Identifying barrierless mechanisms for benzene formation in the interstellar medium using permutationally-invariant reaction discovery

Christopher Robertson, Ross Hyland, Andrew J. D. Lacey, Sebastian Havens, and Scott Habershon*

Department of Chemistry, University of Warwick, Coventry, CV4 7AL, United Kingdom

E-mail: S.Habershon@warwick.ac.uk

Abstract

Complex chemical reaction environments, such as those found in combustion engines, the upper atmosphere, or the interstellar medium, can contain large numbers of different reactive species participating in similarly-large numbers of different chemical reactions. In such settings, identifying the most-likely multi-step reaction mechanisms which lead to the production of a particular defined product species is an extremely challenging problem, requiring search and evaluation over a large number of different possible candidate mechanisms while also addressing the permutational challenges posed when considering large number of reaction routes available to sets of identical molecular species. In this article, the problem of generating candidate reaction mechanisms which form a defined product from a diverse set of reactive molecules is cast as a discrete optimization of a permutationally-invariant cost function describing similarity between the target product and the product generated by a trial reaction mechanism. This approach is demonstrated by generating 2230 candidate reaction mechanisms which form benzene from diverse sets of reactive molecules which have been experimentally-identified in the interstellar medium. By screening this set of auto-generated mechanisms, using dispersion-corrected DFT to evaluate reaction energies and activation barriers, we identify several candidate barrierless reaction mechanisms (both previously-proposed and new) for benzene formation which may operate in the low temperatures found in the interstellar medium, and could be investigated further to supplement existing microkinetic models.

1 Introduction

Simulations which enable automated generation of complex reaction networks or proposal of chemical reaction mechanisms are experiencing a significant growth of interest, powered by more accurate and efficient *ab initio* electronic structure methods for evaluating molecular energies,[1–5] development of new algorithms for key computational tasks such as transition-state (TS) finding),[6–14] and the explosion of machine-learning and artificial intelligence tools for computational chemistry.[15–24] To date so-called reaction discovery tools have been developed in an array of different computational flavours. For example, deep learning strategies have been used to assimilate the information found in common databases of organic chemical reaction outcomes, and use it to predict the reaction products of new reactions, in many cases enabling identification of synthetic routes comparable to those formulated by human experts.[15,25,26] Simulations based on accelerated sampling in traditional molecular dynamics (MD) schemes have been used to investigate formation of amino acids from simpler reactants, echoing the classic Urey-Miller experiments, or in the determination of catalytic cycles and thermal molecular decomposition pathways.[9,13,27–29] As a final example of emerging simulation classes, heuristic tools based on graph theory, in combination with *ab initio* calculations, have been used to auto-build reaction networks for a varied set of chemical reaction systems, including nanoparticle-catalyzed industrial reactions, combustion processes and organometallic catalysis.[30–36] Together, these exciting new simulation approaches are increasingly providing a direct connection between atomistic data (*e.g.* optimized geometries, TSs) and macroscopic kinetic observables such as rate laws and product selectivities.

Our recent work has focussed on the development of graph-based tools for discovery of multistep reaction mechanisms.[30,31,37–39] As described below, we formulate the challenge of mechanism proposal as a discrete optimization problem; we seek the set of chemically-sensible reaction steps which transform the connectivity matrix (CM) of the input reactants into the CM of the user-defined products. This search is achieved using a simulated annealing (SA) optimization procedure, and has been shown to be very fast (because it operates purely in

the space of molecular CMs), and also generally applicable. For example, we have shown how our double-ended graph-driven sampling (GDS) scheme can be used to propose reaction mechanisms for carbon monoxide oxidation, hexane aromatization, and the water-gas shift reaction.[31] We have also shown that, when combined with semi-empirical or *ab initio* electronic structure calculations, our GDS scheme can correctly identify the accepted 'correct' mechanism of catalysis by an organometallic complex.[38]

These same successful calculations have also highlighted two important algorithmic issues in our GDS approach. First, our approach currently does not account for permutational invariance amongst reactive atoms and molecules when searching for reaction mechanisms. In particular, our current algorithm seeks the set of reaction steps which lead from a user-defined set of reactant molecules to a user-defined set of product molecules; in other words, there is a requirement that there is a known one-to-one mapping between the atoms in the reactants and those in the products. This requirement demands that users define exactly where each atom in the reactant molecules ends up in the product molecules; of course, the demand that users exactly define the molecular structures *and* component atoms in the product molecules is clearly running counter to the overall goal of automated reaction discovery. Second, a closely-related problem is the fact that our current reaction-discovery algorithm demands that the CMs defining reactants and target products must have an equal number of atoms; this stems from the absence of permutational invariance noted above, as well as the manner in which we quantify the fitness of a candidate reaction mechanism, as described below. Again, this demand of equal numbers of atomic species in reactants and products runs counter to our original goal of automated reaction discovery.

The drawbacks relating to permutational invariance are particularly undesirable when considering mechanism discovery in complex reactant mixtures such as those found in combustion engines (*i.e.* complex mixtures of hydrocarbons, O_2, H_2O, NOx), Earth's atmosphere (*e.g.* volatile organic compounds, O_2, H_2O, HNO_3, H_2SO_4, CO_2 and more) and the interstellar medium (*e.g.* organic radicals and ions, ice and dust, H_2, H_2O). For example, the last

few decades has seen growing interest in the possible mechanisms by which polycyclic aromatic hydrocarbons (PAHs) form in the interstellar medium (ISM);[40–48] such molecules are a key component of the ISM, accounting for a significant portion of the ISM carbon budget, and implicated in the long-term evolution of the gas-phase and surface-based ISM chemistry. In particular, formation mechanisms for the simplest cyclic aromatic hydrocarbon, benzene (C_6H_6) have been proposed and studied by both computational and experimental methodologies,[41,45,47,49,50] leading to a series of postulated mechanisms including ion-molecule reactions[48] and barrierless radical reactions which can yield benzene.[41] However, it is clear that these chemical settings, seeking product formation mechanisms starting from a large number of possible reactant species, are exactly the sorts of systems which automated reaction discovery tools should be able to address.

With this goal in mind, this Article aims to address the challenges associated with permutational invariance in our GDS algorithm. In Section 2, we explain how our double-ended GDS approach can be modified to account for atomic permutational invariance, and how an optimization cost-function can be developed which enables us to seek reaction mechanisms leading to a single target product molecule from a large number of input reactant molecules. In Section 3, we then demonstrate these developments by seeking barrierless mechanisms leading to formation of benzene from neutral organic molecular species which are typically found in low-temperature interstellar environments; to the best of our knowledge, this is the first time that a reaction discovery method has been applied to such a problem. After generating 2230 candidate reaction mechanisms leading to benzene, we use density functional theory (DFT) calculations to calculate the reaction energetics and reaction barriers, enabling us to identify a set of 126 reaction mechanisms which form benzene with either very low or zero effective overall energetic barrier; some of these reactions have been previously suggested from computational or experimental investigations, while other mechanisms, to the best of our knowledge, have not been considered previously. In Section 4, we conclude by highlighting some further possible avenues to improve the overall efficiency of automated

mechanism-discovery tools for complex reactive systems.

2 Theory

We begin by briefly outlining our GDS approach to reaction-mechanism generation, as has been reported previously. Then we highlight the problems introduced by permutational invariance of atomic species in our existing GDS approach; we then go on to show how these problems can be addressed by modification of the optimization cost-function employed in GDS.

2.1 Existing double-ended graph-driven sampling method

Our GDS approach to reaction-mechanism discovery has been described in detail elsewhere;[30,31,37,38] here, we present only a brief outline of the relevant aspects.

GDS is based on the concept of a molecular CM, an $n \times n$ square matrix for any n-atom system whose entries simply identify whether or not two atoms are bonded, regardless of the type of bonding. For our purposes, it is sufficient to define the elements of the CM $\mathbf{G}$ as:

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

Here, r_{ij} is the distance between two atoms i and j, and r_{ij}^{cut} is a fixed cutoff value which depends on the atom types; broadly, it is proportional to the sum of the covalent radii of the atomic elements. Extensions of such CMs to account for intermolecular encounter complexes have also been proposed,[29] but we limit the discussion here to the standard case CM above.

At the start of a GDS calculation, the user must define the input reactant structure (typically a set of reactant molecules) and the target product structure; these are defined by vectors of length $3n$, labelled as $\mathbf{r}^R$ and $\mathbf{r}^P$ respectively. As noted above, we assume for now that these structures must be identically atom-ordered; in other words, the input is such

that the user must make a choice as to which atoms in the input reactant structure map onto which atoms in the target product structure.

In addition to the reactant and target product structures, the user also provides a library of chemically-allowed reaction classes.[30,31,37,38] This library defines which generic types of chemical reactions should be available to GDS in the search for a mechanism connecting $\mathbf{G}^R$ to $\mathbf{G}^P$. Commonly, we employ a broad set of reaction classes, including atomic association/dissociation events, three-atom insertion/elimination, and four-atom shift reactions; of course, the set of library reactions can be made as small or as large as desired, although with implications for the efficiency of the GDS optimization. In addition, we note that it is trivial to include chemical prior knowledge in this scheme; for example, in the context of heterogeneous catalysis, one might require that any reaction takes place at the surface of a metal cluster. Such constraints are simple to implement within the GDS scheme, but are not used here. The set of reaction classes used in the simulations performed here is further described below, and in the *Supplemental Information*.

The final input required before a GDS calculation is the identification of chemical constraints which must be obeyed by any GDS-proposed chemical reaction mechanism. Again, as in the case of the definition of the reaction-class library, the definition of chemical constraints can be kept deliberately broad or can be tuned to the system of interest. Most commonly, we impose simple atomic valence constraints in order to prevent generation of intermediate molecular structures with nonsensical atomic valences; for example, we typically constrain the valence of carbon atoms to lie between one and four. As in the case of the reaction classes noted above, the valence constraints can be readily tightened or loosened to restrict or broaden the mechanism search. The valence constraints employed in the calculations reported here are detailed in the *Supplemental Information*.

With the definition of reactant structure, product structure, reaction classes and chemical constraints in place, GDS then proceeds to locate a reaction mechanism (that is, a series of elementary reaction-steps) which connects $\mathbf{G}^R$ to $\mathbf{G}^P$, subject to the input chemical

constraints and using the set of reaction-classes defined in the input reaction library. To identify a reaction-mechanisms connecting $\mathbf{G}^R$ to $\mathbf{G}^P$, our GDS scheme treats the problem as a challenge in discrete optimization. In particular, we proceed by proposing a reaction-mechanisms comprising n_r elementary reaction step; for each elementary step, we define the reaction-class of the ith elementary step, $k(i)$, and the corresponding atomic indices, $\mathbf{I}_i$, to which this reaction-class is applied. An example of this is shown in Fig. 1, which shows the application of $n_r = 2$ sequential elementary reactions. In the first step, the selected reaction class involves simple bond formation, and the selected atomic indices are $(1, 3)$; in the second step, the three-atom reaction involves dissociation of molecular oxygen, with atoms $(1, 2, 4)$ participating. From this example, it is clear that a given reaction-mechanism of n_r steps can be encoded as a list of reaction-classes $\{k_i\}_{i=1}^{n_r}$, and a corresponding list of atomic indices to which each of these reactions is applied, $\{\mathbf{I}_i\}_{i=1}^{n_r}$. The sets $\mathbf{k}$ and $\mathbf{I}$ therefore represent discrete parameters which can be changed to yield alternative reaction-mechanisms.

After application of a sequence of n_r reactions, the resulting CM $\bar{\mathbf{G}}$ is given by

$$\bar{\mathbf{G}} = \mathbf{G}^R + \sum_{i=1}^{n_r} \mathbf{R}^{k_i}(\mathbf{I}_i), \tag{2}$$

where the summation implies sequential application of the CM operations encoded by the reaction-classes and associated atomic indices; $\mathbf{R}^{k_i}(\mathbf{I_i})$ implies that the CM should be updated by application of reaction-class k_i to atomic indices $\mathbf{I}_i$. Within this scheme, the reaction-mechanism discovery process can be viewed as a discrete optimization challenge in which one seeks to find a sequence of n_r reaction-steps (that is, both reaction-classes $\mathbf{k}$ and atomic indices $\mathbf{I}$) resulting in a final graph-error function $F = 0$, where

$$F = \sum_{j,k<j} (\bar{G}_{jk} - G^P_{jk})^2 \tag{3}$$

The graph-error function F simply enumerates the number of incorrect elements (*i.e.* bonds) in the CM generated after application of n_r proposed reaction steps ($\bar{\mathbf{G}}$) relative to the target

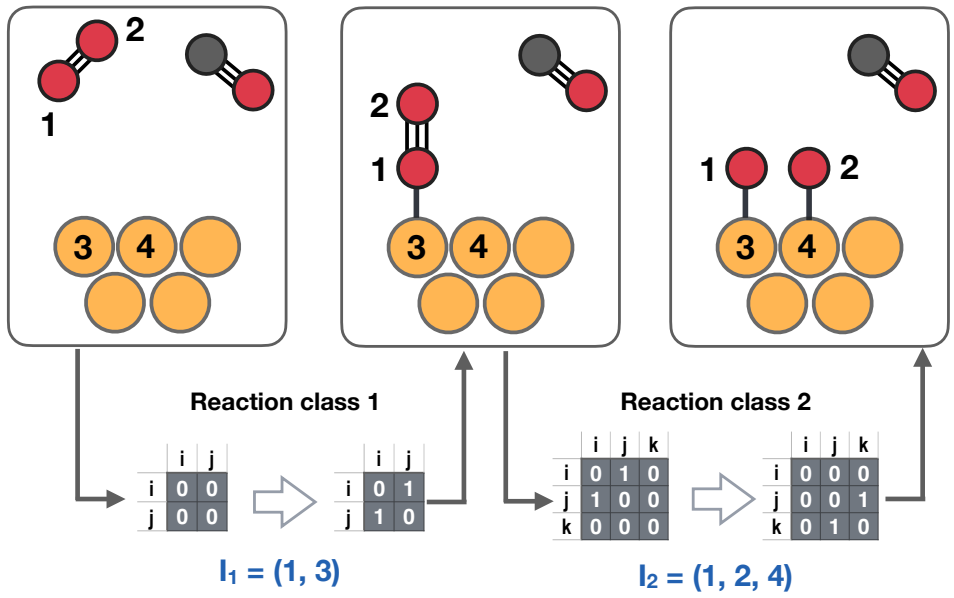


Figure 1: Schematic illustration of a proposed reaction-mechanism in the GDS approach; here, red represents oxygen, grey represents carbon and gold atoms represent platinum. In this example, the mechanism comprises two steps, the first (application of reaction class 1) involving two atoms and the second (application of reaction class 2) involving three atoms; in each case, the corresponding CMs before and after reaction are shown, and the corresponding atomic indices are also defined. This mechanism is therefore defined by a set of discrete integers defining reaction class, $\mathbf{k} = (1, 2)$, and reactive indices, $\mathbf{I} = [(1, 3), (1, 2, 4)]$.

product CM ($\mathbf{G}^P$).

With target optimization function F in hand, we then proceed to find a reaction-mechanism with $F = 0$; in other words, we seek to find the set of reaction-classes $\{k_i\}_{i=1}^{n_r}$ and reactive atomic indices $\{\mathbf{I}_i\}_{i=1}^{n_r}$ such that the final target product is formed after a given number of reaction steps n_r. This search can be performed using any algorithm suited to discrete optimization; in our recent work, and in this Article, we use a simulated annealing (SA) scheme to perform minimization of F. Here, for a fixed number of iterations n_{SA}, we perform Metropolis updates in which one of the following moves are performed with equal probability:

- Change the atomic indices $\mathbf{I}_j$ of a randomly chosen reaction-step j;
- Change both the reaction class k_j **and** the atomic indices $\mathbf{I}_j$ of a randomly chosen

reaction-step j.

After an update of the reaction-sequence, the new value of the graph-error function F is evaluated and the move is accepted or rejected using the usual Metropolis criterion. In our SA calculations, we simply employ a linear cooling regime over the course of n_{SA} iterations, starting at an initial temperature T_i. The success rate of this scheme in finding reaction mechanisms, as noted below, is typically over 90% in the reactions we have considered to date.

The imposition of atomic and molecular valence constraints is straightforward within our reaction-mechanism optimization scheme; such constraints can be used to ensure that mechanism searching is limited to 'chemically-sensible' pathways. This is achieved during our SA calculations by attributing an arbitrarily large F-value to any proposed reaction-mechanism which generates an intermediate structure that disobeys any of the user-defined target constraints. For example, if carbon valences v_C are restricted to lie within the range $v_C \in [1, 4]$, then any proposed reaction sequence which results in a reaction intermediate with v_C lying outside this range will trigger a large F value, such that the proposed reaction sequence will be inevitably rejected by the Metropolis scheme. The actual valence constraints used in the simulations discussed in this paper are highlighted in the *Supplementary Information*; however, we emphasize that these constraints can be as rigid or as flexible as desirable by the user.

To summarize, the GDS algorithm described above is a flexible and efficient scheme to rapidly identify sequences of elementary reaction steps which lead from user-defined reactants to target products. Because the optimization of F operates solely in the space of the CMs of the reactants, reaction intermediates and the products, it is very fast, typically locating a candidate reaction-mechanism with $F = 0$ in a few minutes on a standard desktop computer.

2.2 Generating intermediate molecular structures

The outcome of the reaction-mechanism-finding algorithm above is a sequence of reaction classes and reactive atomic indices, $[\mathbf{R}^{k_1}(\mathbf{I}_1), \mathbf{R}^{k_2}(\mathbf{I}_2), \ldots, \mathbf{R}^{k_{n_r}}(\mathbf{I}_{n_r})]$, which lead to formation of the target product structure. The next task is to convert from 'graph-space' to 'real-space', generating Cartesian atomic coordinates for all molecules along the proposed reaction-path.

To achieve this, we use the concept of graph-restraining potential (GRP), as introduced in our previous work.[30,31,37,38] In brief, we define a simple empirical potential energy function, $W(\mathbf{r}, \mathbf{G})$, which is designed to be a minimum when the set of Cartesian coordinates $\mathbf{r}$ is consistent with a target CM, $\mathbf{G}$; as such, starting from a set of atomic coordinates and a target CM $\mathbf{G}$, minimization of $W(\mathbf{r}, \mathbf{G})$ with respect to the atomic coordinates yields a structure in which $\mathbf{r}$ obeys the bonding pattern encoded in $\mathbf{G}$.

The GRP function has the following (somewhat arbitrary) form:

$$W(\mathbf{r}, \mathbf{G}) = \sum_{j>i} \Big[\delta(G_{ij} - 1)[H(r_{ij}^{min} - r_{ij})\sigma_1(r_{ij}^{min} - r_{ij})^2 + H(r_{ij} - r_{ij}^{max})\sigma_1(r_{ij}^{max} - r_{ij})^2] + \delta(G_{ij})\sigma_2 e^{-r_{ij}^2/(2\sigma_3^2)} \Big] + V_{mol}(\mathbf{r}, \mathbf{G}). \tag{4}$$

Here, the first summation runs over all pairs of atoms, $\delta(x)$ is the Dirac delta function and $H(x)$ is the Heaviside step function. The first term in the parentheses, which is multiplied by $\delta(G_{ij} - 1)$, only operates on pairs of atoms which *should* be bonded (so have $G_{ij} = 1$ in the target CM $\mathbf{G}$, and hence $\delta(G_{ij} - 1) = 1$); the two harmonic terms provide a resulting force which pushes the atoms i and j together until they lie at a distance $r_{ij} \in [r_{ij}^{min}, r_{ij}^{max}]$, where r^{min} and r^{max} are suitably chosen bonding-distance limits which are simply tabulated for all pairs of atoms. Ultimately, we will use geometry optimization on *ab initio* PESs to generate the final molecular structures at each intermediate reaction step, so the exact definition of these limits is somewhat flexible; we typically define these limits as the sum of the atomic covalent radii plus/minus some small flexibility on the order of 0.1-0.2 Å. In contrast, the

Gaussian term, preceded by $\delta(G_{ij})$, operates on atoms which are expected to be non-bonded; the Gaussian function acts as a simple repulsive wall, with a range and strength related to the parameters σ_3 and σ_2, respectively.

The final term in Eq. 4, $V_{mol}(\mathbf{r}, \mathbf{G})$, acts on separate *molecules* within the structure to ensure that they are neither too close or too far from each other (noting that a single atom, not bonded to anything else, is considered to be a "molecule" in this context), and is defined as

$$V_{mol}(\mathbf{r}, \mathbf{G}) = \sum_{J>I} [H(R^{min} - R_{IJ})\sigma_4(R^{min} - R_{IJ})^2, + H(R_{IJ} - R^{max})\sigma_4(R^{max} - R_{IJ})^2]. \quad (5)$$

Here, R_{IJ} is the distance between the centres-of-mass of two molecules I and J, and R^{min} and R^{max} are user-defined minimum and maximum distances between any pair of molecules. Finally, we note that the parameters σ_{1-4} are somewhat arbitrarily chosen to generate sensible molecular structures for the broad class of molecular species investigated to date; these parameters are given in the *Supplementary Information*.

Given a series of CMs generated by our double-ended GDS algorithm above, the GRP function $W(\mathbf{r}, \mathbf{G})$ allows us to generate atomic coordinates which are consistent with each intermediate CM; in other words, the GRP converts from a string of CMs to a sequence of molecular structures, with each structure corresponding to one of the intermediate steps of the proposed reaction-mechanism. With the atomic coordinates of each intermediate structure available, we can subsequently perform standard *ab initio* electronic structure calculations to generate optimized geometries and calculate relative free energies, while application of standard MEP-finding[51–53] and TS-finding algorithms similarly allow evaluation of activation energies and approximate reaction rates (through TST,[54–57] for example).

Although the GRP function leads to a unique set of reaction intermediate structures for a given set of initial atomic coordinates for the reactant structure, we note that some ambiguity remains due to the incomplete nature of CMs. In particular, CMs do not define either

the stereochemistry or the dihedral conformation of the molecular intermediates. Instead, in our approach, these aspects are imposed in the selection of the atomic coordinates of the reactant structure, in the sense that each intermediate structure along a given reaction-path is generated starting from the atomic coordinates of the previous structure. In the ISM reactions noted below, the absence of stereochemistry is not important due to the absence of stereochemical centers. In addition, many of the reactive species have just a small number of dihedral angles, leading to limited conformational flexibility. However, we note that future applications focussing on stereocenters and/or conformational flexibility will require consideration of these factors; for example, conformer searching based on empirical force-fields can help search for different intermediate reactive conformers, and similarly the stereochemistry of newly-formed molecules can also be considered as a conformational parameter to be sampled after initial GDS/GRP generation of a reaction-mechanism. While these extensions to our GRP approach could go beyond our standard CM-based intermediate structure generation, we note that both of these extensions will inevitably impose greater demands on the number of reactions to analyze. Finally, we note that the use of a GRP is just one of many possible routes to generate intermediate molecular structures, but is employed here for compatibility with our graph-based scheme; alternative strategies, such as transformation into SMILES strings or other cheminformatics descriptors, are also possible, but not considered here.

2.3 New error function to account for permutational invariance

Despite the success of our mechanism-finding algorithm in several test applications,[31,38] two important algorithmic issues have also become evident. The first drawback relates to permutational invariance (Fig. 2). As outlined above, our double-ended GDS scheme for reaction-mechanism searching optimizes a graph error function F which quantifies the difference between a target CM (representing the desired products) and the CM generated after applying a series of n_r chemical reactions to an input reactant CM. A graph error function

of $F = 0$ implies that a reaction-path (*i.e.* sequence of elementary chemical reactions) which connect reactants and products has been successfully identified.

Now consider the role of permutational invariance. In Fig. 2(A), we show a schematic GDS-determined reaction-path which has $F = 0$; in other words, the sequence of reactions (defined by the reaction classes $\mathbf{k}$ and reactive indices $\mathbf{I}$) connects the input reactant structure on the left-hand side to the target product structure on the right-hand side. However, we note that, in the current form of our GDS algorithm, the input reactants and target products must exactly define the connectivity of *all* atoms; in other words, in Fig. 2(A), oxygen atoms 1 and 3 must end up in the same product CO_2 molecule, as must the oxygen atoms labelled 2 and 4. If, during the GDS optimization procedure, we generate the string of reactions shown in Fig. 2(B), leading to oxygen atoms 1 and 3 being in *different* CO_2 molecules, our current GDS algorithm would assign this reaction-path an error function value $F > 0$, and would continue the search for an alternative path with $F = 0$. This is obviously undesirable, given that the identities of the product molecules in each case are identical. As such, the requirement that one must currently define both the product molecular structure *and* the atomic indices of each product molecule is an important hurdle in general application of our GDS strategy. In addition, for cases when the mechanism under investigation is not known (*i.e.* the most interesting applications of GDS), this requirement will prevent the ultimate goal of automatic reaction discovery.

The second current drawback of our GDS algorithm to be addressed here relates to the target product structure. In our current implementation of GDS, as described above, we demand that the number of atoms in the reactant and product structures are exactly the same. However, this is again a significant drawback to fully-automated mechanism discovery because it demands that we define at the outset where every single atom resides in the product structure. In contrast, there are many examples of chemically-reactive systems where we would like to investigate all of the possible reaction mechanisms by which a large and diverse collection of molecular species could ultimately lead to formation of a *single* user-defined

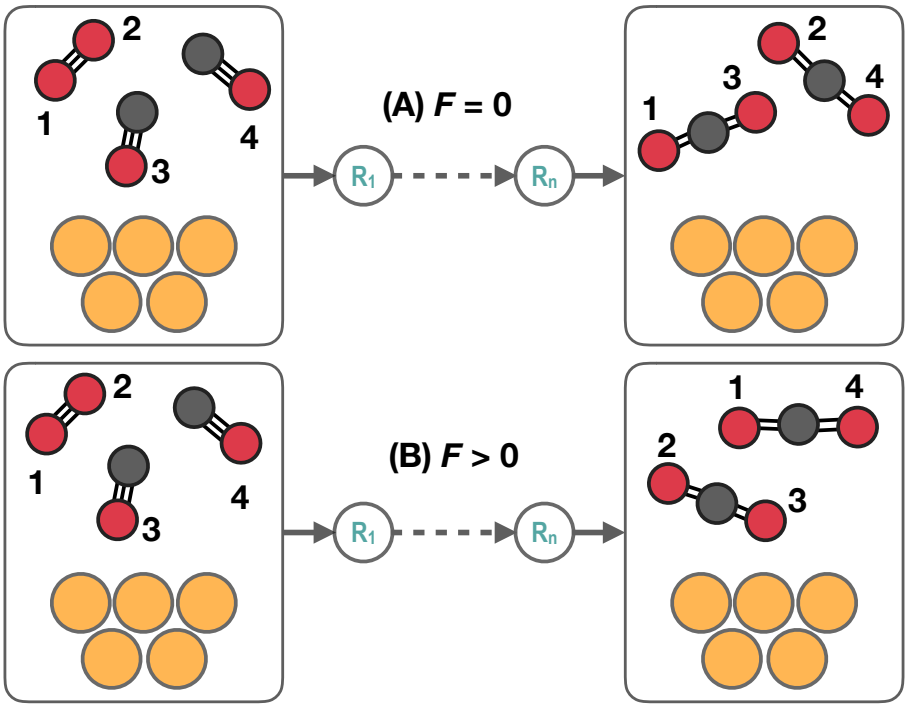


Figure 2: Ilustration of the role of permutational invariance in current GDS algorithm. In (A), we show the reactants (upper-left) and target products (upper-right) in a representative GDS calculation; in the target product, oxygen atoms 1 and 3 are required to be in the same CO_2 molecule in the target product. Any GDS-generated reaction-mechanism which generates such a structure will have a graph-error function $F = 0$. In (B), we show an alternative reaction-mechanism which could be generated during a GDS calculation, leading from the same set of reactant species (lower-left) to the same set of products (lower-right). However, in this case, the indices of the atoms in the product CO_2 molecules are not identical to those of the target product (upper-right); as a result, the standard atom-wise graph-error function F will be greater than zero, and our original GDS algorithm will fail to identify mechanism (B) as a successful reaction-mechanism leading to desired products.

molecular product, without regard for the identities of the remaining spectator species. This is illustrated in Fig. 3, which shows one of the reactive systems to be investigated later in this Article. In particular, as explained below, we would like to propose a large library of reaction mechanisms which lead to formation of benzene from a diverse set of reactant molecules; we do not want to absolutely define where every single atom from the reactant set must reside at the end of the reaction, but we do definitively want to form our target product (benzene, in this case). If our GDS scheme can be adapted to account for this challenge of product definition, it would dramatically expand the scope of our approach.

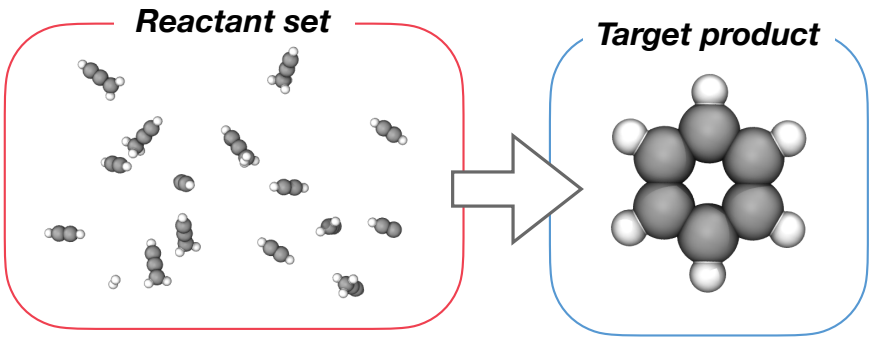


Figure 3: Representative problem set-up which is not currently addressable using GDS. Here, we wish to form the single target product molecule, C_6H_6, from a diverse set of hydrocarbon reactant species; in our current GDS algorithm,[31,38] we are forced to define the final location of *all* atoms in both reactant and product structures. However, in complex reactive systems where we are focussing on formation of a single target molecule, without regard for the remaining spectator molecules, our GDS algorithm needs to be modified, as explained in the text.

We can address both of these algorithmic issues by a change to the graph-error function F, as follows. First, after application of a sequence of n_r proposed reaction steps to the input reactant structure, we split the resulting product CM $\tilde{\mathbf{G}}$, which describes the *total* product structure, into the set of *molecular* CMs $\tilde{\mathbf{g}}^k$, where the index k identifies the molecular species (of which we assume there are M in the product structure resulting from the proposed n_r reaction steps). Each of these molecules CMs is defined as in Eq. 1, but only contains the atoms in each molecular species; identification of molecular CMs from the total product CM is straightforward using the Floyd-Warshall shortest-path algorithm.[58,59]

Second, for each molecular CM $\tilde{\mathbf{g}}^k$, we assign atomic masses to each vertex, yielding a new set of weighted CMs $\tilde{\mathbf{a}}^k$, with elements

$$\tilde{a}_{ij}^k = \begin{cases} \tilde{g}_{ij}^k & i \neq j, \\ m_i & i = j, \end{cases} \tag{6}$$

where m_i is the atomic mass of atom i. This modification of the CM is a common approach in graph-based cheminformatics studies, where incorporating vertex (*i.e.* atomic) features is often used to provide additional descriptor information.[60–62] Using the mass-weighted matrices $\tilde{\mathbf{a}}^k$, we then calculate the corresponding eigenvalues, $\tilde{\lambda}^k$, for each molecule k. As is well-known, the eigenvalues of a matrix are permutationally-invariant, and the addition of mass-weighting ensures that molecules with the same bonding pattern, but with different elements (*e.g.* O_2 and CO) will return different eigenvalues.

The same procedure as above, namely evaluation of the eigenvalues of the mass-weighted molecular CM, can also be applied to a *target* molecular structure which is desired as the reaction product; this target molecular structure does not necessarily need to contain the same number of atoms as the reactant structure, as suggested in Fig. 3. Identifying the number of atoms in the target product structure as n, and referring to the eigenvalues of the target molecular structure as $\mathbf{\Omega}$, we now define a new graph-error function as

$$F_p = \min_k \left[\delta(n - n^k) \sum_{i=1}^{n} (\lambda_i^k - \Omega_i)^2 + \left[1 - \delta(n - n^k)\right] \Delta \right]. \tag{7}$$

The function F_p returns the minimum value calculated over all M molecules identified in the proposed CM $\tilde{\mathbf{G}}$. The first term in the parentheses is only evaluated if molecule k has the same number of atoms n^k as the number of atoms in the target structure (n); in this case, we then evaluate the error between the two structures as the sum of the squared differences between the eigenvalues of the mass-weighted molecular CMs. However, if the number of atoms in molecule k is not the same as the number of atoms in the target structure, the first

term is ignored due to the action of the Dirac delta function, and the second term comes into operation; the parameter Δ is a penalty term which penalizes formation of structures which differ in the number of atoms from the target. After evaluating these two terms for all molecules in the product structure, the final error function is taken to be the minimum value amongst all of the M molecules considered.

The new graph-error function F_p of Eq. 7 has the following properties. First, it is invariant to permutation of atomic indices, which means that we do not need to explicitly define where each individual atom must find itself in the target product structure. Second, mass-weighting of the connectivity matrix ensures that molecules with the same bonding pattern but different elements compared to the target structure will be flagged with larger F_p values. Finally, the introduction of a penalty function based on number of atoms ensures that F_p can always return a numerical value, regardless of the number of atoms in the target and product CMs; this is an important aspect for the SA optimization approach adopted in our GDS scheme, which requires that a real-valued number be attached to any newly-proposed reaction mechanism.

To summarize, we have described our original double-ended GDS algorithm for reaction-mechanism finding, and we have subsequently shown how our scheme can be generalized to account for permutational invariance and formation of single target molecules from diverse sets of reactant species. We now demonstrate that this new scheme is a powerful strategy for exploring reaction mechanisms in complex chemical environments.

3 Application, results and discussion

As a challenging application of our computational mechanism-search scheme, as well as a demonstration of the improvements afforded by considering permutational invariance, we consider the formation of benzene (C_6H_6) in low-temperature ($\sim$10 K) environments in the ISM. This system is prime example of a complex chemical reaction environment; spectro-

scopic observations of low-temperature interstellar environments have detected vibrational signatures attributed to benzene, yet the large number of possible reactive species and possible chemical reaction mechanisms have led to some uncertainty in pinning down the operating mechanism(s) of benzene formation in such environments.[41,45,47–50,63,64]

In particular, our simulations will focus on investigating the formation of benzene from neutral molecular species commonly found in the ISM; the possible formation routes to benzene from neutral-molecule reactions has been investigated by a series of both experimental (*e.g.* molecular beams) and computational studies over the last couple of decades, leading to proposal of a number of different formation mechanisms.[41,47,49,50,64] Of particular importance is the benzene formation mechanism proposed by Kaiser and coworkers,[41] which posits that the initial reaction-step is a radical addition of C_2H to *trans*-1,3-butadiene; this initial addition is followed by ring-closure, hydrogen migration and hydrogen dissociation to form C_6H_6. Perhaps most importantly, *ab initio* calculations indicate that the proposed mechanism is energetically barrierless with respect to the reactants; this barrierless nature is a key requirement for any neutral-molecule reaction mechanism occurring at the very low temperatures of the relevant regions of the ISM. Beyond this barrierless radical mechanism described above, alternative schemes have also been proposed and investigated computationally, including addition of acetylene (C_2H_2) to diacetylene (C_4H_3),[45] and reaction of propargyl radicals (C_3H_3);[50,65] indeed, several of these proposals were originally based on kinetic models from combustion chemistry,[66–68] although the recent investigation of the importance of barrierless reaction routes has focussed attention on reactivity in these systems at low-temperatures.

Based on this previous body of work, we chose to investigate the formation of benzene in low-temperature environments, using mechanism-finding calculations starting from a varied set of reactant molecular species. We focus only on reactions involving neutral molecular species here; in particular our aim is to use our new mechanism-search scheme to build a large library of reaction-mechanisms which lead to *barrierless* formation of benzene in low-temperature environments comparable to those studied previously. Using *ab*

initio calculations of thermodynamic and kinetic properties, we will then investigate the plausibility of our mechanistic library for benzene formation, enabling us to explore which of the previously-proposed benzene formation routes are most likely, and if any alternative mechanisms (which might not have been assessed before) might also be plausible.

Electronic structure, TS location and free-energy calculations. Before proceeding to discuss the results of our mechanism-search simulations, we highlight the methods used to evaluate molecular energies, to find transition-states for selected reaction-steps, and to calculate molecular free energies. All electronic structure calculations used the *ORCA* quantum chemistry package.[69,70]

All calculations reported below used dispersion-corrected density functional theory (DFT) to perform geometry optimization and to calculate molecular energies. Specifically, we employed the revPBE exchange-correlation functional[71] with the Ahlrichs def2-SVP basis set;[72] dispersion corrections were calculated using the DFT-D3 scheme, along with the Becke-Johnson damping.[73,74]

Previous investigations of barrierless mechanisms leading to benzene in the ISM have used computationally-demanding electronic structure schemes, notably CCSD(T) or the G3//B3LYP composite scheme, to evaluate high-accuracy relative molecular energies.[41,49] In our case, we note that the large number of geometry optimization calculations and TS-finding calculations which must be performed to seek out barrierless reaction mechanisms in the set of reaction-mechanisms proposed by GDS means that a compromise between computational cost and accuracy has to be made; this led to our choice of the DFT approach noted above. Specifically, as detailed below, we perform ∼23,000 separate DFT geometry optimization calculations in order to evaluate molecular energies along the ∼2250 mechanisms proposed by our mechanism-finding scheme; this large number of calculations is not readily amenable to CCSD(T) or G3//BLYP methods, with DFT representing a good compromise.

However, it is of course important to benchmark our approach against previous calcu-

lations in order to validate our chosen computational scheme. Fig. 4 shows the reaction energies and barrier heights, predicted by our DFT calculations and by previous CCSD(T) and G3//B3LYP schemes, for a series of molecular structures along a barrierless mechanism to benzene which has been proposed previously[41,49] Very encouragingly, we find that our predicted energies (given relative to the energies of the reactant molecules) calculated at the revPBE/D3BJ DFT level are in very good qualitative agreement with these previous calculations. The root-mean-square difference in the DFT-calculated energies of the stationary points shown in Fig. 4, relative to the previous CCSD(T) and G3//B3LYP calculations, is around 16 kJ/mol (or 3.8 kcal/mol); furthermore, it is clear from Fig. 4 that the DFT calculations capture the qualitative trend in the relative energies of the series of intermediate structures along a key proposed reaction mechanism for this system. Bearing in mind the large number of geometry optimization calculations and TS-finding calculations required to screen the large set (> 2200) of multi-step mechanisms proposed by GDS, we conclude that revPBE/D3BJ DFT offers a good compromise between computational cost and accuracy.

TSs for selected reactions were located using climbing-image nudged-elastic band (CI-NEB[51]), followed by TS optimization using standard Hessian-based methods. Once located, the validity of TS structures were assessed by intrinsic reaction-coordinate (IRC) calculations to check that the desired reactant and product structures were obtained. In a few instances (as highlighted below), we found that the CI-NEB and TS optimization calculations seemingly converged, but the IRC calculations revealed pathways which did not lead to the desired product structure. In these few cases, we chose the pragmatic approach of approximating the reaction-barrier using information from the CI-NEB calculation; this is obviously not expected to be wholly accurate, but the large number of reactions generated by GDS necessitated this practical approach.

Finally, we note that free energies at 10 K were calculated for all geometry-optimized molecular species and TSs using the rigid rotor/harmonic oscillator approach,[75] with the Hessian calculated using the same level of DFT as noted above.

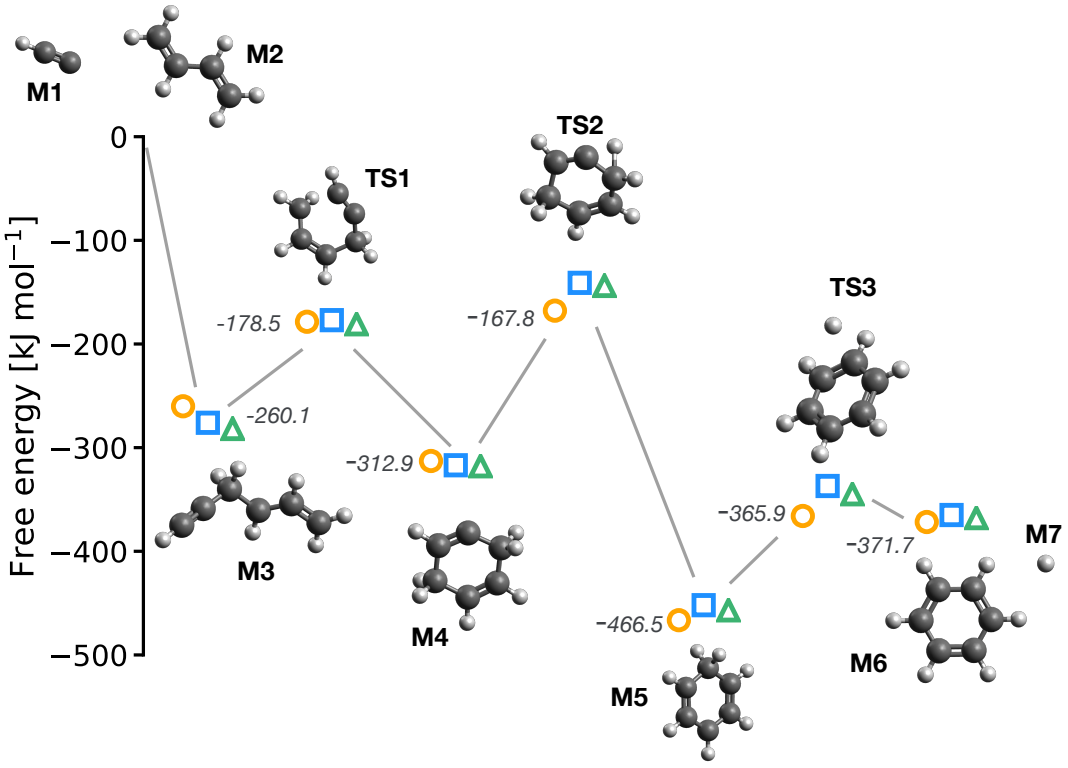


Figure 4: Barrierless reaction mechanism for formation of benzene from addition of C_2H to *trans*-1,3-butadiene. The values shown (in kJ mol^{-1}) are the free energies of molecular species relative to the reactants; the orange circle indicates the energies given by our revPBE DFT calculations, the blue squares are values from the G3//B3LYP composite method,[49] and the green triangles illustrate values from CCSD(T) calculations;[41] all free energy value are calculated at 298 K to enable comparison across methods.

3.1 Initial GDS results

We performed a total of 2250 GDS simulations in order to identify reaction-mechanisms leading to formation of benzene. After some initial tests, the maximum allowed number of reaction steps considered in our mechanism search calculations was $n_r = 8$. These simulations were performed for a mixture of different initial reactant species, as well as different total numbers of initial reactant molecules; to add further variability to the sampled reaction mechanisms, we also performed some of these GDS calculations using alternative sets of allowed chemical reactions, particularly removal of four-atom reaction events. The different initial system definitions and the standard move classes used in these simulations are given in the *Supplementary Information*. The sets of initial hydrocarbon species (*e.g.* C_2H, C_2H_2, C_3H_3, C_4H_3, C_3H_4, C_4H_6, C_4H_5) selected for the initial reactants were chosen to be representative of those hydrocarbons which are commonly discussed in the context of benzene formation in the ISM; our aim is to test whether our modified GDS algorithm can successfully determine known mechanisms for C_6H_6 formation, and whether any new reaction mechanisms can be identified.

Of the 2250 GDS simulations performed, we found that 2230 successfully located a reaction mechanism forming C_6H_6, giving a 99.2% success-rate of our GDS algorithm, and the average number of simulated annealing iterations required was 270×10^3. Depending on the number of iterations required, a typical calculation (including SA search plus DFT geometry optimization of all intermediate molecular species) took 1-2 hours on a single standard computing node; of course, each mechanism search is independent of the others, so many such calculations can be run simultaneously.

For the 2230 successful GDS calculations which formed benzene within the maximum number of allowed reaction steps n_r, we subsequently screened the reaction mechanisms using DFT calculations. Here, we performed DFT geometry opimization calculations for all of the molecular structures formed at each intermediate reaction-step for each proposed reaction-mechanism. In total, this required 24,632 DFT geometry optimizations to be performed.

Of these, we found that 1940 DFT calculations failed due to non-convergence of either self-consistent field iterations or geometry optimization; in such cases, we find that these convergence problems are typically a sign of distorted intermediate geometries generated along the proposed mechanisms, and we choose to remove the corresponding mechanisms from further consideration.

Following DFT calculations, we applied a series of further filtering steps to focus attention on mechanisms which meet validity criteria demanded for benzene formation in the ISM. First, we note that our goal is to seek out *barrierless* mechanisms for benzene formation, most relevant to formation in low-temperature ISM regions. As such, we subsequently remove from further consideration all of those reaction-mechanisms which exhibit intermediate molecular structures which are greater in energy than the reactant molecular species; this screening is based solely on the energies of optimized molecular geometries, rather than kinetic activation barriers, but provides a convenient first screening of candidate barrierless mechanisms. This energetic filtering leaves a total of 224 'thermodynamically barrierless' reaction mechanisms for further consideration.

Next, we compared all pairs of barrierless reaction mechanisms based on: (i) the number of reaction steps (excluding allowed 'null' reactions which do not progress the reaction), and (ii) the sum of the energies of reactant and product molecules at all steps in the proposed reaction mechanisms. In particular, two reaction-mechanisms were judged to be the same if they possessed the same number of (non-null) reaction-steps, and if the difference between the energies of the molecular structures at every reaction step was less than 5×10^{-3} E_h (or ~ 13 kJ mol^{-1}). We find that these criteria mean that reactions which involve the same molecular species but different conformations are kept in the set of unique reaction-mechanisms. The final post-processing approach applied to the remaining reaction mechanisms was simple visual inspection; this allowed elimination of any remaining mechanisms which were deemed to be closely related to others.

As shown below, the simple screening of reaction mechanisms based on comparison of the

energetics of reaction intermediates allows us to identify sets of unique reaction mechanisms; however, we note that alternatives to this procedure could also be used (and might be preferable for even larger reaction networks). In particular, direct comparison of complete reaction mechanisms based on evaluation of structural molecular fingerprints of all reactive intermediates could be employed, and will be explored in the near-future.

The outcome of the filtering procedures is a final set of 126 reaction mechanisms which are flagged as being unique and barrierless (at least from the point of view of the energies of the intermediate molecular structure - kinetic aspects are discussed further below). Figure 5 gives an overview of the energies of the different molecular species at each reaction step as a simple schematic way of illustrating the diversity in reaction mechanisms generated. Here, the radii of the different circles indicate the frequency with which molecules of each particular energy occur at each reaction-step; we note that a reaction number of zero corresponds to the initial set of reactant molecules. Figure 5(a) shows that the initial reactant sets comprise molecular species with between two and four carbon atoms, as well as molecular hydrogen. As the reaction steps progress, we see gradual decreases in the number of C_2-C_4 species, as demonstrated by the shrinking radii of the circles with increasing reaction-steps. The decrease in the frequency of these species is matched by the increase in the number of C_6 species and, to a lesser extent, C_5 and C_7 species. Furthermore, we note that none of the benzene-forming mechanisms generated here form any intermediates with more than seven carbon atoms; this emphasises the fact that our mechanism search scheme is not free to combinatorially generate molecular species indiscriminately, but is instead constrained by the search for a well-defined product species. Figure 5(b) shows the same distribution of molecular species, but only for those reactions judged to be unique. The pattern of behaviour is similar to that shown in Fig. 5(a), with C_2-C_4 species rapidly being converted into C_6 species. The final observation here is that C_6 species are typically formed after just one reaction step, which is possible because of the initial distribution molecular species in the C_2-C_4 range; of course, starting with alternative reactant sets with, for example, only C_1-C_2

species, would require more reaction-steps to generate more complex C_6 species.

3.2 Barrierless mechanisms to benzene formation

The filtering of reaction mechanisms noted above, followed by visual filtering, demonstrated that a varied series of different reaction mechanisms leading to benzene formation were successfully identified. Ultimately, the combination of clustering based on molecular energy characteristics and visual analysis, as well as exploratory NEB calculations used to investigate potentially high-energy reaction barriers, resulted in a set of ~ 12 reaction mechanisms which led to benzene formation and: (i) were predicted to be thermodynamically barrierless (based solely on the molecular energies at different reaction-steps), (ii) did not proceed through chemically-unusual reaction intermediates, and (iii) did not proceed by chemically-unusual reaction mechanisms.

In what follows, we focus our attention on describing the 'best' candidate reaction mechanisms for barrierless formation of benzene in the ISM. In particular, we focus our attention on seven candidate mechanisms, and describe the results of further mechanistic analysis based on NEB calculations, TS optimization and free-energy calculations; we also note below where we identify reaction-mechanisms which have been reported previously, but did not warrant further detailed investigation due to the availability of previous calculations. All reported calculations were performed using the same DFT scheme as described above for initial evaluation of molecular energies.

3.2.1 RM-1: Addition of C_2H to *trans*-1,3-butadiene

The first reaction mechanism of note determined in our GDS simulations, referred to hereafter as **RM-1**, is exactly the mechanism which was previously proposed by Jones *et al.*[41] The mechanism, shown in Fig. 4, begins with barrierless addition of the radical C_2H to *trans*-1,3-butadiene (C_4H_6), followed by closure of the six-membered ring. Following a further intramolecular hydrogen transfer step and dissociation of hydrogen from the cyclic C_6H_7

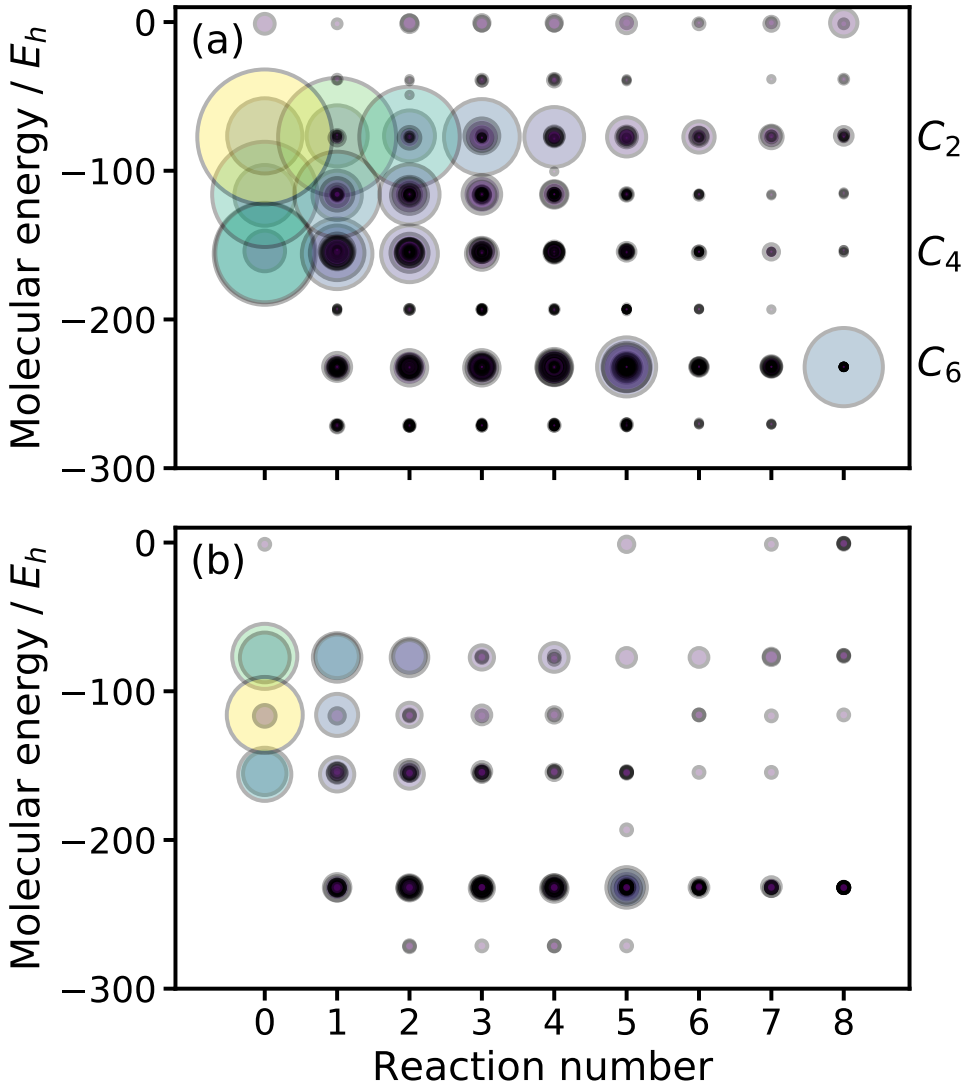


Figure 5: Frequencies of molecular energies as a function of reaction-step in GDS simulations; the size and color of each circle represent the frequency of observation, and molecules which have a calculated energy (at optimized geometry) which differ by less than $10^{-3} E_h$ (or 2.6 kJ mol^{-1}) were clustered together for this analysis. Panel (a) shows the molecular energy distribution calculated for all successful GDS calculations, and panel (b) shows the molecular energy distribution only for the 126 reaction mechanisms judged to be unique after filtering.

species, the final benzene product is formed. The largest free-energy barrier in this reaction is 145 kJ mol^{-1} for the third reaction-step, involving hydrogen transfer from a CH_2 site to an adjacent 'bare' carbon atom in the C_6H_7 ring (see Fig. 4). Importantly, as noted previously, the entire reaction mechanism leading to formation of benzene is barrierless with respect to the total energy of the reactants; as such, it would be expected that **RM-1** could proceed in the low-temperature environment of the ISM.

3.2.2 RM-2: Ring closure following hydrogen transfer

A first alternative reaction mechanism is illustrated in Fig. 6. Here, the initial addition of C_2H to C_4H_6 proceeds in the same way as in **RM-1**, forming product **M3** through a barrierless addition reaction. However, rather than direct ring closure as in **RM-1**, **RM-2** proceeds through a hydrogen-shift reaction to form **M8**, which subsequently undergoes ring-closure to form the benzene product and a hydrogen atom.

While the initial addition of C_2H is barrierless, and the barrier to hydrogen shift (159.9 kJ mol^{-1}) is comparable to the greatest barrier in **RM-1**, this reaction mechanism can be ruled out as feasible due to the very high-barrier to concerted ring closure and hydrogen dissociation. This barrier is estimated to be around 626 kJ mol^{-1}.

In an attempt to study this reaction further, we sought to locate a TS for the ring-closure reaction starting from **M8**, but this time leading to **M6** without concerted hydrogen dissociation. However, for this ring-closure, the TS located after NEB and TS-optimization did not correspond to the targeted ring-closure reaction. Instead, the TS corresponded to formation of an intermediate structure containing a four-membered ring, with a barrier of around 96 kJ mol^{-1}. Starting from this four-membered ring intermediate, we then sought a TS connecting to the benzene product **M6**, although this proved unsuccessful, leading again to a false TS which connected to **M8**. In summary, the challenge of locating the TS for ring-closure without concerted hydrogen dissociation suggested that this mechanism was not a competitor of **RM-1**.

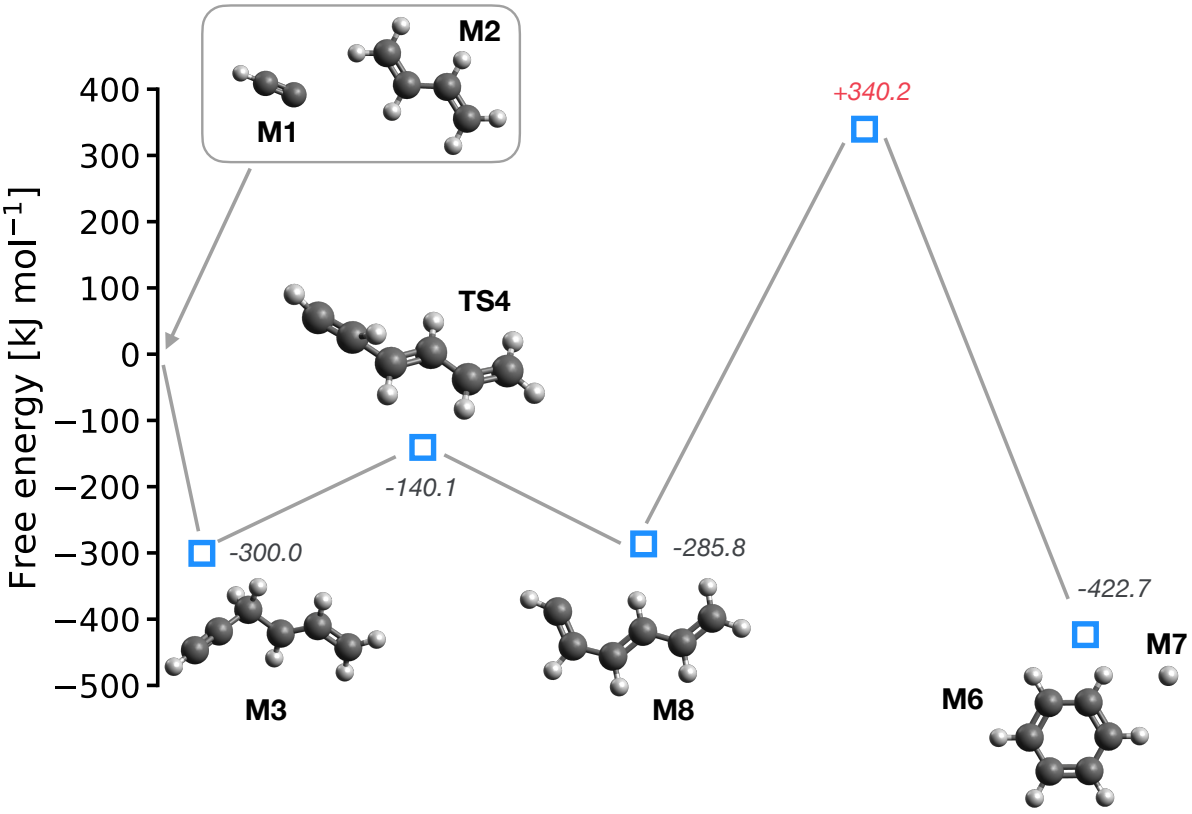


Figure 6: Summary of mechanism **RM-2**. Reactant species are shown in the box in the upper-left corner, and free energies (in kJ mol^{-1}, calculated at 10 K) are shown relative to the reactants. The final barrier (in red) is an estimate for the illustrated reaction, given that the TS for this reaction proved difficult to locate (as discussed in the text).

3.2.3 RM-3: Addition of C_2H to C_4H_5

The next reaction of interest selected for study (**RM-3**) is shown in Fig. 7. As in **RM-1** and **RM-2**, the initial step of this reaction is barrierless addition of C_2H, but this time to the C_4H_5 species (**M9**). The resulting C_6H_6 species **M10** then proceeds to form a six-membered ring; two subsequent hydrogen-shifts, proceeding through **TS11** and **TS12**, then lead to formation of the benzene product. As shown in Fig. 7, this entire reaction path is barrierless relative to the reactant species, and the maximum reaction barrier is about 120 kJ mol^{-1}, which is lower than the maximum barrier in **RM-1**.

However, mechanism **RM-3** exhibits an important feature which likely rules it out as a further candidate for consideration as a new reaction mechanism for benzene formation in the ISM. In particular, the ring-closure reaction leading from **M10** to **M11** is found to be endothermic and barrierless in the reverse direction; as such, one might expect that intermediate **M11** would not have sufficient lifetime to act as a route towards benzene, especially given the fact that the forward reaction from **M11** to **M12** exhibits a barrier of $\sim$120 kJ mol^{-1}.

However, it is interesting to note that **RM-3** does, at least to the best of our knowledge, represent a new barrierless mechanism for benzene formation; finding such mechanisms was, after all, the main focus of our algorithm and this paper. Further analysis of the energetic and kinetic characteristics unfortunately suggests that the likely impact of this mechanism in ISM benzene formation is negligible.

3.2.4 RM-4 and RM-5: Small barriers to initial addition of C_2H_2

Two further mechanisms of interest which were located by our GDS algorithm (**RM-4** and **RM-5**) are shown in Fig. 8. Both of these reactions proceed through initial addition of acetlyene (C_2H_2) to radical species; in the case of **RM-4**, the initial reaction is addition of C_2H_2 to *iso*-C_4H_3 whereas, in the case of **RM-5**, the reaction proceeds through addition of C_2H_2 to C_4H_5. Both reactions form C_6H_7 intermediate species which subsequently cyclise

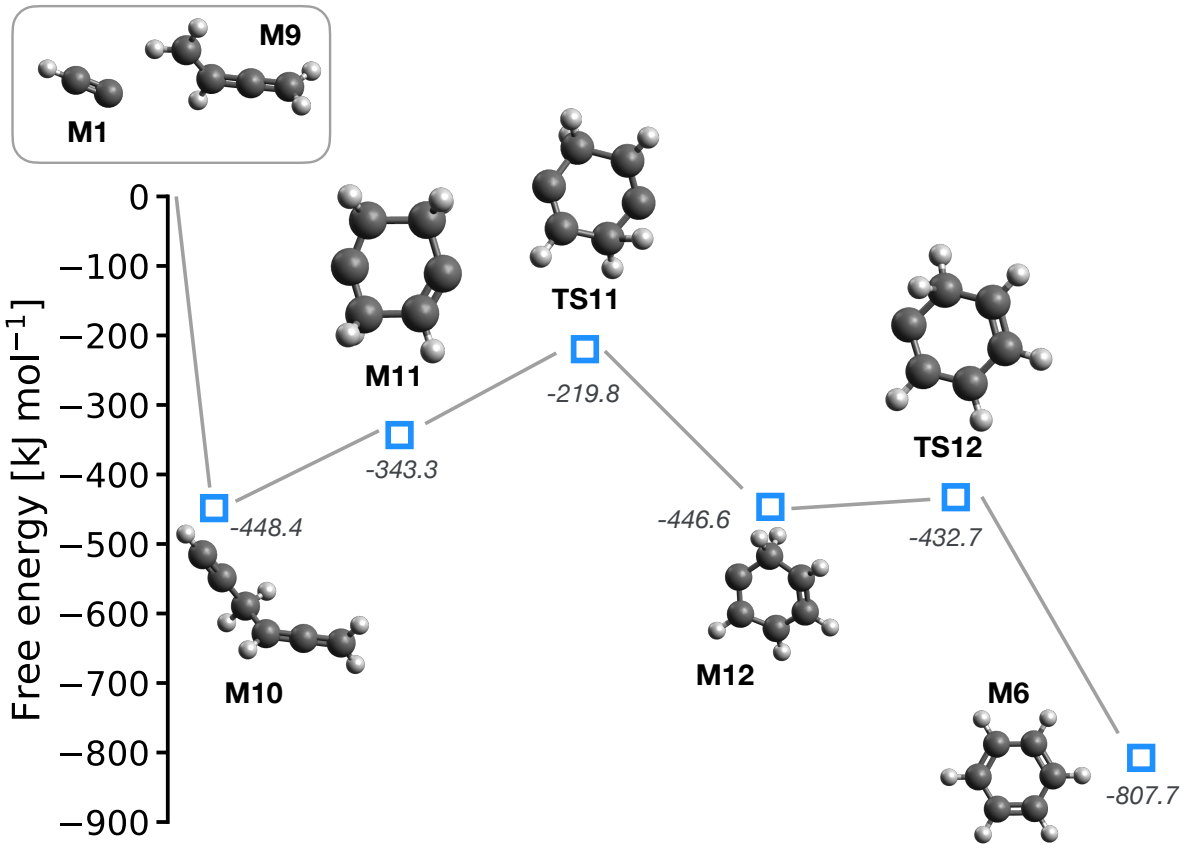


Figure 7: Summary of mechanism **RM-3**. Reactant species are shown in the box in the upper-left corner, and free energies (in kJ mol^{-1}, calculated at 10 K) are shown relative to the reactants.

to form six-membered rings, although the two reactions then differ in the details of the later stages of benzene formation. Specifically, **RM-4** involves hydrogen addition, followed by a hydrogen-shift reaction (moving a hydrogen atom from a CH_2 moiety in the six-membered ring to an adjacent carbon atom with no bound hydrogen atoms); in our GDS calculation, the hydrogen addition reaction involved a single given hydrocarbon species which was present in the reactant set, but we note that the identity of the donor species is somewhat arbitrary (in the sense that it could effectively be any available species). In contrast, **RM-5** proceeds to benzene through a hydrogen-shift reaction, forming species **M5** as in **RM-1**, which then dissociates atomic hydrogen to form benzene.

The initial screening calculations performed for these mechanisms, where the energies of the intermediate molecular species were evaluated using DFT calculations, suggested that both of these reactions were 'thermodynamically barrierless' in the sense that no intermediate species has an energy which is greater than the reactant species. However, further NEB calculations indicate an important feature in both reactions; the initial addition of C_2H_2 is found to have a small barrier to reaction which is 19.7 kJ mol^{-1} for **RM-4** and 30.6 kJ mol^{-1} for **RM-5**. Under thermal conditions at room-temperature or higher, these barriers to reaction would not be to be the rate-limiting steps in these reaction mechanisms for benzene formation; however, in the low-temperature environment of the ISM, even these low barriers present a significant factor counting against both mechanisms as routes for benzene formation. At a representative ISM temperature of 10 K, the available thermal energy for each degree-of-freedom is around 0.08 kJ mol^{-1}, suggesting that the normal thermal reaction rate associated with even these low association barriers would be very slow indeed.

As a result of the detection of small energetic barriers to C_2H_2 association in both **RM-4** and **RM-5**, we chose not to pursue further analysis of the later reaction steps, instead ruling out the feasibility of these mechanisms for benzene formation in the ISM based on the initial barriers alone. However, it is worth noting that these reactions appear to be related to the recent proposal of resonance-stabilized radical reactions which are implicated in the

formation of soot pre-cursors;[76] in the setting of high-temperature hydrocarbon combustion, the small initial barriers to reaction for mechanisms **RM-4** and **RM-5** would be much less important, suggesting that these routes might play a part in benzene formation in flames (assuming that the reactive C_4H_3 and C_4H_5 species are present). The application of our GDS approach to study combustion processes, obviously related to the processes under study here, is a clear avenue for further investigation.

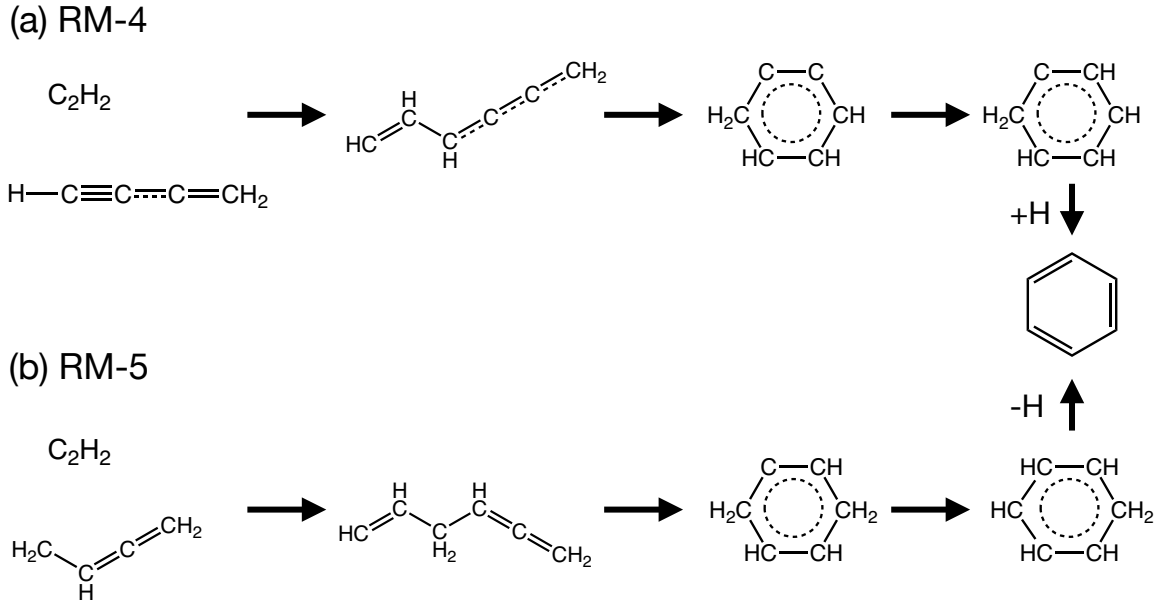


Figure 8: (a) Summary of mechanism **RM-4**, starting from reaction of C_2H_2 with C_4H_3. (b) Summary of mechanism **RM-5**, starting from reaction of C_2H_2 with C_4H_5. Both mechanisms exhibit small energetic barriers (20 - 30 kJ mol^{-1}) to the initial addition step.

3.2.5 RM-6 and RM-7: Addition of C_2H to C_2H_2

The final reaction mechanism which is detailed here from the set of GDS-determined results is **RM-6**, as shown in Fig. 9. This reaction proceeds with the barrierless addition of C_2H to acetylene, in a similar manner to the initial addition observed in **RM-1**. Subsequently, a second C_2H_2 molecule adds onto the intermediate C_4H_3 to form **M17**, which subsequently undergoes ring-closure and addition of hydrogen (from any available reactant species) to

form the benzene product.

While the initial addition of C_2H_2 is barrierless, and the addition of further C_2H_2 is associated with only a small barrier of around 7 kJ mol^{-1}, we found that the later ring-closing reaction proved problematic in seeking to find a reliable TS; as a result, we instead approximate the barrier to this reaction as being around 220 kJ mol^{-1} according to the results of our CINEB calculation. This barrier is larger than the largest barrier encountered in **RM-1**, but it is notable that the entire reaction mechanism **RM-6** is barrierless as expected, and so appears to be a valid alternative to **RM-1** (assuming, of course, that a TS connecting **M17** and **M18** could eventually be determined).

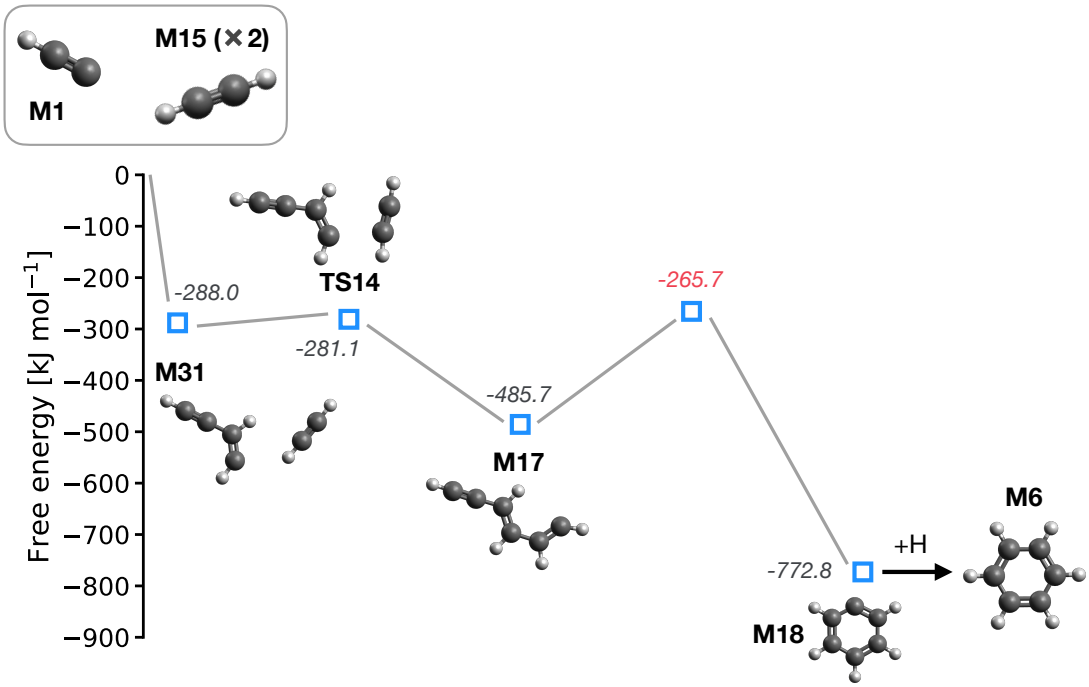


Figure 9: Reaction mechanism **RM-6**. Reactants are shown in the upper-left corner, and energies are given relative to the reactants in kJ mol^{-1}. The value given in red is a CINEB estimate of the reaction barrier shown; the corresponding TS approximation from CINEB did not converge to the desired TS, as detected by subsequent IRC calculation.

Further investigation of **RM-6** gives further insight into the reason for the high barrier to the ring-closure reaction. In particular, initial addition of C_2H to C_2H_2 in **RM-6** leads

(in this case) to formation of *Z*-*n*-C_4H_3, with addition of a further C_2H_2 forming a *trans*-double-bonded intermediate (**M17**). The geometry of this intermediate means that the ring-closure reaction leading to **M18** demands significant intramolecular distortion, resulting in the relatively high observed energetic barrier.

This insight leads to proposal of a further reaction mechanism which proceeds in the same way as **RM-6**, but *via* different *conformers* of the reactive species along the pathway. In particular, as shown in Fig. 10, C_2H can in principle add to C_4H_3 in different orientations to form a different intermediate C_6H_5 structure, namely **M29**. The ring-closure reaction from **M29** to **M18** proceeds straightforwardly through **TS8**, and addition of hydrogen through reaction with an available species can subsequently form the benzene product **M6**. We note that reaction-mechanism **RM-7** contains the same set of reactions as **RM-6**, but consideration of different available isomers leads to a different pathway.

Compared to both **RM-6** and **RM-1**, mechanism **RM-7** certainly presents as a plausible route to benzene formation in the ISM. It is barrierless, just as mechanism **RM-1** is; in addition, the barriers to the different elementary reaction steps which form **RM-7** are much lower than those encountered in **RM-1**, suggesting that **RM-7** would be expected to proceed at a faster overall rate. As such, **RM-7** presents itself as a good candidate for a barrierless route to benzene in the ISM, complementary to the previously-suggested **RM-1**.

However, as a final point, it is worth noting another factor which will impact the ultimate feasibility of this reaction pathway in the ISM (as well as in other settings such as combustion). In particular, the *n*-C_4H_3 species which is required to react with acetylene to form species **M29** is known to undergo a 1,2-hydrogen-shift reaction to form the *iso*-C_4H_3 species. As noted previously,[45,47,64] the *iso*-C_4H_3 is thermodynamically more stable than the *n*-C_4H_3 isomers by around 42 kJ mol^{-1}, and the barrier to isomerization from *n*-C_4H_3 to *iso*-C_4H_3 is 222 kJ mol^{-1}. As a result of this isomerization channel, reaction **RM-7** might generally be expected to be prevented from forming benzene in the ISM. However, as above, it is also interesting to note that **RM-7** could potentially operate in higher-temperature combustion

settings, especially where the relative concentration of the n-C_4H_3 can be maintained.

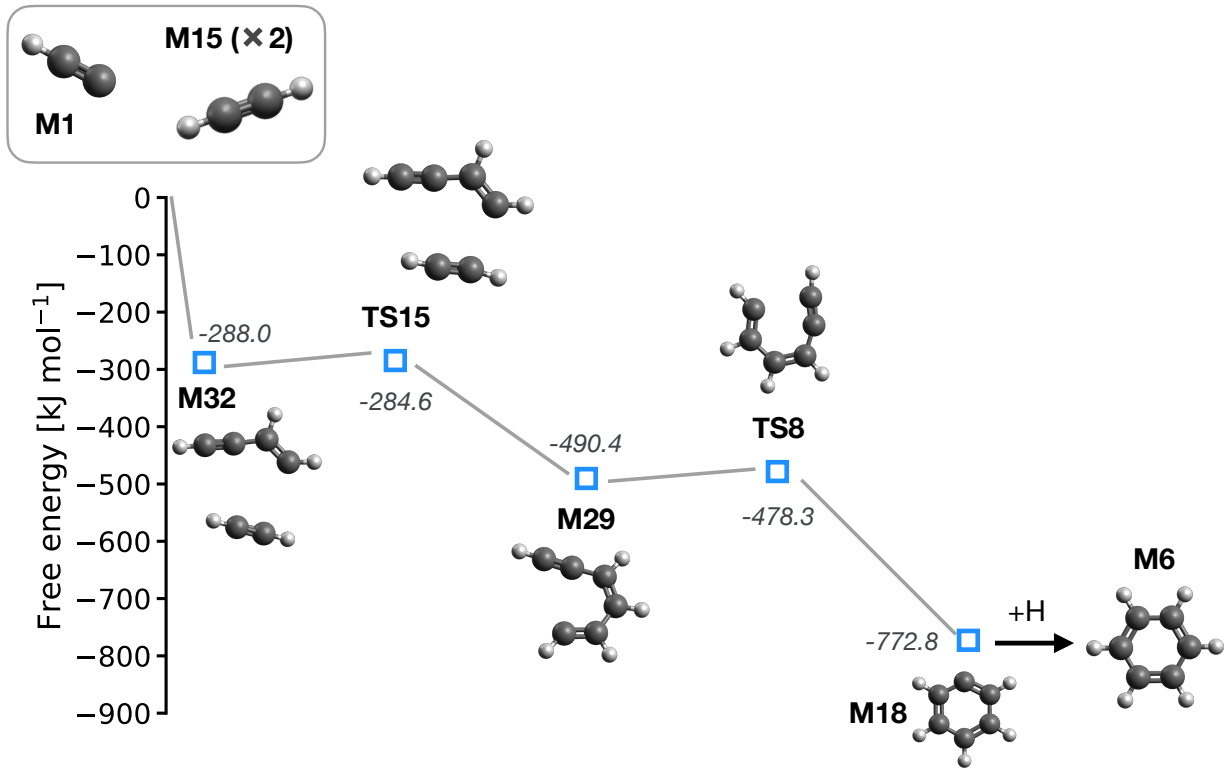


Figure 10: Reaction mechanism **RM-7**. Reactants are shown in the upper-left corner, and energies are given relative to the reactants in kJ mol^{-1}.

3.3 Further comments

The analysis above has focussed attention on seven reaction mechanisms which, according to initial DFT screening, presented themselves as viable candidates for barrierless formation of benzene in the ISM. We have shown that one of these reactions corresponds to the previously-proposed mechanism by Jones and co-workers, we have identified further reactions proceeding *via* reaction of C_2H and C_2H_2, and we have also observed reactions of C_2H_2 with C_4 species.

As noted previously, however, we initially located 126 reaction mechanisms which were flagged as potentially barrierless, with no intermediate structures having energy greater than

the corresponding reactant species. Further inspection of many of these reactions reveal interesting chemical routes to benzene, albeit routes that can usually be excluded after visualization or comparison to the more standard mechanisms noted above. In this section, we simply note some of these other reaction mechanisms as a route to better understanding possible improvements of our GDS algorithm.

Perhaps most importantly in this context is the finding that the majority of the 126 unique reaction mechanisms can be ruled out based on the over-emphasis of concerted reactions. For example, we find a number of proposed reaction mechanisms which involve *simultaneous* addition and dissociation reaction-steps, or related steps requiring insertion of molecules into bonds which would otherwise be expected to be non-reactive. Such reactions are difficult to screen out on the basis of the DFT energies of the intermediate structures alone; after all, such concerted reactions can often lead to reaction products which are exothermic relative to reactants, but the same such reactions would usually be ruled out as viable on the basis of the large activation energies which would generally be expected of concerted reactions for such systems. Instead, an additional screening based on predicted activation energies for the elementary reaction-steps in a proposed mechanism would prove to be very useful in removing such reactions from further consideration, thereby dramatically improving the efficiency of mechanism search and reducing the need to visually check reaction-mechanisms before choosing to proceed with further analysis. Another way to ameliorate the impact of such mechanisms is to remove them entirely by placing further constraints on the atoms which can participate in each reaction-class. At present, both screening based on predicted activation energies and valence-based constraints on reactive atoms are areas of focus in our research.

In discussing **RM-6** above, we have also touched upon another important challenge in our current GDS approach. In particular, our approach focusses on finding sequences of elementary reaction-steps which lead to product formation, but we have seen that different competing mechanisms (*i.e.* **RM-7**) can be determined which have the same CM updates but

differ in their reactive conformers. In such cases, we anticipate that sampling over molecular conformers using fast empirical force-fields might be a viable future route to automatically generating conformational alternatives.

As a final point, it is worth noting that the self-reactions of propargyl radicals C_3H_3, was also detected in our set of GDS simulations. Here, we identified several propargyl-based reaction mechanisms which formed benzene through different relative approach orientations of the C_3H_3, namely:

$$C_2H_2{-}CH{\cdots}CH{-}C_2H_2 \longrightarrow C_6H_6 \tag{8}$$

$$CH{-}C_2H_2{\cdots}CH{-}C_2H_2 \longrightarrow C_6H_6 \tag{9}$$

$$CH{-}C_2H_2{\cdots}C_2H_2{-}CH \longrightarrow C_6H_6. \tag{10}$$

These reactions have been studied in previous *ab initio* electronic structure calculations,[41,50,65] leading to the conclusion that this set of reactions are not important pathways to benzene in the ISM due to the relative instability of the intermediate species formed by addition of propargyl radicals. As such, we chose not to pursue further electronic structure calculations for these reaction mechanisms, but note that it is pleasing that these previously-proposed mechanisms were automatically proposed by our double-ended GDS algorithm.

4 Conclusions

In this Article, we have shown how our recently-developed double-ended reaction-mechanism search algorithm can be adapted to: (i) define target reaction products which are a sub-set of the total set of reactive molecules, and (ii) account for atomic permutational invariance in the target product structure. These important modifications to our earlier algorithm mean that the resulting approach is much more generally-applicable; one can define a generic set of input reactant molecules and a single target molecule, and then generate an arbitrary number of reaction-mechanisms connecting these two end-points. By accounting for permutational

invariance in the cost-function which is optimized during the our graph-based search, there is no need to define where every atom in the reactant set must reside in the target product; the resulting approach is therefore much more useful in the context of automated reaction discovery.

As a test case for our approach, we generated over 2200 reaction mechanisms which lead from several generic sets of hydrocarbon molecules (containing C_2-C_4 species) to a target benzene product. By calculating energies of intermediate structures, and subsequently clustering reaction mechanisms based on these energies, we find 126 unique reaction mechanisms which form benzene without any of the intermediate structures having higher energy than the reactants. Further visual inspection of this unique set, as well as targeted MEP-finding calculations, were subsequently used to identify a set of around seven reaction-mechanisms which were studied in more detail as candidate barrierless mechanisms for C_6H_6 formation.

Encouragingly, we find that these barrierless mechanisms are a mixture of both previously-suggested mechanisms and mechanisms which represent new routes to C_6H_6 formation. In particular, we identified the reaction initiated by addition of C_2H to *trans*-1,3-butadiene as the most likely barrierless mechanism (**RM-1**), but also note that the reaction between C_2H and C_2H_2 (**RM-7**) also presents itself as a favourable reaction for interstellar benzene formation. As noted above, the issue of isomerization to more stable intermediates like n-C_4H_3 is not accounted for in our direct search approach, and deserves further account. In addition, we have also observed formation of benzene through the well-known propargyl recombination routes, and have also identified further barrierless routes which can, ultimately, be excluded from further consideration due to the presence of significant activation energies. However, these reactions (as well as many of the others we have discounted in our search for barrierless reactions) might be of further interest in the context of hydrocarbon combustion processes, where elevated temperature might favour these alternative reaction routes.

The calculations reported here also serve as a useful guide to further areas of development. The first relates to computational expense. In total, we estimate that around 24,000 DFT

geometry optimizations were performed, in addition to 100-200 NEB optimizations of MEPs. This is clearly a large computational effort, stemming predominantly from the large number of reactions and intermediate structures which are generated in our approach. To address this problem, we need much more efficient methods of *approximating* the characteristics of individual elementary chemical reactions, such as activation energy and reaction energy; machine-learning tools may find a use here. A second challenge relates to MEP- and TS-finding; these are, in general, much more challenging to automate than geometry optimization to local minima, with poor starting MEPs or reactant/product configurations often leading to slow convergence (or, often, no convergence at all). These challenges have been considered extensively in the last decade or so, with the development of methods such as the growing-string approach, offering routes to further automating the reaction discovery process.[11,77–79]

A further remaining algorithmic challenge in our approach is the fact that different mechanisms possessing identical sets of reaction-steps, but arranged in a different sequence, can in principle be generated. Although we anticipate that the imposition of atomic valence constraints will likely have a role in ruling out some of these alternative mechanisms, this cannot be universally guaranteed. As such, identifying these sequence-differing reaction mechanisms and sampling over the different possible permutations is a further topic which needs to be addressed; we leave this detail for future work.

The final area worth further thought relates to completeness of generated reaction-mechanisms: in other words, how many reaction mechanisms do we need to generate to ensure that all relevant mechanisms have been sampled? Although the total number of reactions and molecular species which can be generated for an arbitrary set of reactant molecules is combinatorially large, this number can be reduced to (hopefully) manageable levels by restricting the total number of allowed reaction-steps in any given mechanism (as we do in our approach). As such, one might expect that careful databasing of reactions and intermediate structures, as well as accurate comparison using efficient molecular fingerprints, could enable complete construction of reaction networks. This challenge, as well as those noted above,

are clearly goals for the near future.

Acknowledgments

The authors gratefully acknowledge the support of the Engineering and Physical Sciences Research Council (EPSRC) through award EP/R020477/1. We also acknowledge the Scientific Computing Research Technology Platform at the University of Warwick for providing high-performance computing resources.

Data availability

Molecular structures in mechanisms **RM-1** to **RM-7** are available at wrap.warwick.ac.uk/147082.

References

(1) Gaggioli, C. A.; Stoneburner, S. J.; Cramer, C. J.; Gagliardi, L. Beyond Density Functional Theory: The Multiconfigurational Approach To Model Heterogeneous Catalysis. *ACS Catal.* **2019**, *9*, 8481–8502.

(2) Ghosh, S.; Verma, P.; Cramer, C. J.; Gagliardi, L.; Truhlar, D. G. Combining Wave Function Methods with Density Functional Theory for Excited States. *Chem. Rev.* **2018**, *118*, 7249–7292.

(3) Grimme, S.; Hansen, A.; Brandenburg, J. G.; Bannwarth, C. Dispersion-Corrected Mean-Field Electronic Structure Methods. *Chem. Rev.* **2016**, *116*, 5105–5154.

(4) Kong, L.; Bischoff, F. A.; Valeev, E. F. Explicitly Correlated R12/F12 Methods for Electronic Structure. *Chem. Rev.* **2012**, *112*, 75–107.

(5) Bannwarth, C.; Caldeweyher, E.; Ehlert, S.; Hansen, A.; Pracht, P.; Seibert, J.; Spicher, S.; Grimme, S. Extended tight-binding quantum chemistry methods. *WIRES Comput. Mol. Sci.* e01493.

(6) Govind, N.; Petersen, M.; Fitzgerald, G.; King-Smith, D.; Andzelm, J. A generalized synchronous transit method for transition state location. *Comp. Mat. Sci.* **2003**, *28*, 250 – 258.

(7) Peters, B.; Heyden, A.; Bell, A. T.; Chakraborty, A. A growing string method for determining transition states: comparison to the nudged elastic band and string methods. *J. Chem. Phys.* **2004**, *120*, 7877–7886.

(8) Maeda, S.; Morokuma, K. A systematic method for locating transition structures of A+B $\rightarrow$ X type reactions. *J. Chem. Phys.* **2010**, *132*, 241102.

(9) Martínez-Núñez, E. An automated method to find transition states using chemical dynamics simulations. *J. Comput. Chem.* **2015**, *36*, 222 – 234.

(10) Koslover, E. F.; Wales, D. J. Comparison of double-ended transition state search methods. *J. Chem. Phys.* **2007**, *127*, 134102.

(11) Zimmerman, P. M. Reliable transition state searches integrated with the growing string method. *J. Chem. Theory Comput.* **2013**, *9*, 3043–3050.

(12) Ren, W.; Vanden-Eijnden, E.; Maragakis, P.; E, W. Transition pathways in complex systems: Application of the finite-temperature string method to the alanine dipeptide. *J. Chem. Phys.* **2005**, *123*, 134109.

(13) Rodríguez, A.; Rodríguez-Fernández, R.; A. Vázquez, S.; L. Barnes, G.; J. P. Stewart, J.; Martínez-Núñez, E. tsscds2018: A code for automated discovery of chemical reaction mechanisms and solving the kinetics. *J. Comput. Chem.* **2018**, *39*, 1922–1930.

(14) Koistinen, O.-P.; Dagbjartsdóttir, F. B.; Ásgeirsson, V.; Vehtari, A.; Jónsson, H. Nudged elastic band calculations accelerated with Gaussian process regression. *J. Chem. Phys.* **2017**, *147*, 152720.

(15) Coley, C. W.; Jin, W.; Rogers, L.; Jamison, T. F.; Jaakkola, T. S.; Green, W. H.; Barzilay, R.; Jensen, K. F. A graph-convolutional neural network model for the prediction of chemical reactivity. *Chem. Sci.* **2019**, *10*, 370–377.

(16) Kandathil, S. M.; Fletcher, T. L.; Yuan, Y.; Knowles, J.; Popelier, P. L. A. Accuracy and tractability of a kriging model of intramolecular polarizable multipolar electrostatics and its application to histidine. *J. Comput. Chem.* **2013**, *34*, 1850–1861.

(17) Rupp, M.; Tkatchenko, A.; Müller, K.-R.; von Lilienfeld, O. A. Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning. *Phys. Rev. Lett.* **2012**, *108*, 058301.

(18) Snyder, J. C.; Rupp, M.; Hansen, K.; Müller, K.-R.; Burke, K. Finding Density Functionals with Machine Learning. *Phys. Rev. Lett.* **2012**, *108*, 253002.

(19) Bartók, A. P.; Csányi, G. Gaussian approximation potentials: A brief tutorial introduction. *Int. J. Quantum Chem.* **2015**, *115*, 1051–1057.

(20) Bartók, A. P.; Gillan, M. J.; Manby, F. R.; Csányi, G. Machine-learning approach for one- and two-body corrections to density functional theory: Applications to molecular and condensed water. *Phys. Rev. B* **2013**, *88*, 054104.

(21) Pozun, Z. D.; Hansen, K.; Sheppard, D.; Rupp, M.; Müller, K.-R.; Henkelman, G. Optimizing transition states via kernel-based machine learning. *J. Chem. Phys.* **2012**, *136*.

(22) Vu, K.; Snyder, J. C.; Li, L.; Rupp, M.; Chen, B. F.; Khelif, T.; Müller, K.-R.; Burke, K. Understanding kernel ridge regression: Common behaviors from simple functions to density functionals. *Int. J. Quantum Chem.* **2015**, *115*, 1115–1128.

(23) Nguyen, T. T.; Székely, E.; Imbalzano, G.; Behler, J.; Csányi, G.; Ceriotti, M.; Götz, A. W.; Paesani, F. Comparison of permutationally invariant polynomials, neural networks, and Gaussian approximation potentials in representing water interactions through many-body expansions. *J. Chem. Phys.* **2018**, *148*, 241725.

(24) Welborn, M.; Cheng, L.; Miller, T. F. Transferability in Machine Learning for Electronic Structure via the Molecular Orbital Basis. *J. Chem. Theory Comput.* **2018**, *14*, 4772–4779.

(25) Schreck, J. S.; Coley, C. W.; Bishop, K. J. M. Learning Retrosynthetic Planning through Simulated Experience. *ACS Cent. Sci.* **2019**, *5*, 970–981.

(26) Schwaller, P.; Laino, T.; Gaudin, T.; Bolgar, P.; Hunter, C. A.; Bekas, C.; Lee, A. A. Molecular Transformer: A Model for Uncertainty-Calibrated Chemical Reaction Prediction. *ACS Cent. Sci.* **2019**, *5*, 1572–1583.

(27) Varela, J. A.; Vázquez, S. A.; Martínez-Núñez, E. An automated method to find reaction mechanisms and solve the kinetics in organometallic catalysis. *Chem. Sci.* **2017**, *8*, 3843–3851.

(28) Wang, L.-P.; Titov, A.; McGibbon, R.; Liu, F.; Pande, V. S.; Martínez, T. J. Discovering chemistry with an ab initio nanoreactor. *Nature Chem.* **2014**, *6*, 1044–8.

(29) Kopec, S.; Martínez-Núñez, E.; Soto, J.; Peláez, D. vdW-TSSCDS—An automated and global procedure for the computation of stationary points on intermolecular potential energy surfaces. *Int. J. Quantum Chem.* **2019**, *119*, e26008.

(30) Habershon, S. Automated Prediction of Catalytic Mechanism and Rate Law Using Graph-Based Reaction Path Sampling. *J. Chem. Theory Comput.* **2016**, *12*, 1786–1798.

(31) Ismail, I.; Stuttaford-Fowler, H. B. V. A.; Ochan Ashok, C.; Robertson, C.; Habershon, S. Automatic Proposal of Multistep Reaction Mechanisms using a Graph-Driven Search. *J. Phys. Chem. A* **2019**, *123*, 3407–3417.

(32) Kim, Y.; Choi, S.; Kim, W. Y. Efficient Basin-Hopping Sampling of Reaction Intermediates through Molecular Fragmentation and Graph Theory. *J. Chem. Theory Comput.* **2014**, *10*, 2419–2426.

(33) Kim, Y.; Kim, J. W.; Kim, Z.; Kim, W. Y. Efficient prediction of reaction paths through molecular graph and reaction network analysis. *Chem. Sci.* **2018**, *9*, 825–835.

(34) Gao, C. W.; Allen, J. W.; Green, W. H.; West, R. H. Reaction Mechanism Generator: Automatic construction of chemical kinetic mechanisms. *Comp. Phys. Commun.* **2016**, *203*, 212 – 225.

(35) Simm, G. N.; Reiher, M. Context-Driven Exploration of Complex Chemical Reaction Networks. *J. Chem. Theory Comput.* **2017**, *13*, 6108–6119.

(36) Simm, G. N.; Vaucher, A. C.; Reiher, M. Exploration of Reaction Pathways and Chemical Transformation Networks. *J. Phys. Chem. A* **2019**, *123*, 385–399.

(37) Habershon, S. Sampling reactive pathways with random walks in chemical space: Applications to molecular dissociation and catalysis. *J. Chem. Phys.* **2015**, *143*, 094106.

(38) Robertson, C.; Habershon, S. Fast screening of homogeneous catalysis mechanisms using graph-driven searches and approximate quantum chemistry. *Catal. Sci. Technol.* **2019**, *9*, 6357–6369.

(39) Robertson, C.; Ismail, I.; Habershon, S. Traversing Dense Networks of Elementary Chemical Reactions to Predict Minimum-Energy Reaction Mechanisms. *ChemSystemsChem* **2020**, *2*, e1900047.

(40) Bohme, D. K. PAH [polycyclic aromatic hydrocarbons] and fullerene ions and ion/molecule reactions in interstellar and circumstellar chemistry. *Chem. Rev.* **1992**, *92*, 1487–1508.

(41) Jones, B. M.; Zhang, F.; Kaiser, R. I.; Jamal, A.; Mebel, A. M.; Cordiner, M. A.; Charnley, S. B. Formation of benzene in the interstellar medium. *Proc. Nat. Acad. Sci. USA* **2010**, *108*, 452–457.

(42) Tielens, A. Interstellar Polycyclic Aromatic Hydrocarbon Molecules. *Annu. Rev. Astron. Astrophys.* **2008**, *46*, 289–337.

(43) Parker, D. S. N.; Zhang, F.; Kim, Y. S.; Kaiser, R. I.; Landera, A.; Kislov, V. V.; Mebel, A. M.; Tielens, A. G. G. M. Low temperature formation of naphthalene and its role in the synthesis of PAHs (Polycyclic Aromatic Hydrocarbons) in the interstellar medium. *Proc. Nat. Acad. Sci. USA* **2012**, *109*, 53–58.

(44) Keyte, I. J.; Harrison, R. M.; Lammel, G. Chemical reactivity and long-range transport potential of polycyclic aromatic hydrocarbons – a review. *Chem. Soc. Rev.* **2013**, *42*, 9333.

(45) Lories, X.; Vandooren, J.; Peeters, D. Cycle formation from acetylene addition on C4H3 radicals. *Phys. Chem. Chem. Phys.* **2010**, *12*, 3762–3771.

(46) Kaiser, R. I.; Parker, D. S. N.; Mebel, A. M. Reaction dynamics in astrochemistry: low-temperature pathways to polycyclic aromatic hydrocarbons in the interstellar medium. *Annu. Rev. Phys. Chem.* **2015**, *66*, 43–67.

(47) Walch, S. P. Characterization of the minimum energy paths for the ring closure reactions of C_4H_3 with acetylene. *J. Chem. Phys.* **1995**, *103*, 8544–8547.

(48) Woods, P. M.; Millar, T. J.; Zijlstra, A. A.; Herbst, E. The Synthesis of Benzene in the Proto–planetary Nebula CRL 618. *Astrophys. J.* **2002**, *574*, L167–L170.

(49) Lee, K. L. K.; McGuire, B. A.; McCarthy, M. C. Gas-Phase Synthetic Pathways to Benzene and Benzonitrile: A Combined Microwave and Thermochemical Investigation. *Phys. Chem. Chem. Phys.* **2019**, *21*, 2946–2956.

(50) Wilson, E.; Atreya, S.; Coustenis, A. Mechanisms for the formation of benzene in the atmosphere of Titan. *J. Geophys. Res.* **2003**, *108*.

(51) Henkelman, G.; Uberuaga, B. P.; Jónsson, H. A climbing image nudged elastic band method for finding saddle points and minimum energy paths. *J. Chem. Phys.* **2000**, *113*, 9901.

(52) Smidstrup, S.; Pedersen, A.; Stokbro, K.; Jónsson, H. Improved initial guess for minimum energy path calculations. *J. Chem. Phys.* **2014**, *140*, 214106.

(53) Henkelman, G.; Jónsson, H. Improved tangent estimate in the nudged elastic band method for finding minimum energy paths and saddle points. *J. Chem. Phys.* **2000**, *113*, 9978.

(54) Truhlar, D. G.; Garrett, B. C.; Klippenstein, S. J. Current Status of Transition-State Theory. *J. Phys. Chem.* **1996**, *100*, 12771–12800.

(55) Laidler, K. J.; King, M. C. Development of transition-state theory. *J. Phys. Chem.* **1983**, *87*, 2657–2664.

(56) Henriksen, N. E.; Hansen, F. Y. *Theories of Molecular Reaction Dynamics: The Microscopic Foundation of Chemical Kinetics*; Oxford University Press, 2011.

(57) Laidler, K. J. *Chemical Kinetics*, 3rd ed.; Harper Collins: New York, 1987.

(58) Floyd, R. W. Algorithm 97 shortest path. *Comm. ACM* **1962**, *5*, 345.

(59) Warshall, S. A theorem on Boolean matrices. *J. ACM* **1962**, *9*, 11.

(60) Todeschini, R.; Consonni, V. *Handbook of molecular descriptors*; John Wiley & Sons, 2008; Vol. 11.

(61) Janezic, D.; Milicevic, A.; Nikolic, S.; Trinajstic, N. *Graph-theoretical matrices in chemistry*; CRC Press, 2015.

(62) David, L.; Thakkar, A.; Mercado, R.; Engkvist, O. Molecular representations in AI-driven drug discovery: a review and practical guide. *J. Cheminformatics* **2020**, *12*, 56.

(63) Cernicharo, J.; Heras, A. M.; Tielens, A. G. G. M.; Pardo, J. R.; Herpin, F.; Guélin, M.; Waters, L. B. F. M. Infrared Space Observatory Discovery of C_4H_2, C_6H_2, and Benzene in CRL 618. *Astrophys. J.* **2001**, *546*, L123–L126.

(64) Le, T. N.; Mebel, A. M.; Kaiser, R. I. Ab initio study of C_4H_3 potential energy surface and reaction of ground-state carbon atom with propargyl radical. *J. Comp. Chem.* **2001**, *22*, 1522–1535.

(65) Miller, J. A.; Melius, C. F. Kinetic and thermodynamic issues in the formation of aromatic compounds in flames of aliphatic fuels. *Combust. Flame* **1992**, *91*, 21 – 39.

(66) P. R. Westmoreland, J. B. H., A. M. Dean; Longwell, J. P. Forming benzene in flames by chemically activated isomerization. *J. Phys. Chem.* **1989**, *93*, 8171–8180.

(67) Richter, H.; Howard, J. B. Formation and consumption of single-ring aromatic hydrocarbons and their precursors in premixed acetylene, ethylene and benzene flames. *Phys. Chem. Chem. Phys.* **2002**, *4*, 2038–2055.

(68) Pope, C. J.; Miller, J. A. Exploring old and new benzene formation pathways in low-pressure premixed flames of aliphatic fuels. *Proc. Combust. Inst.* **2000**, *28*, 1519 – 1527.

(69) Neese, F. The ORCA program system. *WIRES Comput. Mol. Sci.* **2012**, *2*, 73–78.

(70) Neese, F. Software update: the ORCA program system, version 4.0. *WIRES Comput. Mol. Sci.* **2018**, *8*, e1327.

(71) Zhang, Y.; Yang, W. Comment on "Generalized gradient approximation made simple". *Phys. Rev. Lett.* **1998**, *80*, 890.

(72) Weigend, F.; Ahlrichs, R. Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for H to Rn: Design and assessment of accuracy. *Phys. Chem. Chem. Phys.* **2005**, *7*, 3297–3305.

(73) Grimme, S.; Antony, J.; Ehrlich, S.; Krieg, H. A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu. *J. Chem. Phys.* **2010**, *132*, 154104.

(74) Grimme, S.; Ehrlich, S.; Goerigk, L. Effect of the damping function in dispersion corrected density functional theory. *J. Comp. Chem.* **2011**, *32*, 1456–1465.

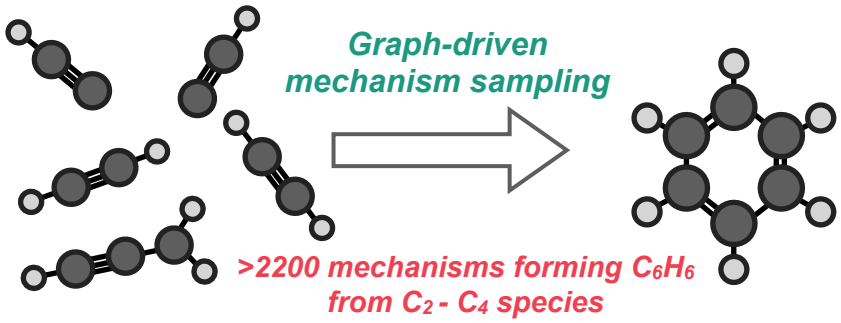
(75) Grimme, S. Supramolecular Binding Thermodynamics by Dispersion-Corrected Density Functional Theory. *Chem. Eur. J.* **2012**, *18*, 9955–9964.

(76) Johansson, K. O.; Head-Gordon, M. P.; Schrader, P. E.; Wilson, K. R.; Michelsen, H. A. Resonance-stabilized hydrocarbon-radical chain reactions may explain soot inception and growth. *Science* **2018**, *361*, 997–1000.

(77) Dewyer, A. L.; Argüelles, A. J.; Zimmerman, P. M. Methods for exploring reaction space in molecular systems. *WIRES Comput. Mol. Sci.* **2018**, *8*, e1354.

(78) Nett, A. J.; Zhao, W.; Zimmerman, P. M.; Montgomery, J. Highly Active Nickel Catalysts for C-H Functionalization Identified through Analysis of Off-Cycle Intermediates. *J. Am. Chem. Soc.* **2015**, *137*, 7636–9.

(79) Zimmerman, P. M. Automated discovery of chemically reasonable elementary reaction steps. *J. Comput. Chem.* **2013**, *34*, 1385–1392.



For Table of Contents Only.