

# Surrogate Hessian Accelerated Structural Optimization for Stochastic Electronic Structure Theories

Juha Tiihonen,<sup>1</sup> Paul R. C. Kent,<sup>2</sup> and Jaron T. Krogel<sup>1</sup>

<sup>1</sup>*Materials Science and Technology Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831, USA*

<sup>2</sup>*Computational Sciences and Engineering Division, Oak Ridge, TN 37831, USA*

(Dated: 18 April 2022)

We present an efficient energy-based method for structural optimization with stochastic electronic structure theories, such as diffusion quantum Monte Carlo (DMC). The method is based on robust line-search energy minimization in reduced parameter space, exploiting approximate but accurate Hessian information from a surrogate theory, such as density functional theory. The surrogate theory is also used to characterize the potential energy surface, allowing simple but reliable ways to maximize statistical efficiency while retaining controllable accuracy. We demonstrate the method by finding the minimum DMC energy structures of selected flake-like aromatic molecules, benzene, coronene and ovalene, represented by 2, 6 and 19 structural parameters, respectively. In each case, the energy minimum is found within 2 parallel line-search iterations. The method is near-optimal for a line search technique and suitable for a broad range of applications. It is easily generalized to any electronic structure method where forces and stresses are still under active development and implementation, such as diffusion Monte Carlo, auxiliary-field Monte Carlo, stochastic configuration interaction, as well as deterministic approaches such as the random-phase approximation. Accurate and efficient means of geometry optimization could shed light into a broad class of materials and molecules showing high sensitivity of induced properties to structural variables.

<sup>a</sup>

## I. INTRODUCTION

Finding energy-minimizing geometries is an essential part of predictive electronic structure simulation. The atomic geometry, that is, a set of ionic coordinates and lattice constants, affects in some capacity all measurable properties of the system. This introduces a foundational conflict: predictive *ab initio* simulation is only fully consistent when done at an energy-minimizing geometry.

Most recent electronic structure predictions rely on geometries predicted from the density functional theory (DFT), which often match experiments to 0.01-0.03 Å for molecules<sup>1</sup> and transition metal complexes<sup>2</sup>, 0.03-0.1 Å for lattice constants of bulk solids<sup>3</sup> and 0.2-0.3 Å for lattice constants of layered materials<sup>4</sup>. However, higher accuracies may be desired or required. For example, in the area of 2D materials, such as van der Waals heterostructures, the properties are highly tunable by structural effects<sup>5</sup>. In some materials, structural sensitivity due to strong electron correlation or dispersive effects can result in substantial variations in not only the predicted geometries but also the resultant band struc-

tures and electronic properties<sup>6-8</sup>. The consistent use of higher accuracy methods for both the geometries and the electronic properties of interest is therefore required in these cases. These calculations need to be performed efficiently because the methods are more computationally expensive, and energy gradients may not be available or have distinct numerical properties that must be accommodated.

In this work we concentrate on quantum Monte Carlo (QMC) methods, though the techniques developed here are applicable to other stochastic electronic structure methods without modification. They can also be applied to deterministic methods should an implementation of energy gradients not be available. QMC accurately describes many-body energy surfaces, and thus, has the potential to inform chemical and materials problems involving high structural sensitivity<sup>6</sup>. The typical approach for energy minimization is the use of energy gradients, namely forces and stresses. Indeed, forces computed from variational Monte Carlo (VMC) have been successfully applied to relax numerous molecules<sup>9-18</sup>, to study, for instance, bond-length alternation in polyacetylene chains<sup>17</sup> with accuracies rivaling or exceeding high-level quantum chemistry. However, while forces from the mean-field methods, such as DFT, face little trouble in terms of cost, implementation, or consistency with the potential energy surface (PES), forces in methods like QMC are more challenging. The estimated mean of the QMC force does not obey the central limit theorem<sup>19</sup>, and thus sophisticated estimators (e.g. Refs. 20-26) are needed to control accuracy and statistical uncertainty. Thus, in QMC the utility of energy-based structural optimization methods can be much higher than for other

---

<sup>a</sup>This manuscript has been authored in part by UT-Battelle, LLC, under contract DE-AC05-00OR22725 with the US Department of Energy (DOE). The publisher, by accepting the article for publication, acknowledges that the US government retains a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this manuscript, or allow others to do so, for US government purposes. DOE will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan (<http://energy.gov/downloads/doe-public-access-plan>).

methods, though in practice substantial optimizations are still needed to make them practical.

A central challenge of directly using QMC energies for structural relaxation is stochasticity: QMC energies come with finite statistical uncertainty and a high cost to reduce the uncertainty. Therefore, the total number of energy evaluations used in the relaxation needs to be low, and the meaningful scope of problems is limited by the complexity of the PES. This was first addressed by Schuetz *et al.*<sup>27</sup>, who optimized a 3-parameter formaldehyde molecule with solution mapping. They used factorial design to construct a data set of 15 energies, which could be fitted to produce the equilibrium geometry in a single iteration with reduced uncertainty. However, this approach would have factorial scaling with the number of parameters. A similar multi-dimensional PES fitting approach was presented in Ref. 28, where a 2-parameter model of water was relaxed with force and energy-based VMC methods. That work and Ref. 29 had particular focus on different fitting grids, their effects on the bias and propagated uncertainty, and means to correct the bias to some degree. Another approach in Ref. 30 optimized up to 9 structural parameters in molecules by performing sequential line minimizations in special directions in the parameter space that were estimated by iteratively fitting a multidimensional quadratic model to the PES. Constructing the Hessian directly from energy information requires many samples, though this was partially mitigated by using Bayesian inference. Work in Ref. 31 relaxed the assumption of a purely quadratic energy surface by using Gaussian process to sample the multi-dimensional PES of a 22 parameter model of benzene statistically, while still requiring many energy samples (quadratic in the number of structural parameters) to resolve the minimum.

In this work, we outline and demonstrate a scheme for energy-based structural optimization that is highly efficient. The method we propose is in the family of conjugate direction methods<sup>32</sup>, where coordinate descent<sup>33,34</sup> (or line search) is used along a set of conjugate directions obtained from an approximate Hessian matrix describing the PES near its minimum. The method differs from previous approaches<sup>30,35</sup> in that the approximate Hessian is taken from a cheaper surrogate theory (density functional theory (DFT) in this case), instead of being constructed expensively from the direct high quality PES. When the surrogate Hessian is sufficiently close to the true Hessian, a rapid convergence to the minimum energy structure is possible without requiring additional information. Additionally, we exploit the independence of the conjugate directions to resolve the descent along all directions simultaneously, which supports parallel evaluation of energies as is well suited for stochastic methods in high performance computing environments. In practice, we find that only a few parallel line-search iterations are needed to relax the structure, independent of parametric complexity. Furthermore, we use the surrogate theory to characterize the propagation of statistical and systematic

errors in the method. This enables us to impose accuracy tolerances on the stochastically estimated structural parameters while explicitly minimizing the computational cost required to obtain them.

A preliminary version of our algorithm was successfully applied to quasi-2D germanium selenide in Ref. 36. In that work, no attempt was made to control systematic or stochastic errors and only the basic concept of a surrogate Hessian was used. Here, we fill out a complete theory for the structural optimization method, including novel techniques to control systematic errors in the relaxed geometry and also to explicitly minimize computational cost, which are critical components of a general method for structural relaxation in the presence of a stochastic PES.

The remainder of this work is organized as follows: In Section II, we lay out the main aspects of the method, both in terms of theory and practice. We begin from the origin of the Hessian and conjugate directions in reduced parameter space, and how to obtain the surrogate Hessian. We then discuss practical ways to consider and optimize individual line search grids, to control parameter errors and crucially to minimize computational cost. In Section III we demonstrate the method with DMC energy minimization of three flake-like molecules – benzene ( $C_6H_6$ ), coronene ( $C_{24}H_{12}$ ), and ovalene ( $C_{32}H_{14}$ ) – with up to 19 symmetry inequivalent parameters and discuss the accuracy and efficiency of the relaxation process. We conclude in Sec. IV with a summary and outlook regarding the method. A complete implementation is made available at [https://github.com/QMCPACK/surrogate\\_hessian\\_relax](https://github.com/QMCPACK/surrogate_hessian_relax).

## II. THE SURROGATE HESSIAN ACCELERATED PARALLEL LINE SEARCH METHOD

A key ingredient to minimize the cost of a conjugate direction method is the acquisition of a sufficiently accurate approximation to the true Hessian,  $H$ , of the PES in order to find efficient search directions. Stochastic approximations to  $H$  may be costly especially in the absence of PES gradients. Density functional theory (DFT) with approximate functionals is in practice quite accurate in a broad array of electronic structure problems, yielding structural parameters within a few percent of those obtained with higher order methods. A DFT PES may therefore be generally expected to provide a high quality approximation to the true Hessian and near optimal search directions at low computational cost.

We use surrogate information from DFT in three important ways to enhance the optimization process. First, the DFT optimal structure is used as the starting point of the stochastic optimization process. Given the overall accuracy of DFT, this greatly increases the chance that the initial structure lies within or very near the quadratic region of the true PES and not far overall from the true minimum. The proximity to the mini-

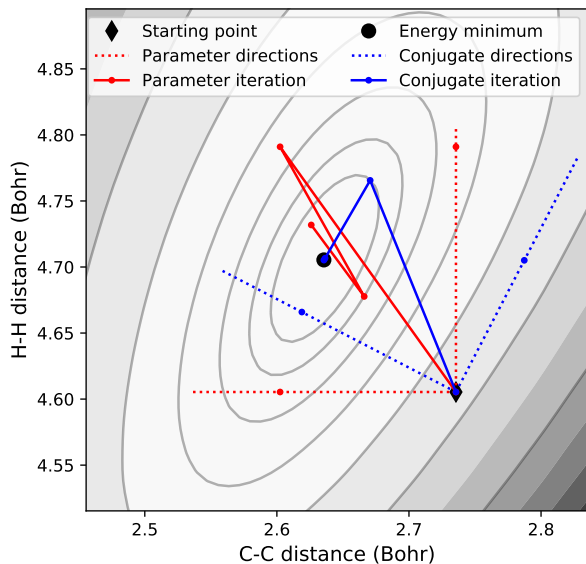


FIG. 1. Example parallel line-search iteration using optimal (blue) and unguided (red) search directions in the DFT PES of 2 benzene parameters: C-C distance and H-H distance. Starting from a displaced point, the directions and the first search results are respectively presented with dotted lines and small dots to mark the found energy minima. After the first iteration the optimal search is in the quadratic region and immediately after finds the target, whereas the unguided search converges in a slow zigzag pattern.

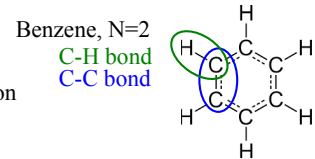
imum of the true PES reduces the number of iterations required in the stochastic optimization process. Second, the search directions are obtained from the DFT Hessian. The high quality of these directions combined with the near quadratic behavior of the PES at the DFT initial structure also contributes to reducing the required number of iterations. In practice, we find that two iterations are often sufficient to reach the optimal structure within reasonable target accuracy, as illustrated in Fig. 1 on the DFT PES of benzene (see Sec. III B for details). Finally, the DFT force constants – and also full anharmonic information – along each approximate conjugate direction are used to estimate and control for statistical and systematic errors. Each direction of the PES is treated in a statistically separate way that is optimally tuned for mode stiffness, with fits along softer modes generally requiring greater statistical sampling and stiff modes requiring and receiving a lower degree of sampling.

The high level algorithm may be summarized as follows: (1) Find the optimal structure of the surrogate theory to use as a starting point for the line search (2) Obtain the Hessian of the surrogate theory in the space of the structural parameters and calculate the surrogate conjugate directions, (3) Estimate the optimal line search grids and statistical sampling (low sampling, high noise limit) of directional fits that remain within the target parameter accuracy, (4) Perform the line search of the

## Surrogate theory: DFT

### 1. Relaxation

parameter couplings  
from phonon calculation

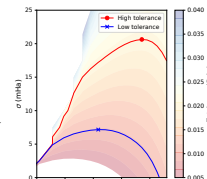


### 2. Parameter Hessian

eigenvectors to  
conjugate directions

### 3. Line-search optimization

- 1) use resampling to predict bias and noise in the line-search
- 2) optimize the line-search to allow maximum noise while meeting an accuracy target



## Stochastic theory: DMC

### 4. Line-search

sample energies  
along all directions

repeat until  
converged

### 5. Find new minimum

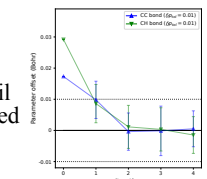


FIG. 2. Overview of the algorithm with illustrations based on the benzene molecule.

stochastic PES along all surrogate conjugate directions in parallel, (5) Obtain the best statistical estimate of the optimal structure by fitting a low order polynomial along each search direction and repeat from (4) until convergence is reached. This overview is also illustrated in Fig. 2

Details of the line search algorithm (step 4 above) are found in section II A. Obtaining the surrogate Hessian (step 2 above) is discussed in more detail in section II B. Methodological developments to enable stochastically efficient structural optimization while controlling for parameter bias (step 3 above) are the subject of sections II C and II D.

### A. Parallel line search

The problem class targeted by our proposed technique is the structural optimization of solid state and molecular systems in the context of stochastic electronic structure methods in the absence of gradients. Thus, our task is to find the minimum of the potential energy surface (PES)

$$R_0 = \arg \min_R E(R) \quad (1)$$

where the coordinates  $R$  can in general contain contributions from both the ionic positions  $R_I$  and the supercell axes  $A$ , if present, so that  $R = [A, R_I]$ . The local PES itself is the mean of a statistical distribution

$$\bar{E}(R) = \int dE P(R, E) E. \quad (2)$$

For convenience, and to apply symmetries, we may consider the molecular/crystal structure to be mapped out by a set of  $N$  parameters without a loss of generality, or

$$\bar{E}(R) = \bar{E}(R(p)) \quad (3)$$

and we refer to the PES in the parameter space simply as  $E(p)$ .

In this context, finite statistical sampling results in a view of the PES that is non-smooth, and so any successful structural optimization method must be robust to statistical noise. Further, it is desirable to maximize the noise tolerated by the method since this reduces the computational cost of the approach. For a similar reason, the method should minimize the total number of energy evaluations needed to find the optimal structure.

Near the energy minimum, the PES may be expanded to second order as

$$E(p) = E_0 + \frac{1}{2}(p - p_0)^T H_p (p - p_0) \quad (4)$$

where the  $N \times N$  matrix  $H_p$  is the force constant matrix, or energy Hessian, in the space of the parameters. The Hessian may be diagonalized to obtain a set of directions that are conjugate to the energy isosurfaces of  $E(p)$  near the minimum. The diagonalization is a rotation into a space of decoupled modes

$$\begin{aligned} E(p) &= E_0 + \frac{1}{2}(p - p_0)^T U^T \Lambda U (p - p_0) \\ &= E_0 + \frac{1}{2}x^T \Lambda x \\ &= E_0 + \frac{1}{2} \sum_{n=1}^N \lambda_n x_n^2 \end{aligned} \quad (5)$$

The  $N$  conjugate directions,  $d_n$ , are given by the columns of  $U^T \equiv D$ , or  $d_n = U_n^T$ .

In the event that the local PES is precisely quadratic in  $p$ –or is even a direct composition of a quadratic function–and  $H$  is exact, the conjugate directions form an optimal basis for line search (coordinate descent). That is, starting from a point  $\bar{p}$  away from  $p_0$ , a set of  $N$  lines may be traced out along each search direction via the independent conjugate parameter ( $x_n$ )

$$p(x_n) = \bar{p}_n + d_n x_n \quad (6)$$

By finding the minimum along each line separately

$$x_{0n} = \arg \min_{x_n} E(\bar{p}_n + d_n x_n) \quad (7)$$

the full minimum is found directly as

$$p_0 = \bar{p} + D x_0 \quad (8)$$

and so the line search converges in a single step<sup>35</sup> when all directions are considered in parallel. Line search/coordinate descent converges regardless of the

(linearly independent) directions employed and also independent of assumptions of the local shape of the PES, but in general more than one iteration is required.

Deviations from quadratic behavior are accommodated in the current method by sampling a regular grid of  $M$  points ( $M \leq 7$ ) along each search direction and then fitting a low order polynomial to that slice of the PES to find a good approximation to the minimum along that direction. This approach is beneficial in the case of a stochastic PES, as the statistical sampling from all points along the grid contribute in a way that can reduce the variance of each respective estimated conjugate parameter value. The total number of energy evaluations required at each iteration is  $M \times N$ , or linear in the number of structural parameters  $N$  since  $M$  is fixed to a low integer.

In the current method, the line search along all search directions is performed simultaneously rather than sequentially. This parallel line search is advantageous, since simulation of large electronic structure problems with stochastic methods such as QMC are often performed at large scale computing facilities which incur large time delays if computational jobs must be run sequentially.

## B. Obtaining the Surrogate Hessian

Optimal line search requires knowledge of the structural parameter Hessian  $H_p$ , to obtain the ideal directions  $d_m$ . The principle of the surrogate Hessian accelerated line-search is to use a low-cost but moderate accuracy Hessian from a surrogate theory ( $H_p^S \approx H_p$ ) for guiding line search along the more accurate and expensive stochastic potential energy surface, where the exact Hessian is unavailable.

The structure is defined by a set of generalized coordinates,  $R$ . For molecules, these are simply the locations of the ionic centers, or  $R = R_I$ . For systems involving partial or full periodicity, the generalized coordinates may be extended to include the axis vectors of the periodic cell, or  $R = [A, R_I]$ . For any given parameterization of the structure, we represent the mapping between the structural parameters and the generalized coordinates  $R$  as

$$R = R_0^S + f(p, p_0^S) \quad (9)$$

where  $R_0^S$  is the optimal structure from the surrogate theory and  $p_0^S$  constitutes the parameterization for the optimal surrogate structure and  $f$  denotes a mapping from parameter displacements to the generalized coordinates, such that  $f(p_0^S, p_0^S) = 0$ .

The surrogate Hessian in parameter space is formally given by

$$H_p^S = \nabla_p \nabla_p^T E^S|_{p=p_0^S} \quad (10)$$

If there are only a few structural parameters of interest, then a finite difference approach may be preferred to

obtain the Hessian as

$$H_p^S(i, j) \approx \frac{E_{\Delta p_i, \Delta p_j}^S - E_{\Delta p_i, -\Delta p_j}^S - E_{-\Delta p_i, \Delta p_j}^S + E_{-\Delta p_i, -\Delta p_j}^S}{4\Delta p_i \Delta p_j} \quad (11)$$

where we have used

$$E_{\Delta p_i, \Delta p_j}^S \equiv E^S(R_0^S + f(p_0^S + \Delta p_i + \Delta p_j, p_0^S)) \quad (12)$$

For a larger number of parameters, obtaining the Hessian in the full generalized space and then mapping to the parametric space may be beneficial. To realize this approach, we first expand the surrogate PES to second order about its minimum and substitute in the structural parameterization, which gives

$$\begin{aligned} E^S &\approx E_0^S + \frac{1}{2}(R - R_0^S)^T H_R^S (R - R_0^S) \\ &= E_0^S + \frac{1}{2}f(p, p_0^S)^T H_R^S f(p, p_0^S) \end{aligned} \quad (13)$$

where  $H_R^S$  represents the surrogate Hessian in the full space of the generalized coordinates. The elements of the parametric Hessian  $H_p^S$  may then be related to the full space Hessian  $H_R^S$  as

$$\begin{aligned} H_p^S(i, j) &= \partial_{p_i} \partial_{p_j} E^S|_{p=p_0^S} \\ &= \partial_{p_i} f^T|_{p=p_0^S} H_R^S \partial_{p_j} f|_{p=p_0^S} \\ &\quad + \partial_{p_i} \partial_{p_j} f^T|_{p=p_0^S} H_R^S f(p_0^S, p_0^S) \\ &= \partial_{p_i} f^T|_{p=p_0^S} H_R^S \partial_{p_j} f|_{p=p_0^S} \end{aligned} \quad (14)$$

where we have made use of the fact that  $f(p_0^S, p_0^S) = 0$  as well as the hermiticity of the Hessian matrices.

When only ionic coordinates enter the parameterization, the full space Hessian is just the dynamical matrix of the surrogate theory, or  $H_R^S = D_{R_I}^S$ , which may be obtained from a single molecular vibration or phonon calculation. For periodic systems with a partial or full parameterization of the cell degrees of freedom, additional work must be done to obtain the full space Hessian. In this case, the full space Hessian can be expressed in block form with respect to the cell and ionic coordinates as

$$H_R^S = \begin{bmatrix} -S_A \nabla_A \sigma_A^S & -\nabla_A F_{R_I}^S \\ -\nabla_A F_{R_I}^{S^T} & D_{R_I}^S \end{bmatrix} \quad (15)$$

Here,  $F_{R_I}^S$  and  $\sigma_A^S$  are the ionic forces and cell stresses in the surrogate theory and  $S_A$  is a vector containing the surface areas of the cell faces. In practice, the cell gradients ( $\nabla_A \cdot$ ) appearing in the above expressions are conveniently evaluated with finite differences since both the ionic forces and cell stress are readily available from DFT.

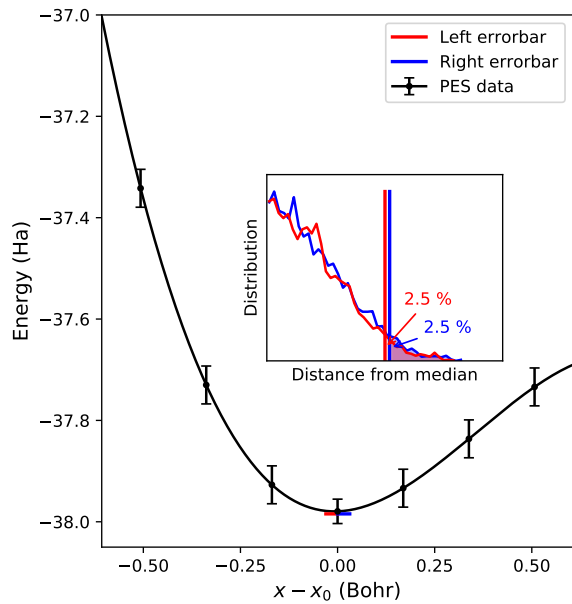


FIG. 3. Stochastic effects in finding the PES minimum are simulated by randomly resampling the fitting procedure on the actual surrogate PES. The solid black line is a cubic polynomial fit to  $M = 7$  grid points based on DFT simulations along one of the conjugate directions of the benzene molecule. The black vertical errorbars refer to the uniform uncertainties  $\sigma$ , which generate horizontal errorbars (red/blue for left/right tails of the distribution) for positioning the minimum. Both of the errorbars in the figure are magnified for visual clarity. The inset illustrates that  $\geq 95\%$  statistical confidence is ensured by choosing the larger out of the left/right horizontal errorbars, each of which contains 47.5% of the nearest samples. The distribution in the inset is a histogram based on 1000 resamplings of the fit.

### C. Controlling bias at maximum efficiency for individual search directions

An important component of the parallel line search method is the minimization of the energy along an individual search direction. It is desirable to achieve this within a specified tolerance on the error in the conjugate parameter. It is also desirable to do so in a way that minimizes computational cost. The cost in this case is driven solely by the degree of statistical sampling required in QMC, and so cost minimization is equivalent to maximizing the statistical noise that can be tolerated along each line search. Within a few simplifying choices, we describe below how this can be effectively achieved by relying on information obtained from the cheaper surrogate theory.

Here, we consider the task of estimating the PES minimum along a single conjugate direction  $d_n$ . For simplicity in discussion, the subscript  $n$  has been suppressed on all relevant quantities such as the discrete search points and corresponding energy values. Given a set of  $M$  points  $\{x_m\}_{m=1}^M$  distributed along the search direction  $d_n$  with

corresponding mean energies  $\{\bar{E}_m\}_{m=1}^M$  and mean statistical errors  $\{\sigma_m\}_{m=1}^M$ , the minimum  $x_0$  may be estimated by fitting a parametric curve  $g(a, x)$  to the data, or

$$a_0 = \arg \min_a \sum_{m=1}^M (g(a, x_m) - \bar{E}_m)^2 \quad (16)$$

$$x_0 = \arg \min_x g(a_0, x) \quad (17)$$

where  $a$  denotes the set of parameters in the fitted curve. We denote this process by a non-linear operation  $G$ , such that  $x_0 = G(E_1, \dots, E_M) \equiv G(\{E_m\})$ . If we denote the location of the exact minimum as  $\tilde{x}_0$  (i.e. in the limit of zero noise), then the error in the estimated minimum,  $x_0 - \tilde{x}_0$ , obeys a probability distribution  $P(x_0 - \tilde{x}_0)$  that may be represented formally as

$$P(x_0 - \tilde{x}_0) = \int d\eta_1 \dots d\eta_M \delta(x_0 - \tilde{x}_0 - G(\{\bar{E}_m + \sigma_m \eta_m\})) \quad (18)$$

where each  $\eta_m$  is a normally distributed stochastic variable with zero mean and unit variance. The distribution of  $P$  is not necessarily normal, but it has a finite variance. Estimates of the location of the minimum are relatively insensitive to the choice of  $M$ , because the effect of choosing high or low  $M$  values is cancelled out by using lower or higher numbers of samples per point, respectively. Low, odd numbers are convenient, because they require less overhead in preparing the trial wavefunction, and they allow a common point between all search directions. Unless otherwise stated, we shall use  $M = 7$  for the remainder of this work.

Given a tolerance on the error in the conjugate parameter,  $\delta x_{tol}$ , we demand that the error in the estimated minimum lie within the tolerance with a high likelihood. This condition may be represented formally by requiring that the left and right percentiles encompassing 95% likelihood both fall below the tolerance in absolute value, or

$$\max(|P_{2.5}|, |P_{97.5}|) \leq \delta x_{tol} \quad (19)$$

where  $P_{2.5}$  and  $P_{97.5}$  represent the 2.5% and 97.5% percentiles of the distribution  $P(x_0 - \tilde{x}_0)$ .

The error distribution  $P(x_0 - \tilde{x}_0)$  depends on the selected conjugate points  $\{x_m\}$ , the energy means  $\{\bar{E}_m\}$ , the mean statistical errors  $\{\sigma_m\}$  and the function space spanned by  $g(a, x)$ . In order to make the analysis and corresponding error control algorithm tractable, we make a few simplifying choices and assumptions. First, we choose  $\{x_m\}$  to lie on a uniform grid of length  $L$  (the grid extent) centered at an estimate of the minimum obtained from the surrogate theory ( $\tilde{x}_0^S$ ). Second, we choose to obtain statistical estimates of the energy values on the grid with uniform uncertainty  $\sigma$ , so that  $\sigma_m = \sigma$ . Finally, we assume that the energy means are well modeled by the energy values from the deterministic surrogate theory, or  $E_m \approx E_m^S$ . These choices are illustrated in Fig. 3. With

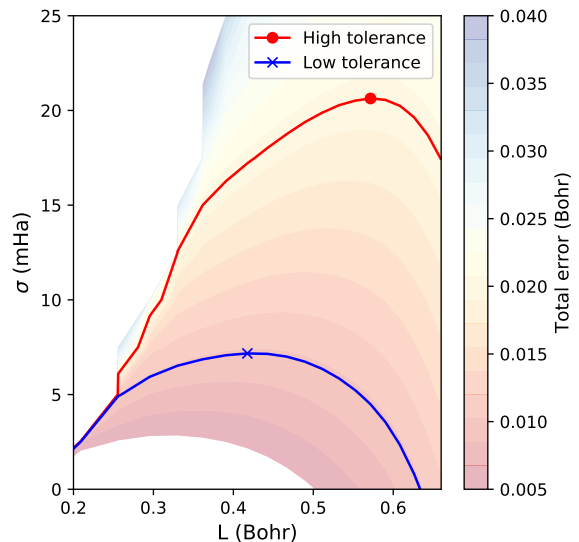


FIG. 4. To optimize the cost of each line-search we find the grid extent  $L$  that allows as large an input noise  $\sigma$  as possible while keeping the error within a tolerance. This is done by mapping the high likelihood edge of the conjugate parameter bias  $\delta x_P$  on the  $L\sigma$ -plane and finding the peak of the isocontour  $(L_{max}, \sigma_{max})$  corresponding to a given error tolerance. Contours from low and high tolerances ( $\delta x_{tol}$ ) are shown, demonstrating that a larger conjugate parameter error tolerance generally allows the use of a larger grid extent  $L$ , which in turn tolerates larger input noise  $\sigma$  (lowers computational cost).

uniform stochastic noise on  $E$ , a distribution of possible minima are obtained, as indicated by the horizontal error bar near the minimum. The error distribution  $P(x - \tilde{x}_0)$  is shown in the inset to Fig. 3 with percentiles  $P_{2.5}$  and  $P_{97.5}$  shown in red and blue, respectively. With these choices, for a given  $g(a, x)$  modeling the PES along search direction, the degrees of freedom determining the error distribution  $P(x - \tilde{x}_0)$  have now been reduced to just two: the grid extent  $L$  and the point-wise statistical noise  $\sigma$ .

We exploit this reduced freedom to directly maximize the statistical noise subject to variations in the grid extent while obeying the high-likelihood error tolerance condition expressed in Eqn. 19. To accomplish this, we map the high likelihood edge of the bias,  $\delta x_P \equiv \max(|P_{2.5}|, |P_{97.5}|)$  as a function of  $\sigma$  and  $L$ . For a given  $\sigma$ ,  $L$ , and  $g(a, x)$  the error distribution of the PES minimum  $P(x_0 - \tilde{x}_0)$ , and hence  $\delta x_P$ , may be explicitly calculated by a Monte Carlo sampling of Eqn. 18, where each Gaussian-distributed stochastic variable  $\eta_m$  is sampled independently. By using correlated sampling across variations in  $\sigma$  and  $L$ , we explicitly obtain a smooth estimate of the bias surface  $\delta x_P(\sigma, L)$ . Any point along the contour  $\mathcal{C}(\sigma, L)$  defined by  $\delta x_P(\sigma, L) = \delta x_{tol}$  satisfies the tolerance condition of Eqn. 19. The point of maximum statistical efficiency is obtained by finding the point  $(\sigma_{max}, L_{max})$  on  $\mathcal{C}$  that has the largest value of  $\sigma$ . This condition determines the line search grid (defined by

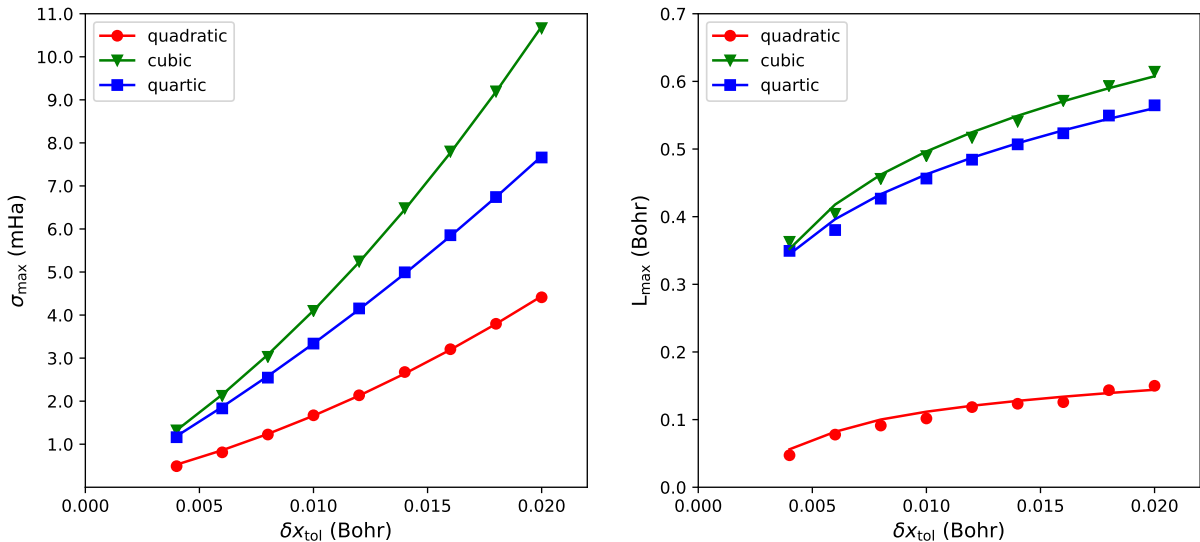


FIG. 5. Maximum noises obtained from error isocontours as functions of tolerance in a benzene line-search (direction  $n = 0$ ) using quadratic (red dots), cubic (green triangles) and quartic (blue squares) polynomial degrees. The quadratic degree is a suboptimal choice, because it tolerates the least noise, making it most expensive, and also the smallest grid extents  $L$ , making it least reliable for bracketing the equilibrium. The cubic and quartic polynomials tolerate approximately similar grid extents, but the cubic emerges optimal for having fewer degrees of freedom and thus tolerating more statistical noise.

$L_{max}$ ) that admits the largest possible statistical noise while still obeying the chosen tolerance on the bias of the PES minimum along the search direction. An example of the error surface  $\delta x_P(\sigma, L)$  obtained via correlated sampling is shown in Fig. 4 along with two contours corresponding to lower (blue; 0.01 Bohr) and higher (red; 0.02 Bohr) conjugate parameter tolerances ( $\delta x_{tol}$ ). In each case the point  $(\sigma_{max}, L_{max})$  is marked, showing the result of the noise maximization procedure.

The final selection we make relates to the PES fitting function  $g(a, x)$ . In the near-minimum regime, low order polynomials are expected to capture well the local behavior of the PES. We consider quadratic, cubic, and quartic polynomials for this purpose. Polynomials that tolerate a high statistical noise and large search grid extent while still yielding a good fit are to be preferred. We use this condition as a means of selecting the best polynomial form for stochastic PES fitting along the search directions. Fig. 5 shows the maximum tolerable noise  $\sigma_{max}$  (left panel), and the corresponding optimal grid extent  $L_{max}$  as a function of the bias tolerance  $\delta x_{tol}$  for quadratic (red circles), cubic (green triangles), and quartic (blue squares) polynomials for an example direction in the benzene PES. As can be seen in the left panel of Fig. 5, cubic polynomials allow for the largest amount of statistical noise at all error tolerances considered. The optimal grid extent is also maximal for cubic polynomials (right panel), independent of the error tolerance, which is advantageous because this allows for the greatest possible search domain in each iteration of the line search.

The error distribution  $P(x - \tilde{x}_0)$  simultaneously accounts for errors arising from multiple possible sources relative to the surrogate PES, including statistical error,

statistical bias, and systematic bias arising from the inevitable mismatch between a given  $g(a, x)$  and the PES at large enough grid extent  $L$ . For quadratic polynomials, the systematic bias dominates early with respect to grid extent, curtailing the possibility of noise insensitivity. At higher polynomial orders, the statistical error and bias come to dominate, with statistical noise contributing to large fluctuations in the fitted polynomial shape and hence the location of the minimum. The cubic form mitigates the effects of these two limits, exposing few fitting parameters to statistical fluctuations while providing a basic description of asymmetry about the minimum which allows for modest systematic bias at larger grid extents.

With these developments, we now have the capability to obtain stochastic estimates of the PES minimum along a given line search direction  $d_n$  with controlled error and high efficiency. Ultimately, however, we wish to control the error of estimated direct structural parameters  $\{p_n\}$  which each mix contributions from all the conjugate directions. Effective control of error in the estimated structural parameters is the topic of the next subsection.

#### D. Controlling error in structural parameter estimates

Controlling the totality of error on each structural parameter according to user specified tolerances requires an accounting of how the various sources of error mix when conjugate parameter values are mapped back into the space of the parameters. Even though the mapping between the conjugate directions and the parameters is simple and linear ( $p = Dx$ ), the error in the minimum

$x_{0n}$  along each direction  $d_n$  is distributed according to Eqn. 18, which is non-Gaussian and contains a complex dependence on both the pointwise statistical noise  $\sigma$  and the grid extent  $L$ . We therefore introduce two approximate, but effective, approaches to impose our desired condition that the parameter error  $\delta p$  remains at or below a given tolerance, or

$$|\delta p| \leq \delta p_{tol} \quad (20)$$

where the inequality and absolute value are applied element-wise. In general, it is desirable to saturate the inequality in Eqn. 20, since obtaining parameters with overly small uncertainty rapidly becomes statistically inefficient.

The first approach approximately solves an  $N$ -dimensional fixed point problem while minimizing cost. The second approach takes physical inspiration from thermodynamics, where the uncertainty in each normal mode (here the parameter Hessian eigenvectors) follows directly from equipartition of energy at thermal equilibrium. We refer to the first approach as the fixed point balancing algorithm and the second as the thermal balancing algorithm. We briefly cover each approach below.

The fixed point error balancing algorithm directly uses the non-linear mapping between the errors in the minima along the conjugate directions and the errors in the equilibrium structural parameters

$$\delta p = \mathcal{M}(\delta x) \quad (21)$$

We obtain the forward mapping via sampling the convolution of the direction-wise error distributions (Eqn. 18) through the linear mixing matrix  $D$ . In general, we wish to determine  $\delta x_{tol}$  from the provided  $\delta p_{tol}$ , which constitutes an inversion of this mapping. We are interested in the set of conjugate parameter tolerances that minimizes cost while still satisfying Eqn. 20. This is expressed formally as

$$\delta x_{tol} = \arg \min_{\delta x} \mathcal{C}(\delta x) , \quad |\mathcal{M}(\delta x)| \leq \delta p_{tol} \quad (22)$$

where  $\mathcal{C}$  represents the computational cost given  $\delta x$ .

Based on the earlier section, each tolerance  $\delta x$  is known to map to a set of optimal mesh  $L_{max}$  and input noise  $\sigma_{max}$ , all of which increase monotonically. Thus, the cost function can be written as

$$\mathcal{C}(\delta x) = \sum_{i=1}^N \frac{1}{\sigma_{max}^2(\delta x_i)}, \quad (23)$$

where  $\mathcal{C}$  is proportional to the sum over inverse squared noises of all the point evaluations of the PES along all conjugate directions.

Assuming that maximum parameter errors relate to minimum cost, the ideal solution for  $\delta x$  is one that saturates Eqn. (20) such that the fixed point  $\mathcal{M}(\delta x) \approx \delta p_{tol}$  is approached. In the limit of purely systematic errors (limit of no noise), the propagation of error

obeys  $\delta x = D^T \delta p$ . With unbiased Gaussian statistical noise the propagation of error instead obeys  $\delta x = (((D^T)^{\circ 2})^{-1} \delta p)^{\circ 1/2}$ , where  $A^{\circ \alpha}$  denotes the Hadamard (element-wise) power for matrices. In the low noise limit, this relation instead becomes  $\delta x \approx |D|^T \delta p$ , where  $|\cdot|$  refers to the element-wise absolute value. In general, neither limit is fully satisfied, but a mixture of the two limits provides an effective means to search for the point of minimum cost, according to

$$\delta x(a, z) = a |z D^T \delta p_{tol} + (1 - |z|) |D|^T \delta p_{tol}|, \quad (24)$$

where  $z$  ranges from  $-1$  to  $1$  and  $a$  is a rescaling factor used to meet Eqn. (21) as closely as possible. With this approach, we obtain a low cost target for the conjugate parameter tolerances as

$$a_0(z) = \arg \min_a (\delta p_{tol} - \mathcal{M}(\delta x(a, z)))^2 \quad (25)$$

$$z_0 = \arg \min_z \mathcal{C}(\delta x(a_0(z), z)) \quad (26)$$

$$\delta x_{tol} = \delta x(a_0(z_0), z_0) \quad (27)$$

while obeying the constraint  $|\mathcal{M}(\delta x(a, z))| \leq \delta p_{tol}$ .

In the thermal balancing algorithm, we instead introduce a fictitious temperature  $T$  and consider the probable range of motion of the normal modes (vibrations along conjugate directions) when all modes are in thermal equilibrium. Via the equipartition theorem, the energy in each mode is  $\frac{1}{2} k_B T$  ( $k_B = 1$  in atomic units), and so the conjugate direction tolerances can be parameterized simply by  $T$  as

$$\delta x_{tol,n}(T) = \sqrt{\frac{T}{\lambda_n}} \quad (28)$$

where  $\lambda_n$  are the mode stiffnesses from Eqn. 5. The temperature is increased until one of the  $N$  components of  $\mathcal{M}(\delta x_{tol}(T))$  saturates the inequality in Eqn. 20. An advantage of the thermal balancing approach is that one does not need to specify a tolerance for each structural parameter. Instead, a subset of parameters tolerances can be specified (including just a single one) and the other parameter tolerances will be set automatically in a physically reasonable way.

Regardless of the cross-parameter tolerance enforcement algorithm used, the resulting cost balance differs strongly among the modes. Soft modes are the hardest to resolve statistically for a given parameter tolerance and thus often dominate the cost of the structural optimization. For this reason alone, it is valuable from an efficiency standpoint to closely tailor the target statistical noise for each respective search direction in the structural optimization process.

### III. DEMONSTRATION OF STRUCTURAL OPTIMIZATION FOR FLAKE-LIKE MOLECULES

We next demonstrate some of the most important properties of the parallel line-search algorithm by finding minimum energy geometries of a series of flake-like

aromatic molecules, namely benzene ( $C_6H_6$ ), coronene ( $C_{24}H_{12}$ ) and ovalene ( $C_{32}H_{14}$ ). The molecules have been chosen because they comprise a related family of increasing complexity.

Here we use DFT to obtain the surrogate Hessian and perform the main line-search with DMC, as detailed in Sec. III A. In all cases, the vibrational approach (Eqns. 13 & 14) was used to obtain the surrogate Hessian. The parameters optimized are inter-atomic distances, or bond lengths, where we target 0.01 or 0.02 Bohr accuracy, depending on the bond type. We compare the results to selected experiments<sup>37–39</sup>, though there are limitations that enter this comparison. The main limitation is that the experimental molecular structures were deduced at finite temperature, and also include zero-point motion effects, while our calculations are not exact and are performed at zero temperature. Additionally, inferring ground state structural parameters from experiments may contain biases, such as overestimating C-H bonds (see e.g. Ref 40), which have not been corrected in any of our reference data. The distances between hydrogen atoms were not estimated in the original experiments, because of their relatively high uncertainty.

### A. Simulation details

The surrogate mean-field simulation is based on DFT calculations with the PBE<sup>41</sup> functional, as implemented in Quantum ESPRESSO<sup>42</sup>. Quantum ESPRESSO is used to compute mean-field energies and force-constant matrices, as detailed in earlier sections. Ultrasoft pseudopotentials<sup>43</sup> are used with plane-wave cutoff energies of 40 Ry and for energy and 160 Ry for density, and with self-consistent field (SCF) convergence accuracy of  $10^{-9}$  Ry. The force cutoff for relaxation is 0.0001 Ry/Bohr. The phonon calculation was done at Gamma point with self-consistency threshold  $1.0 \times 10^{-12}$  Ry<sup>2</sup> (input variable `tr2_ph` to `ph.x` in the Quantum ESPRESSO suite).

The method used to generate the stochastic PES is standard DMC as implemented in QMCPACK<sup>44,45</sup>. The trial wavefunction was based on DFT-PBE orbitals from Quantum ESPRESSO, using correlation-consistent pseudopotentials (ccECP)<sup>46</sup> with cutoff energies of 400 Ry and 800 Ry for energies and densities, respectively, and convergence accuracy of  $10^{-9}$  Ry in the total energy. A two-body Jastrow factor was used with electron-ion and electron-electron terms between the relevant ion and spin species. The Jastrow optimization targeted minimum energy with the linear method<sup>47</sup>. Subsequent DMC calculations based on the optimized Slater-Jastrow trial wavefunction used the T-moves<sup>48</sup> scheme to sample the non-local pseudopotential contributions and maintain the variational quality of DMC. A time-step convergence study was performed to ensure that DMC time-step of 0.01/Ha does not lead to a significant bias in the structural properties. Targeting specific statistical uncertain-

ties  $\sigma$  was done by assuming that the observed uncertainty  $\sigma_N$  depends on the total number of samples  $N$  as  $\sigma_N^2 \sim \sigma_E^2/N = c/N$ , consistent with the central limit theorem, where  $c$  is a characteristic constant depending on the energy variance  $\sigma_E^2$  among other effects, such as sample autocorrelation. The total number of samples can be broken down as  $N = N_w N_b N_s$ , where the number of walkers  $N_w = 2000$  and the number of statistical blocks  $N_b = 200$  are fixed. The number of steps per block  $N_s$  is varied to meet  $N_s \approx c/N_w N_b \sigma^2$  to produce  $\sigma_N \approx \sigma$ . Statistical uncertainties of the structural parameters have 95% confidence based on statistical resampling, as detailed in Sec. II C.

The structural optimization was carried out within a custom Python framework which implements all techniques laid out in this work. The framework makes extensive use of Nexus<sup>49</sup> to manage and streamline standard QMC workflows. The workflow for each DMC simulation involves SCF calculation, Jastrow optimization, and finally a DMC simulation. The Python code and all inputs and outputs used in this work is archived at the Materials Data Facility<sup>50,51</sup> [data link to be provided upon acceptance of this manuscript]. The Python code is also available at [https://github.com/QMCPACK/surrogate\\_hessian\\_relax](https://github.com/QMCPACK/surrogate_hessian_relax).

### B. Low Parameter Molecule: Benzene

Let us start by demonstrating some basic properties of the structural optimization algorithm for the benzene molecule using various parametric mappings and line-search settings. Benzene ( $C_6H_6$ ) is a planar molecule comprising an inner hexagonal ring of 6 carbon atoms and an outer ring of 6 hydrogen atoms. It has 36 degrees of freedom in the coordinate space, but  $D_{6h}$  symmetry reduces the ground state structure to two irreducible parameters. We consider two alternative representations of the parameters: carbon-carbon distance with either hydrogen-hydrogen distance (CC/HH) or carbon-hydrogen distance (CC/CH). The two parameterizations produce different mappings of the PES. CC/CH is more natural in the sense that the parameter directions are closer to the conjugate directions of the surrogate Hessian than is the case for the CC/HH parameterization.

First, we demonstrate the importance of conjugate search directions in the convergence of the parallel line-search (see Figures 1 and 6). For demonstration purposes we optimize the structure relative to the bare DFT PES, which is free of stochastic effects and is of course fully consistent with the surrogate Hessian. The energy-minimizing target geometry is also known exactly, but the line-search is initiated from a displaced starting point, where both parameters have been offset by  $\pm 0.1$  Bohr. We first use the CC/HH parameterization of benzene where the Hessian is more off-diagonal, showing more significant mixing in the conjugate directions. The parallel line-search progresses by simultaneously finding the

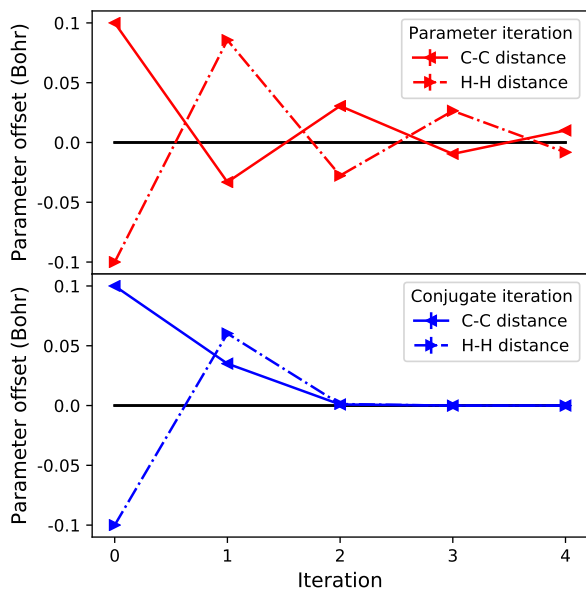


FIG. 6. Convergence of benzene structural parameters for the noise-free DFT PES via parallel line-search. The CC/HH parameters of benzene begin displaced by 0.1 Bohr and are relaxed back to equilibrium based on either the original parameter directions (top, red) or the optimized conjugate directions obtained from the DFT parameter Hessian (bottom, red).

energy minima along each search direction separately and then using the resulting geometry as the starting point for the next iteration. Progress towards the DFT PES minimum is illustrated in Figure 1 using either the bare parameter directions (red, top) or the conjugate directions (blue, bottom) to guide the line search. The corresponding parameter offsets from the targets are shown in Figure 6 as functions of the number of iterations: The parameter directions (red, top) are inefficient, leading to overcompensation and slow, jagged convergence towards the target. The conjugate directions obtained from the Hessian (blue, bottom) allow rapid convergence in effectively two iterations, the first of which is consumed in entering the quadratic region of the PES from the sizable starting displacement. From this example, it is evident that accurate directions are needed to coincidentally optimize multiple strongly coupled parameters efficiently.

To demonstrate the equivalence between the fixed point and thermal balancing algorithms for selecting and enforcing parameter accuracy tolerances, we share optimization results using each method for benzene using DFT with added noise. For the fixed point algorithm, the tolerances for the CC and CH bond lengths are set to 0.01 Bohr. For the thermal balancing algorithm, the largest effective temperature was selected such that both bond length targets remained at or below 0.01 Bohr. The resulting bond lengths are  $r_{CC} = 2.634(4)$  Bohr and  $r_{CH} = 2.061(4)$  Bohr for fixed point, and  $r_{CC} = 2.632(4)$  Bohr and  $r_{CH} = 2.066(4)$  Bohr with thermal balancing.

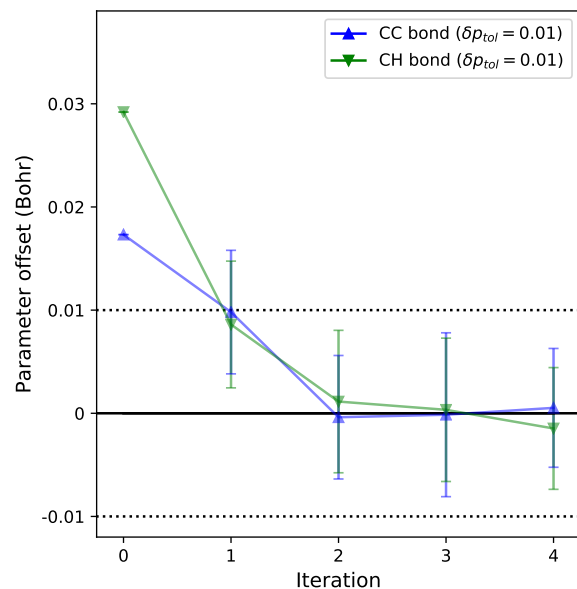


FIG. 7. Convergence of benzene structural parameters for the stochastic DMC PES via parallel line search. The initial CC/CH parameters are obtained from the DFT relaxed structure and converge rapidly under line search along the surrogate conjugate directions to the DMC equilibrium. For the DMC relaxation, a parameter accuracy tolerance of 0.01 Bohr was used for both the CC and CH bond lengths. The zero of the offset is the mean over three last line-search iterations.

The results are in statistical agreement with each other and remain within the 0.01 Bohr accuracy target relative to the noise-free DFT results of  $r_{CC} = 2.636$  Bohr and  $r_{CH} = 2.070$  Bohr.

Next, and in all the remaining examples, we will optimize the structure based on the stochastic DMC PES and use DFT to obtain the initial structure, the surrogate Hessian, and the conjugate directions for the parallel line search. The DMC energy is subject to finite noise and all the optimization and control techniques outlined in Sections II C and II D have been employed.

In all following examples, the fixed point method is used to enforce parameter accuracy targets. We choose to optimize the CC/CH parameterization of benzene to parameter tolerances of 0.01 Bohr. For each line-search we use third order polynomials with 7 grid points, based

Parameter	DFT	DMC <sup>31</sup>	DMC	Expt.
$r_{CC}$	2.636	2.6209(1)	2.618(4)	2.6271(4)
$r_{CH}$	2.070	2.037(2)	2.041(4)	2.053(3)
RMS	0.014(2)	0.017(2)	0.011(4)	

TABLE I. Comparison of estimated CC/CH DFT (PBE) and DMC parameter values for benzene with experiment (Bohr units). The first column of DMC results are obtained by a re-sampling analysis of biquartic fits to the 2D CC/CH uniform grid data obtained in Ref. 31. Root-mean-squared deviations (RMS) are computed relative to experimental results<sup>37</sup>.

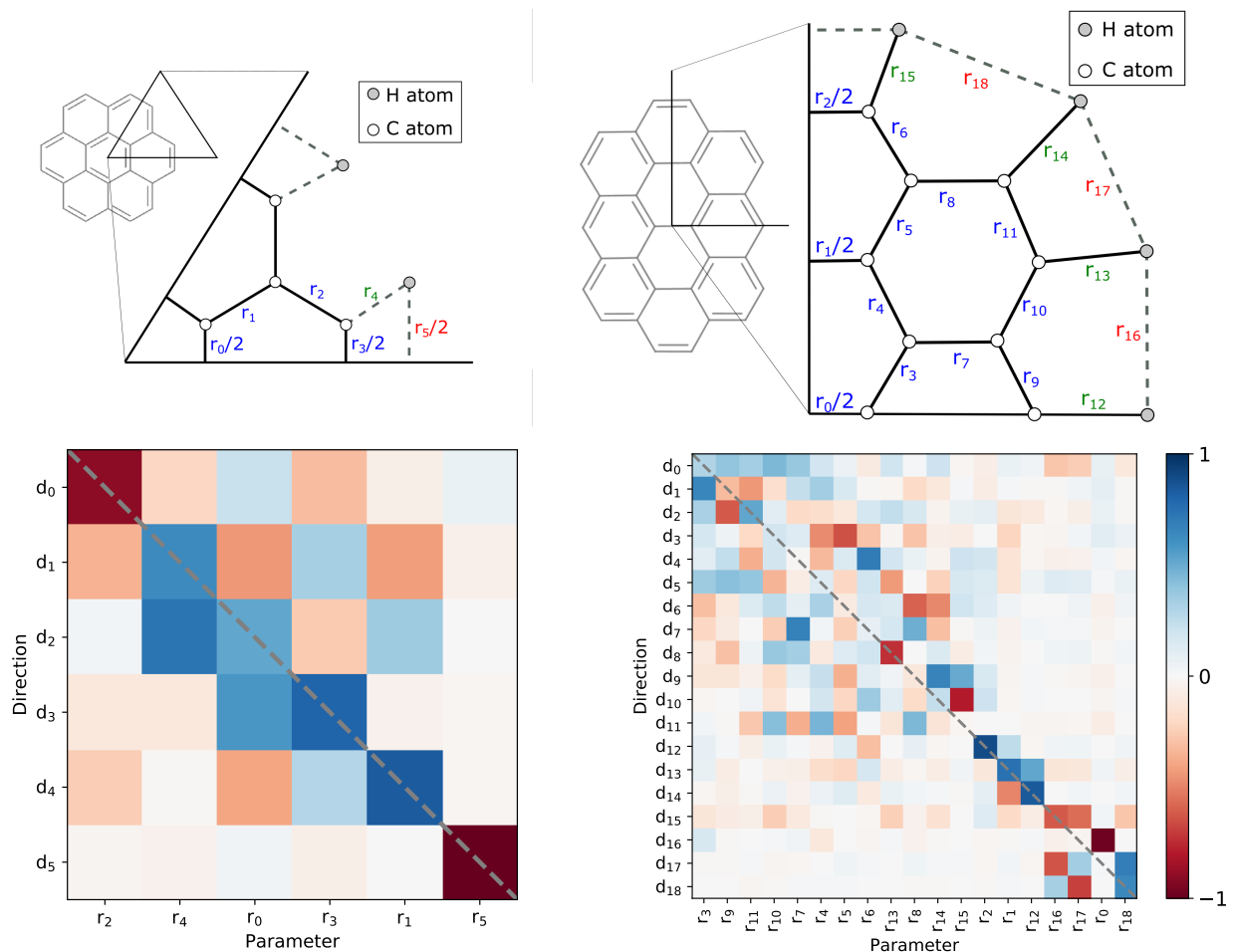


FIG. 8. Parametric structural models (top row) and conjugate directions showing parameter couplings (bottom row) for coronene (left column) and ovalene (right column). The parameters ( $r_i$ ) and the conjugate directions ( $d_i$ ) are sorted in order of descending stiffness.

on the discussion in Sec. II C.

The traces of parameter convergence from the DFT starting point to the estimated DMC equilibrium is shown in Fig. 7. Similar to the noise-free DFT case, the structural parameters for DMC converge in two iterations and fall within the target tolerances after the first iteration, which is near the ideal limit for conjugate direction methods. Subsequent single point, higher precision DMC calculations indicate that the DMC energy is reduced by 0.98(8) mHa between the initial DFT geometry and the estimated minimum. An energy reduction on this order of magnitude is consistent with energy changes observed under similar displacements in a prior DMC study of benzene in Ref. 31 using Gaussian process techniques for structural optimization. Due to the cost optimization algorithm outlined in Section II C, each parallel line-search iteration of this work has a statistical cost equivalent to performing a single total energy calculation resolved to a statistical uncertainty of 0.74 mHa on the total energy. The fact that we can resolve structural

differences with energy uncertainties much larger than the local energy scale of the PES underscores the benefit of using the cost optimized line-search curve fitting approach.

The numerical values of the optimal DMC structural parameters for benzene are found in Table I, alongside values obtained from experiment<sup>37</sup>. The DMC values are statistical means and uncertainties based on the last three line-search iterations. The root-mean squared deviation from the experiment (RMS) of DMC is 0.011(4) Bohr compared to 0.014(2) Bohr for PBE and the DMC errors fall near the target accuracy of 0.01 Bohr we imposed as a constraint for the structural optimization.

### C. More Complex Molecules: Coronene and Ovalene

We now perform DMC structural optimization for two more complicated molecules, coronene ( $C_{24}H_{12}$ ) and ovalene ( $C_{32}H_{14}$ ), whose structures and parameter couplings

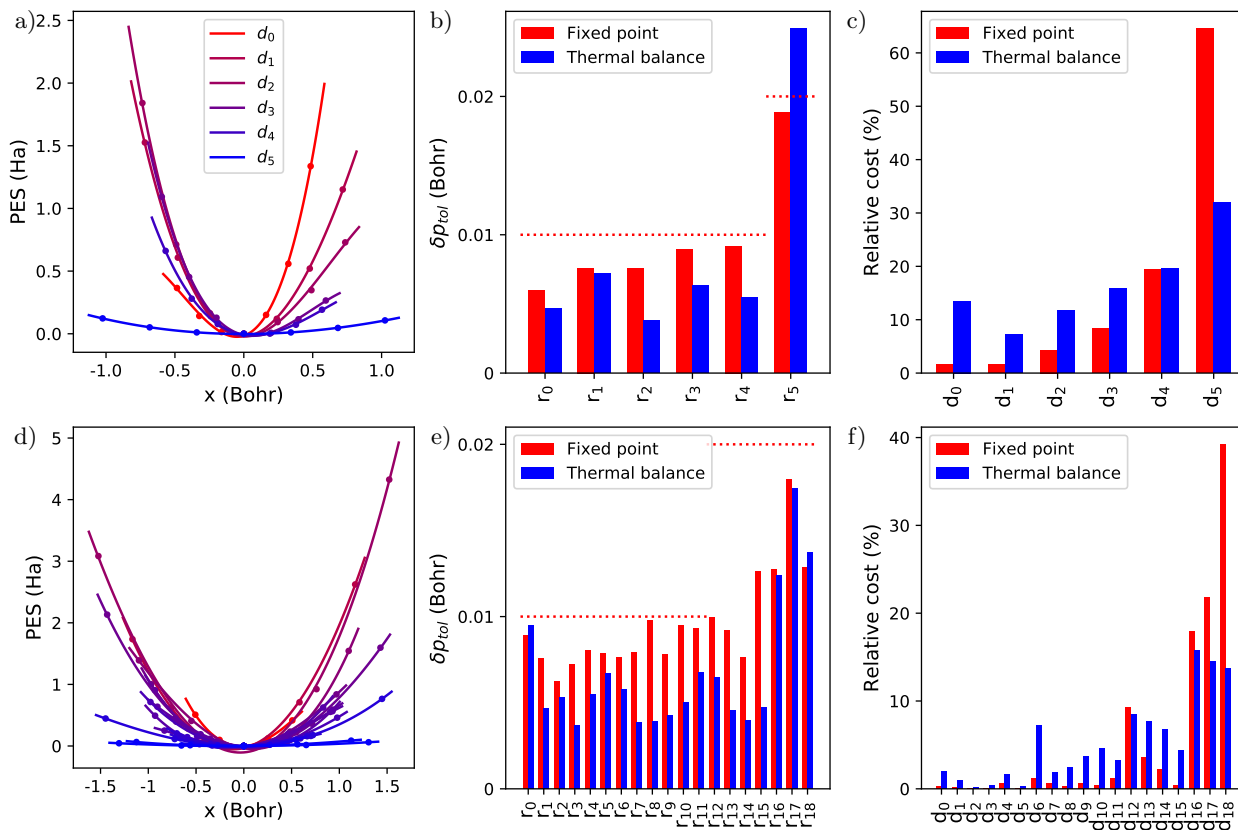


FIG. 9. Results of the surrogate-based cost-optimization procedure for 6-parameter coronene (top row) and 19-parameter ovalene (bottom row). The left panel (a,d) presents DFT PES curves near equilibrium along each conjugate direction with maximum extents ( $L_{max}$ ) optimized for computational cost while satisfying parameter tolerances. The middle panel (b,e) compares per-parameter ( $r_i$ ) tolerances ( $\delta p_{tol}$ ) obtained from the fixed point algorithm with manually set tolerances (red; tolerances given by dotted lines) and the thermal balancing algorithm (blue) with equivalent total cost. The right panel (c,f) shows the relative computational costs attributable to the line search along each respective conjugate direction ( $d_i$ ), as dictated by the maximum tolerable noise ( $\sigma_{max}$ ) for each PES slice shown in panels (a,d).

are illustrated in Fig. 8. Similar to benzene, coronene has  $D_{6h}$  symmetry, but has greater parametric complexity since it is composed of 7 aromatic rings. Symmetry reduces the 108 coordinate degrees of freedom to just 6 structural parameters for a full description of the equilibrium geometry. Ovalene has three additional aromatic rings, which further reduces the symmetry to  $D_{2h}$ . Here the 138 coordinate degrees of freedom are reduced to 19 bond parameters to describe the equilibrium structure. The ovalene structure relaxed here is the largest in terms of atom count and symmetry inequivalent structural parameters yet attempted in QMC with an energy based optimization algorithm<sup>30,31</sup>.

The sets of conjugate directions within each parameterization are presented in the bottom panels of Fig. 8 using heatmaps, to illustrate the nature of the parameter couplings. It is clear that the distance-based parameters of both molecules are highly coupled, and so the Hessian information is critical for fast convergence of the relaxation. The parameter-space surrogate Hessians are obtained by transforming dynamical matrices

from DFT frozen-phonon calculations to the parameter space using numerical Jacobian matrices of the parameter mappings (see Supplemental Information for more details). As usual, the conjugate directions are the normalized eigenvectors of the parameter Hessian.

In Figure 8, the conjugate directions are numbered and the parameters ordered based on decreasing stiffness to compression, bringing out a few key characteristics. First, there is clear, although partial, diagonal correlation. High diagonal correlation signals energetic independence of the parameters from one another. In the case of coronene, the  $r_5$  parameter (H-H distance) is highly independent, and also the softest. Likewise, the last three parameters of ovalene  $r_{16}$ - $r_{18}$  (H-H distances) form a soft and relatively independent subspace of search directions, with significant internal couplings. In general, the C-C and C-H bonds experience significant complex couplings which emphasizes the benefit of determining them automatically via diagonalization of the surrogate Hessian.

Using information from the surrogate Hessian and energy surface, we have cost-optimized the parallel line-

search of each molecule to meet the parameter tolerances, as laid out in Sec. II C. We use the fixed-point balancing algorithm (see Sec. II D) to enforce parameter tolerances of 0.01 Bohr (C-C and C-H bonds of coronene; C-C bonds of ovalene) or 0.02 Bohr (H-H bond of coronene; C-H and H-H bonds of ovalene), allowing looser tolerances in the soft parameters to better focus the computational effort.

The results of the cost-optimization procedure are illustrated in Fig. 9 for coronene (a-c) and ovalene (d-f). The sub-panels illustrate error and cost-related properties of the sets of conjugate directions of coronene and ovalene. The left panels (a, d) show DFT energy surfaces that represent the actual cost-optimized line-search grids, showing variations in the stiffness, the spatial extent, and the characteristic shape due to anharmonicity. The middle panels (b, e) show the enforced total parameter errors (red bars) based on random resampling of the parallel line-search, and the manually requested tolerances (red dotted lines). As required, the parameter errors remain within the tolerances, but not all of them are saturated, because the fixed-point error balancing algorithm samples a constrained space of direction balances and returns the one with the lowest estimated computational cost rather than demanding the closest match between requested and enforced tolerances in all parameters. This algorithm leads to cost-savings of factors of 1.4, 3.0 and 6.9 respectively for benzene, coronene and ovalene. The cost-saving potential is estimated relative to the unguided case (no surrogate information used for cost optimization), where each line-search would instead be executed with uniform statistical sampling according to the lowest tolerated noise.

Figure 9 also includes a comparison to the thermal error balancing algorithm (blue bars; see Sec. II D), where the temperature has been set to match the total computational cost of the fixed-point error balancing algorithm. The intuition of the thermal balancing algorithm is to spread the parameter tolerances more naturally based on atomic motions that may be anticipated with similar probability at a fixed temperature. This approach can be useful because relative parameter tolerances are set in a physical way without requiring detailed knowledge of the system. It also generally results in per-direction costs with lower spread. This second property is useful in the context of parallel line search as it reduces the wall time spent per iteration (and hence the overall time to solution) since the next iteration must wait for all line search computations in the prior iteration to complete before commencing. Based on the thermal balance, most of the target parameter errors of coronene (Fig. 9b) are reduced, except for the bottleneck H-H parameter ( $r_5$ ), where the target error is increased. For ovalene (Fig. 9e), the thermal balancing algorithm happens to outperform the fixed-point balancing algorithm in terms of the target accuracy of the structural parameters, while maintaining a much more uniform balance of costs.

With the line-search parameters optimized for cost and accuracy, we next demonstrate the DMC structural re-

laxation. In Fig. 10 we show rapid convergence of structural parameters and DMC energies for both systems, coronene and ovalene, using the surrogate DFT structures as the starting points. The DMC energies converge in just one iteration and are lowered by approximately 3 mHa in both cases. Statistical convergence of the structural parameters to the manual tolerances (dotted lines) is observed within 1-2 iterations, showing near ideal performance. In both cases, the assessment of converge is done self-consistently against averages over the saturated region (the last three iterations for each molecule). The self-consistent assessment of convergence only covers statistical effects, but not the systematic bias, which is controlled for via the tolerance but remains finite. More accurate evaluations between the DFT starting geometries and the predicted DMC geometries result in 3.3(3) mHa energy lowering for coronene and 3.7(3) mHa for ovalene.

Comparison of the structural parameters between surrogate DFT, DMC and experiments<sup>38,39</sup> are presented in Tables II and III for coronene and ovalene, respectively. The RMS deviation from experiment for C-C bonds in coronene is 0.015(3) Bohr for PBE and 0.011(4) for DMC, following closely the observations for benzene in Sec. III B. For ovalene, the RMS deviations from experiment are 0.017(9) Bohr for PBE and 0.018(9) Bohr for DMC. While the deviation here is possibly larger, the substantially larger uncertainty in the experimental values for ovalene also renders these deviations statistically consistent with those of coronene.

Greater precision can be obtained by comparing averaged C-C and C-H bond lengths across the molecules and theoretical/experimental methods. Bond lengths averaged in this way can be found in Table IV. From the table, we can observe that PBE overestimates the bond lengths by 0.0089(4), 0.010(3), and 0.005(9) Bohr for the C-C bond in benzene, coronene, and ovalene, respectively, and by 0.029(4) Bohr for the C-H bond in benzene. DMC underestimates the C-C bond lengths by -0.009(4), -0.008(4), and -0.007(9) Bohr for the C-C bond in benzene, coronene, and ovalene, respectively, while giving close agreement (0.00(6) Bohr) for the C-H bond in benzene. In general, the finite temperature present in experiment is expected to increase the observed bond lengths

Parameter	DFT	DMC	Expt.
$r_0$	2.692	2.679(3)	2.693(6)
$r_1$	2.689	2.660(3)	2.676(6)
$r_2$	2.685	2.681(4)	2.685(6)
$r_3$	2.594	2.567(4)	2.566(6)
$r_4$	2.072	2.037(3)	
$r_5$	4.676	4.616(9)	
RMS	0.015(3)	0.011(4)	

TABLE II. Comparison of estimated parameter values for coronene in units of Bohr. Root-mean-squared (RMS) deviations are computed relative to experimental results<sup>38</sup>, where available.

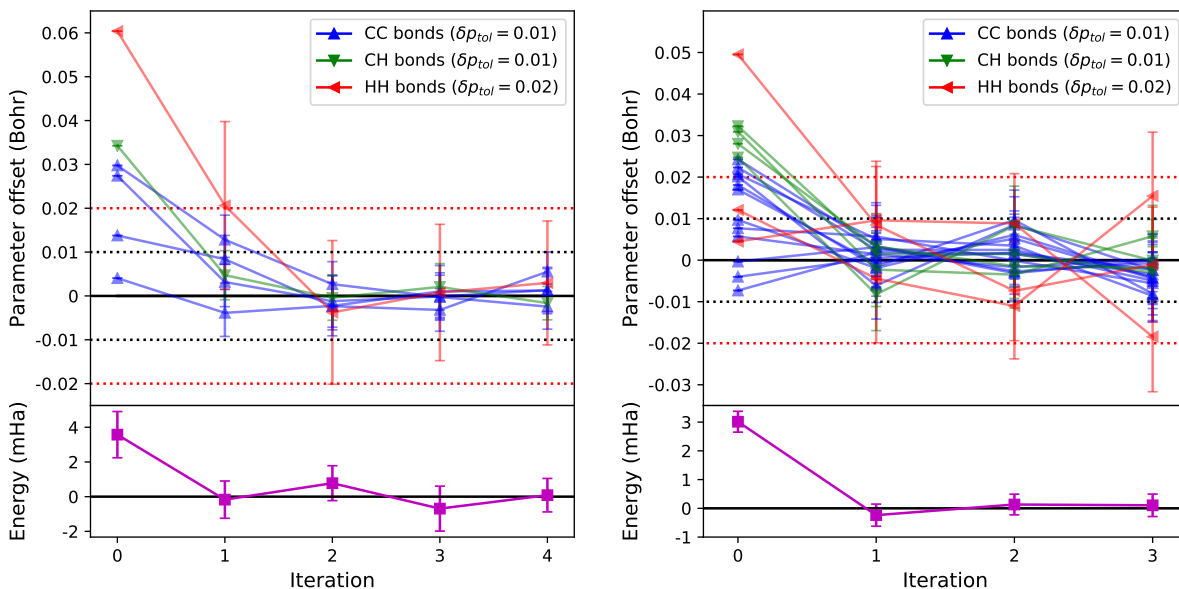


FIG. 10. Convergence of the parameters and energy of coronene (left) and ovalene (right) with respect to the mean of three last iterations of the parallel line search. The convergence of parameter values is shown in the top panels, while the lower panels show the behavior of the total energy.

relative to the exact zero Kelvin results. This is consistent with the reduced bond lengths seen in the zero temperature DMC results and a full account of temperature effects is likely to only increase the observed overestimation on the part of PBE.

Relative changes across structures may be expected to be less sensitive to experimental finite temperature ef-

Parameter	DFT	DMC	Expt.
$r_0$	2.704	2.697(4)	2.70(3)
$r_1$	2.693	2.683(6)	2.66(3)
$r_2$	2.607	2.590(4)	2.62(3)
$r_3$	2.680	2.657(5)	2.70(3)
$r_4$	2.690	2.694(5)	2.70(3)
$r_5$	2.691	2.665(5)	2.67(3)
$r_6$	2.671	2.662(5)	2.66(3)
$r_7$	2.707	2.683(5)	2.71(3)
$r_8$	2.698	2.697(5)	2.68(3)
$r_9$	2.647	2.629(5)	2.64(3)
$r_{10}$	2.702	2.708(6)	2.72(3)
$r_{11}$	2.582	2.565(5)	2.56(3)
$r_{12}$	2.074	2.053(7)	
$r_{13}$	2.072	2.042(6)	
$r_{14}$	2.072	2.042(6)	
$r_{15}$	2.072	2.039(5)	
$r_{16}$	4.685	4.676(9)	
$r_{17}$	4.670	4.617(10)	
$r_{18}$	4.692	4.682(10)	
RMS	0.017(9)	0.018(9)	

TABLE III. Comparison of estimated parameter values for ovalene in units of Bohr. Root-mean-squared deviations (RMS) are computed relative to experiments<sup>39</sup>, where available.

	C-C			C-H		
	$C_6H_6$	$C_{24}H_{12}$	$C_{32}H_{14}$	$C_6H_6$	$C_{24}H_{12}$	$C_{32}H_{14}$
DFT	2.636	2.665	2.673	2.070	2.072	2.073
DMC	2.618(4)	2.647(2)	2.661(2)	2.041(4)	2.037(3)	2.044(3)
Expt.	2.6271(4)	2.655(3)	2.668(9)	2.041(4)		

TABLE IV. Estimated mean C-C and C-H bond lengths for benzene, coronene, and ovalene from DFT (PBE), DMC, and experiment. All experimental values<sup>37-39</sup> were obtained near room temperature, while the theoretical results are at 0K with clamped nuclei.

fects and the description of correlation on the part of theory. The mean experimental C-C bond length increases by 0.028(3) and 0.041(9) Bohr for coronene and ovalene, respectively, when compared to benzene. For DMC, the same respective quantities are 0.029(4) and 0.043(3) Bohr, which agrees very closely with the changes observed experimentally.

#### IV. CONCLUSIONS

We have developed an energy-based method for molecular and solid state structural optimization with stochastic electronic structure theories, such as DMC. The method is based on robust line-search energy minimization in the space of structural parameters. The method exploits inexpensive but reasonably accurate Hessian information from a surrogate theory, such as DFT, in order to accelerate the descent on the stochastic PES. Based on the surrogate Hessian, line-searches are performed in

parallel iterations, with convergence to the energy minimum being achieved in only a few steps in practice. The line search iterations require only  $\mathcal{O}(N)$  stochastic energy evaluations. The use of a surrogate Hessian offers great computational savings since directly constructing the Hessian for  $N$  reduced parameters with the high-accuracy stochastic theory requires  $\mathcal{O}(N^2)$  additional energy evaluations at high cost. The method also employs the surrogate PES to predict and control systematic and statistical errors in the estimated structural parameters while maximizing statistical efficiency.

We have demonstrated the method on the aromatic molecules benzene, coronene and ovalene, with up to 19 symmetry inequivalent structural parameters. In each case, statistical convergence of the structural parameters to 0.01 or 0.02 Bohr accuracy is achieved in two parallel iterations or fewer, which is near the ideal limit for conjugate direction methods applied to an exact, quadratic PES. Based on direct evaluation of energies instead of energy gradients, the method is reliable and simple to use with any stochastic electronic structure method, such as the quantum Monte Carlo family of methods or with the stochastic formulations of the coupled cluster and configuration interaction methods.

## V. SUPPLEMENTAL MATERIAL

See supplementary material for more detail regarding structure parameter mappings, obtaining the parameter derivatives via automatic differentiation, and convergence properties of the line search when using 2nd, 3rd, or 4th order polynomials for local PES fitting.

## VI. ACKNOWLEDGEMENTS

This research has been provided by the US Department of Energy, Office of Science, Basic Energy Sciences, Materials Sciences and Engineering Division, as part of the Computational Materials Sciences Program and Center for Predictive Simulation of Functional Materials. This research used resources of the Compute and Data Environment for Science (CADES) at the Oak Ridge National Laboratory, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC05-00OR22725. An award of computer time was provided by the Innovative and Novel Computational Impact on Theory and Experiment (INCITE) program. This research used resources of the Argonne Leadership Computing Facility, which is a DOE Office of Science User Facility supported under Contract DE-AC02-06CH11357.

## VII. DATA AVAILABILITY

The data that support the findings of this study are openly available within the article, its supplementary material, and in the Materials Data Facility<sup>50,51</sup> at [link to be provided upon acceptance of this manuscript].

## VIII. REFERENCES

- <sup>1</sup>F. Neese, "Prediction of molecular properties and molecular spectroscopy with density functional theory: From fundamental theory to exchange-coupling," *Coordination Chemistry Reviews* **253**, 526–563 (2009).
- <sup>2</sup>M. Bühl, C. Reimann, D. A. Pantazis, T. Bredow, and F. Neese, "Geometries of third-row transition-metal complexes from density-functional theory," *Journal of Chemical Theory and Computation* **4**, 1449–1459 (2008).
- <sup>3</sup>J. Klimeš, D. R. Bowler, and A. Michaelides, "Van der waals density functionals applied to solids," *Physical Review B* **83** (2011), 10.1103/physrevb.83.195131.
- <sup>4</sup>T. Björkman, "Testing several recent van der waals density functionals for layered structures," *The Journal of Chemical Physics* **141**, 074708 (2014).
- <sup>5</sup>V. Zatzko, S. M.-M. Dubois, F. Godel, C. Carrétéro, A. Sander, S. Collin, M. Galbiati, J. Peiro, F. Panciera, G. Patriarche, P. Brus, B. Servet, J.-C. Charlier, M.-B. Martin, B. Dlubak, and P. Seneor, "Band-gap landscape engineering in large-scale 2d semiconductor van der waals heterostructures," *ACS Nano* (2021), 10.1021/acsnano.1c00544.
- <sup>6</sup>D. Wines, K. Saritas, and C. Ataca, "A first-principles quantum monte carlo study of two-dimensional (2d) GaSe," *The Journal of Chemical Physics* **153**, 154704 (2020).
- <sup>7</sup>K. Haule and G. L. Pascut, "Forces for structural optimizations in correlated materials within a DFT+embedded DMFT functional approach," *Physical Review B* **94**, 195146 (2016).
- <sup>8</sup>N. Lanata, Y. Yao, C.-Z. Wang, K.-M. Ho, and G. Kotliar, "Phase diagram and electronic structure of praseodymium and plutonium," *Physical Review X* **5**, 011008 (2015).
- <sup>9</sup>M. Barborini, S. Sorella, and L. Guidoni, "Structural optimization by quantum monte carlo: Investigating the low-lying excited states of ethylene," *Journal of Chemical Theory and Computation* **8**, 1260–1269 (2012).
- <sup>10</sup>M. Barborini and L. Guidoni, "Reaction pathways by quantum monte carlo: Insight on the torsion barrier of 1,3-butadiene, and the conrotatory ring opening of cyclobutene," *The Journal of Chemical Physics* **137**, 224309 (2012).
- <sup>11</sup>E. Coccia and L. Guidoni, "Quantum monte carlo study of the retinal minimal model  $c_{25}h_{66}nh_2^+$ ," *Journal of Computational Chemistry* **33**, 2332–2339 (2012).
- <sup>12</sup>E. Coccia, D. Varsano, and L. Guidoni, "Protein field effect on the dark state of 11-cis retinal in rhodopsin by quantum monte carlo/molecular mechanics," *Journal of Chemical Theory and Computation* **9**, 8–12 (2012).
- <sup>13</sup>R. Guareschi and C. Filippi, "Ground- and excited-state geometry optimization of small organic molecules with quantum monte carlo," *Journal of Chemical Theory and Computation* **9**, 5513–5525 (2013).
- <sup>14</sup>E. Coccia, D. Varsano, and L. Guidoni, "Ab initio geometry and bright excitation of carotenoids: Quantum monte carlo and many body green's function theory calculations on peridinin," *Journal of Chemical Theory and Computation* **10**, 501–506 (2014).
- <sup>15</sup>A. Zen, Y. Luo, S. Sorella, and L. Guidoni, "Molecular properties by quantum monte carlo: An investigation on the role of the wave function ansatz and the basis set in the water molecule," *Journal of Chemical Theory and Computation* **9**, 4332–4350 (2013).

- <sup>16</sup>D. Varsano, E. Coccia, O. Pulci, A. M. Conte, and L. Guidoni, "Ground state structures and electronic excitations of biological chromophores at quantum monte carlo/many body green's function theory level," *Computational and Theoretical Chemistry* **1040-1041**, 338–346 (2014).
- <sup>17</sup>M. Barborini and L. Guidoni, "Ground state geometries of polyacetylene chains from many-particle quantum mechanics," *Journal of Chemical Theory and Computation* **11**, 4109–4118 (2015).
- <sup>18</sup>M. Barborini and L. Guidoni, " $\pi$ -conjugation in trans-1,3-butadiene: Static and dynamical electronic correlations described through quantum monte carlo," *Journal of Chemical Theory and Computation* **11**, 508–517 (2015).
- <sup>19</sup>J. R. Trail, "Heavy-tailed random error in quantum monte carlo," *Phys. Rev. E* **77**, 016703 (2008).
- <sup>20</sup>R. Assaraf and M. Caffarel, "Computing forces with quantum monte carlo," *The Journal of Chemical Physics* **113**, 4028–4034 (2000).
- <sup>21</sup>R. Assaraf and M. Caffarel, "Zero-variance zero-bias principle for observables in quantum monte carlo: Application to forces," *The Journal of Chemical Physics* **119**, 10536–10552 (2003).
- <sup>22</sup>C. Attaccalite and S. Sorella, "Stable liquid hydrogen at high pressure by a NovelAb InitioMolecular-dynamics calculation," *Physical Review Letters* **100** (2008), 10.1103/physrevlett.100.114501.
- <sup>23</sup>A. Badinski, P. D. Haynes, J. R. Trail, and R. J. Needs, "Methods for calculating forces within quantum monte carlo simulations," *Journal of Physics: Condensed Matter* **22**, 074202 (2010).
- <sup>24</sup>S. Sorella and L. Capriotti, "Algorithmic differentiation and the calculation of forces by quantum monte carlo," *The Journal of Chemical Physics* **133**, 234111 (2010).
- <sup>25</sup>S. Moroni, S. Sacconi, and C. Filippi, "Practical schemes for accurate forces in quantum monte carlo," *Journal of Chemical Theory and Computation* **10**, 4823–4829 (2014).
- <sup>26</sup>P. L. Ríos and G. J. Conduit, "Tail-regression estimator for heavy-tailed distributions of known tail indices and its application to continuum quantum monte carlo data," *Physical Review E* **99** (2019), 10.1103/physreve.99.063312.
- <sup>27</sup>C. A. Schuetz, M. Frenklach, A. C. Kollias, and W. A. Lester, "Geometry optimization in quantum monte carlo with solution mapping: Application to formaldehyde," *The Journal of Chemical Physics* **119**, 9386–9392 (2003).
- <sup>28</sup>A. Zen, D. Zhelyazov, and L. Guidoni, "Optimized structure and vibrational properties by error affected potential energy surfaces," *Journal of Chemical Theory and Computation* **8**, 4204–4215 (2012).
- <sup>29</sup>D. M. Cleland and M. C. Per, "Performance of quantum monte carlo for calculating molecular bond lengths," *The Journal of Chemical Physics* **144**, 124108 (2016).
- <sup>30</sup>L. K. Wagner and J. C. Grossman, "Quantum monte carlo calculations for minimum energy structures," *Phys. Rev. Lett.* **104**, 210201 (2010).
- <sup>31</sup>R. Archibald, J. T. Krogel, and P. R. C. Kent, "Gaussian process based optimization of molecular geometries using statistically sampled energy surfaces from quantum monte carlo," *The Journal of Chemical Physics* **149**, 164116 (2018).
- <sup>32</sup>H. Sorenson, "Comparison of some conjugate direction procedures for function minimization," *Journal of the Franklin Institute* **288**, 421–441 (1969).
- <sup>33</sup>W. I. Zangwill, *Nonlinear programming: a unified approach*, Vol. 52 (Prentice-Hall Englewood Cliffs, NJ, 1969).
- <sup>34</sup>S. J. Wright, "Coordinate descent algorithms," *Mathematical Programming* **151**, 3–34 (2015).
- <sup>35</sup>M. J. D. Powell, "An Iterative Method for Finding Stationary Values of a Function of Several Variables," *The Computer Journal* **5**, 147–151 (1962), <https://academic.oup.com/comjnl/article-pdf/5/2/147/975263/5-2-147.pdf>.
- <sup>36</sup>H. Shin, J. T. Krogel, K. Gasperich, P. R. C. Kent, A. Benali, and O. Heinonen, "Optimized structure and electronic band gap of monolayer GeSe from quantum monte carlo methods," *Physical Review Materials* **5**, 024002 (2021).
- <sup>37</sup>J. Plíva, J. Johns, and L. Goodman, "Infrared bands of isotopic benzenes:  $\nu_{13}$  and  $\nu_{14}$  of  $^{13}\text{C}_6\text{D}_6$ ," *Journal of Molecular Spectroscopy* **148**, 427–435 (1991).
- <sup>38</sup>A. S. Filatov, N. J. Sumner, S. N. Spisak, A. V. Zabula, A. Y. Rogachev, and M. A. Petrukhina, "Jahn–teller effect in circulenes: X-ray diffraction study of coronene and corannulene radical anions," *Chemistry – A European Journal* **18**, 15753–15760 (2012), <https://chemistry-europe.onlinelibrary.wiley.com/doi/pdf/10.1002/chem.201202026>.
- <sup>39</sup>D. M. Donaldson and J. M. Robertson, "The crystal and molecular structure of ovalene a quantitative x-ray investigation," *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences* **220**, 157–170 (1953).
- <sup>40</sup>J. Gauss and J. F. Stanton, "The equilibrium structure of benzene," *The Journal of Physical Chemistry A* **104**, 2865–2868 (2000).
- <sup>41</sup>J. P. Perdew, K. Burke, and M. Ernzerhof, "Generalized gradient approximation made simple," *Phys. Rev. Lett.* **77**, 3865–3868 (1996).
- <sup>42</sup>P. Giannozzi, O. Andreussi, T. Brumme, O. Bunau, M. B. Nardelli, M. Calandra, R. Car, C. Cavazzoni, D. Ceresoli, M. Cococcioni, N. Colonna, I. Carnimeo, A. D. Corso, S. de Gironcoli, P. Delugas, R. A. DiStasio, A. Ferretti, A. Floris, G. Fratesi, G. Gallo, R. Gebauer, U. Gerstmann, F. Giustino, T. Gorni, J. Jia, M. Kawamura, H.-Y. Ko, A. Kokalj, E. Küçükbenli, M. Lazzeri, M. Marsili, N. Marzari, F. Mauri, N. L. Nguyen, H.-V. Nguyen, A. O. de-la Roza, L. Paulatto, S. Poncé, D. Rocca, R. Sabatini, B. Santra, M. Schlipf, A. P. Seitsonen, A. Smogunov, I. Timrov, T. Thonhauser, P. Umari, N. Vast, X. Wu, and S. Baroni, "Advanced capabilities for materials modelling with quantum ESPRESSO," *Journal of Physics: Condensed Matter* **29**, 465901 (2017).
- <sup>43</sup>K. F. Garrity, J. W. Bennett, K. M. Rabe, and D. Vanderbilt, "Pseudopotentials for high-throughput DFT calculations," *Computational Materials Science* **81**, 446–452 (2014).
- <sup>44</sup>J. Kim, A. D. Baczewski, T. D. Beaudet, A. Benali, M. C. Bennett, M. A. Berrill, N. S. Blunt, E. J. L. Borda, M. Casula, D. M. Ceperley, S. Chiesa, B. K. Clark, R. C. Clay, K. T. Delaney, M. Dewing, K. P. Esler, H. Hao, O. Heinonen, P. R. C. Kent, J. T. Krogel, I. Kylänpää, Y. W. Li, M. G. Lopez, Y. Luo, F. D. Malone, R. M. Martin, A. Mathuriya, J. McMinis, C. A. Melton, L. Mitas, M. A. Morales, E. Neuscammman, W. D. Parker, S. D. P. Flores, N. A. Romero, B. M. Rubenstein, J. A. R. Shea, H. Shin, L. Shulenburger, A. F. Tillack, J. P. Townsend, N. M. Tubman, B. V. D. Goetz, J. E. Vincent, D. C. Yang, Y. Yang, S. Zhang, and L. Zhao, "QMCPACK: an open source ab initio quantum monte carlo package for the electronic structure of atoms, molecules and solids," *Journal of Physics: Condensed Matter* **30**, 195901 (2018).
- <sup>45</sup>P. R. C. Kent, A. Annaberdiyev, A. Benali, M. C. Bennett, E. J. L. Borda, P. Doak, H. Hao, K. D. Jordan, J. T. Krogel, I. Kylänpää, J. Lee, Y. Luo, F. D. Malone, C. A. Melton, L. Mitas, M. A. Morales, E. Neuscammman, F. A. Reboledo, B. Rubenstein, K. Saritas, S. Upadhyay, G. Wang, S. Zhang, and L. Zhao, "QMCPACK: Advances in the development, efficiency, and application of auxiliary field and real-space variational and diffusion quantum monte carlo," *The Journal of Chemical Physics* **152**, 174105 (2020).
- <sup>46</sup>M. C. Bennett, C. A. Melton, A. Annaberdiyev, G. Wang, L. Shulenburger, and L. Mitas, "A new generation of effective core potentials for correlated calculations," *The Journal of Chemical Physics* **147**, 224106 (2017).
- <sup>47</sup>J. Toulouse and C. J. Umrigar, "Optimization of quantum monte carlo wave functions by energy minimization," *The Journal of Chemical Physics* **126**, 084102 (2007).
- <sup>48</sup>M. Casula, "Beyond the locality approximation in the standard diffusion monte carlo method," *Physical Review B* **74** (2006), 10.1103/physrevb.74.161102.

- <sup>49</sup>J. T. Krogel, “Nexus: A modular workflow management system for quantum simulation codes,” *Computer Physics Communications* **198**, 154–168 (2016).
- <sup>50</sup>B. Blaiszik, K. Chard, J. Pruyne, R. Ananthakrishnan, S. Tuecke, and I. Foster, “The materials data facility: Data services to advance materials science research,” *JOM* **68**, 2045–2052 (2016).
- <sup>51</sup>B. Blaiszik, L. Ward, M. Schwarting, J. Gaff, R. Chard, D. Pike, K. Chard, and I. Foster, “A data ecosystem to support machine learning in materials science,” *MRS Communications* **9**, 1125–1133 (2019).