Potential energy surface fitting by a statistically localized, permutationally invariant, local interpolating moving least squares method for the many-body potential: Method and application to N_4

Jason D. Bender,[1] Sriram Doraiswamy,[1] Donald G. Truhlar,[2,a)] and Graham V. Candler[1,a)]
[1]*Department of Aerospace Engineering and Mechanics, University of Minnesota, Minneapolis, Minnesota 55455, USA*
[2]*Department of Chemistry, Chemical Theory Center, and Supercomputing Institute, University of Minnesota, Minneapolis, Minnesota 55455, USA*

(Received 28 August 2013; accepted 2 January 2014; published online 3 February 2014)

Fitting potential energy surfaces to analytic forms is an important first step for efficient molecular dynamics simulations. Here, we present an improved version of the local interpolating moving least squares method (L-IMLS) for such fitting. Our method has three key improvements. First, pairwise interactions are modeled separately from many-body interactions. Second, permutational invariance is incorporated in the basis functions, using permutationally invariant polynomials in Morse variables, and in the weight functions. Third, computational cost is reduced by statistical localization, in which we statistically correlate the cutoff radius with data point density. We motivate our discussion in this paper with a review of global and local least-squares-based fitting methods in one dimension. Then, we develop our method in six dimensions, and we note that it allows the analytic evaluation of gradients, a feature that is important for molecular dynamics. The approach, which we call statistically localized, permutationally invariant, local interpolating moving least squares fitting of the many-body potential (SL-PI-L-IMLS-MP, or, more simply, L-IMLS-G2), is used to fit a potential energy surface to an electronic structure dataset for N_4. We discuss its performance on the dataset and give directions for further research, including applications to trajectory calculations. © *2014 AIP Publishing LLC*. [http://dx.doi.org/10.1063/1.4862157]

I. INTRODUCTION

Born-Oppenheimer potential energy surfaces (PESs) play a vital role in many subfields of chemistry. For example, a PES is a basic requirement for the calculation of cross sections and reaction rates from classical or quasiclassical trajectory methods in molecular reaction dynamics.[1] Energy transfer and dissociation in high-temperature air must be understood for the accurate simulation of hypersonic aerospace flows using computational fluid dynamics (CFD) and other approaches.[2–5] To investigate these chemical processes, a potential energy surface for $N_2 + N_2$ collisions is required. In a recent paper, a global PES was presented for the N_4 system; the surface was based on electronic structure calculations using complete active space second-order perturbation theory (CASPT2) and a least squares fitting method involving permutationally invariant polynomials.[6] In the present paper, we describe an alternate strategy, which we pursued concurrently, for fitting a six-dimensional surface to the tetranitrogen dataset. Our approach is a new version of the local interpolating moving least squares (L-IMLS) method, as investigated by Dawes *et al.*[7–10] and Guo *et al.*,[11] with three improvements. (1) Our method treats pairwise interactions separately from many-body interactions. (2) It incorporates permutational invariance, both in the basis functions and in the weight functions, by applying ideas pioneered by Braams and Bowman[12] and Xie and Bowman.[13,14] (3) The method employs a statistically correlated cutoff radius for increased computational efficiency. We call our approach *statistically localized, permutationally invariant, local interpolating moving least squares fitting of the many-body potential*, abbreviated SL-PI-L-IMLS-MP or simply L-IMLS-G2, where G2 denotes "second generation."

Numerous techniques have been proposed for constructing potential energy surfaces from analytic functions, given a discrete set of data from quantum chemistry. The motivation for this fitting procedure is well-known: in many chemical problems, it is prohibitively expensive to carry out direct dynamics, i.e., to generate energies and forces as needed directly from electronic structure calculations. Instead, one uses a *fitting function* that is relatively inexpensive to evaluate for energies and gradients. Fitting potential energy surfaces for systems with more than three atoms presents many challenges, and it has yielded a rich literature. Several reviews of fits and fitting methods are available.[15–18] In addition to approaches based on L-IMLS and permutational symmetry, which form the foundation for the present research and which we will discuss (with references) in detail, other schemes include those based on splines,[19–22] the double many-body expansion method (DMBE),[23–25] reproducing kernel Hilbert space interpolation (RKHS),[26,27] the combined valence bond molecular mechanics method (CVBMM),[28] modified

a)Authors to whom correspondence should be addressed. Electronic addresses: truhlar@umn.edu and candler@aem.umn.edu.

Shepard interpolation,[29–31] and the multiconfiguration molecular mechanics method (MCMM).[32–34]

A useful categorization of fitting methods is explained in a recent review.[17] We summarize that discussion briefly, since it formed a conceptual framework for much of this research. Fitting methods can be *global* or *local*. In a global method, each electronic structure data point influences the fitting function uniformly, i.e., in a way that is formally independent of the evaluation location. For a method to be global it is not necessary for each data point to have precisely the same effect on the fitting function; such a condition of constancy is a stronger condition than that of spatial uniformity. In a local method, the influence of a data point on the fitting function may vary with evaluation location. Typically, this means that only those data points that are nearby, with respect to an appropriate multi-dimensional definition of distance, affect the fitted energy at an evaluation point. L-IMLS and L-IMLS-G2 are local methods. We will further explore the distinction between global and local approaches throughout this paper. Note that any fitting method may be classified as global or local; this classification scheme is universal, because it is based only in a general sense on the way in which data points influence the fitting function.

In Sec. II, we discuss several variants of least squares methods, restricting our attention to one dimension. Our purpose there is to briefly review concepts, terminology, and prior research that will form the foundation for our exposition of L-IMLS-G2. In Sec. III, we develop L-IMLS-G2 in six dimensions. We devote particular attention to the separation of pairwise interaction energy from the total energy, the incorporation of permutational invariance in the basis functions, the incorporation of permutational invariance in the weight function distance metric, and the construction of a cutoff radius correlation for reducing computational cost. Finally, we present results from the application of L-IMLS-G2 to the N_4 dataset, draw conclusions, and give suggestions for future work.

II. GLOBAL AND LOCAL FITTING IN ONE DIMENSION

Consider a set of N data points, specified by the sequences $(x_1, x_2, \ldots, x_N)$ and $(f_1, f_2, \ldots, f_N)$. Suppose we wish to fit the data using a quadratic polynomial. Thus, we wish to determine *coefficients* (a_1, a_2, a_3) such that the fitting function $f(x) = p(x) = a_1 + a_2 x + a_3 x^2$ is a reasonable approximation to the data. The factors 1, x, and x^2 are *basis functions*. A simple way to solve this problem is to minimize the following functional:

$$E(p(x)) = \sum_{k=1}^{N} (p(x_k) - f_k)^2. \tag{1}$$

As discussed in textbooks on linear algebra, this leads to the matrix *normal equations*,

$$\mathbf{B}^{\mathsf{T}}\mathbf{B}\mathbf{a} = \mathbf{B}^{\mathsf{T}}\mathbf{f}, \tag{2}$$

where T denotes a matrix transpose, and where we have defined the following matrices:[35]

$$\mathbf{B} = \begin{pmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ \vdots & \vdots & \vdots \\ 1 & x_N & x_N^2 \end{pmatrix}, \quad \mathbf{a} = \begin{pmatrix} a_1 \\ a_2 \\ a_3 \end{pmatrix}, \quad \mathbf{f} = \begin{pmatrix} f_1 \\ f_2 \\ \vdots \\ f_N \end{pmatrix}. \tag{3}$$

The normal equations are solved for the coefficients $\{a_j\}$ using linear algebra. Those coefficients define the quadratic polynomial. This process is the standard *least squares* (LS) approach.

The method can be refined by introducing weights into the normal equations. Lancaster and Šalkauskas give a detailed discussion of a variety of least-squares-based methods, including theoretical issues of differentiability and numerical stability.[36] Suppose that some of the data points are deemed more significant than others, that is, we place a higher priority on the fitting function's accuracy near some data points than near others. Then we seek to minimize the following functional:

$$E(p(x)) = \sum_{k=1}^{N} \omega(x_k)(p(x_k) - f_k)^2. \tag{4}$$

Here, ω is the *weight function*, defined at least on the set $\{x_k\}$. A data point with a larger weight will have a larger influence on the functional. Minimization of E leads to new matrix normal equations,

$$\mathbf{B}^{\mathsf{T}}\mathbf{\Omega}\mathbf{B}\mathbf{a} = \mathbf{B}^{\mathsf{T}}\mathbf{\Omega}\mathbf{f}, \tag{5}$$

where we have defined the following diagonal matrix:

$$\mathbf{\Omega} = \begin{pmatrix} \omega(x_1) & 0 & \cdots & 0 \\ 0 & \omega(x_2) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \omega(x_N) \end{pmatrix} \Bigg/ \left(\sum_{k=1}^{N} \omega(x_k) \right). \tag{6}$$

Observe that $\mathbf{\Omega}$ is normalized such that all of its elements lie between 0 and 1. As before, the normal equations are solved for the coefficients $\{a_j\}$, which define the fitted polynomial. This process is the *weighted least squares* (WLS) approach.

There are many ways to define weights in the WLS method. When the value of the weight function at one of the data point coordinates x_i is relatively large, the resulting WLS polynomial will be especially accurate near that point. With a large enough weight, the polynomial will appear to nearly interpolate through the data point at x_i, although it is important to note that true interpolation is only achieved in the theoretical limit of infinite relative weight at x_i. It is interesting that smooth solutions to Eq. (5) can be obtained even with extremely large relative weights.[36,37] These observations motivate the consideration of a family of WLS polynomials, one associated with each data point. We can construct a set of WLS polynomials $\{p_1(x), p_2(x), \ldots, p_N(x)\}$ using a corresponding set of weight functions $\{\omega_1(x), \omega_2(x), \ldots, \omega_N(x)\}$,

where the polynomial $p_i(x)$ is designed to be especially accurate in the vicinity of x_i. We refer to x_i as the *localization point* of the WLS fit. This procedure, which requires us to solve a total of N matrix equations of the form given in Eq. (5), can be accomplished by equating $\omega_i(x)$ to a generalized, two-parameter weight function of the following form:

$$\omega(x, x_i) \equiv \nu\left(z = \frac{|x - x_i|^2}{R^2}\right), \quad \text{where } \nu(z) \equiv \frac{\exp(-z)}{z^{p_\nu} + \varepsilon_\nu^{p_\nu}}. \tag{7}$$

Similar functions have been studied in the one-dimensional case in the work of Maisuradze *et al.*[38,39] McLain also recommends a similar function in his studies of weighted least squares methods.[40,41] In Eq. (7), p_ν is an integer power, ε_ν is a small number used to ensure that $\omega_i(x)$ is well-defined at $x = x_i$, and R is a constant scaling parameter. We refer to the variable z as the *weight function variable*. In this scheme, a good generalized weight function should have the properties that it is large when $z = 0$ (at the localization point) and that it decreases sharply as z increases.

Having constructed a family of WLS polynomials in this way, we can now define a smooth fit to the entire dataset. We refer to the polynomials as local fitting functions or simply *local fits*. A new fitting function is defined as a weighted average of the local fits:

$$f(x) = \frac{\sum_{i=1}^{N} w(x, x_i)\, p_i(x)}{\sum_{i=1}^{N} w(x, x_i)}. \tag{8}$$

In Eq. (8), $w(x, x_i)$ is not the same as $\omega(x, x_i)$. These two functions serve different mathematical purposes; ω defines values in matrix equations that are solved for the coefficients of the local fits, whereas w defines weights in a weighted average of all local fits. Farwig presented a detailed study of interpolation methods involving equations similar to Eq. (8).[42,43] Furthermore, the modified Shepard interpolation scheme that Collins and Ischtwan presented for PES fitting relies on a similar formula. Note, however, that their method requires the gradient and Hessian at each data point to construct Taylor polynomials.[29,30] (Some success has also been achieved with modified Shepard interpolation schemes that do not require Hessians,[44,45] but further discussion of these methods is beyond our scope.) In the present approach, only energies are needed at the data points.

The purpose of w is to vary the influence of the local fits on $f(x)$. We use a form similar to Eq. (7), with new parameters p_u and ε_u that serve roles similar to those of the parameters p_ν and ε_ν. For simplicity, we use the same value of the scaling parameter R:

$$w(x, x_i) \equiv u\left(z = \frac{|x - x_i|^2}{R^2}\right), \quad \text{where } u(z) \equiv \frac{1}{z^{p_u} + \varepsilon_u^{p_u}}. \tag{9}$$

When the *evaluation point* x is close to one of the data points x_i, the corresponding weight function $w(x, x_i)$ is large, and the value of $f(x)$ is nearly equal to the value of the corresponding local fit $p_i(x)$. In this way, the local fits are recovered near the data points themselves, which had defined the localization points in the WLS constructions. The weighted sum in Eq. (8) combines the local fits into a single smooth function. The process we have described is the *local interpolating moving least squares* (L-IMLS) approach. The term L-IMLS and its application to PES fitting in physical chemistry are due to Dawes *et al.*[7] and Guo *et al.*,[11] who developed the method for systems of higher dimensionality. Their research makes use of weight functions similar to those in Eqs. (7) and (9).

Notice that the evaluation of an L-IMLS fitting function by Eq. (8) does not require any matrix computations. The construction of the local fits should be done in advance, yielding a set of coefficients $\{a_j^{[i]}\}$ for each of the data points x_i. Then, calculating $f(x)$ at any location only requires the evaluation of the local fits, the evaluation of the weight functions, and finally the summation and normalization defined by Eq. (8). We refer to these two distinct phases of the L-IMLS procedure as the *construction step* and the *evaluation step*. The fact that the evaluation step does not involve any work with matrix equations is a vital feature with regard to efficiency, as Dawes *et al.* and Guo *et al.* emphasize in their development and analysis of the method.[7,11]

It is useful to categorize the three methods above in the framework[17] we described in the Introduction. The LS and WLS approaches are global fitting methods. In the context of our one-dimensional discussion, they each yield a single quadratic polynomial, with coefficients determined by the data points via matrix normal equations. L-IMLS is a local approach. It involves the determination of a set of quadratic polynomials, one associated with each data point. With a typical weight function, the fitting function defined by Eq. (8) is more heavily influenced by data points that lie closer to the evaluation point. Thus, the data points do not influence the fitting function in a spatially uniform way. We expect the local method to be both more accurate and more computationally expensive than the global methods, because it uses a larger number of coefficients to define the fitting function. Figure 1 illustrates each of the three methods applied to a set of five data points.

For each of the examples considered in this section, the fitting function was based on polynomials, since we used monomial basis functions. When possible, we will continue to frame our discussion in terms of polynomials for simplicity. However, note that each of the LS, WLS, and L-IMLS methods can be used with other basis functions, such as trigonometric, rational, or exponential functions. Least squares methods are not limited to fitting functions of polynomial form. In this exposition, we do require that all basis function coefficients be determined linearly, via matrix normal equations of the form given by Eq. (5). Nonlinear methods are available for determining fitting function parameters in more general problems, but these are beyond our scope. Moreover, such approaches present additional difficulties, such as sensitivity to initial guesses.[46]

We also remark on a fourth kind of least squares method found in the literature. In the *interpolating moving least squares* (IMLS) approach, the fitting function f is defined as follows. At an evaluation point x, we form a set of matrix normal equations as in Eq. (5), by using the weight function $\omega_x(x') \equiv \omega(x', x)$ in which the localization point is taken

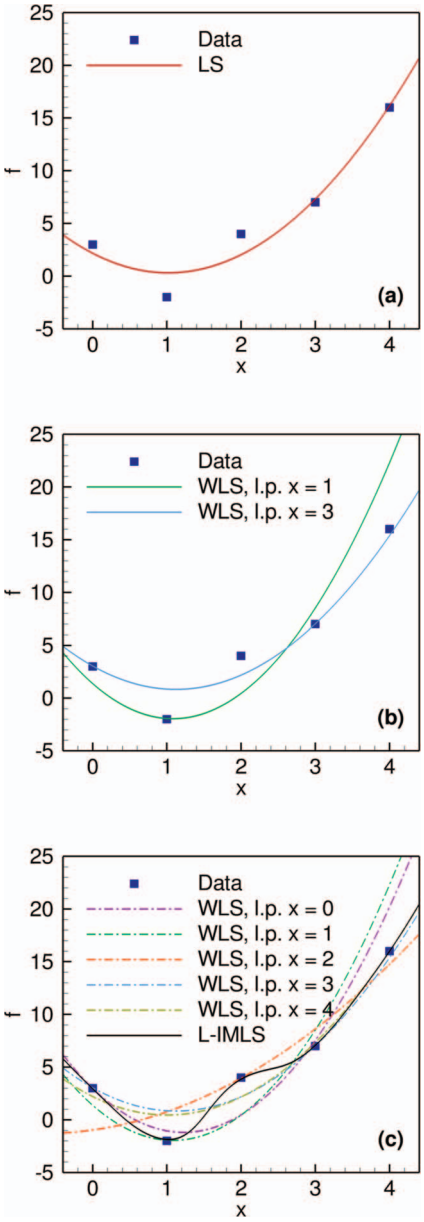

FIG. 1. Illustration of different fitting methods in one dimension using quadratic polynomials, for the data set defined by $(x_i) = (0, 1, 2, 3, 4)$ and $(f_i) = (3, -2, 4, 7, 16)$. Part (a) depicts the LS method. Part (b) depicts the WLS method, with the weight function given by Eq. (7) with $p_v = 1$ and $\varepsilon_v = 10^{-3}$. Results are shown for two different selections of the localization point (l.p.). Part (c) depicts the L-IMLS method. The weight function for the construction step is the same weight function used in part (b). The weight function for the evaluation step is given by Eq. (9) with $p_u = p_v$ and $\varepsilon_u = \varepsilon_v$. The dashed lines are the five local fits. The solid line is the fitting function defined by Eq. (8).

to be the evaluation point itself. The normal equations are solved for a set of coefficients $\{a_j\}$. These define a single polynomial, which is then evaluated at x. The resulting value is $f(x)$. Lancaster and Šalkauskas include a thorough discussion of IMLS methods, both for one-dimensional and multidimensional problems.[36] The IMLS approach was studied for physical chemistry applications by Maisuradze *et al.*[38,39] Tokmakov *et al.* also explored how to incorporate gradient information at the data points into the IMLS method.[47] The work by Guo *et al.* focuses on explicitly comparing the IMLS and L-IMLS approaches.[11]

The difference between L-IMLS and IMLS merits emphasis. Both are local methods by the definitions given above[17] (despite the distinction in nomenclature). IMLS is a simpler method in some respects; it involves only one major computation step and one generalized weight function. However, it is typically much more computationally expensive, since each fitting function evaluation with IMLS requires the solution to a new set of matrix normal equations. In particular, it would be quite difficult to express an IMLS fitting function in an explicit form like Eq. (8). Ultimately, as Guo *et al.* conclude, IMLS is less practical than L-IMLS for multidimensional fitting problems. Moreover, they found that the accuracies of IMLS and L-IMLS are similar.[11]

We conclude this section with several important comments about terminology. As a method for fitting potential energy surfaces, L-IMLS developed out of earlier studies using IMLS.[7,11] However, in organizing this discussion, we made a pedagogical choice to explain L-IMLS in terms of WLS fits, rather than introducing IMLS first. We believe this to be a more direct way to present and understand the method. According to Lancaster and Šalkauskas, the term "moving" in IMLS refers to the use of a new weight function, new matrix normal equations, and a new polynomial for every fitting function evaluation.[36] Strictly speaking, L-IMLS takes a different approach, in which a finite number of polynomial coefficients are computed in advance. Thus, rather than viewing L-IMLS as a variant of a moving least squares method, it might be better characterized as a *local weighted least squares* (L-WLS) method; it relies on a family of WLS local fits. Furthermore, as we noted earlier, the term "local" in L-IMLS could be interpreted as redundant, since IMLS is itself a local method according to the definitions[17] we reviewed. Nevertheless, we will continue to use the term L-IMLS to maintain consistency with prior literature.

Our final remark concerns the term "interpolating." Formally, a method is interpolating only if it produces a surface that passes through every one of the data points. Accordingly, in the exposition of Lancaster and Šalkauskas, IMLS is described as a special case of an ordinary *moving least squares* (MLS) method in which the maximum value of a generalized weight function $\omega(x', x)$ is increased to infinity.[36] We will instead use the term "interpolating" more loosely, to refer to methods that are designed to give highly accurate approximations to data but that do not necessarily achieve the theoretical interpolation limit. Again, this is done for consistency with earlier literature in the physical chemistry community.

III. SL-PI-L-IMLS-MP IN SIX DIMENSIONS

Having explored the conceptual framework of the L-IMLS approach, we now describe how we applied these ideas to construct a six-dimensional potential energy surface for a four-atom system with permutational invariance. As mentioned in Sec. I, we can refer to our method by the long name in the section heading or by the shorter name L-IMLS-G2. Throughout the development of this method, we maintained the following overall goals: (1) Our approach should yield a highly accurate and smooth surface. (2) It should be computationally efficient to evaluate energies. Finally, (3) the

method should be simple enough so that gradients, as needed for molecular dynamics simulations, can be computed analytically and efficiently.

A. Overview

We first give an overview of L-IMLS for the general six-dimensional fitting problem, establishing notation and conventions that will be used when we describe the distinguishing features of L-IMLS-G2. An analogous formulation of multi-dimensional L-IMLS can be found in the article by Guo *et al.*[11] Also instructive are the expositions of the IMLS method by Kawano *et al.* and Maisuradze *et al.* in their studies of the six-dimensional HOOH system.[48,49] Finally, as noted previously, Lancaster and Šalkauskas give an extensive treatment of multi-dimensional IMLS in their text.[36]

Consider a set of N data points, specified by three sequences:

$$(\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_N}), \quad (\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_N), \quad (f_1, f_2, \ldots, f_N).$$

The ith data point is described by two ordered six-tuples, $\mathbf{x}_i = (x_{i,1}, x_{i,2}, \ldots, x_{i,6})$ and $\mathbf{q}_i = (q_{i,1}, q_{i,2}, \ldots, q_{i,6})$, and by the energy f_i. We call $\mathbf{x}_i$ the coordinates of the point in the *basis function coordinate system*, and we call $\mathbf{q}_i$ the coordinates of the point in the *weight function coordinate system*. These two coordinate systems could be the same, but we will discuss the general case where they are not. We assume that we have a well-defined way to convert between the two coordinate systems. The flexibility granted by this framework is an important feature. Indeed, Maisuradze *et al.* have noted the advantages of using a "hybrid" coordinate system, in which the coordinate system for the basis functions is not necessarily the same as the one used for the weight functions.[49]

As in one dimension, L-IMLS in six dimensions consists of two distinct steps: a construction step and an evaluation step. In the first step, we construct a set of N local fits, one for each data point, as follows. Consider one of the points $\mathbf{x}_i \sim \mathbf{q}_i$, where the symbol $\sim$ denotes equivalence across the two coordinate systems. Treating $\mathbf{x}_i \sim \mathbf{q}_i$ as a localization point for a WLS fit, we construct a local fit of the following form:

$$p_i(\mathbf{x}) = \sum_{j=1}^{M} a_j^{[i]}\, b_j(\mathbf{x}). \tag{10}$$

In this equation, $\{b_j(\mathbf{x})\}$ are basis functions and $\{a_j^{[i]}\}$ are coefficients. M is the number of basis functions used for each local fit. We next introduce a weight function defined on the weight function coordinate space:

$$\omega(\mathbf{q}_1, \mathbf{q}_2) \equiv \nu\left(z = \frac{d^2(\mathbf{q}_1, \mathbf{q}_2)}{R^2(\mathbf{q}_1)}\right), \quad \text{where}$$
$$\nu(z) \equiv \frac{\exp(-z)}{z^{p_\nu} + \varepsilon_\nu^{p_\nu}}. \tag{11}$$

Equation (11) is similar to Eq. (7), but now we have defined the weight function variable z more generally; $d^2(\mathbf{q}_1, \mathbf{q}_2)$ denotes the square of the distance between points $\mathbf{q}_1$ and $\mathbf{q}_2$ in the weight function coordinate space. For now, we will assume that this distance metric is well defined; it will be discussed in greater detail in Sec. III D. Also, we allow the scaling parameter R to be a function of $\mathbf{q}_1$, the first argument of ω; this will also be the subject of further discussion.

Minimization of the functional,

$$E(p_i(\mathbf{x})) = \sum_{k=1}^{N} \omega(\mathbf{q}_k, \mathbf{q}_i)(p_i(\mathbf{x}_k) - f_k)^2, \tag{12}$$

leads to the normal equations,

$$\mathbf{B}^\mathsf{T}\boldsymbol{\Omega}_i\mathbf{B}\mathbf{a}_i = \mathbf{B}^\mathsf{T}\boldsymbol{\Omega}_i\mathbf{f}, \tag{13}$$

where we have defined the following matrices:

$$\mathbf{B} = \begin{pmatrix} b_1(\mathbf{x}_1) & b_2(\mathbf{x}_1) & \cdots & b_M(\mathbf{x}_1) \\ b_1(\mathbf{x}_2) & b_2(\mathbf{x}_2) & \cdots & b_M(\mathbf{x}_2) \\ \vdots & \vdots & \ddots & \vdots \\ b_1(\mathbf{x}_N) & b_2(\mathbf{x}_N) & \cdots & b_M(\mathbf{x}_N) \end{pmatrix},$$
$$\mathbf{a}_i = \begin{pmatrix} a_1^{[i]} \\ a_2^{[i]} \\ \vdots \\ a_M^{[i]} \end{pmatrix}, \quad \mathbf{f} = \begin{pmatrix} f_1 \\ f_2 \\ \vdots \\ f_N \end{pmatrix}, \tag{14}$$
$$\boldsymbol{\Omega}_i = \begin{pmatrix} \omega(\mathbf{q}_1, \mathbf{q}_i) & 0 & \cdots & 0 \\ 0 & \omega(\mathbf{q}_2, \mathbf{q}_i) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \omega(\mathbf{q}_N, \mathbf{q}_i) \end{pmatrix} \Bigg/ \left(\sum_{k=1}^{N} \omega(\mathbf{q}_k, \mathbf{q}_i)\right).$$

Observe that the definition of $\mathbf{B}$ in Eq. (14) is a generalization of the definition used in Eq. (3). Solving Eq. (13) for the coefficients $\{a_j^{[i]}\}$ defines the local fit $p_i(\mathbf{x})$. We repeat this process for each of the data points. This ultimately yields an $N \times M$ array of coefficients, describing a family of local fits.

For the evaluation step, we define a second weight function, analogous to Eq. (9),

$$w(\mathbf{q}_1, \mathbf{q}_2) \equiv u\left(z = \frac{d^2(\mathbf{q}_1, \mathbf{q}_2)}{R^2(\mathbf{q}_1)}\right), \quad \text{where}$$
$$u(z) \equiv \left(\frac{s(z)}{z^{p_u} + \varepsilon_u^{p_u}}\right), \tag{15}$$

where

$$s(z) = \begin{cases} (1 - z^4)^4 & \text{if } 0 \le z < 1 \\ 0 & \text{if } z \ge 1 \end{cases}. \tag{16}$$

Note that $s(z)$ forces $u(z)$ to be zero for all $z \ge 1$. The powers of 4 in Eq. (16) are used to achieve a satisfactory degree of differentiability. We use the same scaling factor R in Eq. (15) that we used in Eq. (11). Whenever the distance $d(\mathbf{q}_1, \mathbf{q}_2)$ is greater than or equal to $R(\mathbf{q}_1)$, the weight function $w(\mathbf{q}_1, \mathbf{q}_2)$ is zero; therefore, R will be called the *cutoff radius* for the remainder of the discussion. Our choice in Eq. (16) is based on the work of Guo *et al.*[11] and Kawano *et al.*,[50] who investigated cutoff strategies using a slightly more general form of

Eq. (16). An alternative choice for $s(z)$ involving an exponential function can be found in the paper by Tokmakov *et al.*[47] and in the work of Levin.[51] Yet another choice based on the hyperbolic tangent function can be found in earlier work by Maisuradze *et al.*[49] and by Guo *et al.*[52]

Consider an arbitrary point $\mathbf{x} \sim \mathbf{q}$. We adopt the shorthand notation $d_i^2 = d^2(\mathbf{q}, \mathbf{q}_i)$. Then the fitting function evaluated at $\mathbf{x}$ is defined by the following expression:

$$f(\mathbf{x}) = \frac{\sum_{i=1}^{N} w(\mathbf{q}, \mathbf{q}_i)\, p_i(\mathbf{x})}{\sum_{i=1}^{N} w(\mathbf{q}, \mathbf{q}_i)} = \frac{\sum_{\{i:d_i^2 < R^2(\mathbf{q})\}} w(\mathbf{q}, \mathbf{q}_i)\, p_i(\mathbf{x})}{\sum_{\{i:d_i^2 < R^2(\mathbf{q})\}} w(\mathbf{q}, \mathbf{q}_i)}. \tag{17}$$

As in Eq. (8), $f(\mathbf{x})$ is a weighted, normalized sum of local fits evaluated at $\mathbf{x}$. We only need to consider the local fits p_i whose localization point $\mathbf{q}_i$ lies within a hyper-sphere that is centered at $\mathbf{q}$ and has a radius of $R(\mathbf{q})$. As we noted in our discussion of Eq. (16), all other local fits receive a weight of zero and can be ignored.

B. The separation of pairwise interaction energy

Next we describe how we adapted the multi-dimensional L-IMLS method to construct a PES for a system of four identical atoms, using the example of N_4. Our focus will be on the distinguishing features of L-IMLS-G2. In this subsection we explain the initial step in our approach. As described in an earlier paper[6] and the references contained therein, the potential energy curve for the N_2 molecule has been extensively studied. Consequently, it is desirable to guarantee the following property in the N_4 PES: for geometries that are characterized only by pairwise interactions, the PES should reduce to a sum of pairwise potentials. We can ensure this behavior by fitting the energy difference between the total electronic energy E (which was calculated from quantum chemistry), and the pairwise component of the electronic energy E_{PW} (which was previously fit[6] using an analytic generalized Morse potential). The N_4 fitting procedure is conducted only on the resulting many-body component of the electronic energy E_{MB}:

$$f = E_{\mathrm{MB}} \equiv E - E_{\mathrm{PW}}. \tag{18}$$

We expect E_{MB} to be a smoother function than the total energy E, with favorable consequences for the accuracy of our fitting procedure (and, indeed, our experience bears this out). In particular, the E_{PW} term captures much of the character of the steep repulsive walls of the six-dimensional PES. This concept of fitting the difference from an energy component that is expressible in a simple analytic form was also discussed by Kawano *et al.*; in Eq. (18), E_{PW} plays the role of a "zeroth-order potential function" in the context of their investigation.[48] The basic principle is even older.[53]

C. Permutational invariance and the basis functions

We next turn to the definitions of a six-dimensional coordinate system and a set of basis functions $\{b_j\}$ to use in Eq. (10). A system of four atoms has six internal degrees of freedom. Consider any geometry of four atoms, corresponding to a single point on the potential energy surface. In previous work,[6] this point was identified by the following ordered six-tuple $\mathbf{t}$:

$$\mathbf{t} = (t_1, t_2, t_3, t_4, t_5, t_6) = (r_A, r_B, d, \theta_A, \theta_B, \phi). \tag{19}$$

These variables are depicted in the diagram in Figure 1 of Ref. 6; r_A and r_B are the 1–2 and 3–4 diatomic distances, θ_A and θ_B are the angles that the 1–2 and 3–4 diatomic axes form with the line from the center of mass of 1–2 to the center of mass of 3–4, d is the distance between these two centers of mass, and ϕ is the dihedral angle. (Specifically, ϕ is the angle between the plane containing the 1–2 axis and the line from the center of mass of 1–2 to the center of mass of 3–4, and the plane containing the 3–4 axis and the line from the center of mass of 1–2 to the center of mass of 3–4.) We call $\mathbf{t}$ the coordinates of the point in the *default coordinate system*. While this scheme proved useful for organizing electronic structure calculations, it is not a practical choice for the L-IMLS-G2 procedure: the mixture of distances and angles is cumbersome when defining basis functions and distances. Instead, a common choice is to identify the point by the set of six internuclear distances, which we denote by the ordered six-tuple $\tilde{\mathbf{q}}$,

$$\tilde{\mathbf{q}} = (\tilde{q}_1, \tilde{q}_2, \tilde{q}_3, \tilde{q}_4, \tilde{q}_5, \tilde{q}_6) = (r_{12}, r_{13}, r_{14}, r_{23}, r_{24}, r_{34}), \tag{20}$$

where r_{bc} is the distance in three-dimensional Cartesian space between atom b and atom c. We call $\tilde{\mathbf{q}}$ the coordinates of the point in the *raw internuclear distance coordinate system.* To better reproduce the repulsive walls of the potential energy surface and its asymptotic behavior at large internuclear distances, a modified coordinate system can be defined using Morse variables. We define a new ordered six-tuple $\mathbf{q} = (q_1, \ldots, q_6)$ by the following expression, for $l = 1, \ldots, 6$:

$$q_l = \exp\left(-\frac{\tilde{q}_l - \tilde{q}_{\mathrm{eq}}}{a}\right). \tag{21}$$

In this expression, $\tilde{q}_{\mathrm{eq}}$ is the equilibrium bond distance (equal to 1.098 Å for nitrogen) and a is a parameter. Both have units of distance. We call $\mathbf{q}$ the coordinates of the point in the *raw Morse coordinate system*.

The basis functions should account for the permutational symmetry of the system.[54] More specifically, if a geometry can be obtained by a permutation of the atoms of a second geometry, then the basis functions evaluated for each geometry should be identical. The subject of incorporating permutational invariance in PES fitting has been discussed in detail in the literature, and several approaches have been suggested. Murrell *et al.* recommended the use of symmetry variables, which they constructed by analogy with the irreducible representations of permutation groups.[55] In a recent review, Braams and Bowman described two other approaches,[12] which we summarize here. In the *monomial symmetrization* approach, the PES is expressed as a linear combination of polynomial basis functions in the Morse variables. Each basis function is constructed to ensure permutational invariance, by using the correspondence between permutations of the atoms and permutations of the variables. Xie and Bowman studied this method in detail and developed codes to perform the procedure.[13,14] In the

second approach, the PES is expressed as a sum of products of *primary invariant polynomials* and *secondary invariant polynomials* in Morse variables. The number of primary invariant polynomials is always equal to the number of variables. Braams and Bowman used results from the theory of polynomial invariants to show that, with a finite number of secondary invariant polynomials, all possible invariant polynomials up to an arbitrary order can be expressed in this way in the functional form of the PES. The authors used the computational algebra package MAGMA to assist in determining secondary invariant polynomials.[12,56] Bowman *et al.*, in their recent review, emphasize the importance of accounting for permutational symmetry in PES construction in general.[18] As examples, we note that the methods they describe were applied to PESs for NO_3 and OH_3,[57,58] among many other systems.

To incorporate permutational invariance into the L-IMLS method, we combined various ideas from these previously used approaches. First, we chose a set of six permutationally invariant polynomials $\mathbf{x} = (x_1, \ldots, x_6)$. These define a new six-dimensional coordinate system, which we call the *permutationally invariant Morse coordinate system*. We define the basis functions as simple monomials in those coordinates. Thus, the six coordinates play roles similar to those of the primary invariant polynomials described by Braams and Bowman,[12] although our method does not include corresponding elements that are analogous to the secondary invariant polynomials. Importantly, our goal was not to ensure that all possible permutationally invariant polynomials could be expressed in terms of basis functions. Because the fitting function is expressed as a weighted average of a large number of local fits, per Eq. (17), each one only needs to contain enough terms to adequately approximate a relatively small region of the PES. We will discuss the performance of our method in Sec. VI.

Ultimately, we used the following expressions to define the permutationally invariant Morse coordinate system. These were selected from among the permutationally invariant polynomials generated by the procedure described by Xie and Bowman:[13,14]

$$
\begin{aligned}
x_1 &= \frac{1}{12}\begin{pmatrix} q_1q_2 + q_1q_3 + q_1q_4 + q_1q_5 + q_2q_3 + q_2q_4 \\ +q_2q_6 + q_3q_5 + q_3q_6 + q_4q_5 + q_4q_6 + q_5q_6 \end{pmatrix}, \\
x_2 &= \frac{1}{4}\left(q_1q_2q_4 + q_1q_3q_5 + q_2q_3q_6 + q_4q_5q_6\right), \\
x_3 &= \frac{1}{24}\begin{pmatrix} q_1q_2(q_1+q_2) + q_1q_3(q_1+q_3) + q_1q_4(q_1+q_4) \\ +q_1q_5(q_1+q_5) + q_2q_3(q_2+q_3) + q_2q_4(q_2+q_4) \\ +q_2q_6(q_2+q_6) + q_3q_5(q_3+q_5) + q_3q_6(q_3+q_6) \\ +q_4q_5(q_4+q_5) + q_4q_6(q_4+q_6) + q_5q_6(q_5+q_6) \end{pmatrix}, \\
x_4 &= \frac{1}{4}\left(q_1q_2q_3 + q_1q_4q_5 + q_2q_4q_6 + q_3q_5q_6\right), \qquad (22) \\
x_5 &= \frac{1}{12}\begin{pmatrix} q_1q_3q_4 + q_2q_3q_4 + q_1q_2q_5 + q_2q_3q_5 \\ +q_2q_4q_5 + q_3q_4q_5 + q_1q_2q_6 + q_1q_3q_6 \\ +q_1q_4q_6 + q_3q_4q_6 + q_1q_5q_6 + q_2q_5q_6 \end{pmatrix}, \\
x_6 &= \frac{1}{3}\left(q_1q_2q_5q_6 + q_1q_3q_4q_6 + q_2q_3q_4q_5\right).
\end{aligned}
$$

Each of these six coordinates approaches zero, as a system approaches a geometry exhibiting only pairwise interactions. In this way, we maintain consistency with our strategy of fitting only the many-body component of the energy, as expressed in Eq. (18). Observe that x_1, x_2, and x_3 capture three-body interactions; if a geometry is dominated by pairwise interactions, then each of these coordinates will be small. Likewise, x_4, x_5, and x_6 capture four-body interactions; if a geometry is dominated by pairwise interactions and three-body interactions, then each of these coordinates will be small. Note that a polynomial such as $q_1 + q_2 + q_3 + q_4 + q_5 + q_6$ would not be a suitable choice for this scheme. This polynomial is permutationally invariant, but it does not have the asymptotic behavior required by our strategy of fitting only the many-body component of the energy. Finally, we note that the constant normalizing factors in Eq. (22) were added mainly to assist in interpreting coordinate values. As expected, they do not appear to have any significant impact on PES quality.

Using the coordinate system described by Eq. (22), the monomial basis functions were enumerated based on degree. It is a simple exercise to list all such monomials. There are six of degree one, 21 of degree two, 56 of degree three, and 126 of degree four. Thus, if we wished to use all basis functions of degree two or less, we would use a total of 27 basis functions. To use all basis functions of degree three or less or of degree four or less, we would use, respectively, 83 or 209. Note that a constant term is not allowed. Indeed, to maintain consistency with our strategy of fitting only the many-body component of the energy, all basis functions must vanish as four or more of the six internuclear distances approach infinity.

D. Permutational invariance and the distance metric

The weight functions of Eqs. (11) and (15) require a distance metric, which assigns a real number $d^2(\mathbf{q}_1, \mathbf{q}_2)$ to each pair of points $\mathbf{q}_1$ and $\mathbf{q}_2$. The distance metric plays a critical role in any method based on L-IMLS: it assigns a numerical measure of closeness to a pair of points, which is used both to construct the normal equations in Eq. (13) and to compute the weighted average of local fits in Eq. (17). A good distance metric should return a small value for a pair of points if and only if the two corresponding geometries are similar. In particular, if $\mathbf{q}_1$ and $\mathbf{q}_2$ are different ordered six-tuples, but $\mathbf{q}_1$ describes a geometry that can be obtained by a permutation of the atoms in the geometry described by $\mathbf{q}_2$, then $d^2(\mathbf{q}_1, \mathbf{q}_2)$ should be zero. The need to account for permutational invariance in this way makes the selection of a distance metric a subtle task.

Consider any two points $\mathbf{q}_1$ and $\mathbf{q}_2$, expressed in the raw Morse coordinate system. The most naïve approach is to use a simple two-norm of the vector $\mathbf{q}_1 - \mathbf{q}_2$:

$$
d^2\left(\mathbf{q}_1, \mathbf{q}_2\right) = \sum_{l=1}^{6} |q_{1,l} - q_{2,l}|^2 = |\mathbf{q}_1 - \mathbf{q}_2|^2 . \qquad (23)
$$

However, this definition ignores the permutational symmetry of the four-atom system: it will only return zero if $\mathbf{q}_1$ and $\mathbf{q}_2$ are identical ordered six-tuples, and not in the case that $\mathbf{q}_1$ and

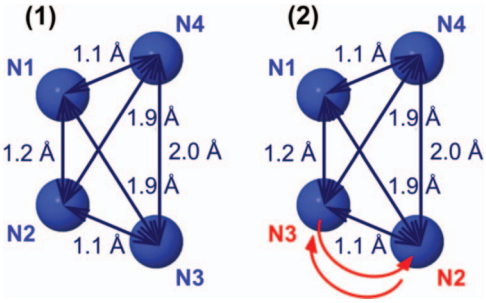

FIG. 2. Example of failure of the naïve distance metric. We consider two geometries, one of which is obtained by a permutation of the atoms of the other. Following the conventions in Eqs. (20) and (21) with $a = 1.0$ Å, the coordinates of geometry (1) are $\tilde{\mathbf{q}}_1 = (1.2, 1.9, 1.1, 1.1, 1.9, 2.0)$ Å and $\mathbf{q}_1 = (0.90, 0.45, 1.0, 1.0, 0.45, 0.41)$. The coordinates of geometry (2) are $\tilde{\mathbf{q}}_2 = (1.9, 1.2, 1.1, 1.1, 2.0, 1.9)$ and $\mathbf{q}_2 = (0.45, 0.90, 1.0, 1.0, 0.41, 0.45)$. Observe that the array $\mathbf{q}_2$ is a reordering of the array $\mathbf{q}_1$. The distance metric given by Eq. (23) yields $d^2(\mathbf{q}_1, \mathbf{q}_2) \approx 0.42 \neq 0$, despite the fact that the geometries are permutationally symmetric.

$\mathbf{q}_2$ are different but correspond to permutationally symmetric geometries. Figure 2 gives an illustration of this difficulty.

We develop a more sophisticated strategy as follows. There are a total of $4! = 24$ possible permutations of four identical atoms. Each of these 24 permutations corresponds to a permutation of the six raw Morse coordinates, according to the conventions of Eqs. (20) and (21). For any ordered six-tuple $\mathbf{q}_2$, let $\chi_k(\mathbf{q}_2)$ be the rearranged ordered six-tuple that corresponds to the kth permutation, where $k = 1, \ldots, 24$. Then we can define the distance metric as a minimum of two-norms over all such permutations:

$$d^2(\mathbf{q}_1, \mathbf{q}_2) = \min\left\{|\mathbf{q}_1 - \chi_k(\mathbf{q}_2)|^2 : k = 1, \ldots, 24\right\}. \quad (24)$$

The use of a minimum function in a distance metric for IMLS-based methods has been explored by other workers.[59] This distance metric meets the "zero distance condition" described above, and we were able to use it to obtain a continuous PES. However, the surface was not smooth: the use of the minimum function led to cusps in the surface, and it destroyed our ability to compute continuous, analytic gradients. Note that problems related to smoothness have also been observed in some potential energy surfaces constructed from splines.[60]

Consequently, we sought a function that has a behavior similar to the minimum function used in Eq. (24) but is also continuously differentiable with respect to the coordinates of the first point $\mathbf{q}_1$. A modified power mean can be used to satisfy both of these conditions.[61] We used

$$d^2(\mathbf{q}_1, \mathbf{q}_2) = \left(\sum_{k=1}^{24}\left(|\mathbf{q}_1 - \chi_k(\mathbf{q}_2)|^2\right)^{-p_d}\right)^{\frac{-1}{p_d}}, \quad (25)$$

where p_d is a power that defines the sharpness of the function. We set $p_d = 2$. Note that the minimum function is recovered in the limit $p_d \to \infty$. This distance metric proved effective, yielding an accurate and continuously differentiable surface. The performance of the method will be discussed in more detail in Sec. VI.

E. The cutoff radius correlation

In the fitting function evaluation step of L-IMLS, we calculate a weighted average of local fits from Eq. (17). The number of local fits is equal to the number of data points that lie within a hyper-sphere that is centered on the evaluation point $\mathbf{q}$ and has a radius $R(\mathbf{q})$. We denote this number by L, and we say that those points *lie within the cutoff radius*. The cost of an evaluation depends strongly on L. Thus, it is desirable to keep this number as small as possible, while still maintaining a highly accurate fit. Evidently, the number of points cannot be zero, since then no local fits would be used to evaluate the fitting function. One typically imposes a condition that L never falls below some threshold value $L_{\min}$ that ensures a desired level of accuracy.

The distribution of data points throughout the six-dimensional coordinate space depends on how those data points were originally obtained. For many chemical datasets, it is not true that the data points are approximately uniformly distributed throughout the relevant portion of space, nor would this property typically be attractive from a chemical point of view. Instead, some regions are populated with data points more densely than others. For example, regions of high density might correspond to equilibrium geometries, transition states, or reaction pathways. If a constant cutoff radius value was used in all evaluations, then the need to maintain $L \geq L_{\min}$ would force us to set R based on the lowest-density region of the coordinate space where calculations are needed. This, in turn, would degrade performance in the high-density regions, where an unnecessarily large number of local fits would be used in Eq. (17).

Several strategies for increasing computational efficiency in IMLS-based methods via the cutoff radius have been explored in the literature, as noted earlier in our discussion of the $s(z)$ function in Eq. (16). Two works describe ideas for varying the cutoff radius at different evaluation locations. Maisuradze *et al.* used a variable cutoff radius and a smooth damping function to improve efficiency in a six-dimensional implementation of IMLS.[49] More recently, Kawano *et al.* compared two cutoff strategies in IMLS, which they termed the *fixed radius cutoff* and *density adaptive cutoff* approaches. The later approach involved a sophisticated iterative scheme to determine the cutoff radius.[50] Also, note that the weight functions used in the work by Dawes *et al.* incorporate a density-adaptive scaling factor, but it was not used as a cutoff radius.[7] These researchers interpret their scaling factor as a type of *confidence radius*, citing earlier work by Thompson *et al.*[62]

We use a new statistical approach, which we call the statistically localized strategy. First, for each data point, we calculate a single *characteristic coordinate* q_{ch}, which is assumed to correlate in some way with the density of data points. Next, we construct a univariate polynomial *cutoff radius correlation* (CRC), which expresses the cutoff radius as a function of the characteristic coordinate. For an arbitrary location in six-dimensional space, the cutoff radius is obtained simply by calculating the characteristic coordinate and then evaluating the CRC. For our problem, we defined the characteristic coordinate q_{ch} by the following simple

expression:

$$q_{\mathrm{ch}} \equiv q_{\mathrm{ch}}(\mathbf{q}) = q_1^2 + q_2^2 + q_3^2 + q_4^2 + q_5^2 + q_6^2. \quad (26)$$

Then, the cutoff radius is expressed in the following form:

$$R\left(q_{\mathrm{ch}}\right) = \sum_{c=1}^{C_{\max}} \alpha_c q_{\mathrm{ch}}^{c-1}. \quad (27)$$

This value is squared for use in Eqs. (11) and (15). In this expression, $C_{\max}$ is the total number of terms in the CRC, and $\{\alpha_c\}$ are coefficients. We used $C_{\max} = 5$ for N_4. This choice and the choice of the characteristic coordinate given in Eq. (26) were based on numerical experimentation; a strategy was deemed effective if it was simple and it yielded a significant gain in computational efficiency.

We determine the coefficients $\{\alpha_c\}$ by a least squares procedure, performed as a pre-processing step before the local fit construction step. Our algorithm consists of the following steps. First, we choose a target value L_{tar} and a minimum allowable value $L_{\min}$ for the number of points within the cutoff radius associated with each data point. We then loop over all data points. For each point $\mathbf{q}_i$, we calculate two values: $R_{\mathrm{tar}}(\mathbf{q}_i)$, the value of the cutoff radius needed to include exactly L_{tar} data points within the cutoff radius, including $\mathbf{q}_i$ itself, and $q_{\mathrm{ch}}(\mathbf{q}_i)$, the characteristic coordinate for $\mathbf{q}_i$, given by Eq. (26). We then use a standard least squares method to fit a polynomial curve to the one-dimensional dataset R_{tar} versus q_{ch}. This gives an initial guess for the coefficients $\{\alpha_c\}$. Next, to ensure reasonable quality of the CRC, we again loop through all data points. For each point $\mathbf{q}_i$, we compute the cutoff radius $R(\mathbf{q}_i)$ using Eq. (27) and determine the number of points L that lie within it. If $L \geq L_{\min}$ for all data points, then the CRC is deemed valid. Otherwise, we increase the value of L_{tar} and repeat the entire procedure. In our implementation, we also used various simple weighting strategies in the least squares procedure to improve the fit of the CRC to the data R_{tar} versus q_{ch}.

For the tetranitrogen system, this statistical approach proved effective at increasing computational efficiency. For the final CRC, we used $L_{\min} = 3$ and $L_{\mathrm{tar}} = 59$, after numerical experimentation. The average number of local fits used in the fitting function evaluations was significantly lower when we used the CRC instead of a constant cutoff radius limited by the lowest-density region of the coordinate space, and high fitting accuracy was maintained. We will discuss the performance of the method in more detail in Sec. VI.

Finally, it is instructive to compare the statistically localized approach with the density adaptive cutoff method designed by earlier researchers.[49,50] In those prior studies, the cutoff radius was defined implicitly by a nonlinear equation (involving a sum over all data points), which was solved using an iterative scheme. In the present work, the cutoff radius is defined explicitly by Eqs. (26) and (27). Its computation during a PES evaluation is generally much less costly, since the coefficients $\{\alpha_c\}$ in Eq. (27) can be determined in advance. Likewise, it is straightforward to calculate the derivatives of Eqs. (26) and (27), a feature that facilitates the analytic evaluation of gradients. A disadvantage of the statistically localized approach is that it is does not provide a strict lower bound on the value of L; as we discussed above, the $L \geq L_{\min}$ condition is verified only at the data points themselves. Consequently, it is possible that this condition will fail to hold at a location far from any data points. By contrast, the density adaptive cutoff method does not present this risk, because it guarantees that L never falls below a user-specified value.[50]

F. Analytic gradients

As we have stated earlier, we were guided in our research by the desire to keep L-IMLS-G2 simple enough to allow gradients to be computed analytically, instead of resorting to finite difference approximations. Our method meets this goal. Indeed, it is a straightforward (though rather tedious) exercise to compute the derivatives of Eq. (17) with respect to each of the 12 Cartesian coordinates describing a four-atom system. This requires calculation of the derivatives of the weight functions given by Eq. (15), the basis functions, the distance metric given by Eq. (25), the cutoff radius correlation given by Eq. (27), the characteristic coordinate given by Eq. (26), the coordinate transformation from raw Morse coordinates to permutationally invariant Morse coordinates given by Eq. (22), the coordinate transformation from raw internuclear distance coordinates to raw Morse coordinates given by Eq. (21), and finally the coordinate transformation from Cartesian coordinates to raw internuclear distance coordinates. Note that the derivatives of the cutoff radius are particularly simple, since it is expressed as a univariate polynomial. We verified the correctness of our implementation of analytic gradients by demonstrating excellent agreement with numerical gradients, which we calculated using a central finite difference scheme.

IV. DATA TO BE FIT

We applied the L-IMLS-G2 method to a set of 16 435 data points for the tetranitrogen system. These were presented in previous work.[6] Note that we used the updated version of the dataset, which was discussed in the erratum[6] to the initial publication. Also note that unit conversion factors were provided with the dataset. For consistency, those same factors should be used in any analysis of the data.

V. CONSTRUCTING THE FIT

In this section, we discuss the procedures we used to construct and analyze a potential energy surface for the N_4 system using L-IMLS-G2. Our focus is on how we selected key parameters in the fitting approach. There are many ways to judge the quality of a potential energy surface. Typically, more than one method of judgment is needed to gain confidence that the surface is suitable for further work. In our research, we used two quantitative tests, a *fitting accuracy test* and a *cross validation test*, to measure the success of the fitting function, and we also used several qualitative tests.

In the fitting accuracy test, the L-IMLS-G2 fitting function is evaluated using Eq. (17) at each of the electronic

structure data points themselves. The evaluated energies are compared with the energies of the data points, which we call the *reference energies*. Statistics, such as the mean unsigned error, are then computed on the results. This is perhaps the simplest test that can be done to measure the success of the L-IMLS-G2 method. However, it would be dangerous to rely only on the fitting accuracy test to claim that a viable fitting function has been obtained. Indeed, recall that L-IMLS-G2 uses each of the data points as localization points for a set of N local fits; the method is explicitly designed to yield good approximations for the data points themselves. Thus, other tests are needed to ensure that surface quality is maintained at points in space that are not extremely close to a data point.

The cross validation test is one such test. Detailed descriptions can be found in Ref. 63. Here, we give a concise summary of the procedure. First, we select an integer K, and we divide the set of data points into K partitions of approximately equal size. For our research, we use $K = 5$; i.e., we use five partitions. (These are constructed by arranging the points in a chemically motivated order, then putting points 1, 6, 11, ... in the first partition, points 2, 7, 12, ... in the second partition, and so on.) Next we iterate over the partitions. For each partition I, we construct a set of local fits using only the data points in the other remaining partitions. Then, we use these local fits to evaluate the fitting function at each of the data points in partition I, and the evaluated energies are compared with the reference energies. In this way, the points in partition I form a *test set*, and the points in the other remaining partitions form a *training set*; the procedure is repeated using each of the partitions as a test set. The cross validation test gives a quantitative measure of the predictive capability of the fitting method: low errors will be returned if the fitting function accurately approximates data points that were not used in its construction.

We also devoted particular attention to the performance of L-IMLS-G2 along the four linear synchronous transit (LST) paths[64] that were considered[6] in the original electronic structure calculations. Further details of these paths can be found in the earlier paper,[6] and specification of the paths' endpoints can be found in the supplemental material.[65] A total of 30 electronic structure data points were obtained along these LST paths. To qualitatively evaluate the predictive capability of our method, we calculated the fitted energy along these paths in two cases. In the *full dataset* case, we included all points in the L-IMLS-G2 construction step. In the *reduced dataset* case, we included all points except the 30 LST points. For a high-quality PES, the fitted energies should closely match the electronic structure energies in both cases, and the curves from each case should be qualitatively similar. Such behavior would support the claim that the PES is accurate even in regions where electronic structure data was not obtained.

Armed with the quantitative fitting accuracy and cross validation tests and the qualitative test using the LST paths, we proceeded to explore different choices of the parameters in the L-IMLS-G2 method. Specifically, we needed to select values for p_v and ε_v in Eq. (11) and for p_u and ε_u in Eq. (15). These weight function parameters can have a dramatic effect on the quality of the fit. To reduce the size of the parameter space, we chose to set $p_v = p_u$ and $\varepsilon_v = \varepsilon_u$ in our work. (However, as we have noted above, this is not a necessary constraint.) We varied p_u from 1 to 4 and varied ε_u from 10^{-3} to 10^{-1}. We note the following qualitative trends from this search. Within reasonable bounds, larger values of p_u and smaller values of ε_u decreased fitting accuracy errors but increased cross validation errors. This is an expected result. Indeed, such changes sharpen the weight functions in Eqs. (11) and (15), which in turn forces the L-IMLS-G2 fitting function to be more strongly dominated at an evaluation point by the local fit associated with the nearest data point. Thus, the fitting function tends to become a better approximation to the data points themselves, at the expense of overall smoothness. The predictive capability of the fitting function suffers, as captured in the larger cross validation errors. We examined various one-dimensional and two-dimensional cuts through the six-dimensional PES to further assess accuracy and smoothness. Ultimately, based on these quantitative and qualitative tests, we chose $p_u = 3$ and $\varepsilon_u = 10^{-1}$ as recommended values. Note that we do not have strong reasons to believe that these choices are optimal for an arbitrary dataset. Rather, they seem to be good choices for the tetranitrogen dataset we considered, and they could be seen as reasonable initial guesses if applying the method to other data.

We also explored different values for the a parameter used to define the raw Morse coordinate system in Eq. (21). Within reasonable bounds, this variation did not appear to significantly affect error statistics from the fitting accuracy or cross validation tests. We chose the value $a = 0.85$ Å for our final fitting function. The error statistics tended to be much more sensitive to the choices of the weight function parameters described above than to the choice of a.

We also studied different values for the maximum order η of the basis functions in the local fit construction step. We found that a large value of η was not necessary to obtain a high quality PES. Indeed, since a PES based on L-IMLS techniques is constructed from a large number of local fits, it does not appear necessary to use a high-order polynomial for each one, especially when we remove the steep pairwise potential beforehand, per Eq. (18). Indeed, it may even be less accurate to use a large value of η. Errors in the fitting accuracy and cross validation tests were generally lower after η was increased from 2 to 3. However, for $\eta = 4$, the errors increased considerably for some points in the cross validation test. This may indicate that, at this value of η, L-IMLS-G2 is *over-fitting* the dataset. We use this language to refer to the situation in which the PES is extremely accurate near the data points used in its construction, but is of unacceptable quality in other regions of space. We therefore used $\eta = 3$ for the final fit. This corresponds to a total of 83 basis functions for each local fit.

A final consideration in our research concerned the possibility of redundancy in the N_4 dataset. The permutationally invariant distance metric in Eq. (25) gave us a mathematical tool to identify the closest pairs of data points in the six-dimensional weight function coordinate space. We noticed that the dataset included some pairs of points that were separated by a very small distance, indicating that the

TABLE I. Mean unsigned errors (MUEs) from the fitting accuracy test, for various values of the order η of the local fits. All energies and MUEs are in kcal/mol.

Data subset	Number of points	MUE $\eta = 2$	MUE Final fit	MUE $\eta = 4$
$E < 100$	516	0.33	0.17	0.084
$100 \le E < 228$	1556	0.57	0.32	0.20
$228 \le E < 456$	9238	0.68	0.41	0.28
$456 \le E < 1000$	1515	3.3	2.0	1.3
$1000 \le E$	323	2.9	1.5	0.80
All data	13148	1.0	0.60	0.39

corresponding geometries were nearly the same, except for permutation of the atoms. Building on the hypothesis that, in such cases, only one of the geometries should be necessary to build a sufficiently accurate L-IMLS-G2 PES, we added a capability to our code to systematically identify and rank the closest pairs of data points with respect to the distance metric in Eq. (25). For each pair up to a user-specified maximum, one of the data points in the pair could be excluded or *masked* from future consideration in the fitting procedure. After experimentation with this tool, we decided to use a 20% masking strategy. That is, we chose to exclude the top 20%, rounded down, of the electronic structure data points that were deemed "most redundant" by the permutationally invariant distance metric. We thus masked a total of 3287 data points out of the total 16 435, for a final recommended dataset of 13 148 data points. Errors from the fitting accuracy and cross validation tests were only slightly affected by this masking strategy, while the average number of local fits used for the fitting function evaluations decreased substantially. Thus, we considered this improvement in efficiency to be worthwhile, and all results shown in this paper used the masked dataset.

VI. FINAL RESULTS – ACCURACY AND EFFICIENCY OF THE FIT

Summarizing the outcomes of the investigations described above, the final potential energy surface for the tetranitrogen system used $p_v = p_u = 3$, $\varepsilon_v = \varepsilon_u = 10^{-1}$, $a = 0.85$ Å, $\eta = 3$, and a 20% masking strategy. A Fortran subroutine of the N_4 PES is in the POTLIB library.[66,67]

Table I shows results from the fitting accuracy test for this fit, and Table II shows results from the cross validation test. For comparison, we also include results from the corresponding test fits using $\eta = 2$ or $\eta = 4$ instead of $\eta = 3$. Mean unsigned errors (MUEs) are given in these tables; root-mean-square errors are provided in the supplemental material.[65] As expected by design of the method, the errors in Table I are quite small, typically a few tenths of a percent or less. High accuracy is also reflected in the cross validation test statistics in Table II, where errors are typically less than 1%. Furthermore, we note that the cross validation test may actually overestimate expected errors, in the sense that the final fit is based on 13 148 data points, while the fits used in this test are each based on only about 10 518 points. The possibility of over-fitting the original data by using $\eta = 4$, as discussed in the Sec. V, is especially evident in the cross validation MUEs for those data points with reference energies E greater than or equal to 1000 kcal/mol. Indeed, notice that the MUE for this data subset drops when η increases from 2 to 3 (the recommended value), but then rises substantially when η is further increased to 4. Observe that this increase in error in the cross validation test occurs even though the corresponding error in the fitting accuracy test decreases. This illustrates the finding, discussed in Sec. V, that using higher-order basis functions with L-IMLS-based methods does not necessarily increase overall surface quality.

TABLE II. Mean unsigned errors (MUEs) from the cross validation test with five partitions, for various values of the order η of the local fits. All energies and MUEs are in kcal/mol.

Data subset	Number of points	MUE $\eta = 2$	MUE Final fit	MUE $\eta = 4$
$E < 100$	516	0.44	0.26	0.13
$100 \le E < 228$	1556	0.89	0.59	0.43
$228 \le E < 456$	9238	1.2	0.98	0.90
$456 \le E < 1000$	1515	5.0	3.9	3.6
$1000 \le E$	323	7.3	6.5	11
All data	13148	1.7	1.4	1.4

TABLE III. Mean unsigned errors (MUEs) for planar and nonplanar geometries from the fitting accuracy test on the final fit. This table was generated from the same results used for Table I. All energies and MUEs are in kcal/mol.

	Planar		Nonplanar	
Data subset	Number of points	MUE	Number of points	MUE
$E < 100$	317	0.19	199	0.15
$100 \le E < 228$	903	0.36	653	0.27
$228 \le E < 456$	5325	0.38	3913	0.45
$456 \le E < 1000$	1013	1.7	502	2.4
$1000 \le E$	290	1.3	33	3.6
All data	7848	0.58	5300	0.63

Tables III and IV show additional statistics, which were gathered using the same test data used for Tables I and II, respectively. We divide the data points into two categories: those corresponding to planar geometries of the four nitrogen atoms and those corresponding to nonplanar geometries.

TABLE IV. Mean unsigned errors (MUEs) for planar and nonplanar geometries from the cross validation test with five partitions on the final fit. This table was generated from the same results used for Table II. All energies and MUEs are in kcal/mol.

	Planar		Nonplanar	
Data subset	Number of points	MUE	Number of points	MUE
$E < 100$	317	0.30	199	0.19
$100 \le E < 228$	903	0.68	653	0.47
$228 \le E < 456$	5325	0.99	3913	0.98
$456 \le E < 1000$	1013	3.9	502	3.9
$1000 \le E$	290	5.6	33	14
All data	7848	1.5	5300	1.2

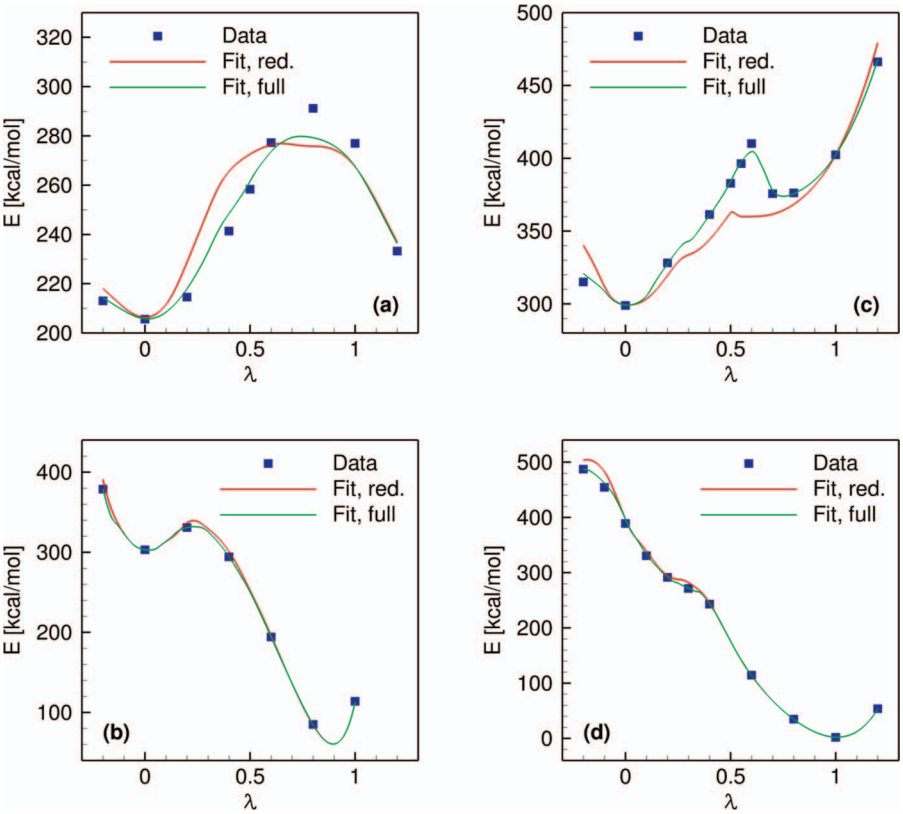

FIG. 3. Fitted energies along the four linear synchronous transit (LST) paths. For each path, a fit is considered based on the full dataset (full) and based on the reduced dataset (red.), which was obtained by eliminating the 30 LST points. As defined in previous work,[6] λ is the LST parameter corresponding to progress along the path. Further information about the endpoints of the paths can be found in prior work[6] and in the supplemental material.[65]

Out of the total 13 148 data points, 7848 correspond to planar geometries and 5300 correspond to nonplanar geometries. Notice that the MUEs associated with each category are comparable; errors associated with the nonplanar geometries are slightly higher for high-energy points but slightly lower for low-energy points. This provides further evidence that the fitting procedure is robust and is capable of representing the full dimensionality of the potential energy surface.

The accuracy attained by the present fit may be compared with that of an earlier fit[6] to essentially the same dataset. This previous research employed a global method based on permutationally invariant polynomials in Morse variables; it yielded a mean unsigned error over 16 435 data points of 4.1 kcal/mol. The local method used in the present article appears to be much more accurate, yielding overall MUEs of 0.60 kcal/mol and 1.4 kcal/mol from the fitting accuracy and cross validation tests, respectively. This is an expected result; as we discussed earlier in this paper, local methods are generally more accurate but more computationally expensive than global methods.

When fitting a potential energy surface for use in dynamics calculations to an electronic structure dataset, little is gained by fitting the data to a higher accuracy than that of the associated electronic structure method. Because the CASPT2 method used to generate the data fitted here was chosen based on the need to describe bond dissociation to highly open-shell species,[6] the data is less accurate than could be attained for a study of a closed-shell singlet system near its equilibrium geometry. Consequently, it is noteworthy that the accuracy of the L-IMLS-G2 PES, as reflected in the tests of Tables I–IV, is generally better than the expected accuracy of CASPT2 itself for energies at least up to the dissociation limit of about 228 kcal/mol.

Figure 3 shows results from the qualitative test of predictive capability based on the LST paths. We show the reference energies and the fitted energies corresponding to the full and reduced dataset cases, as discussed in Sec. V. The fitted energies from the full dataset case agree well with the reference energies along all four paths. Shifting our attention to the reduced dataset case, we see that excellent agreement is maintained along the second and fourth paths. A larger discrepancy is seen along the first and, especially, the third path: the fitted energies from the reduced case deviate both from the reference energies and from the fitted energies from the full case. This deviation is most pronounced in the third path near the cusp at about $\lambda = 0.6$, which is due to a state crossing. Ultimately, we consider the kind of discrepancies shown in Figure 3 (most serious at energies above the dissociation limit) to be acceptable. They are also a reminder of the inherent limitations of constructing the lowest adiabatic PES for a system exhibiting surface crossings (and avoided crossings along cuts). A similar difficulty was discussed in detail in the previous work.[6] Finally, we reiterate that a broader

TABLE V. Statistics on the number of local fits used for the evaluations in the fitting accuracy test. L_μ is the mean number of local fits and L_σ is the standard deviation of the number of local fits. All energies are in kcal/mol.

Data subset	Number of points	L_μ	L_σ
$E < 100$	516	450	119
$100 \leq E < 228$	1556	346	214
$228 \leq E < 456$	9238	460	409
$456 \leq E < 1000$	1515	216	159
$1000 \leq E$	323	110	94
All data	13148	409	368

measure of the ability to fit points away from the original data is provided by the cross validation test discussed above.

We can judge computational efficiency by examining the number of local fits used to evaluate the fitting function via Eq. (17). Recall that the final PES involves 13 148 local fits, and each one of these consists of 83 basis functions with coefficients determined via matrix normal equations of the form given by Eq. (13). By implementing the cutoff radius correlation in Eq. (27), the number of local fits used for any particular fitting function evaluation is typically much less than the total number available. Tables V and VI show the mean and standard deviation of the number of local fits used for each energy evaluation in the fitting accuracy and cross validation tests, respectively. Typically, less than 5% of the available local fits are used for any one case. Note that the results in Tables V and VI also reflect the increased efficiency granted by the 20% masking strategy to reduce data redundancy. For comparison, results from testing on the unmasked dataset are provided in the supplemental material.[65] Also note that the number of local fits is typically larger than the target value L_{tar} that was used to construct the cutoff radius correlation, as described in Sec. III E. This is a consequence of the variable density of data points across the coordinate space and of the least squares weighting schemes that were used to determine the coefficients $\{\alpha_c\}$ of the cutoff radius correlation in Eq. (27). The ultimate measure of success of the cutoff radius correlation was the gain in computational efficiency that it provided.

In Figure 4, we show selected one-dimensional cuts through the final PES. In each one, five of the six default coordinates are fixed. The energy and one component of the gradient with respect to the Cartesian coordinates are presented.

TABLE VI. Statistics on the number of local fits used for the evaluations in the cross validation test with five partitions. L_μ is the mean number of local fits and L_σ is the standard deviation of the number of local fits. All energies are in kcal/mol.

Data subset	Number of points	L_μ	L_σ
$E < 100$	516	360	96
$100 \leq E < 228$	1556	277	171
$228 \leq E < 456$	9238	368	327
$456 \leq E < 1000$	1515	173	128
$1000 \leq E$	323	88	76
All data	13148	328	295

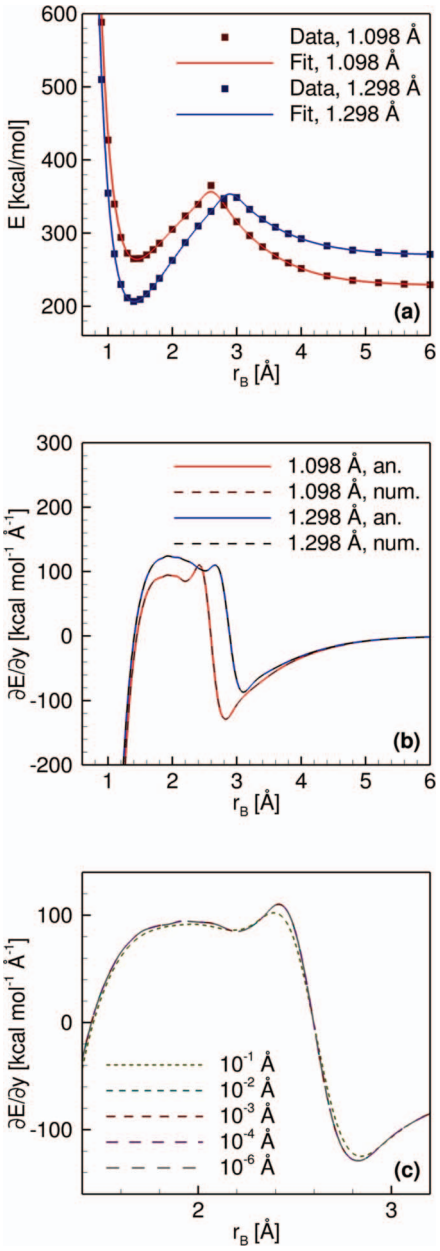

FIG. 4. One-dimensional cuts through the N_4 fitted PES, with fixed parameters $\theta_A = \theta_B = \phi = 90°$, $r_A = 1.098$ Å or $r_A = 1.298$ Å, and $d = 1.2$ Å. (a) shows excellent agreement between the fitted energies and the corresponding electronic structure data. (b) shows the fitted gradient component $\partial E/\partial y$ for the third nitrogen atom, computed analytically (an.) and numerically (num.) using central finite differences with a step-size of 10^{-3} Å. The agreement between the analytic and numerical gradients is excellent. (c) shows a closer view of numerical gradient curves corresponding to $r_A = 1.098$ Å, with various values of the finite difference step-size. The 10^{-3} Å case from (b) is shown again for clarity. As the step-size is decreased below about 10^{-2} Å, the curve does not appreciably change, supporting the claim that the numerical computation in (b) is converged.

Figure 4(a) shows that the agreement between the reference energies and the fitted energies is excellent, even where the PES exhibits sharp features. Figure 4(b) illustrates that the agreement between the analytical and numerical gradients is excellent; inspecting cuts like these was one of the techniques we used to verify that analytic gradients were properly implemented. Figure 4(b) also shows that the gradient curves

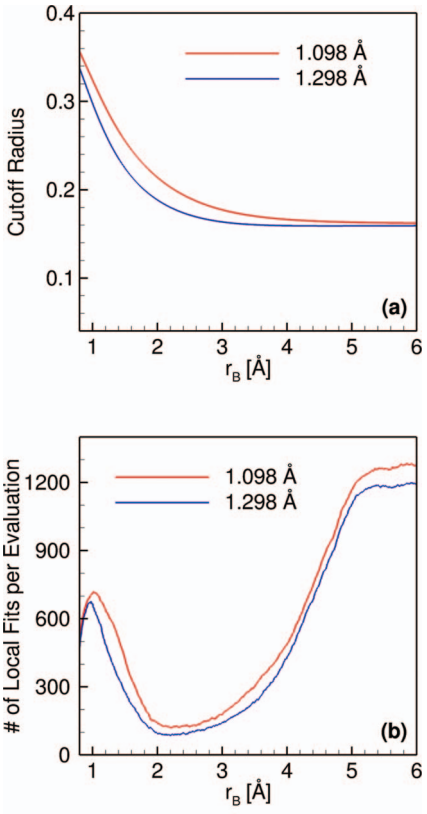

FIG. 5. Performance data corresponding to the one-dimensional cuts of Figure 4. (a) depicts the cutoff radius used in each evaluation, as given by Eq. (27). (b) depicts the number of local fits used for each evaluation.

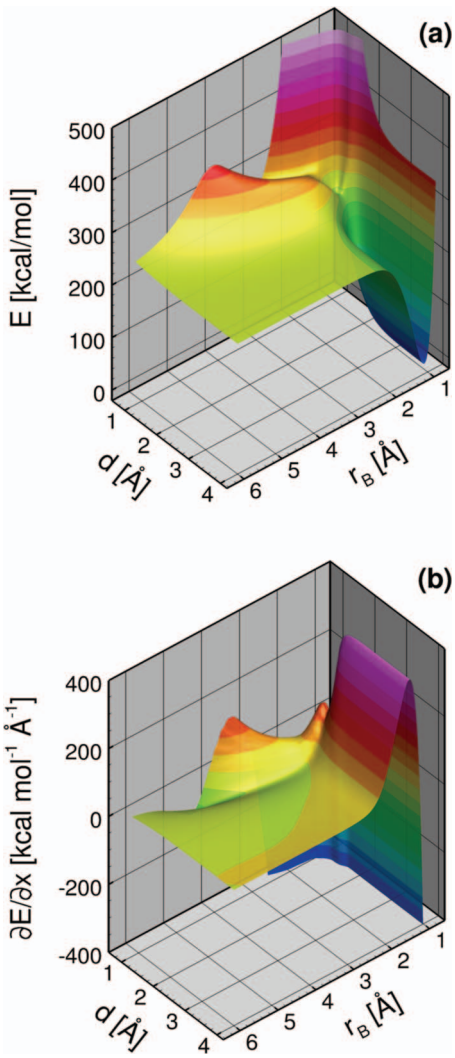

FIG. 6. Two-dimensional cuts with fixed parameters $\theta_A = \theta_B = 90°$, $\phi = 0°$, and $r_A = 1.098$ Å. (a) shows the fitted potential energy, and (b) shows the gradient component $\partial E/\partial y$ for the third nitrogen atom computed analytically from the fit. Each cut was constructed from 80 000 points.

are smooth. Notably, they do not exhibit erratic behavior near the data points themselves; such an undesirable result can occur if the weight functions in Eqs. (11) and (15) are excessively sharp due to poor choices of parameters. Figure 4(c) depicts the effect of the finite difference step-size on one of the numerical gradient curves; in particular, we show the numerical gradient curves computed using step sizes of 10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}, and 10^{-6} Å. The latter four curves are essentially indistinguishable. Generally, we found that the numerical results were converged for step-sizes of 10^{-3} Å or less.

Figure 5 shows the cutoff radius and the number of local fits used in each evaluation, for the same cuts that were considered in Figure 4. This provides one way to visualize the CRC of Eq. (27). The variation in the number of local fits is due to the CRC and to the non-uniform spatial distribution of data points. Importantly, observe that, although the number of local fits varies from about 100 to about 1200 along the cuts, the quality of the energy and gradient curves in Figure 4 does not appreciably change. This provides further evidence that the weight functions and the CRC are operating as intended.

Figure 6 shows examples of two-dimensional cuts through the final energy surface and through one of the gradient-component surfaces. Here, four of the six default coordinates are fixed, and two are varied. The figure shows that both surfaces are smooth.

VII. SUMMARY AND FUTURE WORK

In this paper, we discussed an improved L-IMLS method. After reviewing the least squares, weighted least squares, and L-IMLS methods in one dimension, we described a new version of L-IMLS for a six-dimensional system of four atoms. The method may be called statistically localized, permutationally invariant, local interpolating moving least squares fitting of the many-body potential, or, more simply, L-IMLS-G2. Our approach incorporates permutational invariance in both the basis functions and the weight functions. We treat pairwise interaction energy distinctly from many-body interaction energy, and we use a cutoff radius correlated to data point density to statistically account for the variable density of data points. All elements of the method were designed in such a way to allow gradients to be computed analytically. We applied the method to construct a highly accurate PES for a N_4 dataset developed in previous work.

Like other local fitting methods, L-IMLS-G2 yields very accurate surfaces with relatively low-order polynomial local

fits. Because an energy evaluation using L-IMLS-G2 depends only on nearby data points, the method is especially well-suited to PESs with rugged features, such as those used for modeling high-temperature vibrational energy exchange and dissociation. The method is typically both more accurate and more expensive than, for example, a global approach[6] that relies on a single, high-order polynomial. We discussed several important considerations in using the L-IMLS-G2 method. The quality of the fit is strongly affected by the choice of weight functions, coordinate systems, and the distance metric. We noted that it is possible to obtain a PES that is very accurate near the data points, but is of unacceptably poor quality in other regions; such behavior can be avoided by making proper choices of parameters. The tradeoffs between accuracy and smoothness can be analyzed quantitatively by using a cross validation test to measure predictive capability in regions not populated with data points.

Future research on this topic could focus on several key threads. First, we note that L-IMLS-G2 can be applied to any system of four atoms (i.e., even if those atoms are not identical), by appropriately accounting for the system's permutational symmetry. (While most of the paper focuses on the illustrative case of four identical atoms, L-IMLS-G2 can be adapted to other cases with minor changes. The method, as developed above, is sufficiently general so that other systems could be treated by defining proper analogues to the raw Morse coordinate system given by Eq. (21), the permutationally invariant Morse coordinate system given by Eq. (22), and the permutationally invariant distance metric given by Eq. (25). All core principles of L-IMLS-G2 would remain unchanged. Note, in particular, that all of Sec. III A remains general; it does not make the assumption of identical atoms. Presently, we are planning to adapt the method in this way to construct a PES for N_2O_2. The extension of the method to systems of more than four atoms is also feasible, although further modifications would be needed to account for the higher dimensionality of the PES. Obviously, serious consideration would need to be given to the resulting increase in computational cost.

In the near term, we plan to use the tetranitrogen L-IMLS-G2 PES to conduct quasiclassical trajectory calculations, to study vibrational energy exchange and dissociation in high-temperature environments typical of hypersonic flows. We are interested in further comparing the performance of the L-IMLS-G2 surface with that of the N_4 surface described in previous work,[6] which used a global method based on permutationally invariant polynomials. We wish to more carefully quantify the differences in the two surfaces for dynamics calculations, both in terms of computational cost and trajectory outcomes. Additional work could also focus on the exploration of alternative choices for the weight functions, the permutationally invariant Morse coordinate system, the distribution of data points, and the cutoff radius correlation scheme. Many different choices were explored for this research, but we have by no means exhausted all possibilities. In particular, we are interested in more advanced ways to improve performance by reducing the average number of local fits needed in the fitting function evaluation step. Finally, much deeper consideration can be given to the idea of more tightly integrating the fitting process with the electronic structure calculations and the trajectory calculations. For example, Castillo *et al.*[68] and Dawes *et al.*[69] have considered how to couple trajectory calculations with fitting to automatically identify areas of space for new electronic structure data. Nevertheless, we believe that the present second-generation L-IMLS method is already a step forward, and the enhancements we introduced here should be useful for fitting the potential energy surfaces of other systems.

ACKNOWLEDGMENTS

The authors are grateful to Yuliya Paukku, Antonio Varandas, Zoltan Varga, and Ke R. Yang for stimulating collaboration on overlapping projects. We also thank Bastian J. Braams and Joel M. Bowman for helpful discussions regarding permutationally invariant polynomials. Jason D. Bender was supported in this work by the U.S. Department of Energy Computational Science Graduate Fellowship (DOE CSGF) under Grant No. DE-FG02-97ER25308. Other work was supported by the U.S. Air Force Office of Scientific Research (AFOSR) under the Multidisciplinary University Research Initiative (MURI) Grant No. FA9550-10-1-0563. The views and conclusions contained herein are solely those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the University of Minnesota, DOE, AFOSR, or the U.S. Government.

[1] D. G. Truhlar and J. T. Muckerman, in *Atom-Molecule Collision Theory: A Guide for the Experimentalist*, edited by R. B. Bernstein (Plenum Press, New York, 1979), pp. 505–566.
[2] C. Park, J. Thermophys. Heat Transfer **3**, 233 (1989).
[3] D. Bose and G. V. Candler, J. Chem. Phys. **104**, 2825 (1996).
[4] D. Bose and G. V. Candler, J. Chem. Phys. **107**, 6136 (1997).
[5] I. Nompelis, G. V. Candler, and M. S. Holden, AIAA J. **41**, 2162 (2003).
[6] Y. Paukku, K. R. Yang, Z. Varga, and D. G. Truhlar, J. Chem. Phys. **139**, 044309 (2013); **140**, 019903 (2014) (erratum).
[7] R. Dawes, D. L. Thompson, Y. Guo, A. F. Wagner, and M. Minkoff, J. Chem. Phys. **126**, 184108 (2007).
[8] R. Dawes, D. L. Thompson, A. F. Wagner, and M. Minkoff, J. Chem. Phys. **128**, 084107 (2008).
[9] R. Dawes, A. F. Wagner, and D. L. Thompson, J. Phys. Chem. A **113**, 4709 (2009).
[10] R. Dawes, X.-G. Wang, A. W. Jasper, and T. Carrington, J. Chem. Phys. **133**, 134304 (2010).
[11] Y. Guo, I. Tokmakov, D. L. Thompson, A. F. Wagner, and M. Minkoff, J. Chem. Phys. **127**, 214106 (2007).
[12] B. J. Braams and J. M. Bowman, Int. Rev. Phys. Chem. **28**, 577 (2009).
[13] Z. Xie, Ph.D. thesis, Emory University, Atlanta, GA, 2008.
[14] Z. Xie and J. M. Bowman, J. Chem. Theory Comput. **6**, 26 (2010).
[15] D. G. Truhlar, R. Steckler, and M. S. Gordon, Chem. Rev. **87**, 217 (1987).
[16] G. C. Schatz, Rev. Mod. Phys. **61**, 669 (1989).
[17] A. Fernández-Ramos, J. A. Miller, S. J. Klippenstein, and D. G. Truhlar, Chem. Rev. **106**, 4518 (2006).
[18] J. M. Bowman, B. J. Braams, S. Carter, C. Chen, G. Czakó, B. Fu, X. Huang, E. Kamarchik, A. R. Sharma, B. C. Shepler, Y. Wang, and Z. Xie, J. Phys. Chem. Lett. **1**, 1866 (2010).
[19] N. Sathyamurthy and L. M. Raff, J. Chem. Phys. **63**, 464 (1975).
[20] J. M. Bowman, J. S. Bittman, and L. B. Harding, J. Chem. Phys. **85**, 911 (1986).
[21] J. Rheinecker, T. Xie, and J. M. Bowman, J. Chem. Phys. **120**, 7018 (2004).
[22] M. Patrício, J. L. Santos, F. Patrício, and A. J. C. Varandas, J. Math. Chem. **51**, 1729 (2013).

[23]A. J. C. Varandas, F. B. Brown, C. A. Mead, D. G. Truhlar, and N. C. Blais, J. Chem. Phys. **86**, 6258 (1987).
[24]A. J. C. Varandas, Adv. Chem. Phys. **74**, 255 (1988).
[25]A. J. C. Varandas and S. P. J. Rodrigues, J. Phys. Chem. A **110**, 485 (2006).
[26]T. Hollebeek, T.-S. Ho, and H. Rabitz, Annu. Rev. Phys. Chem. **50**, 537 (1999).
[27]T.-S. Ho and H. Rabitz, J. Chem. Phys. **119**, 6433 (2003).
[28]A. Chakraborty, Y. Zhao, H. Lin, and D. G. Truhlar, J. Chem. Phys. **124**, 044315 (2006).
[29]J. Ischtwan and M. A. Collins, J. Chem. Phys. **100**, 8080 (1994).
[30]M. A. Collins, Theor. Chem. Acc. **108**, 313 (2002).
[31]G. E. Moyano and M. A. Collins, Theor. Chem. Acc. **113**, 225 (2005).
[32]Y. Kim, J. C. Corchado, J. Villà, J. Xing, and D. G. Truhlar, J. Chem. Phys. **112**, 2718 (2000).
[33]T. V. Albu, J. C. Corchado, and D. G. Truhlar, J. Phys. Chem. A **105**, 8465 (2001).
[34]H. Lin, J. Pu, T. V. Albu, and D. G. Truhlar, J. Phys. Chem. A **108**, 4112 (2004).
[35]O. Bretscher, *Linear Algebra with Applications*, 3rd ed. (Pearson Prentice Hall, Upper Saddle River, NJ, 2005).
[36]P. Lancaster and K. Šalkauskas, *Curve and Surface Fitting: An Introduction* (Academic Press, London, 1986).
[37]P. Lancaster and K. Šalkauskas, Math. Comput. **37**, 141 (1981).
[38]G. G. Maisuradze, D. L. Thompson, A. F. Wagner, and M. Minkoff, J. Chem. Phys. **119**, 10002 (2003).
[39]G. G. Maisuradze and D. L. Thompson, J. Phys. Chem. A **107**, 7118 (2003).
[40]D. H. McLain, Comput. J. **17**, 318 (1974).
[41]D. H. McLain, Comput. J. **19**, 178 (1976).
[42]R. Farwig, Math. Comput. **46**, 577 (1986).
[43]R. Farwig, J. Comput. Appl. Math. **16**, 79 (1986).
[44]K. A. Nguyen, I. Rossi, and D. G. Truhlar, J. Chem. Phys. **103**, 5522 (1995).
[45]O. Tishchenko and D. G. Truhlar, J. Chem. Phys. **132**, 084109 (2010).
[46]S. T. Banks and D. C. Clary, Phys. Chem. Chem. Phys. **9**, 933 (2007).
[47]I. V. Tokmakov, A. F. Wagner, M. Minkoff, and D. L. Thompson, Theor. Chem. Acc. **118**, 755 (2007).
[48]A. Kawano, Y. Guo, D. L. Thompson, A. F. Wagner, and M. Minkoff, J. Chem. Phys. **120**, 6414 (2004).
[49]G. G. Maisuradze, A. Kawano, D. L. Thompson, A. F. Wagner, and M. Minkoff, J. Chem. Phys. **121**, 10329 (2004).
[50]A. Kawano, I. V. Tokmakov, D. L. Thompson, A. F. Wagner, and M. Minkoff, J. Chem. Phys. **124**, 054105 (2006).
[51]D. Levin, Math. Comput. **67**, 1517 (1998).
[52]Y. Guo, A. Kawano, D. L. Thompson, A. F. Wagner, and M. Minkoff, J. Chem. Phys. **121**, 5091 (2004).
[53]A. J. C. Varandas and J. N. Murrell, Faraday Discuss. Chem. Soc. **62**, 92 (1977).
[54]A. J. C. Varandas and J. N. Murrell, Chem. Phys. Lett. **84**, 440 (1981).
[55]J. N. Murrell, S. Carter, S. C. Farantos, P. Huxley, and A. J. C. Varandas, *Molecular Potential Energy Functions* (John Wiley & Sons, Chichester, 1984).
[56]H. Derksen and G. Kemper, *Computational Invariant Theory* (Springer-Verlag, Berlin, 2002).
[57]B. Fu, J. M. Bowman, H. Xiao, S. Maeda, and K. Morokuma, J. Chem. Theory Comput. **9**, 893 (2013).
[58]B. Fu, E. Kamarchik, and J. M. Bowman, J. Chem. Phys. **133**, 164306 (2010).
[59]R. Sivaramakrishnan, J. V. Michael, A. F. Wagner, R. Dawes, A. W. Jasper, L. B. Harding, Y. Georgievskii, and S. J. Klippenstein, Combust. Flame **158**, 618 (2011).
[60]J. C. Corchado, J. Espinosa-Garcia, and M. Yang, J. Chem. Phys. **135**, 014303 (2011).
[61]*Means and Their Inequalities*, edited by P. S. Bullen, D. S. Mitrinović, and P. M. Vasić (Reidel, Dordrecht, Holland, 1988).
[62]K. C. Thompson, M. J. T. Jordan, and M. A. Collins, J. Chem. Phys. **108**, 8302 (1998).
[63]S. Geisser, *Predictive Inference: An Introduction* (Chapman & Hall, New York, 1993).
[64]T. A. Halgren and W. N. Lipscomb, Chem. Phys. Lett. **49**, 225 (1977).
[65]See supplementary material at http://dx.doi.org/10.1063/1.4862157 for further details about the tests used to analyze potential energy surface quality.
[66]R. J. Duchovic, Y. L. Volobuev, G. C. Lynch, T. C. Allison, J. C. Corchado, D. G. Truhlar, A. F. Wagner, and B. C. Garrett, Comput. Phys. Commun. **144**, 169 (2002); D. G. Truhlar, A. F. Wagner, and B. C. Garrett, *ibid.* **156**, 319 (2004) (erratum).
[67]See http://comp.chem.umn.edu/potlib/ for the latest version of POTLIB that includes the present N_4 potential energy surface as well as that of Ref. 6. The PESs are given as Fortran subroutines. Each one takes as inputs the Cartesian coordinates of four nitrogen atoms and returns as outputs the energy and the gradient. Other codes to conduct the various auxiliary tests and procedures discussed in this paper are included in a larger program named FALCONS (Fitting Algorithm for the Local Construction Of eNergy Surfaces). FALCONS was created specifically for this research and is being prepared for wider release.
[68]J. F. Castillo, M. A. Collins, F. J. Aoiz, and L. Bañares, J. Chem. Phys. **118**, 7303 (2003).
[69]R. Dawes, A. Passalacqua, A. F. Wagner, T. D. Sewell, M. Minkoff, and D. L. Thompson, J. Chem. Phys. **130**, 144107 (2009).