THE JOURNAL OF CHEMICAL PHYSICS **143**, 094106 (2015)

Sampling reactive pathways with random walks in chemical space: Applications to molecular dissociation and catalysis

Scott Habershon[a)]
Department of Chemistry and Centre for Scientific Computing, University of Warwick, Coventry CV4 7AL, United Kingdom

(Received 22 June 2015; accepted 21 August 2015; published online 3 September 2015)

Automatically generating chemical reaction pathways is a significant computational challenge, particularly in the case where a given chemical system can exhibit multiple reactants and products, as well as multiple pathways connecting these. Here, we outline a computational approach to allow automated sampling of chemical reaction pathways, including sampling of different chemical species at the reaction end-points. The key features of this scheme are (i) introduction of a Hamiltonian which describes a reaction "string" connecting reactant and products, (ii) definition of reactant and product species as chemical connectivity graphs, and (iii) development of a scheme for updating the chemical graphs associated with the reaction end-points. By performing molecular dynamics sampling of the Hamiltonian describing the complete reaction pathway, we are able to sample multiple different paths in configuration space between given chemical products; by periodically modifying the connectivity graphs describing the chemical identities of the end-points we are also able to sample the allowed chemical space of the system. Overall, this scheme therefore provides a route to automated generation of a "roadmap" describing chemical reactivity. This approach is first applied to model dissociation pathways in formaldehyde, H_2CO, as described by a parameterised potential energy surface (PES). A second application to the $HCo(CO)_3$ catalyzed hydroformylation of ethene (oxo process), using density functional tight-binding to model the PES, demonstrates that our graph-based approach is capable of sampling the intermediate paths in the commonly accepted catalytic mechanism, as well as several secondary reactions. Further algorithmic improvements are suggested which will pave the way for treating complex multi-step reaction processes in a more efficient manner. © *2015 AIP Publishing LLC*. [http://dx.doi.org/10.1063/1.4929992]

I. INTRODUCTION

The most common experimental methods to determine chemical reaction mechanisms generally rely on inference from *indirect* observations. For example, measurements of reaction rates, and the corresponding reaction enthalpies and entropies, as a function of macroscopic variables such as temperature and pressure can give an insight into the likely reaction pathways.[1] The effects of changing functional groups, isotopic identities, or solvent characteristics can similarly provide valuable information which can often be subsequently used to deduce reaction mechanisms.[2–4] Furthermore, in favourable cases, kinetic trapping may allow determination of selected structures lying along the pathway from reactants to products.[5–7] Of course, several modern experimental approaches can be viewed as providing a (almost) direct view of chemical reactions, including time-resolved diffraction[8–10] and ultrafast spectroscopy;[11–13] however, these methods still rely on extensive data processing[14–16] and, often, chemical intuition in elucidating the underlying chemical dynamics. Furthermore, these direct investigations of the structural or electronic properties of molecules are significantly hindered as molecular complexity increases or as a solvent environment is introduced.[17]

In contrast, computer simulations provide a *direct* route to determining chemical reaction mechanisms with atomic resolution.[18,19] Over the last few decades, a wide variety of computational strategies have been developed specifically with the purpose of determining reaction pathways and the associated energetic barriers. For example, the Nudged Elastic Band (NEB) method[20,21] and related approaches,[22–24] the zero- and finite-temperature string (FTS) methods,[25,26] and the growing string approach[27] all aim to search for a reaction path given an initial guess path connecting *specified* reactant and product configurations; the constraint of generating a good initial guess is removed in methods such as Gradient Extremal Following (GEF),[28,29] Scaled Hypersphere Searching (SHS[30–32]), and reduced gradient following (RGF).[33] Rather than focussing on the search for single reaction paths, methods such as transition path sampling[34–43] and Onsager-Machlup path sampling[44,45] can instead generate ensembles of reaction pathways; subsequent analysis of the path ensemble, for example, by calculation of commitor probabilities, allows further identification of important features associated with transition states (TSs). More recently, the Artificial Force-Induced Reaction (AFIR) method,[46,47] in which a biasing force is used to drive chemical bond breaking and formation, has been developed as an

a)Electronic mail: S.Habershon@warwick.ac.uk

automated approach to generating reaction pathways and stationary points. This wide variety of reaction-path-finding methods has been employed to model a similarly wide variety of chemical processes; representative examples include rearrangements in molecular clusters,[37] metal-catalyzed hydroformylation reactions,[47] protein folding problems,[40,48] and enzyme-catalyzed reactions.[49] Furthermore, these examples demonstrate that such approaches are readily applied to model chemical reaction paths in both gaseous and (implicit or explicit) condensed-phase systems.

Despite these clear successes, several challenges remain for computational approaches to predicting mechanistic chemistry. Several methods, particularly those based on a chain-of-states representation, require well-defined configurations for the initial and final points in the reaction process, as well as sensible initial guesses for the initial reaction path; this restriction can represent a barrier to the discovery of previously unknown reaction chemistry. Many string-based approaches are also incompatible with modelling systems which can exhibit multiple different reaction pathways; in such cases, an initial guess for all of the various reactive pathways would be required. Furthermore, methods such as GEF are computationally demanding, often as a result of requiring the Hessian as input, presenting challenges in modelling large systems.

In this article, we propose and test a method which attempts to address these difficulties. The approach outlined here has the following features: (i) a chain-of-states (or string-based) description of a reaction path is employed, where the string is parameterised by a set of Fourier coefficients,[45,50] (ii) the initial and final configurations of the reaction path, as well as the Fourier coefficients defining the path, are dynamic, facilitating sampling of multiple reactive pathways, (iii) the chemical connectivity of the initial and final configurations is defined in terms of mathematical graphs, and a random walk in the space of allowed chemical graphs is employed to allow description of multiple reactive pathways, and finally (iv) periodic application of reaction-path refinement methods, in this case NEB, is used to generate representative reaction paths for all allowed chemical reactions sampled in the system. Overall, this approach allows sampling of multiple reactive pathways between multiple reactant and product types and, by constraining the chemical identities (i.e., connectivity graphs) of reactants and products, we also have a simple route to constraining the reaction-path search to a desired region of *chemical* space.

As the first tests of our methodology, we consider two different systems. First, our approach is applied to model dissociation and isomerization in formaldehyde. In the significantly more challenging second application, we consider the catalytic cycle associated with hydroformylation of ethene by $HCo(CO)_3$,[47,51,52] a reaction involving multiple molecular fragments and multiple reactive steps. In both cases, we find that the chemically relevant reaction processes are sampled, although it is found some pathways, such as *cis-trans* HCOH isomerization in formaldehyde, are not sampled because generation of these specific end-points is a "rare" event. These results point the way towards further refinements of our approach, such as enhanced sampling methods for the reaction-path end-points.

II. METHODOLOGY

An overview of the simulation approach developed here is shown in Fig. 1. This approach combines a dynamic description of the reaction path with stochastic changes in the chemical connectivity of the reactants and products; this algorithm therefore allows sampling of multiple pathways connecting multiple different product and reactant states. The individual parts which constitute this approach are outlined below.

A. Reaction path definition and Hamiltonian sampling

Following on from chain-of-states methods for modelling reaction paths such as NEB, we define a reaction pathway as a string leading from reactants $\mathbf{r}_0$ to products $\mathbf{r}_P$; along this string, we have M intermediate configurations. Each state along the string represents a configuration in the $(3N_a - 6)$-dimensional configurational space of the system, where N_a is the number of atoms. However, in contrast to standard NEB-type methods, the string here is parameterised by a set comprising P Fourier coefficients for each of the $(3N_a - 6)$ degrees-of-freedom, $\{\mathbf{a}_k\}_{k=1}^{P}$, such that[45,50]

$$\mathbf{r}_i = \mathbf{r}_0 + \lambda_i(\mathbf{r}_P - \mathbf{r}_0) + \sum_{k=1}^{P} \mathbf{a}_k \sin(k\pi\lambda_i). \tag{1}$$

Here, $\lambda_i \in \{0,1\}$ is a variable which describes position along the string between reactants and products, and $\mathbf{r}_i$ represents the position vector of the ith configuration along the reaction string. Overall, the reaction-path system is defined by $(3N_a - 6)$ coordinates for each of the reaction path end-points, $\mathbf{r}_0$ and $\mathbf{r}_P$, as well as $P \times (3N_a - 6)$ Fourier coefficients which define the path through Eq. (1). We note that we favour the Fourier parameterization of the reaction path here, compared to a standard chain-of-states representation as commonly employed in NEB-type methods, because the resulting pathways tend to be smoother in configurational space; for example, initial tests of the approach described below using a standard chain-of-states representation demonstrated that the reaction pathway can develop "kinks" which can cause failures of reaction-path refinement methods such as NEB.

As well as being defined by coordinates $\mathbf{r}_0$ and $\mathbf{r}_P$, the product and reactant states also have associated with them connectivity graphs $\mathbf{G}^0$ and $\mathbf{G}^P$, respectively. These graphs are $N_a \times N_a$ matrices with elements

$$G_{ij} = \begin{cases} 1 & \text{if } r_{ij} < r_{ij}^{cut}, \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

Here, r_{ij} is the distance between atoms i, j and r_{ij}^{cut} is a cutoff distance; in this work, the cutoff distances depend only on the atomic types of i,j, and are given in Table I. The graphs $\mathbf{G}^0$ and $\mathbf{G}^P$ therefore define the chemical connectivity of reactant and product states; these graphs play an important part in exploring the space of allowed chemical reaction pathways in our approach. At this point, it is important to emphasize that graphs in our approach are used to drive the search of chemical reaction space, as described below; in a recent report using basin-hopping Monte Carlo (BHMC) to determine reaction intermediates,[52] graphs were simply used as a post-processing

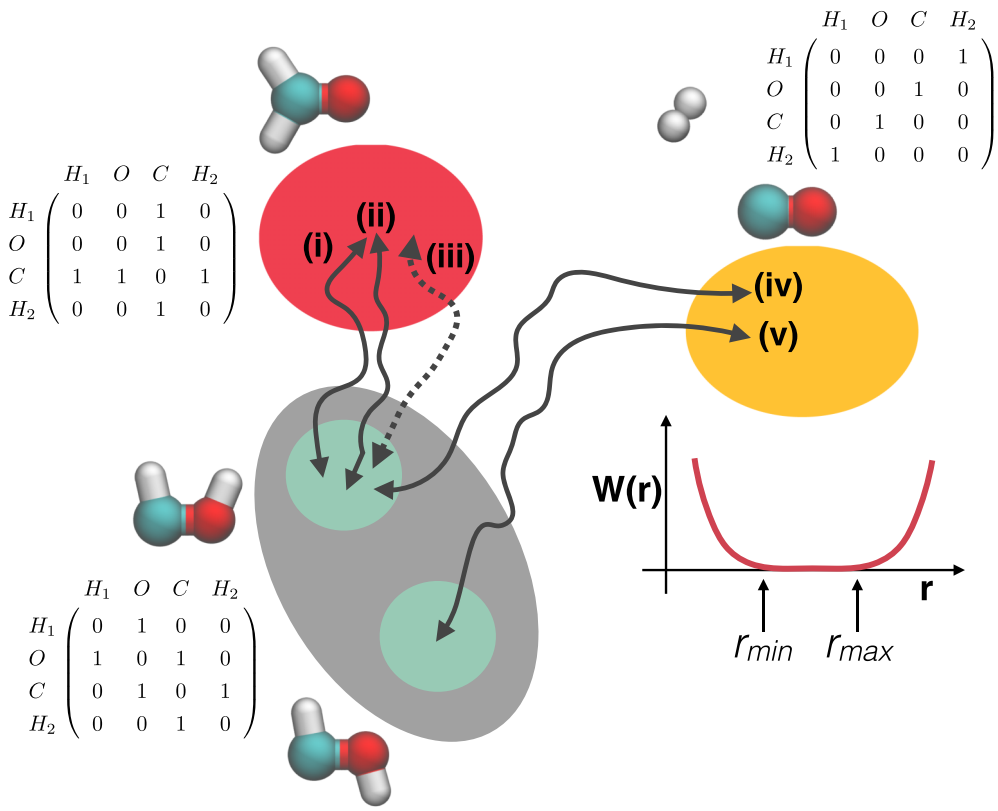

FIG. 1. Schematic illustration of a typical graph-based reaction-path sampling simulation. The labels (i)-(v) indicate possible states of the system at different times during the simulation. Initially (path (i)) the reaction path connects the *cis* isomer of the HCOH to H_2CO; the respective connectivity graphs for these end-points are shown on the left-hand side of the figure. Following dynamic evolution of the reaction path under the Hamiltonian of Eq. (3), an alternative path (ii) is sampled; however, note that the graphs $\mathbf{G}^0$ and $\mathbf{G}^P$ describing the chemical bonding of the end-points have not changed, so the path is restricted to sample the same end-points as path (i). After a period of time t_{NEB}, NEB refinement starting from the current reaction path is performed, resulting in a reaction path labelled (iii). As the simulation proceeds, graph moves will update the chemical connectivity graph of the reaction end-points; in path (iv), the graph for one of the end-points has been changed from one representing to H_2CO to a graph representing $H_2 + CO$ and, because the constraint surface $W(\mathbf{r}, \mathbf{G})$ (illustrated in the lower-right corner) enforces this connectivity on the sampled end-points, the system now naturally samples reaction paths connecting HCOH and $H_2 + CO$. Finally, path (v) has the same graphs at each end-point as path (iv), but illustrates an alternative situation where the *trans* isomer has been sampled as the HCOH end-point.

analysis tool, rather than being part of the simulation protocol itself.

To facilitate sampling of multiple reactive pathways, we now define a Hamiltonian which describes the reaction string. In particular, we assign fictitious masses and momenta to the Fourier coefficients which parameterize the reaction path, and use the following Hamiltonian:

$$H(\mathbf{r}_0, \mathbf{r}_P, \mathbf{p}_0, \mathbf{p}_P, \mathbf{a}, \mathbf{G}^0, \mathbf{G}^P) = \sum_{i=1}^{N_a} \frac{|\mathbf{p}_0^{(i)}|^2}{2m_i} + \sum_{j=1}^{N_a} \frac{|\mathbf{p}_P^{(j)}|^2}{2m_j} + \sum_{k=1}^{P} \frac{|\mathbf{b}^{(k)}|^2}{2\mu} + V_s(\mathbf{r}_0, \mathbf{r}_P, \mathbf{a}, \mathbf{G}^0, \mathbf{G}^P). \quad (3)$$

The first two terms in Eq. (2) are, respectively, the total kinetic energy (KE) of the reactant and product configurations while the third term represents the KE contribution from the momenta $\{\mathbf{b}_k\}_{k=1}^P$ conjugate to the Fourier coefficients $\mathbf{a}$; the Fourier coefficient momenta are assigned an arbitrary mass μ, which is here assumed to be the same for all coefficients. The potential energy of the string system, $V_s(\mathbf{r}_0, \mathbf{r}_P, \mathbf{a}, \mathbf{G}^0, \mathbf{G}^P)$, is defined as

$$V_s(\mathbf{r}_0, \mathbf{r}_P, \mathbf{a}, \mathbf{G}^0, \mathbf{G}^P) = V(\mathbf{r}_0) + V(\mathbf{r}_P) + W(\mathbf{r}_0, \mathbf{G}^0) + W(\mathbf{r}_P, \mathbf{G}^P) + \frac{1}{M} \sum_{k=1}^{M} \left[V(\mathbf{r}_k) + \gamma_1 |\mathbf{r}_k - \mathbf{r}_{k-1}|^2 \right], \quad (4)$$

TABLE I. Parameters used in the reaction string Hamiltonian MD simulations and NEB optimization.

μ	10^5 m$_e$
γ_1	0.05 $E_h a_0^{-2}$
σ_1	0.01 $E_h a_0^{-2}$
σ_2	0.01 $E_h a_0^{-2}$
σ_3	4.0 a_0
σ_4	5×10^{-3} $E_h a_0^{-2}$
R^{max}	10.0 Å
R^{min}	5.0 Å
P_u	5×10^{-4}
k_s	0.05 $E_h a_0^{-2}$

Atom pair	r^{cut} (Å)	r^{min} (Å)	r^{max} (Å)
O–H	1.2	0.9	1.15
C–H	1.2	0.9	1.15
C–O	1.5	1.15	1.45
H–H	1.0	0.8	0.95

where $V(\mathbf{r})$ is the potential energy surface (PES) of the system evaluated at $\mathbf{r}$. The summation over the M intermediate states includes contributions from the PES at each intermediate configuration given by Eq. (1); we note that the configurations at each of the M intermediate states are given by Eq. (1), giving rise to a dependence on the Fourier coefficients $\mathbf{a}$. Equation (4) also contains harmonic "spring" terms with force constant γ_1, which act to minimise the distance between adjacent configurations along the string. These harmonic terms are familiar from path integral (PI) simulations,[53–59] although the force constant here is arbitrarily chosen to maintain a stable and continuous reaction pathway, and is not related to either the mass or temperature of the system as in standard PI treatments. However, as an aside, we note that methodologies for enhanced PI simulations, such as normal-mode transformations or advanced thermostats,[58,60,61] could in the future be exploited to improve sampling. Furthermore, we note that an alternative formulation of Eq. (4) based on Fourier PI simulations[62,63] could include a term which is a simple sum-of-squares in the Fourier coefficients instead of the position-based spring terms here; in simple test simulations, we found that the current approach resulted in a numerically more stable simulation overall.

The function $W(\mathbf{r},\mathbf{G})$ is a constraint surface which enforces the chemical connectivity graph $\mathbf{G}$ for the configuration $\mathbf{r}$. As a result, this function ensures that the chemical structure of configuration $\mathbf{r}$ cannot be changed beyond that defined in the graph $\mathbf{G}$. In other words, the only way in which the chemical identity of $\mathbf{r}$ can change is by changes in $\mathbf{G}$; this fact is exploited to explore chemical reaction space, as described below. The choice of constraint terms in $W(\mathbf{r},\mathbf{G})$ is somewhat arbitrary so, for simplicity, we use harmonic terms to enforce chemical bonding, with a simple repulsive Gaussian term to enforce non-bonding between atom pairs for which $G_{ij} = 0$. Overall, the constraint function used in this work is

$$W(\mathbf{r},\mathbf{G}) = \sum_{j>i}\left[\delta(G_{ij}-1)[H(r_{ij}^{min}-r_{ij})\sigma_1(r_{ij}^{min}-r_{ij})^2 + H(r_{ij}-r_{ij}^{max})\sigma_1(r_{ij}^{max}-r_{ij})^2] + \delta(G_{ij})\sigma_2 e^{-r_{ij}^2/(2\sigma_3^2)}\right] + V_{mol}(\mathbf{r},\mathbf{G}). \quad (5)$$

The summation in Eq. (5) runs over all pairs of atoms, $\delta(x)$ is the Dirac delta function, $H(x)$ is the Heaviside step function, and σ_1, σ_2, and σ_3 are variables describing the constraint interactions. The effect of the potential $W(\mathbf{r},\mathbf{G})$ can be seen to constrain *bonded* atoms to lie at bond lengths approximately between r^{min} and r^{max}, while a repulsive Gaussian potential acts between pairs of *non-bonded* atoms. In this way, the chemical graph $\mathbf{G}$ is straightforwardly imposed on the configuration $\mathbf{r}$. Similarly, the *molecular* constraint term $V_{mol}(\mathbf{r},\mathbf{G})$ ensures that graphs describing one or more distinct molecules cannot sample configurations which are incompatible with the expected number of molecules; in short, distinct molecules are kept sufficiently far apart that there is no possibility for them to form new chemical bonds which are not consistent with the desired graphs describing reactants or products. In this work, we use simple harmonic constraints to enforce separation of unique molecules, such that

$$V_{mol}(\mathbf{r},\mathbf{G}) = \sum_{\substack{j>i \\ m_i(\mathbf{G})\neq m_j(\mathbf{G})}} [H(R^{min}-r_{ij})\sigma_4(R^{min}-r_{ij})^2, + H(r_{ij}-R^{max})\sigma_4(R^{max}-r_{ij})^2], \quad (6)$$

where R^{min} and R^{max} are, respectively, minimum and maximum allowed distances between a pair of atoms, each in a different molecule. In calculating this constraint term, the label m_i identifies the molecule which each atom is assigned to; these labels can be readily determined given the relevant connectivity graphs. Finally, we emphasise that the constraint potentials of Eqs. (5) and (6) enforce only "loose" constraints on the system in order to restrict sampling to the correct region of chemical space, as defined by the relevant connectivity graph. For example, in the calculations illustrated below, the constraint potential of Eq. (5) is chosen to maintain bonded O–H distances between 0.9 Å and 1.15 Å; in other words, the constraint potential does not act until the bond distances start to drift significantly from the average value of around 1 Å.

Overall, Eq. (3) describes a reaction-path system which can be sampled using Hamilton's equations-of-motion;[19,64] analytical derivatives are readily calculated for the configurations $\mathbf{r}_0$ and $\mathbf{r}_P$, as well as for the set of Fourier coefficients $\mathbf{a}$, allowing straightforward application of standard molecular dynamics (MD) integration algorithms. By assigning arbitrary temperatures to the reactant, product, and Fourier coefficient momenta, it is therefore possible to explore the space of possible chemical reaction pathways. We note that the configurations sampled by the reactant and product states will not correspond exactly to the expected thermal distributions as a result of the additional constraint and harmonic spring terms of Eq. (3). However, this Hamiltonian approach *will* provide a means to sample approximate reaction paths which can be further refined by methods such as NEB.[20,21]

As a final point, we note that the method as presented here operates exclusively in Cartesian coordinates; this is predominantly for reasons of computational simplicity. However, there is no reason that the same method cannot be applied in alternative sets of coordinates, as long as one could easily transform to coordinates which would allow one to calculate the required PES and derivatives required in Eqs. (3) and (4). Exploiting alternative coordinate choices may help improve sampling of reactive pathways; this point is discussed later.

B. Random walks in chemical space

Sampling according to the Hamiltonian of Eq. (3) allows one to generate approximate representative reaction paths between a range of reactant and product *configurations* (or conformations); however, in order to fully explore the chemistry of a reactive system, this is not sufficient. In particular, the *chemical identities* (i.e., chemical isomers) of reactants and products must also be allowed to change; in other words, we need a method for exploring the chemical space spanned by $\mathbf{G}^0$ and $\mathbf{G}^P$. In this work, we choose to use a simple Monte Carlo (MC)-like procedure to move around the chemical space defined by the graphs $\mathbf{G}^0$ and $\mathbf{G}^P$. However, we note that the usual restriction of detailed balance which applies in standard MC sampling of atomic configurations is not essential here; all we are interested in is moving around configurational and chemical space in order to generate approximate reaction paths to be used as input for further refinement. As a result, the most important consideration in moving around the chemical space of reactants and products is the restriction to chemically relevant species; this is reasonably straightforward to achieve given the connectivity graphs $\mathbf{G}^0$ and $\mathbf{G}^P$.

Our approach to updating the graphs $\mathbf{G}^0$ and $\mathbf{G}^P$ is summarised as follows.

1. At each time step, for both $\mathbf{G}^0$ and $\mathbf{G}^P$, compare a uniform random number $\eta \in [0,1]$ to a user-defined update probability P_u. If $\eta \leq P_u$, then the end-point graph is updated; if $\eta > P_u$, then we resume time-propagation of the reaction string.
2. Store the current graphs, string Fourier coefficients, and end-point coordinates.
3. To update the graph $\mathbf{G}$, we apply one of two graph moves with equal probability:
   - Select a random off-diagonal element G_{ij}; if $G_{ij} = 1$, replace it with 0, and if $G_{ij} = 0$ replace it with 1.
   - Select two unique off-diagonal elements G_{ij} and G_{kl} with different values, and swap them: $G_{ij} \rightarrow G_{kl}$ and $G_{kl} \rightarrow G_{ij}$.

   We note that these graph moves are appropriate for our initial application to formaldehyde reactivity; updated graph moves in the study of the catalytic hydroformylation cycle are discussed below.
4. Check that the new end-point graphs are chemically allowed (see below); if not, revert updated graphs to their stored values and resume time-propagation.
5. Optimise the coordinates of the reactant and product configurations such that the constraint potential terms $W(\mathbf{r},\mathbf{G})$ (calculated with the new graphs) are minimized.
6. Optimise the coordinates of the reactant and product configurations such that the PES terms $V(\mathbf{r})$ are minimized.
7. Reset the path coefficients $\mathbf{a}$ and check whether a new initial linear path between reactants and products results in any atomic close contacts along the reaction pathway with $r_{ij} < 0.7$ Å. If so, revert the updated graphs, coordinates, and string coefficients to their stored values and carry on with time-propagation.
8. With fixed reactant and product configurations, minimize the PES of Eq. (4) by varying the Fourier coefficients $\mathbf{a}$.
9. Resample the momenta of the reactant state, product states and Fourier coefficients, and resume time-propagation.

The outcome of successful graph updates is the generation of new reactant or product configurations with modified chemical bonding; furthermore, the coordinates of the reactant and product states, as well as the Fourier coefficients, are all relaxed such that the overall system is in a reasonable, low potential energy state appropriate as a starting point for further time-evolution.

As noted in the outline above, there are two points at which the new graphs may be rejected. First, if an initial linear path between reactants and products causes severe atomic overlaps the graph move is ignored. This requirement is not an essential part of our approach; for example, subsequent relaxation of the Fourier coefficients should remove any steric clashes. However, in practice, we find that this optimization process can become unstable due to large potential energy gradients in cases where atoms experience close contacts at some point along the reaction pathway; as a result, we adopt the simpler approach of simply rejecting these graph moves.

Second, we reject reactant or product graphs if they are not "chemically sensible." Given typical atomic valences, it is straightforward to ensure that unfeasible bonding configurations are not imposed on the system. For example, we know that the number of covalent chemical bonds formed by hydrogen should be no greater than one, so graphs which fail to comply with this valence sum can be rejected; similar valence rules are simple to impose for heteroatoms. Note that this approach does not necessarily restrict us to exploring only "known" chemical pathways; one can always relax the constraints on atomic valence types to allow formation of chemically exotic bonding arrangements. However, we note that a significant advantage of encoding bonding arrangements as graphs is the fact that one can "tune" the types of chemistry which can be explored during the path-sampling simulations. For example, it is straightforward to sample only those pathways in which selected carbon atoms remain sp^2 hybridized while others remain sp^3 hybridized; in future applications, we hope to exploit this flexibility in searching for reactive pathways in complex systems.

C. Nudged elastic band method

The approach outlined above allows sampling of approximate reaction pathways between two different chemical species, as well as changes in the identities of the chemical species at either end of the reaction pathway. These reaction paths are good candidates to use as initial guesses for further refinement. Combining sampling of the reaction paths and the associated chemical graphs describing the reactants and products with a method for refining reaction paths therefore gives an approach capable of automatically exploring allowed chemical reaction mechanisms for a given system. However, we emphasise that the focus of this work is firmly on methods for generating approximate reaction paths which can be used as input for further refinement and analysis.

In this work, we employ the standard NEB method[20,21] to refine the initial paths generated by the sampling methodology

described above; the relevant equations outlining the NEB approach are given in the Appendix. While NEB may not be particularly accurate in locating TSs, it is computationally simple and is straightforward to implement within the framework outlined here. We emphasise that our overall approach is *not* tied to the use of NEB; any of the standard approaches for analysing reaction paths, including improved methods such as climbing image NEB (CI-NEB),[23] could be employed. Furthermore, we are not limited to considering "zero temperature" methods; the initial reaction pathways could also be employed as input to methods such as the finite-temperature string approach.[26] As an aside, we note that one might be more interested in generating so-called minimum-action paths which correspond to (classical) dynamical trajectories between defined end-point conformations, rather than the NEB-type paths generated in this work. In such cases, the NEB refinement employed here could be replaced by methods based on optimizing the Onsager-Machlup or Maupertuis action;[45,50] these alternative reaction "curves" will be investigated in the future.

Finally, we note that some correlation between sampled reaction pathways may be present in our approach (except in the obvious case when t_{NEB} is very large). While correlation between pathways is undesirable from a sampling point-of-view, it is not as important in this work as it is in MD simulations. This is because we are interested in generating starting points for NEB-type refinement; if several similar pathways are sampled, the only drawback is wasted computer time. Finally, we note that the graph rearrangement moves which periodically take place also serve to reset the configurations of the system, further avoiding stagnation.

III. APPLICATIONS

In this section, we consider application of our graph-based reaction-path sampling methodology to two different chemical systems. In the first case, we investigate dissociative reactions of formaldehyde using a parameterized PES.[65] In a second application, we consider the hydroformylation of ethene catalyzed by $HCo(CO)_3$;[47,51,52] this complex reaction, involving multiple fragments and four main reaction steps in the accepted catalytic mechanism, represents a strong challenge to any reaction-path finding methodology.

In both systems, the aim of the simulations reported here is to investigate whether the graph-sampling scheme outlined above is capable of extracting good first approximations to the main chemical reactive pathways in the systems under consideration. While sampled pathways are refined using NEB, we are not interested in explicitly locating TS configurations. Indeed, it is intended that the reactive pathways generated with the current methodology are used as input to a second simulation step in which a much higher level of theory is used to determine TS configurations (starting from the approximate structures obtained by NEB) and transition-state theory rate constants for each sampled reactive pathway. These parameters would then allow direct kinetic modelling of the multiple reactive pathways discovered here, affording a route towards understanding mechanism, particularly in the case of catalytic cycles. These developments will be reported in a future publication; the aim of this work is to evaluate the potential of our graph-sampling approach and highlight areas for future development.

A. Formaldehyde decomposition

As the first application of our reaction path sampling approach, we model the chemistry of gas-phase formaldehyde, H_2CO. This molecule has been used as a test example for several reaction path finding methods, including SHS,[30] RGF,[33] and GEF;[66] these approaches have been used to map out the whole range of chemical reactivity, including dissociation of molecular and atomic hydrogen, as well as *cis-trans* isomerization of HCOH.

Previous calculations on the formaldehyde system have employed a variety of *ab initio* approaches to model the PES, including Density Functional Theory[30] and Hartree-Fock calculations.[33,66] In this work, we employ the PES developed by Bowman and coworkers;[65] this is a fit to around 80 000 CCSD(T) calculations and 53 000 multireference configuration interaction (MRCI) calculations. The functional form itself is comprised of five local fits to different regions of the PES, with smooth switching functions interpolating between them; in general, each of the local fits is based on expansion in a set of Morse variables. Overall, this PES provides a very accurate description of much of the chemistry associated with rearrangements of formaldehyde, with typical RMS relative errors in fitting of the order of 3%. The formaldehyde PES has also been shown to be able to describe the dissociation dynamics observed experimentally, including identification of a "roaming" mechanism.[67]

Our choice of PES imposes some constraints on the reaction channels which can be accurately modelled. The formaldehyde PES employed here[65] describes the H_2CO minimum, the molecular dissociation (H_2 + CO), the radical dissociation channel (H + HCO), and both *cis* and *trans* isomers of HCOH; it does not describe other reaction channels, such as CH + OH. As a result, the following constraints on the reactant and product graphs were imposed to ensure that the reaction pathway stays within the region of configuration space in which the PES is reliable:

- The number of covalent bonds formed by hydrogen can be no greater than 1: $\sum_j G_{ij} \leq 1$, where the atomic type of atom index i is hydrogen.
- The number of covalent bonds formed by oxygen can be no greater than 2: $\sum_j G_{ij} \leq 2$, where the atomic type of atom index i is oxygen.
- The total number of molecules described by the reactant and product graphs must be no greater than two; the number of distinct molecules encoded by a graph can be straightforwardly determined by application of the Floyd-Warshall shortest-path finding algorithm.[68,69]
- Graphs which describe (CH + OH), (H + COH), and dissociation of monatomic heteroatoms are forbidden.

These graph constraints are simple to impose in our sampling approach; during sampling of the chemical graphs, new product or reactant graphs which violate these constraints are rejected, the graph is reverted to its original form and sampling continues as normal.

The remaining calculation details were as follows. The reaction string comprised a total of 50 configurations, including reactant and product states (so $M = 48$). The time-evolution of the positions and momenta of the reactant and product, as well as the Fourier coefficients and their associated momenta, was performed using the standard velocity Verlet algorithm[19,64,70] with a time step of 0.1 fs. The total simulation time was 50 ps, and three separate calculations were performed to sample the space of chemical reaction pathways. Initial physical and fictitious (Fourier) momenta were drawn from the Boltzmann distribution, and an Anderson thermostat[19,70] was employed throughout time-evolution to maintain the physical temperatures of the reactant and product configurations, and the fictitious temperature of the Fourier coefficients, at 100 K. The fictitious masses associated with the Fourier coefficients were set to a value of 10^5 a.u.; the results obtained seem to be quite insensitive to the actual value of this mass parameter. To prevent overall translation or rotation of the system, we fix the position of the carbon atom at the origin, constrain the oxygen to lie along the x-axis, and constrain one of the hydrogen atoms to move in the xy-plane. The values of the remaining parameters of the Hamiltonian employed in this work are shown in Table I. NEB optimization starting from the current reaction string was performed every 1 ps. We used a steepest descents optimization to obtain adequately converged reaction paths; this approach could, of course, be refined to include the CI-NEB method to locate the TS more accurately, or alternative reaction-path optimization methods such as the string method could similarly be employed. Optimization under the graph constraints, as required during our chemical graph sampling approach, was also performed using a steepest descents approach; accurate minimisation is not required in this case because the aim is to simply generate a reasonable molecular geometry which satisfies the desired graph constraints. As a final point, we note that all simulations were initialized with formaldehyde as the reactant and the molecular channel H_2 + CO as the products; at no point was the system "seeded" with information about any other reaction pathways or PES minima.

Figure 2 illustrates a sample of the reaction paths determined by our path-generation approach; in particular, three paths originating from formaldehyde are shown. As expected, these paths cover the range of chemical transformations for formaldehyde, given that some reaction channels are not allowed on the PES employed here. Figures 2(a) and 2(b) illustrate the molecular and radical dissociation channels, respectively; consistent with previous calculations, the molecular dissociation channel displays a TS around 371 kJ mol^{-1} higher in energy than formaldehyde, while the energy of the radical products lies at about 395 kJ mol^{-1} higher in energy. The (approximate) TS determined for the molecular channel [Fig. 2(a)] displays the expected asymmetric geometry; the accuracy of the transition-state geometry and energy could be improved by using, for example, eigenvector following methods or CI-NEB.

Figure 2(c) illustrates one of the strengths of the sampling approach outlined here. In particular, Fig. 2(c) shows an alternative reaction path leading from formaldehyde reactant to radical products; in other words, the reactants and products

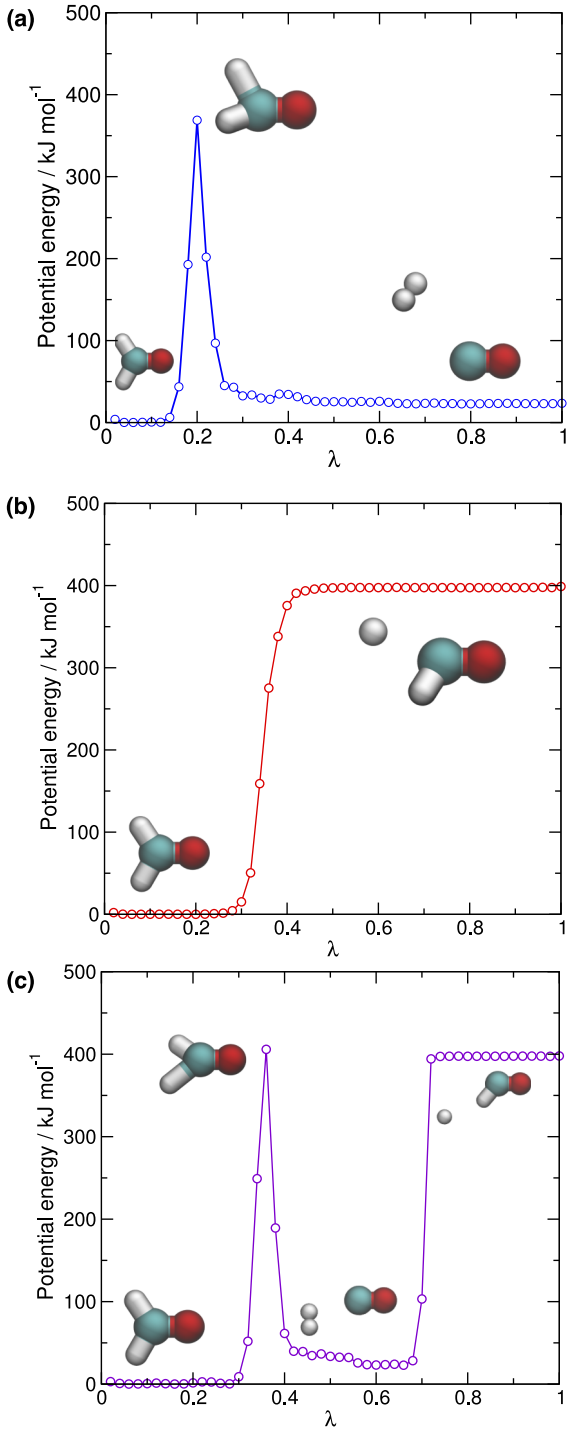

FIG. 2. Reaction paths determined by the graph-based sampling methodology outlined here; each panel illustrates a reaction path originating from formaldehyde. Panel (a) shows the molecular dissociation channel leading to H_2 + CO, panel (b) shows the radical dissociation channel leading to H + HCO, and panel (c) illustrates an "indirect" radical dissociation mechanism in which a (presumably) low kinetic energy hydrogen molecule dissociates and is recaptured by the parent CO moiety, ultimately leading to H + HCO products.

are the same as in Fig. 2(a), but our sampling approach has *automatically* located an alternate pathway. In fact, it is difficult to imagine how one might choose an initial NEB path which would lead to the "indirect" radical channel shown in Fig. 2(c). In particular, we observe dissociation of molecular hydrogen as a first step; however, instead of direct dissociation, the

molecular hydrogen remains trapped in the attractive well of the carbon monoxide, such that the maximum observed distance between carbon and hydrogen is around 3.1 Å. The hydrogen molecule subsequently returns towards the carbon monoxide, forming the HCO product and a dissociated hydrogen atom. This "indirect" mechanism for producing the radical products presumably results from a set of initial momenta which lead to an initial molecular hydrogen product with low kinetic energy; however, chain-of-states methods such as NEB work in configuration space, such that it is difficult to attribute and analyse momentum distributions during the course of a given reaction pathway. Further analysis could be carried out using the initial paths generated here, for example, by using action-based methods[44,50] which aim to generate realistic Verlet-like paths between fixed reactant and product configurations; this will be an area for future work. While the reactive pathway shown in Fig. 2(c) could in principle be considered as a combination of those shown in Figs. 2(a) and 2(b), the important point is that the path-sampling methodology outlined here can generate non-trivial pathways which would be difficult to locate using, for example, simple linear interpolation for the initial guess of the reaction path.

We note that Fig. 2(c) is not claimed as a "new" reaction mechanism in this system; it simply serves to emphasize that non-trivial reaction paths between defined end-points can be automatically generated in our approach. As an aside, it is also important to emphasize that the sampled reaction pathways, as in all string-based methods such as NEB, are not *dynamic* in nature, but simply represent a set of configurations along an approximate reaction path between defined end-points. In contrast, the novel "roaming" mechanism[67] was observed in direct quasi-classical dynamic trajectories. That said, one might expect that a reaction pathway corresponding to the roaming mechanism would be accessible in our approach. The fact that no such path is observed must point to incomplete sampling in *path-space* during our simulations; this is discussed further below.

Figure 3 illustrates further reaction paths sampled in this work; in this case, the reaction paths shown either start or end at the HCOH isomer. As in Fig. 2(c), Fig. 3(a) shows a multi-step reaction process which leads from the *trans* HCOH isomer to the molecular dissociation products; again, this multi-step path would be difficult to locate using a simple linearly interpolated reaction path. Figures 3(b) and 3(c) illustrate, respectively, direct hydrogen atom dissociation from the *cis* and *trans* isomers of HCOH. As expected, based on previous SHS simulations of the formaldehyde PES,[30] dissociation from the *cis* isomer is effectively barrierless, while dissociation from the *trans* isomer is not; clearly, the initial reaction paths generated by the path sampling process described here are capable of providing physically relevant initial guesses for further refinement.

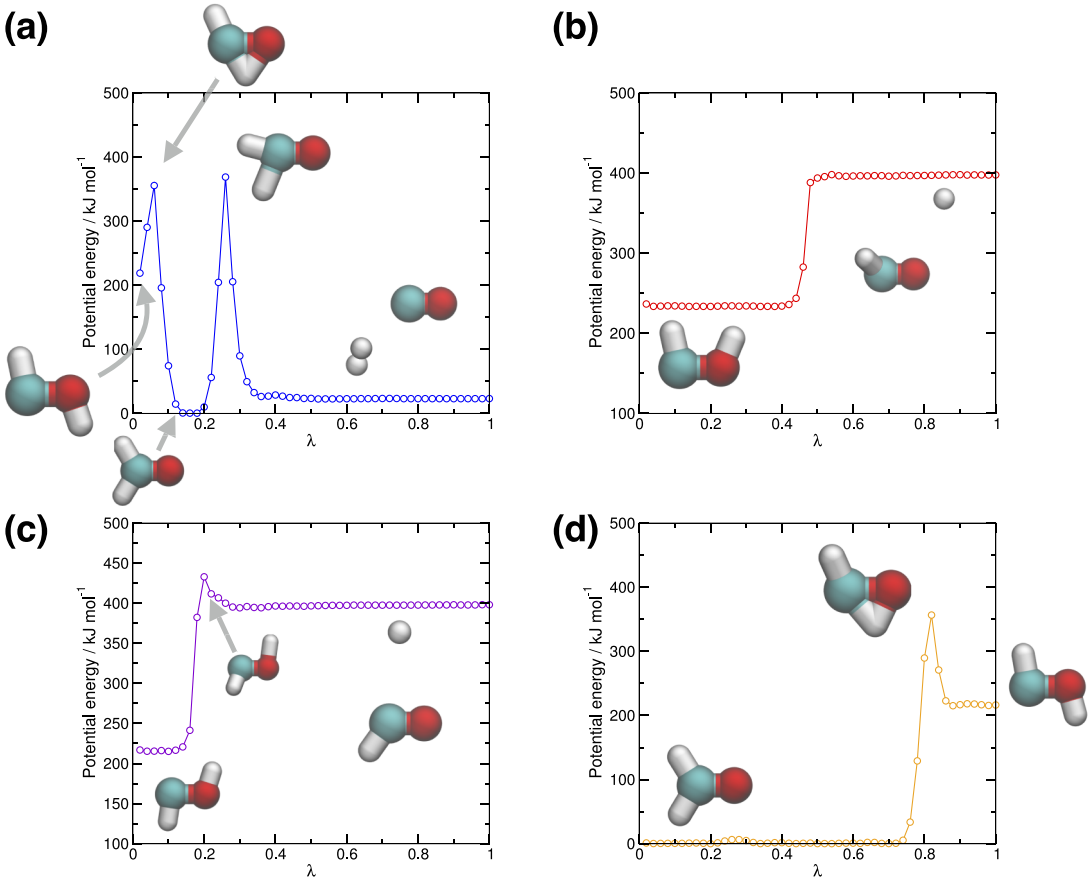

FIG. 3. Further reaction paths determined in graph-based sampling simulations; in this case, all reaction paths involve HCOH. Panel (a) illustrates a pathway involving isomerization to formaldehyde, followed by dissociation of molecular hydrogen. Panels (b) and (c) show, respectively, hydrogen atom dissociation from the *cis*- and *trans*-isomers of HCOH, and panel (d) illustrates isomerization from formaldehyde to *trans*-HCOH, without further dissociation of hydrogen products.

Figure 3(d) shows a simple hydrogen transfer reaction leading from formaldehyde to the *trans*-HCOH product. We note that we did not observe the corresponding transfer resulting in the *cis*-HCOH isomer. Furthermore, we did not observe the direct *cis-trans* isomerization reaction of HCOH in any of the three calculations performed here. This is not particularly surprising; the barrier to *cis-trans* isomerization is around 96 kJ mol^{-1},[65] much larger than the available thermal energy in either the reactant or product momenta, so spontaneous isomerization of the path end-points is a rare event at the simulation temperatures employed here. An alternative pathway to generating *cis*- and *trans*-isomers at each end-point is to create these configurations as part of the graph rearrangement step outlined above; evidently in the calculations performed here this did not occur. The probability of generating these specific molecular arrangements during graph rearrangement steps is simply dependent on the reactant and product configurations before graph changes are attempted; specifically, the minimization under constraints which is part of the graph rearrangement process must result in *cis*- and *trans*-isomers at the reaction path end-points. Clearly, leaving this sampling of isomers to chance in the graph rearrangement step is not desirable. Instead, we are now exploring the use of temperature-accelerated sampling methods for the end-point configurations to circumvent this problem in order to fully sample both different chemical isomers and different conformations;[71,72] for example, in the present case, it is straightforward to see that enhanced sampling along the HCOH torsional coordinate will increase the sampling of the *cis-trans* isomerization. Furthermore, a similar temperature-accelerated scheme can be formulated for the path variables; this is one route to improved sampling in *path-space* which would be expected to aid in the sampling of alternative reaction paths, such as the "roaming" mechanism noted above.

Finally, the information of Figs. 2 and 3 can be easily synthesized into a reaction "roadmap" which describes the chemistry of the formaldehyde PES studied here; this is shown in Fig. 4. In principle, given the graph definitions of reactant and starting products of all sampled reaction paths, as well as the transition states and associated energies along these paths, the generation of a global map of the PES can be straightforwardly automated; however, in the current work this map has been generated by simply cataloguing the minima and transition states observed. As expected, we observe all major reaction channels accessible for the particular formaldehyde PES employed here; the exceptions are those pathways which involve the *cis*-isomer of HCOH, notably the *cis-trans* isomerization, as highlighted above. The calculated TS geometries (determined approximately in this work), potential energy barriers, and end-point structures are all in agreement with previous calculations; as noted above, all of these details could, in principle, be further refined using, for example, a higher level electronic structure theory along with zero-point vibrational energy corrections to determine relative energies along the reaction paths.

B. Catalytic hydroformylation of ethene

While the application of our graph-based reaction sampling methodology to formaldehyde decomposition has confirmed its utility, as well as highlighting some avenues for improvement, the current state-of-the-art in reaction path finding methods lies towards determining mechanisms for catalytic cycles.[47,52] Addressing this challenge is particularly difficult due to the multiple-step nature of the catalytic process, the presence of multiple chemical species which can be required to react at different stages of the catalytic cycle, and the fact that computationally expensive treatment of the PES is required to correctly describe the reactivity of the system. However, developing methods to predict reaction mechanisms and associated rates for catalytic reaction would represent an important step forward, particularly given the widespread role of catalysis in much of industrial chemistry.

As a first demonstrative step in this direction, we consider here the hydroformylation (oxo) process which converts alkenes into aldehydes by addition of carbon monoxide and hydrogen. This industrially important process can be catalyzed by the cobalt species $HCo(CO)_3$ which is formed *in situ* in a reaction vessel containing a soluble cobalt salt (or fine cobalt powder) with carbon monoxide and hydrogen at

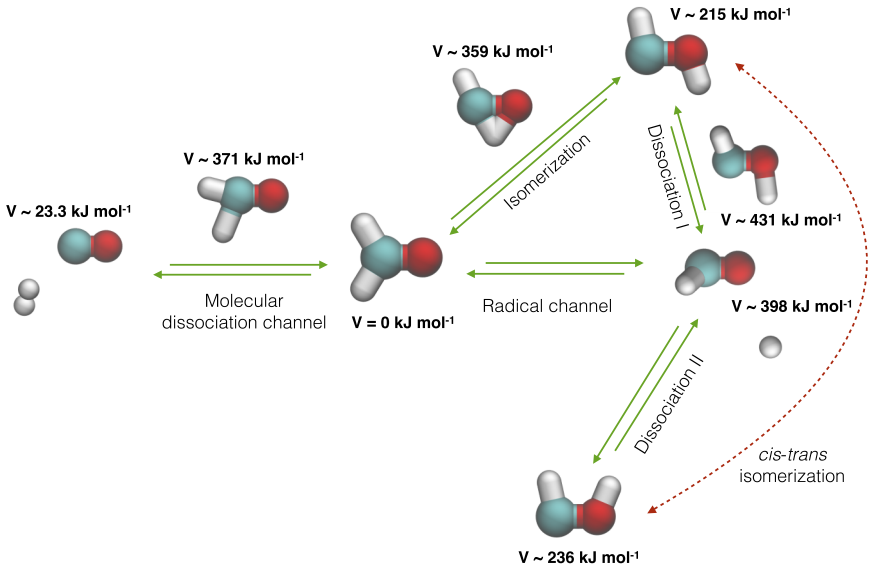

FIG. 4. Reaction network for formaldehyde, as described by the Bowman PES used in the current work.[65] The reaction paths located by the graph sampling methodology outlined here are shown, along with relevant TSs. Note that the energies of minima and TSs are approximate; further refinement by methods such as CI-NEB would yield more accurate geometries and energies. As noted in the text, the physical isomerization between *cis*- and *trans*-HCOH is not observed in our simulations; our graph-based scheme is aimed at sampling different *chemistry* in the reactants and products, although use of temperature-accelerated sampling[64,71,72] would help improve sampling of the configurational space for a given set of chemical species at either end-point of the reaction string.

high temperature and high pressure. According to the common mechanism,[47,51,52] the reaction proceeds by four main steps: (i) coordination of the alkene to the cobalt centre followed by alkene insertion into the Co–H bond, (ii) insertion of CO into the resulting Co–C bond, (iii) oxidative addition of H_2 to the cobalt centre, and (iv) reductive elimination of the aldehyde product, resulting in reformation of the catalytic species. It is clear that modelling the Co-catalyzed hydroformylation is a challenging example for reaction path-finding methodologies, involving step-wise reactions of four different molecular fragments. Here, we perform direct simulations of Co-catalyzed hydroformylation of ethene to demonstrate that our graph-based methodology is equally applicable to complex catalytic systems as it is to simpler systems such as formaldehyde decomposition; furthermore, the same catalytic process has been recently studied by AFIR[47] and BHMC methods,[52] providing useful benchmark results against which to compare.

The catalytic cycle of ethene hydroformylation by $HCo(CO)_3$ is shown in Fig. 5. In the case of formaldehyde decomposition described above, a high-quality PES describing different reaction channels was available; in the case of ethene hydroformylation by $HCo(CO)_3$, this is not the case. Instead, we employed self-consistent charge density functional tight binding (SCC-DFTB) calculations to calculate potential energies and forces for the catalytic system, using standard parameterizations.[73–75] We note that previous BHMC simulations have confirmed that this computational approach is capable of describing the main intermediates in the catalytic cycle to be studied here.[52]

The calculation details for the description of the catalytic hydroformylation process were generally the same as described above; minor changes to the atomic pair cutoff distances in Table I were introduced such that optimised molecular structures at the DFTB level gave the correct corresponding molecular graphs in our simulations, and values for the new parameters associated with cobalt were derived in a similar manner. However, due to their computational expense, these simulations used 20 configurations along the reaction pathways ($M = 18$) and, in order to increase the rate at which reactive pathways are sampled and reduce the total calculation time, the probability of graph rearrangements in each step was increased to 5×10^{-3}. To initialise the calculations here, the end-points of the reaction string were simply set to be the same configuration of the four molecular fragments, namely, $HCo(CO)_3$, C_2H_4, H_2, CO; starting from this initial configuration, conformational sampling and chemical graph moves then update the system and its bonding such that other chemical products are automatically located (see Fig. 1). The four initial fragments were placed in arbitrary orientations, with the centre-of-mass of each fragment no more than around 10 Å apart. After every 200 simulation time steps, the current reaction pathway was used as input to NEB minimization. Furthermore, we periodically start new sampling calculations which used newly encountered molecular structures (corresponding to newly sampled chemical graphs) as the initial configuration at the start- and end-points; in this manner, our algorithm enables parallel reaction path searching in a trivial fashion. However, in total, we ran five independent simulations starting from the separated fragments, plus an additional simulation starting from a system with cobalt-bound ethene (generated in one of the first five simulations); each of these simulations was run for around 2×10^3-5×10^3 simulation steps. These simulations, which were completed in less than 48 h, were sufficient to extract the main steps of the accepted catalytic cycle, as described below.

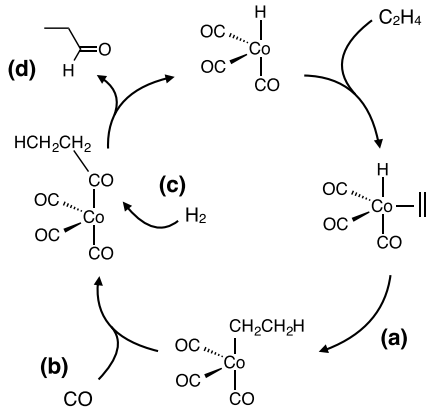

FIG. 5. Accepted mechanism for hydroformylation of ethene by $HCo(CO)_3$. (a) Following initial coordination, ethene is inserted into the Co–H bond, (b) subsequent coordination and insertion of CO into the resulting Co–C bond is then followed by (c) oxidative addition of molecular hydrogen, and (d) reductive elimination of the product aldehyde and regeneration of the cobalt catalyst.

The simulation of metal-catalyzed chemical reactions requires that the allowed graph moves introduced in our original graph sampling algorithm above must be expanded. For example, the graph moves employed in the case of formaldehyde decomposition, in which at most two graph elements are changed, are inadequate in describing commonly observed chemical processes such as insertion into metal-heteroatom bonds: M–X + Y → M–Y–X. In this example, *one* bond is broken (M–X) while *two* bonds are formed (M–Y and Y–X). To allow such changes to occur in our simulations, we expanded the allowed graph moves to include bond rearrangements which are commonly observed in metal-catalyzed reactions:

1. 1,2-insertion: M–X + Y–Z → M–Y–Z–X. Note that the Y–Z bond remains intact.
2. Metathesis: M–X + Y–Z → M–Y + X–Z.
3. 1,1-insertion: Z–X + Y → Z–Y–X. Here, note that Y represents an arbitrary species, not just a single atom; the important point to note is that both Z and X are bonded to the *same* atom in Z–Y–X.

Furthermore, the inverse of each of the above reactions was included in the allowed move set, as was the standard graph element swap introduced in the simulation of formaldehyde; in total, eight different graph moves were employed with equal probability throughout the simulations reported here. It is worth emphasizing three further points here: (i) the processes above, and their inverses, are common chemical reactions associated with metal centres which can be found in any standard textbook; they are not specific to reactivity of $HCo(CO)_3$, (ii) these graph moves do not preclude the discovery of novel chemical reactions in the system to be studied; indeed, as will be shown below, some "exotic" reactions (such as direct

insertion of carbon monoxide across a carbon-carbon double bond) are observed in our simulations, and (iii) building on the introduction of standard chemical intuition, we note that bond rearrangements which invoke energetically unfeasible bonding rearrangements, such as the direct dissociation $C_2H_4 \rightarrow 2CH_2$, are forbidden in these simulations. In essence, this approach limits the sampling to low-energy structures, a strategy which is commonly employed in other path-finding methods. As an aside, we note that although these energetically unfavourable moves were simply rejected in our simulations, it is straightforward to imagine a probabilistic scheme in which graph moves are accepted or rejected based on energetic changes, similar to the standard Monte Carlo procedure; this idea will be explored in the future. In summary, these adaptations of our approach highlight the fact that chemical intuition can be readily incorporated into our sampling scheme in a trivial manner yet, by allowing non-standard chemical moves such as direct bond "flips," we also allow a broad sampling of the allowed reactivity of the system.

Figure 6 clearly demonstrates that our path-sampling method can determine reaction pathways representing the key steps in the catalytic cycle. In Fig. 6(a), the initial insertion of ethene into the Co–H bond is illustrated. In our simulations, the sampled configuration corresponding to ethene coordination finds the C–C double bond lying perpendicular to the direction of the Co–H bond, while recent AFIR simulations[47] have determined that the configuration in which the C–C bond lies parallel to the Co–H bond is higher in energy by around 20 kJ mol^{-1}; the difference between our sampled configuration and that determined by AFIR is another example in which generation of the end-points would benefit from enhanced sampling methods, as in the *cis-trans* isomerization of formaldehyde above. In Fig. 6(b), we illustrate a reaction path corresponding to insertion of CO into the Co–C bond; note that this reaction proceeds via an intermediate in which the incoming CO molecule first binds to the cobalt centre before insertion, a feature which arises automatically in our sampling simulations. In Fig. 6(c), we see oxidative addition of the hydrogen molecule at the cobalt centre and finally, in Fig. 6(d), we observe reductive elimination of the aldehyde product and regeneration of the $HCo(CO)_3$ catalytic species. These steps in the catalytic cycle are consistent with previous investigations by AFIR;[47] however, we note that although the relative potential energy values are qualitatively correct, they are not expected to be as accurate as the AFIR simulations, which employed full DFT at the B3LYP/6-31G level. A clear route for future development of our methodology is to investigate the extent to which the sampling of reaction paths and the

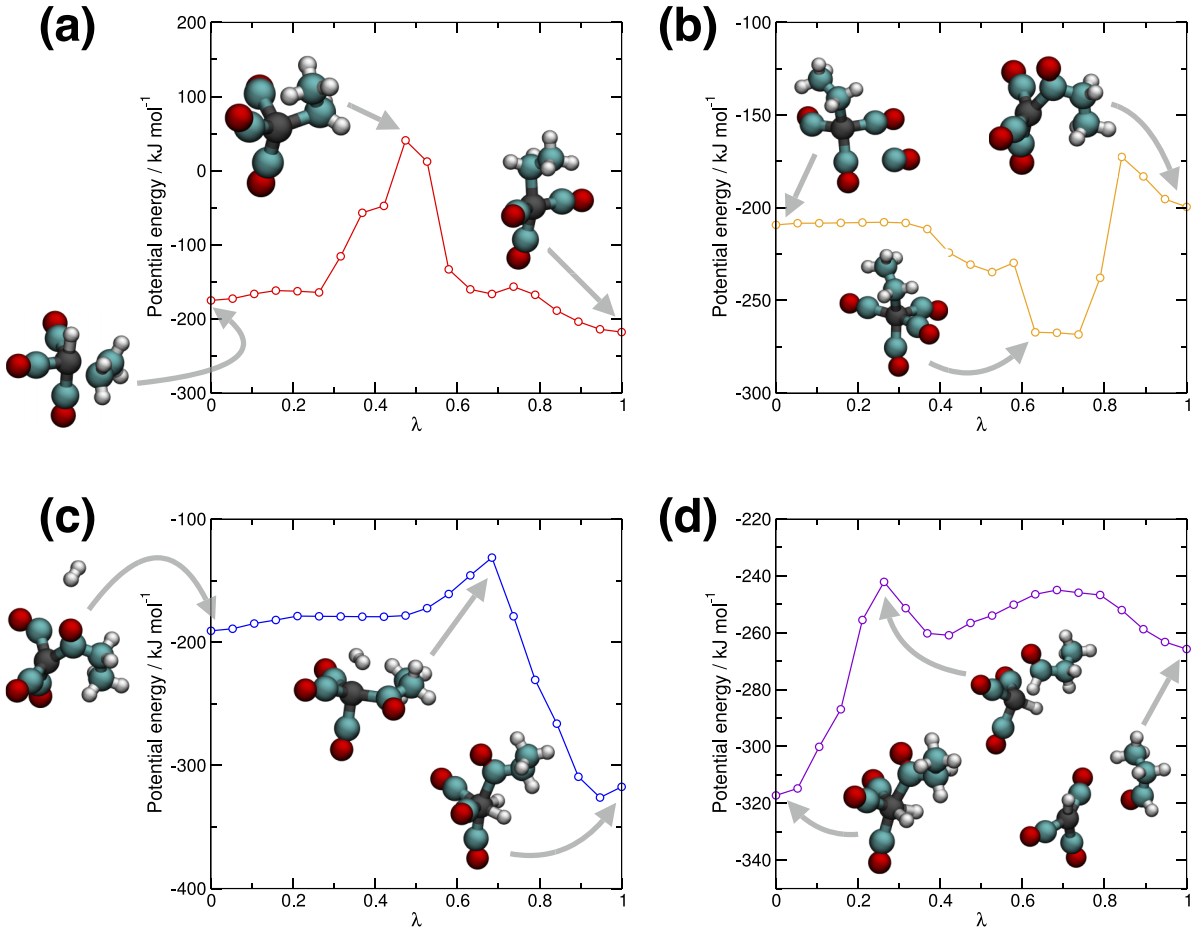

FIG. 6. Representative reaction paths for individual steps in the $HCo(CO)_3$ catalyzed hydroformylation of ethene. (a) Insertion of ethene into the Co–H bond, (b) coordination of CO and insertion into the Co–C bond, (c) oxidative addition of H_2 at the cobalt centre, and (d) reductive elimination of the aldehyde product and regeneration of the catalyst. Note that all simulations included the full set of atoms in the system; however, in cases where molecules act as spectators only, these have been removed from the representative diagrams for clarity. The energies are plotted relative to the total energy of the individual starting molecular fragments following geometry optimization.

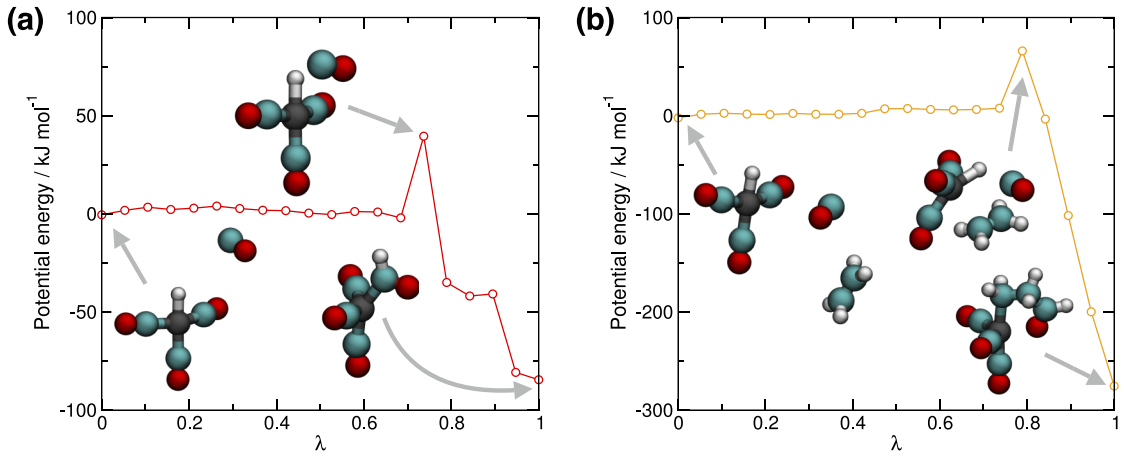

FIG. 7. Two additional reaction paths determined during graph-sampling simulations of the $HCo(CO)_3 + CO + H_2 + C_2H_4$ system. Panel (a) illustrates direct insertion of CO into the Co–H bond, and panel (b) illustrates a direct three-body reaction between $HCo(CO)_3$, carbon monoxide, and ethene, resulting in a cobalt-bound aldehyde species. As above, in cases where molecules act as spectators only, these have been removed from the representative diagrams for clarity; however, all fragments were included in the simulations.

NEB-type optimization of the paths can use different levels of theory.

As well as the accepted hydroformylation mechanism, it is interesting to note that several more "exotic" reaction pathways are also observed during the course of our path-sampling simulations. For example, as shown in Fig. 7, we observe reactions corresponding to (i) direct insertion of CO into the Co–H bond and (ii) formation of a cobalt-aldehyde intermediate by direct three-body reaction of $HCo(CO)_3$, C_2H_4, and CO. The spontaneous observation of these processes, not associated with the standard catalytic mechanism, demonstrates the potential of our methodology for discovering novel reactive pathways in system containing multiple chemically active molecules. As a related point, we note that visual analysis of our simulations shows that around 20 independent reaction paths were generated; although the aim of this work was to confirm that our path sampling approach generates the basic steps in a known catalytic mechanism, determination of unknown catalytic cycles could be automated given the individual reaction steps as input, an aspect of this work which we aim to explore in the near-future.

As a final point, it is interesting to comment on the computational expense of the graph-sampling calculations reported here. The *total* number of force evaluations to sample steps in the catalytic cycle was around 4×10^5, with a further 3×10^5 force evaluations used in NEB refinement. However, it is interesting to note that the four main steps in the catalytic hydroformylation cycle of Fig. 6 were sampled after around 2×10^5 force evaluations. For comparison, BHMC simulations of the same system required around 2.3×10^5 force evaluations,[52] although it is not clear whether this value pertained to the determination of stationary points alone or includes determination of reaction paths. Furthermore, it was only reported that BHMC found the three main intermediate structures of the catalytic cycle shown in Fig. 5[52] and it is not clear whether any further reactive pathways were sought or found. The AFIR approach, which has also been employed to model the same catalytic cycle as considered here, required fewer force evaluations than both BHMC and the graph-based approach outlined here, using around 6×10^4 force evaluations.[47] In the case of AFIR, it is noted that some reactive pathways were not explored; also, we note that, rather than searching for stationary points in which all relevant molecular fragments were present, the AFIR simulations explored reactions between subsets of fragments in order to build up a picture of the reaction profile. The same approach could easily be employed in our work to facilitate exploration of reactive space, but this idea was not used in the present work. Overall, the best that can be said is that the computational expense of the current graph-based method is at least comparable to related approaches, even without considering further avenues for optimization. Perhaps most importantly, the catalytic cycle results here were obtained in less than 48 h total simulation time without special considerations such as parallel computing, demonstrating that our approach has great potential in describing and rationalising complex catalytic cycles.

IV. CONCLUSIONS

In this article, we have demonstrated a simple approach to sampling chemical reaction pathways; the method outlined here allows both sampling of different conformations at the reaction end-points, as well as sampling of different reactant and product chemistry. The key novel features of our methodology are: (i) introduction of a Hamiltonian for sampling the chain-of-states configuration space leading from reactants to products with defined chemical arrangements, and (ii) introduction of a graph-based scheme for sampling different chemical species at the end-points of the reaction pathways. Together, these two features should, in principle, allow automated sampling of the complete reactivity of a given PES; for complex, multi-step chemical reactions such as the catalytic cycle studied here the automation of the reaction-path searching methodology is a particular advantage over traditional methods which might treat each step individually using user-defined start-points, end-points, and intermediates.

In the initial application to formaldehyde, we have seen that the major reactive channels are reproduced as expected; however, this application also highlighted some problems with

our approach as it is currently implemented. In particular, standard MD sampling of configurations at the reaction path end-points can prove insufficient to sample the entire reactive space; in the case of formaldehyde, this issue manifested itself in the fact that we did not observe *cis-trans* isomerization of HCOH due to the large barrier associated with this process relative to the available thermal energy. Furthermore, our simulations did not appear to sample a reactive pathway which corresponded to the interesting "roaming" mechanism. Both of these problems are symptomatic of limited sampling in both end-point configurational space and path-variable space, and are not inherent issues with our overall graph-based scheme; for example, the sampling of end-point configurations could be straightforwardly improved by using temperature-accelerated sampling methodologies,[64,71,72] and similar enhanced sampling can be envisaged for path variables. These are both avenues which will be explored in the near future.

In a second, more challenging application, we modelled the cobalt-catalyzed hydroformylation of ethene, a system involving four different molecular fragments. Our approach was capable of determining the main reaction steps and intermediates in this system, and also automatically generated some secondary reactions which are not part of the accepted catalytic cycle. Current work is focussed on the challenge of combining our graph-based methodology with methods for accurately determining TS geometries and reaction rate constants; this combined approach is expected to provide a useful route to investigating catalytic mechanisms, particularly once the sampling issues outlined above have been further considered.

The real strength of the approach outlined here is in the use of connectivity graphs to define the chemical identities of reactant and product states. This idea can be used to straightforwardly bias the reaction-path search towards user-defined chemical pathways; in other words, our approach can be either used "blindly" such that any reactants and products are allowed, permitting searching of the complete reactive space, or in a "guided" manner, such that only selected chemical reactions are studied. This flexibility could be exploited to facilitate the search for novel reaction pathways involving multiple steps, as commonly observed in homogeneous catalytic reactions like the cobalt-catalyzed hydroformylation studied here; this is a further area we are now exploring.

ACKNOWLEDGMENTS

This work was supported by the award of start-up funding from the University of Warwick. The author is grateful for computational resources provided by the Centre for Scientific Computing at the University of Warwick, and to Professor Joel Bowman (Emory University) for providing computer code to calculate the formaldehyde PES.

APPENDIX: NUDGED ELASTIC BAND CALCULATIONS

In the NEB method, the reactant and product configurations ($\mathbf{r}_0$ and $\mathbf{r}_P$, respectively) are fixed and the M configurations along the chain-of-states which comprise the reaction pathway are refined under the action of instantaneous forces which minimize the total potential energy along the string whilst also imposing the restraint that the configurations along the reaction path should remain evenly spaced. The total force acting on the ith configuration along the reaction pathway is

$$\mathbf{F}_i = -\nabla V(\mathbf{r}_i)|_\perp + \mathbf{F}_i^s|_\parallel, \tag{A1}$$

where the derivative perpendicular to the string is

$$\nabla V(\mathbf{r}_i)|_\perp = \nabla V(\mathbf{r}_i) - (\nabla V(\mathbf{r}_i) \cdot \boldsymbol{\tau}_i)\,\boldsymbol{\tau}_i, \tag{A2}$$

and the parallel force arising due to spring interactions is

$$\mathbf{F}_i^s|_\parallel = (k_s\,[|\mathbf{r}_{i+1} - \mathbf{r}_i| - |\mathbf{r}_i - \mathbf{r}_{i-1}|] \cdot \boldsymbol{\tau}_i)\,\boldsymbol{\tau}_i. \tag{A3}$$

Here, τ_i is a tangent vector[24] parallel to the string at configuration i and k_s is a spring constant. In contrast to more sophisticated methods, there is no guarantee that one of the M intermediate configurations lies at the TS, although we find that NEB is sufficiently accurate to clearly discern between different reactive pathways in this work. In all calculations reported here, we use a simple steepest descent minimization under the NEB forces on the M intermediate reaction path configurations, and NEB minimization is carried out starting from the current reaction path definition every t_{NEB} time steps (see Fig. 1).

[1]N. Isaacs, *Physical Organic Chemistry*, 2nd ed. (Addison Wesley Longman, 1995).
[2]J. S. Hill and N. S. Isaacs, J. Phys. Org. Chem. **3**, 285 (1990).
[3]P. F. Cook, N. J. Oppenheimer, and W. W. Cleland, Biochemistry **20**, 1817 (1981).
[4]S. B. Karki, J. P. Dinnocenzo, J. P. Jones, and K. R. Korzekwa, J. Am. Chem. Soc. **117**, 3657 (1995).
[5]I. Schlichting and K. Chu, Curr. Opin. Struct. Biol. **10**, 744 (2000).
[6]M. Jacox, Chem. Soc. Rev. **31**, 108 (2002).
[7]T. Rosenau, A. Potthast, and P. Kosma, in *Polysacharrides II*, Advances in Polymer Science Vol. 205, edited by D. Klemm (Springer, 2006), pp. 153–197.
[8]A. H. Zewail, Annu. Rev. Phys. Chem. **57**, 65 (2006).
[9]S. T. Park, A. Gahlmann, Y. He, J. S. Feenstra, and A. H. Zewail, Angew. Chem., Int. Ed. **47**, 9496 (2008).
[10]I.-R. Lee, A. Gahlmann, and A. H. Zewail, Angew. Chem., Int. Ed. **51**, 99 (2012).
[11]A. H. Zewail, J. Phys. Chem. A **104**, 5660 (2000).
[12]A. H. Zewail, Science **242**, 1645 (1988).
[13]A. Douhal, S. K. Kim, and A. H. Zewail, Nature **378**, 260 (1995).
[14]S. Habershon and A. H. Zewail, ChemPhysChem **7**, 353 (2006).
[15]B. M. Goodson, C.-Y. Ruan, V. A. Lobastov, R. Srinivasan, and A. H. Zewail, Chem. Phys. Lett. **374**, 417 (2003).
[16]C.-Y. Ruan, V. A. Lobastov, R. Srinivasan, B. M. Goodson, H. Ihee, and A. H. Zewail, Proc. Nat. Acad. Sci. U. S. A. **98**, 7117 (2001).
[17]S. J. Greaves, R. A. Rose, T. A. A. Oliver, D. R. Glowacki, M. N. R. Ashfold, J. N. Harvey, I. P. Clark, G. M. Greetham, A. W. Parker, M. Towrie, and A. J. Orr-Ewing, Science **331**, 1423 (2011).
[18]D. Chandler, *Introduction to Modern Statistical Mechanics* (Oxford University Press, Oxford, UK, 1987).
[19]D. Frenkel and B. Smit, *Understanding Molecular Simulation: From Algorithms to Applications* (Academic Press, San Diego, USA, 2002).
[20]G. Mills and H. Jónsson, Phys. Rev. Lett. **72**, 1124 (1994).
[21]G. Mills, H. Jónsson, and G. K. Schenter, Surf. Sci. **324**, 305 (1995).
[22]G. Henkelman and H. Jónsson, J. Chem. Phys. **111**, 7010 (1999).
[23]G. Henkelman, B. P. Uberuaga, and H. Jónsson, J. Chem. Phys. **113**, 9901 (2000).
[24]G. Henkelman and H. Jónsson, J. Chem. Phys. **113**, 9978 (2000).
[25]W. E, W. Ren, and E. Vanden-Eijnden, Phys. Rev. B **66**, 052301 (2002).
[26]W. E, W. Ren, and E. Vanden-Eijnden, J. Phys. Chem. B **109**, 6688 (2005).
[27]B. Peters, A. Heyden, A. T. Bell, and A. Chakraborty, J. Chem. Phys. **120**, 7877 (2004).
[28]H. B. Schlegel, Theor. Chim. Acta **83**, 15 (1992).
[29]M. Basilevsky and A. Shamov, Chem. Phys. **60**, 347 (1981).

[30]K. Ohno and S. Maeda, Phys. Scr. **78**, 058122 (2008).
[31]S. Maeda and K. Ohno, J. Phys. Chem. A **109**, 5742 (2005).
[32]K. Ohno and S. Maeda, Chem. Phys. Lett. **384**, 277 (2004).
[33]W. Quapp, M. Hirsch, O. Imig, and D. Heidrich, J. Comput. Chem. **19**, 1087 (1998).
[34]C. Dellago, P. Bolhuis, F. Csajka, and D. Chandler, J. Chem. Phys. **108**, 1964 (1998).
[35]F. Csajka and D. Chandler, J. Chem. Phys. **109**, 1125 (1998).
[36]C. Dellago, P. Bolhuis, and D. Chandler, J. Chem. Phys. **108**, 9236 (1998).
[37]P. Geissler, C. Dellago, and D. Chandler, Phys. Chem. Chem. Phys. **1**, 1317 (1999).
[38]C. Dellago, P. Bolhuis, and D. Chandler, J. Chem. Phys. **110**, 6617 (1999).
[39]P. Geissler, C. Dellago, and D. Chandler, J. Phys. Chem. B **103**, 3706 (1999).
[40]P. Bolhuis, C. Dellago, and D. Chandler, Proc. Nat. Acad. Sci. U. S. A. **97**, 5877 (2000).
[41]P. Bolhuis, D. Chandler, C. Dellago, and P. Geissler, Annu. Rev. Phys. Chem. **53**, 291 (2002).
[42]M. Hagan, A. Dinner, D. Chandler, and A. Chakraborty, Proc. Nat. Acad. Sci. U. S. A. **100**, 13922 (2003).
[43]P. Varilly and D. Chandler, J. Phys. Chem. B **117**, 1419 (2013).
[44]D. Passerone and M. Parrinello, Phys. Rev. Lett. **87**, 108302 (2001).
[45]H. Fujisaki, M. Shiga, and A. Kidera, J. Chem. Phys. **132**, 134101 (2010).
[46]S. Maeda and K. Morokuma, J. Chem. Theory Comput. **7**, 2335 (2011).
[47]S. Maeda and K. Morokuma, J. Chem. Theory Comput. **8**, 380 (2012).
[48]E. F. Koslover and D. J. Wales, J. Chem. Phys. **127**, 134102 (2007).
[49]L. Xie, H. Liu, and W. Yang, J. Chem. Phys. **120**, 8039 (2004).
[50]D. Passerone, M. Ceccarelli, and M. Parrinello, J. Chem. Phys. **118**, 2025 (2003).
[51]R. F. Heck and D. S. Breslow, J. Am. Chem. Soc. **83**, 4023 (1961).
[52]Y. Kim, S. Choi, and W. Y. Kim, J. Chem. Theory Comput. **10**, 2419 (2014).
[53]R. P. Feynman and A. R. Hibbs, *Quantum Mechanics and Path Integrals* (McGraw-Hill, New York, 1965).
[54]M. Parrinello and A. Rahman, J. Chem. Phys. **80**, 860 (1984).
[55]A. Wallqvist and B. Berne, Chem. Phys. Lett. **117**, 214 (1985).
[56]R. A. Kuharski and P. J. Rossky, J. Chem. Phys. **82**, 5164 (1985).
[57]S. Habershon and D. E. Manolopoulos, J. Chem. Phys. **135**, 224111 (2011).
[58]S. Habershon, D. E. Manolopoulos, T. E. Markland, and T. F. Miller, Annu. Rev. Phys. Chem. **64**, 387 (2013).
[59]S. Habershon, Phys. Chem. Chem. Phys. **16**, 9154 (2014).
[60]S. Habershon, G. S. Fanourgakis, and D. E. Manolopoulos, J. Chem. Phys. **129**, 074501 (2008).
[61]M. Ceriotti, M. Parrinello, T. E. Markland, and D. E. Manolopoulos, J. Chem. Phys. **133**, 124104 (2010).
[62]J. D. Doll, J. Chem. Phys. **81**, 3536 (1984).
[63]J. D. Doll, R. D. Coalson, and D. L. Freeman, J. Chem. Phys. **87**, 1641 (1987).
[64]M. E. Tuckerman, *Statistical Mechanics: Theory and Molecular Simulation* (Oxford University Press, 2012).
[65]X. Zhang, S. Zou, L. B. Harding, and J. M. Bowman, J. Phys. Chem. A **108**, 8980 (2004).
[66]K. Bondensgøard and F. Jensen, J. Chem. Phys. **104**, 8025 (1996).
[67]D. Townsend, S. A. Lahankar, S. K. Lee, S. D. Chambreau, A. G. Suits, X. Zhang, J. Rheinecker, L. B. Harding, and J. M. Bowman, Science **306**, 1158 (2004).
[68]R. W. Floyd, Comm. ACM **5**, 345 (1962).
[69]S. Warshall, J. ACM **9**, 11 (1962).
[70]M. P. Allen and D. J. Tildesley, *Computer Simulation of Liquids* (Oxford University Press, Oxford, 2005).
[71]J. B. Abrams and M. E. Tuckerman, J. Phys. Chem. B **112**, 15742 (2008).
[72]C. F. Abrams and E. Vanden-Eijnden, Proc. Nat. Acad. Sci. U. S. A. **107**, 4961 (2010).
[73]M. Elstner, D. Porezag, G. Jungnickel, J. Elsner, M. Haugk, T. Frauenheim, S. Suhai, and G. Seifert, Phys. Rev. B **58**, 7260 (1998).
[74]B. Aradi, B. Hourahine, and T. Frauenheim, J. Phys. Chem. A **111**, 5678 (2007).
[75]G. Zheng, H. A. Witek, P. Bobadova-Parvanova, S. Irle, D. G. Musaev, R. Prabhakar, K. Morokuma, M. Lundberg, M. Elstner, C. Köhler, and T. Frauenheim, J. Chem. Theory Comput. **3**, 1349 (2007).