Title

Automated Discovery and Refinement of Reactive Molecular Dynamics Pathways

Journal

Journal of Chemical Theory and Computation, 12(2)

ISSN

1549-9618

Authors

Wang, Lee-Ping
McGibbon, Robert T
Pande, Vijay S
et al.

Publication Date

2016-02-09

DOI

10.1021/acs.jctc.5b00830

Peer reviewed

Automated Discovery and Refinement of Reactive Molecular Dynamics Pathways

Lee-Ping Wang,[†,‡] Robert T. McGibbon,[†] Vijay S. Pande,[†] and Todd J. Martinez[*,†,‡,¶]

† *Department of Chemistry, Stanford University, Stanford, CA 94305.*

‡ *The PULSE Institute, Stanford University, Stanford, CA 94305*

¶*SLAC Linear Accelerator Laboratory, Menlo Park, CA 94025*

E-mail: todd.martinez@stanford.edu

November 28, 2015

Abstract

We describe a flexible and broadly applicable energy refinement method, "nebterpolation," for identifying and characterizing the reaction events in a molecular dynamics (MD) simulation. The new method is applicable to *ab initio* simulations with hundreds of atoms containing complex and multi-molecular reaction events. A key aspect of nebterpolation is smoothing of the reactive MD trajectory in internal coordinates to initiate the search for the reaction path on the potential energy surface. We apply nebterpolation to analyze the reaction events in an *ab initio* nanoreactor simulation that discovers new molecules and mechanisms, including a C-C coupling pathway for glycolaldehyde synthesis. We find that the new method, which incorporates information from the MD trajectory that connects reactants with products, produces a dramatically distinct set of minimum energy paths compared to existing approaches that start from information for the reaction endpoints alone. The energy refinement method

described here represents a key component of an emerging simulation paradigm where molecular dynamics simulations are applied to discover the possible reaction mechanisms.

1 Introduction

Chemical reactions occur when atoms move through space and rearrange their chemical bonds, and the study of thermodynamics (i.e. reaction energies) and kinetics (reaction rates) of elementary reactions is at the heart of experimental and theoretical chemistry. The essential features of many reactions can be understood by studying the adiabatic potential energy surface (PES) of the electronic ground state, a parametric function of the atomic positions in the high-dimensional configuration space in which the reactant, product, and transition state (TS) structures are stationary points. The transition state is the highest energy point on the minimum energy path (MEP) connecting reactants and products, characterized by having exactly one negative eigenvalue in the Hessian matrix. The TS and MEP provide a starting point for understanding the reaction rate of elementary reactions using rate theories such as transition state theory.[1,2] Numerical methods such as transition path sampling[3] and transition-path theory[4] can incorporate the influence of the thermodynamic ensemble, dynamical effects,[5] and nuclear quantum effects on the reaction rate. In many mechanistic problems of interest, the reactants and product structures may be known from experiment, but the transient intermediates are relatively difficult to detect and the transition state (TS) structures can only be found by supporting experimental data with theoretical calculations.[6,7] This sets the stage for theory and simulation to find and characterize the intermediates and transition states.

The search for TS structures is an important challenge in theoretical and computational chemistry that has motivated the development of diverse theoretical methods. The first practical calculations of polyatomic TS structures by Komornicki and coworkers[8,9] were carried out by first using constrained energy minimizations along manually chosen reaction

coordinates to locate an initial guess structure, then minimizing the Euclidean norm of the gradient to locate the TS. However, the gradient is zero at any stationary point on the PES and not just the desired TS, which makes the final answer highly dependent on the initial guess. This difficult problem has motivated the development of methods for generating accurate initial guesses as well as improved optimization algorithms for locating the TS.

The procedure for generating an initial guess from known reactant and product structures often begins by constructing a reaction pathway that connects reactant and product (the *endpoints*) using purely geometric methods such as the linear synchronous transit[10] or related interpolation methods.[11–13] Following this, the pathway is iteratively optimized using information from the PES,[14] either by minimizing an energy functional of the pathway as in the nudged elastic band method,[15,16] or by minimizing the normal component of the force along equally spaced structures as in the string method.[17,18] The growing string,[18–22] freezing string,[23] searching string[24] and quadratic-PES[25] are methods that build and optimize the reaction pathway simultaneously, leading to improved efficiency. We shall refer to these methods collectively as *reaction path-finding methods* that provide a sequence of structures connecting reactant and product along an approximate MEP as well as an initial guess for the TS structure. The TS guess can then be used to initiate a numerical search for the stationary point, using *transition state optimization methods* such as partitioned rational function optimization,[26–29] the dimer method[30,31] and other types of minimum mode following algorithms.[32–35] The computational efficiency may be further improved by using internal coordinate systems.[36–38] Once the TS structure is found, the intrinsic reaction coordinate (IRC) method enables the calculation of the MEP by following the energy downhill (forward and backward) along the mass-weighted steepest descent direction.[39–43] This is useful to verify that the resultant TS indeed connects the reactant and product structures.

Generally speaking, reactivity may be characterized by multiple elementary steps where the intermediates are not known *a priori*, requiring a broader exploration of the configuration space to find new minima on the PES and the pathways connecting them. Global optimiza-

tion methods for systematically finding low-energy configurations and the pathways between them include the basin-hopping method by Wales and coworkers,[44,45] the scaled hypersphere search / global reaction route mapping method by Maeda, Ohno and coworkers,[46–49] and the recent minima hopping method.[50] Reaction discovery may be aided by chemical knowledge by applying heuristic rules (e.g. bond breaking) to generate the intermediates[51,52] or by introducing forces to press reactant and product together.[53,54] Reaction events can also be observed in *ab initio* molecular dynamics (AIMD) simulations, but a major challenge is the simulation time required to escape from deep free energy minima.[55] The frequency of reaction events can be greatly accelerated by applying bias potentials that push the system away from the free energy minima along a collective variable,[56–60] which assumes some knowledge of the reaction coordinate or collective variable along which to apply the biasing potential. Reactivity is similarly enhanced when studying extremely high-temperature[61,62] or high-pressure regimes[63,64] because that shifts the chemical equilibrium towards products with higher entropy or lower volume, respectively.

We recently introduced the *ab initio* nanoreactor, a special type of AIMD simulation focused on exploring and discovering new reaction pathways.[65] The nanoreactor is a high-temperature AIMD simulation with many reactant molecules in which reactivity is accelerated by several means: (a) accelerating the electronic structure calculation on graphics processing units,[66,67] (b) employing an approximate level of electronic structure theory, and (c) a virtual piston which pushes molecules toward the center of the simulation, greatly increasing the frequency of collisions and barrier crossings. The nanoreactor MD simulation trajectory contains a great number of *reaction events* that lead from the starting reactants to new and unexpected intermediates and products. This motivates a detailed study of the reaction events to assess their mechanistic relevance for different experimental conditions, including but not limited to high temperature and pressure regimes. Intuitively, a reaction event in the trajectory is a promising initial guess to search for an underlying minimum energy path; however, this calculation is highly challenging because the trajectory usually

contains myriad high-frequency and large-amplitude motions that are either orthogonal to the reaction mechanism or involve repeated crossing of the activation barrier. In order to address these challenges, we developed an energy refinement method to calculate minimum energy paths from the nanoreactor MD trajectory. This method, which we call "nebterpolation," recognizes and extracts reaction events from the simulation trajectory, applies a new internal-coordinate smoothing algorithm to remove high-frequency motions, then applies established reaction path-finding methods and transition state optimizations to afford the transition state and minimum energy path. In this paper, we describe the development of the energy refinement method and its application to several novel examples as well as a large data set of reaction events from a nanoreactor MD simulation.

We start by defining and describing the various types of pathways in the stages of energy refinement, starting from the *reaction event* as observed in the nanoreactor, and progressing through the *initial pathway* with energy-minimized endpoints, the *smoothed pathway* generated by internal-coordinate smoothing, and the *final pathway* (also the IRC or MEP) from locating the transition state and reconnecting it with the endpoints (Figure 2). Next, we present some interesting representative examples of reactions from the nanoreactor, including a C-C coupling pathway yielding glycolaldehyde (Figure 3) and some examples of multiple pathways connecting the same reactant and product (Figures 5 and 6). Finally, we investigate the effect of including information from the MD trajectory to find the TS by comparing the outcomes of a *path-based approach* that uses the smoothed pathway to initialize the string method,[18] and an *endpoint-only approach* that uses the freezing string method to build the pathway from just the endpoints.[23] We find that both approaches are able to find transition states for a substantial fraction of reactions but with limited overlap, which indicates that using the pathway to supply the initial guess can provide distinct mechanistic information compared to starting from only the endpoints.

2 Methods

2.1 Identifying reaction events

The input to our analysis procedure is an *ab initio* molecular dynamics (MD) trajectory, consisting of a discrete time series of atomic coordinates separated in time by a fixed sampling interval (our coordinates are saved every time step, or 0.5 fs). In principle, any reactive MD method could be used to generate these trajectories, such as reactive force fields[68] or tight-binding density functional theory[69] (DFTB). The starting structure in the simulation consists of many different molecules with more than 100 total atoms. The trajectory contains many individual reaction events, but we start with no information describing when the reactions occur or which atoms participate. We shall assume in the following that the MD trajectory contains individual *reaction events* involving a small subset of atoms and taking place in a small time interval; thus, the first task is to identify and extract these reaction events from the simulation to characterize them in greater detail.

We represent molecules using a connectivity graph data structure where nodes represent atoms and edges represent covalent bonds. For each frame $\mathbf{Q}_t$ in the MD simulation (where time is measured in units of the sampling interval), we construct a connectivity graph G_t by assigning bonds to pairs of atoms (i, j) separated by distance $r_{ij;\,t} < 1.4(R_i + R_j)$, where R_i is the covalent radius[70] of atom i and the factor of 1.4 helps to compensate for transient bond stretching during the MD simulation. The graphs describing individual molecules are obtained by separating the overall graph into its connected component subgraphs. The nodes and edges also contain attributes of the corresponding atoms and bonds (e.g. node attributes include the chemical element and the global index of the atom in the simulation.)

Since we are interested in analyzing trajectories with a large number of reaction events, it is useful to define the following binary time series for each molecular graph m within the

overall simulation:

$$E_t^m = \begin{cases} 1 & \text{if } m \in G_t \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

where the $m \in G$ notation signifies that m is isomorphic to one of the connected component subgraphs of G. In the test for isomorphism, the nodes are labeled by the atom indices and edges are not labeled; thus, geometric isomers and stereoisomers are considered to be isomorphic and reaction events that preserve the atomic connectivity (e.g. cis-trans isomerization) are not detected. This is done for convenience in the present work and is not a fundamental limitation - it is also possible to further test graphs in order to distinguish between isomers. The existence of a molecule m in the simulation at time t is thus measured by the binary state of E_t^m.

In practice, the MD simulation contains large-amplitude molecular vibrations and collisions that briefly perturb the connectivity graphs without any reactivity taking place, manifesting as noise in E_t (we have omitted the superscript m for brevity). We addressed this problem using a two-state hidden Markov model (HMM), in which our observed time series E_t is modeled as a probabilistic function of a two-state Markov chain X. The Markov chain is a stochastic binary sequence of states X_t characterized by a *transition probability* $\mathbf{T} \equiv P(X_t|X_{t-1})$ describing the frequency of transitions in the sequence.[71] Although the states X_t are not directly observable, they generate the observed values of E_t according to the *output probability* $\mathbf{O} \equiv P(E_t|X_t)$ describing how often a given value of X_t produces a particular observation of E_t. The HMM is thus parameterized by the transition probability, the output probability, and the initial probability $P(X_0)$. The end goal is to calculate the most likely sequence of states $\{V_0, V_1, ..., V_t\}$ through the Markov chain, which contains a reduced amount of noise due to the parameterization of the HMM.

We set the initial probabilities to a uniform distribution (i.e. $P(X_0 = 0) = P(X_0 = 1) =$

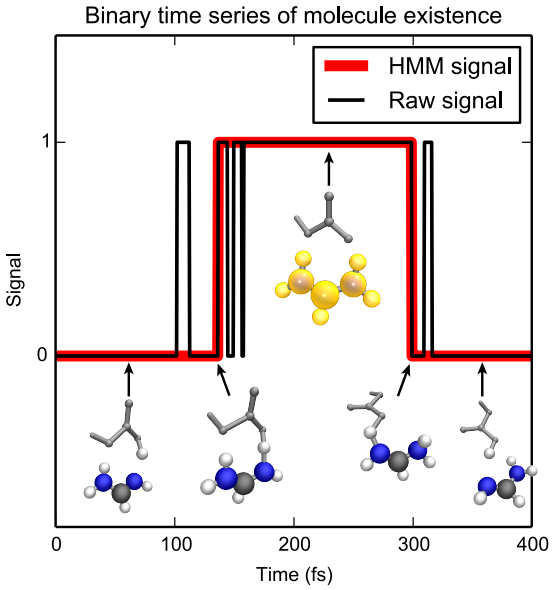

Figure 1: Using a two-state HMM to model the time series of the existence of a molecule. The raw signal (black line) is calculated by building a connectivity graph of the atoms as a function of time and testing for the subgraph corresponding to the molecule highlighted in yellow ($CH(NH_2)_2^+$, center). High-amplitude vibrations in the MD trajectory introduce significant noise into the raw signal, and the noise is removed by constructing a two-state HMM and estimating the most likely sequence of hidden states (red line).

0.5) and parameterized the transition probability as:

$$\mathbf{T} = \begin{pmatrix} 0.999 & 0.001 \\ 0.001 & 0.999 \end{pmatrix} \tag{2}$$

which indicates that X_t has a 0.001 probability of making a $0 \rightarrow 1$ or $1 \rightarrow 0$ transition in any frame. Choosing smaller values in the off-diagonal of $\mathbf{T}$ increases the strength of the noise filter and forces X_t to have fewer transitions. Here, the parameter is chosen so that the frequency of transitions is roughly consistent with the piston interval of 4000 frames in the nanoreactor simulation. We similarly parameterized the output probability as:

$$\mathbf{O} = \begin{pmatrix} 0.6 & 0.4 \\ 0.4 & 0.6 \end{pmatrix} \tag{3}$$

which signifies that E_t has a 0.6 probability of being equal to X_t at any given time. Choosing larger values in the off-diagonal of $\mathbf{O}$ allows the Markov process to deviate more often from the observed signal and behave according to its intrinsic transition probability. The HMM provides the joint probability distribution over the observed and hidden variables as:

$$P(X_{0:t}, E_{1:t}) = P(X_0) \prod_{i=1}^{t} P(X_i|X_i - 1)P(E_i|X_i). \tag{4}$$

where $E_{1:t} \equiv \{E_1, ..., E_t\}$ represents the inclusive sequence of observed values of E from the initial value to any time t. We apply the Viterbi algorithm to compute the most likely sequence of states over the Markov chain, namely:

$$V_{0:T} = \underset{X_0, X_1, ..., X_T}{\mathbf{max}} P(X_{0:T}|E_{1:T}) \tag{5}$$

where T is the length of the whole MD trajectory. As shown in Figure 1, V_t^m accurately models the existence of molecule m at time t and removes the noise from the distance-based

measurement. The transitions of V_t^m are effective indicators of reaction events involving the molecule m.

A reaction event includes the complete set of all molecules that participate in rearranging their bonds. In order to extract a reaction event involving molecule m, we start at a $1 \rightarrow 0$ transition of V^m at t_{event} and trace its atoms a_m forward in time to t_{after}, where they have converted to one or more product molecules $\{n; V^n_{t_{\text{after}}} = 1\}$. If the molecular graphs $\{n\}$ contain more atoms than a_m, it means that molecule m does not contain all of the product atoms, and multiple reactant molecules must have been involved. The set of atoms involved in the reaction is then expanded to a_n containing all atoms in $\{n\}$, and the reactant molecules are found by tracing the atoms a_n backward in time to t_{before}, prior to the detected reaction event. This process is iterated back and forth until complete and consistent sets of reactant and product molecules are found.

The resulting reaction event $\mathbf{q}_t$ contains the coordinates for the reactant and product atoms over the time interval of the reaction $(t_{\text{before}}, t_{\text{after}})$. $\mathbf{q}_t$ is extracted from $\mathbf{Q}_t$ for the subsequent refinement calculations, assuming that only the electrons of A_m are directly involved in the reaction. The net charge and spin polarization are approximated by averaging the Mulliken charge and spin populations over the time interval, summing over the atoms, and rounding to the nearest integer. This is a good approximation for most of the organic reactions considered in this paper, and future work will extend this method to include electron donors and acceptors that participate in the reaction without undergoing changes in their connectivity.

Since $\mathbf{q}_t$ contains far fewer atoms and frames than the whole trajectory $\mathbf{Q}_t$, we may carry out the refinement calculations at a higher level of electronic structure theory than was used to run the MD simulation. In this paper, the MD simulations were carried out using Hartree-Fock (both restricted and unrestricted) and the 3-21G basis set, while the refinement calculations were carried out using unrestricted B3LYP and the 6-31+G(d,p) basis set. Even higher level electronic structure methods could be employed if desired, but this suffices for

illustration of the method.

2.2 Determining path endpoints

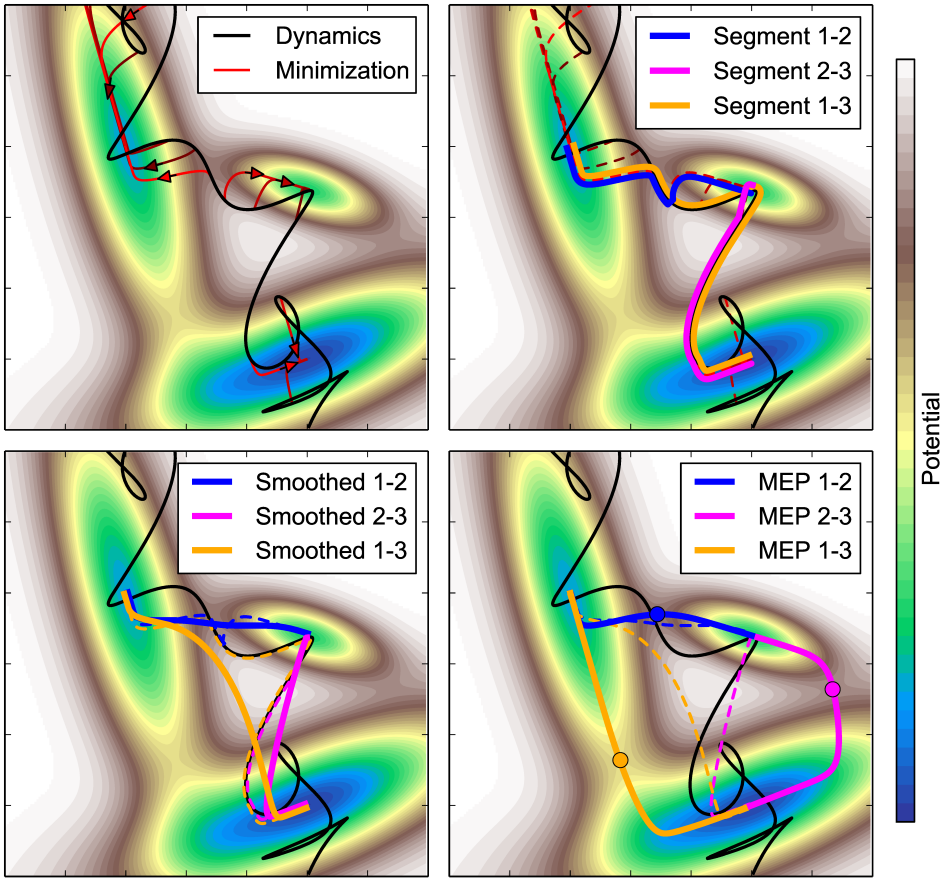

Figure 2: Illustration of the reaction path refinement method. *Top left*: A molecular dynamics trajectory (black curve) traverses the potential energy surface and passes through several energy basins. After identifying this trajectory as a reaction event, evenly sampled points on the trajectory are energy-minimized (red lines with arrows) in order to identify the distinct basins (chemical species). *Top right*: The sequence of structures from the minimization are concatenated with intervals of the molecular dynamics trajectory to form initial pathways connecting the energy basins. *Bottom left*: The connecting segments are smoothed in internal coordinates to decrease their total arc length and curvature. *Bottom right*: Reaction path-finding methods such as nudged elastic band and the string method may be applied to find the transition state, followed by a transition state optimization and intrinsic reaction coordinate calculation to obtain the final MEP.

The next step after identifying and extracting the reaction event is to search for energy-minimized structures which will serve as initial guesses to the endpoints of the minimum

energy path. This is done by initializing energy minimizations from evenly sampled time points in $\mathbf{q}_t$; we used a 10 frame sampling interval (5 fs simulation time), which is rather fine-grained and slightly shorter than an O-H classical vibrational period. Each energy minimization initialized from $\mathbf{q}_t$ provides a sequence of structures that proceeds downhill on the potential energy surface in small steps (on the order of 0.3 Å) and ends at a local minimum. The set of configurations used to initiate the minimizations typically converge to a much smaller set of distinct local minima, allowing us to partition the time series into a small number of intervals - namely, $(t^a_{\text{begin}}, t^a_{\text{end}})$ for all frames that minimize to chemical species a. An individual reaction event may contain two or more intervals corresponding to reactants, products and intermediates, or only one interval if the minimizations all converge to a single chemical species; the latter may be due to differences in the level of theory between the MD and refinement calculations, the removal of noncovalent interactions when extracting the reactant and product atoms from their environment, or the absence of an energy barrier as is often the case for homolytic bond dissociation to two radicals. In this work, we will focus on reaction events that have at least two energy basins.

Because two intervals are likely to be separated by a barrier crossing, we select the sequence of frames in between $(t^a_{\text{end}}, t^b_{\text{begin}})$ and concatenate it with the two sequences of structures from their associated energy minimizations. This provides a closely spaced sequence of structures connecting the chemically distinct energy basins (Figure 2, top right), which we call the *initial pathway* $\mathbf{x}^{ab}$. The initial pathway contains closely spaced frames taken from MD simulations and energy minimizations, which may contain useful information for locating the TS.

2.3 Internal coordinate smoothing

The initial pathway is kinked at the concatenation points and contains high-frequency motions from the MD simulation that render it unsuitable for direct initiation of a minimum energy path search. In this section, we describe a method for smoothing the coordinates

in redundant internal coordinates that preserves the energy-minimized endpoints and essential features of the connecting pathway while removing kinks and high-frequency motions. The goal is to transform the initial pathway into a *smoothed pathway* that serves as a good initial guess for reaction path-finding methods. In principle, the Cartesian coordinates can be smoothed directly by simply taking a running average, but since the potential energy surface is highly nonlinear and anharmonic, a linear smoothing could generate unphysical structures with very high energies (e.g. a linear interpolation may result in unphysical structures where atoms pass through each other.) This motivates the use of internal coordinates (e.g. interatomic distances, angles, and dihedrals) for the smoothing procedure.

We employed a redundant internal coordinate system in the smoothing procedure, because nonredundant internal coordinate systems are well-known to exhibit singularities in the Jacobian — especially when large changes in the Cartesian coordinates are involved. Our choice of redundant internal coordinates is the union of: (1) all pairwise interatomic distances, and (2) the interatomic distances, angles, and dihedral angles from the union of all bonds that occur in the initial pathway. Since this is an overdetermined coordinate system, the smoothed internal coordinates cannot be uniquely inverted to obtain a set of Cartesian coordinates. We thus define our *inverse Cartesian coordinates* in terms of a least-squares objective function, where the minimum solution is a set of Cartesian coordinates that closely corresponds to the smoothed internal coordinates in a least-squares sense.

Our smoothing procedure follows these steps:

1. Prior to smoothing, the structures along the initial pathway are linearly spaced along the total arc length measured using the RMS displacement between adjacent frames. The following procedure is applied to each frame in sequence:

2. Calculate the redundant internal coordinates $\mathbf{z}_i \equiv IC(\mathbf{x}_i)$ for each structure along the initial pathway.

3. Calculate smoothed redundant internal coordinates for each structure $\tilde{\mathbf{z}}_i$ by convolution

with windowing function: $\tilde{\mathbf{z}}_i = \sum_{j=-\Delta}^{+\Delta} \mathbf{z}_{i+j} w(j)$, where the windowing function $W(j)$ is defined on the interval $(-\Delta, +\Delta)$ and normalized to one.

4. Find the inverse Cartesian coordinates that minimize the least-squares error for each set of smoothed redundant internal coordinates: $\tilde{\mathbf{x}}_i = \min_{\mathbf{x}'}(IC(\mathbf{x}') - \tilde{\mathbf{z}}_i)^2$, using $\mathbf{x}_i$ as the initial guess.

5. The sequence of inverse Cartesian coordinates is the smoothed pathway.

We found that introducing a repulsive pseudo-energy term in the objective function was helpful for preventing close contacts between pairs of atoms during the minimization. With the repulsive term, the total function to be minimized takes the following form:

$$\begin{aligned}\chi^2(\mathbf{x}) = &\sum_{i,j \in N} \left| w_r (r_{ij}(\mathbf{x}) - \tilde{r}_{ij}) \right|^2 + w_V V_{ij}(r_{ij}(\mathbf{x})) \\ &+ \sum_{\{i,j,k\} \in \text{angles}} \left| w_\theta (\theta_{ijk}(\mathbf{x}) - \tilde{\theta}_{ijk}) \right|^2 + \sum_{\{i,j,k,l\} \in \text{torsions}} \left| w_\phi (\phi_{ijkl}(\mathbf{x}) - \tilde{\phi}_{ijkl}) \right|^2 ,\end{aligned} \tag{6}$$

where

$$V_{ij}(\mathbf{x}) = \begin{cases} D_{ij}(1 - e^{-a_{ij}(r_{ij} - r_{ij}^0)})^2 & : x < r_{ij}^0 \\ 0 & : x \geq r_{ij}^0. \end{cases} \tag{7}$$

Here, $r_{ij}(\mathbf{x})/\theta_{ijk}(\mathbf{x})/\phi_{ijkl}(\mathbf{x})$ are interatomic distances/angles/dihedrals (respectively) calculated from the trial coordinates $\mathbf{x}$, and $\tilde{r}_{ij}/\tilde{\theta}_{ijk}/\tilde{\phi}_{ijkl}$ are the smoothed internal coordinates which are the target values in the minimization. V_{ij} is the repulsive part of the Morse potential which goes smoothly to zero at r_{ij}^0, and the parameters D_{ij}, a_{ij} and r_{ij}^0 are taken from standard tables of bond dissociation energies, bond lengths and vibrational force constants. The weights $w_r = 1.0$ Å^{-1}, $w_\theta = (\frac{\pi}{6})^{-1}$, $w_\phi = \pi^{-1}$ and $w_V = (0.01 \text{ kJ/mol})^{-1}$ are chosen such that all contributions to the objective function are the same order of magnitude in numerical tests.

The independent minimizations for each $\tilde{\mathbf{x}}_i$ are nonlinear and the objective function depends on the smoothed internal coordinates $\tilde{\mathbf{z}}_i$; thus, it is possible to find adjacent pairs of inverse Cartesian coordinates $(\tilde{\mathbf{x}}_{i-1}, \tilde{\mathbf{x}}_i)$ with large deviations, resulting in a undesirable discontinuity in the smoothed pathway. When this happens, we avoid discontinuities by introducing a restraint term into the minimization:

$$\overline{\chi}_i^2(\mathbf{x}) = \chi_i^2(\mathbf{x}) + \overline{w}|\mathbf{x} - \tilde{\mathbf{x}}_{i-1}|^2, \tag{8}$$

where $\overline{\chi}^2$ is the objective function with the restraint, and $\overline{w}$ is the weight associated with the restraint that biases the solution towards the previous frame. The minimizations are carried out sequentially so that $\tilde{\mathbf{x}}_{i-1}$ is known when starting the search for $\tilde{\mathbf{x}}_i$. Starting with a restraint term of zero, a discontinuity is detected using the following condition:

$$\frac{\max \text{ abs}(\tilde{\mathbf{x}}_i - \tilde{\mathbf{x}}_{i-1})}{\max \text{ abs}(\mathbf{x}_{i+1} - \mathbf{x}_i)} > 2$$

, and the search for $\tilde{\mathbf{x}}_i$ is restarted with a small weight factor $\overline{w} = 0.02$. The weight $\overline{w}$ is increased by successive factors of 1.5 until the discontinuity is suppressed.

Introducing the restraint term can lead to problematic behavior in a very small number of cases where the ending structure on the smoothed pathway is different from that of the initial pathway. When this happens, we recalculate a sequence of inverse Cartesian coordinates starting from the ending structure and going in reverse. The search is finished when a structure from the reversed pathway coincides with a structure on the forward pathway. We found that this method was robust in smoothing a diverse set of initial pathways where linear interpolation gives unphysical interatomic distances of less than 0.5 Å.

The *smoothed pathway* consists of a set of Cartesian coordinates $\{\tilde{x}_1, \tilde{x}_i, ..., \tilde{x}_n\}$ which we use to initialize the minimum energy path search described in the next section. The code for carrying out the internal coordinate smoothing operation, *nebterpolator*, is open-source and freely available on the Web.

2.4 Searching for the minimum energy path

The fourth and final step is to locate the minimum energy path (MEP). We employ standard methods, such as nudged elastic band (NEB) or the string method, using the smoothed pathway as an initial guess. These methods iteratively search for a discretized pathway that minimizes an energy functional of the path (in the case of NEB) or the component of the gradient perpendicular to the path (in the case of the string method). The output of path optimization is a discretized pathway that approximates the MEP and a transition state (TS) estimate corresponding to the highest energy structure along the pathway. The TS estimate is then used to initialize a TS geometry optimization, which searches for a critical point with exactly one imaginary mode, and the character of the TS is verified using a frequency calculation. Finally, the optimized TS geometry connects back to the reactant and product structures via the imaginary mode that leads downhill in both directions; a pair of intrinsic reaction coordinate (IRC) calculations provide the final pathway connecting the TS to the reactant and product energy basins.

Our proposed approach uses the smoothed pathway to initiate the search for a minimum energy path; this contrasts with methods such as growing string and freezing string that proceed from knowledge of only the endpoints of the initial pathway. Since our approach provides the endpoints as well as the connecting pathway, we compared our *pathway-based approach* (using the smoothed pathway to initiate a string method search for the TS) with an *endpoint-only approach* (using the endpoints of the initial pathway to initiate a freezing string search for the TS). Our results compare the pathway-based approach and the endpoint-only approach to assess whether the connecting pathway provides additional information for locating the TS.

3 Computational Details

The calculations in this paper fall into two categories. In the first category, the nanoreactor AIMD simulations generate trajectories with hundreds of atoms and a large number of reaction events; these simulations were carried out using the TeraChem quantum chemistry software package.[67] In the second category, the refinement calculations include smaller-scale quantum chemistry calculations such as geometry optimizations, reaction path-finding and transition state searches carried out using the Q-Chem quantum chemistry software package.[72] In particular, the freezing string calculations in the endpoint-only approach were performed in a delocalized internal system with 21 nodes, 3 gradients per step, LST interpolation and a quasi-Newton optimizer, following the suggested settings from the Q-Chem documentation. The refinement calculations also involve a large amount of geometric analysis and manipulations, as well as automation for carrying out the calculations on more than a thousand reaction events. The automation was implemented in a Python module that communicates with Q-Chem and also includes *nebterpolator*. The parallel refinement calculations used the Work Queue distributed computing framework (Figure S2).

Table 1: Summary information of nanoreactor simulations analyzed in this work. The parameters of the boundary potential are described in the main text. Also listed are the number of reaction events found in each simulation and the corresponding number of final pathways; each reaction event can lead to any number of final pathways.

Label	Length (ps)	Method	Basis	Charge	t_1 / t_2 (ps)	r_1 / r_2 (Å)	k_1 / k_2 (kcal mol^{-1} Å^{-2} amu^{-1})	Reaction Events	Final Paths
1	59.69	RHF	STO-3G	0	1.0 / 1.0	14.4 / 7.2	1.0 / 1.0	62	90
2	17.94	RHF	STO-3G	0	2.0 / 0.5	14.4 / 7.2	1.0 / 1.0	29	19
3	15.24	RHF	3-21G	0	2.0 / 0.5	14.4 / 7.2	1.0 / 1.0	19	3
4	253.40	RHF	STO-3G	0	1.5 / 0.5	14.0 / 8.0	1.0 / 0.5	47	22
5	45.94	UHF	STO-3G	-1	1.5 / 0.5	14.0 / 9.0	1.0 / 0.5	31	64
6	78.17	UHF	STO-3G	-1	3.5 / 0.5	14.0 / 9.0	1.0 / 0.5	35	7
7	436.16	RHF	3-21G	0	1.5 / 0.5	14.0 / 8.0	1.0 / 0.5	317	713
8	277.00	UHF	3-21G	-1	1.5 / 0.5	14.0 / 9.0	1.0 / 0.5	127	47
9	37.11	UHF	3-21G	0	1.5 / 0.5	14.0 / 8.0	1.0 / 0.5	53	62
10	17.94	UHF	STO-3G	0	1.5 / 0.5	14.0 / 8.0	1.0 / 0.5	8	14
11	155.29	RHF	3-21G	0	1.5 / 0.5	14.0 / 8.0	1.0 / 0.5	249	59
12	65.80	RHF	3-21G	0	1.5 / 0.5	14.0 / 9.0	1.0 / 0.5	43	17

Table 1 provides the details of the nanoreactor simulations that are analyzed in this paper. The simulations used either the restricted or unrestricted Hartree-Fock (RHF/UHF) electronic wavefunction and small Gaussian basis sets (STO-3G or 3-21G) to calculate the Born-Oppenheimer potential energy surface. Four different initial configurations were used containing the same molecules: 14 H_2O, 14 CH_4, 14 NH_3, 14 CO, 16 H_2. The equations of motion were numerically integrated using Langevin dynamics with a time step of 0.5 fs, an equilibrium temperature of 2000 K (also the starting temperature), and a friction coefficient of 7 ps^{-1}. A flat-bottom spherical potential was used to contain the molecules within a finite volume as $U(r) = mk(r - r_0)^2\theta(r - r_0)$, where r_0 is the sphere radius, k a force constant, m the atomic mass, and $\theta(r - r_0)$ is the Heaviside step function. This potential is zero out to r_0 and quadratic for larger values of r. The parameters of $U(r)$ are modulated as a rectangular pulse waveform; the parameters are held at r_1, k_1 for time t_1, and then "pulsed" to r_2, k_2 for time t_2; importantly, the smaller value of r_2 forces molecules with radial distance $r > r_2$ to accelerate towards the center and collide at high velocities, which drives the reactivity. The spherical potential is proportional to the atomic mass to ensure equal inward accelerations of molecules that are outside of the sphere during these compression phases. A total of 1020 reaction events were found in 1460 ps of aggregate simulation time; the single longest simulation ran for 436 ps and produced 317 reaction events.

The energy refinement calculations used the B3LYP hybrid density functional and two Gaussian basis sets; the 6-31G(d) basis (here referred to as the small basis) was used in all calculations leading up to and including the transition state search. The 6-31+G(d,p) basis (here referred to as the large basis) was used to reoptimize the transition state from the small basis calculation and calculate the IRC leading from the transition state to the nearest energy minimum.

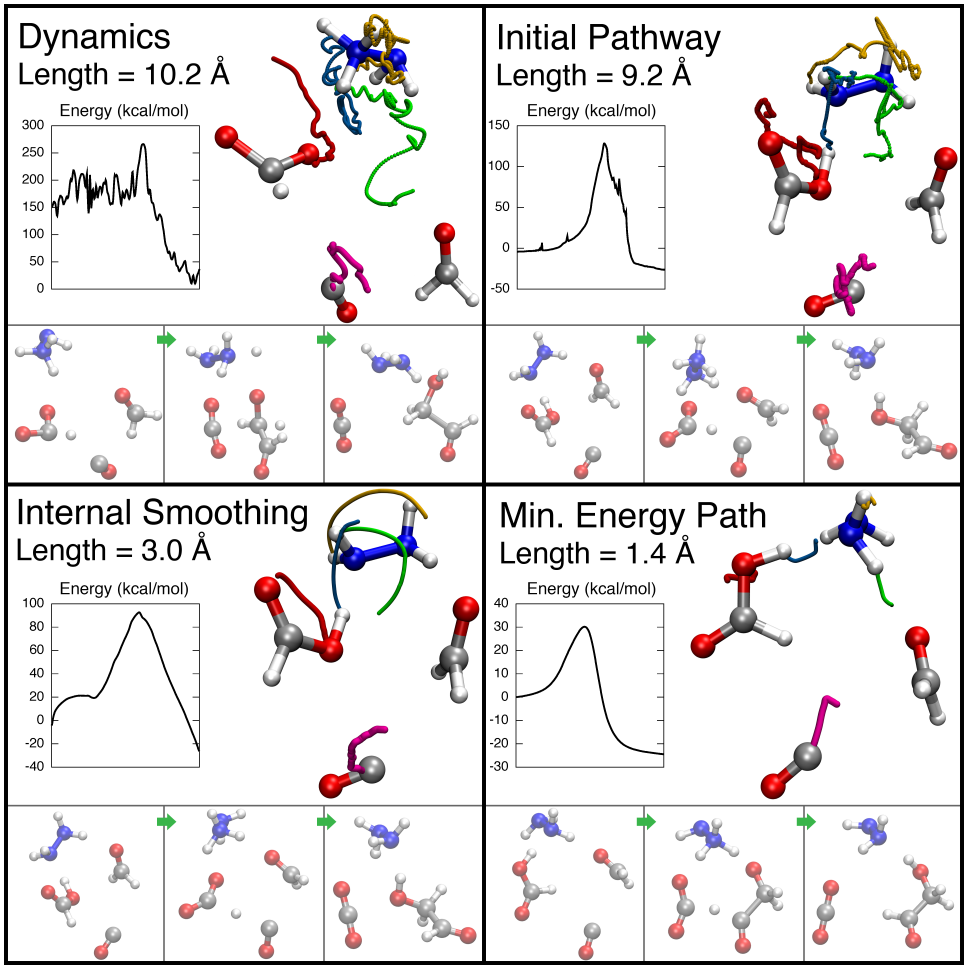

Figure 3: Demonstration of reaction pathway refinement applied to the glycolaldehyde example reaction. Each panel indicates one stage of the refinement. The initial structure of the pathway is shown in ball-and-stick representation (C, gray; O, red; H, white), and colored tracks indicate the positions of selected atoms along the pathway (magenta, C; red, O; gold/green/blue, H). Three frames taken from the start, midpoint, and end of the pathway are rendered at the bottom of each panel. The path length is given by the cumulative sum of the RMS displacement between frames. The inset shows the potential energy calculated at the B3LYP/6-31+G** level along the pathway. *Top left*: Frames containing reaction event selected from the nanoreactor MD simulation. *Top right*: Initial pathway constructed by connecting structures from energy minimization and molecular dynamics. *Bottom left*: Internal coordinate smoothing of the initial pathway. *Bottom right*: Final minimum energy path, optimized using the smoothed pathway as an initial guess. The reaction energy and activation energy are -24.4 kcal/mol and 30.1 kcal/mol.

4 Results and Discussion

4.1 Example reaction: Glycolaldehyde retrosynthesis

In this section, we describe an example of the energy refinement procedure applied to a reaction event observed in the nanoreactor, where glycolaldehyde ($HOCH_2-CH=O$) and carbon dioxide (CO_2) molecules collide to form carbon monoxide (CO), formaldehyde ($H_2C=O$) and formate ion ($HCOO^-$). Hydrazine (H_4N_2) participates in the collision and accepts a proton to form hydrazinium ion ($H_5N_2^+$). Since the reaction pathway is equally valid in either direction, we will treat glycolaldehyde as the product in the subsequent analysis because the discussion is more chemically intuitive. Thus, the (reversed) reaction event is given by:

$$\begin{aligned} H_5N_2^+ + HCOO^- + H_2CO + CO \\ \longrightarrow H_4N_2 + CO_2 + HOCH_2-CH=O. \end{aligned} \tag{9}$$

The reaction event is 340 frames (170 fs) in length, with 17 atoms and an overall formula of $C_3H_8N_2O_4$. The sum of the Mulliken charge populations over the atoms has a mean of 0.117 over the frames and a standard deviation of 0.140, and the Mulliken spin populations were 0.0 throughout (the MD simulation used restricted HF). Based on this, we assigned neutral charge and singlet multiplicity for subsequent calculations on this pathway. Energy minimizations of the endpoints cause hydrazinium to transfer a proton to formate, so the reaction event after minimization is:

$$\begin{aligned} H_4N_2 + HCOOH + H_2CO + CO \\ \longrightarrow H_4N_2 + CO_2 + HOCH_2-CH=O. \end{aligned} \tag{10}$$

The endpoints of the initial pathway are sufficient to search for the transition state via the endpoint-only approach. The endpoint-only approach did not produce a viable reaction pathway or transition state estimate, as the highest energy along the path was > 10 a.u. higher than the energy of the reactants, and much higher than any frame along the dynamics

pathway (< 0.5 a.u. higher than the reactants). The high energies were due to close contacts along the freezing string pathway, as the closest approach between any pair of atoms was 0.23 Å in the calculation output. It should be stressed that the freezing string calculations were carried out on thousands of reactions with the same settings as described in Section 3, indicating the success of the method for many other cases. Similarly, a linear interpolation of the Cartesian coordinates from the initial to the final structure resulted in a close contact of 0.17 Å (Figure S1). The complex motions of atoms in the reaction event indicate a high degree of difficulty in finding the reaction pathway using the endpoints alone.

We next applied the path-based approach; internal-coordinate smoothing led to the pathway in Figure 3, bottom left. The smoothed pathway connects the reactant and product through a curved and much shorter pathway, with a total arc length of 3.0 Å(vs. 9.2 Å for the initial pathway.) The energies along the pathway are also lower, with a maximum energy of 92 kcal/mol (vs. 128 kcal/mol for the initial pathway.) The reduction in maximum energy is an encouraging result and shows that internal coordinate smoothing can generate a chemically reasonable connecting pathway without the oscillations and kinks from the initial pathway.

The smoothed pathway was used to initiate a string method calculation, which provides a starting structure to the TS optimization. The IRC calculation starting from the TS is shown in Figure 3, bottom right, and provides the final pathway connecting the TS with its reactant and product basins. Examination of the final pathway reveals some interesting comparisons with the earlier stages of refinement. The endpoints of the final pathway are chemically identical to those of the initial / smoothed pathway, and the atomic indices of the transferred hydrogen atoms are the same, yet we observed two major differences:

First, the smoothed pathway has a greater arc length than the MEP as it contains a tumbling motion of hydrazine and formic acid. This is likely due to the potential energy surfaces of the reactant and product basins containing numerous local minima from nonbonding interactions that are well-separated in space but close in energy. In other words, the MEP

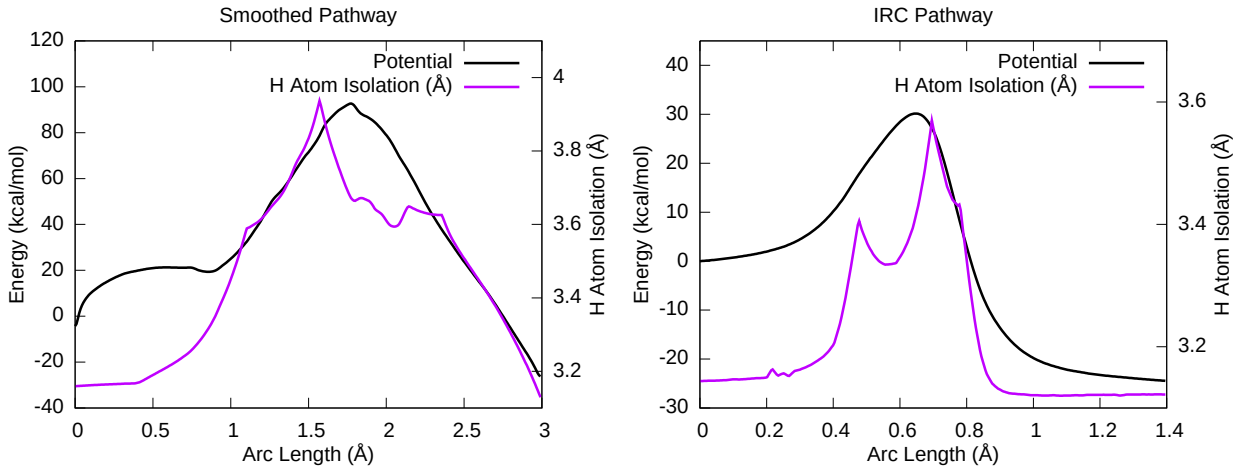

Figure 4: Plot of the potential energy along the smoothed pathway and final pathway from Figure 3, and a simple metric describing the degree of isolation of transferred hydrogen atoms, defined in the main text. The coincidence of the curves in the high-energy region indicates that the highest energies along the smoothed pathway may be due to the occurrence of isolated hydrogen atoms, a kind of chemically unfavorable structure resulting from smoothing.

that connects the endpoints of the initial pathway actually crosses several small barriers corresponding to covalent and noncovalent rearrangements, and the IRC calculation only finds the two basins nearest the TS. In this example we successfully found the chemically relevant transition state; we address the frequency of success in Section 4.3.

Second, the transition state energy of the final pathway is much lower at just 30 kcal/mol, compared to 92 kcal/mol for the smoothed pathway. We observed that the structures in the smoothed pathway contained H atoms that were distant from both the donor and acceptor site (i.e. r_{DH} and r_{HA} were both large), whereas in the final pathway the donor and acceptor approached each other prior to H-transfer such that r_{DH} and r_{HA} are never both large. Figure 4 plots the energy on the smoothed and IRC pathways, along with a simple measure of the hydrogen atom isolation, given by the sum of $\min(r_{\mathrm{DH}}, r_{\mathrm{HA}})$ for each of the three H atoms being transferred. The qualitative similarity between the hydrogen atom isolation and the potential in the high-energy regions indicates that the smoothing procedure may benefit from having even more chemical information, for example from a lower level of QM theory or a reactive force field. We leave this possibility open for future research.

4.2 Single reaction, multiple pathways

The final pathway found by the refinement procedure is a local MEP in the space of all possible pathways that connect two minima on the potential energy surface. Since the nanoreactor simulations often finds several reaction events with the same reactant and product molecules, it could discover several distinct pathways for the same reaction. In what follows, we discuss some reaction pathways that connect two minima separated by a single barrier where the pathways are qualitatively different.

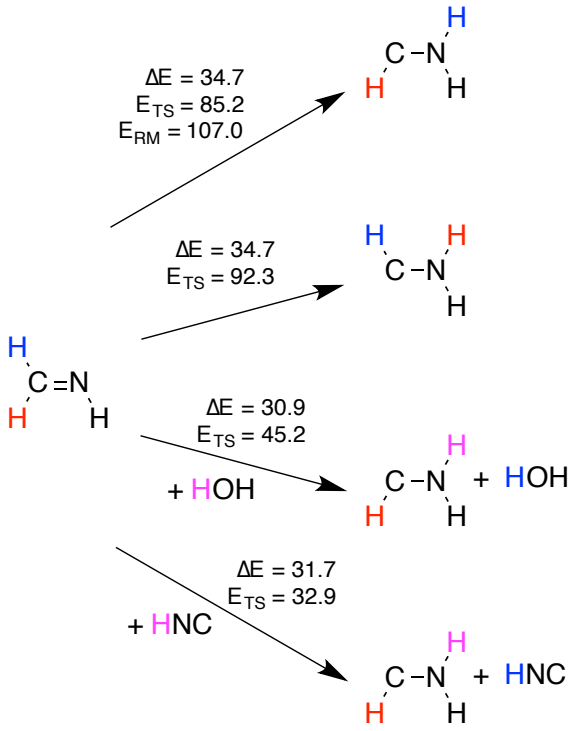

Figure 5: Several isomerizations of formaldimine to aminomethylene. Hydrogen atoms are colored to indicate the pathway that is followed. E_{RM} denotes the plateau energy of the roaming pathway in which H dissociates before attaching to N. The first and second pathway differ by which hydrogen migrates; in the top pathway the initial and final H positions are in a *cis* configuration, and has a lower barrier than the *trans* configuration. In the bottom two pathways, a proton is exchanged with H_2O and HNC.

An interesting example is the isomerization of singlet formaldimine to aminomethylene: $CH_2{=}NH \longrightarrow CH{-}NH_2$, involving the migration of a single H atom with a calculated

reaction energy of 34.7 kcal/mol. This reaction has several distinct pathways (Figure 5), two of which involve migration of the H in the *cis* or the *trans* position with respect to its final position. The *cis* pathway has a barrier of 85.2 kcal/mol and H moves across in a linear path, whereas the *trans* pathway has a higher barrier of 92.3 kcal/mol and H crosses over the C=N bond in a highly curved path. A "roaming" pathway[73,74] was also found where the H atom completely dissociates before reattaching; the highest energy is 107.0 kcal/mol above the reactants, and the energy along the pathway resembles a broad plateau rather than a well-defined transition state. The roaming pathway is perhaps better treated as two separate barrier-less reactions involving dissociation and association of H, and we relegate further discussions of such reactions to a future paper.

Many reaction pathways are affected by the presence of water, ammonia or other proton-labile molecules that may participate as a bridge in proton / hydrogen atom transfer. In these cases, the pathway involving the participating proton bridge always has a lower barrier in our Urey-Miller nanoreactor data set. When water participates in formaldimine isomerization, the barrier is dramatically lower at 45.2 kcal/mol, and the reaction energy changes to 30.9 kcal/mol due to noncovalent interactions. When hydrogen isocyanide (HNC) participates, the barrier decreases even further (32.9 kcal/mol) and is almost completely diminished as the reaction energy is 31.7 kcal/mol. By contrast, hydrogen cyanide (HCN) reacts in a very different way by adding to formaldimine to form $H_2N{-}CH_2{-}N{\equiv}C$, which can then isomerize to form 2-aminoacetonitrile ($H_2N{-}CH_2{-}C{\equiv}N$).

In most of the cases we examined, multiple pathways connecting the same two endpoints differ by a proton or H-atom bridge. Examples involving differences in heavy atom behavior are less common; here we provide such an example of how heavy atoms may behave differently along a reaction pathway. The nucleophilic addition of carbonic acid to formaldimine leads to aminomethyl hydrogen carbonate:

$$(HO)_2C{=}O + HN{=}CH_2 \longrightarrow H_2N{-}CH_2{-}O{-}COOH, \tag{11}$$

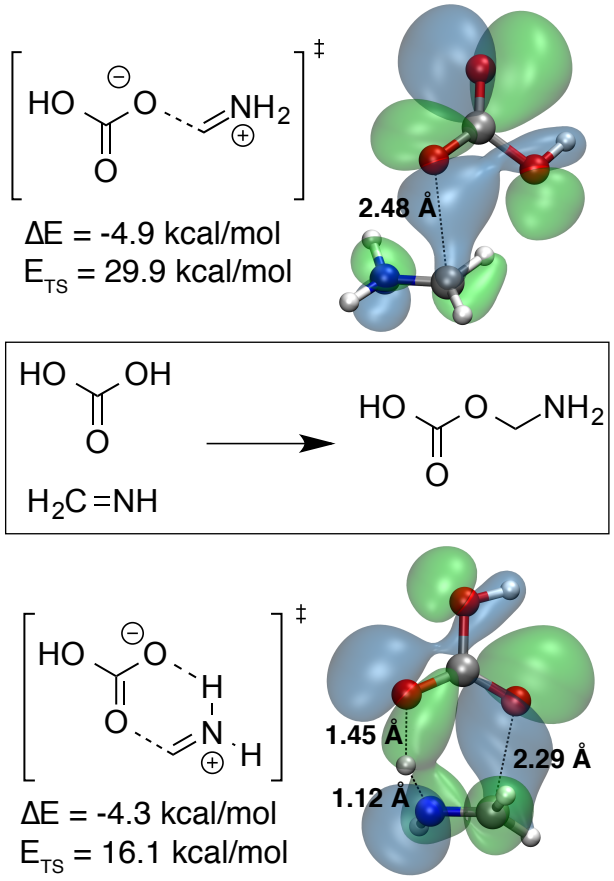

Figure 6: Transition states for two different reaction pathways for the nucleophilic addition of carbonic acid to formaldimine. The upper transition state has an energy that is more than 10 kcal/mol higher than the lower transition state. Differences in reaction energies are due to small conformational differences in the reactant and product molecules. The highest occupied Kohn-Sham orbital is rendered at an isosurface value of ±0.02. Key distance measurements are given in Angstrom. Atoms are colored as C, gray; O, red; N, blue; H, white.

in which a C−O bond is formed and one hydroxyl proton is transferred to the imine. Two qualitatively different pathways were found with chemically equivalent endpoints (Figure 6): a "lower" pathway with $\Delta E = -4.3$ kcal/mol, $E_a = 16.1$ kcal/mol and an "upper" pathway with $\Delta E = -4.9$ kcal/mol, $E_a = 29.9$ kcal/mol. In the lower pathway, the proton donor and the nucleophile are two different O atoms on carbonic acid, whereas in the upper pathway the same O-atom first transfers the proton then migrates over to make the C−O bond. The difference in barrier heights may be rationalized by looking at the highest occupied Kohn-Sham orbital of the transition states; the HOMO in the upper pathway has C−O bonding character, whereas the HOMO in the lower pathway also has O−H and N−H bonding character. Thus, the transition state in the lower pathway avoids complete dissociation of the O−H bond until the C−O bond is formed.

In the preceding reactions with multiple distinct pathways, the products at the end of the pathways are chemically identical but different in terms of the atomic indices; an example is the two isomerization pathways of formaldimine on the top of Figure 5, where two H atoms are swapped. It may be difficult to experimentally measure the relative participation of the pathways, but isotope labeling studies could offer a possible route towards distinguishing them. In any case, it is important to consider all of the pathways when deriving a rate expression for any elementary reaction.

4.3 Statistical trends in method behavior

We compared the performance and overall behavior of the pathway-based and endpoint-only approaches by performing calculations on a diverse set of systems. We investigated a set of 2652 initial pathways; the results are summarized in Figure 7, where the colored squares indicate the frequencies of the different possible outcomes in each approach. The blue square indicates the number of systems where both approaches found a final pathway containing a reaction, and the orange (resp. green) square counts the systems where only the path-based (resp. endpoint-based) approach produced a positive result. The bottom right quadrant

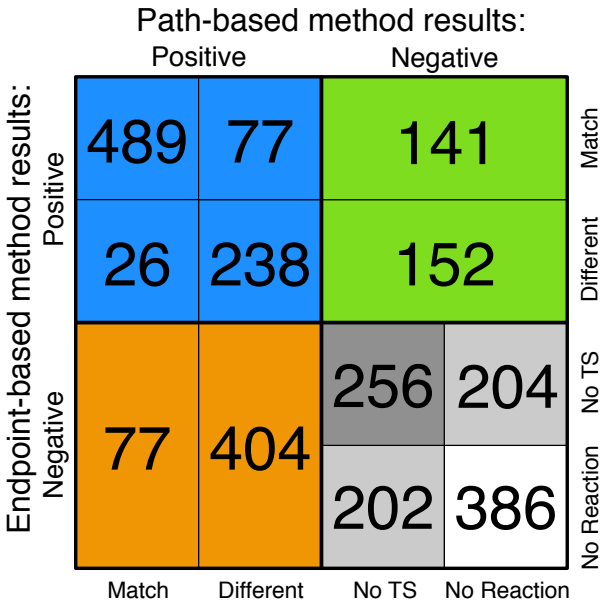

Figure 7: Overall outcome of energy refinement carried out on a data set of 3282 initial pathways from the Urey-Miller nanoreactor simulation, using the path-based approach and the endpoint-only approach. The blue, orange and green boxes respectively count the systems where both approaches, or only one (path-based or endpoint-based) produced a final pathway containing a reaction. The final pathways are further divided into those that recover the same endpoints of the initial pathway, and those that lead to different endpoints. The gray and white squares count negative results where neither approach could find a final pathway containing a reaction, either because a transition state could not be found or because the final pathway contained no reactions.

counts the systems where both approaches produced negative results, either due to failing to find a transition state (gray square) or finding a final pathway with no reactivity (white square).

Overall, 1604 calculations out of the total 2652 resulted in reactive final pathways, 810 of which correctly led back to the reactant and product species of the initial pathway. 1048 calculations produced negative results where no reactive final pathway was found. The number of transition states found using the automated procedures is encouraging given the high difficulty in locating transition states. We observed some interesting comparisons; the pathway-based approach was more often able to find a final pathway containing a reaction (1311 vs. 1123 for the endpoint-based method), but conversely the endpoint-based method was more likely to find a final pathway that matches the endpoints of the initial pathway (707 vs. 592 for the path-based method). Our hypothesis is that the path-based method sometimes provides an initial guess that crosses more than one barrier, indicating that more fine-grained division of the pathways may be possible. On the other hand, the endpoint-based method cuts the corners of the MD pathway by connecting the endpoints more directly, increasing the likelihood of finding a single barrier but also may fail when the energy landscape is more complex.

The collection of final pathways contains some redundancies in the chemical space because many reactions are observed to happen more than once. Figure 8 is a histogram of distinct final pathways, binned by the mean arc length and number of atoms. There are 1157 distinct pathways in all, and the overlap of the two methods is quite small; only 334 distinct pathways are found by both methods, whereas the path-based (resp. endpoint-based) method alone could find 517 (resp. 306) of the pathways. Furthermore, the path-based method produces a greater number of pathways with long arc lengths (> 6 Å), and the relative success of endpoint-based methods increases with the number of atoms. The different distributions of pathways found by either method indicates that both are helpful when the goal is to broadly explore the chemical space.

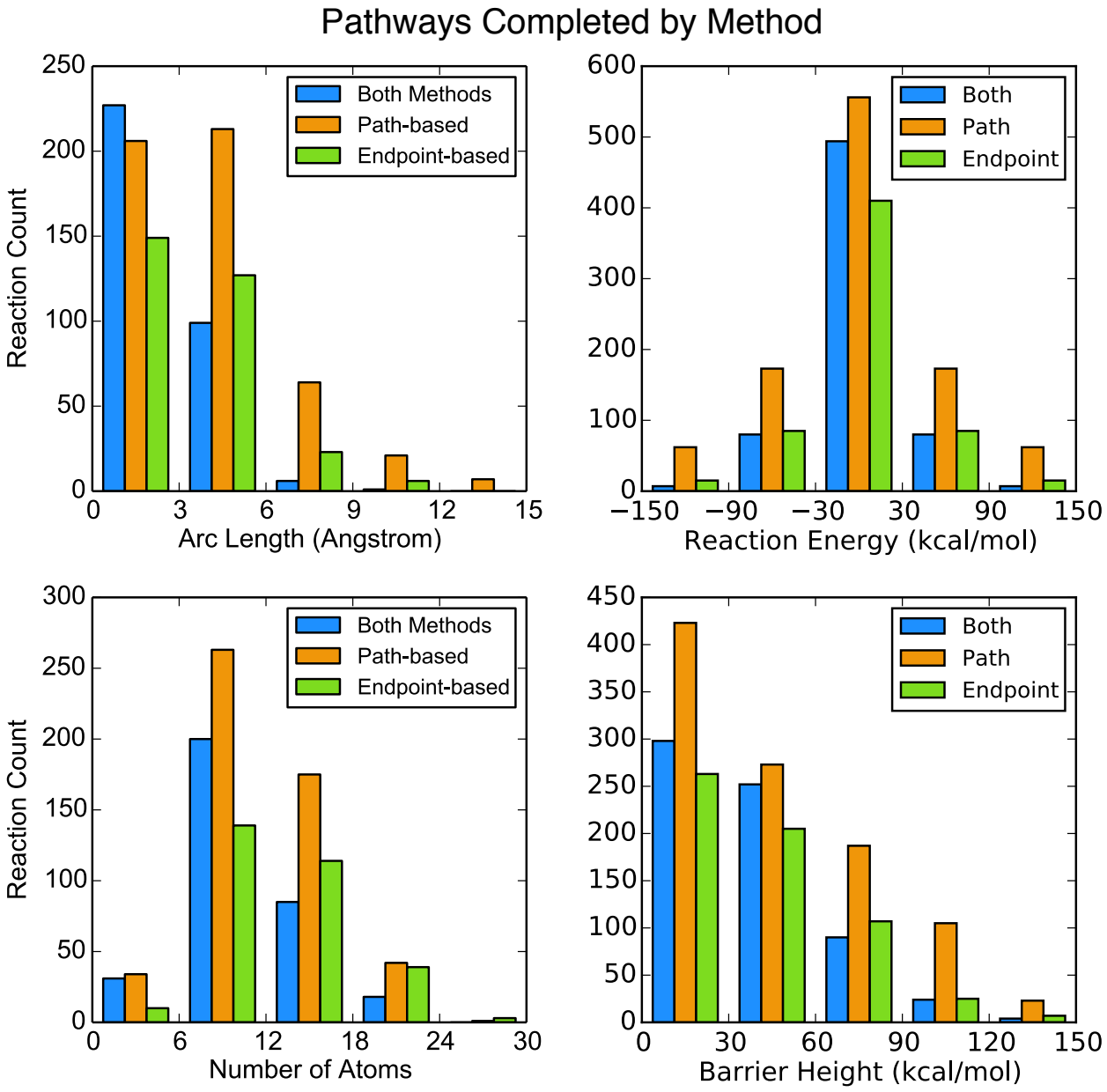

Figure 8: Statistics on distinct final pathways resulting from reaction events in the Urey-Miller nanoreactor simulation. There are a total of 1157 final pathways in the data set with different pairs of endpoints. Green (resp. orange) bars denote IRC pathways that were only found by the endpoint-based method and not the path-based method (resp. vice versa), and blue bars denote IRC pathways that were found by both methods.

5 Conclusion

In the past ten years, *ab initio* molecular dynamics and other kinds of reactive MD simulations have emerged as a promising approach for exploring chemical reactivity in complex systems with many reaction intermediates and elusive transition states. This work describes a systematic approach for connecting the reactive molecular dynamics simulation with automated approaches for mapping individual reaction events in the MD simulation to zero-temperature minimum energy paths. The key components of the method are to recognize the reaction event, locate the energy minima corresponding to the reactant and product, and leverage the MD trajectory to construct a smoothed pathway used to initialize a reaction path-finding calculation. Our calculations indicate that including the trajectory information in a pathway-based approach leads to a greater number of distinct reaction pathways compared to using an endpoint-only approach using only reactant and product structures, indicating that the MD trajectory provides valuable information for finding transition states and reaction pathways in the chemical space.

The method and calculations presented in this paper highlight many promising directions for future work. The application of more detailed electronic structure methods is certainly needed to verify and/or improve the quantitative accuracy of the results. Complete active space approaches[75–78] are especially relevant when static correlation plays an important role in the reactants or the transition state, but the systematic selection of an active space for such a large number of reactions is a major challenge. Including solvent effects could qualitatively alter the PES and lead to new intermediates and pathways, but the heavy reliance on geometry optimization requires solvent models with a well-behaved potential energy surface; switching-Gaussian implicit solvent models or cluster-continuum models may provide a good first-order description.[79–81] Hundreds of the initial pathways did not lead to a transition state or MEP containing a reaction; these could serve as a "challenge" data set for further improving reaction path-finding or TS optimization methods. More rigorous techniques for extracting the reaction events from the MD simulation could be helpful for

identifying nearby molecules in the simulation that have nontrivial effects on the pathway.

Notably, the majority of trajectories used in this paper used restricted HF because we could attain much longer simulation lengths compared to unrestricted HF (UHF). This is mostly due to the relative ease of converging the self-consistent field in RHF, which decreases the computational cost of each time step. The choice of wavefunction affects the reactions discovered; the UHF simulations contained more reaction events involving homolytic bond cleavage and radical species, whereas the RHF simulations were more heterolytic in nature. This is presumably due to the inability of RHF to describe the open-shell electronic state for homolytic bond breaking reactions and overestimating the dissociation energy. Although RHF prevents some reactions from happening, we did not observe a significant decrease in overall reactivity; rather, this seemed to bias the simulation toward finding reactions that did not involve radical or open-shell species, such as nucleophilic attack and proton transfer. These results motivate future studies that investigate how the choices of electronic wavefunction, basis set, and/or DFT approximation could be used to shift the distribution of reaction events for refinement, toward achieving our objective of discovering and characterizing the broad landscape of reactions in complex experimental conditions.

Supporting Information Available

Expanded version of Figure 3, comparing results from different minimum energy path optimization methods and flow chart describing distributed algorithm for path refinement. This information is available free of charge via the Internet at http://pubs.acs.org This material is available free of charge via the Internet at `http://pubs.acs.org/`.

Acknowledgement

This work was supported by the Office of Naval Research (N00014-14-1-0590) and the DOE Office of Basic Energy Science through the Predictive Theory of Transition Metal Oxide

Catalysis Grant. TJM is grateful to the Department of Defense (Office of the Assistant Secretary of Defense for Research and Engineering) for a National Security Science and Engineering Faculty Fellowship (NSSEFF). We are grateful to Matt Harrigan for providing software that assisted in the drawing of Figure 2.

References

(1) Berne, B. J.; Borkovec, M.; Straub, J. E. *J. Phys. Chem.* **1988**, *92*, 3711.

(2) Cukier, R. *J. Phys. Chem. A* **1999**, *103*, 5989.

(3) Bolhuis, P. G.; Chandler, D.; Dellago, C.; Geissler, P. L. *Ann. Rev. Phys. Chem.* **2002**, *53*, 291.

(4) E, W.; Vanden-Eijnden, E. *Ann. Rev. Phys. Chem.* **2010**, *61*, 391.

(5) Ammal, S.; Yamataka, H.; Aida, M.; Dupuis, M. *Science* **2003**, *299*, 1555.

(6) Williams, I. H. *Chem. Soc. Rev.* **1993**, *22*, 277.

(7) Ziegler, T.; Autschbach, J. *Chem. Rev.* **2005**, *105*, 2695.

(8) McIver, J. W.; Komornicki, A. *J. Am. Chem. Soc.* **1972**, *94*, 2625.

(9) Komornicki, A.; Ishida, K.; Morokuma, K.; Ditchfield, R.; Conrad, M. *Chem. Phys. Lett.* **1977**, *45*, 595.

(10) Halgren, T. A.; Lipscomb, W. N. *Chem. Phys. Lett.* **1977**, *49*, 225.

(11) Bauer, M. S.; Strodel, B.; Fejer, S. N.; Koslover, E. F.; Wales, D. J. *J. Chem. Phys.* **2010**, *132*, 054101.

(12) Wales, D. J.; Carr, J. M. *J. Chem. Theory Comput.* **2012**, *8*, 5020.

(13) Smidstrup, S.; Pedersen, A.; Stokbro, K.; Jonsson, H. *J. Chem. Phys.* **2014**, *140*, 214106.

(14) Sheppard, D.; Terrell, R.; Henkelman, G. *J. Chem. Phys.* **2008**, *128*, 134106.

(15) Henkelman, G.; Jonsson, H. *J. Chem. Phys.* **2000**, *113*, 9978.

(16) Henkelman, G.; Uberuaga, B.; Jonsson, H. *J. Chem. Phys.* **2000**, *113*, 9901.

(17) Weinan, E.; Ren, W. Q.; Vanden-Eijnden, E. *Phys. Rev. B* **2002**, *66*, 052301.

(18) Peters, B.; Heyden, A.; Bell, A.; Chakraborty, A. *J. Chem. Phys.* **2004**, *120*, 7877.

(19) Goodrow, A.; Bell, A. T.; Head-Gordon, M. *J. Chem. Phys.* **2008**, *129*, 174109.

(20) Behn, A.; Zimmerman, P. M.; Bell, A. T.; Head-Gordon, M. *J. Chem. Theory Comput.* **2011**, *7*, 4019.

(21) Zimmerman, P. *J. Chem. Theory Comput.* **2013**, *9*, 3043.

(22) Zimmerman, P. M. *J. Chem. Phys.* **2013**, *138*, 184102.

(23) Behn, A.; Zimmerman, P. M.; Bell, A. T.; Head-Gordon, M. *J. Chem. Phys.* **2011**, *135*, 224108.

(24) Chaffey-Millar, H.; Nikodem, A.; Matveev, A. V.; Krueger, S.; Roesch, N. *J. Chem. Theory Comput.* **2012**, *8*, 777.

(25) Plessow, P. *J. Chem. Theory Comput.* **2013**, *9*, 1305.

(26) Cerjan, C. J.; Miller, W. H. *J. Chem. Phys.* **1981**, *75*, 2800.

(27) Simons, J.; Jorgensen, P.; Taylor, H.; Ozment, J. *J. Phys. Chem.* **1983**, *87*, 2745.

(28) Banerjee, A.; Adams, N.; Simons, J.; Shepard, R. *J. Phys. Chem.* **1985**, *89*, 52.

(29) Baker, J. *J. Comput. Chem.* **1986**, *7*, 385.

(30) Henkelman, G.; Jonsson, H. *J. Chem. Phys.* **1999**, *111*, 7010.

(31) Heyden, A.; Bell, A.; Keil, F. *J. Chem. Phys.* **2005**, *123*, 224101.

(32) Munro, L. J.; Wales, D. J. *Phys. Rev. B* **1999**, *59*, 3969.

(33) Machado-Charry, E.; Beland, L. K.; Caliste, D.; Genovese, L.; Deutsch, T.; Mousseau, N.; Pochet, P. *J. Chem. Phys.* **2011**, *135*, 034102.

(34) Shang, C.; Liu, Z.-P. *J. Chem. Theory Comput.* **2012**, *8*, 2215.

(35) Zeng, Y.; Xiao, P.; Henkelman, G. *J. Chem. Phys.* **2014**, *140*, 044115.

(36) Peng, C. Y.; Ayala, P. Y.; Schlegel, H. B.; Frisch, M. J. *J. Comput. Chem.* **1996**, *17*, 49.

(37) Baker, J.; Chan, F. R. *J. Comput. Chem.* **1996**, *17*, 888.

(38) Billeter, S. R.; Turner, A. J.; Thiel, W. *Phys. Chem. Chem. Phys.* **2000**, *2*, 2177.

(39) Kato, S.; Fukui, K. *J. Am. Chem. Soc.* **1976**, *98*, 6395.

(40) Ishida, K.; Morokuma, K.; Komornicki, A. *J. Chem. Phys.* **1977**, *66*, 2153.

(41) Fukui, K. *Acc. Chem. Res.* **1981**, *14*, 363.

(42) Gonzalez, C.; Schlegel, H. B. *J. Chem. Phys.* **1989**, *90*, 2154.

(43) Taketsugu, T.; Gordon, M. S. *J. Chem. Phys.* **1995**, *103*, 10042.

(44) Wales, D.; Doye, J. *J. Phys. Chem. A* **1997**, *101*, 5111.

(45) Wales, D. J.; Scheraga, H. A. *Science* **1999**, *285*, 1368.

(46) Ohno, K.; Maeda, S. *Chem. Phys. Lett.* **2004**, *384*, 277.

(47) Maeda, S.; Ohno, K. *J. Phys. Chem. A* **2005**, *109*, 5742.

(48) Watanabe, Y.; Maeda, S.; Ohno, K. *Chem. Phys. Lett.* **2007**, *447*, 21.

(49) Ohno, K.; Maeda, S. *Phys. Scr.* **2008**, *78*, 058122.

(50) Schaefer, B.; Mohr, S.; Amsler, M.; Goedecker, S. *J. Chem. Phys.* **2014**, *140*, 214102.

(51) Zimmerman, P. M. *J. Comput. Chem.* **2013**, *34*, 1385.

(52) Rappoport, D.; Galvin, C. J. .; Zubarev, D. Y.; Aspuru-Guzik, A. *J. Chem. Theory Computation* **2014**, *10*, 897.

(53) Maeda, S.; Morokuma, K. *J. Chem. Theory Comput.* **2011**, *7*, 2335.

(54) Maeda, S.; Ohno, K.; Morokuma, K. *Phys. Chem. Chem. Phys.* **2013**, *15*, 3683.

(55) Laio, A.; Parrinello, M. *Proc. Natl. Acad. Sci. USA* **2002**, *99*, 12562.

(56) Carter, E. A.; Ciccotti, G.; Hynes, J. T.; Kapral, R. *Chem. Phys. Lett.* **1989**, *156*, 472.

(57) Voter, A. *Phys. Rev. Lett.* **1997**, *78*, 3908.

(58) Iannuzzi, M.; Laio, A.; Parrinello, M. *Phys. Rev. Lett.* **2003**, *90*, 238302.

(59) Ensing, B.; De Vivo, M.; Liu, Z.; Moore, P.; Klein, M. *Acc. Chem. Res.* **2006**, *39*, 73.

(60) Pietrucci, F.; Andreoni, W. *Phys. Rev. Lett.* **2011**, *107*, 085504.

(61) Sorensen, M.; Voter, A. *J. Chem. Phys.* **2000**, *112*, 9599.

(62) Martinez-Nunez, E. *J. Comput. Chem.* **2015**, *36*, 222.

(63) Bernasconi, M.; Chiarotti, G. L.; Focher, P.; Parrinello, M.; Tosatti, E. *Phys. Rev. Lett.* **1997**, *78*, 2008.

(64) Goldman, N.; Reed, E. J.; Fried, L. E.; Kuo, I.-F. W.; Maiti, A. *Nature Chem.* **2010**, *2*, 949.

(65) Wang, L.-P.; Titov, A.; McGibbon, R.; Liu, F.; Pande, V. S.; Martinez, T. J. *Nature Chem.* **2014**, *6*, 1044.

(66) Titov, A. V.; Ufimtsev, I. S.; Luehr, N.; Martinez, T. J. *J. Chem. Theory Comput.* **2013**, *9*, 213.

(67) Ufimtsev, I. S.; Martinez, T. J. *J. Chem. Theory Comput.* **2009**, *5*, 2619.

(68) van Duin, A. C. T.; Dasgupta, S.; Lorant, F.; Goddard III, W. A. *J. Phys. Chem. A* **2001**, *105*, 9396.

(69) Elstner, M.; Porezag, D.; Jungnickel, G.; Elsner, J.; Haugk, M.; Frauenheim, T.; Suhai, S.; Seifert, G. *Phys. Rev. B* **1998**, *58*, 7260.

(70) Cordero, B.; Gomez, V.; Platero-Prats, A. E.; Reves, M.; Echeverria, J.; Cremades, E.; Barragan, F.; Alvarez, S. *Dalton Trans.* **2008**, *37*, 2832.

(71) Russell, S.; Norvig, P. *Artificial Intelligence: A Modern Approach*, 3rd ed.; Pearson Education: Upper Saddle River, New Jersey 07458, 2010.

(72) Shao, Y.; Molnar, L. F.; Jung, Y.; Kussmann, J.; Ochsenfeld, C.; Brown, S. T.; Gilbert, A. T. B.; Slipchenko, L. V.; Levchenko, S. V.; O'Neill, D. P.; Jr., R. A. D.; Lochan, R. C.; Wang, T.; Beran, G. J. O.; Besley, N. A.; Herbert, J. M.; Lin, C. Y.; Voorhis, T. V.; Chien, S. H.; Sodt, A.; Steele, R. P.; Rassolov, V. A.; Maslen, P. E.; Korambath, P. P.; Adamson, R. D.; Austin, B.; Baker, J.; Byrd, E. F. C.; Dachsel, H.; Doerksen, R.; Dreuw, A.; Dunietz, B. D.; Dutoi, A. D.; Furlani, T. R.; Gwaltney, S. R.; Heyden, A.; Hirata, S.; Hsu, C.-P.; Kedziora, G.; Khalliulin, R. Z.; Klunzinger, P.; Lee, A. M.; Lee, M. S.; Liang, W.; Lotan, I.; Nair, N.; Peters, B.; Proynov, E. I.; Pieniazek, P. A.; Rhee, Y. M.; Ritchie, J.; Rosta, E.; Sherrill, C. D.; Simmonett, A. C.; Subotnik, J. E.; III, H. L. W.; Zhang, W.; Bell, A. T.; Chakraborty, A. K.; Chipman, D. M.; Keil, F. J.; Warshel, A.; Hehre, W. J.; III, H. F. S.; Kong, J.; Krylov, A. I.; Gill, P. M. W.; Head-Gordon, M. *Phys. Chem. Chem. Phys.* **2006**, *8*, 3172.

(73) Suits, A. G. *Acc. Chem. Res.* **2008**, *41*, 873.

(74) Bowman, J. M. *Mol. Phys.* **2014**, *112*, 2516.

(75) Roos, B. O. *Adv. Chem. Phys.* **1987**, *69*, 399.

(76) Hohenstein, E. G.; Luehr, N.; Ufimtsev, I. S.; Martinez, T. J. *J. Chem. Phys.* **2015**, *142*, 224103.

(77) Hohenstein, E. G.; Bouduban, M. E. F.; Song, C.; Luehr, N.; Ufimtsev, I. S.; Martinez, T. J. *J. Chem. Phys.* **2015**, *143*, 014111.

(78) Shu, Y.; Hohenstein, E. G.; Levine, B. G. *J. Chem. Phys.* **2015**, *142*, 024102.

(79) Tomasi, J.; Mennucci, B.; Cammi, R. *Chem. Rev.* **2005**, *105*, 2999.

(80) Liu, F.; Luehr, N.; Kulik, H. J.; Martinez, T. J. *J. Chem. Theory Comput.* **2015**, *11*, 3131.

(81) Lange, A. W.; Herbert, J. M. *J. Chem. Phys.* **2010**, *133*, 244111.

Graphical TOC Entry