Molecular Potential Energy Surfaces by Interpolation

Michael A. Collins

Research School of Chemistry, Australian National University, Canberra. ACT 0200 Australia. collins@rsc.anu.edu.au

Abstract. The molecular potential energy surface governs the motion of the atomic nuclei for a molecule in an isolated electronic state. For a molecule of N atoms, this surface is a function of 3N-6 internal coordinates which determine the shape of the molecule. For molecules undergoing chemical reaction, the surface is a relatively complicated function of these many coordinates. Methods have now been developed which allow us to construct this surface as an interpolation of Taylor expansions of the surface around molecular configurations scattered throughout the accessible space.

1 Introduction

Many important chemical reactions occur for molecules in a single electronic state; that is where the wavefunction for the electrons is given by a single eigenfunction of the time-independent Schrödinger equation with the positions of the nuclei fixed in space. The total electronic energy is the eigenvalue. This energy depends on the position of the nuclei. The motion of the atomic nuclei is determined by a Hamiltonian which is just the sum of the nuclear kinetic energy and this total electronic energy – the potential energy of the nuclei. The motion of the nuclei can be evaluated by solving the associated Schrödinger Equation, or approximately, by solving the classical Hamilton's equations. In either case, we must evaluate the electronic energy as a function of the nuclear positions.

Why do we want to evaluate the motion of the nuclei? Because this motion tells us everything there is to know about the mechanism of the chemical reaction, about what new molecules are formed, and about how fast the reaction occurs. Reaction rates are important facts to possess, because the outcome of any practical chemical system is the result of competing processes, and far from equilibrium, speed is more important than stability. Chemists strive to understand the mechanism of reactions in the expectation that this knowledge will provide the means of designing new reaction sequences for novel chemical syntheses.

Unfortunately, the molecular potential energy surface (PES) has proved to be very difficult to evaluate. The value of the PES at any molecular configuration can now be evaluated with adequate accuracy for many small to medium-sized molecules using the methods of *ab initio* quantum chemistry. However, the cost of such calculations can be high and the number of configurations at which the energy must be evaluated

P.M.A. Sloot et al. (Eds.): ICCS 2003, LNCS 2660, pp. 159–167, 2003.

is very large for a molecule undergoing chemical reaction. Pessimistically, this number is of the order of $d^{(3N-6)}$, where d is the number of configurations needed for each configurational degree of freedom.

So-called direct dynamics methods attempt to evaluate the classical dynamics of chemical reactions using Monte Carlo sampling to minimise the number of electronic structure calculations. However, up to now, only very low levels of *ab initio* calculations have been possible with inadequately small samples. The traditional route around this $d^{(3N-6)}$ problem has been to represent the PES as an explicit function of the internal molecular coordinates. This function contains many adjustable parameters. *Ab initio* calculations are performed at a small number of significant molecular configurations, including any known stationary points, and the parameters in the PES function are adjusted to fit this data. Some success has been achieved with this method for reactions involving three atoms, but little progress has been made for larger systems.

Recently, substantial progress has been made in evaluating PES by interpolation of *ab initio* data. For triatomic molecules, very accurate PES have been constructed by Ho, Rabitz and coworkers using a "reproduciing kernel Hilbert space" method.[1-4] Our group has pursued a modified form of Shepard interpolation and applied this successfully to reactions involving several atoms. In order to construct a PES in this way, we need both an accurate interpolation method and an efficient method for deciding where in configuration space the (inevitably) limited number of *ab initio* calculations are performed.

This paper presents a brief description of this method, illustrated by some recent applications to polyatomic reactions. More details can be found elsewhere.[5-11]

2 The Method

The PES is a function of the internal coordinates of the molecule. Any complete set of coordinates would suffice, but for convenience and accuracy we employ the atom-atom distances, $\{R_i\}$. In practice, we actually use the reciprocal distances, $Z_i = 1/R_i$. There are N(N-1)/2 such Z_i and only 3N-6 of them can be independent. We will discuss this "redundant coordinate" problem below. For the moment, let's assume that we can use *ab initio* quantum chemistry to evaluate the energy E(Z) at any molecular configuration we like. Moreover, if we have evaluated E at a set of N_{data} configurations {Z(n)}, we can write the energy at any configuration as a weighted average of these energies:

$$E(Z) \quad = \quad \sum_{n=1}^{N_{data}} w_n\big[Z - Z(n)\big]E(n) \ . \tag{1}$$

If the weights are chosen so that E(Z) = E(n) when Z = Z(n), then Eq.(1) is an interpolation formula, called a Shepard interpolation. A suitable choice of weight function would be

$$w_n = \frac{v_n}{\sum_{j=1}^{N_{data}} v_j}; \quad (2)$$

$$v_n[Z - Z(n)] = \|Z - Z(n)\|^{-p} . \quad (3)$$

The so-called primitive weight v_n diverges with some positive power p as Z → Z(n) to ensure that the relative weight $w_n \rightarrow 1$ as Z → Z(n), and that Eq.(1) is an interpolation of the *ab initio* energies. Equation (1) suffers from the drawback that its gradient is zero at all the data point configurations (when p is as large as it needs to be). To avoid this, we use a modified form of Shepard interpolation where the energies E(n) are replaced by Taylor expansions of the energy:[12, 13]

$$E(Z) = \sum_{n=1}^{N_{data}} w_n[Z - Z(n)]T_n[Z - Z(n)]. \quad (4)$$

We have shown that Eq.(4) can provide an accurate description of the PES and its gradient if the Taylor expansion is at least second order in the deviation from the data point, Z - Z(n). This requires that we evaluate *ab initio* the energy gradient and second derivative matrix at Z(n) in addition to the energy. Higher order Taylor expansions are even more accurate, but at present this does not justify the additional *ab initio* computational cost.[7]

We cannot implement Eq.(4) if N > 4: Since the Z_i are not independent, the partial derivatives of the PES (which appear in the Taylor expansions) cannot be evaluated. The reciprocal bondlengths include redundant coordinates. This problem has been discussed in detail elsewhere.[8, 11, 14] The relevant solution here is to choose an independent set of 3N-6 linear combinations of the Z_i as internal coordinates at each data point Z(n). This can be achieved as follows. The deviation of the N(N-1)/2 reciprocal bondlengths from their values at a data point, δZ = Z - Z(n), is related to the corresponding displacement of the 3N Cartesian coordinates, δX, by

$$\delta Z = B\delta X , \quad (5)$$

where the matrix B (a variant of the Wilson B matrix [15]) is given by

$$B_{ij} = \frac{\partial Z_i}{\partial X_j}, \; i = 1,..,N(N-1)/2; \; j = 1,..,3N. \quad (6)$$

A singular value decomposition of B reveals that there are only 3N-6 non-vanishing singular values, which reflects the fact that the internal structure of a molecule can only be distorted in 3N-6 independent ways. The singular vectors, u(m), associated with these non-zero singular values, λ(m), provide the means to define 3N-6 independent coordinates:

$$\xi_m \quad = \quad \sum_{i=1}^{N(N-1)/2} u_i(m) Z_i / \lambda(m), \;\; m = 1,..3N-6. \tag{7}$$

The singular value decomposition is carried out at each data point Z(n) to define a new set of independent coordinates for use in each Taylor expansion. Eq.(4) is then replaced by a form which can be evaluated for any number of atoms:

$$E(Z) \quad = \quad \sum_{n=1}^{N_{data}} w_n[Z - Z(n)] T_n[\xi - \xi(n)] \,. \tag{8}$$

One very attractive feature of Eq.(8) is the ease with which symmetry is implemented. Once we have calculated the energy, energy gradient and second derivatives at some molecular geometry, these quantities are also known at all other geometries which are related by the permutation of indistinguishable particles. So, we can easily add all these "indistinguishable geometries" to the data set. This ensures that Eq.(8) gives the same energy at all permuted versions of a molecular configuration, as symmetry requires.

An important refinement of Eq.(8) has been achieved by implementing a more sophisticated weighting function than that in Eq.(3). The primitive weight in Eq.(3) is based on the reasonable premise that the closer a data point Z(n) is to the point where the PES is evaluated, the more accurate should be the estimate T_n. So the primitive weight increases as an inverse power of the distance ||Z - Z(n)||. However, an analysis of the convergence of Eq.(8) with increasing data set size, shows that the simple form of Eq.(3) ensures convergence of the PES but exaggerates the divergence of the weight function as $Z \rightarrow Z(n)$. A more accurate primitive weight function is given by

$$v_n \quad = \quad \left\{ D[Z, Z(n)]^q + D[Z, Z(n)]^p \right\}^{-1}, \text{ where} \tag{9}$$

$$D[Z, Z(n)] \quad = \quad \left\{ \sum_{i=1}^{N(N-1)/2} \left[\frac{Z_i - Z_i(n)}{d_i(n)} \right]^2 \right\}^{\frac{1}{2}}.$$

The power q > 2, but q << p which must be greater than 3N-3. The $d_i(n)$ are "confidence lengths". If the Taylor series, T_n, is acceptably accurate for $Z_i - Z_i(n) < d_i(n)$, then the weight function varies only slowly for such cases. The confidence lengths can be estimated from a Bayesian analysis of the error that T_n makes in estimating the energy gradients at nearby data points.[10] Eqs (2), (9), and (8) provide us with an accurate interpolation formula if the data point locations are well chosen.

Traditionally, the locations of stationary points on the PES (equilibrium reactants and products and saddle points) have been seen as particularly significant for the reaction dynamics. The simplest picture of the reactive dynamics is that of motion along a "reaction path" or "minimum energy path" linking reactants, the saddle point (barrier), and products. So, in this tradition, we choose a set of molecular configurations on such a reaction path as an initial data set to describe the PES via Eq.(8). Already, we have just about exhausted our intuition as to where data points should be located. Henceforth, the important decisions are automated. The automation

is based on the following reasoning. Ultimately the PES is used to simulate the chemical reaction to calculate observables such as the reaction rate or the distribution of energy in the reaction products, or qualitative and quantitative information about the reaction mechanism. The relationship between the PES and such observables is rather obscure. However, it is certain that the value of the PES at configurations which the molecule passes through during the dynamics must play some role in determining the observed dynamics. It is only at these configurations where the PES must be known, *not* throughout the whole 3N-6 dimensional configuration space.

Given an initial set of data points on the minimum energy path for the reaction, the PES of Eq. (11) is defined, albeit inaccurately at configurations far from this path. We simulate the chemical reaction in a standard way by solving the classical (Newtonian or Hamiltonian) equations of motion with initial conditions corresponding to the collision of the reacting molecules. The choice of initial conditions should reflect the observable properties we wish to simulate; for example, the initial vibration-rotation-translational energies or temperature should be at least as high as the values appropriate to any relevant experiments. Typically, a small number (say 10) of trajectories (collisions) are calculated, and the molecular configuration is periodically written to a file during these trajectories. This set of configurations represents a small (typically of order 10^3) sample of the dynamically important portion of configuration space. We choose one of these configurations to be a new data point. Two criteria for this choice are used. One is purely geometrical in that it chooses the sampled point which is optimally near many other sampled points but far from existing data points. The second criterion chooses the sampled point for which the uncertainty, $\sigma(Z)$, in E(Z) is highest:

$$\sigma(Z)^2 \;=\; \sum_{n=1}^{N_{data}} w_n \left[T_n(Z) - E(Z) \right]^2 . \tag{10}$$

We assume that if the uncertainty in the PES is large at some Z(k), then it is more likely that the estimated value of E[Z(k)] is inaccurate, and by this criterion we hope to reduce the largest errors in the PES.

Having chosen a new data point, we have a new version of the PES. We repeat the cycle of simulating the reaction and choosing a new data point from a sample obtained by exploring the new PES. The process is iterated until convergence. The PES is deemed to be "converged" when the values of the dynamical observables of interest do not change significantly with increasing data set size. These values are measured by performing large scale classical simulations of the reaction from time to time - say after every 100 new data points are added to the data set.

There are two important aspects of "automated learning" in this process of iteratively constructing a PES (we say we "grow" PES). As each new data point is added the shape of the PES changes. The computer program "learns" more about the shape of the PES at each cycle. In the small set of classical trajectories, the molecule may move to new configurations not previously allowed by the PES, so the program "learns" about new regions of configuration space. In practice, previously unknown reactions and reaction paths can be discovered. The second type of automated learning is achieved by the iterative improvement in the confidence lengths used in the weight function. As the data set grows, the estimation of the confidence lengths becomes more accurate. The program "learns" more accurate values of the PES not

only by accumulating more *ab initio* data, but by learning how to use all its accumulated data more accurately.

The complete process for constructing a PES is automated in a program package called *Grow* which is available from the author at no cost to noncommercial researchers. At present, the package is compatible with the GAUSSIAN program. The interface with GAMESS will be available shortly.

3 Examples

PES have now been constructed for many reactions involving four or more atoms. In this section, results from a few of these systems are used to illustrate the method.

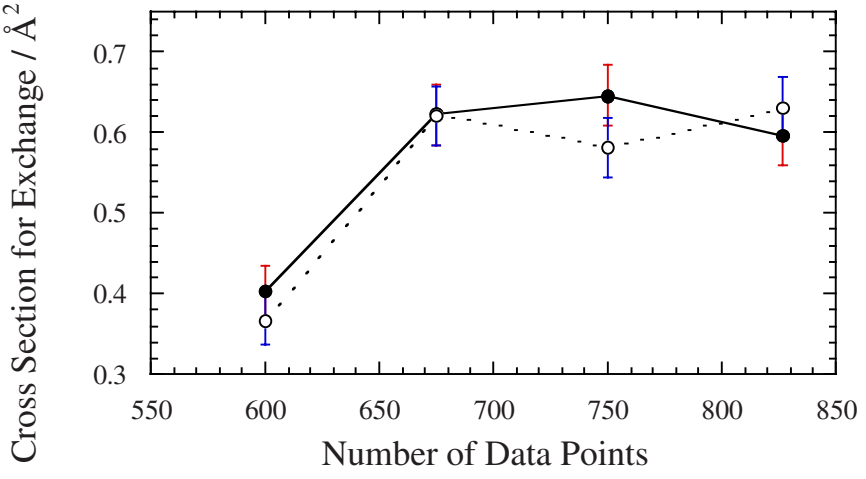

Fig. 1. The classical cross section for the exchange reaction is shown for PESa (o) and PESb (•) as a function of the size of the data set. The error bars correspond to two standard deviations in the trajectory simulations.

The abstraction reaction, H + $H_2O \rightarrow H_2$ + OH, is an important rate determining step in the overall process of hydrogen combustion. This reaction, and its reverse, have also become benchmark systems in the development of quantum reactive scattering calculations. Highly accurate *ab initio* PESs have been constructed for this system and employed to test quantum scattering calculations against experiment.[16-19] This system, and its deuterated analogues, also support the exchange process: H + $H_2O \rightarrow$ HOH + H. The PES for this system was initially grown at the QCISD(T) level of theory using a basis set correction approach: E = E[QCISD(T)/6-311G(d,p)] + E[MP2/6-311++G(3df,2pd)] - E[MP2/6-311G(d,p)]. We denote this surface as PESa. Later the energy (only) of each data point was replaced by the QCISD(T)/6-311++G(3df,2pd) value. We denote this surface as PESb. Figure 1 shows that the

classical cross section for the exchange reaction is converged with respect to the data set size for both surfaces.

This apparent convergence in a dynamical observable is consistent with our general observation, for over ten systems, that the average interpolation error decreases as a function of data set size; usually in a fashion similar to that shown in Figure 2 for the $ArCOH^+$ system.[20] Here 900 molecular configurations were sampled from classical trajectories simulating the collision of Ar and COH^+ and the energy at each geometry was evaluated at the same MP2/6-311G(d,p) level of theory used to construct the PES. The interpolated energy for each of the 900 configurations was evaluated for data sets of increasing size and compared with the directly calculated value. The mean absolute error is shown in Figure 2. The mean error of about 0.8 kJ mol^{-1}, for the largest data set, represents less than 0.3 % of the energy range of these 900 configurations.

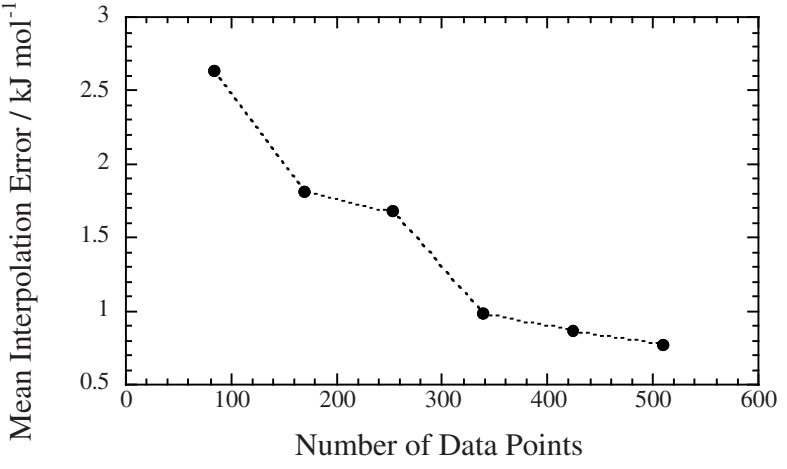

Fig. 2. The mean absolute error in the PES for $ArCOH^+$, estimated from 900 sampled configurations, is shown as a function of the number of data points.

Note that the PES for H_3O was initially constructed with no knowledge of the exchange reaction, and no data points on a pathway for exchange were included in the initial data set. The automated construction process "discovered" the exchange reaction and included data points in that region as required. Figure 3 depicts the data point locations which feature both abstraction and exchange reaction paths in the final data set.

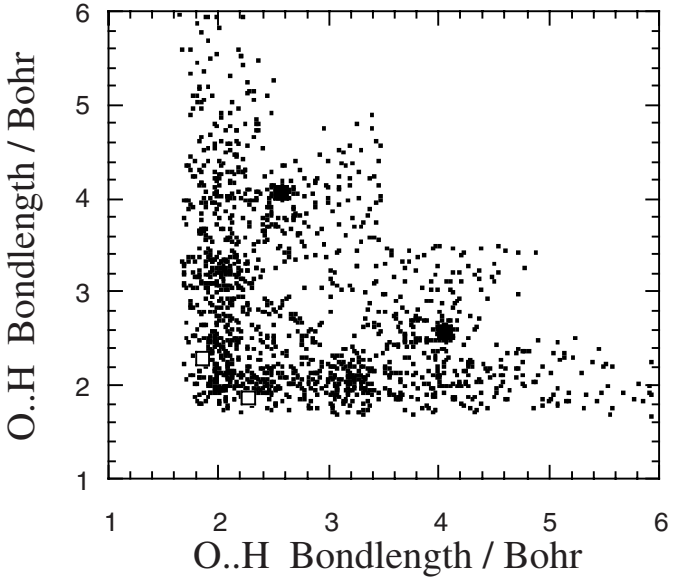

Fig. 3. Part of the data set for H + H_2O is represented as points in a projection onto the two longest OH bondlengths for each data point geometry. The large symbols indicate saddle points for abstraction and exchange.

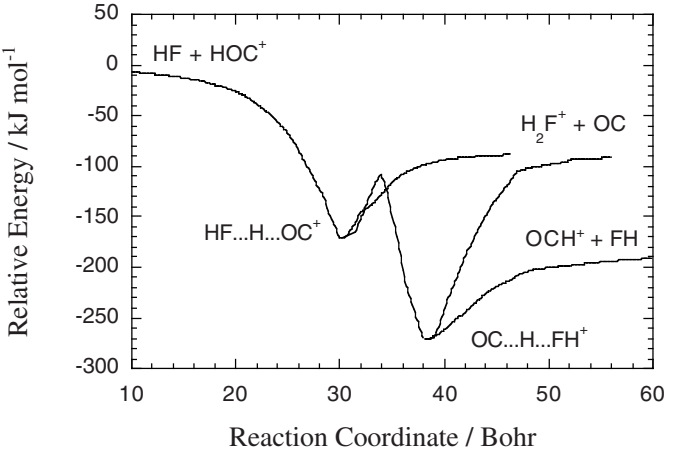

Fig. 4. The energy profile for minimum energy paths on the PES for FH + HOC^+ is shown as a function of a reaction coordinate.[21]

Multiple pathways and competing reactions are common features in the PES for molecules of even as few as four atoms. It seems reasonable to suggest that even greater complexity will be discovered in the PES for larger molecules. Figure 4 depicts the energy profiles for reactions paths on a PES for the five-atom system FH +

COH^+, constructed at the MP2/6-311G(d,p) level of theory. Using a deuterated analogue for clarity, it is clear that four different processes are possible in this system; abstraction, FH + $DOC^+ \rightarrow FHD^+$ + OC; isomerisation, FH + $DOC^+ \rightarrow$ FH + OCD^+, exchange, FH + $DOC^+ \rightarrow$ FD + HOC^+; and isomeration with exchange, FH + $DOC^+ \rightarrow$ FD + OCH^+. Construction of the interpolated PES for this system has provided the means to calculate, classically, the reaction cross sections for all four processes.[21]

References

1. T. Ho, and H. Rabitz, J. Chem. Phys. 104, 2584 (1996).
2. T. Hollebeek, T. Ho, and H. Rabitz, J. Chem. Phys. 106, 7223 (1997).
3. G. C. Schatz, A. Papaioannou, L. A. Pederson, L. B. Harding, T. Hollebeek, T.-S. Ho, and H. Rabitz, J. Chem. Phys. 107, 2340 (1997).
4. T. Hollebeek, T.-S. Ho, H. Rabitz, and L. B. Harding, J. Chem. Phys. 114, 3945 (2001).
5. J. Ischtwan and M. A. Collins, J. Chem. Phys. 100, 8080 (1994).
6. M. J. T. Jordan, K. C. Thompson, and M. A. Collins, J. Chem. Phys. 102, 5647 (1995).
7. M. J. T. Jordan, K. C. Thompson, and M. A. Collins, J. Chem. Phys. 103, 9669 (1995).
8. K. C. Thompson, M. J. T. Jordan, and M. A. Collins, J. Chem. Phys 108, 8302 (1998).
9. K. C. Thompson and M. A. Collins, J. Chem. Soc., Faraday Trans. 93, 871 (1997).
10. R. P. A. Bettens and M. A. Collins, J. Chem. Phys. 111, 816 (1999).
11. M. A. Collins, Theor. Chem. Acc. 108, 313 (2002).
12. R. Farwig, Mathematics of Computation 46, 577 (1986).
13. R. Farwig, in *Algorithms For Approximation,* edited by J.C. Mason and M.G. Cox (Clarendon Press, Oxford, 1987)p. 194.
14. M. A. Collins and D. F. Parsons, J. Chem. Phys. 99, 6756 (1993).
15. E. B. Wilson, J. C. Decius, and P. C. Cross, *Molecular Vibrations* (Dover, New York, 1955)p. 140.
16. M. Yang, D. H. Zhang, M. A. Collins, and S.-Y. Lee, J. Chem. Phys. 115, 174 (2001).
17. D. H. Zhang, M. A. Collins, and S.-Y. Lee, Science 290, 961 (2000).
18. M. Brouard, I. Burak, S. Marinakis, D. Minayev, P. O'Keeffe, C. Vallance, F. J. Aoiz, L. Banares, J. F. Castillo, D. H. Zhang, D. Xie, M. Yang, S.-Y. Lee, and M. A. Collins, Phys. Rev. Lett., in press (2003).
19. M. Brouard, I. Burak, D. Minayev, P. O'Keeffe, C. Vallance, F. J. Aoiz, L. Banares, J. F. Castillo, D. H. Zhang, and M. A. Collins, J. Chem. Phys. 118, 1162 (2003).
20. A. J. Chalk, S. Petrie, L. Radom, and M. A. Collins, J. Chem. Phys. 112, 6625 (2000).
21. M. A. Collins and L. Radom, J. Chem. Phys., in press (2003).