce the runtime of estimation more dramatically.

B. Robust amplitude estimation

The robust amplitude estimation algorithm serves to speed up the estimation of expectation values. A detailed description of this method can be found in the following reference [14]. We introduce the RAE algorithm as a solution to the measurement problem discussed in the previous section. In contrast to the estimation method typically used in VQE, RAE enables a reduction of estimation runtime proportional to improvements in the quality of the quantum hardware. Accordingly, we expect that for quantum devices of sufficient quality, we can use the RAE algorithm to carry out energy estimation in a reasonable amount of time.

The robust amplitude estimation algorithm is used to speed up the estimation of each expectation value

$$\Pi = \cos\theta = \langle A| P |A\rangle, \tag{4}$$

where $|A\rangle = A|0^n\rangle$ in which A is the ansatz circuit, P is an n-qubit Hermitian operator with eigenvalues ± 1, and $\theta = \arccos\Pi$ is introduced to facilitate Bayesian inference. The only substantially new circuit operation that is required for this method is a reflection about the initial state $R_0 = \mathbb{I} - 2|0\rangle\langle 0|^{\otimes N}$.

RAE uses the quantum circuit shown in Figure 1 to generate measurement outcomes as follows: prepare the ansatz state $|A\rangle = A|0^n\rangle$, apply L RAE circuit layers $U = AR_0A^\dagger P$, and then measure the Pauli observable P. In the noiseless setting, the likelihood of the outcomes $d = 0, 1$ depends on the parameter of interest $\Pi = \cos\theta$ as

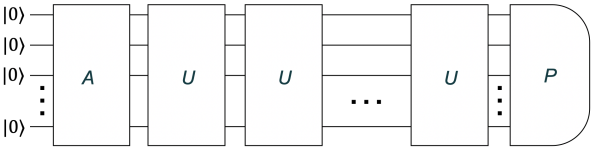

FIG. 1. This figure depicts the operations used for generating measurement outcomes in the robust amplitude estimation algorithm. A is the state preparation circuit, P is the observable of interest (the final circuit operation indicating it's measurement), and U is the Grover iterate comprised of $AR_0A^\dagger P$, with R_0 the reflection about the state $|0\rangle^{\otimes N}$.

$$\mathbb{P}(d|\theta; L) = \frac{1}{2}\Big(1 + (-1)^d \cos((2L+1)\theta)\Big). \tag{5}$$

RAE uses the outcomes from a sequence of such measurements to infer the true value of θ, and hence $\Pi = \langle A|P|A\rangle$. This inference can be implemented in a variety of ways including filtering techniques [25], numerical maximum likelihood estimation [26], and adaptive grid refinement [27]. At the core of each of these methods is the Bayes update rule, whereby a prior distribution $p(\theta)$ capturing initial beliefs about the parameter of interest is updated to a posterior distribution $p(\theta|d)$ by multiplying by the likelihood function and dividing by the model evidence,

$$p(\theta|d) = \frac{\mathbb{P}(d|\theta)p(\theta)}{\int d\theta \mathbb{P}(d|\theta)p(\theta)}. \tag{6}$$

In essence, RAE reduces the estimation runtime by drawing measurement outcomes whose likelihoods depend sensitively on the parameter of interest.

In practice, quantum computation is subject to errors. These errors derive from several sources, including decoherence to the thermal environment and limitations on the calibration of quantum gates. Such errors effect the relationship between the parameter of interest Π and the likelihoods of observed data; the actual likelihood function differs from the idealized likelihood function. To accommodate this, the robust amplitude estimation algorithm incorporates a model for the impact of noise on the inference process. We employ the noise model of [14] to account for the effect of error in RAE. The purpose of the noise model is to predict how errors in the circuit implementation influence the likelihood function from which measurement outcomes are drawn. By composing L noisy RAE circuit layers and measuring the Pauli observable, we model the resulting likelihood function using the *exponential decay* model,

$$\mathbb{P}(d|\theta, \lambda, \overline{p}; L) = \frac{1}{2}\Big(1 + (-1)^d \overline{p}\exp(-\lambda L)\cos((2L+1)\theta)\Big), \tag{7}$$

where $\overline{p}$ denotes the initial ansatz state preparation and measurement error and λ denotes the exponential decay parameter. These two parameters can either be fixed (possibly learned from prior experiments) or can be treated as *nuisance* parameters to be learned during the inference process [28].

In [14], the authors propose a model for the runtime (converted here to be measured in total number of queried RAE circuit layers) for robust amplitude estimation to reach a target mean squared error ε^2, given the noise parameters $\overline{p}$ and λ:

$$t_\varepsilon \approx \frac{e^2}{e-1}\frac{e^{-\lambda}}{2\overline{p}^2}\left(\frac{\lambda}{\varepsilon^2}+\frac{1}{\sqrt{2}\varepsilon}+\sqrt{\left(\frac{\lambda}{\varepsilon^2}\right)^2+\left(\frac{2\sqrt{2}}{\varepsilon}\right)^2}\right), \tag{8}$$

The model shows an interpolation between the traditional scalings known as the shot-noise-limit scaling $O(1/\varepsilon^2)$ and the Heisenberg-limit scaling $O(1/\varepsilon)$. As described in the following section, one of the contributions of this work is the analysis of simulation data which validates this model with respect to a weaker set of assumptions than those used to derive the model. As explained in the following section, this weaker set of assumptions still assumes that the influence of error is perfectly modeled by the exponential decay model likelihood function, leading to over-optimistic conclusions of algorithm performance. Importantly, this suffices for our purposes because we aim to understand the minimal resources necessary, but not sufficient, for achieving quantum advantage.

III. METHODS

The main objective of this work is to predict the resources that are necessary, but not necessarily sufficient, for achieving quantum advantage for the problem of estimating molecular ground state energies. A critical bottleneck in using quantum variational methods to determine the ground state energy is the large number of statistical samples needed for accurate estimates[1]. Thus, a necessary condition for achieving quantum advantage for the ground state energy estimation problem is to carry out the estimation task in a reasonable amount of time. The methods we detail in this section aim to predict the total duration of time needed to ensure an energy estimation to within chemical accuracy for the molecules of interest.

A. Runtime prediction strategy

The task which we analyze is the estimation of the energy expectation value $\langle A|\, H\, |A\rangle$ using RAE, where $|A\rangle = A\,|0\rangle^{\otimes N}$ is the ansatz state. The analysis of this task using VQE was described in [1]. Each molecular Hamiltonian of interest is converted into a weighted sum of Pauli terms as $H = \sum_j \mu_j P_j$. The estimation method first estimates the expectation value of each Pauli term individually as $\hat{\Pi}_j$ and then takes the weighted sum of such estimates as the final energy expectation value estimate according to Equation (1). We consider the case of having a target accuracy ε for the estimate of the final energy, and we wish to minimize the total time required to achieve this target accuracy. Excluding the possibility of estimating Pauli expectation values in parallel across different QPUs from our analysis, we must choose how to optimally allocate time to each estimation so as to minimize the total time. In particular, we should spend proportionally more time, and thus achieve a better accuracy ε_j, on terms with larger coefficients μ_j.

For a single-term estimate $\hat{\Pi}_j$, the total runtime depends on the number of ansatz and phase flip reflection operations used for each estimate and the time taken to implement each of these operations. Since each RAE circuit consists of repeated layers of $U = AR_0A^{\dagger}P_j$, we will first count the total number of such layers used in the estimation process, and then determine the time needed to implement each layer. That time depends on the compilation of the component operations into elementary logical gates. Finally, the time needed to implement each logical gate depends on the quantum error correction resources used and on the runtime of each underlying physical gate. Qualitatively, a larger target fidelity of the gate requires a larger time overhead for its fault-tolerant implementation. Finally, the modality used for the quantum computer can have a significant impact on the time needed for each elementary gate; as a rule of thumb, elementary gates implemented with superconducting-qubit quantum computers tend to be several orders of magnitude faster than those of ion trap quantum computers. Putting all of this together, we arrive at an estimate of the time required to reach a target accuracy in the energy estimate.

We introduce some notation and outline this quantitative strategy. The inputs to the runtime prediction are

- H: Hamiltonian
- ε: Target accuracy, where $\varepsilon^2 = \mathbb{E}(\hat{E} - \langle H\rangle)^2$
- N: Number of logical qubits
- r_g: Elementary physical gate error rate
- $\overline{r}_g$: Elementary logical gate error rate
- T_g: Elementary physical gate time

These inputs determine a number of dependent quantities, which are used to carry out the final runtime prediction. The steps of this process are enumerated as follows:

1. Validate the runtime model of Equation (8) for single-term estimation to target accuracy ε_j: $T_j = \tau_l \cdot t_j(\varepsilon_j, \lambda, \overline{p})$, where τ_l is the duration of a RAE circuit layer and t_j is the total count of the number of RAE layers queried (see Section II B for ε_j, λ, and $\overline{p}$).

2. Establish a method for allocating runtime among terms, amounting to determining optimal target accuracies for each term ε_j^*.

3. Determine circuit depths of the ansatz and phase flip operations D_A and D_R, respectively. Then determine the required logical gate error rate $\overline{r}_g$ from $e^{-\lambda} = (1 - \overline{r}_g)^{(2D_A + D_R)N/2}$. The layer runtime is determined from $\tau_l = (2D_A + D_R)\tilde{T}_g$.

4. Determine fault-tolerant overhead $F(\overline{r}_g, r_g)$ needed to achieve logical gate error rate $\overline{r}_g$ with physical gate error rate of r_g, giving logical gate time $\tilde{T}_g = F(\overline{r}_g, r_g)T_g$.

With these in place the final expression for the runtime to target accuracy is

$$T = \tau_l(D_A, D_R, \tilde{T}_g) \sum_j t_j(\varepsilon_j^*, \lambda, \overline{p}). \tag{9}$$

We note that the time needed for carrying out the estimates in parallel is simply the maximum of the times among the individual terms,

$$T_{\|} = \tau_l(D_A, D_R, \tilde{T}_g) \max_j t_j(\varepsilon_j^*, \lambda, \overline{p}). \tag{10}$$

The methods used for the step of validating the runtime model are reported in Appendix A, and the corresponding results in Appendix A 2. Appendix B describes the compilation model and the assessment of circuit characteristics; Appendix C describes the method for allocating shots across the Pauli expectation values; and Appendix D describes the fault-tolerance cost model used in the final results of Section IV. In the remainder of this section we detail the methods used to make the runtime predictions for standard sampling and RAE.

B. Resource estimation methods

In this subsection we describe the methods for generating energy estimation runtime predictions for standard sampling and RAE.

The problem instances defined in [1] comprise a set of small hydrocarbons for which combustion energies should be calculated to an accuracy comparable to that achievable in experiments. For this purpose, between 104 and 260 qubits are necessary due to the large basis sets involved. Since obtaining and manipulating Hamiltonians for problems of this size is cumbersome with the currently available software, we instead generated two series of Hamiltonians for up to 80 qubits for each molecule, and used the results to extrapolate to the large qubit numbers. The two series of Hamiltonians are generated with two different orbital types, i.e. two different discretizations for the problem, like in [1]. As detailed in [1], we built Hamiltonians so that the number of qubits used is an integer multiple of the number of active electrons, to facilitate extrapolation. For example, for H_2O with 8 active electrons, we built Hamiltonians with 16, 24, 32, 40, 48, 56, 64, 72 and 80 qubits. The Hamiltonian with only 8 qubits is omitted since it would trivially yield the mean-field Hartree-Fock energy.

To obtain the predicted runtime for RAE, we use the following steps:

1. Estimate the total number of Ansatz queries required by RAE to reach chemical accuracy for each Hamiltonian, based on the runtime model validated in Appendix A and on the allocation detailed in Appendix C. This was done for final RMSEs of 10^{-3} and 10^{-4} Ha and for circuit layer fidelities $e^{-\lambda}$ comprised between $(1 - 10^{-3})$ and $(1 - 10^{-6})$ on an approximately logarithmic scale.

2. For a fixed molecule, orbital type, RMSE and layer fidelity, the number of Ansatz queries was fitted using SciPy as a function of the number of qubits with $aN^b + c$ where a, b, and c are fitting coefficients. In the few cases where only two data points were available, the coefficient c was fixed to zero.

3. Plug in the appropriate number of qubits (between 104 and 260) in the function resulting from the fit to estimate the number of Ansatz queries necessary to reach chemical accuracy.

4. Convert the number of Ansatz queries to runtimes in seconds using Equation (9).

The runtime predictions for standard sampling follow those of [1] with a few modifications. Compared to [1], we consider the runtime of standard sampling energy estimation for varied gate error rates. Accordingly, we introduce two competing factors in the runtime predictions. The gate error rates are decreased using quantum error correction; as mentioned above, improving gate error incurs a time overhead that we factor into the runtime predictions. This reduction in gate error rate is helpful, though, in that it reduces the degree to which any error mitigation technique will adjust the expectation value estimates. We will assume an idealistic error mitigation technique that simply rescales the expectation value estimates so as to invert the attenuation factor in the observed expectation value $\langle P_i \rangle_{obs} = f \langle P_i \rangle_{ideal}$, where f is the circuit fidelity. An example of such an estimator is probabilistic error cancellation [29]. In rescaling the observed expectation value by $1/f$, the variance in the estimate is scaled by $1/f^2$ (c.f. Eq. 15 in [29]). Accordingly, we model the error mitigation overhead as a factor in the standard sampling energy estimation runtime

$$\frac{1}{f^2} = \frac{1}{(1-\bar{r}_G)^{D_A N}}, \tag{11}$$

where we take the circuit fidelity to be the product of the logical gate fidelities $1-\bar{r}_G$ of all $D_A N/2$ gates in the ansatz circuit. Note that we have made the optimistic assumption that the readout error is negligible compared to that of the gate error. Hence, to obtain the predicted standard estimation runtimes at large qubit numbers, we follow these steps:

1. Compute K for each Hamiltonian according to the method presented in [1], except that *no grouping* of Pauli terms or variance reduction technique is applied to the Hamiltonians.
2. For a fixed molecule and orbital type, fit K to $aN^b + c$ where a, b, c are fitting coefficients and N the number of qubits. c is set to zero if only two data points are present.
3. Evaluate $M = K/\varepsilon^2$ with K obtained from the formula fitted above applied to large qubit numbers, for ε of 10^{-3} and 10^{-4} Ha.
4. Compute the error mitigation overhead. The logical gate error rate is chosen to be consistent with the RAE circuit fidelities chosen above. Factor the error mitigation overhead with the execution time obtained from D_A and the logical gate execution time $\tilde{T}_g$ that includes the appropriate error correction overhead.

In Section IV we present plots and summarize the findings of these extrapolations.

IV. RESULTS

We now describe the results of the runtime prediction described in Section III B. These results are depicted in detail for two specific molecules in Figure 2 and summarized for the full set of molecules in Table I.

Figure 2 shows the predicted runtimes for the two estimation methods (standard sampling of VQE and robust amplitude estimation) as a function of logical gate error rate. We have chosen the molecules CH_4 and CO_2 to represent the smallest (13×) and largest (64×), respectively, relative improvements of RAE over standard sampling. The figure shows the result of how reducing the logical gate error rates affords the RAE algorithm to run deeper quantum circuits, increasing the degree of quantum amplification and subsequently reducing the runtime of the estimation task. The shape of the runtime vs error rate curve reveals an important phenomenon: for a target estimation accuracy, the runtime has a minimum at an optimal logical gate error rate. We observe the presence of a minimum for both RAE and standard sampling. In the case of RAE, this phenomenon is due to the balance of increased quantum amplification with the cost of error correction overhead. In the case of VQE, the phenomenon is due to the balance of reduced cost of error mitigation overhead with the cost of error correction overhead.

This figure also compares the predicted runtimes of estimation when using compilation to different device connectivities: all-to-all (A2A) and two-dimensional (2D). We observe that the best predicted runtimes are achieved by RAE when all-to-all compiling is used. When the operations are compiled according to a 2D connectivity, the predicted runtime of RAE is larger than that of VQE with any connectivity; this is due to the phase flip operation requiring many gates when compiled according to a 2D connectivity (see Table II).

Table I shows physical and logical qubit resources, gate error rates, and runtime data for each of the eleven molecules studied. We compare these resources and runtimes for standard sampling and RAE. In each case, the data is presented for the optimal logical error rate (minimal runtime) unless otherwise specified. In the case of standard sampling, the number of physical qubits (accounting for error correction overhead) ranges from tens of thousands to roughly a hundred thousand qubits. RAE can take advantage of further reducing the logical error rate. This comes at the cost of an additional qubit overhead that is three to four times that of standard sampling. This additional overhead results in more efficient error correction, bringing the logical error gate for optimal RAE operation five to six orders of magnitude below that of standard sampling. This reduction in logical gate error rate enables a factor of 13 to 64 reduction in the RAE runtime compared to standard sampling.

Despite the predicted runtime improvements of RAE, we observe that these times are still too high to be practical; the lowest predicted runtime is more than *one millenium*. However, like standard sampling, this estimation method is highly parallelizable; running on multiple quantum processing units gives a proportional reduction in runtime. Furthermore, we highlight once again that these runtime estimates do not include any grouping or variance reduction methods. In the case of standard sampling, such methods can reduce the estimated runtime by three to five orders of magnitude. However, RAE operates quite differently from standard sampling, and thus it remains to be seen if a similar improvement can be obtained and how methods used in standard sampling can be adapted.

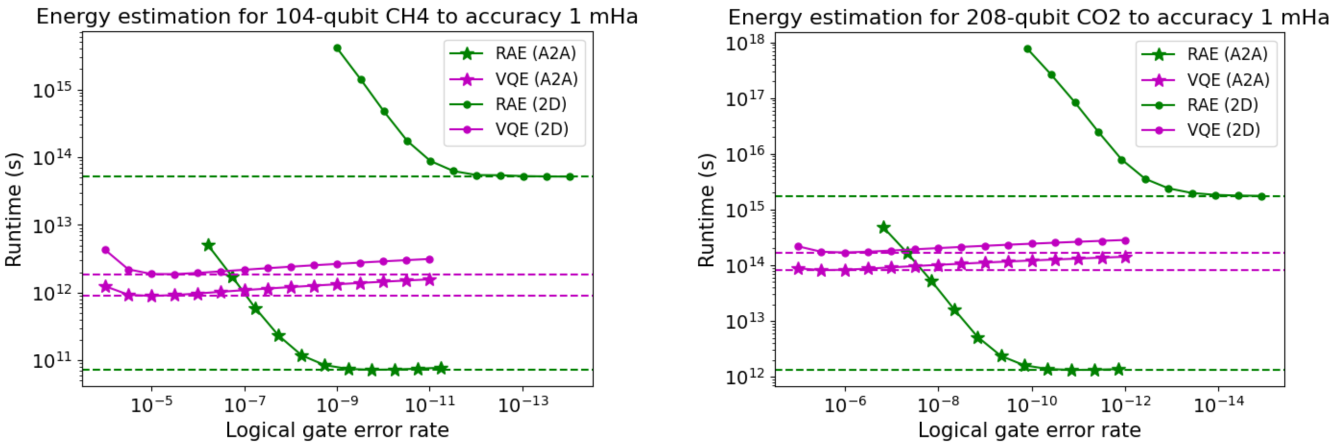

FIG. 2. These figures compare the predicted runtimes of standard sampling (magenta), as typically used in VQE, and robust amplitude estimation (green) for estimating the energies of the CH_4 and CO_2 molecules to within accuracy of 0.001 Hartrees as the logical gate error rate is improved through error correction. The star and dot symbols indicate the use of compilation from all-to-all (A2A) connectivity and two-dimensional (2D) connectivity, respectively. The horizontal dotted lines mark the minimal predicted runtimes for each method and connectivity. We have chosen the CH_4 and CO_2 molecules because they yield the smallest and largest speedups of RAE over standard sampling, respectively. For both molecules, using RAE with A2A connectivity yields the lowest predicted runtime. However, the predicted runtimes are still too high to be practical.

Molecule	C_2H_2	C_2H_4	C_2H_6	C_2H_6O	C_3H_4	C_3H_6	C_3H_8	CH_4	CH_4O	CO_2	H_2O
Number of logical qubits	130	156	182	260	208	234	260	104	182	208	104
Number of VQE physical qubits	43,900	61,200	71,300	117,000	81,500	105,000	117,000	35,200	71,300	81,500	35,200
Number of RAE physical qubits	162,000	195,000	228,000	325,000	260,000	292,000	352,000	120,000	228,000	281,000	130,000
Optimal VQE gate error rate	1e-5	4e-6	4e-6	2e-6	4e-6	2e-6	2e-6	1e-5	4e-6	4e-6	1e-5
Crossover RAE gate error rate	4e-8	3e-8	2e-8	9e-9	1e-8	1e-8	9e-9	6e-8	2e-8	1e-8	6e-8
Optimal RAE gate error rate	4e-11	3e-11	6e-11	3e-11	4e-11	4e-11	9e-12	2e-10	6e-11	1e-11	6e-11
Estimation runtime VQE (10^9s)	1,700	3,900	22,100	477,000	98,200	125,000	189,000	910	32,700	83,900	1,400
Estimation runtime RAE (10^9s)	70	130	1,100	20,000	5,100	6,200	7,000	72	1,800	1,300	41
Runtime ratio (VQE/RAE)	25	30	20	25	19	20	27	13	18	64	34

TABLE I. This table shows the predictions of resources needed to estimate the energy of the prepared ground state using standard sampling (denoted as VQE) or RAE to below chemical accuracy (1.0mHa<1.3mHa) for each molecule (represented using canonical orbitals). For both standard sampling and RAE (for each molecule), we choose the logical gate error rate so that the runtime is minimized, and we report it as the optimal error rate. In addition, for RAE we indicate the crossover error rate, at which RAE yields lower runtimes than standard sampling. The phase flip operation used in RAE is compiled assuming an all-to-all connectivity and we assume surface code cycle times of 1μs.

V. DISCUSSION AND OUTLOOK

This work contributes to identifying the most promising near-term methods for achieving quantum advantage in molecular simulation. In previous work [1], we identified the “measurement problem” as a bottleneck in the variational quantum eigensolver (VQE) algorithm, making the subroutine of energy estimation prohibitively slow even for small molecules of industrial relevance. Here, we have investigated the extent to which quantum amplification helps to reduce the runtime needed to estimate the energy expectation value in VQE. Specifically, we carried out runtime predictions for the energy estimation subroutine when using robust amplitude estimation (RAE) [14]. We used a performance model for RAE (see Appendix A 2) to carry out runtime predictions for a range of system sizes and molecules. We then extrapolated these results to the larger system sizes of interest.

Using a coarse approach to fault-tolerant compilation and making optimistic assumptions about the operation speeds of the fault-tolerant architecture, we arrive at runtime estimations in terms of number of seconds. We find that using RAE gives a 13 to 64 factor speedup over standard sampling, with RAE requiring just a few times as many physical qubits as VQE from the error correction overhead. Although the predicted runtimes for both methods are still too high to be of practical value, we note that they were obtained without any grouping of the Hamiltonian terms or variance reduction techniques. In the case of standard sampling, such methods can reduce the predicted runtime by up to 5 orders of magnitude. Some grouping methods are compatible with RAE, but it remains to be seen whether the associated cost will yield practical methods. We leave this investigation to future work.

Beyond Hamiltonian grouping techniques, we discuss several other ways in which predictions with more reasonable runtimes might be obtained:

1. Circuit depth reduction: Any reduction in circuit depth saves on time and error; both of which reduce the runtime of RAE. It is possible to reduce circuit depth through improved compilation schemes (e.g. in the phase flip [30]) or improved ansatz circuits [31]. Furthermore, the cost of the input-state reflection in RAE can be greatly reduced by exploiting symmetry in the ansatz circuit and Pauli operators. In particular, for number-conserving ansatz and diagonal Pauli operators, it is possible to reduce the costly reflection to a k-qubit operation (where k is the number of electrons). We estimate the time savings here to be roughly an order of magnitude.

2. Improved fault-tolerant resource estimation: the current resource accounting is quite coarse. A fine-grained resource estimation could lead to either more or less optimistic predictions. However, we did not account for optimizations in the fault-tolerant compilation. For example, one can optimally allocate spacetime volume between qubit counts and time (see, e.g. Appendix C of [32]). Moreover, recent advances in quantum error correction [33–35] may eventually lead to further reductions in predicted runtimes.

3. Quantum-apt application instances: alternative sets of molecules that are still of industrial relevance may serve as better candidates for achieving quantum advantage. Preliminary resource estimations indicate that, for the set of molecules considered, even modern quantum phase estimation techniques take on the order of hours to days to run. We believe that a critical effort towards realizing quantum advantage will be the development and identification of problem instances which are "easy" for a quantum computer and "hard" for state-of-the-art classical methods, while still being of practical value.

Thus far, we have considered robust amplitude estimation as a tool to speed up the energy estimation subroutine of VQE. Yet there are other aspects of VQE which stand to be improved. In the outer loop of the VQE algorithm, a classical optimization process is used to find the ansatz parameters for which the corresponding circuit well approximates ground state preparation. This process can be improved by designing better ansatz circuits [31, 36] and finding more effective methods for parameter optimization [37]. Recent work introduced the concept of state preparation boosters [38] as a method to reliably increase ground state overlap at the cost of using deeper quantum circuits.

Given the remaining challenges for VQE, it is likely that additional quantum algorithm methods will be needed to solve problems of industrial value [39–41]. This is consistent with the perspective that VQE might be used to get a "head start" in the state preparation subroutine for more powerful quantum algorithms, i.e. the output state of the VQE calculation provides a rough approximation of the ground state that can be used as an input for another quantum algorithm.

The next decade is sure to bring the value of quantum computing more clearly into view. We hope this work will help guide the community towards identifying the first quantum computing use cases in quantum chemistry and we expect insights from this benchmark study to inform future benchmarks for molecules that show promise for early quantum advantage.

ACKNOWLEDGEMENTS

The authors would like to acknowledge support from BP. The authors also acknowledge insightful scientific discussions and suggestions from the team of scientists and engineers at Zapata Computing. Numerical results were generated using the Orquestra® platform by Zapata Computing Inc.

Appendix A: Validation of single-term estimation runtime model

In this section of the appendix we present the methods and results for validating the single-term estimation runtime model of Equation (8).

1. Methods

An accurate theoretical analysis of RAE's performance is difficult to obtain due to the adaptivity of the algorithm. In each step of inference, L is chosen to maximize the expected gain in Fisher information per time spent for that sample. However, by making several approximations and assumptions, [14] arrived at upper and lower bounds on the runtime to target estimation. While these bounds were derived for the case of using "engineered likelihood functions" (c.f. [14]), we will find that the bounds mostly capture the performance of estimation with Chebyshev likelihood functions, which we are simulating. Our goal is to test the accuracy of these bounds through simulation of the inference process, while relaxing some assumptions used in the runtime model. The question we aim to answer is: how accurate are the runtime bounds in the case where the true sample rates $\mathbb{P}(d|\theta; L)$ *match* those of the likelihood function $\mathbb{P}(d|\theta, \lambda; L)$ (we assume $\overline{p} = 1$)? In practice, we expect a discrepancy between the true sample rates and the likelihood function used for inference because of the discrepancy between our noise model and actual noise. This discrepancy will, in general, make the expectation value estimates less accurate, leading to worse performance of the algorithm. Accordingly, we will view our results as roughly establishing a lower bound on performance: we expect estimation runtimes in practice to be longer than the ones we obtain.

The three main assumptions used in the simulations are:

1. The effect of noise is described by the exponential decay model of Equation (7) with a known decay parameter λ.
2. The decay parameter λ is determined by the number of effective two-qubit gates in a single Grover iterate (accordingly, λ is independent of the Pauli term involved in the Grover iterate).
3. The duration of the estimation process is determined by the cumulative duration of the circuit implementation time.

Regarding assumptions 1 and 2, in practice the likelihood function is simply a model for the relationship between the parameter of interest and the outcome likelihoods. Therefore, our runtime model does not fully capture the case of likelihood function inaccuracies. However, the present evaluation provides a baseline for further analysis including such inaccuracies.

Regarding assumption 3, the total estimation time in practice will include measurement time and a latency between each measurement due to subsequent reinitialization. However, as the duration of the quantum circuits increases, the relative proportion of time spent on measurement and latency will decrease. Because we will be considering circuits of considerable depth in the regime where some quantum error correction is used, we will take this measurement and latency time to be negligible.

In our numerical experiments, we run simulated inference under a number of different settings and estimate the median error over many trials. In each trial, the prior distribution is chosen as a Gaussian distribution over the phase angle $\theta = \arccos(\Pi)$ in a randomized fashion as follows. The standard deviation of the prior is set to $\sigma = 0.01$ and the mean is drawn from a Gaussian distribution centered around the true phase angle with standard deviation $\sigma = 0.01$. This choice reflects a practical strategy for energy expectation value estimation: means of the prior would be set to values derived from the best classical methods (in this case, coupled cluster) and the prior standard deviations initialized to be larger than the typical errors for the classical method used. We have found that 0.01 is typically larger than the error between the true ground state expectation values and the classical method expectation value.

We simulate the adaptive inference process for robust amplitude estimation using Chebyshev likelihood functions as described in [14]. We vary the true expectation value and the layer fidelity as:

- Π = [0, 0.15, 0.3, 0.45, 0.6, 0.75, 0.9]
- $e^{-\lambda}$ = [0.9, 0.99, 0.999, 0.9999, 0.99999, 0.999999]

Note that $\pm\Pi$ should yield the same performance due to symmetry of the likelihood function; accordingly, we only consider $\Pi \geq 0$. Ultimately we are aiming to understand the relationship between accuracy and runtime. For the higher-fidelity trials, the change in accuracy per additional time is far greater than the lower-fidelity trials. Accordingly, we use fewer steps of Bayesian inference in the higher-fidelity trials. For each setting, we simulate between $274 - 354$ trials and track the error between the current estimate and the true value at each step of Bayesian inference.

The quantity we use to assess the performance of the estimation process is the mean squared error. However, direct use of the mean squared error does not reflect the quality of the estimators. This is due to occasional estimates far from the true value. Accordingly, we have chosen to plot an estimate of the mean-squared error of the 10th percentile of the estimates. That is, we first exclude the worst 10% of the estimates and then compute the sample MSE from the remaining estimates. In the following section, we compare these adjusted MSEs to the MSEs predicted by the runtime model.

2. Results

We present the results validating the runtime model of Equation (8) in Figure 3. The model predicts the relationship between the estimation accuracy and the accumulated runtime of the estimation. The main observation is that a majority of the data points lie within the runtime model bounds. These data points cover a wide range of layer fidelities, accuracies, and expectation values. Accordingly, we conclude that the runtime model is sufficiently accurate to be used in coarse resource estimations, such as those generated in the main text.

We discuss some of the observed discrepancies between the model and the data. In some of the settings (layer fidelity and expectation value), the simulated runtime deviates from the bounds of the runtime model for either high- or low-accuracy. The deviations for the high accuracy (i.e. small ε) regime are simpler to rationalize. The Bayesian nature of the RAE approach means that an estimate can deviate far from the true value with small probability. These bad estimates tend not to improve with further high-L samples due to the phenomenon of aliasing, and they can cause the outliers mentioned above. In cases where the runtime grows without improvements in accuracy, we expect that there has been aliasing leading to a bad estimate in the data. Assuming that this is the cause of these deviations, the likelihood of their occurrence can be exponentially suppressed by repeating the estimation procedure.

A possible explanation for the deviations in the model for the low-accuracy (large ε) cases is as follows. The analytical derivation of the runtime model bounds in [14] approximates the runtime as growing continuously as the posterior distribution variance decreases. This approximation will hold better in the large-runtime (high accuracy) regime, but will tend to underestimate the runtime in the small-runtime regime. The reason we expect the simulated runtime to be higher than the predicted runtime is that we expect the discretization in the simulated setting to lead to suboptimal choice of layer-number L during each step of the inference process. This suboptimal choice of L would cause the accuracy to not improve as much as in the idealized continuous case of the model.

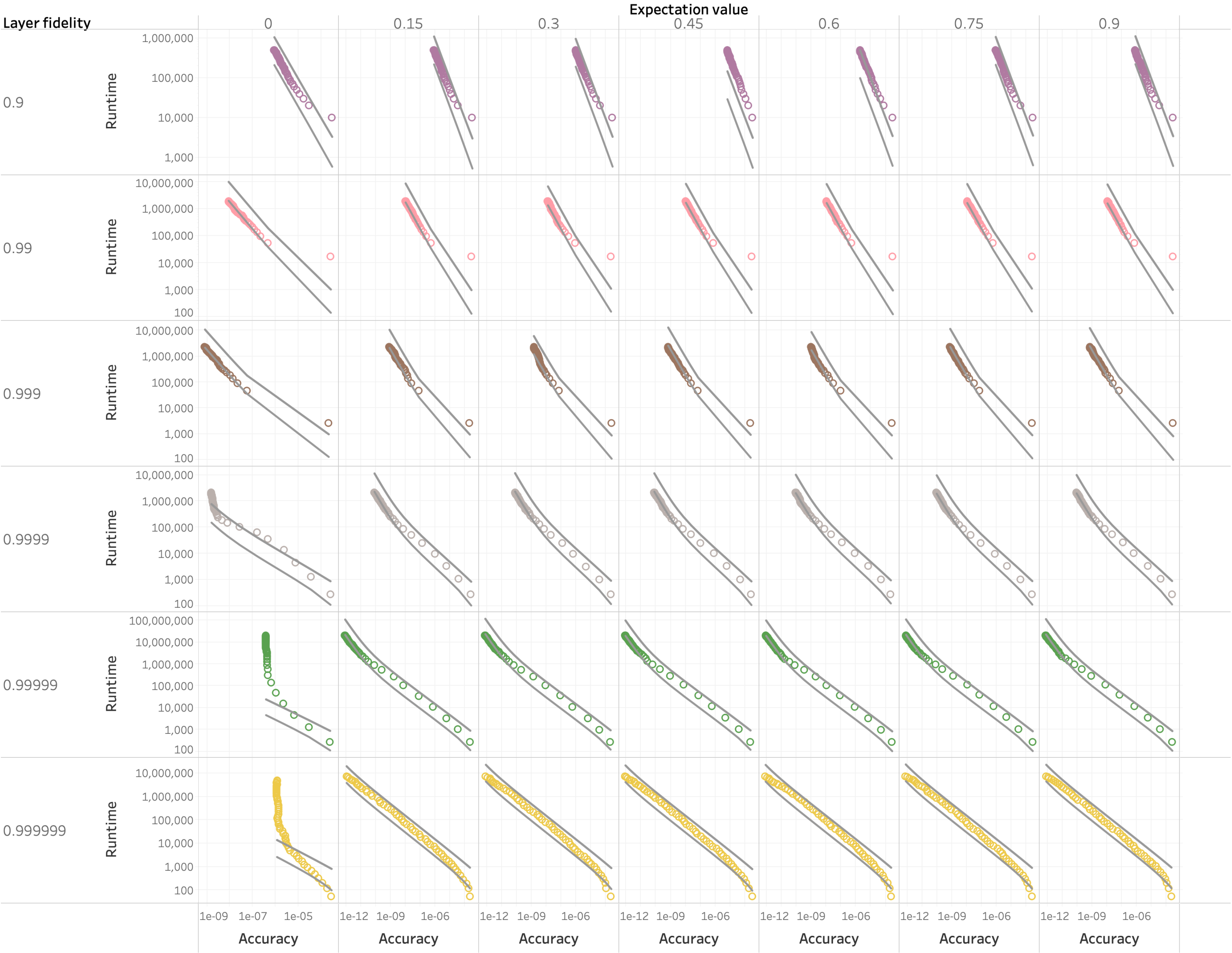

FIG. 3. This figure compares the runtime model of Eq. (8) to the estimation runtimes found in simulation. Each inset plot shows the accumulated runtime in terms of total number of circuit layers queried (computed as the average runtime of the best 90% of trials) as a function of the empirically estimated estimation accuracy (computed as the average squared error of the best 90% of trials). Each row corresponds to a different circuit layer fidelity and each column corresponds to a different true expectation value. For each inset, the upper and lower bounds of the runtime model are plotted as grey curves.

Appendix B: Quantification of circuit compilation overhead

In this section we describe the methods used to calculate the circuit depths of the operations used in RAE. The circuit depths are then used to compute the circuit fidelities as follows. We model the circuit layer fidelity f as the product of the fidelities of all two-qubit gates in the circuit. The effective two-qubit gate count is modeled as the depth times half the number of qubits. The underlying assumption is that, even if a qubit is not subject to an operation in a given layer of the circuit, the decoherence it experiences effectively translates into a loss of fidelity comparable to an imperfect two-qubit gate.

The RAE circuit is composed of a single ansatz circuit A for the state preparation, followed by L layers of RAE iterates $U = AR_0A^\dagger P$. The Pauli gates of P are single-qubit operations so we ignore their contribution to the circuit depth and thus their contribution to reducing the circuit fidelity. Furthermore, in the accounting of fault-tolerant resources, these Pauli gates contribute negligible error and negligible time relative to other gates because they are Clifford gates. Therefore, the important accounting we need is of the circuit depths for the ansatz and phase flip operations D_A and D_R, respectively. We determine these quantities for various compilation strategies that arise from the device's connectivity of the qubits. These quantities are then used to determine the required logical gate error rate $\bar{r}_g$ and the runtime of each layer τ_l. We assess the circuit costs associated to the different compilation strategies enabled by the all-to-all connectivity of ion trap devices and the limited 2D connectivity of superconducting qubit devices.

a. Ansatz circuit The ansatz circuit we use in our analysis is the hardware-efficient ansatz (HEA) described in Section III D of [1]. In the case of a linear connectivity, a single layer of this ansatz comprises two layers of nearest-neighbor two-qubit gates. We make the optimistic assumption that the number of HEA layers needed to prepare a sufficiently accurate approximation to the ground state is twice the number of qubits (or spin orbitals). In the case of a two-dimensional connectivity, we will assume that the ansatz is implemented in a “snake-like” arrangement in the order of the Jordan-Wigner encoding of the spin-orbitals. In the case of all-to-all connectivity we will assume that the device enables an ansatz compilation strategy that affords a reduction in gate count and depth by a factor of two. We believe this to be a conservative assumption because the compilation from a circuit with all-to-all connectivity into a circuit with 2D connectivity typically incurs an overhead cost that grows with the number of qubits.

b. Phase flip operation In our analysis, the most substantial variability in the compilation is in the phase flip operation $R_0 = \mathbb{I} - 2\left|0\right\rangle\left\langle 0\right|^{\otimes N}$. Devices with more connectivity will allow for a compilation of the phase flip operation which uses fewer elementary gates and has shorter depth.

Connectivity	Two-dimensional		All-to-all	
Circuit component	Ansatz	Phase flip	Ansatz	Phase flip
Two-qubit circuit depth	N	192(N-3)(N-1)	N/2	32N-96
Effective number of gates	$N^2/2$	96(N-3)(N-1)N	$N^2/4$	32(N-3)N/2

TABLE II. Summary of compilation circuit costs. Since we are interested in the decay of coherence due to quantum operations (including the idle operation), we will compute the two-qubit circuit depth and then multiply by $N/2$ to get the effective number of two-qubit gates.

The connectivity of the qubits is determined by the elementary multi-qubit gates the device can implement, which, in some cases, is determined by the physical layout of the qubits. We consider two connectivity-dependent compilations of the phase-flip operation as given in [42]. We name the compilations according to their assumed connectivity:

- Two-dimensional: a planar array of qubits connected in a square grid, similar to the qubit layout of some superconducting qubit architectures [43].
- All-to-all: any pair of qubits can be coupled via an elementary gate, which is similar to gates on ion trap devices [44].

The characteristics of these various compilation strategies are summarized in Table II.

Appendix C: Allocation of samples over Pauli terms

We now describe the method of allocating samples over the different Pauli expectation value estimates. The Hamiltonian of interest is decomposed into a linear combination of Pauli terms

$$\sum_i \mu_i P_i, \tag{C1}$$

with coefficients μ_i and Pauli observables P_i. The objective is to estimate the expected energy of a quantum state $\left|A\right\rangle = A\left|0^N\right\rangle$. The true expectation value is a linear combination of Pauli expectation values

$$\left\langle A\right| H \left|A\right\rangle = \sum_i \mu_i \left\langle A\right| P_i \left|A\right\rangle = \sum_i \mu_i \Pi_i \tag{C2}$$

where we have introduced $\Pi_i = \left\langle A\right| P_i \left|A\right\rangle$. The estimation strategy estimates the expectation value of each Pauli operator $\hat{\Pi}_i$ separately, giving the energy estimate $\hat{E}$ as

$$\hat{E} = \sum_i \mu_i \hat{\Pi}_i. \tag{C3}$$

For a fixed unbiased estimation strategy used to obtain the $\hat{\Pi}_i$, the total runtime T_i and mean squared error ε_i^2 of each estimator determine the total runtime T and total mean squared error ε^2 of the estimator $\hat{E}$,

$$T = \sum_i T_i \tag{C4}$$

$$\varepsilon^2 = \sum_i \mu_i^2 \varepsilon_i^2. \tag{C5}$$

We fix the target mean squared error (MSE) at chemical accuracy $\varepsilon^2 = \bar{\varepsilon}^2$ and aim to determine the estimation strategy for each $\hat{\Pi}_i$ which minimizes the total runtime T.

Using the single-term estimation runtime model validated in Appendix A, for each Pauli expectation estimation, we model the runtime T_i to target MSE ε_i^2 as

$$T_i = \frac{\omega}{2}\left(\frac{\lambda}{\varepsilon_i^2} + \frac{1}{\sqrt{2}\varepsilon_i} + \sqrt{\left(\frac{\lambda}{\varepsilon_i^2}\right)^2 + \left(\frac{\sqrt{8}}{\varepsilon_i}\right)^2}\right). \tag{C6}$$

With the above relationships established, we determine the optimal allocation of runtime using the following numerical optimization:

$$\begin{aligned}&\min_{\bar{\varepsilon}^2 = \sum_i \mu_i^2 \varepsilon_i^2} \sum_i T_i \\ &= \min_{\bar{\varepsilon}^2 = \sum_i \mu_i^2 \varepsilon_i^2} \sum_i \frac{\omega}{2}\left(\frac{\lambda}{\varepsilon_i^2} + \frac{1}{\sqrt{2}\varepsilon_i} + \sqrt{\left(\frac{\lambda}{\varepsilon_i^2}\right)^2 + \left(\frac{\sqrt{8}}{\varepsilon_i}\right)^2}\right),\end{aligned} \tag{C7}$$

where ω is proportional to the duration of each layer in the RAE circuit (c.f. Eq. 77 in [14]), μ_i are the coefficients in the Pauli decomposition of the Hamiltonian, and λ is the fidelity decay parameter in the noise model of the RAE likelihood function.

We introduce a Lagrange multiplier Λ to incorporate the constraint and solve for the extreme point

$$0 = \frac{d}{d\varepsilon_i}\left(\sum_i T_i + \Lambda\left(\sum_i \mu_i^2 \varepsilon_i^2 - \bar{\varepsilon}^2\right)\right) \tag{C8}$$

$$= \frac{\omega}{2}\left(\frac{-2\lambda}{\varepsilon_i^3} + \frac{-1}{\sqrt{2}\varepsilon_i^2} + \frac{d}{d\varepsilon_i}\sqrt{\left(\frac{\lambda}{\varepsilon_i^2}\right)^2 + \left(\frac{\sqrt{8}}{\varepsilon_i}\right)^2}\right) \tag{C9}$$

$$+ \Lambda \mu_i^2 \varepsilon_i$$

In order to arrive at an analytic solution for the ε_i, we approximate the hypotenuse expression above as simply the sum of the two legs of the hypotenuse, which upper bounds the contribution to the runtime via the triangle inequality. The consequence of this approximation is that the allotment of runtime to each term will be suboptimal leading to an increased overall runtime (relative to that of the optimal allotment).

$$0 = \frac{d}{d\varepsilon_i}\left(\sum_i T_i + \Lambda\left(\sum_i \mu_i^2 \varepsilon_i^2 - \bar{\varepsilon}^2\right)\right) \tag{C10}$$

$$\approx \frac{\omega}{2}\left(\frac{-2\lambda}{\varepsilon_i^3} + \frac{-1}{\sqrt{2}\varepsilon_i^2} + \frac{-2\lambda}{\varepsilon_i^3} + \frac{-\sqrt{8}}{\varepsilon_i^2}\right) + \Lambda\mu_i^2\varepsilon_i \tag{C11}$$

$$= \frac{-2\omega\lambda}{\varepsilon_i^3} + \frac{-\alpha}{\varepsilon_i^2} + \Lambda\mu_i^2\varepsilon_i, \tag{C12}$$

where $\alpha = \frac{1}{2}(\sqrt{1/2} + \sqrt{8})$. The MSEs are determined by solutions to

$$-2\omega\lambda - \alpha\varepsilon_i + \Lambda\mu_i^2\varepsilon_i^4 = 0. \tag{C13}$$

Letting $a = \alpha/\Lambda\mu_i^2$ and $b = 2\omega\lambda/\Lambda\mu_i^2$, we must solve the quartic equation $\varepsilon_i^4 = a\varepsilon_i + b$. The solutions in the shot-noise limit and Heisenberg limit extremes correspond to $a \approx 0$, giving $\varepsilon_i^2 = b^{1/2}$, and $b \approx 0$ giving $\varepsilon_i^2 = a^{2/3}$, respectively.

We approximately solve the quartic to obtain $\varepsilon_i^2 \approx b^{1/2} + a^{2/3}$. Plugging back in the relevant values gives

$$\varepsilon_i^2 = \frac{\sqrt{2\omega\lambda}}{\Lambda^{1/2}|\mu_i|} + \frac{\alpha^{2/3}}{\Lambda^{2/3}|\mu_i|^{4/3}} \tag{C14}$$

To obtain Λ we plug the above expression into the constraint equation,

$$\bar{\varepsilon}^2 = \frac{\sqrt{2\omega\lambda}}{\Lambda^{1/2}}\sum_i |\mu_i| + \frac{\alpha^{2/3}}{\Lambda^{2/3}}\sum_i |\mu_i|^{2/3}. \tag{C15}$$

Arranging into a polynomial in $\Lambda^{1/6}$, we obtain

$$(\Lambda^{1/6})^4\bar{\varepsilon}^2 = (\Lambda^{1/6})\sqrt{2\omega\lambda}\sum_i |\mu_i| + \alpha^{2/3}\sum_i |\mu_i|^{2/3}. \tag{C16}$$

We will solve this quartic polynomial numerically to ensure that the normalization is correct and that the chemical accuracy constraint is satisfied. Let Λ_* be the numerical solution. The total runtime T_* is then determined by the following procedure. First, we solve for Λ numerically. Then we evaluate each MSE according to $\varepsilon_i^2 = \frac{\sqrt{2\omega\lambda}}{\Lambda^{1/2}|\mu_i|} + \frac{\alpha^{2/3}}{\Lambda^{2/3}|\mu_i|^{4/3}}$. Finally we evaluate the overall runtime as

$$T_* = \sum_i \frac{\omega}{2}\left(\frac{\lambda}{\varepsilon_i^2} + \frac{1}{\sqrt{2}\varepsilon_i} + \sqrt{\left(\frac{\lambda}{\varepsilon_i^2}\right)^2 + \left(\frac{\sqrt{8}}{\varepsilon_i}\right)^2}\right). \tag{C17}$$

Appendix D: Model of fault-tolerant quantum computation overhead

In order to sufficiently reduce the estimation runtimes, we must sufficiently reduce the error rates of the quantum operations. The scalable approach to reducing such error rates is achieved with quantum error correction. While quantum error correction suppresses the error rates of gates and measurements, it incurs an additional cost in terms of physical qubits and processing time. We will analyze these costs of quantum error correction assuming a state-of-the-art implementation of the surface code [45]. Following the analysis in [46], we consider a quantum computer architecture which runs the surface code with physical gate error rates of 10^{-3}.

By increasing the code distance d the logical error rates are reduced as

$$\varepsilon = 10^{-(d+3)/2}. \tag{D1}$$

To protect each operation with a distance-d code requires $N = 2d^2$ physical qubits. Following the assumptions of [46] (that the synthesis of each gate will require, on average $100d$ surface code cycles), then the fidelity of each gate will be

$$f = (1 - 10^{-(d+3)/2})^{100d} \approx 1 - d10^{-(d-1)/2}. \tag{D2}$$

The gates which are enumerated in this model include arbitrary-angle single-qubit rotations as well as all two-qubit gates. Optimistic estimates [45] give surface cycle times of 1μs. Thus, the time needed to implement the above logical gate is $100d\mu$s.

We incorporate these quantum error correction overheads into our runtime estimates by assuming that each single layer of logical gates requires a runtime of $100d\mu$s. Then, both the gate layer runtime and logical gate fidelity are determined by the code distance d.

[1] J. F. Gonthier, M. D. Radin, C. Buda, E. J. Doskocil, C. M. Abuan, and J. Romero, Identifying challenges towards practical quantum advantage through resource estimation: the measurement roadblock in the variational quantum eigensolver, arXiv preprint arXiv:2012.04001 (2020).

[2] M. Reiher, N. Wiebe, K. M. Svore, D. Wecker, and M. Troyer, Elucidating reaction mechanisms on quantum computers, Proceedings of the National Academy of Sciences **114**, 7555 (2017).

[3] T. E. O'Brien, M. Streif, N. C. Rubin, R. Santagati, Y. Su, W. J. Huggins, J. J. Goings, N. Moll, E. Kyoseva, M. Degroote, *et al.*, Efficient quantum computation of molecular forces and other energy gradients, arXiv preprint arXiv:2111.12437 (2021).

[4] J. Alcazar, A. Cadarso, A. Katabarwa, M. Mauri, B. Peropadre, G. Wang, and Y. Cao, Quantum algorithm for credit valuation adjustments, arXiv preprint arXiv:2105.12087 (2021).

[5] N. Stamatopoulos, G. Mazzola, S. Woerner, and W. J. Zeng, Towards quantum advantage in financial market risk using quantum gradient algorithms, arXiv preprint arXiv:2111.12509 (2021).

[6] S. Stanisic, J. L. Bosse, F. M. Gambetta, R. A. Santos, W. Mruczkiewicz, T. E. O'Brien, E. Ostby, and A. Montanaro, Observing ground-state properties of the fermi-hubbard model using a scalable algorithm on a quantum computer, arXiv preprint arXiv:2112.02025 (2021).

[7] J. J. Goings, A. White, J. Lee, C. S. Tautermann, M. Degroote, C. Gidney, T. Shiozaki, R. Babbush, and N. C. Rubin, Reliably assessing the electronic structure of cytochrome p450 on today's classical computers and tomorrow's quantum computers, arXiv preprint arXiv:2202.01244 (2022).

[8] F. Arute, K. Arya, R. Babbush, D. Bacon, J. C. Bardin, R. Barends, R. Biswas, S. Boixo, F. G. Brandao, D. A. Buell, *et al.*, Quantum supremacy using a programmable superconducting processor, Nature **574**, 505 (2019).

[9] A. Kandala, K. Temme, A. D. Córcoles, A. Mezzacapo, J. M. Chow, and J. M. Gambetta, Error mitigation extends the computational reach of a noisy quantum processor, Nature **567**, 491 (2019).

[10] L. Egan, D. M. Debroy, C. Noel, A. Risinger, D. Zhu, D. Biswas, M. Newman, M. Li, K. R. Brown, M. Cetina, *et al.*, Fault-tolerant control of an error-corrected qubit, Nature **598**, 281 (2021).

[11] M. Abobeih, Y. Wang, J. Randall, S. Loenen, C. Bradley, M. Markham, D. Twitchen, B. Terhal, and T. Taminiau, Fault-tolerant operation of a logical qubit in a diamond quantum processor, arXiv preprint arXiv:2108.01646 (2021).

[12] C. Ryan-Anderson, J. Bohnet, K. Lee, D. Gresh, A. Hankin, J. Gaebler, D. Francois, A. Chernoguzov, D. Lucchetti, N. Brown, *et al.*, Realization of real-time fault-tolerant quantum error correction, Physical Review X **11**, 041058 (2021).

[13] A. Peruzzo, J. McClean, P. Shadbolt, M.-H. Yung, X.-Q. Zhou, P. J. Love, A. Aspuru-Guzik, and J. L. O′brien, A variational eigenvalue solver on a photonic quantum processor, Nat. Commun. **5**, 4213 (2014).

[14] G. Wang, D. E. Koh, P. D. Johnson, and Y. Cao, Minimizing estimation runtime on noisy quantum computers, P R X Quantum **2**, 010346.

[15] A. Aspuru-Guzik, A. D. Dutoi, P. J. Love, and M. Head-Gordon, Simulated quantum computation of molecular energies, Science **309**, 1704 (2005).

[16] D. Wecker, M. B. Hastings, and M. Troyer, Progress towards practical quantum variational algorithms, Phys. Rev. A **92**, 10.1103/physreva.92.042303 (2015).

[17] V. von Burg, G. H. Low, T. Häner, D. S. Steiger, M. Reiher, M. Roetteler, and M. Troyer, Quantum computing enhanced computational catalysis, arXiv:2007.14460v1.

[18] V. E. Elfving, B. W. Broer, M. Webber, J. Gavartin, M. D. Halls, K. P. Lorton, and A. Bochevarov, How will quantum computers provide an industrially relevant computational advantage in quantum chemistry?, arXiv:2009.12472v1.

[19] C. Cade, L. Mineh, A. Montanaro, and S. Stanisic, Strategies for solving the Fermi-Hubbard model on near-term quantum computers, Physical Review B **102**, 235122 (2020), arXiv:1912.06007.

[20] A. Kandala, A. Mezzacapo, K. Temme, M. Takita, M. Brink, J. M. Chow, and J. M. Gambetta, Hardware-

efficient variational quantum eigensolver for small molecules and quantum magnets, Nature **549**, 242 (2017).

[21] J. R. Fontalvo, R. Babbush, J. McClean, C. Hempel, P. J. Love, and A. Aspuru-Guzik, Strategies for quantum computing molecular energies using the unitary coupled cluster ansatz, Quantum Sci. Technol. **4**, 14008 (2018).

[22] D. E. Koh, G. Wang, P. D. Johnson, and Y. Cao, A framework for engineering quantum likelihood functions for expectation estimation, arXiv:2006.09349v1.

[23] T. Giurgica-Tiron, S. Johri, I. Kerenidis, J. Nguyen, N. Pisenti, A. Prakash, K. Sosnova, K. Wright, and W. Zeng, Low depth amplitude estimation on a trapped ion quantum computer, arXiv preprint arXiv:2109.09685 (2021).

[24] 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).

[25] N. Wiebe and C. Granade, Efficient bayesian phase estimation, Physical review letters **117**, 010503 (2016).

[26] T. Tanaka, Y. Suzuki, S. Uno, R. Raymond, T. Onodera, and N. Yamamoto, Amplitude estimation via maximum likelihood on noisy quantum computer, arXiv preprint arXiv:2006.16223 (2020).

[27] R. Tipireddy and N. Wiebe, Bayesian phase estimation with adaptive grid refinement, arXiv preprint arXiv:2009.07898 (2020).

[28] R. Royall, On the probability of observing misleading statistical evidence, Journal of the american statistical association **95**, 760 (2000).

[29] A. Mari, N. Shammah, and W. J. Zeng, Extending quantum probabilistic error cancellation by noise scaling, Physical Review A **104**, 052607 (2021).

[30] J. M. Baker, C. Duckering, A. Hoover, and F. T. Chong, Decomposing quantum generalized toffoli with an arbitrary number of ancilla, arXiv preprint arXiv:1904.01671 (2019).

[31] J. S. Kottmann and A. Aspuru-Guzik, Optimized low-depth quantum circuits for molecular electronic structure using a separable pair approximation, arXiv preprint arXiv:2105.03836 (2021).

[32] I. H. Kim, E. Lee, Y.-H. Liu, S. Pallister, W. Pol, and S. Roberts, Fault-tolerant resource estimate for quantum chemical simulations: Case study on li-ion battery electrolyte molecules, arXiv preprint arXiv:2104.10653 (2021).

[33] P. Panteleev and G. Kalachev, Asymptotically good quantum and locally testable classical ldpc codes, arXiv preprint arXiv:2111.03654 (2021).

[34] M. A. Tremblay, N. Delfosse, and M. E. Beverland, Constant-overhead quantum error correction with thin planar connectivity, arXiv preprint arXiv:2109.14609 (2021).

[35] L. Z. Cohen, I. H. Kim, S. D. Bartlett, and B. J. Brown, Low-overhead fault-tolerant quantum computing using long-range connectivity, arXiv preprint arXiv:2110.10794 (2021).

[36] P.-L. Dallaire-Demers, J. Romero, L. Veis, S. Sim, and A. Aspuru-Guzik, Low-depth circuit ansatz for preparing correlated fermionic states on a quantum computer, Quantum Sci. Technol. **4**, 045005 (2019).

[37] S. Sim, J. Romero, J. F. Gonthier, and A. A. Kunitsa, Adaptive pruning-based optimization of parameterized quantum circuits, Quantum Science and Technology **6**, 025019 (2021).

[38] G. Wang, S. Sim, and P. D. Johnson, State preparation boosters for early fault-tolerant quantum computation, arXiv preprint arXiv:2202.06978 (2022).

[39] L. Lin and Y. Tong, Heisenberg-limited ground-state energy estimation for early fault-tolerant quantum computers, PRX Quantum **3**, 010318 (2022).

[40] R. Zhang, G. Wang, and P. Johnson, Computing ground state properties with early fault-tolerant quantum computers, arXiv preprint arXiv:2109.13957 (2021).

[41] K. Wan, M. Berta, and E. T. Campbell, A randomized quantum algorithm for statistical phase estimation, arXiv preprint arXiv:2110.12071 (2021).

[42] A. Holmes, S. Johri, G. G. Guerreschi, J. S. Clarke, and A. Y. Matsuura, Impact of qubit connectivity on quantum algorithm performance, Quantum Science and Technology **5**, 025009 (2020).

[43] G. Q. AI, Exponential suppression of bit or phase errors with cyclic error correction, Nature **595**, 383 (2021).

[44] S. Debnath, N. M. Linke, C. Figgatt, K. A. Landsman, K. Wright, and C. Monroe, Demonstration of a small programmable quantum computer with atomic qubits, Nature **536**, 63 (2016).

[45] D. S. Wang, A. G. Fowler, A. M. Stephens, and L. C. L. Hollenberg, Threshold error rates for the toric and surface codes, arXiv preprint arXiv:0905.0531 (2009).

[46] W. J. Huggins, S. McArdle, T. E. O'Brien, J. Lee, N. C. Rubin, S. Boixo, K. B. Whaley, R. Babbush, and J. R. McClean, Virtual distillation for quantum error mitigation, arXiv preprint arXiv:2011.07064 (2020).