# Automated Discovery of Chemically Reasonable Elementary Reaction Steps

Paul M. Zimmerman*

Due to the significant human effort and chemical intuition required to locate chemical reaction pathways with quantum chemical modeling, only a small subspace of possible reactions is usually investigated for any given system. Herein, a systematic approach is proposed for locating reaction paths that bypasses the required human effort and expands the reactive search space, all while maintaining low computational cost. To achieve this, a range of intermediates are generated that represent potential single elementary steps away from a starting structure. These structures are then screened to identify those that are thermodynamically accessible, and then feasible reaction paths to the remaining structures are located. This strategy for elementary reaction path finding is independent of atomistic model whenever bond breaking and forming are properly described. The approach is demonstrated to work well for upper main group elements, but this limitation can easily be surpassed. Further extension will allow discovery of multistep reaction mechanisms in a single computation. The method is highly parallel, allowing for effective use of modern large-scale computational clusters.


DOI: 10.1002/jcc.23271

## Introduction

Each year, more and more articles report the investigation of chemical reaction mechanisms using first principles molecular modeling techniques. To retain a low computational cost, most studies utilize density functional theory (DFT) due to its attractive cost to accuracy ratio.[1–5] DFT enables the relatively rapid characterization of the model system's potential energy surface (PES), which spans $3N - 6$ degrees of freedom (DOF) ($N$ is the number of atoms in the system). In almost all cases, the computational cost of navigating this large dimensionality precludes any expansive search of this surface. Instead, key intermediates and transition states (TSs) are chosen using chemical intuition and prior knowledge of the system's reactivity to drastically reduce the search space.[6,7] The result of many of these studies is a proposed mechanism with energies derived from first principles. It is inevitable that this approach (denoted the "manual approach") has a serious disadvantage: there is no fundamental metric to decide whether key reaction intermediates or mechanisms have been missed.

The goal of many of these simulations is to provide atomistic data to support experimental results, and the manual approach suits this purpose much of the time. An expert in chemical simulations can often come up with a mechanism that reproduces and explains known experimental results. However, this procedure is often unsatisfying due to the lack of predictive value. In this regard, predictive methods that could explore a more significant volume of reactive space would prove immensely valuable, especially for the discovery of new types of chemical reactions.

Many approaches have been suggested to determine energetically relevant reaction pathways when only the starting structure is known. These methods fall into two general categories: (1) those that search through predetermined reactive coordinates for TSs and (2) methods that use some system property to approximate reactive coordinates. Prominent in the former category are methods such as metadynamics[8–10] and chemical flooding,[11,12] which are molecular dynamics simulations biased to proceed along predefined coordinates. These simulations can follow up to 4–6 coordinates,[10] but following more coordinates is computationally prohibitive. Although methods that explore reactive paths through coordinate biases in principle could be very useful, designating these coordinates is usually a system-dependent task. Methods in category (2) often follow shallowest ascent coordinates to TSs,[13–16] and can even allow multiple TSs to be found from the same intermediate.[17] Shallowest ascent methods give no guarantee that the most important TSs are located for a given system (these tend to repeatedly locate the same TS over multiple runs), in contrast to type (1) methods that are likely to find the important TSs when the appropriate bias coordinate is chosen. An interesting category (2) method for single-ended reaction path finding presented by Maeda[18–20] induces a force between two molecules to cause them to pass over associative reaction barriers. It is not, however, generally useful for nonassociative reactions (e.g., single complex isomerization, dissociations, etc.). Many of these methods are innovative and useful, but none can yet fully replace the manual approach.

If the most relevant reactive intermediates have already been identified, a diversity of methods are available for locating the relevant TSs.[21–33] While this can also be done by

P. M. Zimmerman
Department of Chemistry, University of Michigan, Ann Arbor, Michigan 48109
E-mail: paulzim@umich.edu

manually interpolating between the two intermediates and using well-known local search algorithms such as eigenvector following[31] and the dimer approach,[26] automated methods are often easier to use and more reliable. Double-ended methods such as synchronous transit,[34,35] nudged elastic band,[36–41] the string method,[42–46] and the growing[47–50] and freezing string methods[51] can be used to this effect. The latter two methods are particularly efficacious due to their combination of low computation cost and high reliability.

In contrast to methods that rely on DFT are machine learning approaches designed to predict reaction mechanisms based on analogies to known reactivity.[52–56] While these methods can in principle predict many chemical reactions at great efficiency, they require extensive training sets that are not generally available for all types of reactions. Substantial efforts have been applied to kinetic modeling of reaction networks,[57–62] where sequences of known elementary steps (usually generated using a specific rule system for the chemistry of interest) are studied to determine product distributions at varying conditions. Automated determination of elementary reaction steps could greatly support this effort by identifying unknown elementary steps along with their rate constants.

This article suggests an alternative to existing reaction discovery approaches that is both computationally tractable and not heavily reliant on human intuition. This approach uses principles of atomic connectivity to systematically determine elementary reaction steps in chemical systems. By applying simple rules that provide a system-independent basis of possible elementary reactions, the human effort for finding TSs can be cut down dramatically. The method allows a systematic search for intermediates that may form after bond breaking and forming events and gives a straightforward procedure for locating the related exact TSs. This approach is flexible and allows for changes to any of its components, suggesting it can be improved in the future.

## Theory

### Atomic connectivity definitions

The cornerstone of understanding chemical reactions is the breaking and forming of connections between individual atoms. For the purposes of this study, we will not consider conformational changes or other isomerizations that result in no changes in atom connectivity. Along these lines, atomic coordination number is a simple metric that counts the local bonding environment around each atom. For instance in carbon, the coordination number should range between 1 and 4, and sometimes increases to 5 in special cases. This metric provides a useful tool to measure whether an additional bond may form or whether bond dissociation is possible. To this effect, one can imagine multiple ways to define coordination number.[8,63] Herein, a connected atom pair is specified when the distance between the two is less than the sum of the covalent radii times a constant factor (usually $\sim$1.1). This definition mirrors a typical procedure for specifying internal coordinates.[34,63]

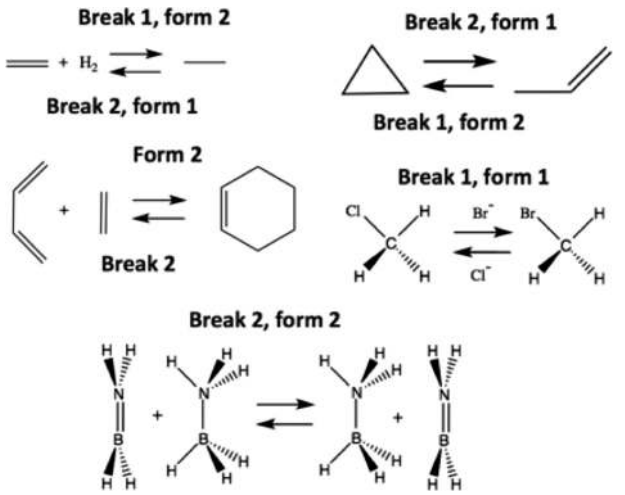


**Figure 1.** Representative examples where atomic connectivity rules can apply to a variety of chemical reactions.

While this strategy could be applied to the entire periodic table, additional effort will be required to develop meaningful connectivity definitions for transition metals. In transition metals, geometries are not always well defined by simply choosing a coordination number. For instance, the heterogeneity of equatorial and axial positions of a trigonal–bipyramidal complex will not be treated by assuming such a species as having five equivalent connections. A suitable force field that captures such structural features would be useful to this effect. In the test cases later, we demonstrate that this approach works well for upper main group elements, though the approach could work equally well for other main group elements in its current form. No attempt is made to systematically capture chirality or isomerization with constant atomic connectivity.

### Allowed changes in connectivity

Rules for connection breaking and forming can be proposed to cover a broad span of chemical reactions. For the majority of chemical reactions, only a small number of atomic connections are made or broken in any one elementary step. Therefore, a simple protocol would be to sample all possible chemical rearrangements with no more than two connections breaking and two connections forming simultaneously, while maintaining the upper and lower limits for coordination number of each atom. This yields hypothetical reactions where at least one connection is formed or one connection breaks in an elementary step, and prevents unnecessary formation of atoms that are under- or over-coordinated. Figure 1 shows a sampling of how these rules may work in real chemical systems. The rules are expected to cover a large variety of reactions, but can be extended if needed for complex reaction types.[54,55]

An important feature of these rules is that they produce a great variety of possible elementary steps, while at the same time providing a search space that is narrow compared to sampling all possible geometric conformations. The reactive search space is, therefore, reduced from $3N - 6$ coordinates to a polynomial

scaling number of connectivity changes, which means the computational cost for this exploration is relatively low.

In Figure 1, kinetically and thermodynamically feasible reactions are shown. In practice, however, the connection rearrangement rules will also produce high energy intermediates such as radicals. Additionally, discovery of thermodynamically feasible intermediates does not guarantee kinetic feasibility. For instance, these rules make no distinction between single and double bonds in hydrocarbons, where cleaving a double bond likely will result in a kinetically improbable step. For these reasons, screening approaches must be used to reduce the size of the configurational space.

### Enacting connectivity isomerization

The connectivity rules can be implemented using any level of atomistic theory to describe the system. These rules amount to the application of constraints on interatomic distances such that qualitatively correct structures are imposed. While DFT methods can, in principle, be used with constraints to modify the starting geometry to the new connectivity, a faster route will be to use a molecular mechanics (MM) force field. Isomerizations can be achieved by adding and deleting connections between the atoms of interest followed by optimization. Because the qualitative structure is explicitly fixed during optimization, MM allows structural isomers with approximate geometries to be generated at essentially no cost. A CHARMM style force field[64–66] has been implemented to this effect, though any type of force field that allows bond definitions could in principle be used. To achieve this, a list of bonds is generated corresponding to the desired changes in connectivity, and the angles are specified based on this list. Because only rough structures are needed from the MM optimization step, success of the optimization is insensitive to the choice of bonding parameters. Therefore, bond, angle, torsion, and van der Waals parameters are chosen based on typical values from CHARMM.[64–66] Electrostatic interactions are neglected at present, but this could be changed in future implementations if charged species are important.

After the force field optimization, DFT subsequently refines the resulting structure to a true intermediate. Besides adjusting bond lengths and angles to more accurate values, the final structure after DFT optimization can sometimes be significantly different than the MM structure. This occurs because at the MM level, qualitatively poor structures are generated (high energy radicals, dissociated bonds, etc…) at the same time as reasonably correct structures. DFT therefore stabilizes this procedure by converting MM optimized structures into qualitatively correct, lower energy intermediates.

### Reducing the size of the search space

The DFT optimization provides a set of intermediates that range from high energy radicals to chemically stable complexes. At this point, a simple screening protocol based on the energy of the intermediate can be used prior to TS searching. Generally speaking, transitions to intermediates that are significantly uphill in energy also have barriers that exceed their endothermicity,[67–69] and these structures can be removed from the subsequent TS search. Cutoffs in energy can be chosen on a case-by-case basis depending on the system.

### Locating the exact TSs

Having generated a set of potentially relevant reactant/product pairs, double-ended string methods provide an automated method for locating TSs. While any such method could in principle be used, the growing string method (GSM) is applied herein. GSM is chosen because it can quickly form a reasonably accurate string and can thereafter be refined to a high-quality reaction path. The highest energy node along the GSM reaction path is used for a subsequent exact TS search. Eigenvector-following algorithms are, thus, used to refine the TS guess from GSM into the exact TS, and this can be achieved without computing the exact hessian.[47] This procedure is effective because GSM provides not only an excellent guess for the TS structure but it also provides a quality reactive tangent to serve as the TS search direction in the eigenvector-following routine. This step could in principle be replaced by a coordinate driving TS location algorithm,[70] where the driving modes would be the isomerization coordinates used to generate the intermediates. GSM uses no information about the connectivity isomerizations that occurred to provide its input—the connectivity rules no longer apply after the intermediates have been generated. Therefore, GSM finds the best reactive path that it can give two input structures, without constraints.

The endpoints of the string do not always represent a single elementary step, and this could possibly be a problem for TS finding. However, a GSM path that includes two elementary (or more) steps usually has a high barrier. This is especially the case when GSM is operated with relatively few nodes along the string. Furthermore, GSM will have high barriers for kinetically infeasible reactions that have only one elementary step. Therefore, the apparent TS barrier from GSM can serve as a screening criterion prior to the exact TS search. A cutoff for TS barrier in GSM eliminates exact TS searches for multiple elementary step and other high barrier reactions.

### Summary of methodology

All the above subsections provide a systematic procedure for identifying low barrier reaction pathways (see Fig. 2). Both the isomer generator and the overall procedure for locating elementary reactions are significant deviations from previous protocols for single-ended TS location. The approach "spontaneously" discovers potential intermediates without considering kinetic feasibility, and only later determines the barrier for formation. In this way, the difficult step of locating reaction barriers can be completed using efficient double-ended string methods.

The procedure relies on low-cost electronic structure methods such as DFT to provide gradients, intermediate energies, and reaction barriers. In principle, the method is limited by at least two factors: (1) the accuracy of the used DFT functional and (2) the scope of reactions that are captured by the connectivity rules that allow at most two connections breaking and two forming in one reactive step. Furthermore,

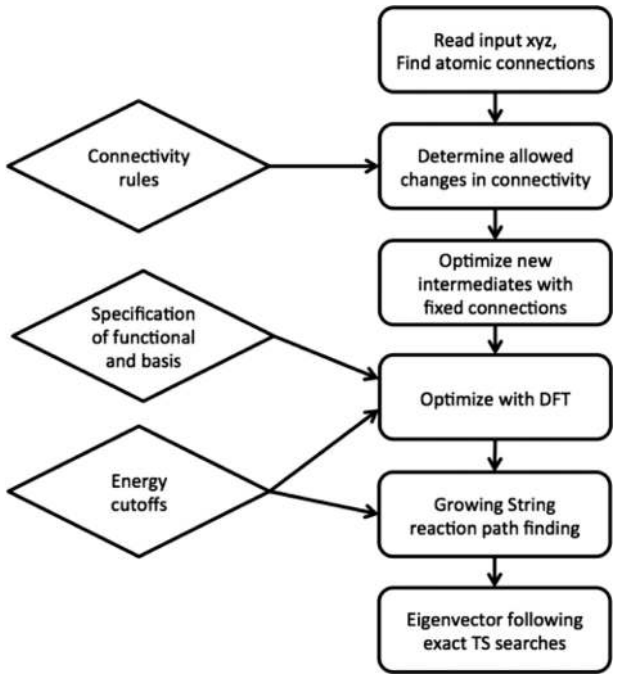


**Figure 2.** Flowchart for the generation of isomers and low barrier transition states (see "Theory" section for details of each step). Required parameters are on the left.

conformational sampling is not attempted, so this method will not be a substitute for methods such as replica exchange Monte Carlo[71] or transition path sampling[72,73] that are useful for high dimensionality, rough PESs. Instead, the method is anticipated to be most useful for which it is designed, that is, for chemical reactions involving bond breaking and forming. Finally, each elementary step location relies on the success of the double-ended string method for finding a TS; this is not a guarantee, but in practice the GSM is quite reliable.

Each DFT structure optimization is independent of the other structures, and each GSM path finding step is similarly independent. The optimization and path finding steps are implemented in embarrassingly parallel fashion to exploit the full potential of large-scale supercomputers. Besides being parallel on this basis, each DFT run can be performed in parallel as well. For instance, $N$ structures can be optimized simultaneously on $M$ cores, allowing the procedure to operate on $N$ times $M$ processor cores at the same time. To this effect, job arrays* can be used, so that all cores run at maximum efficiency.

Finally, the number of generated intermediates scales with polynomial cost in the system size ($\sim N^6$). To arrive at this factor, consider there are about $N^2$ connections that could be added, and therefore $N^4$ combinations of two added connections are possible. There exist order $N$ possible disconnections and $N^2$ possible double disconnections. Overall, two additions and two subtractions of connections total $N^6$ scaling in number of intermediates. Due to the coordination number limits, scaling will depend on the specifics of the system. As atoms with maximum or minimum coordination number will have different numbers of allowed connection addition and subtraction steps compared to atoms that are between the coordination number limits, the number of each type of atom counts in the overall scaling. This scaling reflects a great improvement over naïve PES exploration, where the cost grows exponentially with system size.

In the next section, the procedure will be validated and demonstrated in detail for four test examples. Following validation, reactive studies of two additional complexes will show the versatility and outlook for the method.

## Computational Details

The elementary step locating method is not dependent on the use of any particular density functional or quantum chemical software package. Therefore, the following choices of DFT method and basis are representative of a typical situation but could be tailored to fit the needs of any particular system. The B3LYP density functional[74–76] with the double zeta, polarized 6-31G** basis set is chosen for the DFT computations. A spin unrestricted formulism is used to not bias the results away from radical character. The elementary step finding method is implemented as a stand alone program written in C++, which invokes Q-Chem 4.0[77] to provide the quantum mechanical gradients. The MM optimization uses a conjugate gradient algorithm. The eigenvector-following exact TS search is performed in Q-Chem using the P-RFO method. Frequency computations were performed on all TS structures reported in the text to verify each contains one negative eigenvalue corresponding to the TS normal mode.

A slightly modified version of the GSM[47,48] is used for transition state searches, where linear synchronous transit in Cartesian coordinates was used for the initial interpolation technique.[49] The two input structures are aligned in Cartesian coordinates before the string is started. Eleven nodes were used to characterize the GSM path connecting the endpoint structures. GSM was considered to be completed when the sum of the perpendicular gradient magnitudes[49] reached 0.4 Hartree/Angstrom Reported reaction barriers are potential energy barriers for the true saddle point without zero point correction.

For main group elements in the following examples, the maximum and minimum coordination numbers were fixed as follows. Hydrogen is required to maintain single coordination, carbon and nitrogen must be 1–4 coordinate, and oxygen is 1–2 coordinate. An energy cutoff of 45 kcal/mol above the lowest energy intermediate is used except where noted.

## Verification

To verify the utility of the method, the reactivity of formaldehyde with $NH_3$, $H_2$, $H_2O$, and $CH_3OH$ was investigated. The starting structures are pairs of the reactant molecules optimized using DFT. These simple test cases should yield addition reactions of N—H, H—H, O—H, and O—H across the C=O bond of

*Job arrays are a common feature of computing clusters. These allow each independent process to run as soon as cores become available, and the allocated cores are freed as soon as the job completes. See http://www.adaptivecomputing.com/products/open-source/torque/.

formaldehyde, forming $NH_2CH_2OH$, $CH_3OH$, $HOCH_2OH$, and $CH_3OCH_2OH$, respectively. These reactions are not difficult when examined via chemical intuition, which would narrow the scope to essentially just one or two reactive DOF in each case. However, these examples still are computationally complex, where the majority of the $3N - 6$ DOF are not *a priori* eliminated.

**Table 1.** Structures found during elementary step search with formaldehyde complexes.

| | MM structures | DFT low *E* structures | TS found | Found addition TS? | Walltime[a] (min) |
|---|---|---|---|---|---|
| Ammonia | 69 (23) | 15 (3) | 5 (3) | Yes | 204 |
| Hydrogen | 9 (4) | 3 (2) | 2 (2) | Yes | 86 |
| Water | 34 (13) | 10 (3) | 7 (3) | Yes | 220 |
| Methanol | 154 (39) | 20 (8) | 14 (8) | Yes | 291 |

A cutoff of 45 kcal/mol is used to eliminate high energy structures. The number of unique structures is in parentheses.
[a] Computations were performed using 1 core per DFT process on nodes containing Intel X5650 processors.

For each reactant pair, Table 1 shows the number of isomers generated, the number of low energy intermediates identified by DFT, and the number of exact TSs found. In part, because the method distinguishes between chemically identical atoms (for instance the 3H on $NH_3$ are each considered unique), multiple chemically identical structures can be formed. The remainder of the chemically identical structures are formed via qualitatively different reactions leading to such intermediates. For example, H transfer from N to O may occur in the reaction of $NH_3$ and formaldehyde, which will result in a chemically identical product as simultaneous double H transfer from C to O and N to C. The latter case is arguably not an elementary step, but no such knowledge is available until the reaction pathway has been discovered. Therefore, the corresponding 'duplicate' intermediates remain necessary to study.

As shown in Table 1, not all intermediates lead to the location of a converged TS. This usually occurred because the apparent GSM barrier for these paths was high, and the exact TS search was not performed. For a handful of intermediates, the exact TS was performed but failed to locate the exact TS. Upon examination postcomputation, these runs showed two elementary step behavior, which is not explicitly sought after.

In all four test cases, the expected addition reaction steps were located, with barriers of 31, 80, 35, and 34 kcal/mol, for $NH_3$, $H_2$, $H_2O$, and $CH_3OH$, respectively. With the exception of methanol, the addition reactions had the smallest barrier heights. For methanol, another low barrier reaction was found: isodesmic two hydrogen transfer from methanol to the aldehyde with a barrier of 30 kcal/mol.

For $NH_3$, Figure 3 shows four of the low energy intermediates identified by DFT in order of increasing energy. The rule system produces structures that are qualitatively and energetically reasonable as well as intermediates that are chemically unreasonable. Due to this fact, the high energy intermediates are screened by a cutoff (45 kcal/mol above the lowest energy intermediate), and TS searches are only performed on the remaining species. The two unique TSs resulting from this search are

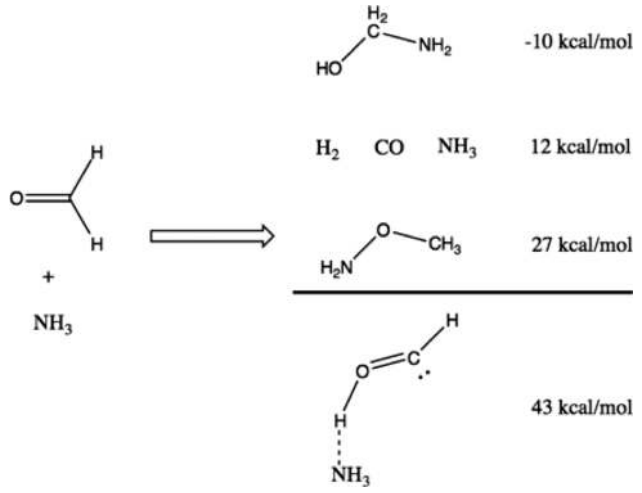


**Figure 3.** Four lowest energy intermediates identified after DFT optimization for the reaction of formaldehyde and ammonia. For clarity, chemically identical structures are not shown. The bottom right intermediate is above the threshold of 45 kcal/mol from the lowest energy intermediate and therefore is removed from the subsequent TS search.

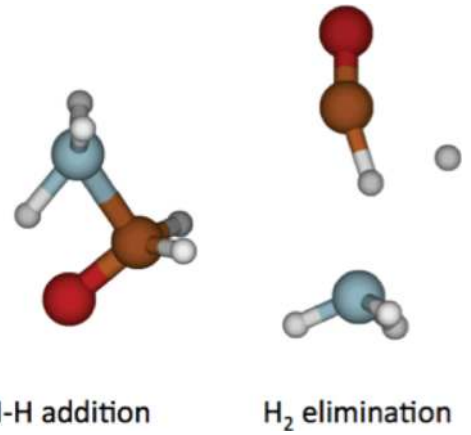


**Figure 4.** Low energy transition states for the reaction of formaldehyde with ammonia.

shown in Figure 4, where the $H_2$ elimination reaction is the second lowest barrier process and has a barrier of 80 kcal/mol. The exact TS search for the reverse N—H addition to yield $CH_3ONH_2$ was not automatically performed, because GSM reported an apparent reaction barrier that was much higher than the standard N—H addition. A search for this TS (outside of the automated procedure) found a barrier of 97 kcal/mol, indicating that neglect of this reaction path was reasonable.

In the four complexes under consideration, a significant number of structures and transition states were located. In practice, the number of structures and TSs requiring human analysis is small due to the energetic ordering provided by DFT. In principle, significant effort could be applied to determine the chemical nature of each structure, and analysis of the resulting elementary reactions could provide interesting information. However, for the purpose at hand this information is used simply as a means to an end for locating the kinetically accessible paths. Examination of the high energy structures shows many radical intermediates, and the high barrier

reactive paths either include multiple elementary steps or proceed through chemically infeasible routes (breaking C=O bonds and various other kinetically unlikely steps).

## Examples

Having verified that the reaction-finding procedure is able to locate the low-barrier paths in simple test systems, the examples that follow will further demonstrate the method's utility beyond addition reactions.

### Propene isomerization

Propene[78,79] offers a relatively simple test case that is related to more complex hydrocarbon transformations. There are several expectations about what may occur in this gas-phase isomerization. First, double bond isomerization is possible and may proceed by H transfer from the terminal $CH_3$ to $CH_2$. Ring closure, which also proceeds through H transfer, could yield cyclopropane. H transfer from the central C to the $CH_2$ carbon is conceivable to form a carbene, but this intermediate is probably high in energy. Methane elimination to yield acetylene is expected to be high barrier. $H_2$ eliminations and formation of radical intermediates are also expected to be unlikely.

Upon performing the reaction-finding procedure, 86 (16 unique) intermediates were generated and 16 (3 unique) energetically feasible structures were found after DFT optimization. Of these, 12 exact TSs were found, including three of the four reactions mentioned in the previous paragraph. The carbene intermediate was found to be too high of energy to be included in the TS search (67 kcal/mol above propene). Likewise, radicals and intermediates that formed $H_2$ were also high in energy. As expected, acetylene formation had a large barrier of over 100 kcal/mol. The relatively low barrier reactions involving ring closure and double bond isomerization were found to have barriers of 69 and 71 kcal/mol, respectively. A summary of the predicted reactivity is given in Figure 5.

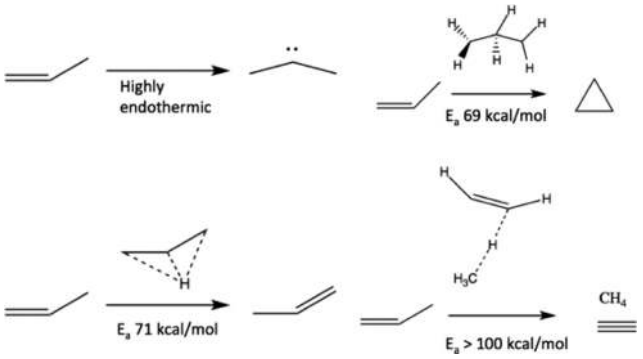


**Figure 5.** Reactivity of propene as determined by automated reaction finding.

### Ethylene and cis-butadiene reactivity

As a final example of the proposed methodology, ethylene and *cis*-butadiene were reacted. In practice, a very large number of intermediates could result due to the many possible single and double H transfers available to this system. To reduce the number of reactions, butadiene's hydrogen atoms are "frozen" and not allowed to change connectivity. Freezing the C—H connections in butadiene stops reaction of the atoms without freezing them in coordinate space. Therefore, the full accuracy of the method is available for the remaining reactions in the reduced reaction space.

The isomer generation procedure with selected frozen connections results in 50 low energy structures from DFT, reduced from over 1300 initial MM structures. Out of these 50, 15 exact TSs were located. Most of the remaining low energy structures were connected to the initial intermediate by a large TS barrier at the GSM level and were thus eliminated from the exact TS search. Six TSs were not converged at the exact TS finding level, and examination of these structures showed that they were not connected to the initial intermediate by a single elementary step. The lowest barrier TS was found to be 4 + 2 Diels–Alder cycloaddition[80,81] at 20 kcal/mol above the reactant complex. No other low barrier (less than 30 kcal/mol) TSs were found.

**Table 2.** Number of elementary steps investigated with different types of frozen coordinates for *cis*-butadiene and ethylene.

| | Number of structures | Number of low energy structures | Found 4+2 cycloaddition? |
|---|---|---|---|
| Nothing frozen | 2946 | 93 | Yes |
| C—H on butadiene frozen | 1316 | 50 | Yes |
| +2 connection only | 25 | 13 | Yes |

A second reaction-finding procedure was run with zero atoms frozen. The total number of proposed elementary steps (Table 2) increased to more than 3000, and unsurprisingly, the same low barrier 4 + 2 cycloaddition was again found. In contrast, the problem can be approached from another extreme: because steps involving addition of two atomic connections are the only expected reactions, the connectivity rules could be set to include just this type. Under this restriction, 25 intermediates were generated, and this set included the expected Diels–Alder reaction. Overall, if reactive DOF are carefully eliminated prior to the procedure, the efficiency of the method can be greatly increased while still being able to locate key elementary steps.

The reactive steps with barriers above 30 kcal/mol involved C—H activations and ring closures. A sampling of these structures, shown in Figure 6 along with their respective activation barriers, suggests the variety of chemistry achievable by the reaction path generation method. In this case, the diversity of low energy structures is relatively large, but most of these structures are kinetically inaccessible. It is important to note that generating so many thermodynamically feasible structures would be tedious without significant automation.

## Discussion

The six examples given above span a reasonably wide space of chemical reactivity. In each case, simple input parameters are used (i.e., one structure, a density functional method, and an energy cutoff), and a wide variety of intermediates were rapidly formed. Rapid intermediate generation is key to this procedure's success because the connection of two

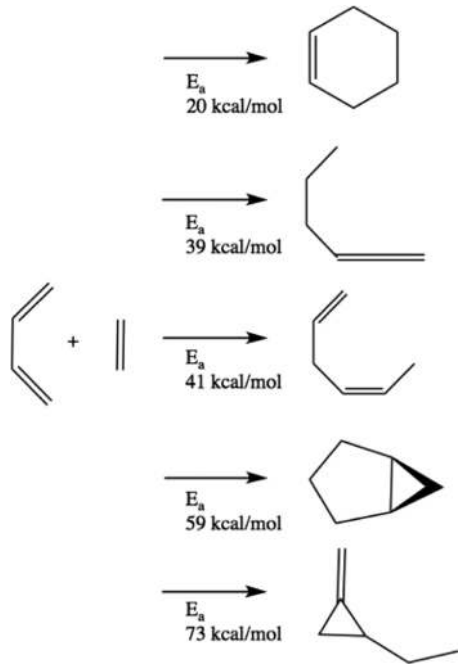


**Figure 6.** Selected elementary steps for the reaction of *cis*-butadiene and ethylene.

intermediates by a string method is relatively fast (as opposed to seeking outward toward TSs from a single intermediate). The GSM reaction paths connecting these intermediates allowed an exact TS search that revealed the energetic ordering of each potential elementary step. The degree of success in locating the kinetically favorable elementary steps was high.

In some cases, the reaction path search using GSM leads to a failure to converge specific TSs. This occurred due to the endpoints of the string not representing a pair of intermediates that are connected by a single elementary step. Although in principle GSM can locate reaction paths that include multiple steps, in practice one must use a sufficiently large number of nodes along the string. Otherwise, the path is ill-defined and has difficultly converging. As the present strategy concentrates on finding single elementary steps, the failure of GSM is often just an indication that the apparent reactive path should be divided into smaller steps. It is easy, however, to visualize the GSM reaction path after the elementary step search and determine whether a single TS connecting reactant to product is likely. In such cases, the GSM string usually shows an intermediate between the two endpoints, and it is obvious that multiple elementary steps are in play. It is conceivable that GSM could fail to find a TS even for a single elementary step reaction, as no double-ended string method is perfectly reliable. While this did not occur in this study, the limitations of the TS finding method could in principle be a challenging aspect of this approach. So far, testing indicates that GSM is reliable enough to capture the most important TSs, but future advancements in double-ended string methods will be welcome to both improve the reliability of this step and reduce the total computational cost.

An important feature of this approach is that TSs can be found using only quantum mechanical gradient computations, while higher-order derivatives are not required. This means that the method could remain efficient using *ab initio* techniques where analytical second derivatives are not available (such as many wave function methods). The success of the exact TS searches relies on the availability of a reasonably accurate vector representing the TS vibrational mode. GSM provides this transition state eigenvector from the direction of the GSM reaction path at the approximate TS.[47]

The proposed procedure for locating reactive paths is modular, because any particular step is performed independently of the others. This means that particular modules could be replaced when improved methodology becomes available. For instance, GSM could be replaced by any other double-ended string method to locate reaction paths. Another example would be the replacement of the rule system for generating intermediates with one that accounted for bond order[82] instead of atomic connectivity, resulting in a more compact set of feasible intermediates. Such changes could not only improve the procedure's efficiency, but allow the study of reactions involving transition metals.

The rule set that allows up to two connections to be formed and two broken in the same elementary step worked well for the present examples. These rules can be extended in special situations where additional connections may be broken or formed, but this will likely only happen in unusual situations (for instance in concerted chain reactions in polymers[83,84]). In larger molecules where the number of possible isomerizations becomes large, atoms that are expected to be unreactive can be frozen out of the isomerization space. This will allow the method to remain viable even with 100s or more atoms in the model.

## Conclusions

Extensive sampling of reactive space is a significant problem for atomistic simulations of chemical reactivity. The method proposed in this article provides a new approach that samples a great variety of reactions in an efficient manner and without human input. Importantly, it operates without prior knowledge of either reaction paths or intermediates beyond a single input structure. The proposed procedure neither relies on molecular dynamics sampling nor TS searches using only local information, making it distinct from previous single-ended TS finding strategies. Because the method only requires the identification of atomic connections, it can operate with any underlying quantum chemical methodology or model. In the future, this procedure can be extended to large systems using a Quantum Mechanics/Molecular mechanics (QM/MM) methodology[85–88] where the QM region is considered the reactive region, and connections are left intact in the MM region.

Current limitations include that only main group elements have been considered, but an appropriate choice of force field and definitions of connectivity changes in transition metals could extend the method's reach. The slowest computational step is applying GSM to find the reaction path, and this might be alleviated by using faster string methods.[51] Looking further forward, careful connection of kinetically feasible elementary steps could yield multistep reaction mechanisms beginning from a single intermediate. Progress in these regards will be presented in future publications.

## Acknowledgments

*P.M.Z. thanks the Center for Advanced Computing at the University of Michigan for computational time and Pavel Nagorny for useful discussions.*

**Keywords:** reaction simulation · elementary reactions · chemical mechanism · double-ended string methods · transition state · chemical automation

How to cite this article: P. M. Zimmerman, *J. Comput. Chem.* **2013,** *34,* 1385–1392. DOI: 10.1002/jcc.23271

[1] A. T. Ziegler, *Chem. Rev.* **1991**, *91,* 651.
[2] W. Kohn, A. D. Becke, R. G. Parr, *J. Phys. Chem.* **1996**, *100,* 12974.
[3] R. G. Parr, W. Yang, *J. Am. Chem. Soc.* **1984**, *106,* 4049.
[4] Y. Zhao, D. G. Truhlar, *J. Chem. Theor. Comput.* **2005**, *1,* 415.
[5] C. J. Cramer, D. G. Truhlar, *Phys. Chem. Chem. Phys.* **2009**, *11,* 10757.
[6] A. T. Bell, M. Head-Gordon, *Annu. Rev. Chem. Biomol. Eng.* **2011**, *2,* 453.
[7] F. J. Keil, *Top. Curr. Chem.* **2012**, *307,* 69.
[8] M. Laio, M. Parrinello, *Proc. Natl. Acad. Sci. USA* **2002**, *99,* 12562.
[9] M. Ianuzzi, A. Laio, M. Parrinello, *Phys. Rev. Lett.* **2003**, *90,* 238302.
[10] B. Ensing, M. de Vivo, Z. Liu, P. Moore, M. L. Klein, *Acc. Chem. Res.* **2006**, *39,* 73.
[11] E. M. Mueller, A. de Meijere, H. Grubmueller, *J. Chem. Phys.* **2002**, *116,* 897.
[12] M. Chen, M. A. Cuendet, M. E. Tuckerman, *J. Chem. Phys.* **2012**, *137,* 24102.
[13] C. J. Cerjan, W. H. Miller, *J. Chem. Phys.* **1981**, *75,* 2800.
[14] J. Simons, P. Jorgensen, H. Taylor, J. Ozment, *J. Phys. Chem.* **1983**, *87,* 2745.
[15] E. Cancès, F. Legoll, M. C. Marinica, K. Minoukadeh, F. Willaime, *J. Chem. Phys.* **2009**, *130,* 114711.
[16] D. Poppinger, *Chem. Phys. Lett.* **1975**, *35,* 550.
[17] B. Peters, W.-Z. Liang, A. T. Bell, A. Chakraborty, *J. Chem. Phys.* **2003**, *118,* 9533.
[18] S. Maeda, K. Ohno, K. Morokuma, *J. Chem. Theory Comput.* **2009**, *5,* 2743.
[19] S. Maeda, K. Morokuma, *J. Chem. Theor. Comput.* **2011**, *7,* 2335.
[20] S. Maeda, E. Abe, M. Hatanka, T. Taketsugu, K. Morokuma, *J. Chem. Theor. Comput.* **2012**, *8,* 5058. DOI: 10.1021/ct300633e.
[21] R. Granot, R. A. Baer, *J. Chem. Phys.* **2008**, *128,* 184111.
[22] S. A. Ghasemi, S. Goedecker, *J. Chem. Phys.* **2011**, *135,* 014108.
[23] H. Chaffey-Millar, A. Nikodem, A. V. Matveev, S. Krüger, N. Rösch, *J. Chem. Theory Comput.* **2012**, *8,* 777.
[24] H. B. Schlegel, *J. Comput. Chem.* **1982**, *3,* 214.
[25] G. Henkelman, H. Jonsson, *J. Chem. Phys.* **1999**, *111,* 7010.
[26] A. Heyden, A. T. Bell, F. J. Keil, *J. Chem. Phys.* **2005**, *123,* 224101.
[27] J. Baker, *J. Comput. Chem.* **1986**, *7,* 385.
[28] D. J. Wales, *J. Chem. Soc., Faraday Trans.* **1992**, *88,* 653.
[29] P. Y. Ayala, H. B. Schlegel, *J. Chem. Phys.* **1997**, *107,* 375.
[30] J. M. del Campo, A. M. Koster, *J. Chem. Phys.* **2008**, *129,* 024107.
[31] H. B. Schlegel, *WIREs Comput. Mol. Sci.* **2011**, *1,* 790.
[32] H. B. Schlegel, *J. Comput. Chem.* **2003**, *24,* 1514.
[33] J. Baker, A. Kessi, B. Delley, *J. Chem. Phys.* **1996**, *105,* 192.
[34] C. Peng, P. Y. Ayala, H. B. Schlegel, M. J. Frisch, *J. Comput. Chem.* **1996**, *17,* 49.
[35] C. Peng, H. B. Schlegel, *Isr. J. Chem.* **1994**, *33,* 449.
[36] G. Mills, H. Jónsson, *Phys. Rev. Lett.* **1994**, *72,* 1124.
[37] G. Henkelman, H. Jonsson, *J. Chem. Phys.* **2000**, *113,* 9978.
[38] G. Henkelman, B. P. Uberuaga, H. Jonsson, *J. Chem. Phys.* **2000**, *113,* 9901.
[39] S. A. Trygubenko, D. J. Wales, *J. Chem. Phys.* **2004**, *120,* 2082.
[40] J. Chu, B. Trout, B. A. Brooks, *J. Chem. Phys.* **2003**, *119,* 12708.
[41] D. Sheppard, R. Terrell, G. Henkelman, *J. Chem. Phys.* **2008**, *128,* 134106.
[42] W. E, W. Ren, E. Vanden-Eijnden, *Phys. Rev. B* **2002**, *66,* 052301.
[43] W. E, W. Ren, E. Vanden-Eijnden, *J. Phys. Chem. B* **2005**, *109,* 6688.
[44] W. Ren, E. Vanden-Eijnden, *J. Chem. Phys.* **2007**, *126,* 164103.
[45] S. K. Burger, W. Yang, *J. Chem. Phys.* **2006**, *24,* 054109.
[46] S. K. Burger, W. Yang, *J. Chem. Phys.* **2007**, *127,* 164107.
[47] S. M. Sharada, P. M. Zimmerman, A. T. Bell, M. Head-Gordon, *J. Chem. Theor. Comput.* **2012**, *8,* 5166.
[48] B. Peters, A. Heyden, A. T. Bell, A. Chakraborty, *J. Chem. Phys.* **2004**, *120,* 7877.
[49] A. Behn, P. M. Zimmerman, A. T. Bell, M. Head-Gordon, *J. Chem. Theor. Comput.* **2011**, *7,* 4019.
[50] W. A. Quapp, *J. Chem. Phys.* **2005**, *122,* 174106.
[51] A. Behn, P. M. Zimmerman, A. T. Bell, M. Head-Gordon, *J. Chem. Phys.* **2011**, *135,* 224108.
[52] M. H. Todd, *Chem. Soc. Rev.* **2005**, *34,* 247.
[53] J. H. Chen, P. Baldi, *J. Chem. Ed.* **2008**, *85,* 1699.
[54] J. H. Chen, P. Baldi, *J. Chem. Inf. Model.* **2009**, *49,* 2034.
[55] M. A. Kayala, C.-A. Azencott, J. A. Chen, P. Baldi, *J. Chem. Inf. Model.* **2001**, *51,* 2209.
[56] M. A. Kayala, P. Baldi, *J. Chem. Inf. Model.* **2012**, *52,* 2526.
[57] R. Vinu, L. J. Broadbelt, *Ann. Rev. Biomol. Eng.* **2012**, *3,* 29.
[58] S. S. Khan, Q. Zhang, L. J. Broadbelt, *J. Atmos. Chem.* **2009**, *63,* 125.
[59] S. S. Khan, L. J. Broadbelt, *J. Atmos. Chem.* **2009**, *63,* 157.
[60] T. M. Kruse, O. S. Woo, H.-W. Wong, S. S. Khan, L. J. Broadbelt, *Macromolecules* **2002**, *35,* 7830.
[61] L. P. Hillewart, J. L. Dierickx, G. F. Froment, *AIChE J.* **1988**, *34,* 17.
[62] R. Sumathi, W. H. Green, *Theor. Chem. Acc.* **2002**, *108,* 187.
[63] V. Bakken, T. Helgaker, *J. Chem. Phys.* **2002**, *117,* 9160.
[64] N. Foloppe, A. D. Mackerell, *J. Comput. Chem.* **2000**, *21,* 86.
[65] D. Yin, A. D. Mackerell, *J. Comput. Chem.* **1998**, *19,* 334.
[66] K. Vanommeslaeghe, E. Hatcher, C. Acharya, S. Kundu, S. Zhong, J. Shim, E. Darian, O. Guvench, P. Lopes, I. Vorobyov, A. D. Mackerell, *J. Comput. Chem.* **2009**, *31,* 671.
[67] B. C. Gates, Catalytic Chemistry; Wiley: New York, **1992**.
[68] J. E. Sutton, D. G. Vlachos, *ACS Catal.* **2012**, *2,* 1624.
[69] V. Pallassana, M. Neurock, *J. Catal.* **2000**, *191,* 301.
[70] J. M. Bofill, J. M. Anglada, *Theor. Chem. Acc.* **2001**, *105,* 463.
[71] Y. Sugita, Y. Okamoto, *Chem. Phys. Lett.* **1999**, *314,* 141.
[72] C. Dellago, P. G. Bolhuis, *Adv. Polym. Sci.* **2011**, 1.
[73] P. G. Bolhuis, D. Chandler, C. Dellago, P. L. Geissler, *Annu. Rev. Phys. Chem.* **2002**, *53,* 291.
[74] A. D. Becke, *Phys. Rev. A* **1998**, *38,* 3098.
[75] C. Lee, W. Yang, R. G. Parr, *Phys. Rev. B* **1988**, *37,* 785.
[76] A. D. Becke, *J. Chem. Phys.* **1993**, *98,* 5648.
[77] Y. Shao, L. Fusti-Molnar, Y. Jung, J. Kussmann, C. Ochsenfeld, S. T. Brown, A. T. B. Gilbert, L. V. Slipchenko,S. V. Levchenko, D. P. O'Neill, R. A. Distasio Jr., R. C. Lochan, T. Wang, G. J. O. Beran, N. A. Besley, J. M., Herbert, C. Y. Lin, T. Van Voorhis, S. H. Chien, A. Sodt, R. P. Steele, V. A. Rassolov, P. E. Maslen, P. P. Korambath, R. D. Adamson, B. Austin, J. Baker, E. F. C. Byrd, H. Dachsel, R. J. Doerksen, A. Dreuw, B. D. Dunietz, A. D. Dutoi, T. R. Furlani, S. R. Gwaltney, A. Heyden, S. Hirata, C.-P. Hsu, G. Kedziora, R. Z. Khaliullin, P. Klunzinger, A. M. Lee, M. S. Lee, W. Liang, I. Lotan, N. Nair, B. Peters, E. I. Proynov, P. A. Pieniazek, Y. M. Rhee, J. Ritchie, E. Rosta, C. D. Sherrill, A. C. Simmonett, J. E. Subotnik, H. L. Woodcock III, W. Zhang, A. T. Bell, A. K. Chakraborty, D. M. Chipman, F. J. Keil, A. Warshel, W. J. Hehre, H. F. Schaefer III, J. Kong, A. I. Krylov, P. M. W. Gill, M. Head-Gordon, *Phys. Chem. Chem. Phys.* **2006**, *8,* 3172.
[78] T. S. Chambers, G. B. Kistiakowsky, *J. Am. Chem. Soc.* **1934**, *56,* 399.
[79] H. F. Bettinger, J. C. Rienstra-Kiracofe, B. C. Hoffman, H. F. Schaefer, J. E. Baldwin, P. v. R. Schleyer, *Chem. Commun.* **1999**, *1999,* 1515.
[80] O. Diels, L. Alder, *Liebigs Ann. Chem.* **1928**, *460,* 98.
[81] L. F. Tietze, G. Kettschau, *Top. Curr. Chem.* **1997**, *189,* 1.
[82] E. D. Glendening, C. R. Landis, F. Weinhold, *WIREs Comput. Mol. Sci.* **2012**, *2,* 1.
[83] R. A. Yoder, J. N. Johnston, *Chem. Rev.* **2005**, *105,* 4730.
[84] I. Vilotijevic, T. F. Jamison, *Science* **2007**, *317,* 1189.
[85] P. M. Zimmerman, M. Head-Gordon, A. T. Bell, *J. Chem. Theory Comput.* **2011**, *7,* 1695.
[86] H. Lin, D. G. Truhlar, *Theor. Chem. Acc.* **2007**, *117,* 185.
[87] D. Bakowies, W. Thiel, *J. Phys. Chem.* **1996**, *100,* 10580.
[88] A. H. de Vries, P. Sherwood, S. J. Collins, A. M. Rigby, M. Rigutto, G. J. Kramer, *J. Phys. Chem. B* **1999**, *103,* 6133.

Received: 11 December 2012
Revised: 15 January 2013
Accepted: 18 February 2013.
Published online on 18 March 2013