Interpolating Moving Least-Squares Methods for Fitting Potential Energy Surfaces: Analysis of an Application to a Six-Dimensional System

Gia G. Maisuradze, Akio Kawano, and Donald L. Thompson

Department of Chemistry, Oklahoma State University, Stillwater, OK 74078

Albert F. Wagner

Chemistry Division, Argonne National Laboratory, Argonne, IL 60439

Michael Minkoff

Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, IL 60439

Abstract

The basic formal and numerical aspects of different-degree interpolated moving least-squares (IMLS) methods are applied to a six-dimensional potential energy surface (PES) of the HOOH molecule, for which an analytic ("exact") potential is available in the literature. We report the results of systematic investigations of the effects of weight function parameters, the degree and partial degree of IMLS, the number of data points allowed, and the optimal automatic point selection of data points up to full third-degree IMLS (TD-IMLS) fits. With partial reduction of cross terms and automatic point selection, the full 6D HOOH PES can be fit over a range of 100 kcal/mol to an accuracy of less than 1 kcal/mol with ~1350 *ab initio* points.

I. Introduction

The progress made during the past decade in electronic structure theory allows for direct use of *ab initio* forces in molecular dynamics simulations. Although most potential energy surfaces (PESs) have been obtained by empirical fitting, interest in using *ab initio* methods has grown in recent years as a result of the increased reliability of electronic structure calculations and enhanced computing capabilities. This direction is especially significant for PESs that describe chemical reactions because surfaces derived from relatively unsophisticated electronic structure calculations are notoriously inaccurate for describing bond breaking and formation. It is fairly routine now to perform high-quality *ab initio* calculations for hundreds to thousands of geometries. Many fitting methods have been studied in an effort to develop a broadly reliable approach to fitting *ab initio* points. Among the approaches used are cubic splines, least squares fitting, and hybrid methods.[1-3] The construction of a PES by these methods can be tedious, however, with the level of difficulty increasing rapidly with the size of the reacting system. Direct dynamics methods circumvent this problem, but they are computationally demanding, especially if high-level quantum chemistry calculations are used.

During the past decade the local fitting method introduced by Ischtwan and Collins,[4] which is based on modified Shepard interpolation, has become widely accepted. The unmodified Shepard method suffers from the flat-spot phenomenon; that is, the derivative of the interpolated surface is zero at each data point.[5] This difficulty is avoided, however, by using a Taylor expansion that includes the first and second derivatives at each data point. An attractive feature of the modified Shepard approach is its mathematical simplicity. It can be coupled with dynamical simulations to bias the fit, but the need for

derivatives, usually up to second order, cannot be readily or inexpensively satisfied by highest-level *ab initio* calculations.

Recently, we have helped introduce interpolating moving least-squares methods (IMLS)[6-9] for fitting PESs. The IMLS methods involve polynomials of any desired degree. The Shepard method is in fact a zero-degree IMLS method. Since the IMLS methods do not need gradients and Hessians, they are efficient for fitting PESs obtained by high-level *ab initio* calculations. As in both the Shepard and Ischtwan-Collins methods, the IMLS method uses a weighted least-squares fitting procedure where the weights are functions of both where the potential is to be evaluated (i.e., the evaluation point) and where the *ab initio* potential values have already been calculated. However, Ischtwan and Collin use the gradients and Hessians at each *ab initio* point to get a force field estimate of the potential at the evaluation point. They then use a Shepard fit to the set of potential estimations. Because the Shepard fit is of such low order, the weighted least-squares procedure implicit in the fit has an analytic solution that reduces to a weighted average and is thus trivial to calculate. In contrast, IMLS generalizes the Shepard approach to higher-order fits where the least squares procedure typically requires matrix algebra. Thus the procedure is not as computationally trivial as the Shepard method. However, IMLS directly fits the *ab initio* potentials, instead of their force field estimates of the potential at the evaluation point. Consequently, no gradients or Hessians at the *ab initio* points are required. In principle a zeroth-order IMLS fit (i.e., a Shepard fit) could also be applied to the *ab initio* potentials without the use of gradients or Hessians. However, Ischtwan and Collins[4] and we[7] have shown that in practice and in principle the derivative properties of this approach are poor and the resulting fitted

potential has undesirable characteristics. In principle, the weights could be discarded and a regular least squares procedure could be used once to define a fixed, namely, not *moving*, least-squares fit to the PES. However, many studies over the years have shown that such a procedure is accurate only if the *ab initio* data set is dense. The *moving* part to IMLS, that is, the locally varying weights, give a nonlinear character and accuracy to the fit that allows IMLS fits to retain high accuracy with relative sparse *ab initio* data sets.[6-9]

Our earlier work focused on features of IMLS for a 1D case[7] and a 3D case.[6] These studies highlighted the improved accuracy in values and derivatives obtainable with higher-degree IMLS. To improve the accuracy and efficiency of interpolation methods, we recently introduced a dual-level approach[10] that employs a zeroth-order PES as a reference surface.[8] This approach was tested on a 6D PES for HOOH with two interpolation methods: modified Shepard and second-degree IMLS. The results show that with the dual-level approach the IMLS and modified Shepard methods give comparably accurate fits for the same number of *ab initio* points but the IMLS requires only the values, not the gradients and Hessians.[8]

The present paper reports a study that is a continuation of that work. Here we explore the effects both of different degrees and mixed-degree polynomials in the IMLS and of *ab initio* point selection by automatic PES generation schemes for the 6D PES of the HOOH dissociation reaction. Different ensembles of data points were used, obtained by different sampling methods. In order to assess the global fitting error, the "*ab initio*" points were calculated from the analytic potential PCPSDE developed by Kuhn et al.[11] and the fitting error was determined by global samplings of the IMLS fit to the PCPSDE PES.

The rest of the paper is arranged as follows. Section II contains a brief review of IMLS methods. The weights and sampling that are used in the IMLS applications are discussed in Sec. III. The results for applications of various degree IMLS are presented in Sec. IV. The results for automatic surface generation are given in Sec. V. A summary and conclusions are given in Sec. VI.

II. Methods

Detailed descriptions of the IMLS methods are available in our previous papers[6,7] and earlier standard references,[5] so we will only briefly outline the approach here for 1D applications. The generalization to many dimensions is straightforward.

Suppose that m linearly independent functions $b_j(\mathbf{X})$ ($j=1,\ldots,m$) are given and defined on the surface and that we are given data values $f_1,\ldots,f_N$. The fitted surface is then a linear combination of these basis functions b_j,

$$u(\mathbf{X}) = \sum_{j=1}^{m} a_j(\mathbf{X})\, b_j(\mathbf{X}), \tag{1}$$

where $a_j(\mathbf{X})$ are the coefficients. To evaluate the fit of function $u(\mathrm{X})$ to the data values, we use the error functional

$$\mathrm{E}(u) = \sum_{i=1}^{N} w_i(\mathbf{X})\, [\mathrm{u}(\mathbf{X}) - \mathrm{f}(i)]^2 \ . \tag{2}$$

We assume that $m \leq N$, and we have introduced $w_i(\mathbf{X})$, a weight function that is a rapidly decreasing function of the distance $||\mathbf{X} - \mathbf{X}(i)||$. The solution to obtain the coefficients $a_j(\mathbf{X})$ follows the standard formulation of the normal equations for least-squares fitting:

$$B^T W(\mathbf{X}) B \mathbf{a}(\mathbf{X}) = B^T W(\mathbf{X}) \mathbf{f}, \qquad (3)$$

where **a** and **f** are column vectors, B is N×m matrix with elements $B_{ij} = b_i(\mathbf{X}(j))$, B^T is transpose matrix, and $W(\mathbf{X})$ is an N×N diagonal matrix whose element $W_{ii} = w_i(\mathbf{X})$. The solution $\mathbf{a}(\mathbf{X})$ to Eq. (3) provides the coefficients to the u at point **X**. Since B is a Vandermonde matrix, it tends to be ill-conditioned,[5] and for higher-degree IMLS we use either singular value decomposition (SVD) or QR-factorization for more stable factorization approaches than directly treating the normal equations.

To apply the IMLS method, one needs to define the coordinates **X**, the weights $w_i(\mathbf{X})$, and the basis functions $b_j(\mathbf{X})$. In our previous study of HOOH[8] the fitting coordinates were taken to be reciprocal interatomic distances $\mathbf{X}=1/\mathbf{R}$, whereas the weight function coordinates were the interatomic distances **R**. This hybrid coordinate system has been shown to be more efficient than only interatomic distances for related problems.[4,8] As in our previous studies, the weights have the form

$$w_i = \frac{1}{\| R - R(i) \|^n + \varepsilon} \quad , \qquad (4)$$

where n is a small positive integer and ε is a small positive real value that forces w_i to be finite at $\mathbf{R}=\mathbf{R}(i)$. Tests show that minor changes in the form of the weights have relatively little effect on performance and behavior.[6]

The basis set is taken to be a direct product of monomials in each degree of freedom in the vector **X**. Varying the powers of the monomial represented in the full basis set gives zero-degree (ZD), first-degree (FD), second-degree (SD), and third-degree (TD) IMLS. As the dimension of the system and the degree of IMLS increase, the number of the basis functions increases dramatically, mostly because of cross terms where the basis functions contains the products of monomials of different nonzero powers of different coordinates. For example, the total number of basis functions for SD-IMLS in the 6D case is 28, but if cross-terms are removed, the number of basis functions drops to 13. The TD-IMLS with and without cross-terms has 84 and 19 basis functions, respectively. Since the cost of an IMLS evaluation goes as the square of the number of basis functions, for higher-degree IMLS the cross-terms are the principal cost in the calculation.

III. Characterization of Weights and Samples

A. Sampling

Because we use an analytical PES, we can test the accuracy of the IMLS fits. We employ two different data sets drawn from the analytic PES: Set 1 is a restricted data set that is the collection of "*ab initio*" points used in the fitting, and Set 2 is a much larger set of points used for the evaluation of the global RMS errors of the fits.

The sampling scheme plays an important role in the efficiency of fitting multidimensional PESs. Previously[8] we used efficient microcanonical sampling (EMS)[12]

for both Set 1 and Set 2 and a symmetrized function as the "exact" function. For Set 1, 89 symmetry distinct predetermined data points were selected to cover the low-energy region of the PES, that is, the equilibrium and the vicinity of the intrinsic reaction coordinates (IRC). We added 10 more sets of data points sampled by EMS with an upper bound to the energy of 100 kcal/mol but with no restriction on the total angular momentum. All the atoms were moved in Cartesian coordinates for each Markov step with a step size of $0.5a_0$ for an acceptance/rejection ratio of approximately unity. In order to reduce the correlation of the sampled points, one point was picked from the Markov sequence every one-hundredth step. The O–O distance was restricted to $r_{OO}<6a_0$ during the walk. The singularity problem that occurs for planar geometries was avoided by using the "distortion" technique proposed by Yonehara et al.[13] For Set 2 we used 10 different ensembles of 5,000 data points sampled by the EMS method.

For the present study, we tested three other sampling methods in addition to EMS: a purely random sampling (RANDOM), a combination of EMS and RANDOM sampling (COMB), and an unequally-spaced grid of data points (GRID). We used the same 89 symmetry distinct predetermined data points used on our early study[8] for Set 1 in all the sampling methods. However, we did not symmetrize the "exact" potential; rather, we employed the "exact" potential as given in Ref. 11. The RANDOM method adds 10 different sets of data points obtained by Monte Carlo sampling with the same upper bound on E. In order to avoid trapping in the RANDOM sampling, one point was picked from the Markov sequence every one-thousandth step. In the COMB scheme, we added 10 different sets of data points, half picked by EMS and half picked by RANDOM sampling. In Set 2 for RANDOM sampling we used 10 different ensembles of 5,000 data

points selected by Monte Carlo, whereas in COMB we used 10 different ensembles of 5,000 data points, half sampled by EMS and RANDOM techniques.

In the GRID method we augmented the 89 predetermined points for Set 1 with points selected in a grid built from a geometric progression:

$$\begin{bmatrix} r_1^{(n1)} \\ r_2^{(n2)} \\ \vdots \\ r_6^{(n6)} \end{bmatrix} = \begin{bmatrix} f^{n1} & & & 0 \\ & f^{n2} & & \\ & & \ddots & \\ 0 & & & f^{n6} \end{bmatrix} \cdot \begin{bmatrix} r_1^{(0)} \\ r_2^{(0)} \\ \vdots \\ r_6^{(0)} \end{bmatrix} , \tag{5}$$

where f>1, each n_i is an integer, r_1 is the H_1-H_2 distance, r_2 is H_1-O_1, r_3 is H_1-O_2, r_4 is O_1-H_2, r_5 is H_2-O_2, r_6 is O_1-O_2, and

$$\begin{bmatrix} r_1^{(0)} \\ r_2^{(0)} \\ r_3^{(0)} \\ r_4^{(0)} \\ r_5^{(0)} \\ r_6^{(0)} \end{bmatrix} = \begin{bmatrix} 4.45a_0 \\ 1.82a_0 \\ 1.82a_0 \\ 1.82a_0 \\ 1.82a_0 \\ 2.75a_0 \end{bmatrix} . \tag{6}$$

We selected the points such that E < 100 kcal/mol and $r_6^{(0)} < 6a_0$. For Set 2 we used only the geometric progression to select points but with values of f closer to unity to produce a finer grid. We calculated the RMS error for values of f that produced from ~3,000 to ~25,0000 points in Set 2, all with E < 100 kcal/mol. Relative to the largest Set 2, we found differences of 2–9% in the RMS errors for ~5,00 points in Set 2, differences of

0.6–1.5% for ~40,000 points, and no differences for greater than 70,000 points. In the calculations presented below, we used 40,192 data points for Set 2 generated by f = 1.1108.

These sampling methods are compared in Fig. 1, where we show the distributions of data points as a functions of energy for 5,000 points select by EMS, RANDOM, and COMB and for 5,134 points selected by GRID. A cut-off in energy of 100 kcal/mol was used in all the sampling methods. As the results in Fig. 1 show, EMS sampling mainly covers the mid-range energies region of the PES and only partially includes points in the high- and low-energy ranges. The RANDOM sampling method covers mainly the high-energy region, only partially covers the mid-range of energies, and poorly defines the low-energy region. The GRID and RANDOM methods provide similar coverages for both high- and mid-energy regions, but the GRID method provides much greater coverage in the low-energy region. The COMB method provides comparable low-energy coverage and in effect averages the differences between the EMS and RANDOM methods in the mid- to high-energy regions. The results in Fig. 1 do not include the 89 symmetry distinct predetermined data points from the low-energy region that significantly improve the coverage of all methods in that region. Because the low-energy region has small spatial extent, it is expected that the 89 data points along with the sampling seen in Fig. 1 are sufficient to determine this area.[8]

By definition, Set 1 and Set 2 essentially encompass all the parts of the PES that are chemically interesting for HOOH $\rightarrow$ OH+OH dissociation and association reactions. Only extraordinary high energies (>100 kcal/mol) or exceptionally large OH+OH separations (>6 a_o) are excluded. Thus a satisfactory RMS error implies that the

spectroscopy and the dissociation/association reactivity of the HOOH system can be computed with confidence. Of all the calculated results presented in this paper, Set 2 (defining the RMS error) is much larger than Set 1 (the *ab initio* set). Hence, both interpolation and extrapolation conditions are found in evaluating the IMLS potential at the coordinates in Set 2. The mixture of these conditions depends on the sampling methods used; nonetheless, the RMS test should be a broad measure of the reliability of IMLS fits.

B. Weight Parameters

In a study of 1D fitting[7] we investigated the dependence of the accuracy of different degrees of IMLS on the weight function parameters ε and n, where we used an underlying grid for Set 1 and Set 2. To determine the dependence for multidimensional PESs and compare it to the 1D case, here we examined RMS errors of different degree IMLS as functions of ε and n. As in the 1D case, Set 1 and Set 2 were constructed by the GRID method and do not include the predetermined 89 data points in Set 1, although for other applications discussed below the predetermined points were used. The total number of data points N is 3,200 (f = 1.1758).

Figure 2 demonstrates the dependence of the RMS error of E on ε for FD-, SD-, and TD-IMLS fits for a fixed value of n. As in the 1D case,[7] when ε is too large, the accuracy of all fits is degraded. As ε decreases, the RMS error reaches a minimum that persists essentially for all further decrease in ε. Also, increasing the degree of IMLS for fixed ε decreases the RMS error. For TD-IMLS an RMS error of ~1 kcal/mol requires $\varepsilon < \sim 10^{-2}$. We observed in the 1D case[7] that with increasing N the range of optimal values of

ε becomes smaller; therefore, in the results for larger N, discussed below, we used $\varepsilon = 10^{-24}$, which guarantees optimal performance of the weights for any tractable value of N.

Figure 3 shows the dependence of RMS error for first derivatives on ε for FD-, SD-, and TD-IMLS. These results are similar to those found for 1D.[7] Since the range of values of the first derivatives is much larger than the range of energy values, the scale of Fig. 3 is much larger than the scale of Fig. 2. As in the 1D case the higher-degree IMLS have a less-pronounced minimum (almost impossible to see on figure).

Figure 4 illustrates the behavior of RMS error of energy for FD-, SD-, TD-IMLS as a function of n for a fixed value of ε. As in the 1D case[7] the RMS error dependence on n is much less severe than that on ε. Comparing the RMS error behavior as a function of n for 1D and 6D PESs shows that the range of optimal n for the 6D PES starts from slightly larger values of n ($n = 8$) than in the 1D case ($n = 6$) and depends only weakly on N. The behavior of the RMS errors for the derivatives is similar to that shown in Fig. 4 for energy.

IV. Various Degrees of IMLS

A. SD-IMLS Results

In Sec. III.A we defined the RANDOM, COMB, and GRID sampling methods that we have investigated. Here, we discuss the effect of the sampling method on the accuracy of SD-IMLS fits; the results are summarized in Table I for 12 different combinations of (Set 1, Set 2). We report the global RMS error and the uncertainty in that error due to the underlying statistical basis of the EMS, RANDOM, and COMB schemes (i.e., the variances in the 10 ensembles used to obtain the RMS error). This

information is given for Set 1 *N* values from 189 to 6,489, which include the 89 predetermined points. Set 2 comprises 50,000 points sampled by EMS, RANDOM, and COMB and 40,192 points sampled by GRID.

The first nine rows of Table I are the combinations of (Set 1, Set 2) possible from the three statistical methods EMS, COMB, and RANDOM. The first, fourth, and seventh rows show the diagonal combinations, for example, (COMB,COMB). For each value of *N*, the RMS error increases in the order of EMS, COMB, and RANDOM. This ordering is expected based on the results in Fig. 1 because this is the order in which higher-energy points are included in the samplings. For each *N*, the off-diagonal combinations, for example, (RANDOM,COMB), display RMS errors that generally increase from the diagonal value in Set 1 and Set 2. When Set 1 is obtained by EMS, the order of the increase is (EMS,EMS), (EMS,COMB), and (EMS,RANDOM). When Set 1 is obtained by RANDOM, the order of the increase is similar, namely, (RANDOM,RANDOM), (RANDOM,COMB), and (RANDOM,EMS). When Set 1 is obtained by COMB, it is unclear which Set 2 is the more dissimilar, EMS or RANDOM. The results in Table I show that the error decreases for (COMB,EMS), probably because EMS weakly samples high-energy values.

Combinations involving GRID are given in the last three rows of Table I. The diagonal (GRID,GRID) combination for most of the values of N produce RMS errors that generally fall between those for the (COMB,COMB) and (RANDOM,RANDOM) diagonal combinations, although for the smallest and largest number of data points (GRID,GRID) gives the lowest and biggest RMS errors, respectively. Off-diagonal combinations, for example, (COMB,GRID), gives larger the RMS errors compared to the

corresponding diagonal contributions–to an extent generally larger than that found between other off-diagonal/diagonal comparisons.

It is difficult to make general recommendations concerning which sampling combinations to use, other than to use a diagonal combination. The particular use of the fitted PES will dictate the most appropriate diagonal combination to use. The higher-energy portion of the PES has the largest spatial extent in multidimensional systems, but at the same time low-energy regions, including equilibrium points and reaction paths, have a large influence on the dynamics. The GRID approach is more likely to maintain a certain density of points per unit spatial "area" on the surface and in this way balance high- and low-energy demands on a PES representation. This sampling method is further highlighted by the results discussed below.

Independent of the behavior of RMS error with sampling techniques, all the results in Table I show a decline in the SD-IMLS error with N. Diagonal combinations produce ~0.65 to ~1.25 kcal/mol RMS errors for on the order of 6500 *ab initio* points.

B. Higher-Degree IMLS

The RMS error versus N for FD-, SD-, and TD-IMLS fits for (GRID,GRID) and (COMB,COMB) samplings are shown in Figs. 5 and 6, respectively. These results are taken from Table I supplemented by an additional calculation at N = 12889 for Set 1. In general the behaviors of the RMS errors for FD-, SD-, and TD-IMLS are similar to that for the 1D case[7] for which increasing the number of data points and the degree of IMLS systematically improves the fit. The overall the dependence on the sampling combination is weak. Both sampling methods show only the TD-IMLS method reaching ~1 kcal/mol RMS error and for values of N that exceed ~3,000 points. The FD-, SD-, and TD-IMLS

RMS errors in Figs. 5 and 6 show for larger values of N an approximately linear dependence on N on a log-log scale. This is a power law behavior that settles in at higher values of N as the IMLS degree increases. Fits show the power increases from about -1/4 for FD-IMLS to about -1/3 for TD-IMLS. Note that the number of *ab initio* points per dimension goes as N^6. Thus the reduction in RMS error goes as the number of points per dimension raised to the -1.5 to -2 power. In the 1D case[7] a power law convergence of RMS error also applies at large values of N. As in HOOH, the inverse powers involved increase with the degree of IMLS. For FD-IMLS to TD-IMLS, the range of powers for the 1D case are about –1.75 to –3.0, a range that largely overlaps the HOOH range. The 1D and 6D results suggest that higher degree IMLS RMS error converges for any dimensional system with some higher than unit inverse power of the number of points per dimension.

As the dimensions of the PES increase, a full higher-degree IMLS fit is increasingly dominated by cross-terms: direct product basis functions where the monomials in at least two coordinates have nonzero powers. For a 6D PES, an SD-IMLS has 28 basis functions of which 15 are cross terms. A TD-IMLS has 84 basis functions of which 65 are cross terms. Since the time to solution for an IMLS evaluation goes as the square of the number of basis functions, the elimination of negligible cross-terms can have a major effect on the computation time. In Figs. 5 and 6, the RMS error versus N is plotted for SD- and TD-IMLS with completely eliminated cross-terms (SD-IMLS^{reduced} and TD-IMLS^{reduced}). The RMS errors for SD-IMLS^{reduced} and TD-IMLS^{reduced} are very close to each other in both sampling methods. This implies somewhat disappointing results for TD-IMLS^{reduced} because TD-IMLS for most values of N has a noticeably lower RMS

error. Going to TD-IMLS^{reduced} considerably reduces the quality of the fit. For SD-IMLS and SD-IMLS^{reduced}, the penalty for eliminating cross-terms is not particularly severe. One possible reason for the impact of cross-term elimination on fit quality is that the optimal values of the weight parameters (ε, n) change with the details of cross-term retention. However, test variations in weight parameters about the values used in Figs. 5 and 6 produce no change in the results.

A better explanation for the results in the figures is that chemical intuition would suggest that not all cross-terms are comparable and negligible. Rather, only certain classes of cross-term may be insignificant depending on the nature of the PES. We have investigated this explanation, and the resulting RMS errors are in Table II (for SD-IMLS derived results) and Table III (for TD-IMLS derived results) for three values of *N* with (GRID,GRID) sampling. (The results in these tables were obtained with an approximation of a variable cutoff radius, which is discussed fully in the Appendix. The basis of this approximation is that *ab initio* points far from the point of PES evaluation have negligible influence on the IMLS result because the weights decay so rapidly with distance. The elimination of the influence of the far points by a cutoff radius retains the accuracy of fit and noticeably reduces CPU time.) The first two rows in both tables are, respectively, for the full-degree IMLS fits and for the reduced fits, where the use of variable cutoff slightly changes the RMS errors displayed in Fig. 5 and Table I. The remaining rows are the results for elimination of certain classes of cross-terms. The third row in each table eliminates all cross-terms except not containing r_6, on the supposition that coupling to the HO–OH reaction coordinate r_6 is the only substantial effect cross-terms should retain. The results in the third row of each table indicate degradations in the

RMS error from the full fit of less than 20% for both SD- and TD-IMLS. (For TD-IMLS, this partial elimination of cross-terms dramatically improves the RMS error for the lowest value of *N*.) This partial elimination produces results much superior to the reduced fits at the cost of the retention of 5 of 15 (SD-IMLS) or 25 of 65 (TD-IMLS) cross-terms.

With the retention of only a few more terms beyond those involving r_6, important angular dependencies can be retained. A zeroth-order potential of HOOH can be written in internal coordinates as

$$V \approx V_\phi(\alpha_1) + V_\phi(\alpha_2) + V_{OH}(r_2) + V_{OH}(r_5) + V_{OO}(r_6) + V_\tau(\tau), \qquad (7)$$

where α_1 is the angle between H_1O_1 and O_1O_2, α_2 is the angle between H_2O_2 and O_2O_1, and τ is a dihedral angle. In our interatomic coordinates the angles α_1 can be described by r_2, r_6, and r_3 and α_2 by r_5, r_6, and r_4. Thus, cross-terms in interatomic coordinates within each triplet express α_1 and α_2 dependencies. For SD-IMLS, of the 10 cross-terms discarded for not having an r_6 dependence, only two have to be retained to also include a full (α_1, α_2) dependence. For TD-IMLS, 12 of the 40 terms discarded have to be retained. The fourth row of each table shows the results. For both cases, for the two lower values of N, the RMS error is below the RMS error for the full fit. For the highest value of *N*, the RMS error is degraded by only ~4–9% from the full fit.

The success of eliminating all cross-terms except those involving r_6, α_1, and α_2 suggests eliminating all terms except those involving α_1 and α_2. Such a strategy is more indiscriminate, resulting in more terms being eliminated. The results are in the fifth row of each table. For both SD- and TD-IMLS, comparing rows three, four, and five, one

clearly sees that retention of terms that describe α_1, and α_2 is somewhat more important than retention of terms that describe r_6.

Unlike SD-IMLS fits, TD-IMLS contains enough terms of enough degrees that one can consider eliminating terms based on degree alone. The last two rows in Table III are the results of a purge of either second-degree or third-degree terms from those terms retained to represent r_6, α_1, and α_2. Purging second-degree terms produces an excellent result, within 3% of the results in the fourth row for no purge. Purging third-degree terms produces a generally unsatisfactory result.

Overall, the results in Table II and III provide the guidance necessary to match the fit to the task. Time to solution goes as the square of the number of terms in the fit. The total number of terms in a full SD-IMLS fit is 28 and in a full TD-IMLS fit is 84. In the calculations of the next section we will feature SD-IMLS and TD-IMLS, in which all the cross terms are eliminated not containing r_6, α_1, and α_2. From the fourth columns in Tables II and III, that will result in 20 terms for SD-IMLS and 56 terms for TD-IMLS, with minimal if any degradation in the RMS error for the full fit. The evaluation costs will be reduced by ~50% for SD-IMLS and by about ~40% for TD-IMLS. If less accuracy can be tolerated, bigger savings in evaluation costs can be achieved.

V. Automatic Point Selection

It is important to generate an accurate PES with the fewest number of data points. In our 1D study[7] we introduced an automatic point selection strategy that significantly reduced the number of points needed to reach a fit of given accuracy. Since multidimensional surfaces deal with a large number of data points, the application of this

kind of strategy is critical. The basis of our automatic point selection scheme is that, for fixed N, IMLS fits of different degrees are most different from one another where the PES is poorly defined but are essential identical to each other and the true PES in the near vicinity of an *ab initio* point. Thus, additional *ab initio* points should be calculated where the contending IMLS fits of different degrees are maximally different. In the present study we use as contending fits SD- and TD-IMLS with partially reduced cross-terms (fourth rows in Tables II and III) as bases sets to select one additional *ab initio* point at a time. As in Tables II and III, the variable cutoff is used in the calculations. As in Figs. 2–6 and all the tables, the same Set 2 of 40,192 points is used to evaluate the RMS error. In principle, new *ab initio* data points could be selected from Set 2. However, it is not practical to survey such a dense set of points to extract the point where contending IMLS fits are maximally different. Thus, in addition to Set 1 (the included *ab initio* points for the current fit) and Set 2, we define a Set 3 with a grid of 6,489 points from which we select the new *ab initio* points to be added one at a time to Set 1.

The results are expressed in Fig. 7. This figure shows the GRID RMS error of the partially reduced TD-IMLS fit where the points are automatically selected as described above from a starting collection of 189, 289, 489, 889, and 1689 *ab initio* points. The figure also shows the RMS errors of TD-IMLS calculated at 189, 289, 489, 889, 1,689, 3,289, 6,889, and 12,889 *ab initio* points selected via the GRID method. The results in the figures display two important properties. First, no matter how many seed points one begins with, the RMS errors for large N all fall on essentially the same curve. Second, the "universal" curve has essentially the same power law dependence as the GRID selection scheme for the *ab initio* points. However, the prefactor is reduced by

approximately 50%. In effect, automatic point selection halves the RMS error for larger values of N. The result is a TD-IMLS fit that represents with a 1 kcal/mol RMS error a full 6D PES everywhere below 100 kcal/mol based on only ~1350 *ab initio* points. The SD-IMLS RMS error derived from the same automatic point selection has a very similar relationship to the SD-IMLS RMS error derived from GRID sampling; in other words, the automatic point selection RMS error is about half that of the grid method.

The results in Fig. 7 do not address the issue of how to terminate the automatic point selection. In an actual application, we do not know the true PES and cannot compute the RMS error. All we know is information about the SD- and TD-IMLS fits. In Fig. 8 we plot as a function of N the RMS difference between SD-IMLS and TD-IMLS fits. Also in the figure as a function of N is the RMS difference of TD-IMLS and the true PES. This last curve is information that in an actual application we cannot know. We seek a similarity between SD-TD information and TD-exact information that allows us to estimate the later by knowledge of the former.

Several obvious ways of representing SD-TD information fail to track the dependence of the TD-IMLS RMS error with N. The maximum difference between SD-IMLS and TD-IMLS is an oscillatory function of N that generally decreases more rapidly with N than the TD-IMLS RMS error does. The RMS difference between SD-IMLS and TD-IMLS on Set 3 is a generally smooth function of N but it also decreases more rapidly with N than the TD-IMLS RMS error does. The underlying problem with both approaches is that Set 3 is selected to be significantly smaller than Set 2 (on which TD-IMLS RMS error is measured) for reasons of practicality. When almost all the points in

Set 3 have been selected for *ab initio* points, SD-TD differences by any measure will be minimal. However, the TD-IMLS RMS error will not.

The successful representation of SD-TD information that tracks TD-IMLS RMS error is to periodically calculate the RMS difference of SD-IMLS and TD-IMLS on the full Set 2. If this were done with every addition of an *ab initio* point, the computational cost would be excessive. However, we are not calculating the RMS difference for *ab initio* point selection but rather for termination of the selection procedure. A calculation of the RMS difference, say, every increase in N by 100 points, will allow termination with a 100-point precision. The ability of the RMS difference of SD-IMLS from TD-IMLS on Set 2 to track TD-IMLS RMS error is displayed in Fig. 8 for Set 3 sizes of 1,689, 3,289, and 6,489 points. All the results are for an initial seed Set 1 of 489 points. For each Set 3, there is a pair of curves with the upper curve displaying the RMS difference and the bottom curve displaying the RMS error. Although there is more initial raggedness in the comparison at small values of N, at the larger value of N the curves track one another with the RMS difference being a 10% to 15% overestimation of the TD-IMLS RMS error. The figure also shows that the smaller the Set 3, the larger the RMS error that selecting all the points in Set 3 will obtain. This is obvious. Set 2 was selected to be the size that it is because quadrupling the size of Set 2 did not appreciably change the TD-IMLS RMS error. Of course, if all the points in Set 2 were selected for *ab initio* points, the TD-IMLS RMS error would be extremely small.

Another consideration about Fig. 8 concerns the general lack of inverse power dependence on N. Generally after N exceeds about half the size of Set 3, the TD-IMLS RMS error flattens out on a log-log plot and thus loses its inverse-power dependence. At

the same time, selecting all the points in Set 3 and then doubling the Set 3 size produces results as in Fig. 6 where, on a coarser scale, an inverse-power dependence still appears to hold. In other words, automatic point selection out of a Set 3 will degrade the inverse-power dependence at large enough values of N as well as limit the final accuracy that can be obtained. Nonetheless, automatic point selection one point at a time out of a Set 3 can be terminated to a desired accuracy knowing only RMS differences between SD-IMLS and TD-IMLS on a Set 3 large enough to converge the TD-IMLS error with the exact PES. Figures 7 and 8 show that such a procedure can arrive at RMS errors of less than 1 kcal/mol with about half the number of points required by a systematic increase in grid density.

IV. Conclusions

In the context of a 6D application to HOOH, we have presented the basic formal and numerical aspects of IMLS methods of different degree with total, completely reduced, and partially reduced cross terms. We have included details of the weights and the sampling procedures used to select *ab initio* points and estimate RMS fitting errors. The major conclusions of this study are as follows:

- The RMS fitting error converges with N, the number of *ab initio* points, in an inverse-power-law fashion with powers that increase with the degree of the IMLS fit. For FD-IMLS to TD-IMLS the inverse powers range from 1.5 to 2 with respect to $N^{1/6}$, the average number of data points per dimension. These results are consistent with our earlier 1D study.[7]

- Higher-degree IMLS fits are dominated by cross terms that contribute in only a minor way to the quality of the fit. Neglect of all cross terms that do not include the approximate reaction coordinate (r_{OO}) produces RMS error increases of less than 15% for SD-IMLS and TD-IMLS. The retention of the cross terms related to r_{OO}, α_1, and α_2 angles gives better or almost the same results as full SD- and TD-IMLS. Moreover, TD-IMLS with elimination of second-order cross-terms along with all cross terms that do not include the reaction coordinate and angles fits PES as good as full TD-IMLS.
- Automatic point selection that optimally selects additional *ab initio* points approximately halves the RMS error relative to blind point-selection by changing grid increments or by Monte Carlo sampling. The automatic point selection scheme is that first developed and tested in 1D applications. Over a 100 kcal/mol range, 1 kcal/mol RMS error can be reached for the full 6D HOOH PES by a TD-IMLS fit of ~1,350 automatically selected *ab initio* points. Reliable termination procedures of automatic point selection allow preselected fit accuracy.

Future studies will examine the application of IMLS in classical trajectory simulations, to higher-dimensional systems, and to systems where gradient information is available.

Acknowledgments

This work was supported by the U.S. Department of Energy, Office of Basic Energy Sciences, Division of Chemical Sciences, and the Mathematical, Information, and Computational Sciences Division subprogram of the Office of Advanced Scientific Computing Research, Office of Science, U.S. Department of Energy under Contract No. W-31-109-Eng-38 (Argonne) and Contract No. DE-FG02-01ER15231 (OSU).

Table I. Comparisons of the RMS errors in the energy calculated by SD-IMLS for different sampling methods, Set 1 and Set 2 (see text). The results were obtained by using 10 different ensembles for each pair. Average values with deviations (in kcal/mol) are listed. Only in the (GRID;GRID) case was used one pair of Set 1 and Set 2.

(Set 1; Set 2)	189	289	489	889	1689	3289	6489
(EMS;EMS)	7.89 ±1.23	4.11 ±0.58	2.48 ±0.22	1.60 ±0.03	1.20 ±0.04	0.87 ±0.05	0.67 ±0.05
(EMS;COMB)	10.09 ±1.59	5.45 ±0.55	3.53 ±0.07	2.46 ±0.07	2.15 ±0.35	1.72 ±0.27	1.37 ±0.11
(EMS;RANDOM)	12.40 ±2.48	6.55 ±0.69	4.53 ±0.39	3.12 ±0.09	2.64 ±0.19	2.20 ±0.13	2.00 ±0.35
(COMB;COMB)	9.78 ±0.99	5.42 ±0.55	3.66 ±0.58	2.36 ±0.17	1.70 ±0.08	1.30 ±0.09	0.99 ±0.07
(COMB;EMS)	9.24 ±1.25	4.92 ±0.56	3.28 ±0.51	1.91 ±0.06	1.38 ±0.06	1.00 ±0.05	0.74 ±0.03
(COMB;RANDOM)	10.61 ±1.19	5.98 ±0.80	4.29 ±0.98	2.80 ±0.41	2.13 ±0.14	1.59 ±0.13	1.16 ±0.05
(RANDOM;RANDOM)	10.41 ±1.99	6.75 ±0.92	4.15 ±0.25	2.75 ±0.30	2.10 ±0.18	1.49 ±0.08	1.16 ±0.07
(RANDOM;COMB)	10.96 ±1.30	6.81 ±1.16	4.29 ±0.25	2.79 ±0.31	2.16 ±0.09	1.52 ±0.05	1.31 ±0.17
(RANDOM;EMS)	11.74 ±1.37	7.17 ±1.25	4.47 ±0.36	2.96 ±0.33	2.26 ±0.08	1.60 ±0.05	1.37 ±0.06
(GRID;GRID)	7.68	4.72	3.72	2.53	2.00	1.59	1.25
(GRID;COMB)	8.53 ±0.19	7.16 ±0.13	3.65 ±0.09	3.20 ±0.04	2.77 ±0.07	1.99 ±0.04	1.44 ±0.08
(COMB;GRID)	35.33 ±7.80	17.77 ±4.10	12.73 ±3.84	5.38 ±0.58	4.30 ±0.64	2.96 ±0.30	2.24 ±0.23

Table II. The RMS error in the energy (in kcal/mol) for SD-IMLS with all cross-terms and with completely and partially reduced cross terms.

Omitted Cross-Terms	Number of Omitted Cross-Terms	289	889	3289
None	0	4.72	2.50	1.51
All	15	5.37	3.20	2.11
All not containing r_6	10	5.08	2.90	1.82
All not containing α_1, α_2 and r_6	8	4.32	2.53	1.58
All not containing α_1, α_2	9	4.30	2.52	1.60

Table III. The rms errors of energy in kcal/mol for TD-IMLS with all cross terms and with completely and partially reduced cross-terms.

Omitted Cross-Terms	Number of Omitted Cross-Terms	289	889	3289
None	0	6.12	1.80	1.03
All	65	6.24	3.46	2.23
All not containing r_6	40	4.39	2.15	1.20
All not containing α_1, α_2 and r_6	28	3.77	1.84	1.13
All not containing α_1, α_2	34	4.03	1.86	1.15
All not containing α_1, α_2, r_6 + all remaining second-order cross-terms	35	3.21	1.85	1.09
All not containing α_1, α_2, r_6 + all remaining third-order cross-terms	58	3.92	2.28	1.48

Appendix

The data points farther from an evaluation point R have a smaller contribution to the fitting because of the reduced weight function. The cost of the IMLS procedure is linear in the number of data points explicitly included in the least-squares procedure. Therefore we can decrease the IMLS computational cost by removing distant data points before solving the normal equations.

One removal method is to exclude data points of whose distances from R are farther than a fixed threshold r_{cut}. However, in applications where the number of data points N will change (such as in automatic point selection discussed in Sec. V), r_{cut} can be a significant function of N. An additional problem is that excluding the ith point whose distance $\delta_i(R)$ from the evaluation point is larger than r_{cut} in effect truncates the weight function for the ith point to zero. This introduces a discontinuity in the fitted energy and derivatives whenever a data point crosses r_{cut}. Both problems can be solved[9] by a weight modified by a damping function S defined by

$$\varsigma(R, r_{cut}) = \sum_i S[\delta_i(R), r_{cut}], \tag{A1}$$

where S is a smooth damping function of the form

$$S(\delta,r) = \begin{cases} 1 - \tanh^3\left(\alpha\left(1 - \dfrac{\delta^2}{r^2}\right)\right), & if\ \ 0 \le \delta < r, \\ 0 & if\ \ r \le \delta, \end{cases} \tag{A2}$$

where α>0 and where r_{cut} is determined by

$$\varsigma(R, r_{cut}) = \varsigma_{const}, \tag{A3}$$

where ς_{const} is a given parameter. The damping function in effect fully counts a data point whose distance δ from the evaluation point at R is well within r_{cut}. At the boundary of r_{cut} and beyond, the damping function has the correct derivative and limiting properties to avoid any discontinuities. $\varsigma(R, r_{cut})$ becomes a fractional count of the number of data points "smoothly" contained within r_{cut}. Selecting a fixed ς_{const} value as input ensures that no matter what the value of N, sufficient data will be available for the IMLS procedure.

Once $S(\delta, r_{cut})$ has been determined, the new weights have the form

$$w_i'(R) = w[\delta_i(R)]S[\delta_i(R), r_{cut}] \tag{A4}$$

with derivatives

$$\frac{\partial w_i}{\partial R_k} = S_i \cdot \frac{\partial \delta_i}{\partial R_k}\frac{\partial w}{\partial \delta_i} + w_i \cdot \left(\frac{\partial \delta_i}{\partial R_k}\frac{\partial S_i}{\partial \delta_i} + \frac{\partial r_{cut}}{\partial R_k}\frac{\partial S_i}{\partial r_{cut}}\right), \tag{A5}$$

where S_i stands for $S[\delta_i(R), r_{cut}]$ and $\partial r_{cut}/\partial R_k$ is

$$\frac{\partial r_{cut}}{\partial R_k} = -\frac{\sum_i \frac{\partial \delta_i}{\partial R_k} \frac{\partial S_i}{\partial \delta_i}}{\sum_i \frac{\partial S_i}{\partial r_{cut}}} \quad . \tag{A6}$$

Figure Captions

Fig. 1 The distribution of data points sampled by EMS (---), RANDOM (⋯), COMB (-·-), and GRID (—) methods as a function of energy.

Fig. 2 The RMS error for the potential energy in kcal/mol versus ε for FD-IMLS $(\cdots)$, SD-IMLS (–), and TD-IMLS (–·–); n = 10, N = 3200. The RMS error is based on 40,192 points.

Fig. 3 Same as Fig. 2 only for the RMS error of the first derivatives in kcal/mol/a_0.

Fig. 4 Same as Fig. 2 only for the RMS error versus n for $\varepsilon = 1\times10^{-24}$. See the text for details.

Fig. 5 The RMS error for the potential energy in kcal/mol versus N for (GRID;GRID) sampling. The points are connected by straight lines for clarity; the dotted line denotes FD-IMLS, dash-dot line denotes SD-IMLS, dash-double-dot line denotes TD-IMLS, solid line is SD-IMLS^{reduced}, and dash line is TD-IMLS^{reduced}. The weight function parameters are n = 10 and $\varepsilon = 1\times10^{-24}$.

Fig. 6 Same as Fig. 5 only for (COMB;COMB) sampling. See the text for details.

Fig. 7 The RMS error versus N for TD-IMLS fits that vary with the selection of *ab initio* points. The solid line with circles is for GRID selection and the solid lines with some minor noise are for automatic point selection from an expanding collection of *ab initio* seed points. See text for details.

Fig. 8 The RMS difference between SD-IMLS and TD-IMLS on Set 2 (symbols) and the TD-IMLS RMS error on Set 2 (solid line) versus N for three different grids for Set 3. The RMS difference is ▲, ■, and ● for Set 3 grids of 1689, 3289, and 6489 points, respectively. The associated TD-IMLS RMS error in each case is shown by the curves below the corresponding points.

Figure 1

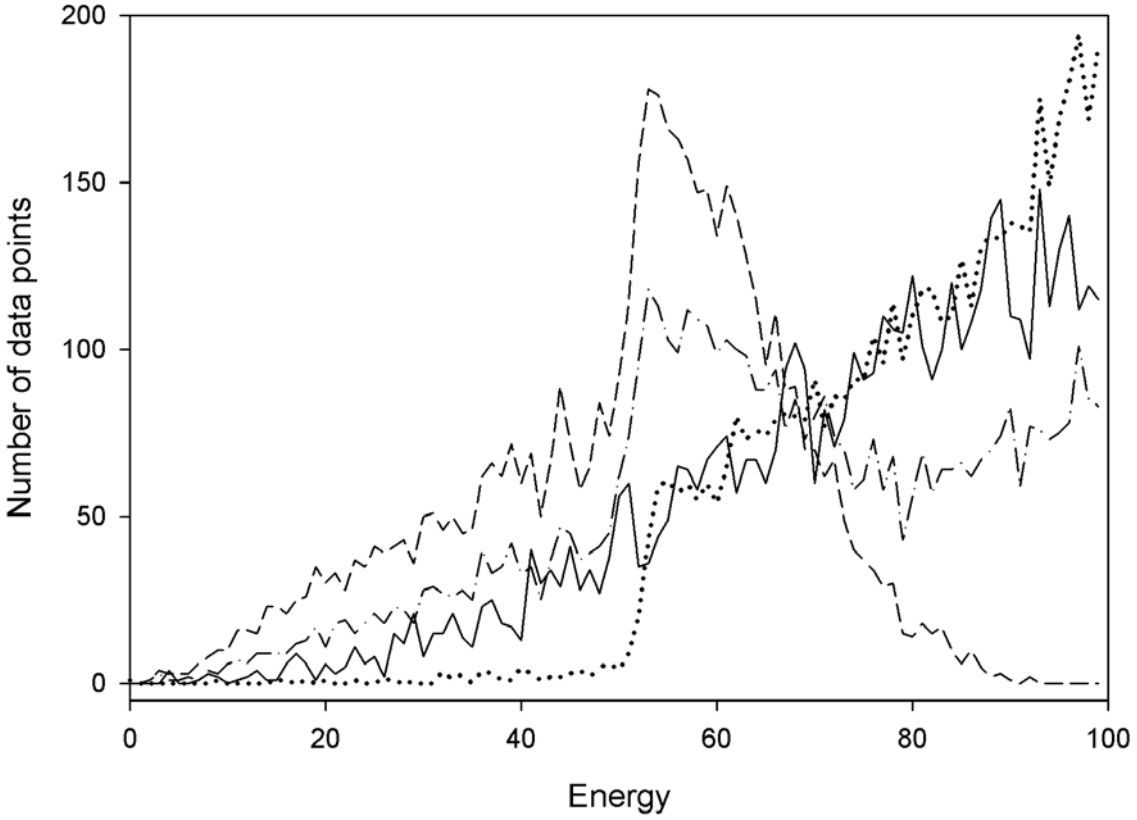


Figure 2

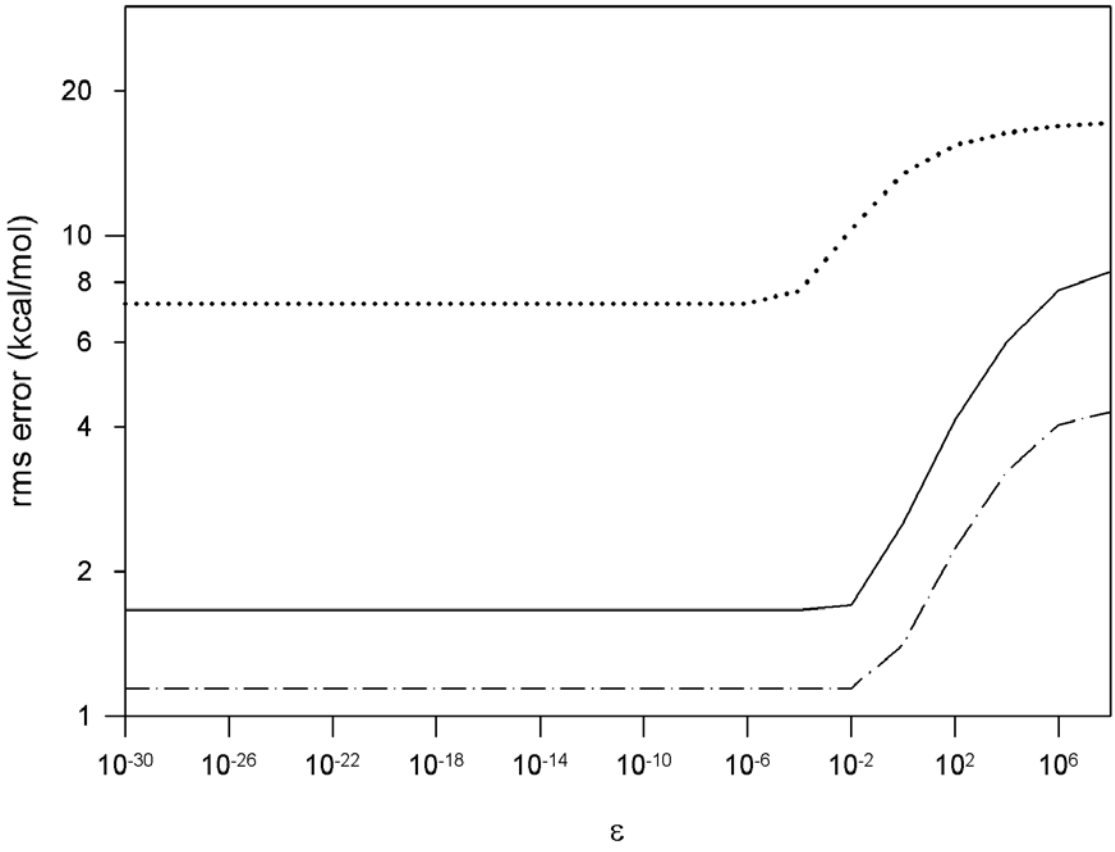

rms error (kcal/mol)
20
10
8
6
4
2
1
10-30
10-26
10-22
10-18
10-14
10-10
10-6
10-2
102
106
ε

Figure 3

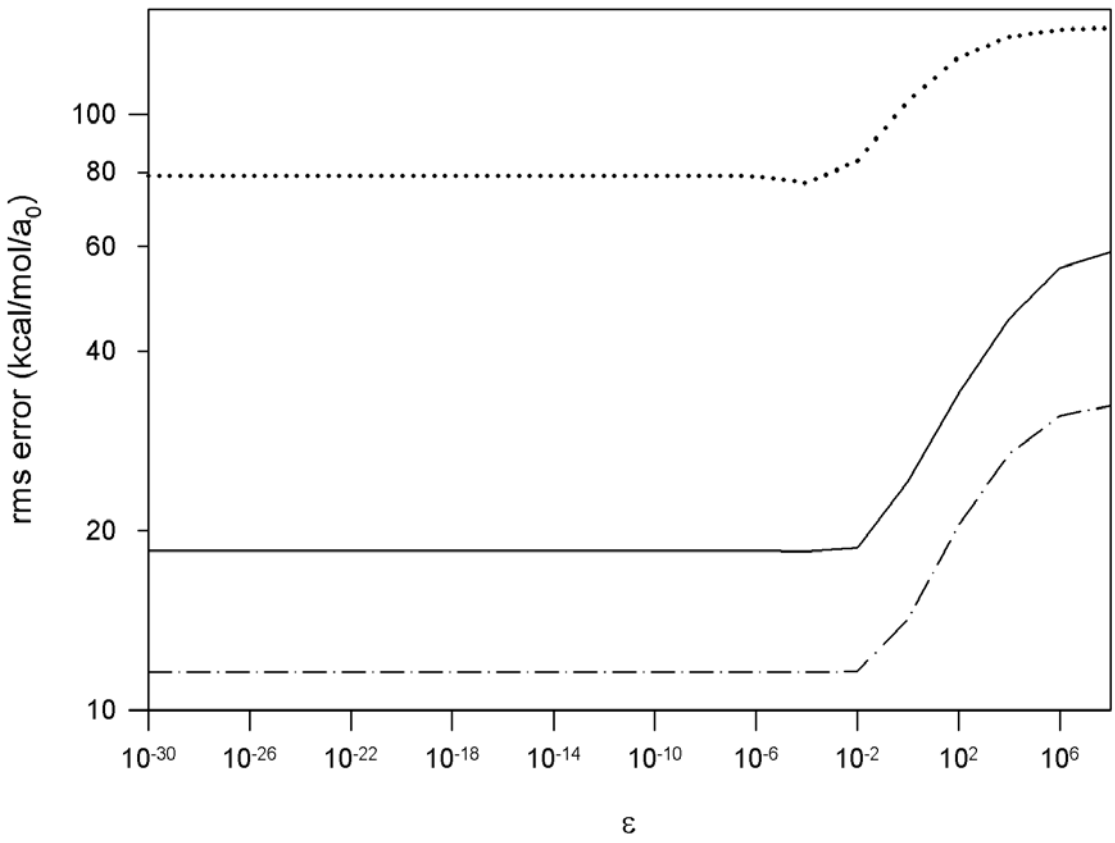

100
80
60
40
20
10
rms error (kcal/mol/a0)
10-30
10-26
10-22
10-18
10-14
10-10
10-6
10-2
102
106
ε

Figure 4

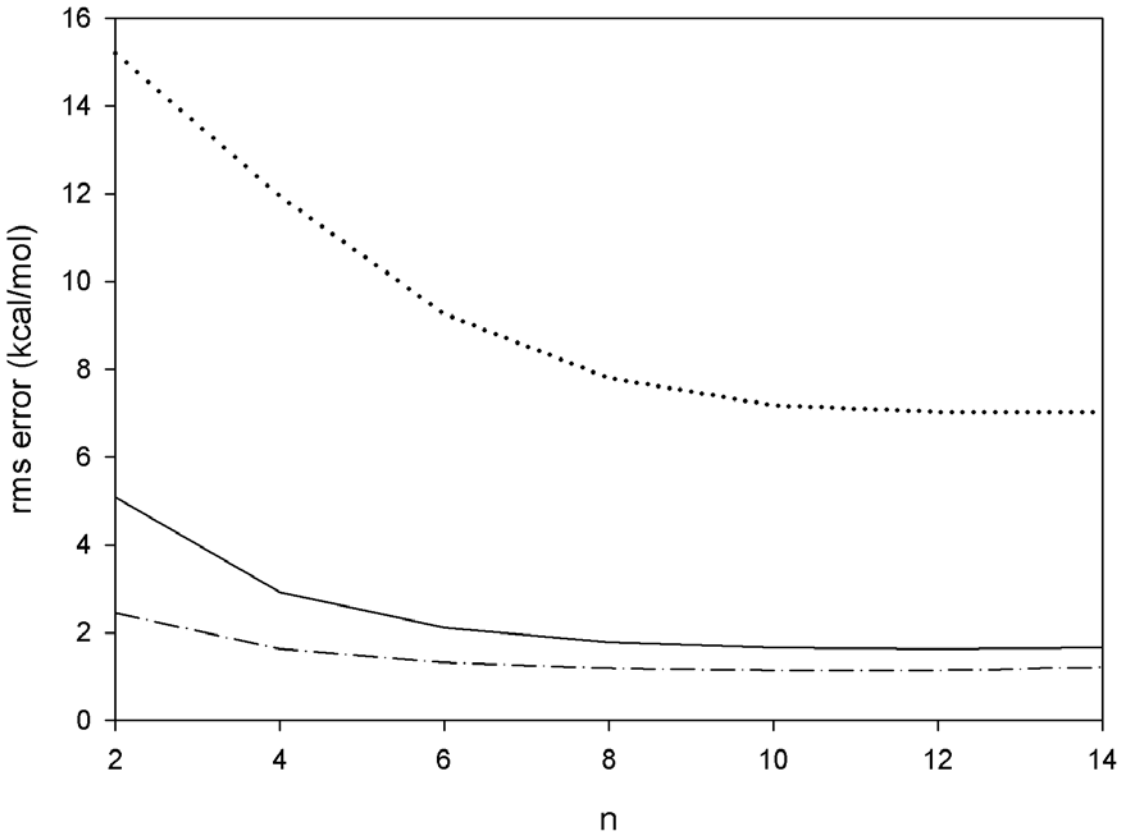


Figure 5

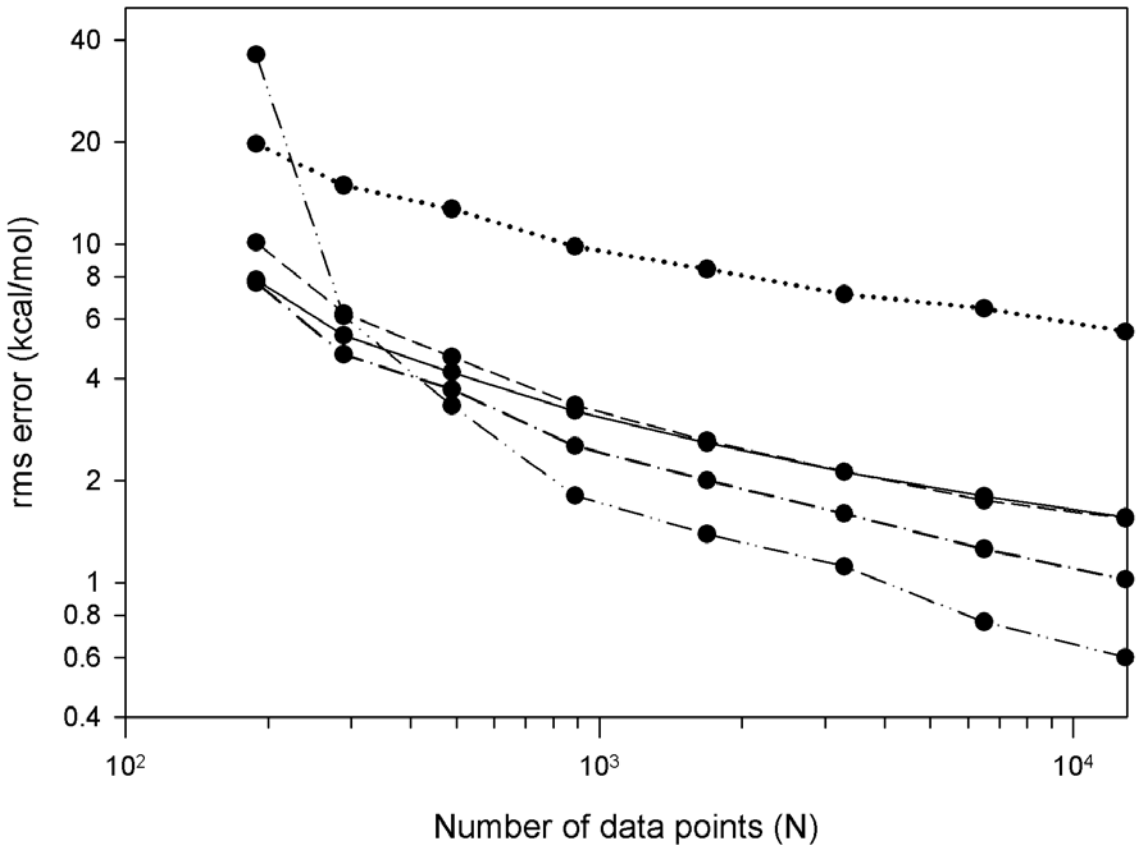

40
20
10
8
6
4
2
1
0.8
0.6
0.4
rms error (kcal/mol)
10^2
10^3
10^4
Number of data points (N)

Figure 6

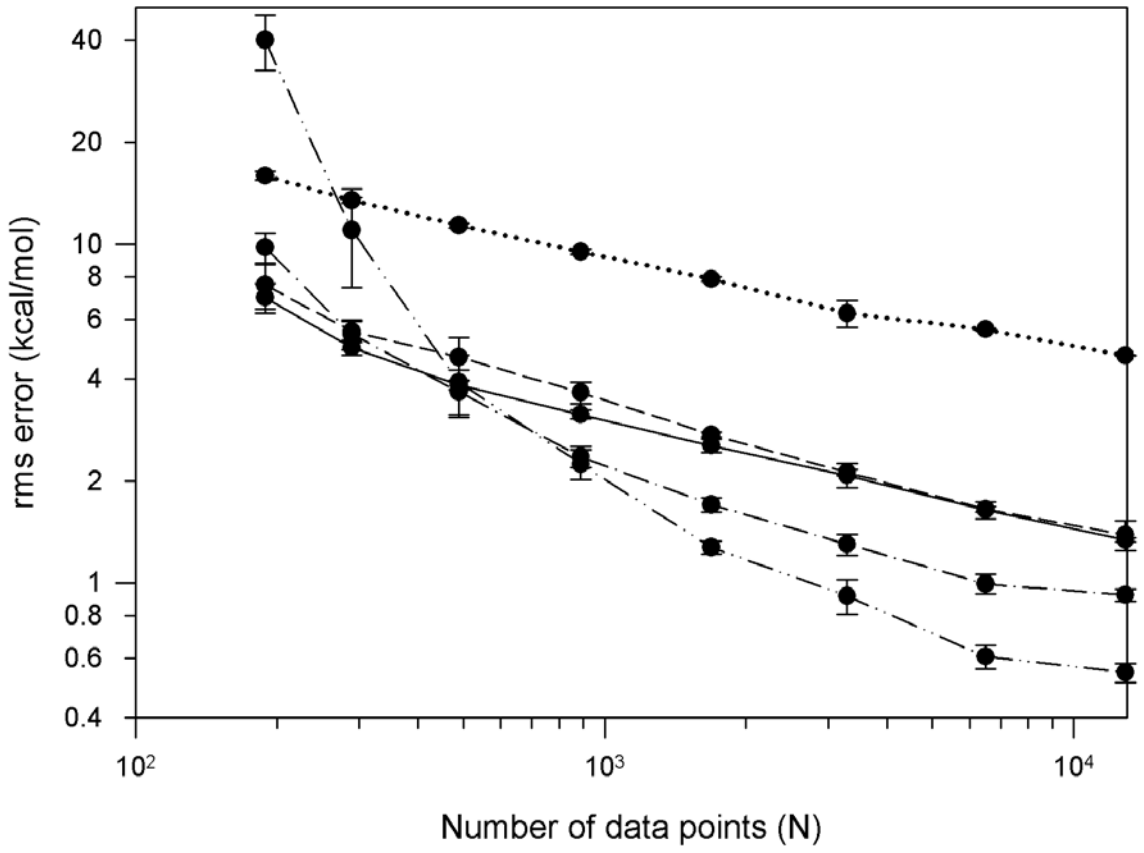

40
20
10
8
6
4
2
1
0.8
0.6
0.4
rms error (kcal/mol)
10^2
10^3
10^4
Number of data points (N)

Figure 7

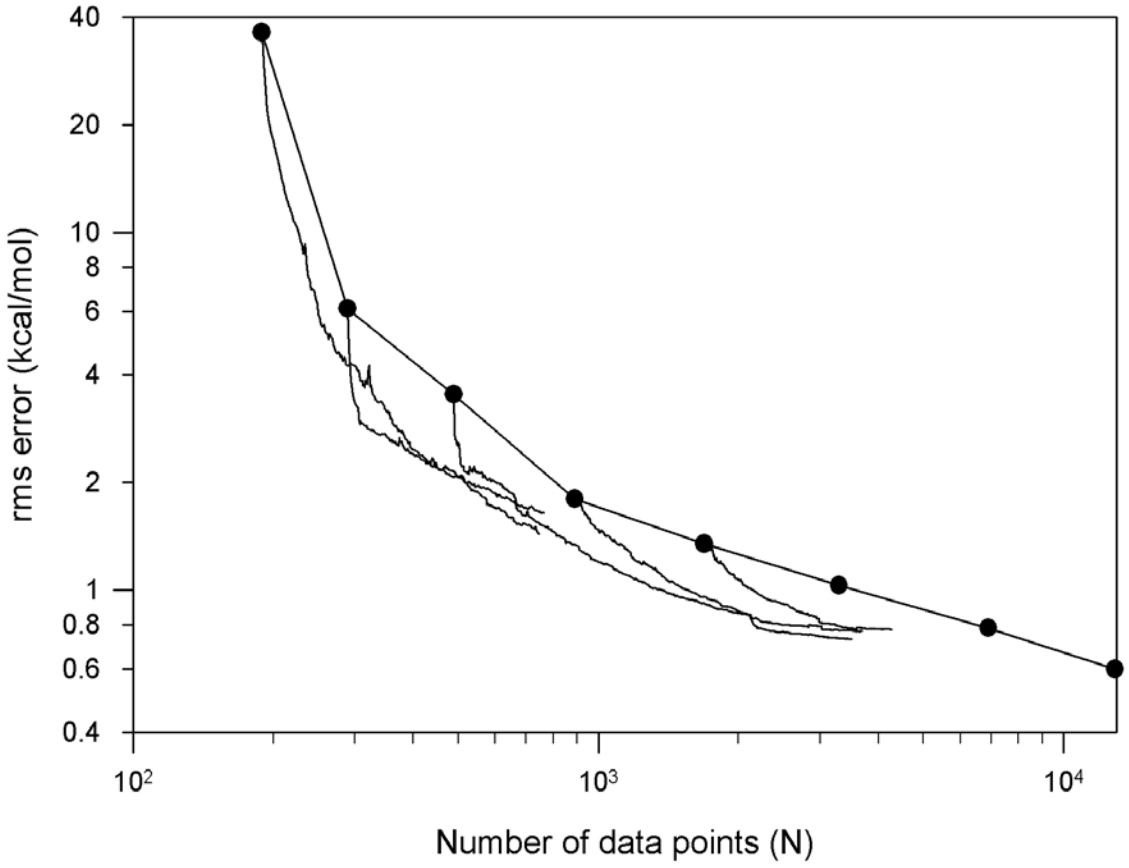
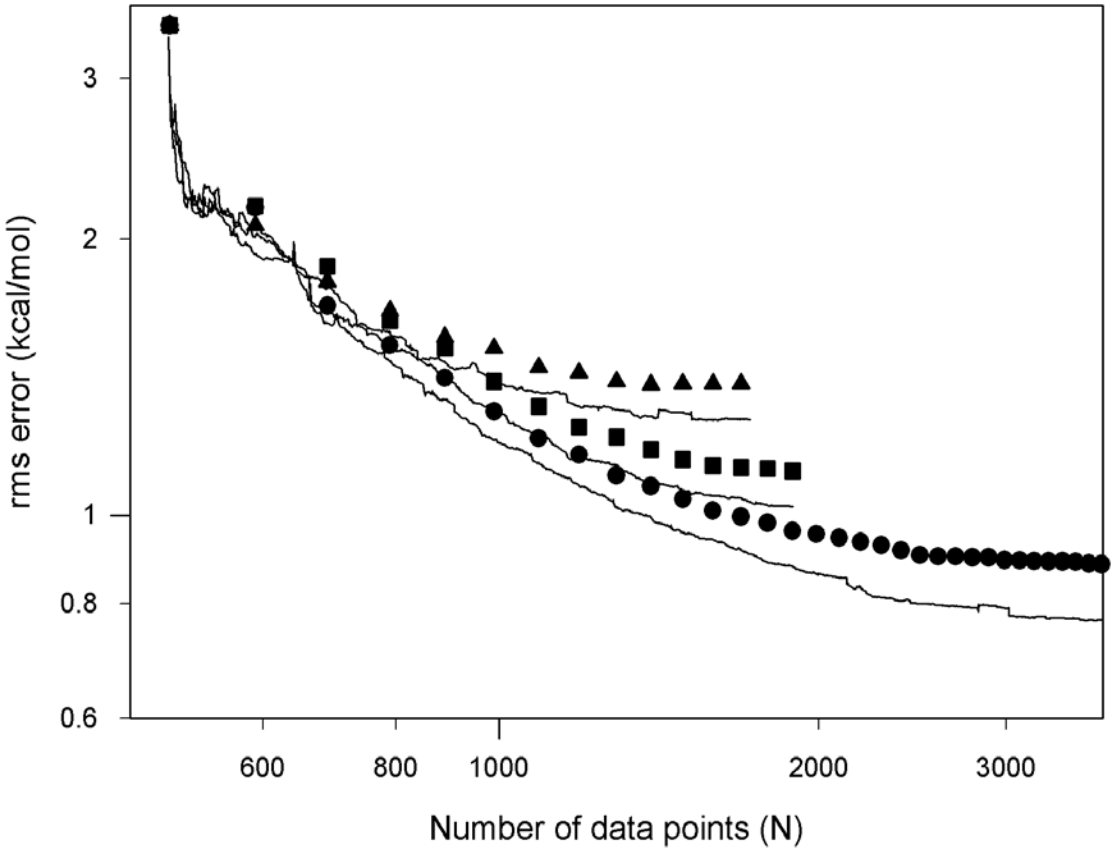

40
20
10
8
6
4
2
1
0.8
0.6
0.4
rms error (kcal/mol)
10^2
10^3
10^4
Number of data points (N)

Figure 8


3
2
1
0.8
0.6
rms error (kcal/mol)
600
800
1000
2000
3000
Number of data points (N)

References

[1] J. N. Murrel, S. Carter, S. C. Farantos, P. Huxley, and A. J. Varandas, *Molecular Potential Energy Functions*, Wiley, Chichester, 1984.

[2] G. C. Schatz, Rev. Mod. Phys. **61**, 669 (1989).

[3] L. M. Raff, and D. L. Thompson, in *Theory of Chemical Reaction Dynamics*, edited by M. Baer (Chemical Rubber, Cleveland, 1985), Vol. III, Chap. 1.

[4] J. Ischtwan and M. A. Collins, J. Chem. Phys. 100 (1994) 8080.

[5] P. Lancaster and K. Salkauskas, *Curve and Surface Fitting, an Introduction*, Academic, London, 1986, Chap. 10.

[6] G. G. Maisuradze and D. L. Thompson, J. Phys. Chem. A, **107**, 7118 (2003).

[7] G. G. Maisuradze, D. L. Thompson, A. F. Wagner, and M. Minkoff, J. Chem. Phys. **119**, 10002 (2003).

[8] A. Kawano, Y. Guo, D. L. Thompson, A. F. Wagner, and M. Minkoff, J. Chem. Phys. **120**, 6414 (2004).

[9] Y. Guo, A. Kawano, D. L. Thompson, A. F. Wagner, and M. Minkoff, J. Chem. Phys. (*in press*).

[6] G. G. Maisuradze and D. L. Thompson, J. Phys. Chem. A, **107**, 7118 (2003).

[7] G. G. Maisuradze, D. L. Thompson, A. F. Wagner, and M. Minkoff, J. Chem. Phys. **119**, 10002 (2003).

[8] A. Kawano, Y. Guo, D. L. Thompson, A. F. Wagner, and M. Minkoff, J. Chem. Phys. **120**, 6414 (2004).

[9] Y. Guo, A. Kawano, D. L. Thompson, A. F. Wagner, and M. Minkoff, J. Chem.

Phys. (*in press*).

[10] K. A. Nguyen, I. Rossi, and D. G. Truhlar, J. Chem. Phys. **103**, 5522 (1995).

[11] B. Kuhn, T. R. Rizzo, D. Luckhaus, M. Quack, and M. A. Suhm, J. Chem. Phys, **111**, 2565 (1999).

[12] (a) H. W. Schrantz, S. Nordholm, and G. J. Nyman, J. Chem. Phys. **94**, 1487 (1991). (b) G. Nyman, S. Nordholm, and H. W. Schrantz, J. Chem. Phys. **93**, 6767 (1990).

[13] T. Yonehara and S. Kato, J. Chem. Phys. **117**, 11131 (2002).