Journal of Theoretical and Computational Chemistry, Vol. 2, No. 2 (2003) 179–191


World Scientific
www.worldscientific.com

# SELECTING METHODS TO SOLVE MULTI-WELL MASTER EQUATIONS

TERRY J. FRANKCOMBE* and SEAN C. SMITH†
*Department of Chemistry, University of Queensland,*
*Brisbane, Queensland, 4072, Australia*
**T.Frankcombe@chemistry.uq.edu.au*
†*S.Smith@uq.edu.au*

Received 20 November 2002
Accepted 12 February 2003

There are several competing methods commonly used to solve energy grained master equations describing gas-phase reactive systems. When it comes to selecting an appropriate method for any particular problem, there is little guidance in the literature. In this paper we directly compare several variants of spectral and numerical integration methods from the point of view of computer time required to calculate the solution and the range of temperature and pressure conditions under which the methods are successful. The test case used in the comparison is an important reaction in combustion chemistry and incorporates reversible and irreversible bimolecular reaction steps as well as isomerizations between multiple unimolecular species. While the numerical integration of the ODE with a stiff ODE integrator is not the fastest method overall, it is the fastest method applicable to all conditions.

*Keywords*: Spectral solution; numerical integration; stiff ODE solver; CPU time; matrix methods.

## 1. Introduction

The master equation (ME) formulation for solving gas-phase chemical kinetics problems is well known and commonly employed.[1–5] The most common and oldest application of ME methods identifies the smallest magnitude eigenvalue of the energy grained ME matrix as the macroscopic rate constant of a unimolecular dissociation or isomerization reaction, treated irreversibly. Beyond this long-time, single exponential decay mode corresponding to the macroscopic rate constant, with a little more computational effort the ME description yields the total concentration and energy-resolved population distribution at all times, following the specification of some initial population distribution. A two-dimensional (2D) ME, resolved in both energy and angular momentum, yields similar information to its smaller one-dimensional (1D) counterpart in terms of both macroscopic rate constants and time-dependent population distributions. 2D MEs incorporate the effect of angular momentum conservation which is neglected by 1D ME descriptions and which may be important in high accuracy calculations.[6–11]

Increasingly, ME methods are being employed to investigate systems beyond single unimolecular reactions. Complex networks of interconverting isomers and bimolecular reactions proceeding through a long-lived collision complex can readily be modelled using ME matrix techniques.[12–19] ME simulations of these so-called multi-well systems can yield similar information to the unimolecular case. The determination of rate constants from the ME is often not as straightforward as simply identifying the smallest eigenvalue of the ME matrix. The evolution in time of a particular initial distribution of population can be readily calculated, and several types of analyses are available

to determine classical chemical rate constants from such time-dependent populations.[20] Bimolecular reaction paths are usually incorporated by assuming that the reaction is occurring under pseudo-first-order conditions.

The matrices arising from 2D and multi-well MEs are significantly larger than unimolecular ME matrices resolved in energy only. While the discretization in energy of a 1D unimolecular ME is typically of the order of hundreds of energy grains (less than a thousand), multi-well MEs resolved in energy only are discretized with hundreds of energy grains within each species, or well. Similarly, 2D MEs are discretized with hundreds of energy grains within each rotational level. One can easily construct a 2D or multi-well ME discretized over tens of thousands of points. The potential also exists to construct 2D multi-well MEs, with a corresponding further increase in the size of the discretization.

Clearly, the order of the matrix describing the discretized ME is equal to the number of points in the discretization. Increasing the size of the discretization (and hence the ME matrix) significantly increases the computational effort and amount of computer time required to solve the ME. Hence using ME methods effectively in applications where 2D or multi-well descriptions are required, with the corresponding larger matrices, needs the fastest solution methodologies available. Additionally, common modes of using ME solutions are in the calculation of falloff curves, requiring repeated solutions for different pressures, or fitting of parameters to the functional forms defining the kinetic or collisional quantities,[21–25] for which a fast solution method is essential to enable many iterations of the fitting procedure.

At least three different approaches to solving the ME for time-dependent population distributions exist and are commonly used. The most common methods are spectral methods, based on finding an eigendecomposition of the ME matrix. The exponential operator is then expanded in the eigenbasis, with a truncated expansion being valid for a restricted time range. A truncation of this type is the origin of identifying the smallest eigenvalue of the unimolecular ME matrix as the macroscopic rate constant. The discretized ME is a stiff ordinary differential equation (ODE), so that stiff ODE integrators can be used to propagate in time an initial population distribution. The classification of the ME as a stiff ODE depends on the range of the eigenvalue spectrum, reinforcing the significance of spectral analysis of the ME matrix. The integration of the ODE can also be achieved by Monte Carlo methods.[26–29] This approach appears to be good at simulating very complex dynamics, but can be slow to converge.

The range of methods available and the importance of the speed of finding the solution makes the question of selecting a solution methodology an important one. The selection is further complicated by numerical properties of the solution, rendering methods that are highly successful under one set of conditions ineffective under other conditions. Usually low temperatures and pressures lead to numerical difficulties.

Despite the importance of the relative speed of solution methods, these are rarely directly compared. In this paper we aim to somewhat redress this situation, giving a direct comparison between several competing methods for a test system. Our primary focus will be on the speed of each method in calculating the solution accurately. The robustness (the ability to solve difficult cases) of each of the selected methods must also be considered.

The structure of the paper is as follows. In the next section we revise the ME used as the test case. Section 3 gives an overview of the solution methods being compared in this work. General descriptions of the solution behaviour and comparisons are given in Sec. 4, while the dependence on the architecture of the processor used to perform the calculations are highlighted in Sec. 5. Section 6 concludes.

## 2. The Master Equation

The ME is well known and described in detail elsewhere,[1–5,13,14] so only some details pertinent to the current case shall be pointed out here. The energy grained multi-well ME discretized over a set of energy grains $p_i$ (with each isomer described by a subset of the $n$ grains $p_i$) can be written as a series of equations of the form

$$\frac{dp_i}{dt} = \omega\delta E \sum_j P_{ij}p_j - \omega p_i - p_i \sum_r k_i^{(L,r)} + \sum_r k_i^{(G,r)} p_{i_{ri}} \qquad (1)$$

where $\omega$ is the collision frequency, $\delta E$ is the energy grain size, $P_{ij}$ describes collisional energy transfer

within each species, $k_i^{(L,r)}$ and $k_i^{(G,r)}$ are microscopic rate constants for the interconversion reactions and $i_{ri}$ is an indexing function. The sum over $j$ is over all energy grains belonging to the same species as grain $i$ while the sums over $r$ are over all reactive channels. For notational simplicity the explicit time-dependence of $p_i$ has not been shown.

Bimolecular reactions are easily incorporated if they are modelled under pseudo-first-order conditions (which makes the reaction linear in $p_i$).[19] The first two terms on the right of Eq. (1) do not apply in the bimolecular case if the reactant not in excess is assumed to maintain its equilibrium distribution, which is a reasonable assumption. The $k_i^{(L,r)}$ and $k_i^{(G,r)}$ terms for reactions from bimolecular states are then formed by the microscopic rate constant for the reaction multiplied by the total population of the bimolecular species assumed to be in excess and the Boltzmann population of the reactant not in excess. As usual, detailed balance can be invoked to determine the rate constants for the reverse reaction.

Equation (1) can be written as a matrix ODE as

$$\frac{d\mathbf{p}}{dt} = A\mathbf{p} \qquad (2)$$

where $A$ is an $n \times n$ matrix if there are $n$ energy grains in total.

Clearly, altering the order of the $p_i$ alters the structure of the matrix $A$. As described in detail elsewhere,[12,13] collecting grains from each isomer together yields a well-structured block matrix with dense blocks on the main diagonal and diagonal blocks elsewhere, plus non-zero rows and columns corresponding to the bimolecular terms. This block structure can be utilized to produce a matrix-vector product routine which takes less time than a generic matrix-vector product, scaling better than the normal order $n^2$.

As a test system we use a multi-well ME similar to one we have studied previously.[13,14] The ME describes the reaction between singlet methylene and acetylene, which proceeds through a multi-well collision complex to form propargyl:

$$^1\mathrm{CH_2} + \mathrm{C_2H_2} \rightleftharpoons \mathrm{C_3H_4} \rightarrow \mathrm{C_3H_3} + \mathrm{H}\,. \qquad (3)$$

This reaction is thought to play an important role in the formation of soot in combustion.[30,31] The $C_3H_4$ species exists as three interconverting isomers:

$$\text{propyne} \rightleftharpoons \text{cyclopropene} \rightleftharpoons \text{allene}\,. \qquad (4)$$

The $^1CH_2 + C_2H_2$ reaction produces the cyclopropene isomer, which must isomerize to allene or propyne before irreversibly decomposing to the propargyl product. This reaction scheme is summarized in Fig. 1.

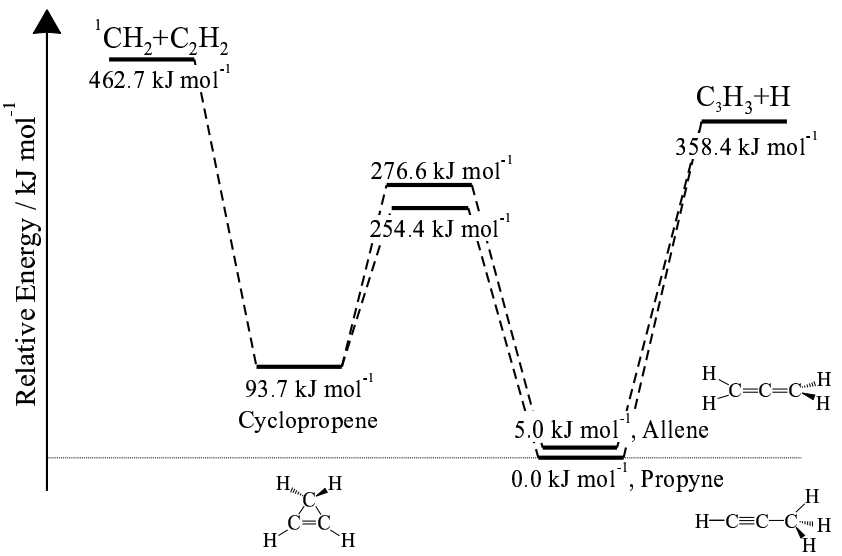


Fig. 1. Schematic reaction scheme for the modelled $^1CH_2 + C_2H_2$ reaction.

The ME studied here differs from that reported in Refs. 13 and 14. The previous studies included only one propargyl-forming reaction, propyne → propargyl + H. That is, decomposition of allene was not considered. The current model treats formation of propargyl from both allene and propyne, with microscopic rate constants based on the inverse Laplace transform[32–35] of the rate expressions of Harding and Klippenstein.[36]

An energy grain size of 200 $cm^{-1}$ was used throughout, giving a matrix of order 714. The collision frequency was taken as the Lennard–Jones value. The rotational constants and vibrational frequencies were taken from Karni *et al.*[37] The $^1CH_2 + C_2H_2$ microscopic rate constants were derived from the data of Blitz *et al.*[38] In the present case, one should not become overly concerned with the accuracy of the input data as this ME is used as a test problem only; detailed predictions of the model are not important in this context.

While the methylene plus acetylene channel was linearized and treated reversibly under pseudo-first-order conditions, our previous work shows that at low temperatures treating the propargyl formation reaction in a similar manner significantly alters the dynamics through reformation of $C_3H_4$.[14] Hence the propargyl population was not included in the modelled state space and was calculated by consideration of conservation of total population. For a discussion of the implications for the available solutions methods on the treatment of bimolecular sinks see our previous paper.[13]

## 3. Solution Methods

This paper compares several varieties of spectral propagation and direct time integration. That is, Monte Carlo integration methods were not considered at this time.

The efficiency and stability of the solution can be improved by symmetrizing the matrix before solving the ME (though under certain circumstances the symmetric form of the ME is not automatically the best choice[39]). If $\mathbf{f}$ is the vector describing the Boltzmann population of the system (with the same ordering as the vector $\mathbf{p}$), then the transformation

$$\boldsymbol{\rho} = S\mathbf{p}\,, \quad B = SAS^{-1} \tag{5}$$

giving the ODE

$$\frac{d\boldsymbol{\rho}}{dt} = B\boldsymbol{\rho} \tag{6}$$

yields a symmetric matrix $B = B^T$ if the diagonal matrix $S$ is given by

$$S_{ii} = f_i^{-1/2}\,. \tag{7}$$

The eigenvalues $\lambda_i$ and eigenvectors $\mathbf{y}_i$ (so that $B\mathbf{y}_i = \lambda_i\mathbf{y}_i$) of a symmetric matrix form an orthogonal set, allowing the formal solution of Eq. (6),

$$\boldsymbol{\rho}(t) = \exp(Bt)\boldsymbol{\rho}(0) \tag{8}$$

to be expressed as

$$\boldsymbol{\rho}(t) = \sum_i \langle\boldsymbol{\rho}(0), \mathbf{y}_i\rangle \exp(\lambda_i t)\mathbf{y}_i \tag{9}$$

allowing the population distribution $\mathbf{p}(t) = S^{-1}\boldsymbol{\rho}(t)$ (and hence the total concentration of each species as an appropriate sum over elements of $\mathbf{p}(t)$) to be calculated for any time $t \geq 0$ from the eigendecomposition of $B$ and the initial population distribution, $\boldsymbol{\rho}(0)$. The eigenvalues $\lambda_i$ are constrained on physical grounds to be strictly non-positive, $\lambda_i \leq 0$. A truncation of the sum in Eq. (9) to include only terms in which the eigenvalue is smaller than some cutoff $c$ yields a truncated expansion valid for times $t \gg 1/c$. Hence, if the eigenvalue spectrum allows, we can get away with looking at a subset of eigenpairs if we are only interested in the evolution of the population at chemically relevant times.

In this work we have used two different approaches to finding the eigenpairs required by Eq. (9). Direct diagonalization using the QR method implemented in the routine DSYEV of the LAPACK library[40] produces the full set of $n$ eigenvalues and eigenvectors for the complete sum in Eq. (9). To truncate the sum to include only contributions from eigenpairs with small eigenvalues, the Lanczos method[41] was used, as implemented in the ARPACK library.[42] While the normal Lanczos iteration converges to extremal eigenpairs faster than internal ones, large magnitude eigenvalues converge much more quickly than small magnitude eigenvalues. It was found that with the Lanczos method applied to the matrix $B$ convergence of the desired small magnitude eigenvalues and the corresponding eigenvectors was impossible to achieve,

particularly for the critical element corresponding to the bimolecular methylene state. Hence the Lanczos method in shift-and-invert mode was used with a zero shift, to transform the desired small eigenvalues into large magnitude eigenvalues. This was at the cost of requiring a linear system solve with the full matrix $B$ at each Lanczos iteration.

The straight application of the Lanczos method to $B$ is not the only method notable by its absence from the timing results to be presented below. Methods based on the drift-determined diffusion approximation[43] which have shown promise previously[12,13,22,44] and are extremely attractive for large systems due to linear scaling, were not successful. The mode of failure was similar to the straight (non-shift-and-invert) Lanczos case, in that the element of the eigenvectors corresponding to the bimolecular methylene state — which completely determines the projection coefficient in Eq. (9) in the current case of an initial methylene population — was not accurately determined.

It is well known that loss of relative accuracy in the eigenvalues and eigenvectors hampers spectral decomposition of MEs at low temperatures and pressures. Following earlier work,[14,38,45] we have overcome this difficulty by performing the calculations in extended precision. The LAPACK and ARPACK routines were ported to quadruple precision, and to arbitrary precision in software arithmetic using the MPFUN package of Bailey.[46] A total of three precisions were used: approximately 16 decimal digits (double precision), approximately 34 decimal digits (quadruple precision) and 50 decimal digits (MPFUN).

A second route to the time-dependent population distribution comes from observing that the solution to Eq. (6) can also be written as

$$\boldsymbol{\rho}(t) - \boldsymbol{\rho}(0) = \int_0^t \frac{d\boldsymbol{\rho}(\tau)}{d\tau} d\tau = \int_0^t B\boldsymbol{\rho}(\tau) d\tau \qquad (10)$$

which fully specifies $\boldsymbol{\rho}(t)$ once one sets $\boldsymbol{\rho}(0)$. Explicit numerical integration of Eq. (6) can in principle be achieved by the simplistic first-order formula

$$\boldsymbol{\rho}(t + \delta t) = \boldsymbol{\rho}(t) + \delta t B\boldsymbol{\rho}(t) \qquad (11)$$

however in practice such a formula, and any related *non-stiff* integration scheme, requires an impractically small $\delta t$ to maintain accuracy. No matter how sophisticated the non-stiff integration and how much recent history of the integration is taken into account (generally reflected in the order of the integration scheme), an impractically large number of matrix-vector products are required to integrate to chemically relevant times, if such integration is achievable at all. Stiff ODE integrators, on the other hand, use less explicit derivative information and more trajectory and eigenvalue information, to achieve more accurate integration over widely varying behaviour. This comes at the expense of using Newton's method to solve a non-linear system of equations involving the Jacobian matrix, the matrix of first derivatives. In the linear first-order case of Eq. (6), the Jacobian matrix is just the ME matrix, $B$.

Two different numerical integration routines were used in this work, both developed by Hindmarch and coworkers: VODE[47] is a variable-coefficient stiff integrator that has been used in previous ME studies, notably by Miller and coworkers.[48–51] LSODA from the ODEPACK package[52] is notable for its ability to automatically switch between stiff and non-stiff integration, as appropriate.[53]

## 4. General Results and Discussion

The two spectral approaches, each in three precision levels, and the two integrators in double precision yield a total of eight variants in the solution method. For each method an initial population of singlet methylene was propagated from $t = 0$ to $t = 1$ second, sampled at 76 discrete times. This time range spans all of the behaviours of the system. Each calculation was performed for MEs describing a range of temperatures from 300 K to 2000 K and at 1 Torr (0.133 kPa) and 1000 Torr (133 kPa).

The quantitative behaviour of the system, as previously reported,[13,14,38] is not significantly altered by the more accurate modelling of the propargyl formation. Representative population profiles are shown in Fig. 2. These population profiles show clearly the different behaviour of the population in different time regimes and indicates that the behaviour is temperature and pressure dependent, particularly in terms of the branching ratios to the $C_3H_4$ stable species.

The shift-and-invert Lanczos and numerical integration approaches require further specification to that which has been given so far. In the Lanczos/ ARPACK case, the number of eigenpairs desired and

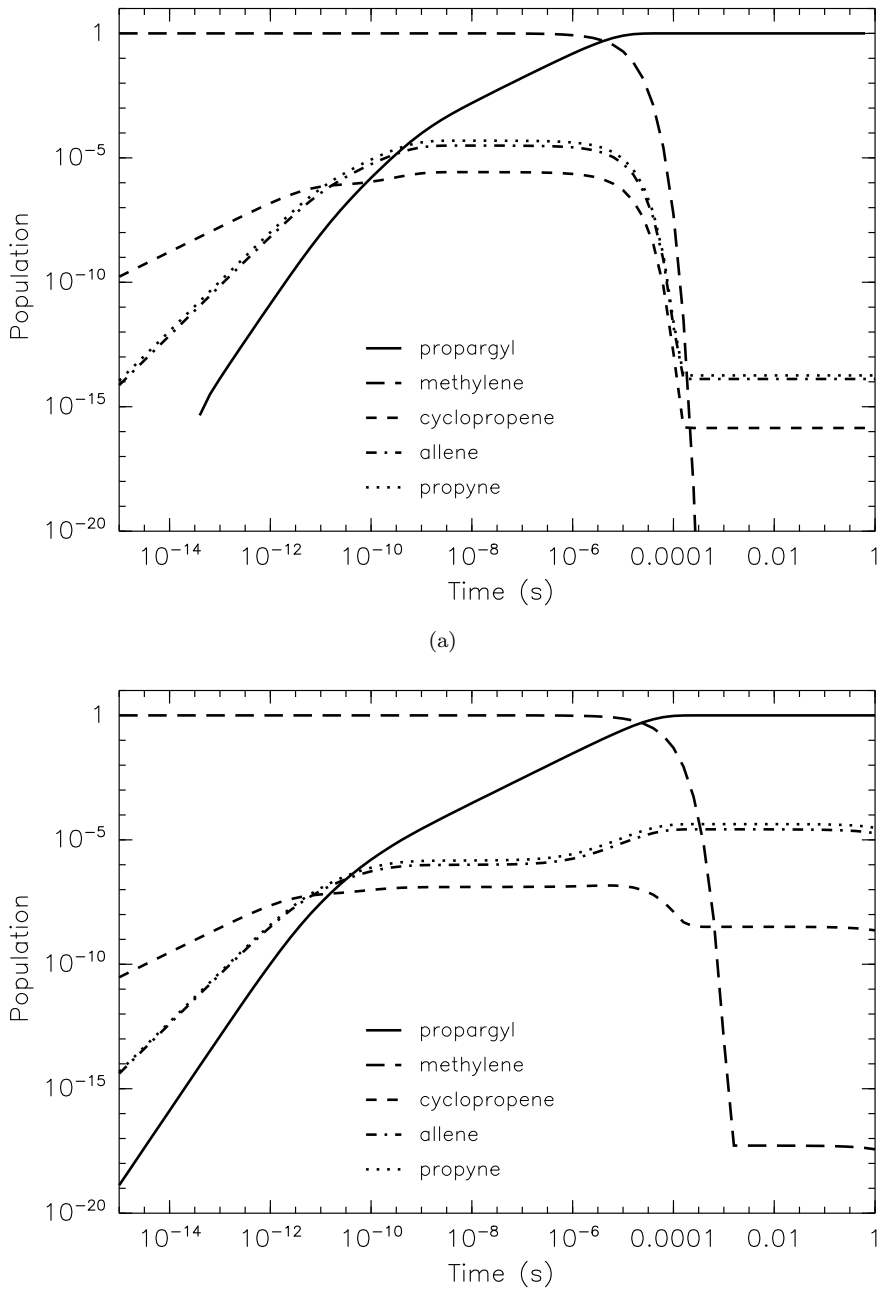


Fig. 2. Population profiles for the five species involved in the modelled $^1CH_2 + C_2H_2$ reaction at (a) 300 K and 1 Torr and (b) 1200 K and 1000 Torr.

the maximum subspace dimension need to be specified. It has been found that for this particular system under the conditions being modelled here, five eigenpairs are sufficient to yield reasonable populations over a decent range of chemically relevant times for the 1000 Torr case. In the 1 Torr case, on the other hand, 25 eigenpairs are required to model the population evolution reasonably. There was little difference in performance observed with changes in the maximum subspace size. Generally a subspace

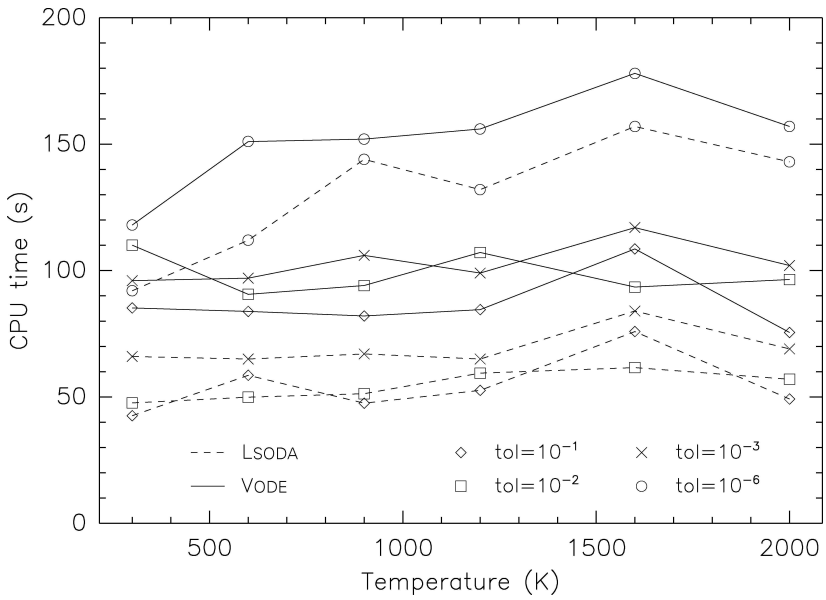


Fig. 3. Sample CPU times for numerically integrating the ME solution with the LSODA and VODE integrators as a function of the modelled temperature. The modelled pressure was 1000 Torr and the relative error tolerances varied from $10^{-6}$ to 0.1. Intel Pentium 4 1.9 GHz CPU.

of dimension 20 would lead to around twice as many restarts before convergence as a subspace of dimension 40.

In the numerical integration case, the error tolerance of the integration needs to be specified. In this work we used a range of tolerances, specifying an allowable relative error in the calculated $\rho_i(t)$ values from $10^{-6}$ to 0.1, with an accurate integration threshold of $10^{-30}$ (so that little effort is made to accurately calculate elements of $\boldsymbol{\rho}(t)$ smaller than $10^{-30}$).

### 4.1. *Numerical integration*

We shall deal first with results from the numerical integration of the ODE. Figure 3 shows sample timings from integrating the initial population (comprised solely of methylene) to one second of simulation time. The integration was performed on an Intel Pentium 4 1.9 GHz processor. While the data shown in Fig. 3 is for modelling a pressure of 1000 Torr, the times arising from modelling 1 Torr are similar. Generally there was a slight trend toward requiring more CPU time to simulate higher temperatures. Such systematic variations of the required CPU time with the conditions being modelled are not particularly surprising. It has been shown, particularly in the work of Miller and Klippenstein,[17,50,54,55] that the eigenvalues controlling the timescales of the different physical processes being modelled change in an unpredictable way as the conditions are varied. Correspondingly the changing behaviour of the system effects the time-step needed to maintain accuracy in the integration, and hence the required number of matrix-vector products and linear system solves. This variation is evident in Fig. 4, which shows the number of matrix-vector products required to integrate the population to a particular time.

Figure 3 clearly shows that the VODE integrator required significantly more CPU time to calculate the solution than the LSODA integrator. This is despite the assertion by the author that VODE is a more sophisticated integrator than LSODA,[47] which is supported by the fact that the VODE integrator required significantly less matrix-vector products to calculate the solution than the LSODA integrator. The timing difference can be attributed to the ability of LSODA to integrate up to around $t = 10^{-10}$ second in non-stiff mode which does not require expensive linear system solves.

In terms of both the total species population and the energy-resolved population distribution, accurate results did not rely on a particularly stringent relative

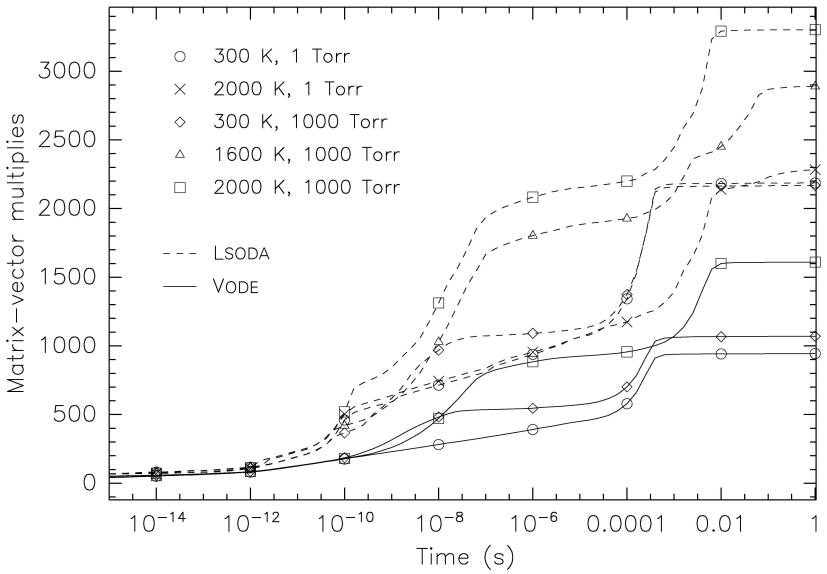


Fig. 4. Number of matrix-vector products required to integrate the ME to particular simulation times with the LSODA and VODE integrators with relative error tolerance $10^{-6}$.

error tolerance for either of the integration routines. Setting the error tolerance to 0.01 yielded significant errors only in very small elements of $\boldsymbol{\rho}(t)$ (below the $10^{-30}$ absolute threshold), and small errors in the element of $\boldsymbol{\rho}(t)$ corresponding to the methylene reactant population. Decreasing the relative error tolerance to 0.1 produced only a moderate increase in the error, significantly smaller than the order of magnitude increase one might expect. The errors produced by the VODE and LSODA integration routines were similar.

Table 1. Average CPU times to calculate the ME solution by direct diagonalization, relative to the total time for the double precision calculation. Intel Pentium 4 1.9 GHz CPU.

| | Double Precision | Quadruple Precision | 50 digit MPFUN |
|---|---|---|---|
| Overhead | $< 0.01$ | 0.37 | 0.16 |
| Diagonalization | 0.95 | 230 | 330 |
| Propagation | 0.05 | 6.3 | 8.0 |
| Total | 1.0 | 240 | 340 |

### 4.2. *Direct diagonalization*

As pointed out in the previous section, spectral methods require high precision arithmetic to combat numerical cancellation errors for low pressure and, particularly, low temperature cases. In terms of required calculation time, the cost of the added precision is high. This is revealed in Table 1 which gives sample timings for the direct diagonalization approach, relative to the total time for the double precision diagonalization and propagation. The data presented in Table 1 is averaged over calculations for the six temperatures considered (from 300 K to 2000 K) at both 1 Torr and 1000 Torr. Systematic variation of calculation time with modelled conditions was once more evident, but was not as simple as in the numerical integration case (where higher temperatures generally took longer), nor as pronounced. The times were averaged as the variation in calculation times generally spanned less than 5% of the average time, making the systematic variation far less significant than in the numerical integration case.

In Table 1, the total calculation times are split into contributions from the matrix diagonalization, from the calculation of the propagated solution via Eq. (9), and the remaining "Overhead". Overhead is primarily the time taken to construct the matrix. Microscopic rate constants and densities of states were precomputed and read from disk, so are not included in the timing of Overhead. Splitting the calculation time in this case shows clearly that the matrix diagonalization is the most time-consuming element of the

calculation of the propagated populations, and that the matrix construction time, which is independent of the solution method within each precision, is an insignificant component of the calculation. The total CPU time required for the total calculation in double precision (the reference time) was 9.54 seconds.

At 1000 Torr, the double precision calculations produced accurate total population profiles over all the timescales of the modelled behaviour at temperatures of 900 K or greater. At 600 K most of the evolution of the total populations was reproduced, losing unimportant detail at times shorter than $t = 10^{-10}$ seconds. The double precision direct diagonalization failed completely at 300 K. The quadruple precision direct diagonalization reproduced the population profiles at the temperatures tested from 600 K and above, at all times. At 300 K the quadruple precision results are inaccurate for times shorter than $t = 10^{-9}$ seconds, with some inaccuracies remaining in the calculated allene population to around $t = 10^{-7}$ seconds. That means that at this temperature and pressure, the quadruple precision calculations do not accurately resolve the populations until well after the high energy $C_3H_4$ populations enter an approximately steady state, with stabilization and reactive loss matching the reactive gain.[14]

At 1 Torr, the double precision direct diagonalization and spectral propagation could not reproduce the accurate behaviour at long times, once the methylene population dropped significantly. That is, even at 2000 K the long-time branching ratios cannot be accurately calculated. The double precision method could correctly calculate the $C_3H_4$ intermediate populations for shorter times, down to 900 K. At 600 K the short time population was not reproduced accurately, though this is a reasonably minor problem. A far greater issue is the fact that the cyclopropene population was significantly over-predicted for all times, which also had an effect on the calculated propargyl population. At 300 K the double precision method failed completely. The quadruple precision version of the calculation calculated accurate populations down to 600 K, but also failed completely for the 300 K case.

### 4.3. *Shift-and-invert Lanczos*

In the 1000 Torr case, five eigenpairs provide details of the evolution of the species populations consistently from around $t = 10^{-9}$ seconds. 25 eigenpairs provides similar resolution in the 1 Torr case. As in the direct diagonalization case, double precision shift-and-invert Lanczos calculations failed at lower temperatures and pressures, requiring higher numerical precision to accurately calculate the solution. Sample relative times for the Lanczos-based calculations are given in Table 2, which shows that in addition to the two orders of magnitude slow-down associated with moving to higher precision, the larger number of eigenpairs required introduced a significant difference between the calculations at the two different pressures considered. Again unlike the direct diagonalization case, there was a significant change in the calculation time with the temperature, with the calculation time generally decreasing with increasing temperature. Typically the calculation time for the system at 600 K was a little more than 20% greater than the 1200 K times represented in Table 2, while the 2000 K timings were a little less than 20% smaller. The five eigenpairs, 1000 Torr case used a subspace of dimension 20 while the 25 eigenpairs, 1 Torr case used a subspace of dimension 40. The matrix construction and solution propagation (in this case greatly truncated from the direct diagonalization case) are again small components of the total calculation time, which is the time represented in Table 2. The reference time was very short compared to the methods previously considered, just 0.52 seconds.

In terms of the required precision, the numerical behaviour of the calculated solutions were very similar to those calculated by direct diagonalization. At 1000 Torr the double precision calculation could produce accurate results down to 900 K, while quadruple precision can readily calculate the time-resolved populations at 600 K, within the limitations of the five eigenpair expansion. Likewise, for the 1 Torr case the long-time populations were not calculated accurately

Table 2. CPU times to calculate the ME solution by shift-and-invert Lanczos using ARPACK for a temperature of 1200 K, relative to the 1000 Torr double precision case. Intel Pentium 4 1.9 GHz CPU.

| | Double Precision | Quadruple Precision | 50 digit MPFUN |
|---|---|---|---|
| 1 Torr | 1.4 | 150 | 205 |
| 1000 Torr | 1.0 | 98 | 125 |

in double precision, even at 2000 K. The intermediate-time populations (from the truncated expansion cut-off around $t = 10^{-9}$ seconds to the depletion of the methylene reactant at around $t = 10^{-3}$ to $t = 10^{-4}$ seconds) could be calculated reasonably accurately in double precision for temperatures of 900 K and above. At 600 K the quadruple precision calculations were accurate until the system approaches the long-time limit as the methylene population drops off. 50 digit arithmetic was required to fully resolve the populations at 1 Torr and 600 K. In most cases the method was not applied at 300 K due to difficulties with the Cholesky decomposition used for the inversion operation.

### 4.4. *Direct comparison*

Having dealt with each of the three methods in turn, we now do some more direct comparisons between the methods. One of the most significant aspects of the comparison is that the numerical integration with a stiff integrator can accurately model the evolution of an initial population, with complete energy resolution, under all the conditions tested here *without* having to resort to high precision calculations in difficult cases. While there is nothing fundamentally wrong with performing the calculations in high precision, it is something of an inelegant solution and requires a high price to be paid in terms of CPU time.

Direct comparison of the times required for all the methods and variations considered is shown in Table 3. The times are quoted relative to the double precision direct diagonalization, the most common approach to solving MEs of this type. The two times quoted for the Lanczos methods are for the five eigenvector, 1000 Torr case and the 25 eigenvector, 1 Torr case, respectively.

Clearly, the double precision spectral approaches require much less CPU time than any of the other methods, with the Lanczos-based method requiring 13 and 18 times less time than the direct diagonalization method. Both methods are standard, general approaches that require no exceptional software. The shift-and-invert Lanczos method has the disadvantage that the number of eigenpairs required to elucidate the desired dynamics is not generally known beforehand.

A much more serious drawback of the double precision spectral methods is that they fail for "low" temperature cases. In this case, low can be anything below 900 K. As the temperature of the modelled system is reduced the failure of the propagation may not be obvious, producing total population profiles that may appear to be reasonable. Errors can be detected by comparing with a more accurate method, though this somewhat defeats the purpose of developing a fast method. A better way of detecting numerical failure, at least the mode of failure caused by catastrophic cancellation in the small elements of the eigenvectors, is by examining the energy resolved populations. Figure 5 demonstrates this. Clearly, the highly irregular and unphysical population at low energy is numerical noise. The magnitude of this noise is the result of the transformation from the symmetrized representation of the ME ($B$ and $\boldsymbol{\rho}$) used to simplify the eigenproblem, back to the original representation ($A$ and $\mathbf{p}$) required to calculate physically meaningful populations. This transformation magnifies numerical noise, initially smaller that $10^{-16}$, so that it dominates the meaningful population at higher energies. In certain situations it is possible to rescue the results by identifying and removing this numerical noise by excluding it from the total population sum.

Table 3. Average CPU times to calculate the ME solution, relative to direct diagonalization in double precision. Intel Pentium 4 1.9 GHz CPU.

| Method | Variant | Relative Time | |
|---|---|---|---|
| Direct | double precision | 1.0 | |
| Diagonalization | quadruple precision | 240 | |
| | 50 digit MPFUN | 340 | |
| Numerical | LSODA $10^{-6}$ tol | 13 | |
| Integration | LSODA $10^{-2}$ tol | 6.3 | |
| | VODE $10^{-6}$ tol | 15 | |
| | VODE $10^{-2}$ tol | 10 | |
| Lanczos | double precision | 0.056 | 0.077 |
| | quadruple precision | 5.6 | 8.6 |
| | 50 digit MPFUN | 6.9 | 12 |

## 5. The Effect of Word-Size

Contrary to popular belief, there is much more to the ability of a computer to perform calculations quickly than the CPU clock speed. Issues such as memory bandwidth and cache hit rate can significantly alter the ability of the CPU to maintain high efficiency.

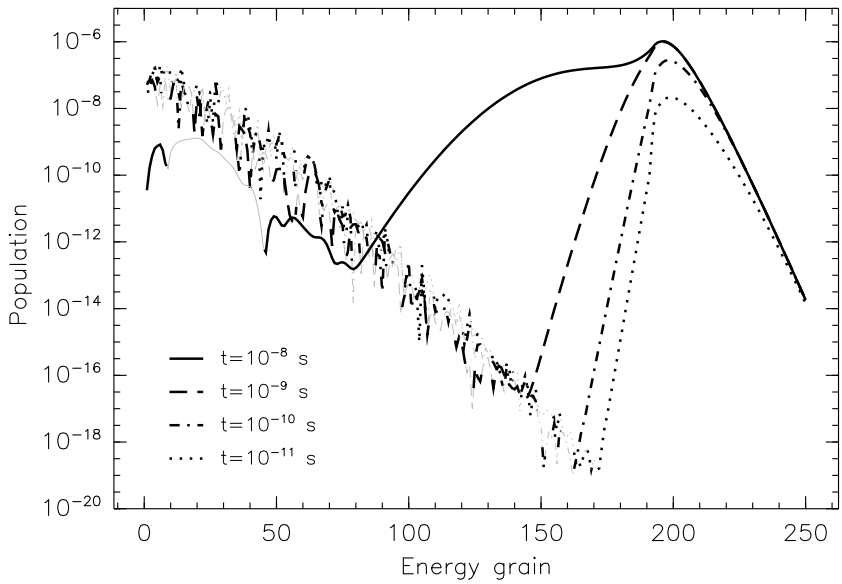


Fig. 5. Energy resolved populations for allene at various times after initiation, calculated by direct diagonalization in double precision for the ME modelling 600 K and 1000 Torr. Population elements calculated to be negative shown with light lines.

Table 4. CPU times to calculate the ME solution at 1200 K and 1000 Torr using CPUs of various architecture, relative to the direct diagonalization in double precision calculation on each architecture. Selected methods are direct diagonalization in double precision (DP DSYEV), quadruple precision (QP DSYEV) and 50 digit arithmetic (50d DSYEV) and numerical integration with LSODA with relative error tolerance of $10^{-2}$. Intel Pentium 4 1.9 GHz, Compaq Alpha 667 MHz and SGI R14000 600 MHz CPUs.

| Method | Pentium 4 | Alpha | R14000 |
|---|---|---|---|
| DP DSYEV | 1.0 | 1.0 | 1.0 |
| LSODA | 6.3 | 6.8 | 4.1 |
| QP DSYEV | 240 | 41 | 62 |
| 50d DSYEV | 340 | 745 | 700 |

The architecture of the processor also has a significant effect, as will be demonstrated next.

Table 4 shows some sample relative timings on three different systems, based on Intel Pentium 4 1.9 GHz, Compaq Alpha 667 MHz and SGI R14000 600 MHz processors. These timings are quoted relative to the double precision direct diagonalization times on each processor, which were 9.54 seconds, 6.46 seconds and 5.51 seconds, respectively. Two things stand out from the timings above and in Table 4. The first is that despite the clock speed being around a factor of three higher for the Pentium 4, the calculation on this processor took significantly longer than for the other two processors. The second is that the quadruple precision calculations took, relatively, almost an order or magnitude longer on the Pentium 4 based machine, almost as long as the 50 digit software arithmetic calculation.

Both of these observations can be rationalized by noting that while the Pentium 4 processor is based on 32 bit arithmetic, the Alpha and R14000 processors are 64 bit chips. The "double precision" calculation on the 64 bit processors are actually calculations in the processor's natural word size, whereas the Pentium 4 is indeed performing precision doubling. The "quadruple precision" calculations on the 64 bit processors are analogous to performing 64 bit arithmetic on 32 bit processors (the origin of the term double precision). On the 32 bit Pentium 4 the quadruple precision calculations involving 128 bit numbers must be decomposed twice into 32 bit calculations, an operation involving considerable overhead and disallowing many code optimizations and access to special features of the processor instruction set and architecture.

## 6. Conclusion

Two things must be considered when selecting a method to solve a large ME problem. The first is one must select a method that is actually capable of solving the problem. Selecting a method satisfying this condition is complicated by the fact that the applicability of the spectral methods (specifically the level of numerical precision required) depends on the temperature and pressure being modelled by the ME and may not be known before attempting the calculation. Even once the calculation has been performed, whether sufficient precision has been used may not be obvious. Clearly, in many applications this is not acceptable, which is why we recently proposed high precision direct diagonalization as a "black box" type method for solving all manner of ME problems without further thought.[14]

The second major consideration is the time taken to calculate the solution. The double precision spectral methods win hands down on this criteria, provided the modelled conditions are sufficiently high in temperature and pressure. While the shift-and-invert Lanczos procedure was much faster than the direct diagonalization with DSYEV, it should be pointed out that the task of determining the required number of eigenpairs for a reasonable description of the dynamics is not trivial. In this work the number of required eigenpairs was determined by altering the number of eigenpairs used in the sum of Eq. (9), starting with a complete high precision eigendecomposition. There is no indication in the eigenvalue spectrum that a subset of 25 eigenpairs yields a significantly improved solution over a subset of the smallest 20 eigenpairs.

Among the next fastest methods are the numerical integration methods, particularly when the error tolerance is reduced. Integration with LSODA with an error tolerance of around $10^{-2}$ is the fastest method capable of resolving the population profiles under all the temperature and pressure conditions tested. Additionally, numerical integration allows non-linear systems to be solved, removing the need for linearization of bimolecular reactions by invoking pseudo-first-order conditions. Hence this must become our new recommended method for solving the ME without regard for the conditions being modelled. Given the relatively loose requested error tolerance, the LSODA integrator implemented in single precision may allow significantly faster execution on 32 bit processors (which currently dominate the desktop computer market).

Spectral propagation may still offer significant advantages over numerical integration, not the least of which is the faster execution under conditions for which it is accurate. In particular if the ME application requires propagations of many initial populations under set conditions, spectral propagation may be much faster as a single eigendecomposition can quickly propagate any initial population, while the numerical integration must be rerun for each.

There still does not exist a scalable method for solving very large problems involving bimolecular reactions. All the methods used here require the explicit formation of the ME matrix and the computational effort is dominated by $n^3$ terms. The shift-and-invert Lanczos method and numerical integration do offer possible routes to scalability if the dense system solve required to calculate the effect of $B^{-1}$ is replaced with an iterative solution method. This possibility is yet to be explored.

## Acknowledgments

We gratefully acknowledge the support of the Australian Research Council in funding this work. The numerical packages LAPACK, ARPACK, MPFUN, VODE and ODEPACK used in this work are available from the Netlib Repository, http://www.netlib.org.

## References

1. R.G. Gilbert and S.C. Smith, *Theory of Unimolecular and Recombination Reactions* (Blackwell Scientific, Oxford, 1990).
2. I. Oref and D.C. Tardy, *Chem. Rev.* **90**, 1407 (1990).
3. S. Nordholm and H.W. Schranz, "Collisional energy transfer in unimolecular reactions: Statistical theory and classical simulation," in *Advances in Chemical Kinetics and Dynamics*, ed. J.R. Barker (JAI, Greenwich, 1995), Vol. 2A.
4. G.D. Billing and K.V. Mikkelsen, *Introduction to Molecular Dynamics and Chemical Kinetics* (John Wiley & Sons, New York, 1996).
5. K.A. Holbrook, M.J. Pilling and S.H. Robertson, *Unimolecular Reactions*, 2nd edn. (John Wiley & Sons, Chichester, 1996).
6. S.C. Smith and R.G. Gilbert, *Int. J. Chem. Kinet.* **20**, 307 (1988).
7. S.C. Smith and R.G. Gilbert, *Int. J. Chem. Kinet.* **20**, 979 (1988).

8. S.H. Robertson, M.J. Pilling and N.J.B. Green, *Mol. Phys.* **89**, 5131 (1996).
9. S.J. Jeffrey, K.E. Gates and S.C. Smith, *J. Phys. Chem.* **100**, 7090 (1996).
10. J.A. Miller, S.J. Klippenstein and C. Raffy, *J. Phys. Chem.* **A106**, 4904 (2002).
11. P.K. Venkatesh, A.M. Dean, M.H. Cohen and R.W. Carr, *J. Chem. Phys.* **107**, 8904 (1997).
12. K.E. Gates, S.H. Robertson, S.C. Smith, M.J. Pilling, M.S. Beasley and K.J. Maschhoff, *J. Phys. Chem.* **A101**, 5765 (1997).
13. T.J. Frankcombe, S.C. Smith, K.E. Gates and S.H. Robertson, *Phys. Chem. Chem. Phys.* **2**, 793 (2000).
14. T.J. Frankcombe and S.C. Smith, *Faraday Discuss.* **119**, 159 (2002).
15. N. Snider, *J. Chem. Phys.* **42**, 548 (1964).
16. M. Quack, *Ber. Bunsen-Ges. Phys. Chem.* **88**, 94 (1984).
17. J.A. Miller, S.J. Klippenstein and S.H. Robertson, *J. Phys. Chem.* **A104**, 9806 (2000).
18. W. Tsang, V. Bedanov and M.R. Zachariah, *Ber. Bunsen-Ges. Phys. Chem.* **101**, 491 (1997).
19. M.A. Hanning-Lee, N.J.B. Green, M.J. Pilling and S.H. Robertson, *J. Phys. Chem.* **97**, 860 (1993).
20. S.J. Klippenstein and J.A. Miller, *J. Phys. Chem.* **A106**, 9267 (2002).
21. S.J. Klippenstein, D.L. Yang, T. Yu, S. Kristyan, M.C. Lin and S.H. Robertson, *J. Phys. Chem.* **A102**, 6973 (1998).
22. S.H. Robertson, M.J. Pilling, D.A. Baulch and N.J.B. Green, *J. Phys. Chem.* **99**, 13452 (1995).
23. V.D. Knyazev and I.R. Slagle, *J. Phys. Chem.* **A105**, 3196 (2001).
24. V.D. Knyazev and I.R. Slagle, *J. Phys. Chem.* **A105**, 6490 (2001).
25. P.W. Seakins, S.H. Robertson, M.J. Pilling, I.R. Slagle, G.W. Gmurczyk, A. Bencsura, D. Gutman and W. Tsang, *J. Phys. Chem.* **97**, 4450 (1993).
26. J.R. Barker, *Chem. Phys.* **77**, 301 (1983).
27. J. Shi and J.R. Barker, *Int. J. Chem. Kinet.* **22**, 187 (1990).
28. J.R. Barker and K.D. King, *J. Chem. Phys.* **103**, 4953 (1995).
29. L. Vereecken, G. Huyberechts and J. Peeters, *J. Chem. Phys.* **106**, 6564 (1997).
30. J.A. Miller and C.F. Melius, *Combust. Flame* **91**, 21 (1992).
31. J.D. Adamson, C.L. Morter, J.D. DeSain, G.P. Glass and R.F. Curl, *J. Chem. Phys.* **100**, 2125 (1996).
32. W. Forst, *J. Phys. Chem.* **76**, 342 (1972).
33. W. Forst, *J. Phys. Chem.* **83**, 100 (1979).
34. W. Forst, *J. Phys. Chem.* **86**, 1771 (1982).
35. J.W. Davies, N.J. Green and M.J. Pilling, *Chem. Phys. Lett.* **126**, 373 (1986).
36. L.B. Harding and S.J. Klippenstein, *Proc. Combust. Inst.* **28**, 1503 (2000).
37. M. Karni, I. Oref, S. Barzilai-Gilboa and A. Lifshitz, *J. Phys. Chem.* **92**, 6924 (1988).
38. M. Blitz, M.S. Beasley, M.J. Pilling and S.H. Robertson, *Phys. Chem. Chem. Phys.* **2**, 805 (2000).
39. T.J. Frankcombe and S.C. Smith, *J. Comput. Chem.* **21**, 592 (2000).
40. E. Anderson, Z. Bai, C. Bischof, S. Blackford, J. Demmel, J. Dongarra, J. Du Croz, A. Greenbaum, S. Hammarling, A. McKenney and D. Sorensen, *LAPACK Users Guide*, 3rd edn. (SIAM, Philadelphia, 1999).
41. J.K. Cullum and R.A. Willoughby, *Lanczos Algorithms for Large Symmetric Eigenvalue Computations* (Birkhäuser, Boston, 1985), Vol. I.
42. R.B. Lehoucq, D.C. Sorensen and C. Yang, *ARPACK Users' Guide*: *Solution of Large Scale Eigenvalue Problems with Implicitly Restarted Arnoldi Methods* (SIAM, Philadelphia, 1998).
43. N.J.B. Green, S.H. Robertson and M.J. Pilling, *J. Chem. Phys.* **100**, 5259 (1994).
44. S.H. Robertson, A.I. Shushin and D.M. Wardlaw, *J. Chem. Phys.* **98**, 8673 (1993).
45. T.J. Frankcombe and S.C. Smith, *Comput. Phys. Commun.* **141**, 159 (2001).
46. D.H. Bailey, *ACM Trans. Math. Software* **21**, 379 (1995).
47. P.N. Brown, G.D. Byrne and A.C. Hindmarsh, *SIAM J. Sci. Stat. Comput.* **10**, 1038 (1989).
48. J.A. Miller and D.W. Chandler, *J. Chem. Phys.* **85**, 4502 (1986).
49. D.W. Chandler and J.A. Miller, *J. Chem. Phys.* **81**, 4105 (1984).
50. J.A. Miller and S.J. Klippenstein, *J. Phys. Chem.* **A105**, 7254 (2001).
51. D.K. Hahn, S.J. Klippenstein and J.A. Miller, *Faraday Discuss.* **119**, 79 (2001).
52. A.C. Hindmarsh, "ODEPACK, a systematized collection of ODE solvers," in *Scientific Computing*, ed. R.S. Stepleman (North-Holland, Amsterdam, 1983).
53. L.R. Petzold, *SIAM J. Sci. Stat. Comput.* **4**, 136 (1983).
54. J.A. Miller and S.J. Klippenstein, *Proc. Combust. Inst.* **28**, 1479 (2000).
55. J.A. Miller and S.J. Klippenstein, *Int. J. Chem. Kinet.* **33**, 654 (2001).