# UC Davis

## UC Davis Electronic Theses and Dissertations

### Title

First-Principles Discovery of Reaction Mechanisms and High-Performance Computational Tools for Minimum Energy Pathway Tracing

### Permalink

https://escholarship.org/uc/item/1c43k4b8

### ISBN

9798193266426

### Author

Park, Heejune

### Publication Date

2026-09-08

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

First-Principles Discovery of Reaction Mechanisms and High-Performance Computational Tools for Minimum Energy Pathway Tracing

By

HEEJUNE PARK

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Chemistry

in the

OFFICE OF GRADUATE STUDIES

of the

UNIVERSITY OF CALIFORNIA

DAVIS

Approved:

_____

Lee-Ping Wang, Chair

_____

Davide Donadio

_____

Kyle Crabtree

Committee in Charge

2026

# Contents


Abstract v

Acknowledgments vi

Chapter 1. Introduction 1

1.1. Quantum Mechanics 1

1.1.1. Schrödinger's Equation 2

1.1.2. Hartree-Fock Method 5

1.1.2.1. Self-Consistent Field Method 7

1.1.2.2. Post Hartree-Fock Methods 9

1.1.3. Density Functional Theory 11

1.1.3.1. The Hohenberg-Kohn Theorems 12

1.1.3.2. The Kohn-Sham Theory 13

1.2. Ab Initio Nanoreactor Molecular Dynamics Simulation 17

1.3. Automated Reaction Pathway Refinement Workflow 20

1.3.1. Energy Minimization 21

1.3.1.1. Coordinate System 23

1.3.2. Minimum Energy Pathway Search 25

1.3.2.1. The Intrinsic Reaction Coordinate Method 26

1.4. Scope and Overview of the Research 28

Chapter 2. Application: N-heterocycle formation pathway discovery 30

2.1. Introduction 30

2.2. Methods 33
2.3. Results and Discussion 38
2.3.1. Nanoreactor Simulations 38
2.3.1.1. Pyridine 38
2.3.1.2. Pyrrole 41
2.3.2. Pathway Refinement and Additional Potential Energy Surface Exploration 44
2.3.2.1. Azete + Cyanomethyl radical 44
2.3.2.2. Ethenimine + $\beta$-cyanovinyl radical 47
2.3.2.3. 1-Pyrrolyl + Cyanomethylene diradical 48
2.4. Conclusion 50
2.5. Supporting Information 52

Chapter 3. Translation-rotation internal coordinates with splicing (TRICS): A new approach to molecular geometry interpolation 53
3.1. Introduction 53
3.2. Methods 57
3.2.1. TRICS Interpolation Method 57
3.2.1.1. Algorithmic Details 62
3.2.2. Computational Details 64
3.2.3. Interpolation Endpoints 65
3.3. Results 66
3.3.1. Benchmark Results 66
3.3.1.1. Interpolation Success Rates 66
3.3.1.2. Analysis of excess energy above the barrier 69
3.3.1.3. Detailed comparison of TRICS, ZM, and IDPP 71
3.3.2. Applications in characterizing barrierless reactions and excited state potential energy surfaces 77

3.4. Discussion 81
3.5. Conclusion 82
3.6. Supporting Information 83
3.6.1. Initial Reaction Pathways 83
3.6.2. Diagnostic for Well-definedness of Internal Coordinate Systems 85
3.6.3. Negative Excess Energy Cases 86

Chapter 4. High-Throughput Approach for Minimum Energy Pathway Search Using the Nudged Elastic Band Method with Efficient Data Handling and Parallel Computing 89
4.1. Introduction 89
4.2. Methods 94
4.2.1. Reactions 94
4.2.1.1. RGD1 reactions 96
4.2.1.2. BH reactions 97
4.2.1.3. Selected reactions 98
4.2.2. Nudged Elastic Band Method 101
4.2.3. GeomeTRIC and QCArchive Project 105
4.3. Results and discussion 108
4.3.1. RGD1 dataset reactions 108
4.3.2. BH dataset reactions 115
4.3.3. Selected reactions 116
4.4. Conclusion 118
4.5. Supplementary Information 119

Chapter 5. Summary and Future Work 120

Bibliography 123

## Abstract

The exploration of potential energy surfaces (PESs) is essential for understanding chemical reaction mechanisms, yet the high dimensionality and complexity of PESs make locating minimum-energy pathways (MEPs) challenging. This thesis develops computational tools to explore PESs and discover reaction pathways. Chapter 1 outlines the theoretical foundations of electronic structure methods, from Hartree–Fock and post-Hartree–Fock approaches to density functional theory, and how the energies from these methods are used to evaluate interatomic forces to run ab initio molecular dynamics (AIMD) simulations; it also describes how AIMD trajectories are refined to characterize reaction mechanisms. Chapter 2 uses a retrosynthetic AIMD simulation nanoreactor approach to identify two previously unknown low-barrier pathways for pyrrole formation and one barrierless pathway for pyridine formation, offering new insights into the prebiotic synthesis of nitrogen-containing heterocycles. Chapter 3 introduces a novel interpolation method, TRICS (translation–rotation internal coordinates with splicing), to generate high-quality MEP approximations by employing multiple internal coordinate systems and iterative splicing; the method performs comparably or better than existing approaches across 679 reactions from three benchmark datasets. Chapter 4 develops a high-throughput nudged elastic band (NEB) framework on cloud infrastructure, enabling efficient parallel calculations and systematic evaluation of protocol parameters. This implementation achieves high convergence rates across diverse reactions and successfully reproduces results for transition-metal-catalyzed systems. Together, these tools advance the automated exploration and characterization of chemical reaction pathways on complex PESs.

## Acknowledgments

I thank my Ph.D. advisor, Professor Lee-Ping Wang, for his guidance, patience, and generosity. I learned a great deal from you—not only science, but also how to be a good scientist. You never failed to appreciate my work while gently nudging me to progress further. I greatly value all of your advice, which will serve me well throughout my future career.

I also thank Professor Chenchen Song for supporting me during a summer session when I implemented the IRC method in geomeTRIC. In addition, I learned a great amount of quantum chemistry while serving as a TA for your courses.

I thank all of my QE committee members—Professor Alexei Stuchebrukhov, Professor C. William McCurdy, Professor Kyle Crabtree, and Professor Ambar Kulkarni—for their patience and guidance throughout the exam process. I further thank Professor Davide Donadio and Professor Kyle Crabtree for reviewing this thesis and providing suggestions that improved it.

I want to thank Dr. Benjamin Pritchard from MolSSI for guiding me through the software aspects of my major projects. Without your help, I would not have been able to complete my main project.

I am grateful to the university and to the United States as a whole for welcoming a foreigner like myself to pursue a Ph.D. degree tuition-free while providing a stipend. I still find it hard to believe that I am completing this degree while receiving support.

I am thankful for my colleagues, especially my group members: Zhecheng, Nathan, Jesi, Lisa, Hyesu, Marshall, and Nanhao. I genuinely enjoyed interacting with all of you and appreciate the help you offered when I first joined the group. I also thank my cohort member

Adam for struggling through coursework and problem sets together. I thank Zekun for his kindness and hospitality toward me. Finally, I extend my gratitude to Sommer Johansen, who helped me with QE preparation and generously allowed me to join her nanoreactor project.

I thank my wife, Carol Thao, for her support throughout my Ph.D. journey. You were always there for me and asked how my work was going—not because you were interested in the details of my research, but because you cared about how I was doing. Thank you for being an amazing partner; I look forward to overcoming many challenges and enjoying many blessings together. I also thank my two daughters, Chloe K. Park and Anna M. Park, for entering this world as my children. You two are the greatest achievement I have accomplished so far. I thank my parents and my younger brother for supporting me throughout the Ph.D. program, and I thank Patrick and Beth, as well as Tony and Grace, for their patience and generosity toward my family.

Most importantly, I thank God, who created this world and provided the physical laws that I can study and by which I am continually awed. I also thank Jesus for undergoing the most agonizing and humiliating death in my place. You are the Lord who deserves all praise.

# Chapter 1

# Introduction

In computational chemistry, accurate prediction of molecular behavior, such as whether a reaction will occur, how a protein will fold, or how atoms will vibrate and move, is essential for understanding chemical processes. Central to these predictions is the relationship between molecular structures (i.e., nuclear configurations) and their energies, which defines a landscape known as the potential energy surface (PES). The PES is the foundational concept governing molecular structure and dynamics in quantum mechanics. Computing even small portions of the PES is often computationally intensive and time-consuming, frequently requiring high-performance resources; yet the resulting insights are invaluable. Ultimately, accurate predictions depend on a precise characterization of the PES landscape at a reasonable cost.

In this chapter, we investigate how computational chemistry, grounded in first-principles calculations, can predict and confirm molecular reaction pathways. First, we discuss the use of quantum mechanics to compute molecular energies and gradients (forces). Next, the ab initio nanoreactor molecular dynamics (MD) simulation is introduced, along with its ability to induce molecular collisions and accelerate reaction discovery. Finally, an automated workflow for refining reaction pathways, capable of discovering chemical reactions from MD simulation trajectories, is presented.

## 1.1. Quantum Mechanics

The wave-particle duality opened a new domain of study called quantum mechanics. Human intuition, shaped by classical experience, struggles to reconcile how an object can behave as both a particle and a wave. In classical mechanics, assuming that the objects of interest are particles yields accurate enough results. However, in the much smaller microscopic world, this assumption fails as they exhibit more wave-like behavior. A new paradigm was required

to properly describe the quantum world. In 1926, the physicist Erwin Schrödinger published his work "Quantisierung als Eigenwertproblem", [1] demonstrating that the energy of wave-like matter can be solved via an eigenvalue problem. This represented a breakthrough, as Schrödinger's wave mechanics offered a more intuitive wave-based alternative to the earlier, abstract matrix mechanics developed by Heisenberg, Born, and Jordan. [2,3] The mathematical equivalence of the two approaches was demonstrated shortly thereafter. [4,5] First, we will explore the equation discovered by Schrödinger and various approaches to its approximate solutions.

### 1.1.1. Schrödinger's Equation

In the infinite-dimensional Hilbert space, mathematical functions describing matter waves—known as wavefunctions, $\psi(\mathbf{r})$, where $\mathbf{r}$ denotes position—that retain their functional form when an operator called the Hamiltonian $\hat{H}$ acts upon them are called eigenfunctions, and the corresponding scaling factors (eigenvalues) represent the possible energy levels ($E$). The time-independent Schrödinger equation is thus expressed as:

$$\hat{H}\,|\psi_k(\mathbf{r})\rangle = E_k\,|\psi_k(\mathbf{r})\rangle \tag{1.1}$$

where $k$ indexes the different stationary states. The Hamiltonian $\hat{H}$ describes the total energy of a quantum system and can be written as:

$$\hat{H} = \hat{T}_n + \hat{T}_e + \hat{V}_{nn} + \hat{V}_{ne} + \hat{V}_{ee} \tag{1.2}$$

The first two terms are kinetic energies, with the subscripts $n$ and $e$ indicating nuclei and electrons, respectively. The last three terms capture interactions between nuclei ($nn$), between nuclei and electrons ($ne$), and between electrons ($ee$). The mass of a proton or neutron is approximately 1836 times greater than that of an electron. Since an atomic nucleus typically contains multiple protons and neutrons, its mass is at least several thousand times greater than that of an electron. This large mass difference enables the decoupling of

nuclear and electronic motions, allowing the electrons to be treated quantum mechanically while the nuclear positions are held fixed.

This approach is known as the Born-Oppenheimer (BO) approximation. [6] Within this approximation, the nuclear kinetic energy operator $\hat{T}_n$ is neglected (effectively set to zero), while the nuclear-nuclear repulsion term $\hat{V}_{nn}$ becomes a constant scalar (dependent on the fixed nuclear geometry). The resulting decoupled non-relativistic electronic Hamiltonian ($\hat{H}_e$) can be expressed in atomic unit ($\hbar = e = m_e = 4\pi\epsilon_0 = 1$) as

$$\hat{H}_e = \hat{T}_e + \hat{V}_{ne} + \hat{V}_{ee} + V_{nn} = -\frac{1}{2}\sum_i \nabla_i^2 - \sum_{i,A} \frac{Z_A}{r_{iA}} + \sum_{i<j} \frac{1}{r_{ij}} + \sum_{A<B} \frac{Z_A Z_B}{R_{AB}} \tag{1.3}$$

where $i$ and $j$ run over electrons and $A$ and $B$ iterate through nuclei. $r_{ij} = |\mathbf{r}_i - \mathbf{r}_j|$ is the distance between electrons $i$ and $j$, $r_{iA} = |\mathbf{r}_i - \mathbf{R}_A|$ is the distance between electron $i$ and nucleus $A$, and $R_{AB} = |\mathbf{R}_A - \mathbf{R}_B|$ is the fixed distance between nuclei $A$ and $B$; and $Z_A$ denotes the charge of nucleus $A$.

One important physical property that the electronic Hamiltonian in Eq. (1.3) fails to account for is the intrinsic angular momentum of electrons, known as spin ($s_i$). [7] The Stern-Gerlach experiment in 1922 demonstrated this by showing that a beam of silver atoms splits into two discrete components in an inhomogeneous magnetic field. [8] All particles with non-zero spin (including nuclei) exhibit magnetic moments, which introduce additional terms into Eq. (1.3), but these effects will be neglected in this discussion. To incorporate spin, we introduce two functions, $\alpha(\omega)$ and $\beta(\omega)$, corresponding to spin up and spin down, respectively, where $\omega$ is an unspecified spin variable. The wavefunction in Eq. (1.1) can then be expressed as $\psi(\mathbf{r}, \omega)$. Using the four degrees of freedom (three spatial coordinates and one spin coordinate), the wavefunction for an $N$-electron system can be written as $\Psi(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N)$, where $\mathbf{x}_i = \{\mathbf{r}_i, \omega_i\}$. Since electrons are fermions, the wavefunction must be antisymmetric under the exchange of any two electrons, i.e., interchanging the coordinates of any two electrons changes the sign of the wavefunction. [7]

$$\Psi(\mathbf{x}_1 ... \mathbf{x}_i ... \mathbf{x}_j ... \mathbf{x}_N) = -\Psi(\mathbf{x}_1 ... \mathbf{x}_j ... \mathbf{x}_i ... \mathbf{x}_N) \tag{1.4}$$

This antisymmetry property automatically enforces the Pauli exclusion principle, which states that no two electrons can occupy the same spin-orbital (i.e., the same set of quantum numbers). [7, 9] In particular, if two electrons have the same spatial and spin coordinates ($\mathbf{x}_i = \mathbf{x}_j$), the wavefunction must vanish.

$$\Psi(\mathbf{x}_1 ... \mathbf{x}_i ... \mathbf{x}_i ... \mathbf{x}_N) = -\Psi(\mathbf{x}_1 ... \mathbf{x}_i ... \mathbf{x}_i ... \mathbf{x}_N) = 0 \tag{1.5}$$

The required antisymmetry of the wavefunction under particle exchange also guarantees their indistinguishability. Specifically, exchanging two electrons flips the sign of $\Psi$, the probability density $|\Psi|^2$ remains invariant, ensuring that all physical observables are unaffected by the permutation. [7]

An antisymmetric many-electron wavefunction for an $N$-electron system can be constructed from a set of $N$ one-electron functions (spin-orbitals) using a Slater determinant, which can be written as

$$\Psi(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N) = \frac{1}{\sqrt{N!}} \begin{vmatrix} \chi_1(\mathbf{x}_1) & \chi_1(\mathbf{x}_2) & \cdots & \chi_1(\mathbf{x}_N) \\ \chi_2(\mathbf{x}_1) & \chi_2(\mathbf{x}_2) & \cdots & \chi_2(\mathbf{x}_N) \\ \vdots & \vdots & \ddots & \vdots \\ \chi_N(\mathbf{x}_1) & \chi_N(\mathbf{x}_2) & \cdots & \chi_N(\mathbf{x}_N) \end{vmatrix} \tag{1.6}$$

where $\{\chi_k(\mathbf{x}_i)\}_{k=1}^{N}$ are the occupied spin-orbitals and the factor $1/\sqrt{N!}$ ensures normalization when the spin-orbitals are orthonormal. [7] One can verify that this form is antisymmetric: interchanging any two electron coordinates (i.e., swapping two columns) changes the sign of the determinant, while assigning the same spin-orbital to two electrons (i.e., making two rows identical) yields a zero determinant, automatically enforcing the Pauli exclusion principle.

Although a single Slater determinant satisfies antisymmetry and the Pauli exclusion principle, it represents only an independent-particle model and neglects electron correlation. Furthermore, obtaining the exact non-relativistic wavefunction within the BO approximation generally requires full configuration interaction (FCI)—a linear combination of all possible Slater determinants constructed by distributing the electrons among the available orbitals. This approach, however, remains computationally intractable for systems with more than a few electrons. Various approximations, including truncations of full CI, stem from the practical and highly successful Hartree–Fock method, which restricts the wavefunction to a single optimized Slater determinant. This approach provides a great independent-particle description and serves as the starting point for various modern quantum chemical methods. [7]

### 1.1.2. Hartree-Fock Method

The Hartree-Fock (HF) method aims to minimize the total electronic energy $E_0 = \langle\Psi|\,\hat{H}\,|\Psi\rangle$ by optimizing the spin-orbitals $\chi_k$ within a single Slater determinant ($|\Psi\rangle$). In general, the spin-orbitals can be expressed as linear combinations:

$$\chi_k(\mathbf{x}) = c_\alpha \psi_k^\alpha(\mathbf{r})\alpha(\omega) + c_\beta \psi_k^\beta(\mathbf{r})\beta(\omega) \tag{1.7}$$

where $\psi_k^\alpha(\mathbf{r})$ and $\psi_k^\beta(\mathbf{r})$ are spatial orbitals, $c_\alpha$ and $c_\beta$ are mixing coefficients, and $\alpha(\omega)$ and $\beta(\omega)$ are the spin functions. These orbitals are expressed using a basis set: a collection of mathematical functions, typically centered on atomic nuclei, that approximate atomic orbitals. Molecular orbitals are then expanded as linear combinations of these basis functions. The spin functions are eigenfunctions of the z-component of the spin operator $\hat{S}_z$:

$$\hat{S}_z\,|\alpha(\omega)\rangle = \frac{1}{2}\,|\alpha(\omega)\rangle \quad \text{and} \quad \hat{S}_z\,|\beta(\omega)\rangle = -\frac{1}{2}\,|\beta(\omega)\rangle \tag{1.8}$$

The coefficients $c_\alpha$ and $c_\beta$, along with the coefficients in the spatial functions ($\psi_k = \sum_\mu C_{\mu k}\phi_\mu(\mathbf{r})$, where $\phi(\mathbf{r})$ is an atomic orbital function), are variationally optimized. In most cases, the

spin-orbitals in Eq. (1.7) are simplified to contain only one spin state (i.e., $c_\alpha = 1$, $c_\beta = 0$ or vice versa). This simplification allows the computation to be simpler, as there are fewer parameters to optimize.

The spatial orbitals, $\psi_k^\alpha(\mathbf{r})$ and $\psi_k^\beta(\mathbf{r})$, can be restricted to be identical for both spins; this is known as the restricted Hartree–Fock (RHF) method and is commonly used for closed-shell systems with no unpaired electrons. Conversely, they can be allowed to differ depending on the spin in open-shell systems; this is the unrestricted Hartree–Fock (UHF) method.

The HF energy of a single Slater determinant constructed from orthonormal spin-orbitals can be expressed as the expectation value:

$$E_0 = \langle\Psi|\,\hat{H}_e\,|\Psi\rangle = \langle\Psi|\,\hat{T}_e\,|\Psi\rangle + \langle\Psi|\,\hat{V}_{ne}\,|\Psi\rangle + \langle\Psi|\,\hat{V}_{ee}\,|\Psi\rangle + V_{nn} \tag{1.9}$$

where

$$\langle\Psi|\,\hat{T}_e\,|\Psi\rangle = -\frac{1}{2}\sum_i \int \chi_i^*(\mathbf{r},\omega)\nabla^2\chi_i(\mathbf{r},\omega)d\mathbf{r}d\omega \tag{1.10}$$

$$\langle\Psi|\,\hat{V}_{ne}\,|\Psi\rangle = -\sum_{i,A} \int \chi_i^*(\mathbf{r},\omega)\frac{Z_A}{r_{iA}}\chi_i(\mathbf{r},\omega)d\mathbf{r}d\omega \tag{1.11}$$

$$\langle\Psi|\,\hat{V}_{ee}\,|\Psi\rangle = \frac{1}{2}\sum_{i,j}(J_{ij} - K_{ij}) \tag{1.12}$$

$$J_{ij} = \int \chi_i^*(\mathbf{r}_1,\omega_1)\chi_j^*(\mathbf{r}_2,\omega_2)\frac{1}{r_{12}}\chi_i(\mathbf{r}_1,\omega_1)\chi_j(\mathbf{r}_2,\omega_2)d\mathbf{r}_1 d\omega_1 d\mathbf{r}_2 d\omega_2 \tag{1.13}$$

$$K_{ij} = \int \chi_i^*(\mathbf{r}_1,\omega_1)\chi_j^*(\mathbf{r}_2,\omega_2)\frac{1}{r_{12}}\chi_i(\mathbf{r}_2,\omega_2)\chi_j(\mathbf{r}_1,\omega_1)d\mathbf{r}_1 d\omega_1 d\mathbf{r}_2 d\omega_2 \tag{1.14}$$

and $V_{nn}$ is the nuclear–nuclear repulsion energy, which is constant within the BO approximation. The indices $i$ and $j$ run over the occupied spin-orbitals which are orthonormal.

The one-electron contributions ($\langle\Psi|\,\hat{T}_e\,|\Psi\rangle$ and $\langle\Psi|\,\hat{V}_{ne}\,|\Psi\rangle$) are simply sums over the occupied spin-orbitals. The two-electron repulsion is described by Coulomb ($J_{ij}$) and exchange ($K_{ij}$) integrals. The Coulomb integral $J_{ij}$ represents the classical electrostatic repulsion between the charge distributions of spin-orbitals $\chi_i$ and $\chi_j$. The exchange integral $K_{ij}$, which

arises from the antisymmetry of electrons, is non-zero only when $\chi_i$ and $\chi_j$ have the same spin (opposite-spin contributions vanish upon spin integration). This exchange term creates a Fermi hole, keeping same-spin electrons farther apart on average and thereby lowering the total energy relative to the classical Coulomb repulsion. [10] Notably, the unphysical self-repulsion of an electron with itself is precisely canceled by its self-exchange contribution ($J_{ii} = K_{ii}$).

#### 1.1.2.1. *Self-Consistent Field Method*

The HF method optimizes the molecular orbitals iteratively using the self-consistent field (SCF) procedure. The important quantity in this approach is the Fock operator $\hat{F}(i)$, an effective one-electron operator whose eigenfunctions are the optimized spin-orbitals $\chi_k(\mathbf{x})$ and whose eigenvalues are the orbital energies $\epsilon_k$:

$$\hat{F}\chi_k(\mathbf{x}) = \epsilon_k \chi_k(\mathbf{x}) \tag{1.15}$$

The Fock operator is defined as

$$\hat{F}(1) = \hat{h}(1) + \sum_{j}^{\text{occ}} \left[\hat{J}_j(1) - \hat{K}_j(1)\right] \tag{1.16}$$

where $\hat{h}(1) = -\frac{1}{2}\nabla_1^2 - \sum_A \frac{Z_A}{r_{1A}}$ is the one-electron core Hamiltonian, and the Coulomb ($\hat{J}_j$) and exchange ($\hat{K}_j$) operators are defined analogously to the integrals $J_{ij}$ and $K_{ij}$ (Eqs. (1.13) and (1.14)). In practice, the spatial parts of the spin-orbitals are expanded in the atomic orbital basis $\{\phi_\mu(\mathbf{r})\}$:

$$\psi_k^\alpha(\mathbf{r}) = \sum_\mu C_{\mu k}^\alpha \phi_\mu(\mathbf{r}), \quad \psi_k^\beta(\mathbf{r}) = \sum_\mu C_{\mu k}^\beta \phi_\mu(\mathbf{r}) \tag{1.17}$$

Because the Fock operator itself depends on the occupied spin-orbitals through the Coulomb and exchange operators, the expansion coefficients $C^{\sigma}_{\mu k}$ must be optimized iteratively. Substituting these linear-combination-of-atomic-orbitals (LCAO) expansions into the HF equations and projecting onto the basis functions converts them into the matrix Roothaan–Hall equations:

$$\mathbf{F}^{\sigma}\mathbf{C}^{\sigma} = \mathbf{S}\mathbf{C}^{\sigma}\boldsymbol{\epsilon}^{\sigma} \qquad (\sigma = \alpha, \beta) \tag{1.18}$$

where $\mathbf{S}$ is the overlap matrix with elements $S_{\mu\nu} = \langle \phi_\mu | \phi_\nu \rangle$, $\boldsymbol{\epsilon}^{\sigma}$ contains the orbital energies, and $\mathbf{C}^{\sigma}$ is the coefficient matrix being optimized.

For closed-shell systems treated within the RHF framework, the Fock matrix elements take the explicit form

$$F_{\mu\nu} = H^{\text{core}}_{\mu\nu} + \sum_{\lambda\sigma} P_{\lambda\sigma} \left[ (\mu\nu|\lambda\sigma) - \frac{1}{2}(\mu\lambda|\nu\sigma) \right] \tag{1.19}$$

where

$$H^{\text{core}}_{\mu\nu} = \langle \phi_\mu | \hat{h} | \phi_\nu \rangle, \qquad P_{\lambda\sigma} = 2 \sum_{i}^{\text{occ}} C^{*}_{\lambda i} C_{\sigma i} \tag{1.20}$$

and $(\mu\nu|\lambda\sigma)$ are the two-electron repulsion integrals in chemist notation. The density matrix $P_{\lambda\sigma}$ is constructed from the occupied molecular-orbital coefficients. (In the unrestricted HF case the $\alpha$ and $\beta$ Fock and density matrices are constructed separately.) The SCF procedure begins with an initial guess for the coefficient matrices, typically obtained by diagonalizing the core Hamiltonian matrix. The density matrix is then constructed from these coefficients and used to build the Fock matrix. Solving the generalized eigenvalue problem yields new coefficients (eigenvectors) and orbital energies (eigenvalues). These updated coefficients are used to rebuild the density and Fock matrices, and the cycle is repeated until the total energy or the density matrix converges to within a predefined threshold. Upon convergence, the final optimized coefficients define the molecular orbitals $\psi_k(\mathbf{r})$ and the corresponding spin-orbitals $\chi_k(\mathbf{x})$.

#### 1.1.2.2. *Post Hartree-Fock Methods*

The HF method successfully incorporates several key quantum properties of electrons, such as exchange and the Pauli exclusion principle, but its mean-field treatment of electrostatic repulsion remains a classical approximation that fails to fully capture dynamical electron correlation. Moreover, as a single-determinant approach, HF cannot describe static (or non-dynamic) electron correlation. To achieve higher accuracy, more advanced methods are required that explicitly account for these electron correlations beyond the HF reference.

As discussed earlier, FCI provides the exact solution to the Schrödinger equation within a given finite basis set, properly addressing both dynamic and static electron correlation, but it is computationally impractical for most systems. A practical approach to recovering electron correlation at reduced cost involves defining an active space comprising a subset of orbitals and electrons; the wavefunction is then expressed as a linear combination of all Slater determinants generated by excitations within this active space. This forms the foundation of multi-configurational methods, such as complete active space self-consistent field (CASSCF) theory. [10, 11] CASSCF is particularly effective for systems dominated by strong static (non-dynamic) correlation, as it exactly treats the near-degeneracy effects within the chosen active space. Its major drawback is the exponential growth of computational cost with active-space size, which rapidly becomes prohibitive for larger active spaces.

Single-reference perturbation theory offers a complementary route, primarily targeting the dynamic correlation missing from the HF reference. It expands the correlation energy as a power series using Rayleigh–Schrödinger perturbation theory, with the most common variant being second-order Møller–Plesset perturbation theory (MP2), which is relatively inexpensive. [7,10,12] Higher-order extensions (MP3, MP4, etc.) are possible but significantly more costly. A particularly powerful and systematic single-reference approach for dynamic correlation is the coupled cluster (CC) method, in which the wavefunction is approximated using an exponential ansatz: $|\Psi\rangle = e^{\hat{T}} |\Phi\rangle$, where $|\Phi\rangle$ is the HF reference determinant and

$\hat{T}$ (distinct from the kinetic energy operator) is the cluster operator that generates excited determinants. [10, 13] The operator is expressed as $\hat{T} = \hat{T}_1 + \hat{T}_2 + \hat{T}_3 + \cdots$, corresponding to single, double, triple, and higher excitations. The exponential form automatically includes disconnected higher-order excitations, providing a compact and highly accurate description of dynamic correlation. Truncating $\hat{T}$ at singles and doubles yields CCSD; the widely used CCSD(T) variant (often regarded as the "gold standard" for single-reference ground-state calculations on small- to medium-sized molecules) adds a perturbative treatment of triples, offering excellent accuracy at manageable computational cost. [10, 14]

The bottleneck of the HF method is the two-electron integrals ($J_{ij}$ and $K_{ij}$) as they scale as $O(N^4)$ where $N$ is the number of basis functions. By taking advantage of the sparsity of the Fock matrix, the scaling can be reduced to $O(N^2)$ and even to $O(N)$. [15, 16, 17] The cost of post-HF methods scales more rapidly. In particular, MP2 scales as $O(N^5)$, CCSD and CISD as $O(N^6)$, and CCSD(T) as $O(N^7)$, while full CI scales factorially. This unfavorable scaling arises fundamentally from the many-electron wavefunction, which depends on $4n$ variables, where $n$ is the number of electrons. In the late 1920s, Thomas, [18] Fermi, [19] and later Dirac [20] proposed the Thomas–Fermi–Dirac (TFD) model, which approximates the kinetic and exchange energies by treating the many-electron system locally as a uniform electron gas. [21] This reformulates the problem in terms of the electron density $\rho(\mathbf{r})$, a function of only three spatial variables that is independent of the number of electrons $N$. However, the TFD approach suffers from significant limitations, including its inability to properly reproduce atomic shell structure or molecular binding (due to its semiclassical nature and crude treatment of inhomogeneity). [21] We will discuss in the following section how Hohenberg, Kohn, and Sham advanced this density-based idea into a rigorous and practical framework for calculating electronic energies.

### 1.1.3. Density Functional Theory

The TFD approach constructs the energy as:

$$E[\rho] = T[\rho] + E_x[\rho] + V_{\text{ext}}[\rho] + V_{nn} + \frac{1}{2}\int\int \frac{\rho(\mathbf{r})\rho(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}\, d\mathbf{r}\, d\mathbf{r}' \tag{1.21}$$

where

$$T[\rho] = \frac{3}{10}(3\pi^2)^{2/3}\int(\rho(\mathbf{r}))^{5/3}d\mathbf{r} \tag{1.22}$$

$$E_x[\rho] = -\frac{3}{4}\left(\frac{3}{\pi}\right)^{1/3}\int(\rho(\mathbf{r}))^{4/3}d\mathbf{r} \tag{1.23}$$

$$V_{\text{ext}}[\rho] = -\int \rho(\mathbf{r})v_{\text{ext}}(\mathbf{r})d\mathbf{r} \tag{1.24}$$

Here, $T[\rho]$ is the kinetic energy functional derived independently by Thomas [18] and Fermi [19] by treating the electrons as a locally uniform electron gas (UEG). The exchange term $E_x[\rho]$ is the Dirac exchange functional, [20] obtained by applying the exact HF exchange energy of the UEG in the local-density approximation. The term $V_{\text{ext}}[\rho]$ represents the attractive electron-nucleus interaction, $V_{nn}$ is the nuclear-nuclear repulsion energy, and the final integral is the classical electrostatic electron-electron repulsion.

Although the TFD model was innovative in expressing energies in terms of the electron density $\rho(\mathbf{r})$, it lacked a rigorous theoretical foundation. Two key issues that Hohenberg and Kohn addressed in 1964 were the absence of proof that the ground-state density uniquely determines the external potential $v_{\text{ext}}(\mathbf{r}) = -\sum_A Z_A/|\mathbf{r}-\mathbf{R}_A|$ and thus the Hamiltonian and all ground-state observables, as well as the lack of a variational principle ensuring that minimizing an energy functional of the density yields the true ground-state energy. [22] Prior to their work, it was unclear whether different external potentials could produce the same ground-state density (the question of non-uniqueness) or whether the density alone suffices to determine the system's properties. Furthermore, there was no formal justification that an "optimized" density corresponds to the exact ground-state energy. [22, 23, 24]

#### 1.1.3.1. *The Hohenberg-Kohn Theorems*

The first Hohenberg–Kohn (HK) theorem proves the uniqueness of the external potential (up to a constant) as a functional of the ground-state density via *reductio ad absurdum*. [22] Assume that there exists another external potential $v'_{\text{ext}}(\mathbf{r})$ whose ground-state wavefunction $\Psi'$ yields the same density $\rho(\mathbf{r})$ as in Eq. (1.24). Let $\Psi'$ be its ground-state wavefunction with energy $E' = \langle\Psi'|\hat{H}'|\Psi'\rangle$, where $\hat{H}'$ is the Hamiltonian incorporating $v'_{\text{ext}}(\mathbf{r})$. By the variational principle:

$$E' = \langle\Psi'|\,\hat{H}'\,|\Psi'\rangle < \langle\Psi|\,\hat{H}'\,|\Psi\rangle = \langle\Psi|\,\hat{H} + \hat{V}'_{\text{ext}} - \hat{V}_{\text{ext}}\,|\Psi\rangle \tag{1.25}$$

yielding

$$E' < E + \int [v'_{\text{ext}}(\mathbf{r}) - v_{\text{ext}}(\mathbf{r})]\rho(\mathbf{r})\,d\mathbf{r} \tag{1.26}$$

Applying the same argument with primed and unprimed quantities interchanged gives

$$E < E' + \int [v_{\text{ext}}(\mathbf{r}) - v'_{\text{ext}}(\mathbf{r})]\rho(\mathbf{r})\,d\mathbf{r} \tag{1.27}$$

Adding Eqs. (1.26) and (1.27) leads to the contradiction

$$E' + E < E' + E \tag{1.28}$$

Therefore, the external potential $v_{\text{ext}}(\mathbf{r})$ is a unique functional of the ground-state density $\rho(\mathbf{r})$, up to a constant.

The second HK theorem establishes a variational principle for the ground-state density. The total energy can be expressed as the functional:

$$E_v[\rho] = F[\rho] + \int v_{\text{ext}}(\mathbf{r})\rho(\mathbf{r})\,d\mathbf{r} \tag{1.29}$$

where the universal functional $F[\rho]$ contains the kinetic energy and electron–electron interaction for any density $\rho$. This functional is minimized by the true ground-state density $\rho_0(\mathbf{r})$,

and the minimum value is the ground-state energy $E_0$

$$E_v[\rho] \geq E_0 = E_v[\rho_0] \tag{1.30}$$

for any trial density $\rho$ with $\int \rho(\mathbf{r})d\mathbf{r} = N$ ($N$ being the number of electrons) and $\rho \geq 0$.

To prove this, assume a trial density $\rho' \neq \rho_0$ that is the ground-state density for some other potential $v'_{\text{ext}}(\mathbf{r})$, with ground-state wavefunction $\Psi'$ and energy $E'$. Applying the wave-function variational principle to the original Hamiltonian using $\Psi'$ as a trial function yields:

$$E_0 \leq \langle \Psi' | \hat{H} | \Psi' \rangle = E' + \int [v_{\text{ext}}(\mathbf{r}) - v'_{\text{ext}}(\mathbf{r})]\rho'(\mathbf{r})\, d\mathbf{r} = E_v[\rho'] \tag{1.31}$$

Equality in this relation holds only if $\rho' = \rho_0$, because otherwise $\Psi'$ would be the ground state for the different external potentials, contradicting the first HK theorem. Thus, the ground-state energy and density are obtained by minimizing $E_v[\rho]$ over all valid densities.

These two theorems further solidified the density-based approach to approximating the ground-state energy, but several challenges remained. The universal functional $F[\rho]$ introduced by Hohenberg and Kohn is unknown, and exactly evaluating it would require searching the full many-body Hilbert space to find the many-electron wavefunction $\Psi$ that yields the true ground-state density $\rho_0(\mathbf{r})$, which is computationally intractable. Additionally, direct density-based methods suffered from poor accuracy, particularly due to large errors in the approximate kinetic energy functional. In the following section, we discuss how the Kohn–Sham formulation addressed these challenges and enabled the widespread practical application of modern DFT.

#### 1.1.3.2. *The Kohn-Sham Theory*

Suppose the true interacting Hamiltonian is:

$$\hat{H} = \hat{T} + \hat{V}_{ee} + \hat{V}_{ext} \tag{1.32}$$

The Kohn-Sham (KS) theory proposes an auxiliary system of non-interacting electrons moving in an effective potential $v_{eff}(\mathbf{r})$ with a new Hamiltonian:

$$\hat{H}_{KS} = \hat{T}_{KS} + \sum_i v_{eff}(\mathbf{r}_i) \tag{1.33}$$

where $\hat{T}_{KS}$ is the kinetic energy operator for non-interacting particles (identical to the kinetic operator in the true interacting system and in HF), but evaluated using the KS orbitals. These orbitals were introduced to mitigate the large kinetic energy errors present in earlier direct density functionals. The KS ground-state wavefunction $\Psi^{KS}$ is a single Slater determinant constructed from the occupied spatial orbitals $\{\phi_i(\mathbf{r})\}$, which are used to build the density $\rho(\mathbf{r}) = \sum_i |\phi_i(\mathbf{r})|^2$. Note that there is no $\hat{V}_{ee}$ term, and all many-body interaction effects are incorporated into $v_{\text{eff}}(\mathbf{r})$:

$$v_{\text{eff}}(\mathbf{r}) = v_{\text{ext}}(\mathbf{r}) + v_{\text{H}}[\rho](\mathbf{r}) + v_{\text{xc}}[\rho](\mathbf{r}) \tag{1.34}$$

where $v_{\text{ext}}(\mathbf{r})$ is the external potential. The term $v_{\text{H}}[\rho](\mathbf{r})$ represents the classical Hartree (Coulomb) potential in KS DFT, describing the electrostatic repulsion that an electron at position $\mathbf{r}$ experiences from the total electron density cloud. The exchange-correlation potential $v_{\text{xc}}[\rho](\mathbf{r})$ is the functional derivative of the exchange-correlation energy $E_{\text{xc}}[\rho]$:

$$E_{\text{xc}}[\rho] = F[\rho] - T_s[\rho] - E_{\text{H}}[\rho] \tag{1.35}$$

It captures the remaining contributions after subtracting the non-interacting kinetic energy and the classical Coulomb energy from the universal functional $F[\rho]$. The exact total ground-state energy can then be expressed as:

$$E[\rho] = T_s[\rho] + \int v_{\text{ext}}(\mathbf{r})\rho(\mathbf{r})\, d\mathbf{r} + E_{\text{H}}[\rho] + E_{\text{xc}}[\rho] \tag{1.36}$$

where

$$T_s[\rho] = \sum_{i=1}^{N} \langle \phi_i | -\frac{1}{2}\nabla^2 | \phi_i \rangle \tag{1.37}$$

and

$$E_{\mathrm{H}}[\rho] = \frac{1}{2} \int \int \frac{\rho(\mathbf{r})\rho(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|} \, d\mathbf{r} \, d\mathbf{r}' \tag{1.38}$$

From the two HK theorems, this expression is exact if $E_{\mathrm{xc}}[\rho]$ is exact. In practice, however, $E_{\mathrm{xc}}[\rho]$ must be approximated. Based on their accuracy, these approximations are assigned to different rungs of "Jacob's ladder", where the highest rung represents the "heaven" of exact DFT. [25]

The three lowest rungs of Jacob's ladder correspond to semi-local approximations: the Local Density Approximation (LDA), Generalized Gradient Approximation (GGA), and meta-Generalized Gradient Approximation (meta-GGA), in ascending order. The LDA (rung 1) assumes the system is locally uniform, depending on the density $\rho(\mathbf{r})$. The exchange-correlation energy is:

$$E_{\mathrm{xc}}^{\mathrm{LDA}}[\rho] = \int \rho(\mathbf{r}) \, \varepsilon_{\mathrm{xc}}^{\mathrm{unif}}(\rho(\mathbf{r})) \, d\mathbf{r} \tag{1.39}$$

where $\varepsilon_{\mathrm{xc}}^{\mathrm{unif}}(\rho)$ is the exchange-correlation energy per particle of the uniform electron gas (exact exchange + parametrized correlation). LDA is simple but often overbinds and underestimates bond lengths.

The GGA (rung 2) improves upon this by incorporating the density gradient $\nabla\rho(\mathbf{r})$, better capturing inhomogeneities in molecular and surface systems.

$$E_{\mathrm{xc}}^{\mathrm{GGA}}[\rho] = \int \rho(\mathbf{r}) \, \varepsilon_{\mathrm{xc}}^{\mathrm{GGA}}(\rho(\mathbf{r}), \nabla\rho(\mathbf{r})) \, d\mathbf{r} \tag{1.40}$$

Popular examples include PBE, [26] BLYP, [27, 28] and PW91. [29]

The meta-GGA (mGGA, rung 3) further enhances accuracy by including higher-order information. [30] Specifically, it incorporates ingredients such as the kinetic energy density $\tau(\mathbf{r})$ from the Kohn-Sham orbitals [31] or the Laplacian of the density $\nabla^2\rho(\mathbf{r})$, [32] yielding

superior performance for diverse properties like bond energies and lattice constants. The exchange-correlation energy in mGGAs takes the general form:

$$E_{\mathrm{xc}}^{\mathrm{mGGA}}[\rho] = \int \rho(\mathbf{r})\, \varepsilon_{\mathrm{xc}}^{\mathrm{mGGA}}\big(\rho(\mathbf{r}), |\nabla\rho(\mathbf{r})|, \nabla^2\rho(\mathbf{r}) \text{ or } \tau(\mathbf{r})\big)\, d\mathbf{r} \tag{1.41}$$

where

$$\tau(\mathbf{r}) = \sum_i \frac{1}{2}|\nabla\phi_i(\mathbf{r})|^2 \tag{1.42}$$

(with the sum over occupied orbitals).

Rung 4 is reached by mixing a portion of exact HF exchange with DFT exchange to reduce self-interaction errors. This approach is called hybrid functionals. One of the most widely employed hybrid functionals is PBE0: [33]

$$E_{\mathrm{xc}}^{\mathrm{PBE0}} = (1-c)E_x^{\mathrm{PBE}}[\rho] + c\, E_x^{\mathrm{HF}} + E_c^{\mathrm{PBE}}[\rho] \tag{1.43}$$

where $c = 0.25$, resulting in 25% exact HF exchange and 75% PBE GGA exchange, with full PBE correlation. Another very popular and successful method is B3LYP: [34]

$$E_{\mathrm{xc}}^{\mathrm{B3LYP}} = (1-a-b)E_x^{\mathrm{LDA}} + a\, E_x^{\mathrm{HF}} + b\, E_x^{\mathrm{B88}} + (1-c)E_c^{\mathrm{LDA}} + c\, E_c^{\mathrm{LYP}} \tag{1.44}$$

It is an empirical three-parameter hybrid using 20% exact HF exchange ($a = 0.20$), 72% Becke88 (B88) GGA exchange ($b = 0.72$), [27] 8% LDA exchange (from the $1-a-b$ term), and correlation from 81% LYP GGA + 19% LDA ($c = 0.81$). [28, 35]

A further improvement towards rung 5 (double hybrids) was achieved by incorporating unoccupied orbitals via perturbation theory (MP2 correlation energy):

$$E_{\mathrm{xc}}^{\mathrm{DH}} = (1-\alpha)E_x^{\mathrm{GGA}} + \alpha\, E_x^{\mathrm{HF}} + (1-\beta)E_c^{\mathrm{GGA}} + \beta\, E_c^{\mathrm{MP2}} \tag{1.45}$$

where the added perturbation energy term $\beta E_c^{\mathrm{MP2}}$ is the MP2 correlation energy computed with KS orbitals. [36] The parameters $\alpha$ and $\beta$ are typically determined empirically or semi-empirically (e.g., $\alpha = 0.53$, $\beta = 0.27$ in the B2-PLYP functional). [36]

More recent developments in DFT have addressed the long-range deficiencies of hybrid functionals such as B3LYP. Range-separated hybrids, like $\omega$B97X, overcome the incorrect asymptotic behavior of exchange by splitting the Coulomb operator, treating the short-range contribution primarily with DFT exchange while employing 100% HF exchange at long range. [37] Another long-range correlation problem, the complete absence of London dispersion (van der Waals) interactions in standard functionals, is addressed either by the nonlocal correlation functional VV10, which adds a fully density-dependent dispersion term to the correlation energy, [38] or by the computationally inexpensive empirical correction DFT-D3 developed by Grimme and coworkers. [39] In parallel, another major trend has been the construction of heavily parameterized functionals; examples from the Minnesota series such as M06 and MN15 optimize dozens of empirical coefficients against extensive benchmark databases to achieve high accuracy across thermochemistry, kinetics, and noncovalent interactions. [40, 41]

Thus far, we have discussed methods for calculating the energy of a molecular system. With the energy known, the slope of the PES can be obtained, either numerically or analytically. This slope corresponds to the energy gradients, which yield the forces acting on each atom. Using these forces within the framework of Newtonian mechanics, we can simulate the dynamic behavior of molecules under given conditions over time. Details of such simulation techniques, along with workflows for refining simulation trajectories, will be discussed in the following sections.

## 1.2. Ab Initio Nanoreactor Molecular Dynamics Simulation

Computational simulations serve as powerful tools to complement and interpret experimental results. In particular, ab initio molecular dynamics (AIMD) simulations are widely employed

to predict and rationalize chemically relevant processes observed experimentally. AIMD simulations use classical nuclear dynamics on quantum potential energy surfaces within the BO approximation. In classical Newtonian dynamics, the forces acting on each atom $i$ are obtained as the negative gradient of the potential energy, $\mathbf{F}_i = -\nabla_i V$. The forces are then used to calculate acceleration, $\mathbf{F}_i = m_i \mathbf{a}_i$. The goal of AIMD simulations is to advance positions $\mathbf{r}$ and velocities $\mathbf{v}$ from time $t$ to $t + \Delta t$ until the simulation ends. A widely used approach to propagate positions and velocities is the Velocity Verlet algorithm. [42, 43, 44]

The position is first predicted using a Taylor expansion truncated at the $\Delta t^2$ term:

$$\mathbf{r}(t + \Delta t) = \mathbf{r}(t) + \Delta t\, \mathbf{v}(t) + \frac{\Delta t^2}{2}\mathbf{a}(t) + O(\Delta t^3) \tag{1.46}$$

A naive velocity update using only the current acceleration would be:

$$\mathbf{v}(t + \Delta t) = \mathbf{v}(t) + \Delta t\, \mathbf{a}(t) + O(\Delta t^2) \tag{1.47}$$

At first glance, this velocity update appears reasonable. However, it corresponds to the explicit Euler method, which is only first-order accurate and leads to systematic energy drift. The Velocity Verlet algorithm improves this by incorporating the "future" acceleration $\mathbf{a}(t + \Delta t)$ (computed from the new positions) via trapezoidal averaging:

$$\mathbf{v}(t + \Delta t) = \mathbf{v}(t) + \frac{\Delta t}{2}\left[\mathbf{a}(t) + \mathbf{a}(t + \Delta t)\right] + O(\Delta t^3) \tag{1.48}$$

Figure 1.1 shows the total energy evolution in simulations of a $3 \times 3 \times 3$ xenon cube with initial interatomic separations corresponding to the minimum of the Lennard-Jones potential along the edges. The Velocity Verlet method resolves the energy drift issue with essentially the same computational cost per time step as the Euler method, since the force evaluation at $t + \Delta t$ is reused as the initial acceleration for the subsequent step.

Another important detail is the choice of time step size, which must be small enough to accurately capture the fastest molecular vibrational modes without being excessively

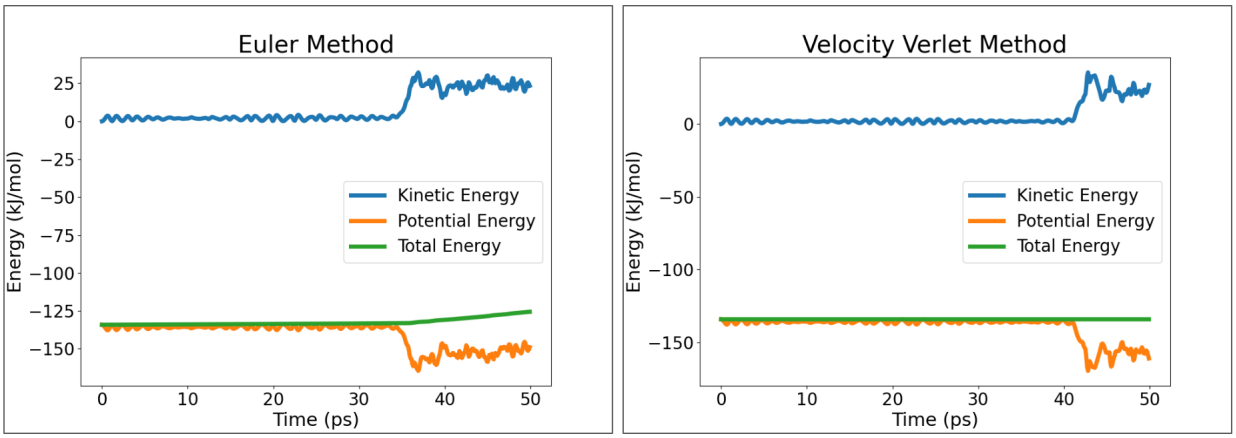



FIGURE 1.1. Comparison of total energy evolution in MD simulations using the explicit Euler (left) and Velocity Verlet (right) integrators with identical time steps (0.002 ps). The explicit Euler method shows the total energy drift due to its reliance on only the current-time acceleration, whereas Velocity Verlet conserves energy well.

small (to avoid unreasonable computational cost). A typical C–H stretching wavenumber is $\sim 3000$ cm$^{-1}$, corresponding to an oscillatory period of $\sim 11$ fs. To ensure stability, the time step is generally chosen to be about 10–20 times smaller than this period (i.e., $\sim 1$ fs). Motions involving heavier atoms are significantly slower, so the constraining factor that leads to high computational cost is primarily the high-frequency C–H vibrations.

To mimic typical experimental conditions at constant temperature and correctly sample the canonical (NVT) ensemble, thermostatting is required. Langevin dynamics, which extends the Newtonian equations by adding frictional damping and stochastic forces that model thermal fluctuations, is commonly used for this purpose. [45]

Molecules in a standard AIMD simulation with a Langevin thermostat and no external forces will simply diffuse and vibrate due to thermal fluctuations. Occasional collisions may lead to chemical reactions, but the probability is extremely low unless the processes are barrierless (e.g., radical recombinations). The AIMD nanoreactor method addresses this limitation by periodically inducing high-energy molecular collisions within a confined spherical volume using a time-dependent external potential. [46] The mathematical details

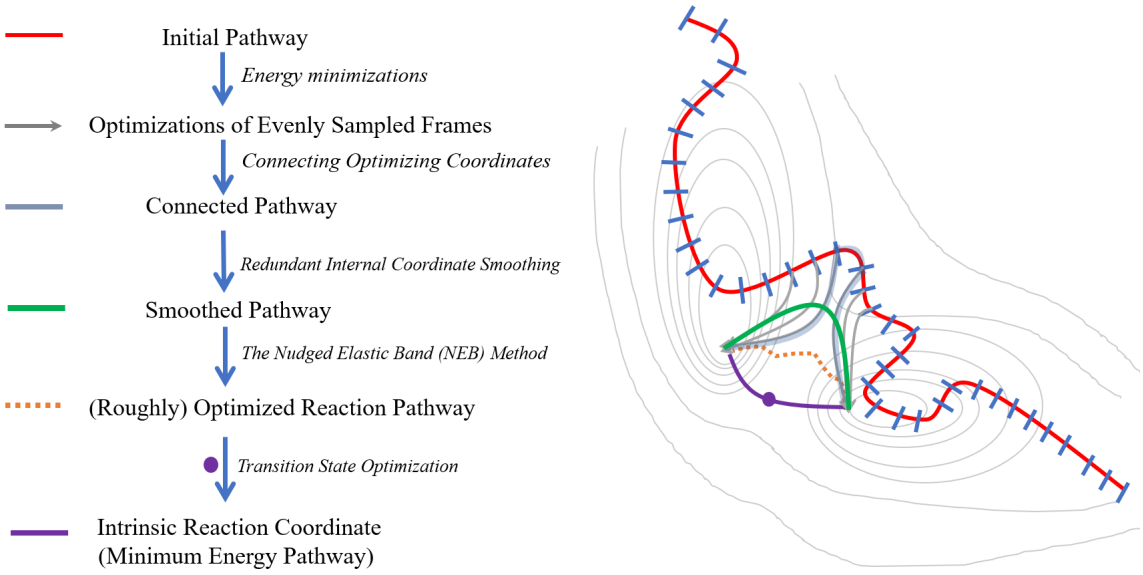



FIGURE 1.2. Illustration of the automated workflow's key processing steps, transforming raw trajectories (red) into refined minimum energy pathways (purple).

of how the external potential is applied to the molecules and how reaction events are detected are discussed in Chapter 2.

## 1.3. Automated Reaction Pathway Refinement Workflow

The nanoreactor approach is particularly effective for discovering non-intuitive and mechanistically interesting reactions; however, the generated trajectories frequently explore high-energy regions of the PES. To compute accurate reaction rates for these discovered processes, minimum energy pathways connecting reactant and product basins on the PES must be determined. This requires refining the raw, high-energy trajectories into well-defined elementary steps. This process quickly becomes impractical to perform manually when processing results from a large number of simulations. Consequently, an automated workflow is essential for efficient post-processing and refinement. Details of such an automated algorithm, tailored specifically to refine these high-energy trajectories, are presented in the following section.

The main goal of this workflow is to locate energy basins in close proximity to an initial molecular trajectory and identify the first-order saddle points that connect these basins on the PES. By determining the energy barriers (i.e., the energy differences between the minima of the basins and the corresponding saddle points), one can apply transition state theory (TST) to compute reaction rate constants. [47, 48, 49]

Locating local minima (energy basins) is relatively straightforward: steepest-descent or similar methods follow the negative gradient of the PES downhill to convergence. In contrast, reliably characterizing transition state (TS) structures (first-order saddle points with exactly one negative Hessian eigenvalue) is much more challenging. Successful optimization of a TS geometry heavily relies on a high-quality initial guess structure that already lies near the desired saddle point.

To generate such robust initial guesses, reliable algorithms play a key role. These typically involve constructing approximate reaction pathways that connect reactant and product basins (e.g., via interpolation methods) and then iteratively refining or guiding those pathways toward the true saddle point region. The resulting pathway maxima often serve as excellent starting geometries for final TS optimization.

This section details the complete workflow, starting with energy minimization techniques for identifying energy basins, followed by strategies for TS search and refinement, and ultimately the acquisition of refined elementary reaction pathways. Figure 1.2 presents a schematic of the overall workflow, which determines the minimum energy pathway connecting two energy basins, starting from a simulation trajectory.

### 1.3.1. Energy Minimization

The geometries corresponding to evenly spaced frames in the initial trajectories (i.e., every 10th frame) are minimized to locate energy minima on the PES. The simplest algorithm for energy minimization is steepest descent, which follows the negative gradient direction at the

current geometry ($-\nabla V(\mathbf{x}_k)$) in Cartesian coordinates $\mathbf{x} = (x_1, \ldots, x_n)$.

$$\mathbf{x}_{k+1} = \mathbf{x}_k - h\,\nabla V(\mathbf{x}_k) \tag{1.49}$$

Here, $h$ is the step size taken from the current geometry ($\mathbf{x}_k$) to the next geometry ($\mathbf{x}_{k+1}$). The PES is often a multidimensional and highly nonlinear landscape, for which steepest descent is not sufficiently efficient. A more advanced and widely used approach is the Newton-Raphson method, which incorporates the curvature of the PES through the second derivatives of the energy. The matrix containing this information is called the Hessian $\mathbf{H}$:

$$\mathbf{H} = \nabla^{(2)} V = \begin{pmatrix} \frac{\partial^2 V}{\partial x_1^2} & \frac{\partial^2 V}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 V}{\partial x_1 \partial x_n} \\ \frac{\partial^2 V}{\partial x_2 \partial x_1} & \frac{\partial^2 V}{\partial x_2^2} & \cdots & \frac{\partial^2 V}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 V}{\partial x_n \partial x_1} & \frac{\partial^2 V}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 V}{\partial x_n^2} \end{pmatrix} \tag{1.50}$$

Once the Hessian is obtained, a Newton-Raphson step is taken as:

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \mathbf{H}^{-1}(\mathbf{x}_k)\nabla V(\mathbf{x}_k) \tag{1.51}$$

The inverse Hessian allows larger steps where the curvature is small (small eigenvalues) and smaller steps where the curvature is large (large eigenvalues). Newton-Raphson methods achieve rapid (typically quadratic) convergence by incorporating curvature information from the Hessian, but the computational cost of evaluating and inverting the full Hessian at each iteration is rather expensive. Quasi-Newton methods address this by starting with an approximate initial Hessian (often constructed from simple empirical rules that reflect molecular connectivity and bond stiffness [50]) and then iteratively improving the approximation using only first-derivative (gradient) information via low-rank updates, most commonly the Broyden–Fletcher–Goldfarb–Shanno (BFGS) formula. [51, 52, 53, 54]

$$\mathbf{H}^{\text{new}} = \mathbf{H}^{\text{old}} + \frac{\Delta\mathbf{g}\,\Delta\mathbf{g}^T}{\Delta\mathbf{g}^T\Delta\mathbf{x}} - \frac{\mathbf{H}^{\text{old}}\Delta\mathbf{x}\,\Delta\mathbf{x}^T\mathbf{H}^{\text{old}}}{\Delta\mathbf{x}^T\mathbf{H}^{\text{old}}\Delta\mathbf{x}} \tag{1.52}$$

The updated Hessian approximation can be pseudo-inverted using singular value decomposition, inverting only the singular values larger than a threshold. The resulting search direction is then used to update the parameters. This process repeats until convergence is achieved.

#### 1.3.1.1. *Coordinate System*

The choice of coordinate system for optimizing molecular structures is critical to the algorithm's performance. Using Cartesian coordinates often leads to a zig-zagging path on the PES due to strong coupling between the coordinates (large off-diagonal elements in the Hessian). A more systematic approach is to define internal coordinates (ICs) based on atomic connectivity (e.g., bond stretches, valence angles, and torsion angles) and adjust them toward energy minima.

The Wilson B matrix [55] $\mathbf{B}$ captures how internal coordinates $q$ change with variations in Cartesian coordinates ($B_{ij} = \partial q_i/\partial x_j$). Using $\mathbf{B}$, the gradient and Hessian in internal coordinates can be obtained as:

$$\mathbf{g}_q = \mathbf{B}^{-T}\mathbf{g}_x \tag{1.53}$$

$$\mathbf{H}_q = \mathbf{B}^{-T}\left(\mathbf{H}_x - \sum_k \frac{\partial\mathbf{B}}{\partial x_k} g_{q,k}\right)\mathbf{B}^{-1} \tag{1.54}$$

where the subscripts $x$ and $q$ denote Cartesian and internal coordinates, respectively, and the second term accounts for the coordinate dependence of $\mathbf{B}$. The Newton-Raphson update in internal coordinates, $\Delta\mathbf{q} = -\mathbf{H}_q^{-1}\mathbf{g}_q$, can then be used to update the Cartesian coordinates:

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \mathbf{B}^T\mathbf{G}^{-1}\Delta\mathbf{q} \tag{1.55}$$

where $\mathbf{G} = \mathbf{B}\mathbf{B}^T$. [56]

One of the early forms of ICs is called the Z-matrix coordinate system, which describes the position of an atom by a bond distance, a valence angle, and a dihedral angle relative to previously defined atoms, starting from a reference atom. [57] Z-matrices often require manual construction, making them error-prone, especially for large systems. A more standardized and robust approach is to automatically define ICs based on intermolecular connectivity as primitive bond stretches, valence angles, and torsion angles. These are referred to here as primitive ICs (PICs). When the number of PICs exceeds the number of internal degrees of freedom (3N-6 for non-linear molecules or 3N-5 for linear molecules, where N is the number of atoms), the **G** matrix becomes singular because the rows of **B** are linearly dependent. Pulay and Fogarasi initially addressed redundancies by constructing non-redundant sets through combinations of primitive functions guided by local pseudosymmetry, rather than selecting individual coordinates. [58, 59] Later, they introduced the use of the generalized inverse of **G** matrix to handle redundancies more systematically while keeping a redundant primitive set. [60] Schlegel and co-workers advanced this further by automatically generating redundant internal coordinates (RICs), typically consisting of all primitive bonds, angles, and torsions, and optimizing directly in this redundant space using the generalized inverse. [61] In redundant optimization, the generalized inverse must be computed in every iteration to project out redundancies and translational/rotational modes, which can be computationally expensive. To reduce this cost, Baker et al. introduced delocalized internal coordinates (DLCs): a non-redundant working set obtained by diagonalizing the **G** matrix built from all PICs at the initial geometry and then using this fixed non-redundant set throughout the optimization. [62, 63, 64] More recently, a well-defined IC system for systems with multiple molecules or fragments has been introduced. [65] It incorporates translational and rotational ICs (TRICs) as additional primitives when constructing the **B** matrix and then applies the same DLC procedure to obtain the working coordinate system. [65]

Once the energy minimization calculations for all selected frames are complete, optimization trajectories (those tracing the path from the initial to the optimized structure) that converge to the same energy minimum are grouped together. Two trajectories from different groups are then connected by inserting a segment of the original trajectory between them. This composite path serves as an initial guess for the minimum energy pathway, which is refined in the following section.

### 1.3.2. Minimum Energy Pathway Search

The minimum energy pathway (MEP) is the steepest-descent path connecting two minima (reactant and product) on the PES while passing through the lowest-energy first-order saddle point, known as the transition state (TS). Identifying a high-quality initial guess for the TS is essential for successful MEP location as discussed above.

The initial concatenated path from the previous step typically contains kinks at the connection points, requiring smoothing before it can be guided toward the true MEP. The smoothing procedure minimizes an objective function that aligns smoothed ICs (achieved by convolution with a windowing function) with the corresponding interpolated Cartesian coordinates derived from the optimization trajectories. This approach can lead to discontinuities or sudden jumps in Cartesian coordinates when the topology of the objective function changes abruptly during the transition from reactant to product.

To address these limitations, a more robust internal-coordinate-based interpolation method, TRICS, is proposed in Chapter 3. TRICS requires only the optimized geometries of the reactant and product, eliminating the need for concatenated optimization trajectories and subsequent smoothing. Its performance is benchmarked against other widely used interpolation techniques, with results presented in the same chapter.

Once a smooth initial path is available, it can be refined using an MEP search algorithm such as the nudged elastic band (NEB) method. NEB optimizes a chain of evenly spaced images while applying spring forces between adjacent images to maintain spacing and prevent

the chain from collapsing into a single minimum. Further details of the NEB method and its performance are provided in Chapter 4.

The highest-energy image along the converged NEB chain serves as an excellent initial guess for TS optimization. Once this guess has been refined to a true first-order saddle point (confirmed by vibrational analysis showing exactly one imaginary frequency aligned with the reaction coordinate) one can follow the negative gradients in both directions along the eigenvector of that imaginary mode. This procedure traces an approximate MEP and verifies connectivity between the saddle point and the intended reactant and product minima.

A widely employed and standard method for rigorously tracing the full MEP from the saddle point is the intrinsic reaction coordinate approach, which follows the steepest-descent path in mass-weighted coordinates. Details of the method will be discussed in the following section.

#### 1.3.2.1. *The Intrinsic Reaction Coordinate Method*

The intrinsic reaction coordinate (IRC) method follows the steepest-descent path on the PES from the TS in mass-weighted coordinates toward lower potential energy. [66, 67] The resulting IRC is widely regarded as the most physically meaningful representation of the reaction path. Here, we discuss the mass-weighted internal coordinate IRC method developed by Gonzalez and Schlegel. [68] Figure 1.3 illustrates a single iteration of the Gonzalez-Schlegel IRC algorithm.

Let $\mathbf{q}$ denote the vector of internal coordinates for the current structure. The mass-weighted Wilson B-matrix elements are defined as $\partial q_i/(\partial x_j\sqrt{m_j})$, where $m_j$ is the mass of atom $j$. The G-matrix is $\mathbf{G} = \mathbf{B}\mathbf{B}^T$. Each iteration employs a controlled step size $s$ (in mass-weighted units). The initial half-step moves to a pivot point $\mathbf{q}^*$ along the normalized negative gradient direction:

$$\mathbf{q}^* = \mathbf{q} - \frac{s}{2} \cdot \frac{\mathbf{G} \cdot \mathbf{g}}{(\mathbf{g}^T \cdot \mathbf{G} \cdot \mathbf{g})^{1/2}} \tag{1.56}$$

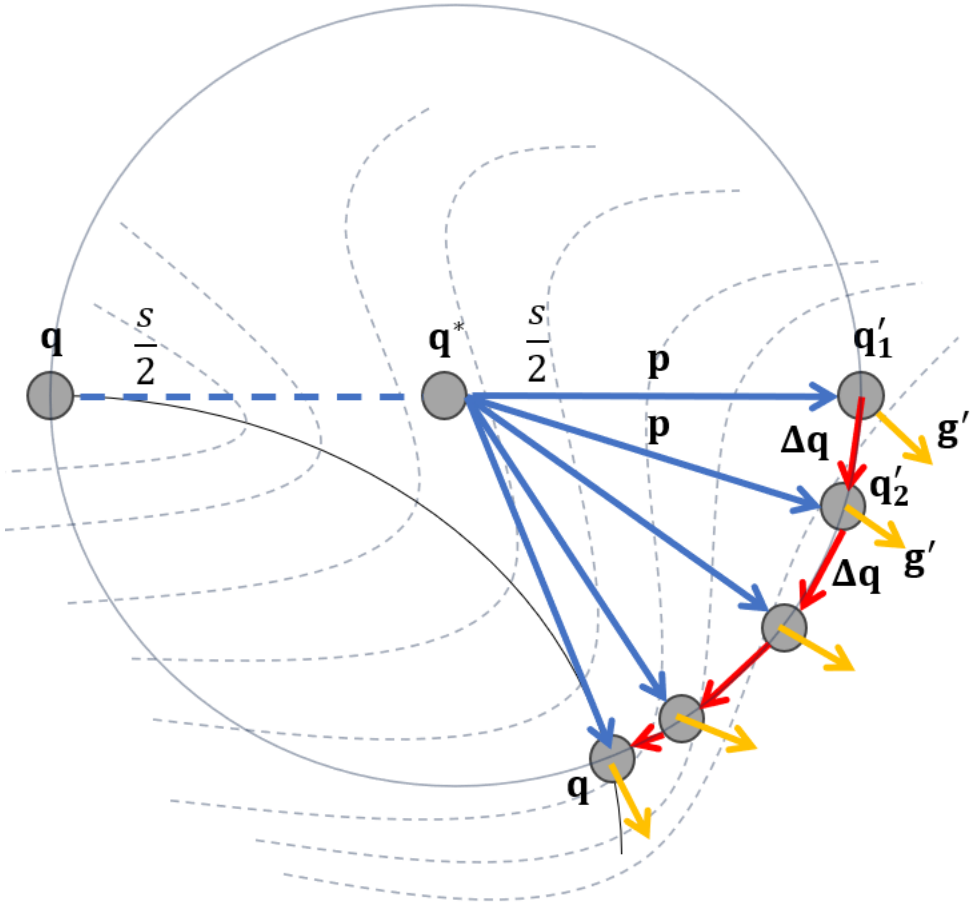



FIGURE 1.3. Schematic of a single iteration in the Gonzalez-Schlegel IRC method. Dotted lines represent contours of the PES, while the solid line denotes the MEP. The circle indicates the hyperspherical constraint surface centered at the pivot point $\mathbf{q}^*$. Refer to Eqs. (1.56)–(1.61) for details on the notations.

where $\mathbf{g}$ is the gradient in internal coordinates. (For the initial step from the TS, the direction follows the eigenvector of the imaginary mode.) The corrector phase begins with a tentative half-step from $\mathbf{q}^*$ in the same direction, reaching a temporary point $\mathbf{q}'_n$. This point is then iteratively adjusted along a hyperspherical arc (centered at $\mathbf{q}^*$ with radius $s/2$) until the predicted negative gradient at the new point aligns with the tangent vector from the pivot point. Define:

$$\mathbf{p} = \mathbf{q}'_n - \mathbf{q}^*, \qquad \Delta\mathbf{q} = \mathbf{q}'_{n+1} - \mathbf{q}'_n. \tag{1.57}$$

Transform to mass-weighted coordinates:

$$\Delta \mathbf{q}_{\mathrm{M}} = \mathbf{G}^{-1/2}\Delta \mathbf{q}, \qquad \mathbf{p}_{\mathrm{M}} = \mathbf{G}^{-1/2}\mathbf{p} \tag{1.58}$$

$$\mathbf{g}_{\mathrm{M}} = \mathbf{G}^{1/2}\mathbf{g}', \qquad \mathbf{H}_{\mathrm{M}} = \mathbf{G}^{1/2}\mathbf{H}\mathbf{G}^{1/2} \tag{1.59}$$

where $\mathbf{g}'$ and $\mathbf{G}$ are evaluated at $\mathbf{q}'_n$, and $\mathbf{H}$ is the Hessian in internal coordinates (approximated quadratically from $\mathbf{q}$). The first constraint is the fixed half-step length (hypersphere radius):

$$(\mathbf{p}_{\mathrm{M}} + \Delta \mathbf{q}_{\mathrm{M}})^T(\mathbf{p}_{\mathrm{M}} + \Delta \mathbf{q}_{\mathrm{M}}) = \left(\frac{s}{2}\right)^2 \tag{1.60}$$

The second constraint is parallelism between the tangent and predicted negative gradient (via Lagrange multiplier $\lambda$):

$$\Delta \mathbf{q}_{\mathrm{M}} = -(\mathbf{H}_{\mathrm{M}} - \lambda \mathbf{I})^{-1}(\mathbf{g}_{\mathrm{M}} - \lambda \mathbf{p}_{\mathrm{M}}) \tag{1.61}$$

Substituting Eq. (1.61) into Eq. (1.60) yields a scalar equation solved iteratively for $\lambda$. The resulting $\Delta \mathbf{q}_{\mathrm{M}}$ updates $\mathbf{q}'$ to $\mathbf{q}'_{n+1}$, and the process repeats until $\|\Delta \mathbf{q}\|$ falls below a predefined threshold (e.g., $10^{-6}$). Upon convergence, the second-half step is taken from $\mathbf{q}^*$ along $\mathbf{p} + \Delta \mathbf{q}$ to obtain the updated $\mathbf{q}$. Iterations continue until convergence to an energy minimum (using standard optimization criteria). The forward direction (along the imaginary mode eigenvector) is computed first, followed by the backward direction (opposite sign). The two segments are then concatenated at the TS to generate the complete IRC path.

## 1.4. Scope and Overview of the Research

Employing nanoreactor AIMD simulations coupled with an automated trajectory refinement workflow, exothermic formation pathways of nitrogen-containing heterocycles in space have been uncovered. The simulations were applied in a retrosynthetic manner, starting from the desired products to identify high-energy reactants. The resulting trajectories were refined using the automated workflow. This project builds on extensive prior work by a previous

graduate student, Sommer Johansen. [69] Details of the method and results are presented in Chapter 2.

The second part of the research focuses on method development. The original automated refinement workflow had several limitations: 1) the smoothing procedure was prone to producing disconnected paths when changes in internal coordinates were large; 2) the algorithm was compatible only with Q-Chem, a quantum chemistry package that is not publicly available; and 3) data from quantum calculations were stored on local disks in a file tree structure, making navigation and sharing difficult.

The first limitation was addressed by developing a new interpolation method that bypasses the smoothing procedure altogether. This new method performs as well as or better than other widely used interpolation methods, with benchmark results presented in Chapter 3.

To resolve the second limitation, key reaction path-finding methods (IRC and NEB) were implemented in the open-source package geomeTRIC. [65] It optimizes molecular structures by requesting energies and gradients from external quantum chemistry packages and is compatible with several publicly available programs, including Psi4. [70].

To overcome the third limitation, systematic management of NEB input and output data was achieved by integrating geomeTRIC with QCFractal, [71] a cloud-based platform that orchestrates large-scale calculations and stores results on a centralized data server. Chapter 4 describes the integration details, the communication between geomeTRIC and QCFractal, and results from over 3,000 NEB calculations performed using this implementation. [72]. Finally, Chapter 5 summarizes the chapters and discusses future work.

# Chapter 2

# Application: N-heterocycle formation pathway discovery

*The content and figures in this chapter are reprinted or adapted from the author's previously published article: Johansen, S. L.; Park, H.; Wang, L.-P.; Crabtree, K. N., ACS Earth Space Chem. 2024, 8 (9), 1771–1783. This open access article is licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0). The author confirmed the reaction pathways between optimized species from molecular dynamics trajectories obtained and refined by Johansen, S. L., generated the potential energy surface plots, and contributed to the manuscript writing, with a primary focus on the methods section.*

## 2.1. Introduction

In 1828, Friedrich Wöhler published the synthesis of urea, a product of the activities of living organisms, from cyanic acid and ammonia. [73,74] This discovery posed a direct challenge to vitalism, which is the belief that chemical compounds related to "life" can only be produced by living organisms. The surprising disclosure opened a new era for exploring the origin of life. [74] Since the journey began, the origin of nucleobases, heterocyclic compounds containing nitrogen atoms in their rings (N-heterocycles), has been receiving great attention as they are the building blocks of genetic materials of all living organisms. [75,76,77]

Astronomers and astrobiologists have been studying the potential interstellar formation of N-heterocycles. [77] A wide variety of species that contain this motif have been identified on meteorites with non-terrestrial isotopic abundances, [78] and there is evidence that the 6.2 $\mu$m unidentified infrared (UIR) band can only arise from the presence of a nitrogen atom in a polycyclic aromatic hydrocarbon (PAH) structure. [79] However, despite the 124 nitrogen-containing species identified in space (per the Cologne Database of Molecular Spectroscopy [80] at the time of this writing) all astronomical searches for N-heterocycles have been unsuccessful. [81,82,83]

Recently, cyano (CN) substituted cyclic carbon rings have been discovered in the Taurus Molecular Cloud-1 (TMC-1). [84] This cold molecular cloud has temperatures near 10 K and number density of order $10^3$ cm$^{-3}$. Despite these conditions, reactions of aromatics and CN have been shown to proceed rapidly due to a barrierless reaction coordinate. [85] The derived abundances of recently detected aromatic and cyclic molecules greatly exceed the values predicted by chemical models, despite their accurate predictions of acyclic carbon chain abundances. This survey included a search for N-heterocycles pyrrole (c-$C_4H_5N$) and pyridine (c-$C_5H_5N$), but they were not found — the upper limits were $\sim 10^{12}$ cm$^{-2}$, comparable to the detected abundances of indene [86] and cyanonaphthalene [87]. If homocyclic carbon rings and N-heterocycles are formed in such a low-temperature environment through bottom-up chemistry, the pathways must proceed through different classes of reactions or require precursors that exist in abundance for the homocyclic but not heterocyclic pathways. Overall, this indicates that there is a wealth of low-temperature astrochemistry that we currently do not understand.

N-heterocycles are also of significant interest in the study of Titan's atmosphere. The Cassini/Huygens mission made several tentative detections of single-ringed species, along with a wealth of other N-containing complex organic molecules that could be potential N-heterocycle precursors. [88,89,90] Titan is considered an analog for early Earth, [91,92,93] so understanding the role N-heterocycles play in this environment could provide insight into potential terrestrial formation of such prebiotic species.

Ultimately, an understanding of how N-heterocycles participate in meteoritic chemistry, nitrogen-containing polycyclic aromatic hydrocarbon (PANH) formation, and Titan's atmosphere requires inclusion of appropriate formation and destruction pathways into chemical kinetics models. These models seek to replicate the chemistry of a particular environment, and when done comprehensively, can offer accurate predictions of molecular abundances. However, the current literature reports few low-temperature reaction pathways involving

N-heterocycles. Additional reactions must therefore be identified to enable their accurate incorporation into chemical kinetics models.

Computational reaction discovery methods, which automate the generation and testing of mechanistic hypotheses, can play an important role in determining new reaction pathways for N-heterocycle formation and other important astrochemical questions. These methods employ a variety of strategies for hypothesis generation, which include the systematic driving of all possible combinations of reaction coordinates derived from graph-theoretical representations, [94, 95, 96, 97, 98, 99, 100, 101, 102, 103, 104, 105, 106, 107, 108, 109] or carrying out *ab initio* molecular dynamics (AIMD) simulations (or other types of reactive MD) which contain "reaction events" in the simulation trajectory. [46, 110, 111, 112, 113, 114, 115, 116, 117, 118, 119] The generated mechanistic hypotheses are tested by computing the potential energy surface (PES) critical points followed by modeling the reaction rates, which filters the discovered pathways down to the kinetically viable ones. Such approaches have found success in gas-phase chemistry including astrochemically relevant reactions [120, 121], organometallic catalysis [122, 123], surface chemistry [124, 125], photochemistry, [126] and prebiotic chemistry [46, 118, 127, 128, 129, 130].

This study uses the *ab initio* nanoreactor, [46, 111] a reaction discovery method that combines an AIMD simulation with a time-dependent external potential that induces high-velocity molecular collisions and provides the kinetic energy needed to cross over activation barriers on the relatively short AIMD-accessible time scales. An advantage of the nanoreactor is that it does not require user-specified reaction coordinates, and thus enables discovery of reactions within a broader space than one defined by rules; the downside is an increased computational cost and the inability to exhaustively sample all pathways. The nanoreactor is an ideal choice for this work because astrochemical reaction pathways can transcend traditional chemical intuition, as demonstrated by the prevalence of radicals, gas-phase ions, highly unsaturated molecules, carbenes and other exotic species not common to terrestrial

chemistry. A key aspect of this work is that the nanoreactor simulations are carried out to discover the pathways of retrosynthesis [131] — that is, to find endothermic and barrierless *fragmentation* pathways — which correspond to kinetically feasible synthesis pathways of N-heterocycles under ultra-cold astrochemical conditions. The resulting trajectories are processed using an automated refinement workflow, supplemented by reaction path optimization methods such as nudged elastic band (NEB) and intrinsic reaction coordinate (IRC) calculations, to identify reaction events and determine the minimum-energy pathways.

Here we detail this search for pyridine and pyrrole formation mechanisms. First, the nanoreactor simulations and the wide range of N-heterocycle breakdown pathways are described. Then we discuss novel reactions found by expanding upon pathways identified in the nanoreactor simulations: first, a pyrrole forming pathway with $\beta$-cyanovinyl (HCCHCN) and ethenimine ($H_2CCNH$) precursors, followed by two reaction schemes whereby an $n$-membered ring is formed through an insertion reaction into an $(n-1)$-membered ring; specifically, the formation of pyrrole from azete (c-$C_3H_4N$) and cyanomethyl ($H_2CCN$), and the formation of pyridine from 1-pyrrolyl (c-$C_4H_4N$) and cyanomethylene (HCCN).

## 2.2. Methods

The computational approach employed here consists of three stages: "discovery", "selection", and "refinement". The discovery stage involves the use of the *ab initio* nanoreactor method, [46] which employs *ab initio* molecular dynamics (AIMD) simulations at high temperature with an additional time-dependent external potential that induces high-velocity molecular collisions, thereby accelerating reactivity. The trajectories of reactions occurring during the simulation are extracted and a subset of reactions are then selected for further analysis on the basis of their plausibility as astrophysical routes to N-heterocycle formation. The refinement stage includes searching for minimum energy paths (MEPs) for selected reactions discovered in the nanoreactor. Following the refinement stage, the results were further expanded by

carrying out potential energy surface exploration starting from model structures inspired by the nanoreactor results.

To begin the discovery stage, initial configurations of molecules were made using the program Packmol, [132] which populates molecules and atoms of interest within spheres of certain radii without overlapping atoms. These spheres contained either pyridine and CN or pyrrole and CN along with a number of inert He atoms (between 10-50) in various initial configurations (see Table S1 of the Supporting Information (SI)). In principle any inert atom could be used as an energy transfer agent in the simulations; we choose helium to minimize the number of electrons that need to be treated in the ab initio calculations. In total, 405 AIMD simulations were carried out with various sets of the parameters ($R_1$, $R_2$, $t_1$, $t_2$, $k_1$, and $k_2$). Table S2 in the SI presents ranges of these parameters for each labeled simulation. All simulations began with either pyrrole or pyridine and CN, in order to search for mechanisms that result in the desired N-heterocycle with CN as a byproduct. Each starting configuration was energy-minimized [28, 34] prior to starting nanoreactor MD simulations. Both the minimizations and nanoreactor simulations were carried out with the B3LYP hybrid density functional, an unrestricted Kohn-Sham wavefunction and the 6-31G(d) Gaussian basis set, as implemented in the TeraChem [133] quantum chemistry software.

The characteristic feature of the nanoreactor simulation is the time-dependent boundary potential which induces reactivity:

$$V(r,t) = f(t)U(r,r_1,k_1) + (1-f(t))U(r,r_2,k_2) \tag{2.1}$$

$$U(r,r_n,k) = \frac{mk}{2}(r-r_n)^2\theta(r-r_n) \tag{2.2}$$

$$f(t) = \theta\left(\left\lfloor\frac{t}{T}\right\rfloor - \frac{t}{T} + \frac{\tau}{T}\right) \tag{2.3}$$

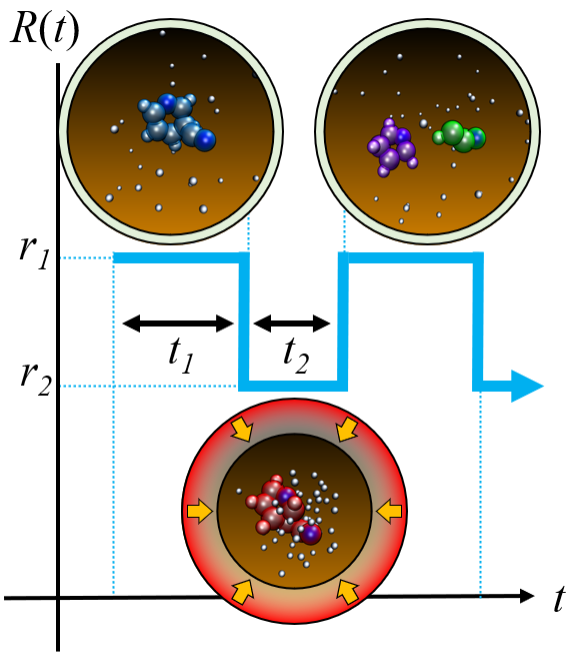



FIGURE 2.1. The visual example of the overall nanoreactor cycle.

where $\theta$ is the Heaviside step function and $\lfloor \cdot \rfloor$ is the floor function. The function $f(t)$ is a rectangular wave that oscillates between 1 (duration $\tau$) and 0 (duration $T - \tau$). Here, $r_1$ is the initial radius of the spherical volume, and $r_2 < r_1$ is the compressed radius. Atoms (mass $m$) located between $r_n$ and the outer boundary experience the harmonic potential $U(r, r_n, k)$ (with force constant $k$), gaining kinetic energy as they are accelerated toward the center. The Heaviside step function in Eq. (2.2) ensures that no potential is applied to atoms inside radius $r_n$. Figure 2.1 provides a visual illustration of a single nanoreactor cycle. The simulation begins in a spherical volume of radius $r_1$. After time $t_1$ ($= \tau$), a compressive shell activates between $r_1$ and $r_2$ for duration $t_2$ ($= T - \tau$), driving captured atoms inward. The volume then expands back to $r_1$, and the cycle repeats until completion. See Figure 2.1 for a visual example of this process.

The time step was set to 0.5 fs and a Langevin thermostat was used with an equilibrium temperature between 1000 - 3500 K and a damping time of 300.0 fs. Table S2 in the SI

details the range of parameters explored for each starting configuration. Each simulation was allowed to run for 24 hours of wall time, yielding several picoseconds of simulation time.

The second stage of this process is the "selection" of reactions to include in the following "refinement" stage. Reaction events in the ab initio molecular dynamics trajectories are identified and refined following the approach of Wang et al. [111] Molecules are represented as connectivity graphs, with covalent bonds assigned between atom pairs separated by distances less than $1.4(R_i + R_j)$, where $R_i$ and $R_j$ are the covalent radii of the respective atoms. The presence or absence of each molecular subgraph is tracked over time, yielding a binary time series for each distinct molecule. Noise arising from high-amplitude vibrations and transient bond stretching is removed using a two-state hidden Markov model, and the most likely sequence of hidden states is obtained via the Viterbi algorithm. Transitions in this denoised sequence reliably identify reaction events. For each detected event, the algorithm traces the involved atoms both forward and backward in time (iteratively expanding the set as needed) to assemble complete and consistent sets of reactant and product molecules. The resulting sub-trajectories, containing the coordinates of these atoms over the reaction interval, are extracted for subsequent refinement to characterize elementary reactions. The results of the trajectory analysis were confirmed by visualization with the VMD program. [134]

Reactions identified in the nanoreactor simulations are then filtered based on their plausibility under astrophysical conditions in an interstellar cloud: low temperature ($\leq 20$ K) and density ($\leq 10^7$ cm$^{-3}$) [135]. The first selection criterion is that the overall reaction must be bimolecular, as three-body collisions are exceedingly improbable in these environments. Similarly, the low density and temperature preclude spontaneous unimolecular rearrangements, so such processes are rejected except those that occur along a bimolecular pathway via submerged barriers. The second criterion is that the overall reaction coordinate is exothermic: the precursors must be higher in energy than the products. Finally, we prioritized reactions

involving precursors that were non-intuitive and/or had not been proposed previously in other studies of N-heterocycle formation.

During the refinement stage, reaction events of interest are further investigated at a more accurate level of theory to determine their kinetic feasibility. The refinement procedure starts with energy minimization of the intermediates extracted from the nanoreactor AIMD trajectory, then transition state (TS) structures on the minimum energy paths connecting the intermediates are located using the nudged elastic band (NEB) method. NEB calculations are initiated by linear interpolation of frames from the nanoreactor simulations connecting two optimized intermediates. The highest-energy structures are then used as inputs to TS optimization calculations to find the precise first order saddle points. Harmonic frequency calculations are performed on all structures to compute the zero point vibrational energy (ZPVE) and confirm the presence of one imaginary mode for the TS structures. The connections between the optimized TS and its adjacent energy basins are verified using the intrinsic reaction coordinate method. The energy minimization, NEB, and harmonic frequency calculations were performed using the range-separated $\omega$B97X-D3 [136] DFT functional, an unrestricted KS wavefunction, and the cc-pVTZ basis set as implemented in the Q-Chem quantum chemistry software. [137] To provide more accurate PES landscapes along the reaction coordinates, single-point energy calculations for all the optimized geometries are carried out using the CCSD(T)/cc-pVTZ level of theory. The ZPEs of CCSD(T) are obtained by adding ZPVEs calculated with $\omega$B97X-D3 DFT functional to the single-point CCSD(T) electronic energies. Special attention was given to pre-reaction complexes and searches for barriers along the entrance channel. The pre-reaction complexes were obtained by optimizing the two reactant molecules using the broken-symmetry approach. Before the optimization, the two reactant molecules were oriented in a way that they have the same relative orientation as the first intermediate. They were separated with a distance larger than 5.00 Angstrom and the optimization was carried with a maximum step size of 0.015 Angstrom.

The initial barriers were then searched through sequential constrained optimizations while scanning along the selected interatomic distance.

## 2.3. Results and Discussion

### 2.3.1. Nanoreactor Simulations

Out of the 405 nanoreactor simulations that were carried out, the simulations started from a single N-heterocycle and one $CN^{\bullet}$ produced the greatest number of bimolecular pathways. In what follows, we will focus exclusively on the results from these simulations and discuss their relevance for astrophysical environments and planetary atmospheres. Generally, the $R_1$ and $k_2$ parameters influenced the simulation outcome most significantly, likely because increasing the expanded sphere size ($R_1$) and force constant of the contracted sphere ($k_2$) both contribute directly to increasing the kinetic energy of molecular collisions.

#### 2.3.1.1. *Pyridine*

A total of 76 pyridine + $CN^{\bullet}$ nanoreactor simulations were carried out using four different initial configurations of the pyridine, CN, and helium atoms. Bimolecular reactions were observed in 19 out of 76 simulations, with several trajectories leading to the same products; the simulation parameters are listed in Table S3 in the SI. Figure 2.2 presents the 16 unique products along with their energies relative to the energy of the reactants, pyridine and CN.

The reaction pyridine + $CN^{\bullet} \rightarrow$ pyrrolyl + HCCN was observed twice in separate simulations. In both cases, $CN^{\bullet}$ first forms a bond with the pyridine C at the ortho C position, then extracts the ortho carbon from the ring; when viewed in reverse, this is a possible mechanism for N-heterocycle ring expansion. This reaction was chosen as an example for further refinement, the results of which are discussed in Section 2.3.2.3.

The other 15 reactions in this group were not further refined in this study, but due to their potential astronomical interest, we describe their general characteristics here. Three of the simulations went energetically downhill and left the pyridine ring intact; two of them are characterized by H transfer from pyridine to $CN^{\bullet}$ resulting in formation of pyridyl radicals **1**

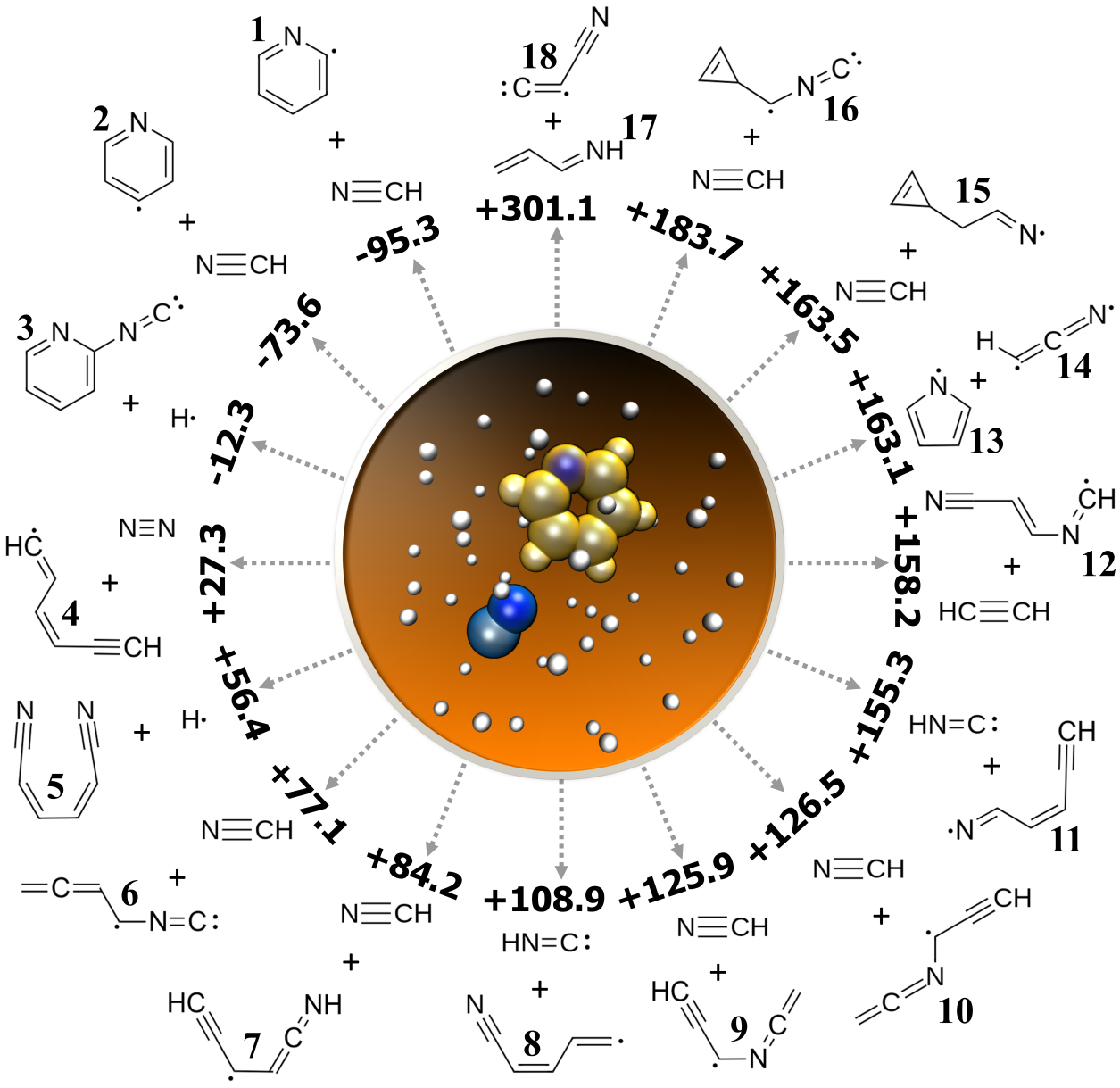



FIGURE 2.2. Products from pyridine + $CN^{\bullet}$ reaction. Electronic energy differences are in kJ/mol.

and **2**, and a third simulation resulted in ortho-isocyanopyridyl **3** + H. The other 12 simulations went energetically uphill and are potential candidates for barrierless, pyridine-forming reactions in reverse. Eight trajectories produced a $C_5H_4N$ radical that was linear (**6-11**) or contained a cyclopropene moiety (**15** and **16**) plus a HCN or HNC coproduct. Closed-shell

molecules have been detected in the ISM with strong structural resemblance to these radicals: most notably, *trans*-cyanovinylacetylene (HCCCHCHCN) [138] differs from **7**, **8** and **11** by the addition of only one hydrogen atom. Additionally, rotational spectra have been measured for three cyanobutadiene isomers, including the parent species of **8**, which could enable further astronomical detections [139]. Other structurally similar molecules found in the ISM include propargyl cyanide ($HCCCH_2CN$) [140], vinylcyanoacetylene ($H_2CCHCCCN$) [138] and the 3-membered rings cyclopropenylidene, [141] cyclopropenone, [142] and ethynyl cyclopropenylidene. Cyclopropenylidene has also been detected in Titan's atmosphere. [89]

The four simulations that went energetically uphill without producing HCN/HNC are described as follows: In one simulation we observed the formation of **5** (1,4-dicyanobutadiene), which differs from the recently detected 1-cyanobutadiene [143] by the addition of one cyano functional group. Another simulation formed the linear radical **12**, a H adduct of 3-isocyanoacrylonitrile, and acetylene $C_2H_2$; [144, 145, 146, 147] the former is a CN adduct of the detected molecule vinyl cyanide. [148] The species **5** and **12** are also similar to other N-substituted unsaturated organic molecules that have been detected in the ISM including isocyanogen (CNCN), [149] E- and Z-cyanomethanimine (HNCHCN), [150, 151] aminoacetonitrile ($H_2NCH_2CN$), [152] and others. [153, 154, 155] One simulation resulted in the pair of species 1-azabutadiene ($H_2CCHCHNH$, **17**) and cyanoethynyl radical ($C_3N$, **18**): the former is an isomer of 2-azabutadiene which was recently characterized spectroscopically and highlighted as a molecule of astrochemical significance [156] and the latter is a known molecule in the ISM. [157, 158] Lastly, one trajectory produced $N_2$ and the unsaturated linear radical **4**, which is a structural isomer of phenyl radical and related to butadienylacetylene ($H_2CHCHCHCCH$) by the removal of H. While butadienylacetylene has not been detected in the ISM to date, it has a similar length and degree of saturation as the known astromolecules vinyl acetylene ($H_2CCHCCH$) [159], allenyl acetylene ($H_2CCCHCCH$) [160] and allenyl diacetylene ($H_2CCCHC_4H$) [161].

Overall, the outcomes of the pyridine + $CN^{\bullet}$ nanoreactor simulations could be characterized in the following ways: (1) The majority of bimolecular products were energetically uphill, indicating they could be candidates for barrierless formation pathways of pyridine + $CN^{\bullet}$ in reverse. (2) A large number of reactions resulted in ring-opening of pyridine to unsaturated linear molecules, which are chemically and structurally very close to molecules detected in the ISM. (3) A smaller number of reactions were energetically downhill which suggest potential reactions of pyridine in the ISM. Interestingly, CN substitution reactions, which have been recently calculated to be barrierless and exothermic [162], were not observed in the nanoreactor simulations. Hydrogen abstraction and NC substitution were observed instead even though these reactions are calculated to have barriers, highlighting the contrast between "reverse" nanoreactor simulations and traditional PES exploration. These results point to multiple possible pathways for pyridine formation from unsaturated linear precursors, and demonstrate the effectiveness of the nanoreactor for generating astrochemical reaction hypotheses.

#### 2.3.1.2. *Pyrrole*

A total of 60 pyrrole + $CN^{\bullet}$ simulations were performed with 6 different initial configurations. Initial minimization resulted in CN bound to the ring: 5 configurations led to binding at the 2- position, adjacent to N (one of which was in an isocyano configuration), and the other was at the 3- position. Among these, 17 of the 60 simulations led to the formation of two products as shown in Figure 2.3 and in Table S4 in the SI. Three simulations led to 1-pyrrolyl (**13**) + HCN/HNC, and the rest led to unique species. Of the reactions observed, 10 were endothermic, meaning that the reverse reactions are energetically favorable for the formation of pyrrole.

Perhaps the two most interesting reverse reactions discovered were ethenimine **26** + $\beta$-cyanovinyl **27** and **14** + 1-H-2-azetyl **31**. **26** + **27** is quite similar to the reaction **27** + vinyl cyanide ($C_3H_3N$) → pyridine + CN that has been observed under single-collision

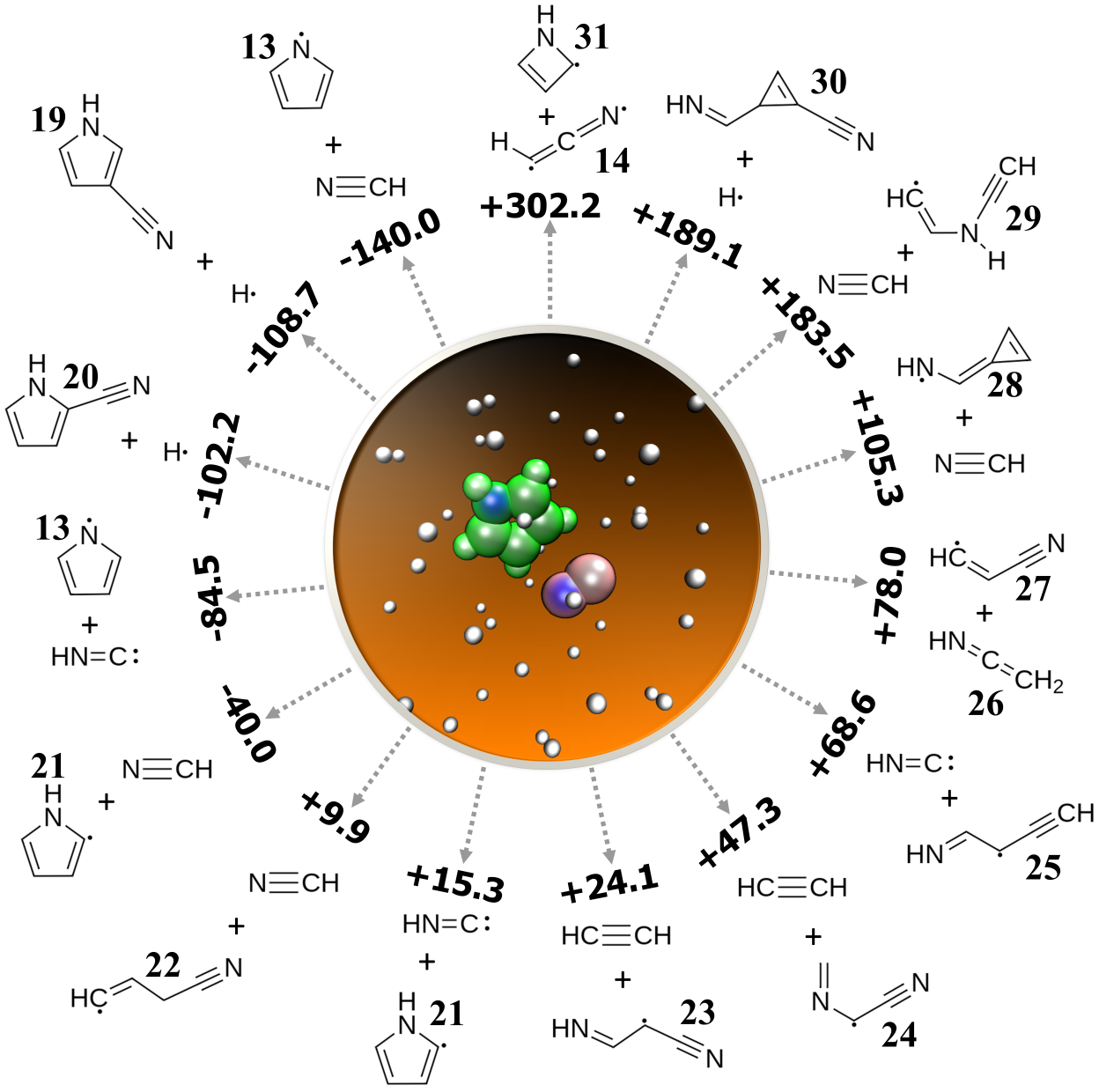



FIGURE 2.3. Products from pyrrole + $CN^{\bullet}$ reaction. Electronic energy differences are in kJ/mol.

conditions in a crossed molecular beam experiment and which was calculated to be barrierless using CCSD(T)/cc-pVTZ//B3LYP/cc-pVTZ. [163] In both cases, the reaction proceeds by addition of the **27** at the $CH_2$ carbon, followed by a rearrangement and cyclization prior to the elimination of CN. The reaction **14** + **31** is a ring expansion reaction similar to the

aforementioned pyridine-forming reaction of **13** + **14**. Both of these pathways were selected for further refinement, discussed in detail below.

Of the remaining pathways, all that proceeded downhill left the 5-membered ring structure intact; they were either hydrogen abstraction reactions leading to HCN/HNC or CN-substitution reactions. Reactions of pyrrole with CN do not appear to have been previously studied, but the analogous formation of benzonitrile from benzene + CN is known to be barrierless, while the H-abstraction channels possess a large barrier. [85] In contrast, only one of the endothermic reactions preserves the ring: the hydrogen abstraction leading to 2-pyrrolyl **21** + HNC. In addition to **31**, two other ring-containing species were found, both forms of functionalized cyclopropene: 3-iminomethyl-1-cyanocyclopropene **30** and cycloprop-2-en-1-ylidenemethanaminyl radical **28**. Neither species has been previously studied; however, the latter is structurally similar to (cyanomethylene)cyclopropane, recently characterized by rotational spectroscopy and an excellent candidate for astronomical searches due to the recent detections of other cyano-substituted species. [164]

The remaining pathways involved breaking the pyrrole structure into acyclic fragments: one of which is closed-shell ($C_2H_2$, HCN, or HNC) and the other is open-shell (**23**, **24**, **25**, and **29**). These open-shell molecules are all related to known astronomical molecules by addition of a small functional group and removal of a hydrogen atom. Two of these are related to ethanimine [165] by a substitution and H-atom removal on the terminal carbon: **23** is functionalized with CN and **25** with $C_2H$, both common functionalization patterns among molecules detected in space. **24** is a CN-functionalized radical derivative of N-methylmethanimine, closely related to methanimine, [166] and **29** is a $C_2H$-functionalized radical derivative of vinylamine. [167] In reverse, each of these pathways presents a potential formation route to pyrrole via cyclization and elimination of CN, and the parent species of **23**, **24**, **25**, and **29** are intriguing targets for astronomical detection. Among these, rotational

spectra for the parent species of **24** [168] and a branched isomer of the parent of **23** [169] have been measured to date.

### 2.3.2. Pathway Refinement and Additional Potential Energy Surface Exploration

Following the initial nanoreactor simulations, three pathways matching the filtering criteria discussed in Section 2.2 were further refined and expanded to explore the mechanisms of the "reverse" reactions and to locate any barrier(s) along the reaction coordinate. For the pyridine + CN simulation, this was the pathway leading to **13** + **14**, and for pyrrole + CN, those leading to **14** + **31** and **26** + **27**. In the case of **14** + **31**, initial exploration of the PES led to identification of a lower-energy bimolecular configuration than the final product in the nanoreactor trajectory. H atom transfer from **31** to **14** leads to azete (c-$C_3H_3N$) + cyanomethyl ($CH_2CN$) approximately 200 kJ/mol lower in energy; this was chosen as the starting point for the detailed PES as discussed further in section 2.3.2.1. The other two pathways were explored starting directly from products identified in the nanoreactor simulations: **26** + **27** → pyrrole + CN is discussed in section 2.3.2.2 and **13** + **14** → pyridine + CN in section 2.3.2.3.

#### 2.3.2.1. *Azete + Cyanomethyl radical*

The PES for the reaction azete + cyanomethyl → pyrrole + CN is shown in Figure 2.4. This calculation was performed on a doublet surface to reflect the spin multiplicity of the products: azete was optimized as a singlet and cyanomethyl as a doublet. Azete is an example of an antiaromatic $4\pi$-electron system like cyclobutadiene, but the nitrogen substitution in azete breaks the $D_{4h}$ symmetry and lifts the degeneracy of the two nonbonding $\pi$ orbitals. Early multireference configuration interaction calculations with a 4-31G basis set found that the ground-state electronic structure of azete is well-described by singlet closed-shell wavefunction with little diradical character owing to the electronegativity difference between the C and N atoms located at opposite vertices of the ring. [170] This is reflected by the much

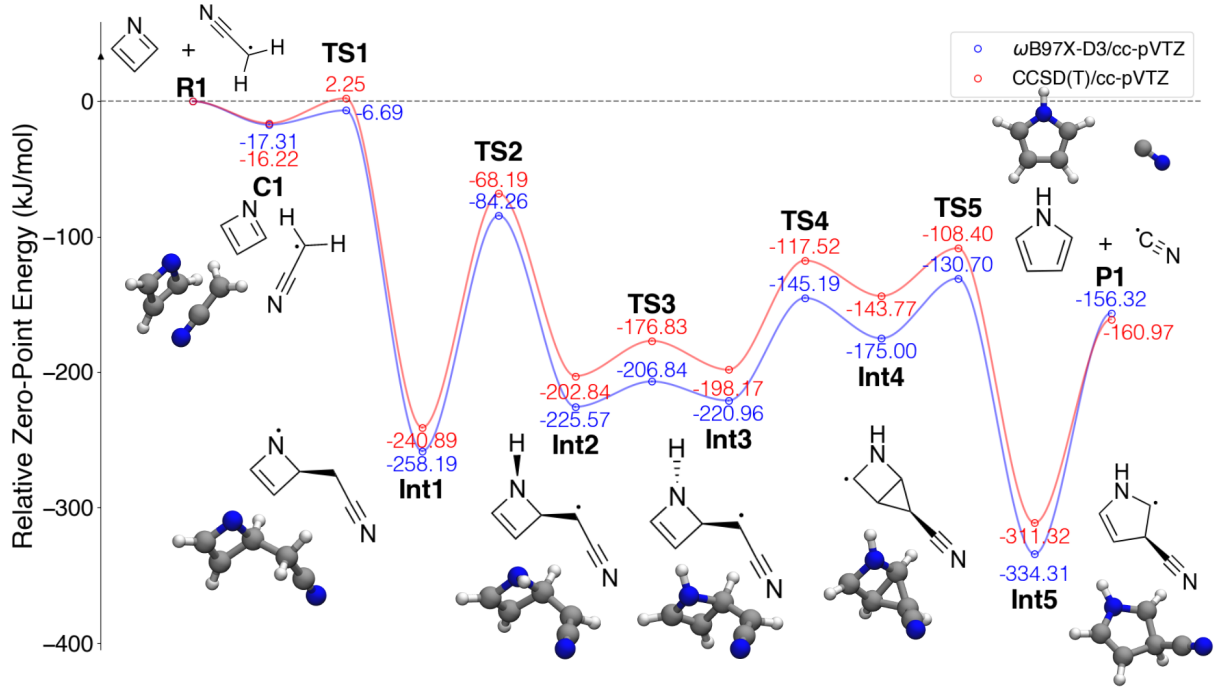



FIGURE 2.4. The PES detailing the reaction of azete (c-$C_3H_3N$) with the cyanomethyl radical ($H_2CCN^\bullet$). Curves connecting points are fitted lines, not an actual shape of the PES.

shorter N=C bond compared with the C=C bond across the ring. A more recent CCSD(T) structural study with basis sets cc-pVTZ, cc-pVQZ, cc-pwCVTZ, and cc-pwcVQZ described azete as a diradical, [171] but details of the electronic structure were not explicitly addressed and the structure is consistent with that of Ref. 170. The ωB97X-D3/cc-pVTZ structure here agrees well with that in Ref. 171, with bond lengths differing by no more than 0.01 Å.

In the reaction, azete and $H_2CCN^\bullet$ first form a pre-reaction complex (**C1**). Both the DFT and CCSD(T) calculations show that **C1** is below the zero-point energy (ZPE) of **R1**. The entrance barrier (**TS1**) calculated with DFT is submerged under the ZPE of **R1**, whereas the CCSD(T) result presents a barrier that is 2.25 kJ/mol above. Through the entrance barrier, **C1** forms a bond between the terminal C of $H_2CCN^\bullet$ and the $\alpha$ C of azete (**Int1**). A H-atom transfer then occurs between the newly bonded C of $H_2CCN$ and the azete N to produce **Int2**. In the next step to **Int3** the N-H bond flips downward, away from the HCCN

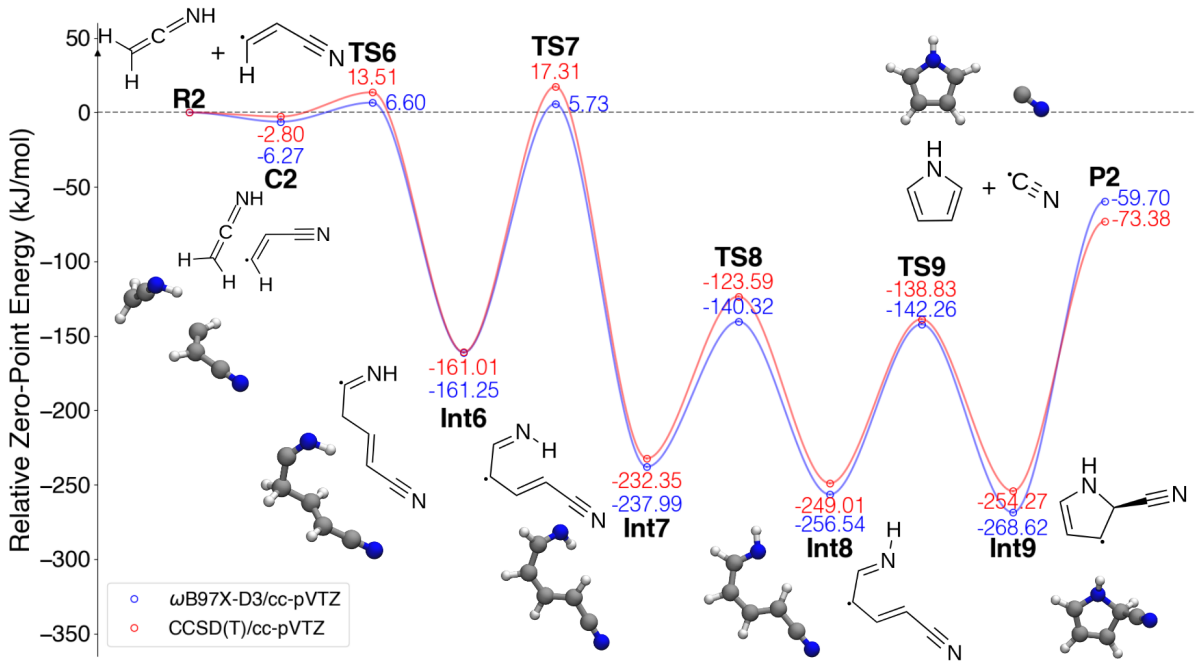



FIGURE 2.5. The PES detailing the reaction of ethenimine($H_2CCNH$) with $\beta$-cyanovinyl(HCCHCN).

group, followed by formation of a second bond between the bonded C in the HCCN chain and the adjacent azete C to make the bicyclic structure **Int4**. Ring expansion leads to the 5-membered ring **Int5**, the cyano-pyrrole radical. The final products pyrrole and $CN^{\bullet}$ are then reached by dissociation.

$H_2CCN$ has been detected in the ISM in TMC-1 and Sgr B2, [172] and is presumed to be present in Titan's atmosphere based on data from the Cassini/Huygens INMS and current chemical models. [173,174,175,176] Azete, on the other hand, has neither been isolated nor detected spectroscopically, potentially due to its high reactivity. Indirect evidence exists for its formation via photolysis of azapyrones in Ar matrices; [177] whether it could be formed via astrochemically relevant gas-phase pathways remains an open question.

#### 2.3.2.2. *Ethenimine + β-cyanovinyl radical*

Figure 2.5 shows the calculated PES for the pathway **26** + **27** → pyrrole + CN. This PES was calculated on a doublet surface with singlet and doublet spin multiplicities for **26** and **27**, respectively. This particular reaction coordinate was discovered by expanding the PES beyond the original nanoreactor trajectory, which proceeded through the formation of a new N − C bond between the reactants and led to several high-energy barriers. The formation of the C − C bond, as shown in Figure 2.5, was manually explored by altering the pathway. First, the intermediates in the nanoreactor pathway that were lower in energy than the reactants were identified. Starting with these structures, bonds were changed iteratively using Avogadro [178] in order to connect the reactants and products. These new intermediates were minimized at uB3LYP/6-31G*, and a Cartesian interpolation of the coordinates was used to generate initial path estimates. These paths were then refined with NEB and TS optimizations as discussed in section 2.2.

Starting from reactants **R2**, the unimolecular intermediate **Int6** is formed by making a C-C bond between the two terminal carbons of the reactants. An initial prereactive complex **C2** followed by a slight barrier **TS6** lies along the entrance channel. Subsequently, a hydrogen migration leads to **Int7** and forces the structure toward planarity owing to an allylic structure. The barrier to this H migration at **TS7** is computed to lie 5.73 and 17.31 kJ/mol above the reactants in the DFT and CCSD(T) calculations, respectively. This is followed by a trans to cis isomerization of the terminal NH to yield **Int8**, followed by ring closure by C-N bond formation (**Int9**) and elimination of $CN^{\bullet}$ to yield pyrrole.

This pathway has a number of similarities to the pyridine formation mechanism from β-cyanovinyl and vinyl cyanide which was calculated to be effectively barrierless at the uB3LYP/cc-pVTZ//CCSD(T)/cc-pVTZ level of theory. [179] In both cases the reaction is initiated with the formation of a C-C bond, the intermediates are all acyclic and unbranched

until ring formation occurs, and the steps with the highest activation barriers involve a hydrogen migration.

Ethenimine has been detected in hot molecular cores (HMCs) within the giant molecular cloud Sagittarius B2 (Sgr B2), where it is believed to form from the tautomerization of methyl cyanide ($CH_3CN$) during shocks. [180] The barrier to tautomerization from ketenimine back to methyl cyanide is calculated to be 262 kJ/mol at the CCSD(T)/aug-cc-pVTZ level, [181] preventing the reverse reaction from occurring spontaneously. An additional formation mechanism of ketenimine is H + cyanomethyl radical ($CH_2CN$), which has been calculated to be barrierless and exothermic. [181] $CH_2CN$ has also been detected in Sgr B2, as well as the Taurus Molecular Cloud 1. [172] Vinyl cyanide, the parent species of $\beta$-cyanovinyl, has been detected in multiple environments in the ISM, [148, 182] including recently within HMC G10.47+0.03. [183] While barrier heights of **TS6** and **TS7** likely inhibit reaction of **26** + **27** in dense molecular clouds where temperatures typically are 10 - 20 K, [184] it may be operative in HMCs as temperatures may reach 300 K. [185]

In the Solar System, vinyl cyanide was detected on Titan in 2017, [90] and the presence of the cyanovinyl radicals is likely because photodissociation processes are rapid in the upper atmosphere. Ethenimine has not been detected directly in Titan's atmosphere but seems likely to be present based on crossed molecular beam experiments of $N(^2D) + C_2H_2$ which saw ethenimine form with a branching ratio between 4 - 5%. [181] At the warmer temperatures of Titan's atmosphere compared with dense clouds, the calculated barrier height for the reaction of **26** + **27** is likely small enough that the reaction is feasible. Additional exploration of the PES and kinetic studies are needed to determine the viability of this pathway in specific environments and its branching ratio relative to other pathways on the PES.

#### 2.3.2.3. *1-Pyrrolyl + Cyanomethylene diradical*

The PES for the reaction **13** + **14** → pyridine + CN is shown in Figure 2.6. The calculation was performed on the doublet surface but unlike the previous two reactions, in this case both

reactants are open-shell. 1-pyrrolyl is well known to be a doublet in its electronic ground state with $\pi$-radical character [186]. HCCN is a carbene which has been experimentally determined to have a triplet electronic ground state and quasi-linear structure. [187, 188] Its HCC bond angle has been calculated to be 145.9° at the CCSD(T)/aug-cc-pVQZ level of theory and the barrier to linearity to be only 0.67 – 0.86 kcal/mol. [189] Here we also observe a quasi-linear structure with an HCC bond angle of 153.8° at the $\omega$B97X-D3/cc-pVTZ level. In the PES, **13** and **14** were calculated with doublet and triplet spin multiplicity, respectively.

From the reactants **R3**, no transition state or pre-reaction complex was along the reaction coordinate leading to **Int10**, in contrast to the previouly discussed reactions. Instead, HCCN bonds directly to the $\alpha$ C in pyrrolyl via the terminal carbon in HCCN. This same carbon atom then forms a second bond to the adjacent pyrrole $\beta$ C to reach the bicyclic structure **Int11**. The bond between the $\alpha$ and $\beta$ C atoms then breaks and the ring widens and flattens to produce the 3-cyanopyridyl radical **Int12**, followed by elimination of CN to form pyridine. The transition states **TS10** and **TS11** are submerged relative to both reactants and products.

This reaction bears similarity to the proposed mechanism of CH insertion into pyrrole. [190] In that case, CH first bonds concertedly to both the $\alpha$ and $\beta$ C atoms in pyrrole, followed by breaking the $C_\alpha$-$C_\beta$ bond and the loss of the nitrogen-bonded H. Because the reaction **13** + **14** starts from pyrrolyl, no H elimination from the N is required. Another similar reaction was detected in the pyrolysis of 3-methoxypyridine, [191] where CO elimination from the pyridoxy radical ($C_5H_4NO$) leads to pyrrolyl. Unlike the insertion reactions with CH and HCCN, however, calculations suggest that cleavage of the O - $CH_3$ bond passes through an overall barrier along the reaction coordinate.

Astronomical detections of HCCN include a definitive detection in circumstellar envelope IRC+10216 and a tentative detection in dense cloud Sgr B2. [192] HCCN has also been suggested to be important in the formation of aerosols in Titan's atmosphere, known as

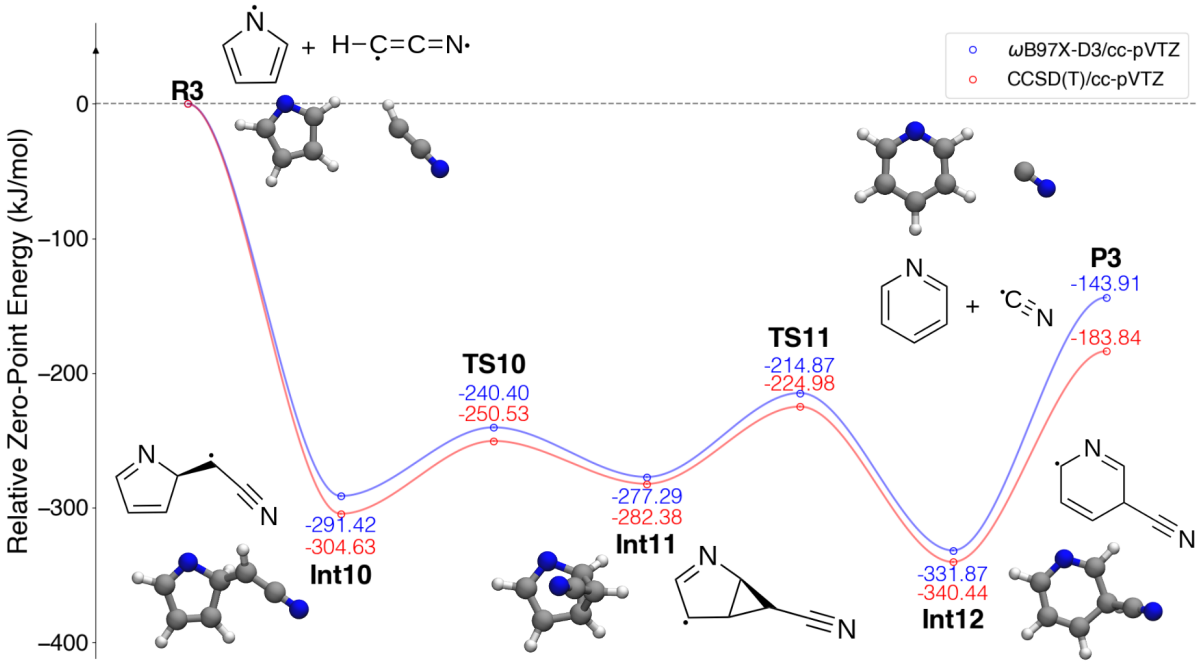



FIGURE 2.6. The PES detailing the reaction of 1-pyrrolyl (c-$C_4H_4N$) with the cyanomethylene diradical ($HCCN^{\bullet\bullet}$).

tholins, that make up the moon's distinctive orange haze. [193] Chemical models [176] predict HCCN to be the third most abundant open-shell species. However, ring insertion/elimination reactions are not currently included in chemical kinetics models of Titan's atmosphere. [175, 176, 194] The 1-pyrrolyl radical has not been measured by rotationally-resolved spectroscopy, precluding astronomical searches to date with radiotelescopes.

## 2.4. Conclusion

Here we have demonstrated the use of an *ab initio* nanoreactor as a tool for the discovery of novel reaction pathways of astrochemical relevance, with a focus on the N-containing heterocycles pyrrole and pyridine. The key aspect of these simulations is that they begin with the heterocycle of interest coupled with an appropriate leaving group, in this case CN, and exploit the nanoreactor's ability to stochastically explore the space of high-barrier/endothermic

pathways to arrive at non-intuitive products. These same pathways, viewed in reverse, serve as potential formation pathways for the heterocycles.

In total, 249 simulations involving pyridine + CN and 156 simulations with pyrrole + CN were performed and analyzed. 19 pyridine and 17 pyrrole simulations resulted in two products. The optimal simulations for the discovery of bimolecular pathways start with only one N-heterocycle and one radical byproduct (in this case, CN) in a spherical volume containing helium atoms. Searching for pathways in this reverse manner requires that a ring breakage occur, but fragmentation into three or more components is avoided. The results suggest that this is best achieved by tuning the parameters that most directly affect the velocities of the sphere contents as they are compressed, notably $r_1$ and $k_2$.

Three pathways have been identified, and the PESs were calculated here at the $\omega$B97X-D3/cc-pVTZ level of theory, with single-point energy calculations also carried out at the CCSD(T)/cc-pVTZ level of theory: **13** + **14** $\rightarrow$ pyridine + CN and **26** + **27** $\rightarrow$ pyrrole + CN which were observed directly in the nanoreactor simulations, and azete + cyanomethyl $\rightarrow$ pyrrole + CN which was inspired by **14** + **31**. The ring expansion reactions **13** + **14** and azete + cyanomethyl are excellent examples of non-intuitive pathways which are barrierless. These are feasible even in low-temperature regions of the ISM if the precursors are initially present, and are likely operative in the atmosphere of Titan where tentative detections of N-heterocycles have been reported. Finally, the reaction **26** + **27** leads to pyrrole from acyclic precursors through a reaction coordinate possessing a small barrier.

This reaction is especially interesting as it represents a potential pathway to N-heterocycle formation directly from acyclic precursors that are likely present in hot cores as well as in Titan's atmosphere, where the temperatures are warmer than cold interstellar clouds.

The *ab initio* nanoreactor approach employed here is readily applicable to other N-heterocycles of interest, such as pyrimidine, purine, imidazole, indole, and others, as well as functionalized derivatives. Additionally, variation of the co-product (CN in the present

study) to, e.g., OH, CH, $C_2H$, or others is worthwhile. Future studies along these lines are envisioned, as well as more detailed computational investigations of the remaining pathways identified in this proof-of-concept study and integration with tools designed for automatic PES exploration.

## 2.5. Supporting Information

Supporting information is available through the journal article.

# Chapter 3

# Translation-rotation internal coordinates with splicing (TRICS): A new approach to molecular geometry interpolation

## 3.1. Introduction

The molecular potential energy surface (PES) defines all of the physicochemical features of a molecular system, such as the stable structures of reactants and products, energetics and rates of reactions, and other thermodynamic and kinetic properties. It is a function defined in $\mathbb{R}^{3N}$, where $N$ is the number of atoms, and maps the nuclear configuration to a potential energy. The Taylor expansion of the PES, including the first-order gradient and second-order Hessian, are used to optimize molecular structures toward critical points that are used to predict reactivity. The transition state (TS) is a first-order saddle point on the PES that sits atop the lowest energy barrier separating reactants from products, while the minimum energy path (MEP) connects the reactants and product structures through the TS at its highest point. The intrinsic reaction coordinate (IRC), a widely used formal definition of the MEP, is given by the steepest descent path in mass-weighted Cartesian coordinates and is everywhere parallel to the instantaneous nuclear acceleration vector. The IRC can also be formulated as a geodesic, or metric-minimizing curve, using a metric tensor that incorporates the PES gradient information. [66, 67, 195, 196] Once the TS and IRC are obtained, reaction rates and mechanisms can be characterized through transition state theory (TST), variational TST, and other methods. [47, 48, 49, 197]

When the TS structure is known, one can trace the minimum energy pathway (MEP) that connects the reactant to the product using the IRC method. This involves taking a small step along the imaginary vibrational mode of the TS and then following the steepest-descent paths downhill in mass-weighted Cartesian coordinates to reach the two energy minima. [66, 67, 195, 196] However, in most cases the TS structure that lies on the MEP

connecting two energy minima is unknown and must be located before the MEP can be traced. Therefore, a range of MEP finding methods have been developed to either refine an initial guessed reaction pathway toward the true MEP or to provide high-quality initial structures for TS optimization.

In most applications, the equilibrium structures of reactant, product, and TS must be located as a prerequisite to calculating reaction rates and other properties. Locating the TS is a challenging calculation that requires a good initial guess of the structure, [198] due to the specific characteristics of the Hessian associated with the correct TS (i.e. one imaginary frequency with the corresponding vibrational mode oriented along the reaction coordinate). Therefore, a range of reaction path finding methods have been developed to characterize the PES along a reaction path approximately and to provide high-quality initial guess structures for TS optimization. The two most widely used MEP finding methods are the nudged elastic band (NEB) [199, 200, 201, 202, 203, 204, 205] and string methods. [206, 207, 208, 209, 210, 211] In addition, various interpolation methods exist for generating a pathway consisting of intermediate structures between two endpoints using only geometric information without any PES information; these can be used to generate input reaction paths for NEB and string calculations, or to characterize the PES in their own right. Linear interpolation of the Cartesian coordinates of the reactant and product has the obvious problem that chemically unreasonable structures and atomic collisions can occur. Performing the interpolation in Z-matrix, or other non-redundant internal coordinates (ICs) is a viable solution, [212] but suffers from the problem that these IC systems are only well-defined in a limited neighborhood, and may break down if the endpoints are substantially different, such as during a reaction; however, it works very well for small geometry changes, and has been successfully incorporated into freezing string reaction path optimization. [213]

The linear synchronous transit (LST) method improves on linear Cartesian interpolation by fitting the coordinates of the intermediate structures (images) to linearly interpolated

interatomic distances between the endpoints. [214] The "Nebterpolator" approach expands on LST by using linearly interpolated values of redundant internal coordinates as the fitting target, including angles and dihedral angles. [111] These methods use least-squares fitting because there are more interatomic distances and redundant internal coordinates than Cartesian degrees of freedom, therefore the target values cannot be satisfied exactly, and undesirable "jumps" in the interpolated path may occur if multiple minima exist in the objective function. Another class of successful methods utilize the NEB approach: the image-dependent pair potential (IDPP) method formulates the objective function as a potential and optimizes the path by applying the NEB method to this potential. [215] More recently, Zhu and Martinez (ZM) demonstrated that a geodesic path connecting two structures on a Riemannian manifold serves as an excellent initial guess for the MEP finding methods. [216] They constructed the required metric tensor from a set of redundant internal coordinates defined by a Morse-like scaling of interatomic distances that penalizes unphysical atomic collisions. [216]

Recent machine learning (ML) models have been developed to directly predict TS structures from reactant and product geometries (or even 2D molecular graphs). [217, 218, 219, 220] Because these models are trained on MEP data, they can bypass the need for an interpolation step in the workflow. Among these, models that use linearly interpolated structures between reactants and products both as input features during training on known TS geometries and as the reference for inference on new reactant–product pairs have demonstrated promising performance. [219, 220] One potential avenue for further improvement of these ML models is to improve the featurization step using higher-quality interpolated structures as input. Furthermore, some classes of reactions do not necessarily pass through first-order saddle points, such as barrierless reactions and ultrafast photoreactions that evolve on the excited state PES. [221, 222] For these reasons, geometry-based methods for structure interpolation will continue to be an important part of the computational chemist's toolkit.

The relation between Cartesian coordinates and ICs is captured by the Wilson B-matrix $\mathbf{B}$, where $B_{ij} = \partial q_i / \partial x_j$. [55] By defining $\mathbf{G} = \mathbf{B}\mathbf{B}^T$, the displacement in Cartesian coordinates, $\Delta\mathbf{x}$, can be calculated from the displacement in ICs $\Delta\mathbf{q}$ as follows:

$$\Delta\mathbf{x} = \mathbf{B}^T\mathbf{G}^{-1}\Delta\mathbf{q} \tag{3.1}$$

For geometry optimizations of structures given in Cartesian coordinates, it is common to build the IC system starting from automatic identification of the chemically relevant bond stretches, valence angles, and dihedral angles from molecular connectivity; we call these individual functions primitive internal coordinates (PICs). [57] When the number of PICs exceeds the internal degrees of freedom (3N-6 or 3N-5), redundancies arise. Pulay and Fogarasi pioneered systematic methods to treat this redundancy, [58, 59, 60] while Schlegel and co-workers popularized direct optimization in redundant internal coordinates (RICs) using the generalized inverse of the $\mathbf{G}$ matrix. [61] Baker et al. later introduced delocalized internal coordinates (DLCs), obtained by diagonalizing the $\mathbf{G}$ matrix at the initial geometry to produce a fixed non-redundant working set. [62, 63, 64] More recently, Wang et al. extended the DLC approach to multi-fragment systems by including new types of PICs for translation and rotation of molecules and fragments, called TRIC (translation-rotation internal coordinates). [65]

One can interpolate between the two endpoints using any non-redundant IC system that spans the Cartesian degrees of freedom, such as DLCs derived from the union of PICs from the reactant and product structures. [212] However, this union-based DLC interpolation is not fully robust, especially when PICs well-defined (e.g., proper angles) in one endpoint become ill-defined or degenerate (e.g., near-linear angles) in the other. Such mismatches can lead to numerical instabilities in the Wilson B-matrix, thereby generating intermediates with unphysically distorted geometries. The primary goal of this work is to overcome the limitations of IC interpolation by using multiple non-redundant IC systems to generate interpolated

paths in their respective valid domains followed by "splicing" the paths together. Our implementation of the new method, called TRICS (Translation-Rotation Internal Coordinate with Splicing), works by partitioning a given pathway into one or more segments according to the configuration spaces where different IC systems are well-defined. The splicing is done iteratively by exchanging target structures between the segments and re-interpolating. The TRICS implementation is available in the open-source geometry optimizer geomeTRIC. [65]

In this manuscript we evaluate the performance of TRICS with a reference dataset of MEPs consisting of 500 randomly selected reactions from the RGD1 dataset, [109] 123 reactions from barrier height quantum chemistry benchmarks (BH), [223] and 56 organic reactions from *ab initio* nanoreactor simulations. [46] In addition to TRICS, several other published interpolation methods were included for comparison: the geodesic interpolation of Ref. 216 (ZM), IDPP, [215] LST, [214] and DLC interpolation. [212] TRICS showed improved performance in path interpolation across the tested benchmarks. It successfully generated evenly spaced frames for all 679 reactions, with the smallest root-mean-square difference (RMSD) fluctuations between adjacent images. TRICS also produced the highest-quality paths, achieving the lowest excess energy of the highest-energy image relative to the optimized reference TS structures across all three datasets. In addition, TRICS successfully interpolated a barrierless path connecting the Franck-Condon point to a MECI, whereas alternative methods introduce an unnecessary barrier. TRICS further demonstrated the capability of extrapolating beyond the given endpoints, a feature not currently supported by NEB-based approaches such as IDPP and ZM.

## 3.2. Methods

### 3.2.1. TRICS Interpolation Method

The intuitive idea behind the TRICS method is to carry out interpolation in one or more patches of configuration space that are each parameterized by its own nonredundant IC

system. The segments are spliced together to ensure smoothness of the interpolated path in the overlapping regions.

We first review the basic features of the IC system, specifically the delocalized internal coordinate (DLC) system that is used in this work. The individual internal coordinates (ICs) $q_i$ are:

$$q_i(\mathbf{x}) = \sum_{j=1}^{N_{\text{cart}}} c_{ij} p_j(\mathbf{x}) \tag{3.2}$$

where $p_j, 1 \leq j \leq N_{\text{prim}}$ form a set of *primitive ICs*, i.e. scalar functions of the Cartesian coordinates such as an individual interatomic distance, angle, dihedral angle, or component of translation / rotation. The set of primitive ICs is determined from the connectivity graph of the initial structure $\mathbf{x}_{\text{init}}$. In TRIC, six translation and rotation PICs are added for each fragment or residue to provide a more natural description of intermolecular degrees of freedom, while the generation of intramolecular ICs follows previous work. The coefficients $c_{ij}$ are the components of the $i$-th eigenvector of the G-matrix of the primitive ICs $\widetilde{\mathbf{G}}$ evaluated at $\mathbf{x}_{\text{init}}$:

$$\widetilde{\mathbf{G}}(\mathbf{x}_{\text{init}}) = \widetilde{\mathbf{B}}\widetilde{\mathbf{B}}^T(\mathbf{x}_{\text{init}}) \tag{3.3}$$

where $\widetilde{\mathbf{B}}$ is the $N_{\text{prim}} \times N_{\text{cart}}$ Jacobian matrix, also called the Wilson B-matrix, relating the PIC and Cartesian coordinates:

$$\widetilde{B}_{ij}(\mathbf{x}_{\text{init}}) = \frac{\partial p_i}{\partial x_j}(\mathbf{x}_{\text{init}}) \tag{3.4}$$

The B and G matrices of the DLC system, denoted as $\mathbf{B}$ and $\mathbf{G}$, are calculated as:

$$\begin{aligned} \mathbf{B}(\mathbf{x}) &= \mathbf{c}\widetilde{\mathbf{B}}(\mathbf{x}) \\ \mathbf{G}(\mathbf{x}) &= \mathbf{B}\mathbf{B}^T(\mathbf{x}) \end{aligned} \tag{3.5}$$

The goal of interpolation is to produce a continuous and smooth function in Cartesian coordinates $\boldsymbol{\chi}(\mathbf{x}_0, \mathbf{x}_1, t_0, t_1; t)$ where $t$ is the reaction coordinate parameter, and $t_0 = 0, t_1 = 1$ are the initial and final values of $t$ corresponding to $\mathbf{x}_0$ and $\mathbf{x}_1$. The function values at the

reaction endpoints are $\boldsymbol{\chi}(\mathbf{x}_0, \mathbf{x}_1, 0, 1; 0) = \mathbf{x}_0$ and $\boldsymbol{\chi}(\mathbf{x}_0, \mathbf{x}_1, 0, 1; 1) = \mathbf{x}_1$. It is intuitive to think of function $\boldsymbol{\chi}$ as being defined by the first four arguments and the chosen IC system, with a single independent variable $t$. An IC displacement $\Delta\mathbf{q}$ between $\mathbf{x}_0$ and $\mathbf{x}_1$ may be computed as:

$$\Delta\mathbf{q}(\mathbf{x}_1, \mathbf{x}_0) \simeq \mathbf{q}(\mathbf{x}_1) - \mathbf{q}(\mathbf{x}_0) \tag{3.6}$$

where the $\simeq$ indicates that displacements of certain PICs such as dihedral angles and the rotation of fragments need to take periodicity into account. Next, $\Delta\mathbf{q}$ is subdivided into a number of intervals as $\delta\mathbf{q} = \Delta\mathbf{q}/n$, then a discretization of the reaction path is generated starting with one of the endpoints as:

$$\begin{aligned}
\boldsymbol{\chi}(\tfrac{1}{n}) &= \mathbf{x}_0 + \Delta\mathbf{x}(\mathbf{x}_0, \delta\mathbf{q}) \\
\boldsymbol{\chi}(\tfrac{2}{n}) &= \boldsymbol{\chi}(\tfrac{1}{n}) + \Delta\mathbf{x}(\boldsymbol{\chi}(\tfrac{1}{n}), \delta\mathbf{q}) \\
&\cdots \\
\boldsymbol{\chi}(1) &= \boldsymbol{\chi}(\tfrac{n-1}{n}) + \Delta\mathbf{x}(\boldsymbol{\chi}(\tfrac{n-1}{n}), \delta\mathbf{q})
\end{aligned} \tag{3.7}$$

where $\boldsymbol{\chi}(\cdot)$ is shorthand for $\boldsymbol{\chi}(\mathbf{x}_0, \mathbf{x}_1, 0, 1; \cdot)$. At the end of the interpolation, $\boldsymbol{\chi}(1) = \mathbf{x}_1$. The interpolated path is a straight line in the internal coordinate parameter space, thus it is naturally a geodesic using the metric $d(\mathbf{x}, \mathbf{y}) = |\mathbf{q}(\mathbf{x}) - \mathbf{q}(\mathbf{y})|$. The Cartesian displacement $\Delta\mathbf{x}$ is calculated from the current structure $\mathbf{x}$ and the IC displacement $\delta\mathbf{q}$ using the following iterative algorithm

$$\begin{aligned}
\mathbf{x}^{[0]} &\equiv \mathbf{x}; & \delta\mathbf{q}^{[0]} &\equiv \delta\mathbf{q}; \\
\mathbf{x}^{[i+1]} &= \mathbf{x}^{[i]} + \mathbf{B}^T\mathbf{G}^\dagger(\mathbf{x}^{[i]}) \cdot \delta\mathbf{q}^{[i]}; & \delta\mathbf{q}^{[i+1]} &= \delta\mathbf{q}^{[i]} - \Delta\mathbf{q}(\mathbf{x}^{[i+1]}, \mathbf{x}^{[i]})
\end{aligned} \tag{3.8}$$

The iterations converge when $|\Delta\mathbf{q}(\mathbf{x}^{[i+1]}, \mathbf{x}^{[i]})|$ falls below a threshold such as $10^{-6}$, then $\Delta\mathbf{x} = \mathbf{x}^{[i]} - \mathbf{x}^{[0]}$. The calculation of $\mathbf{B}$, $\mathbf{G}$, $\Delta\mathbf{q}$ and $\Delta\mathbf{x}$ are routine components of geometry optimization software packages that utilize internal coordinate systems.

A nonredundant IC system has a finite "domain of validity" in the configuration space $\mathcal{D} \subset \mathbb{R}^{3N}$ in which the ICs form a basis that spans the Cartesian coordinates. Naturally, $\mathcal{D}$ contains the Cartesian coordinates used to define the IC system $\mathbf{x}_{\text{init}}$, and it contains a sufficiently large neighborhood around $\mathbf{x}_{\text{init}}$ such that geometry optimization algorithms can take steps to reach the critical points on the PES efficiently. However, for reaction path interpolation, the reactant and product structures $\mathbf{x}_0$ or $\mathbf{x}_1$ can differ significantly in their spatial coordinates and bonding, and accordingly there is no guarantee that $\mathcal{D}$ of the IC system based on $\mathbf{x}_0$ extends to $\mathbf{x}_1$ or vice versa. For configurations / coordinate sets outside of the domain, there is no longer a one-to-one mapping between Cartesian and internal coordinates, and the problem of finding the Cartesian coordinates from an IC displacement becomes underdetermined. An illustrative example is that a displacement in the bond angle can be uniquely mapped back to Cartesian coordinates, but when the bond becomes linear, two orthogonal displacements are needed to map back to Cartesian coordinates. Also, any dihedral angle A—B—C—D becomes undefined if the atoms A—B—C or B—C—D become linear. When these problems occur, the final structure in the sequence of interpolated structures in Eq. 3.7 can be significantly different from the final structure, i.e. $\boldsymbol{\chi}(\mathbf{x}_0, \mathbf{x}_1, 0, 1; 1) \neq \mathbf{x}_1$ and the calculation fails to produce the desired result.

In general, there is no guarantee that both reaction endpoints lie in the valid domain of any one nonredundant IC system. TRICS, short for "translation-rotation internal coordinates with splicing", addresses this problem by generating multiple path segments using IC systems that are valid in different subdomains of the configuration space, which are then spliced together; the procedure is inspired by the physical technique of splicing two ropes or tapes together by overlapping their ends. Suppose the neighborhoods of $\mathbf{x}_0$ and $\mathbf{x}_1$ are exclusively within the domains of IC systems $IC_0$ and $IC_1$ respectively, and the domains have an overlapping region in configuration space (Figure 3.1a). The first segment uses the metric of $IC_0$, begins at $\mathbf{x}_0$, and ends at a "free point" $\mathbf{x}_a$ corresponding to $t_a$; it is denoted

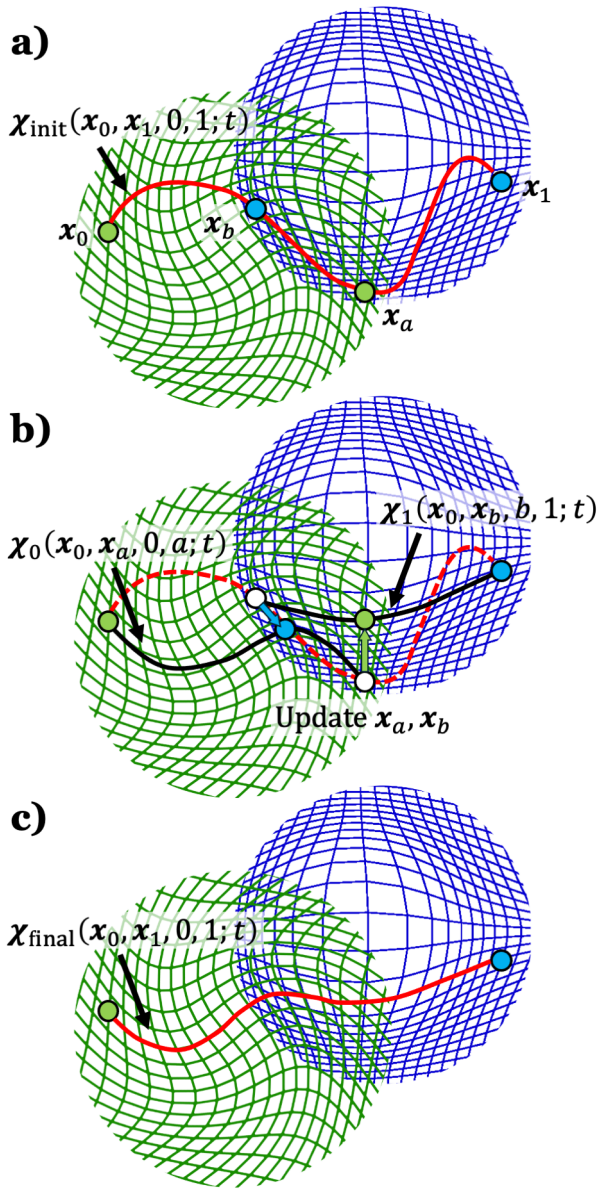



FIGURE 3.1. Illustration of the TRICS splicing algorithm. The green and blue grids represent the isoparametric curves and valid domains of $IC_0$ and $IC_1$ respectively. a) The initial pathway is shown in red, and points $\mathbf{x}_a$ and $\mathbf{x}_b$ on the initial path are chosen. b) The interpolated path segments $\boldsymbol{\chi}_0$ and $\boldsymbol{\chi}_1$, shown in black, are geodesics (straight lines) in the parameter space of $IC_0$ and $IC_1$ respectively, but the segments do not overlap. New $\mathbf{x}_a$ and $\mathbf{x}_b$ points are chosen for the next splicing iteration. c) The final spliced pathway shown in red is continuous and smooth, and its constituent segments are geodesics in the parameter spaces of their respective IC systems.

as $\boldsymbol{\chi}_0(\mathbf{x}_0, \mathbf{x}_a, 0, t_a; t)$. The other segment uses the metric of $IC_1$, begins at a "free point" $\mathbf{x}_b$ corresponding to $t_b < t_a$, and ends at $\mathbf{x}_1$; it is denoted as $\boldsymbol{\chi}_1(\mathbf{x}_b, \mathbf{x}_1, t_b, 1; t)$ Therefore, $\boldsymbol{\chi}_0$ and $\boldsymbol{\chi}_1$ represent ranges of the reaction coordinate parameter $0 \leq t \leq a$ and $b \leq t \leq 1$, where the "overlap region" is given by $b \leq t \leq a$.

The TRICS splicing algorithm aims to solve for the endpoints $\mathbf{x}_a$ and $\mathbf{x}_b$ such that the path segments $\boldsymbol{\chi}_0$ and $\boldsymbol{\chi}_1$ overlap in configuration space. The overlap is ensured by requiring each segment to contain the free endpoint of the other one, i.e. $\boldsymbol{\chi}_0(\mathbf{x}_0, \mathbf{x}_a, 0, t_a, t_b) = \mathbf{x}_b$ and $\boldsymbol{\chi}_1(\mathbf{x}_a, \mathbf{x}_1, t_b, 1, t_a) = \mathbf{x}_a$. This is carried out using an iterative algorithm, in which the endpoint of each path is updated to the image with the same value of $t$ on the other path (Figure 3.1b):

$$\begin{aligned} \mathbf{x}_a &\leftarrow \boldsymbol{\chi}_1(\mathbf{x}_b, \mathbf{x}_1, t_b, 1; t_a) \\ \mathbf{x}_b &\leftarrow \boldsymbol{\chi}_0(\mathbf{x}_0, \mathbf{x}_a, 0, t_a; t_b) \end{aligned} \tag{3.9}$$

The iterations terminate when the displacement between the new and updated values go below a threshold. The final interpolated pathway (Figure 3.1c) can be obtained by concatenating the segments as:

$$\boldsymbol{\chi}_{\text{final}}(\mathbf{x}_0, \mathbf{x}_1, 0, 1; t) = \begin{cases} \boldsymbol{\chi}_0(\mathbf{x}_0, \mathbf{x}_a, 0, a; t) & 0 \leq t \leq \dfrac{a+b}{2} \\ \boldsymbol{\chi}_1(\mathbf{x}_b, \mathbf{x}_1, b, 1; t) & \dfrac{a+b}{2} < t \leq 1 \end{cases} \tag{3.10}$$

$\boldsymbol{\chi}_{\text{final}}$ is a continuous and smooth path at the point where $\boldsymbol{\chi}_0$ and $\boldsymbol{\chi}_1$ are joined because in the limit of a small overlap region ($(a - b) \to 0$), $\boldsymbol{\chi}_0$ and $\boldsymbol{\chi}_1$ are well approximated by a straight line segment in $\mathbb{R}^{3N}$ connecting $\mathbf{x}_a$ and $\mathbf{x}_b$.

#### 3.2.1.1. *Algorithmic Details*

In practice, the number of ICs required to cover the full configuration space along the entire path from $\mathbf{x}_0$ to $\mathbf{x}_1$ is unknown because reaction pathways vary widely in complexity. A single IC often suffices for relatively simple reactions, whereas multiple ICs are needed for more complex ones. To enable TRICS to identify the path segments corresponding to the

valid domains of a chosen IC system and the overlap regions needed to splice them together, an initial pathway must be supplied. We have developed a sufficiently reliable procedure to generate an initial pathway that does not have the geodesic property, detailed in the SI (Section 1: Initial Reaction Pathways). TRICS can use any initial path as input, but initial paths with chemically reasonable primitive ICs along the path lead to the best results.

Once the initial path is available, an IC system is generated for each image along the path using TRIC. Each IC is then evaluated for its numerical stability in all of the other images, resulting in a $N \times N$ matrix of diagnostic criteria $\mathbf{D}$, where $D_{ij}$ predicts the numerical stability of the IC system generated from image $i$ at the Cartesian coordinates of image $j$. The assignment of diagnostic criteria is described in the SI, Section 2 and an example of the diagnostic matrix is shown in Figure 3.10. Lower values of $D_{ij}$ indicate better numerical stability of $IC_i$ at the image $\mathbf{x}_j$; 0 means good stability, 1 means acceptable stability and 2 or higher indicate potential instability. If any row of the matrix has all low values, that indicates the corresponding IC system is numerically stable through the whole pathway and there is no need for splicing; otherwise, the IC segments are determined by finding overlapping sub-rows of $\mathbf{D}$ with low values. The algorithm for finding overlapping sub-rows can be illustrated by navigating a small square sub-matrix, or "car", through the "maze" defined by the areas of $\mathbf{D}$ with low diagnostic values (Figure 3.10); the car starts at the top left and moves across a row, and changes directions to move down a column when an obstacle is encountered. If the bottom row of the matrix or another obstacle is encountered, the car changes direction again to move across the row, and this procedure repeats until the right edge of the matrix is found. The IC segments are then defined as the horizontal segments of the path through the matrix. Since the diagnostic matrix tends to take lower values close to the diagonal, it should always be possible to find a path through the matrix consisting of rightward and downward segments, and the procedure works well for all of our tested reactions.

Within each segment, IC systems corresponding to the segment endpoints as well as the interior are chosen as candidates for interpolation. After the interpolations are completed for the segments, the next splicing iteration begins by replacing the endpoint coordinates of the current segment with the coordinates on the overlapping neighboring segment with the same value of $t$, then re-interpolating the segments. The iterations converge when the maximum displacement of any atom in the endpoint structures on all segments falls below $1.8 \times 10^{-3}$ Å – this threshold was chosen as it is a typical value used for geometry optimization.

When atom clashes are detected along the interpolated trajectories, a new IC system is created that contains an added PIC corresponding to the inverse distance of the clashing pair of atoms and the segment is re-interpolated. The detailed distance threshold used to determine clashing atoms is provided in the SI, Section 1. If the interpolation fails to reach the target structure, the segment is further split into two separate segments. The interpolation terminates without success if a segment becomes too small or remains unresolved after three splits. In all reactions tested in this work, the TRICS algorithm reached successful convergence without triggering the failure criterion.

### 3.2.2. Computational Details

All quantum chemistry calculations in this study were performed at the B3LYP/6-31G(d) level of theory [28, 34, 224, 225] as implemented in Q-Chem quantum chemistry software package. [226] Molecular geometry optimizations and IRC calculations were carried out using geomeTRIC. [65] The TRICS algorithm was implemented in geomeTRIC (version 1.2.0 or newer). LST, IDPP, and DLC interpolations were performed with the open-source software Pysisyphus, [212] while the ZM geodesic interpolation code was obtained from the Supplementary Materials of Zhu et al. [216] An example Q-Chem input file used in this study, together with the optimized structures of reactants, transition states, products, and all interpolated trajectories, are provided as XYZ files in the SI. All three-dimensional molecular structure visualizations were generated using VMD. [134]

| Set | Description | Number of reactions |
|---|---|---|
| BH76 | Atom transfer, nucleophilic substitution, unimolecular and association reactions | 19 |
| BHPERI | Pericyclic reactions | 27 |
| BHDIV10 | Diverse reactions | 8 |
| INV24 | Inversion/racemization processes | 19 |
| BHROT27 | Single bond rotations | 19 |
| PX13 | Proton-exchanges | 13 |
| WCPT18 | Proton-transfer in uncatalyzed and water-catalyzed reactions | 18 |
| | Total | 123 |

TABLE 3.1. Descriptions of the seven BH subsets and the number of reactions used in this study.

### 3.2.3. Interpolation Endpoints

The RGD1 dataset contains over one million neutral, singlet-state organic reactions, each involving up to 10 heavy atoms. [109] TS structures for 500 randomly selected RGD1 reactions were optimized to locate first-order saddle points, followed by IRC calculations. The two IRC endpoints were further optimized and used as the interpolation endpoints. Reactions for which the ICs of the optimized reactant and product structures showed no change (i.e., the two endpoints converged to the same structure) were discarded and replaced until a final set of 500 reactions was obtained. The BH dataset consists of seven subsets. [223] Details of the subsets and the number of reactions used from each in this study are provided in Table 3.1. The 123 BH reactions and the 56 reactions from the NR56 (originally discovered via Nanoreactor AIMD simulations) [46] were processed using the same procedure. For the BH and NR56 reactions, the IRC paths were visually inspected to confirm that the endpoints converged to reasonable energy minima.

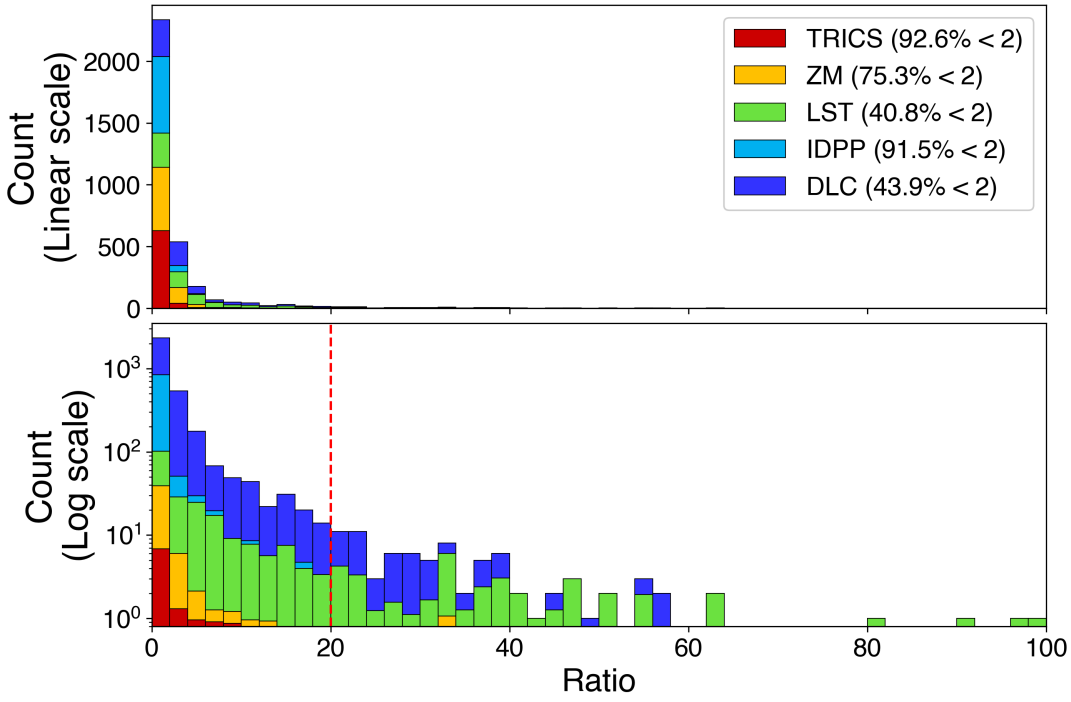



FIGURE 3.2. The top histogram displays the distribution of the ALH values for all 3395 interpolated pathways (679 reactions × 5 methods). The bottom histogram shows the same distribution on a logarithmic scale, with the y-axis lower limit set to 0.8 to make single counts visible. In both panels, the bars are stacked and color-coded by each method. Within each bin, the height of each colored segment is proportional to that method's percentage of pathways in the bin. The legend displays, next to each method's name, the percentage of its pathways that have ALH smaller than 2. A vertical red dashed line on the bottom histogram indicates the threshold ALH value of 20, below which the interpolated trajectories are deemed successful. 11 pathways with ALH values than 100 are omitted from the plot.

## 3.3. Results

### 3.3.1. Benchmark Results

#### 3.3.1.1. *Interpolation Success Rates*

The interpolated trajectories were evaluated for evenness to determine the success rates of each method. Chemically infeasible interpolated trajectories often exhibit large structural gaps between adjacent images. To quantify this, arc length heterogeneity (ALH) was calculated as $\frac{\max(\text{RMSD}(x_i,x_{i+1}))}{\text{median}(\text{RMSD}(x_i,x_{i+1}))}$, where $x_i$ denotes the Cartesian coordinates of the $i$-th image.

| | RGD1 (500) | BH (123) | NR56 (56) | Total (679) |
|---|---|---|---|---|
| TRICS | 100% (1.36±0.43) | 100% (1.56±1.18) | 100% (1.33±0.68) | 100% (1.39±0.66) |
| ZM | 100% (1.70±0.78) | 96.7% (5.55±16.14) | 100% (2.28±1.68) | 99.4% (2.45±7.07) |
| IDPP | 100% (1.34±0.77) | 100% (1.29±1.47) | 100% (1.35±0.67) | 100% (1.33±0.93) |
| DLC | 95.6% (5.09±7.00) | 97.6% (3.18±6.63) | 87.5% (8.73±10.88) | 95.0% (5.04±7.45) |
| LST | 94.6% (5.88±8.56) | 69.9% (20.51±44.54) | 92.9% (8.14±12.68) | 91.5% (8.72±21.40) |

TABLE 3.2. Overall interpolation success rates for each method are reported. For each dataset, the number of reactions is indicated in parentheses next to the dataset name. Averages and standard deviations of the ALH values are provided in parentheses.

Figure 3.2 shows histograms of the ALH values for all three datasets, presented both linear and logarithmic scales. Interpolated paths with ALH below 20 were considered successful. The energies of these trajectory images were then computed to assess their qualities, as discussed in the following section. Table 3.2 reports the overall success rates, together with the average and standard deviation of the ratios across the three datasets. TRICS, ZM, and IDPP exhibit high success rates ($> 99$ %), with TRICS and IDPP also exhibiting particularly low ALH. Interpolation using DLC also yielded reasonably high success rates (95.0 %). In contrast, the LST method showed the lowest success rates (91.5 %) across all three datasets, primarily due to a limitation in solving the objective function. As LST minimizes the function for each image, progressing from the reactant to the product, the landscape of the search space morphs, allowing the possibility for a local minimum to evolve into the global minimum during the interpolation process. This leads to abrupt jumps in the calculated interatomic distances and large structural discontinuities. Nebterpolator, which uses redundant ICs instead of interatomic distances for in its objective function, suffers from the same problem. [111] This disconnectivity issue of LST is especially pronounced in reactions involving large changes in bond angles because LST relies heavily on interatomic distances. Of the 27 BH reactions that LST failed to interpolate, 17 belonged to the INV24 subset and 7 to the BHROT27 subset. [223] These subsets contain reactions involving molecular angular motions, such as inversion, racemization, and rotation. [223]

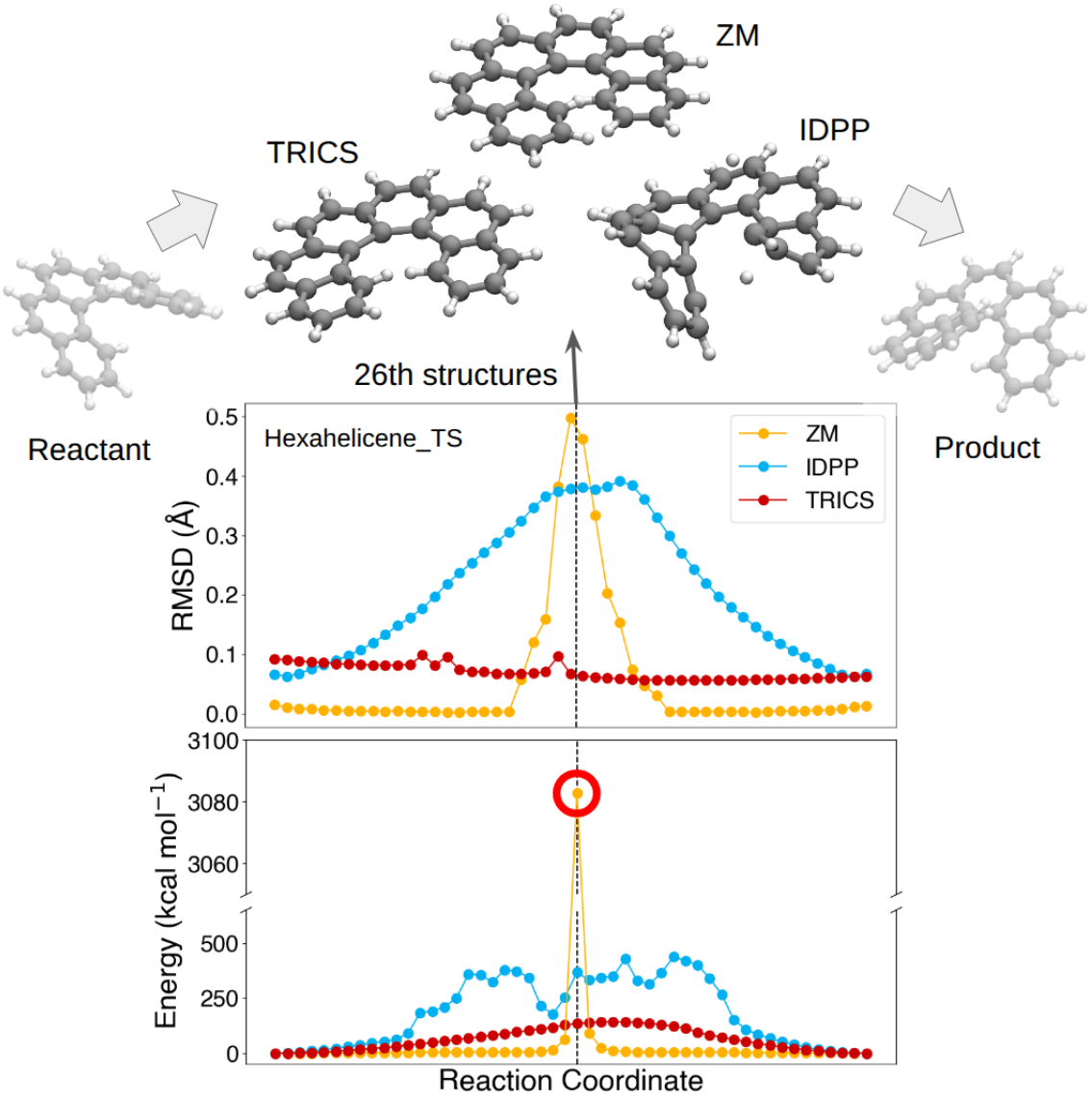



FIGURE 3.3. A plot of RMSD values and energies calculated from trajectories generated by three methods for the Hexahelicene_TS reaction in the INV24 subset. The 3D structures at the top correspond to the 26th image from each method. Atoms are colored as follows: carbon (gray) and hydrogen (white).

The four BH reactions in which ZM failed to produce reasonably continuous trajectories belong to the INV24 subset, and are characterized by sudden, large jumps in RMSD near the TS region. Here, we focus on the Hexahelicene_TS reaction as a representative example and compare trajectories from TRICS and IDPP. Figure 3.3 shows the RMSD values and relative

energies of the images along the path. The RMSDs from TRICS do not fluctuate much and stays near 0.1Å, whereas those form IDPP gradually rise and fall, and the ZM path exhibits a spike in the middle. The corresponding 3D structures provide further insight into the methods' performances. The reactant and product structures are related by a mirror reflection, and the ZM result includes a planar structure in the middle (image 26) where two of the H atoms are very close together, resulting in a energy of >3000 kcal/mol above the endpoints. Meanwhile, IDPP prevents superposition by detaching the two hydrogen atoms and allowing them to drift near the molecule until the product is reached, resulting in an oscillatory energy profile around 500 kcal/mol above the endpoints. In contrast, TRICS successfully generated intermediate structures in which the two hydrogen atoms remained separated from each other while appropriately bonded to their respective carbon atoms, thereby yielding a smooth energy profile. During the TRICS interpolation, multiple clashes were detected between atom pairs, which led to the addition of inverse distance PICs and splitting of the path into multiple segments; the final interpolated path consisted of 5 segments.

#### 3.3.1.2. *Analysis of excess energy above the barrier*

The quality of an interpolated path can be assessed by how closely it reproduces the barrier height of the true MEP. Here, we evaluate the excess energy, $\Delta E_{exs}$, defined by subtracting the optimized TS energy from the highest energy along the interpolated path. A smaller excess energy indicates a high quality interpolated path with maximum energy close to the TS. Some cases of negative excess energy, caused by torsion angle changes along the interpolated pathways in the opposite direction from those in the dataset (as visualized in Figure 3.11). These cases were excluded from further analysis (see SI Section 3: Negative Excess Energy Cases for details). Also, interpolated trajectories with ALH above 20 were excluded from the analysis because they fail to approximate a continuous path.

Figure 3.4a shows the distributions of the excess energies for the five methods across all datasets. TRICS showed the best performance across all the datasets, as evidenced by the

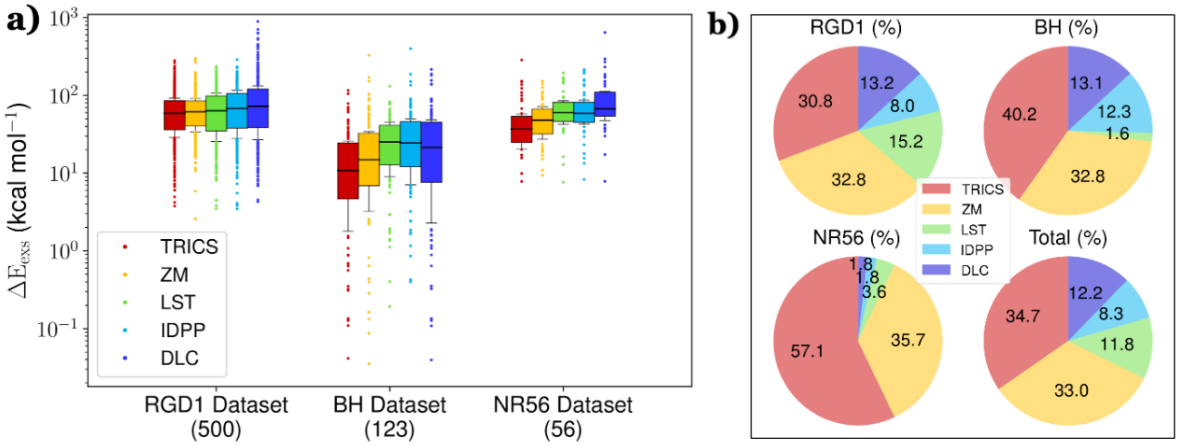



FIGURE 3.4. a) The distributions of excess energy, calculated using B3LYP/6-31G(d), between the optimized transition state structures and the highest-energy structures from the interpolated trajectories are presented as box plots on a logarithmic scale. The solid boxes span from the 25th to the 75th percentile, with a black horizontal line representing the median. The whiskers stretch from the box to the farthest data point within 15 % of the interquartile range from the each edge of the box. Points beyond that distance are shown as individual outliers. A few RGD1 results from DLC showed excess energies exceeding $10^3$ kcal/mol and were excluded from the plot. The total number of reactions for each dataset is shown in parentheses. Reactions with a negative overestimation excess energy (see SI Section 2: Negative Excess Energy Cases) or uneven trajectories were excluded. b) The pie charts illustrate the percentage of reactions for which each method was the best performer.

mean values, followed by ZM and IDPP. Although the mean values for LST in the RGD1 and NR56 datasets are lower than those of IDPP, this does not necessarily indicate better performance for LST, since interpolated trajectories with unevenly distributed images were excluded from the analysis. Similarly, DLC shows a smaller mean value than LST and IDPP in BH due to the exclusion of poor-quality trajectories. We also identified the best-performing method for each reaction by comparing $\Delta E_{exs}$ values across all five methods. Using this analysis, three pie charts were generated for each dataset, along with one for the total, as shown in Figure 3.4b. ZM claimed the largest percentage for the reactions in RGD1, while TRICS accounted for the largest percentages in the other two datasets.

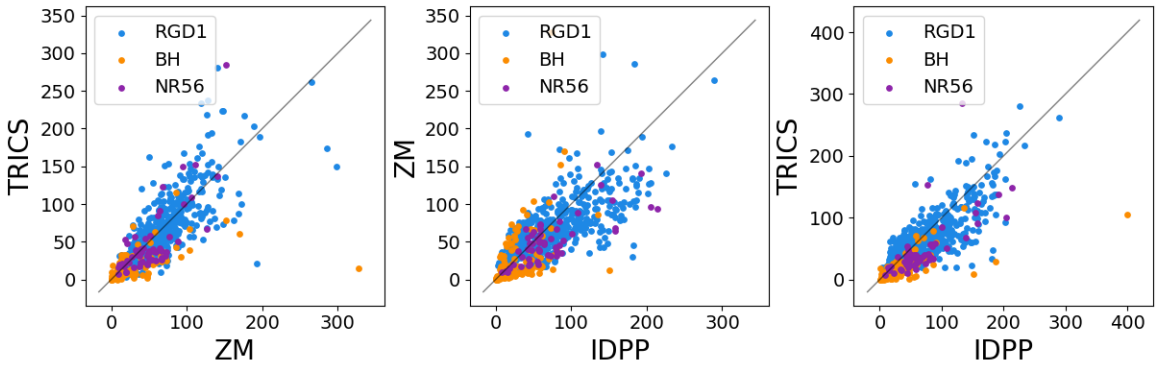



FIGURE 3.5. Three pairwise comparison plots between TRICS, ZM, and IDPP across all three datasets. Each axis shows $\Delta E_{exs}$ (in kcal/mol) computed from trajectories of the respective method. Points falling below the y=x diagonal indicate superior performance of the method on the y-axis.

Overall, TRICS performs best, although the large number of reactions where other methods outperformed TRICS indicates that there is still room for improvement.

Based on the interpolation success rate and the energy analysis, the three best-performing methods (TRICS, ZM, and IDPP) were selected for a closer comparison in the following section.

#### 3.3.1.3. *Detailed comparison of TRICS, ZM, and IDPP*

In this section, we compare the performances of the three best-performing methods (TRICS, ZM, and IDPP) based on trajectory evenness and the energy of the highest-energy images. Figure 3.5 presents pairwise comparisons between the three methods. In each plot, the method with the lower $\Delta E_{exs}$ in the majority of reactions is plotted on the y-axis, so there are more points below the y-axis than there are above. The actual counts of data points above and below the trend line are listed in Table 3.3. This analysis shows that TRICS has the lower barrier for the majority of reactions in a head-to-head comparison vs. ZM or IDPP for the reactions tested. More significantly, for all of the pairwise comparisons, distribution has a significant spread and there are many points far from the diagonal in either direction.

| | **TRICS vs ZM** | | **ZM vs IDPP** | | **TRICS vs IDPP** | |
|---|---|---|---|---|---|---|
| | Above | Below | Above | Below | Above | Below |
| RGD1 | 237 | 263 | 209 | 291 | 193 | 307 |
| BH | 44 | 74 | 38 | 80 | 23 | 99 |
| NR56 | 22 | 34 | 13 | 43 | 8 | 48 |
| Total | 303 | 371 | 260 | 414 | 224 | 454 |

TABLE 3.3. Number of data points lying above and below the y=x line in the scatter plots shown in Figure 3.5. A greater number of data points below the diagonal line indicates better performance of the method plotted on the y-axis.

This means that no single method is uniformly better than the others and there is significant room for improvement in all three.

To investigate cases where TRICS significantly outperformed ZM and vice versa, the differences in $\Delta E_{exs}$ of TRICS ($\Delta E_{exs}^{TRICS}$) and that of ZM ($\Delta E_{exs}^{ZM}$) were calculated as $\Delta\Delta E_{exs} = \Delta E_{exs}^{TRICS} - \Delta E_{exs}^{ZM}$. The ten reactions with the largest positive $\Delta\Delta E_{exs}$ values—corresponding to cases where TRICS performed substantially worse than ZM—were selected, as these highlight opportunities for methodological improvement. Conversely, the ten reactions with the largest negative $\Delta\Delta E_{exs}$ values, where TRICS significantly outperformed ZM, were also examined to understand the factors contributing to its success. These reactions, along with their $\Delta E_{exs}$ values for both methods and the resulting $\Delta\Delta E_{exs}$, are listed in Table 3.4. Reactions whose names begin with "MR" belong to the RGD1 dataset, whereas the reaction labeled "48" originates from the NR56 dataset. All remaining reactions are from BH, with names formatted as "subset_TS structure name". For brevity, the ten reactions with the most positive $\Delta\Delta E_{exs}$ values are denoted P1–P10 (in descending order), and the ten reactions with the most negative $\Delta\Delta E_{exs}$ values are denoted N1–N10 (in ascending order).

The reactants, intermediates, and products for the interpolated trajectories of reactions P1–P10 are shown in Figure 3.6. Due to the absence of an interatomic potential in TRICS, fragments in several trajectories are positioned farther apart than ideal (most notably in

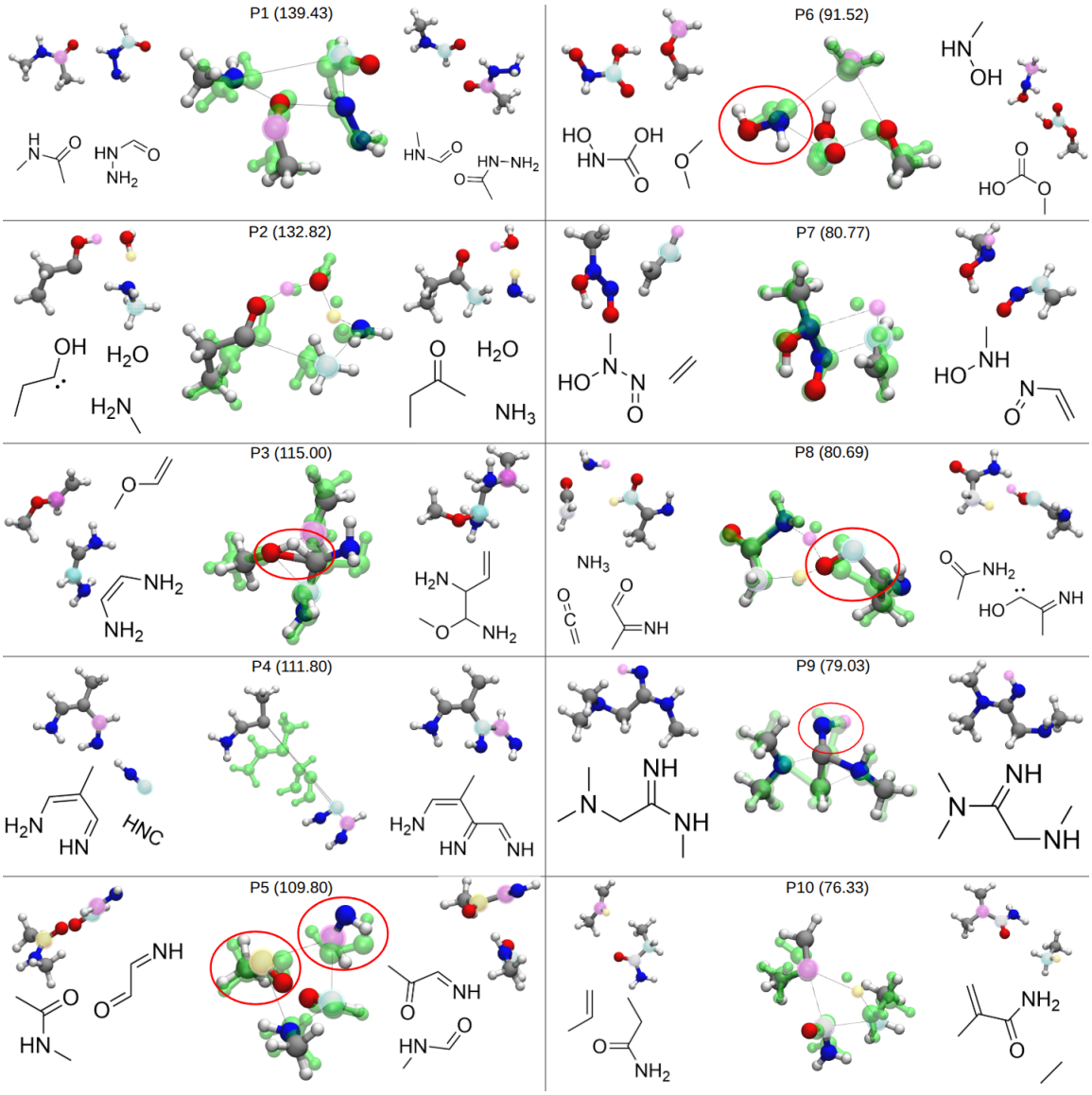



FIGURE 3.6. The ten reactions (P1–P10) in which ZM outperformed TRICS. The $\Delta\Delta E_{exs}$ values (in kcal/mol) are given in parentheses next to each reaction label. For each case, the left and right panels show the reactant and product, respectively, while the central structure directly below the reaction label is the corresponding intermediate structures. For each reaction, the two central intermediates are taken from the same frame number (between the 18th and 30th) of the respective ZM and TRICS interpolated trajectories. The ZM structures are highlighted in green, and the corresponding TRICS structures are shown in standard colors. Atom coloring: carbon (gray), hydrogen (white), oxygen (red), and nitrogen (blue). Selected carbon and hydrogen atoms are additionally color-coded to facilitate visual tracking of the molecular transformation.

| Reactions | $\Delta E_{exs}^{TRICS}$ (kcal/mol) | $\Delta E_{exs}^{ZM}$ (kcal/mol) | $\Delta\Delta E_{exs}$ (kcal/mol) |
|---|---|---|---|
| MR_632778_2 (P1) | 280.23 | 140.79 | 139.43 |
| 48 (P2) | 285.07 | 152.25 | 132.82 |
| MR_697497_2 (P3) | 234.12 | 119.13 | 115.00 |
| MR_379969_2 (P4) | 162.30 | 50.50 | 111.80 |
| MR_467165_1 (P5) | 237.65 | 127.86 | 109.80 |
| MR_617340_1 (P6) | 218.55 | 127.04 | 91.52 |
| MR_651726_0 (P7) | 151.02 | 70.25 | 80.77 |
| MR_434983_1 (P8) | 120.98 | 40.29 | 80.69 |
| MR_37630_1 (P9) | 154.11 | 75.08 | 79.03 |
| MR_150793_2 (P10) | 222.92 | 146.59 | 76.33 |
| INV24_Methinecyanine_TS (N1) | 15.44 | 327.39 | -311.96 |
| MR_119249_1 (N2) | 21.48 | 192.40 | -170.92 |
| MR_333248_2 (N3) | 150.04 | 298.41 | -148.37 |
| MR_547193_0 (N4) | 173.26 | 285.63 | -112.37 |
| BH76_hoch3fts (N5) | 60.76 | 169.48 | -108.73 |
| MR_348404_0 (N6) | 77.95 | 161.52 | -83.57 |
| MR_88942_1 (N7) | 87.84 | 168.08 | -80.25 |
| MR_128523_2 (N8) | 73.95 | 147.70 | -73.75 |
| BHPERI_13ts_8a (N9) | 78.45 | 152.10 | -73.65 |
| MR_161275_0 (N10) | 100.59 | 172.07 | -71.47 |

TABLE 3.4. The table lists the ten reactions with the largest positive $\Delta\Delta E_{exs}$ values and the ten with the largest negative values, where $\Delta\Delta E_{exs}$ is defined as $\Delta E_{exs}^{TRICS} - \Delta E_{exs}^{ZM}$. Among these 20 reactions, only one (labeled "48") is from NR56, and those with names starting with "MR" are from the RGD1 dataset. The remaining reactions are from the BH dataset, with names beginning with the subset followed by the TS structure name.

P1, P2, and P4). In contrast, the intermediate structures obtained with ZM (highlighted in green) keep the fragments at reasonable distances. Similarly, in P7 and P10, the hydrogen atom transferred by TRICS travels through empty space, whereas the ZM pathway brings the fragments close together, enabling a direct "hand-off" of the hydrogen. The P3 intermediate from TRICS places the oxygen atom excessively close to both the hydrogen and its attached carbon. This also stems from the absence of interatomic forces that would otherwise maintain optimal interatomic distances. Another clear difference appears in the torsional and rotational degrees of freedom (P5, P6, P8, and P9). For example, the two

circled fragments in P5 differ by approximately 180° in their relative rotational orientation. Cases P6, P8, and P9 exhibit clear torsion-angle discrepancies, which likely stem from the linear interpolation of internal coordinates in TRICS. This can drive dihedral angles along the longer, higher-energy path. For instance, in P9, ZM correctly leaves the dihedral angle containing the –NH group unrotated, while TRICS performs an unnecessary full rotation.

The other ten cases (N1–N10), where TRICS performed better than ZM, are shown in Figure 3.7. When the reaction involves rotation of a -$CH_3$ group, ZM temporarily breaks the tetrahedral coordination of the central carbon, making it planar, before restoring the tetrahedral geometry to reach the product (N1, N2, and N5). Because ZM has no force that preserves the H–C–H angles within the methyl group, it allows these angles to become planar. In other cases, ZM either disrupts the planar shape of a fragment (N3) or flattens a fragment that should remain non-planar (N4, N10) during the transformation. Another common issue is the fluctuation of H–C–H angles in ZM intermediates: it fails to maintain reasonable H–C–H angles during the transformation and excessively pinches these angles (N6–N9), which increases the energy. The results from Figures 3.6 and 3.7 are consistent with expectations, because ZM primarily focuses on maintaining appropriate interatomic distances that enables smooth fragment transfer, whereas TRICS linearly interpolates internal coordinates while preserving chemically reasonable bond angles and dihedral angles throughout the reaction pathway.

Finally, the evenly interpolated trajectories from the three methods were refined toward the MEPs using our high-throughput NEB workflow. [72] All NEB records contributing to these statistics are publicly accessible via a Python script that queries the database with an internet connection, and example script is provided in the SI. Of the total 2037 trajectories ((500 RGD1 + 123 BH + 56 NR56 reactions) × 3 methods), seven were excluded: four cases from the ZM method with uneven image spacing and three additional cases (one from each method) in which the highest-energy image had a lower energy than the optimized reference

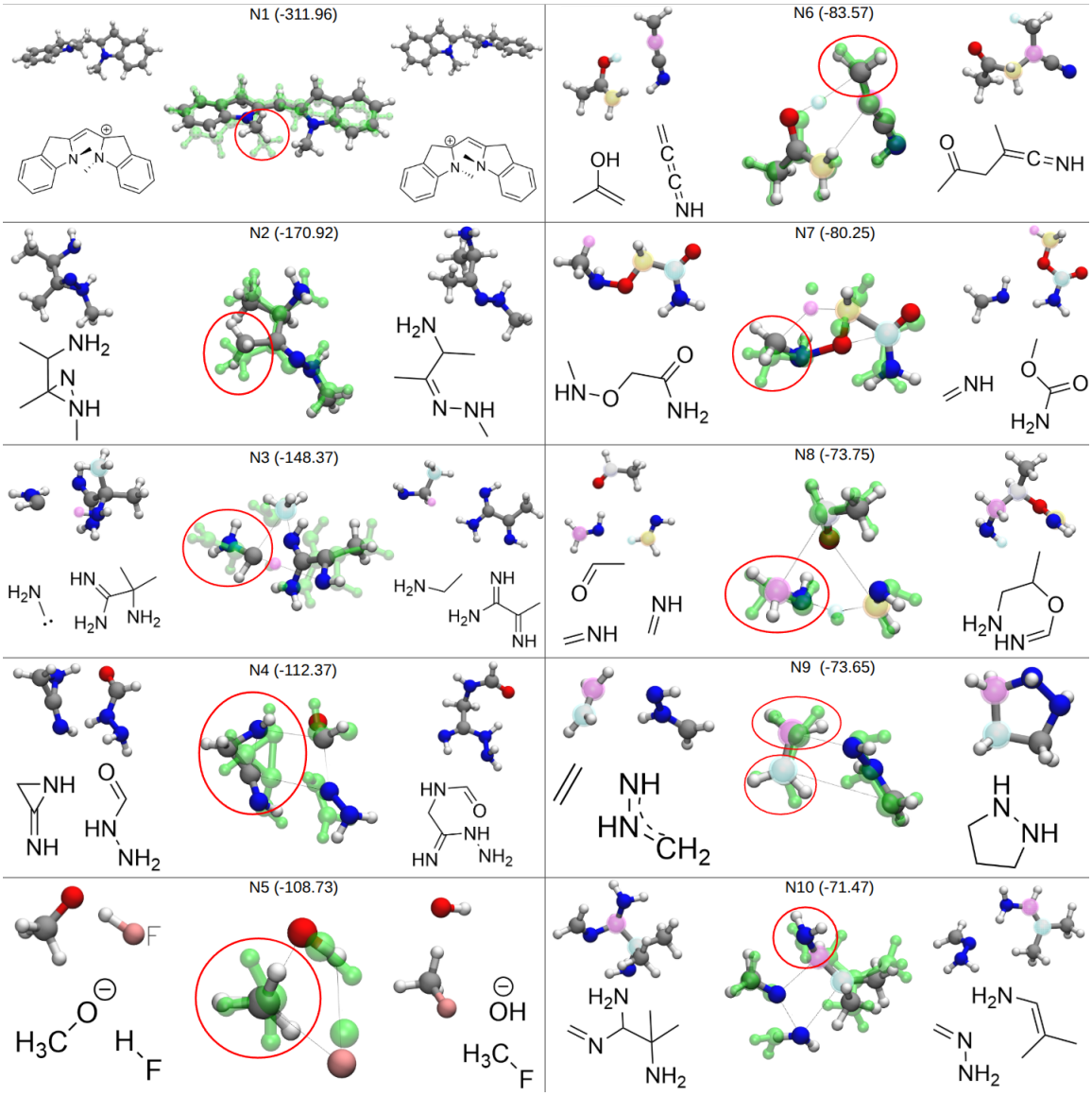



FIGURE 3.7. The ten reactions (N1–N10) in which TRICS outperformed ZM. See Figure 3.6 for other details.

TS structure. The remaining 2030 trajectories were optimized with the hybrid band type and 21 images. Convergence was achieved when both the average and maximum RMS Cartesian gradients for all images fell below $4.86 \times 10^{-4}$ Hartree/Bohr and $9.72 \times 10^{-4}$ Hartree/Bohr, respectively (equivalent to 0.025 and 0.05 eV/Å), within a maximum of 300 iterations.

| | TRICS | ZM | IDPP |
|---|---|---|---|
| **RGD1** | 101.7 ± 43.0 (94.2 %) | 98.9 ± 40.7 (96.4 %) | 105.8 ± 46.2 (94.6 %) |
| **BH** | 38.8 ± 26.6 (99.2 %) | 39.5 ± 25.6 (98.3 %) | 42.6 ± 37.2 (98.4 %) |
| **NR56** | 85.2 ± 26.0 (92.9 %) | 87.2 ± 31.1 (98.2 %) | 92.9 ± 37.0 (92.9 %) |
| **Total** | 88.6 ± 46.2 (95.0 %) | 87.3 ± 43.9 (96.9 %) | 93.0 ± 50.2 (95.1 %) |

TABLE 3.5. Number of iterations required for convergence of the interpolated trajectories using the HT-NEB method (convergence rates shown in parentheses).

Table 3.5 reports the average and standard deviation of the number of iterations required to reach convergence (considering only the cases that converged within the maximum number of 300 iterations) together with the corresponding convergence rates. All three methods achieved high convergence rates (> 90 %), across the three datasets. On the RGD1 dataset, ZM outperformed the others, achieving both the lowest average number of iteration and the highest convergence rate. On the other hand, TRICS required the fewest iterations on the BH and NR56 datasets, although ZM still exhibited the highest convergence rate on NR56. Overall, ZM delivered the lowest average iteration count and the highest convergence rate, primarily due to its strong performance on RGD1, which accounts for the majority of reactions (73.64 % of the total).

### 3.3.2. Applications in characterizing barrierless reactions and excited state potential energy surfaces

The potential energy surface has many diverse types of features that go beyond first-order saddle points, especially in the context of photochemistry. Such studies often involve identifying the molecular structures of the conical intersections, and characterizing the potential energy curves that connect the Franck-Condon point to the conical intersections. A barrierless potential energy curve often implies an ultrafast photoreaction mechanism and efficient population transfer between the excited and ground states. [221, 222] Alternatively, for applications where strong fluorescence is desired, a steep barrier or uphill potential energy

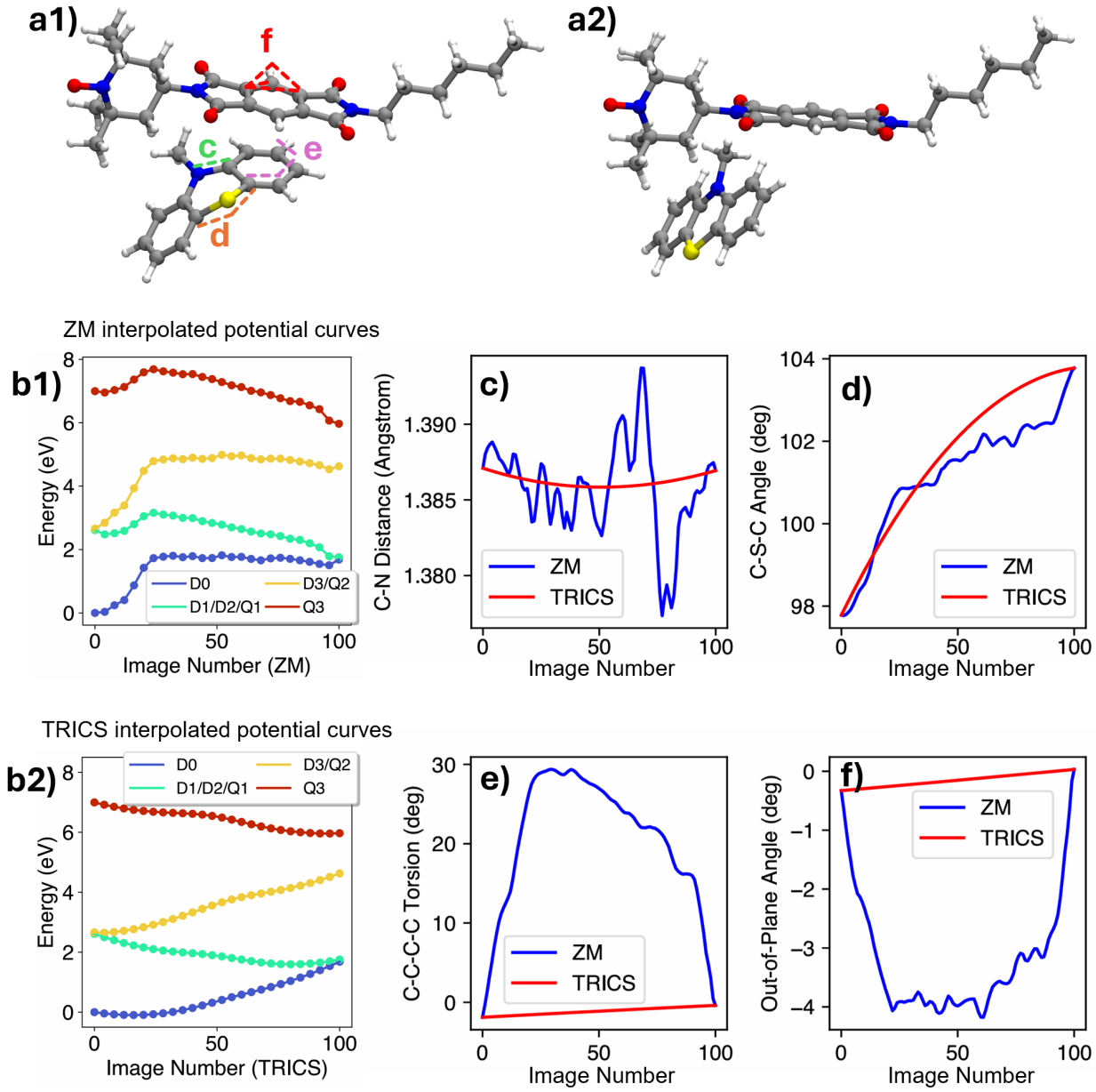



FIGURE 3.8. A comparison of the TRICS and ZM interpolation methods for excited state potential energy curves. a1): The optimized structure at the Franck-Condon point. The primitive ICs that are plotted in parts "c" through "f" are labeled. a2) The optimized MECI structure. b1) and b2): The excited state PECs along the interpolated path generated using the ZM and TRICS methods respectively. The green and yellow curves contain degeneracies between the states indicated. c) through f): Plots of selected primitive coordinates along the interpolated path.

curve is preferable to suppress the non radiative transition pathway and prolong the excited state lifetime. In this section, we will present two examples to demonstrate how the new methods offer new capabilities for photochemical studies.

In the first example, the TRIC interpolation is used as high-performance screening method for barrierless reactions, since it only requires single point energy calculations along the pathway. In particular, if the interpolated potential energy curve has no barrier, then a barrierless reaction has been identified and no further calculation is needed. Alternatively, if a barrier is present along the interpolated path, then one would need to proceed with NEB and transition state optimization to check whether the barrier persists. [227] Therefore, a high-quality interpolation method will help avoid the more expensive NEB calculations when possible. Here the screening method is demonstrated on an experimentally studied photoreaction [228] between 10-methylphenothiazine and nitroxide-linked acceptor. As described previously, [229] the path connects the optimized D2/D3 and D0/D1 MECIs. The structures were optimized using SA-CASSCF using active space (3e,4o), averaging 7 states (4 doublets + 3 quartets), and with a PCM model ($\epsilon = 18.2$) for methanol solvent.

Figure 3.8 shows a comparison of the potential energy curves along the ZM and TRICS interpolated paths. Small barriers are present on all PECs along the ZM path (Figure 3.8b1), so further path optimizations are needed to establish whether the reaction is barrierless. On the other hand, the barriers are absent along the TRICS path and no further optimization is needed (Figure 3.8b2). The TRICS PECs are also more smooth with fewer undulations and bumps. We think this is because TRICS uses linear interpolation in the parameter space of the delocalized ICs, and this leads to smoother and less oscillatory behavior in the primitive ICs. On the other hand, the metric used by ZM is a nonlinear function of pairwise interatomic distances, and the optimal path may involve making compromises between smoothing out intramolecular distances vs. intermolecular ones. The values of many primitive ICs exhibit a smooth monotonic behavior from the TRICS path but a rather uneven behavior from the

ZM path (Figure 3.8c-f). In one case a planar C-C-C-C dihedral angle within an aromatic group increases to 30 degrees in the middle of the ZM path, which may contribute to the barrier; the same angle stays near 0 degrees in the TRICS path (Figure 3.8e). This suggests that TRICS is a preferred approach for high-throughput screening of barrierless reaction pathways.

A further advantage of TRICS is the capability to extrapolate a reaction path beyond the endpoints, as it is straightfoward to generate Cartesian coordinates corresponding to $t < 0$ or $t > 1$. Figure 3.9 shows an example where this feature is useful. The path is the same as Figure 3.8b2 but only the Q1 and Q2 energy levels are shown. The D2/D3 MECI optimized at the SA-CASSCF level is also a Q1/Q2 MECI due to degeneracies; the energy gap between Q1 and Q2 is 0.0217 eV. However, when the energy is recalculated using the highly accurate XMS-CASPT2 method that includes dynamic correlation effects, the gap increases to 0.234 eV. Although the location of the CI seam is expected to shift when changing the level of theory, it is still desirable to obtain a structure with a smaller energy gap at the XMS-CASPT2 level without performing expensive XMS-CASPT2 geometry optimizations.

We carried out extrapolation of the reaction path past the CASSCF-optimized endpoints, and found a point where the XMS-CASPT2 energy gap decreased to 0.068 eV located at $t = -0.10$ on the reaction coordinate. The extrapolated structure is a chemically reasonable approximation of the XMS-CASPT2 MECI because its energy is increased only modestly from the interpolated part of the path; the average energy of Q1+Q2 is 1.789 eV above the Q1 minimum located at $t = 0.78$, whereas at the CASSCF level, the MECI is 1.02 eV above the Q1 minimum located at $t = 0.80$. Therefore, reaction path extrapolation can be useful for estimating the MECI energy and geometry at the CASPT2 level without the need for CASPT2 MECI optimization.

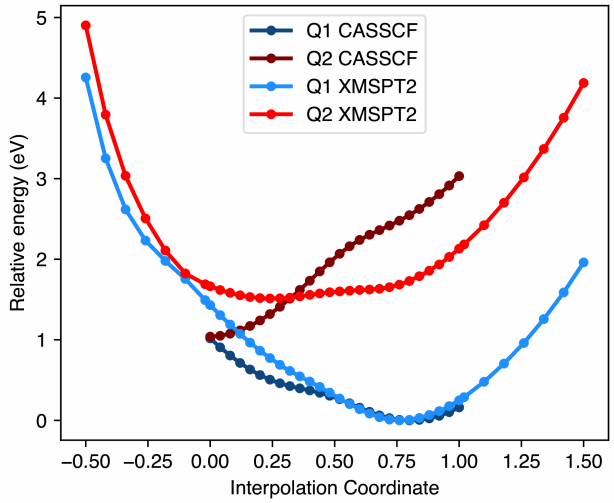



FIGURE 3.9. A plot of excited potential energy surfaces of the excited state photoreaction shown in Fig. 9. The original interpolated path between optimized CASSCF structures is plotted between $x = 0$ and $x = 1$. The extrapolation extends the path by half of the path length on either side. The decrease in the CASPT2 energy gap can be seen at $x = -0.1$, slightly to the left of the original path endpoint.

## 3.4. Discussion

The benchmarks in the preceding section show that TRICS can match or often surpass the performance of published approaches to reaction path interpolation, in terms of path smoothness as measured by arc length heterogeneity, as well as small excess energy above the MEP. The ability of TRICS to generate barrierless interpolation pathways for barrierless reactions shows that it minimizes the perturbations to orthogonal degrees of freedom that increase the energy. Extrapolation of reaction pathways beyond the reactant and product endpoints is straightforward. The method is freely available in geomeTRIC (version 1.2.0 and later).

Despite these advances, there are still major opportunities for improvement of geometry interpolation. The excess energies are still rather large for all of the methods tested, and each method can claim a share of reactions where it gives the lowest excess energy. One possible way to further improve the interpolation performance is via the internal coordinate system. The iterative splicing procedure described here may be used with any nonredundant IC

system. Here we used translation-rotation internal coordinates to perform the interpolation within a segment; while it is good at describing the molecular translations and rotations of non-bonded fragments, it is not designed to span the changes in connectivity when bonds are formed and broken. Therefore, as internal coordinate systems continue to improve to describe more complex structural changes, [230] the interpolation approach described here will also benefit.

It remains to be seen whether a truly general IC system could be devised that does not require any splicing. Indeed, in differential geometry, it is often not possible to parameterize an entire manifold with a single coordinate chart. Even so, the design of IC systems that can cover greater areas of configuration space is a worthwhile pursuit, and can reduce the need for splicing. We are also optimistic that the geometry-based methods presented here can further enhance AI and ML-based approaches to characterize the PES. Ultimately, AI approaches are derived from high-quality data and our understanding of molecular geometry encoded in the featurization. Therefore, further efforts to advance the study of reaction pathways through the lens of geometry are expected to play an important role in the next generation of computational chemistry tools.

## 3.5. Conclusion

In this work, we have introduced a novel interpolation method, TRICS (Translation–Rotation Internal Coordinates with Splicing), that linearly interpolates internal coordinates (ICs) between reactants and products. The challenge that a set of well-defined ICs within one segment of configuration space may become ill-defined in another was addressed by employing multiple translation–rotation internal coordinates (TRICs) for each segment. Trajectories were interpolated within each segment using the corresponding ICs, and the disconnection points between segments were iteratively smoothed by exchanging the target structures with those of the adjacent trajectory and re-interpolating (splicing) the segments together. The

performance of TRICS was benchmarked against four widely used interpolation methods (ZM, IDPP, DLC, and LST) on 679 reactions from three datasets (RGD1, BH, and NR56).

The interpolated trajectories were first filtered according to the evenness of the images. Because poor-quality pathways typically exhibit large gaps between images, the quantified arc length heterogeneity (ALH) was used to assess success; trajectories with ALH below 20 were considered successful. All five methods achieved high success rates ($> 90$ %), with TRICS and IDPP both reaching 100 %. Next, the energy gaps between the highest-energy image of each trajectory and the corresponding reference optimized transition state structure were evaluated, with smaller gaps indicating higher-quality pathways. TRICS produced the smallest gaps across all three datasets. In addition, TRICS generates smooth, barrierless paths on excited-state potential energy surfaces from the Franck–Condon point to a conical intersection, whereas the other leading method introduces an unnecessary barrier. TRICS also offers a clear advantage over ZM and IDPP through its ability to extrapolate beyond the endpoint structures.

TRICS is available in the open-source package geomeTRIC (version 1.2.0 and later). It is a meaningful addition to the theorist's toolkit for discovering reaction pathways. It can generate high quality minimum energy pathway estimates and be used to explore reactions that do not necessarily involve energy barriers. Because it performs linear interpolation in ICs, the method can also extrapolate beyond the reactant and product endpoints. Future work involves fine-tuning the atomic-clash threshold to further lower trajectory energies and integrating it into automated reaction discovery workflows.

## 3.6. Supporting Information

### 3.6.1. Initial Reaction Pathways

The initial trajectories are generated from interpolated paths connecting the starting point $\mathbf{x}_0$ to the destination $\mathbf{x}_1$ using ICs created from four different sets of RICs defined at $\mathbf{x}_0$. The four sets are: (1) PICs plus translational-rotational internal coordinates, (2) PICs with only

the common bonds that exist at both ends plus translational-rotational internal coordinates, (3) PICs with the common bonds plus Cartesian coordinates, and (4) inverse distances and Cartesian coordinates. In the fourth set, all detected bond lengths are inverted, along with additional pairwise inverse distances defined within the angles. Specifically, distances between the two terminal atoms of proper, linear, and dihedral angles and among the three terminal atoms of out-of-plane angles are inverted. Redundancies from each mode are removed via the delocalization process. The resulting non-redundant IC is used to calculate $\Delta\mathbf{q}(\mathbf{x}_0, \mathbf{x}_1)$ and the step size. After each step, the IC is rebuilt based on the new Cartesian coordinates along with the remaining IC differences $\Delta\mathbf{q}(\mathbf{x}_n, \mathbf{x}_1)$. The step size is recalculated based on the number of remaining steps, and the geometry is updated until the final structure is reached. This process is repeated backward (starting from $\mathbf{x}_1$ to reach $\mathbf{x}_0$) to obtain a total of eight candidate pathways. Note that the backward pathways are not necessarily identical to the forward ones, as they are constructed from different sets of RICs.

The initial trajectories are analyzed using a collection of PICs that are well-defined at both endpoints. First, the union of PICs from the endpoints is evaluated on both structures to remove those with diagnostic values greater than 1 (see Section 2 for the diagnostic value calculations). The remaining PICs are then used to compute the changes between adjacent frames $\Delta\mathbf{q}(\mathbf{x}_n, \mathbf{x}_{n+1})$ for each candidate trajectory. Angle coordinates are re-evaluated on every frame; those with diagnostic values larger than 1 are assigned a negligible value (i.e., $-1\times10^{-6}$) in $\Delta\mathbf{q}(\mathbf{x}_n, \mathbf{x}_{n+1})$ to ignore them. For each PIC, the maximum and median changes across frames are calculated, along with the ratio of the maximum to the median. The largest of these ratios among all PICs is stored for each candidate. This metric identifies lower-quality pathways, since a trajectory containing a sudden jump produces a disproportionately large maximum change relative to the median.

The frames are then re-spaced based on the largest changes in the PICs. Candidates for which the maximum Cartesian difference between the final structure and the destination

exceeds $1 \times 10^{-2}$ Å are discarded. To mitigate potential atomic collisions along the trajectories, distance thresholds are calculated and used to replace bond ICs with inverse-distance ICs. The thresholds are the larger of 1.0 Å and the sum of the covalent radii of the corresponding atoms plus 0.9 Å. If any bond distances fall below their respective thresholds, the corresponding distance ICs are replaced with inverse-distance ICs. The new set of PICs is then used to build non-redundant ICs, and the interpolation process is repeated until all distances exceed their thresholds or a maximum of four additional iterations have been performed. Among the final candidates, the one with the lowest maximum-to-median IC change ratio is selected as the initial trajectory and is further refined through TRICS.

### 3.6.2. Diagnostic for Well-definedness of Internal Coordinate Systems

We used the following procedure to select a minimal set internal coordinate systems that span the reaction pathway. First, we create IC systems corresponding to each coordinate set (image) on the initial pathway. A diagnostic value $D_{ij}$ is then calculated to estimate the well-definedness of the $IC_i$ in the neighborhood of $\mathbf{x}_j$. A diagnostic value $d$ is computed for each primitive IC to indicate whether it falls outside of an acceptable range. For example:

(1) Dihedral angles:
   (a) $\phi \geq 175 : d = 3$
   (b) $175 > \phi \geq 165 : d = 2$
   (c) $165 > \phi \geq 155 : d = 1$
   (d) Otherwise: $d = 0$

(2) Linear angles:
   (a) $|\cos(\theta)| < 0.4 : d = 3$
   (b) $0.4 \leq |\cos(\theta)| < 0.6 : d = 2$
   (c) $0.6 \leq |\cos(\theta)| < 0.8 : d = 1$
   (d) Otherwise: $d = 0$

(3) Out-of-plane angles (where $\theta$ is formed by three atoms containing the central atom):

(a) $|\cos(\theta)| > 0.99 : d = 3$

(b) $0.98 < |\cos(\theta)| \leq 0.99 : d = 2$

(c) $0.95 < |\cos(\theta)| \leq 0.98 : d = 1$

(d) Otherwise: $d = 0$

For each out-of-plane angle, another proper angle formed by the three terminal atoms is also examined using the same criteria. The two diagnostic values are then added together, so the total diagnostic value for an out-of-plane coordinate can reach a maximum of 6.

### 3.6.3. Negative Excess Energy Cases

We found three BH reactions where the highest-energy trajectory images having lower energies than the optimized TS structures. All three of them are from BHROT27, and the interpolation methods resulted in trajectories that changed the dihedral angle in the opposite direction of the reaction pathway. Figure 3.11 presents one of the three reactions, h2o2_cis, from BHROT27. The straight arrows indicate the reaction pathway, whereas the curved arrows represent the interpolated pathways. The dihedral angle of h2o2_cis reaction starts at 118° and decreases to −118° passing 0°, undergoing the cis conformation. All four interpolation methods, except LST (since it failed to interpolate even trajectories for all three BHROT27 reactions), interpolated through the "Interpolated TS" structure, increasing the initial dihedral angle and reaching the final value −118°, passing −180°, forming the trans TS conformation. This essentially generated identical results to those from interpolations of the h2o2_trans reaction. The other two BHROT27 reactions, n2h4_ecl2 and h2s2_cis, exhibit the same issue, although only the DLC method resulted in an interpolated TS with lower energy than the optimized TS, while other methods were able to interpolate correctly.

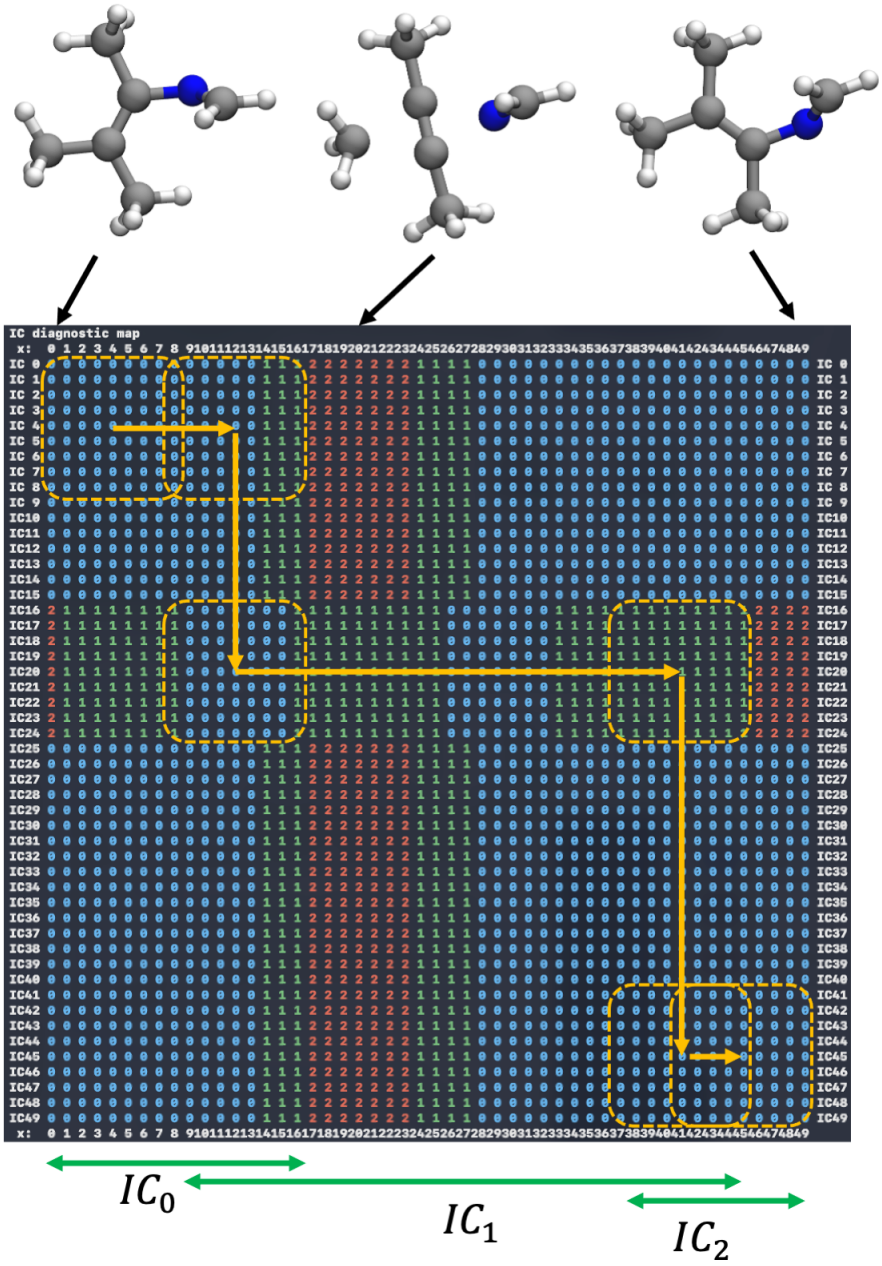



FIGURE 3.10. Diagnostic matrix of an initial pathway for a reaction and algorithm for choosing path segments for TRICS interpolation. The regions in blue, green, and red correspond to diagnostic values of 0, 1, and 2, which indicate the amount of numerical instability of the internal coordinate system generated from image $i$ (row index) evaluated at the coordinates of image $j$ (column index). Representative structures are shown at the top (C, gray; N, blue; H, white). The numerical problems indicated by the cross-shaped region centered around image 20 are due to the linear arrangement of four carbon atoms. The orange box shows how the path segments are determined by moving a square sub-matrix (the "car") from the top left to the bottom right while avoiding regions of high diagnostic values (the "maze"). Final selections of overlapping path segments for interpolation are shown at the bottom.

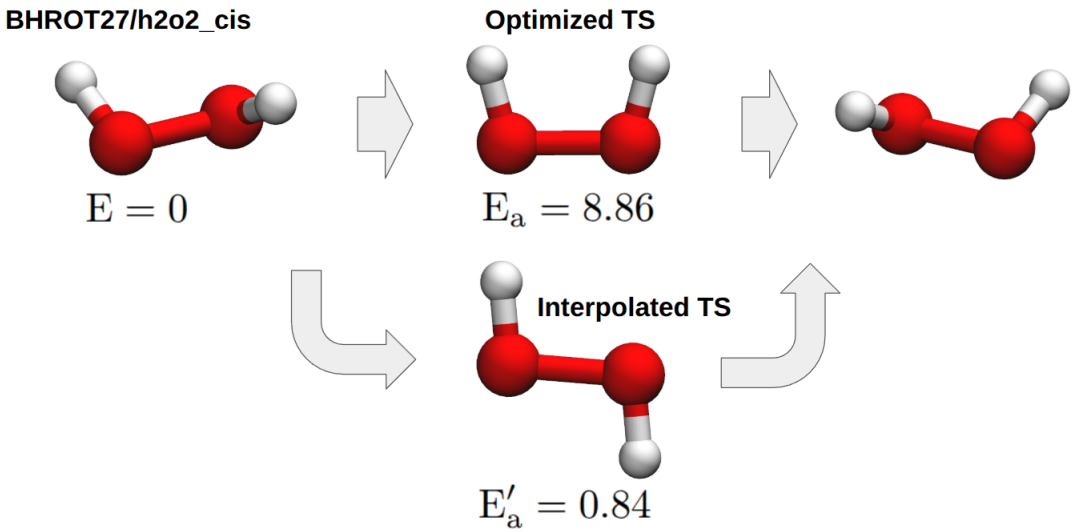



FIGURE 3.11. An example reaction (h2o2_cis: $H_2O_2$ molecule going through cis-dihedral angle change) where the highest-energy interpolated trajectory images ("Interpolated TS") have lower energies than the optimized TS structures ("Optimized TS"). The interpolated TS structures shown are from the TRICS results. The straight arrows show the direction of the reaction, whereas the curves arrows represent the interpolated pathway. Energies were calculated at the B3LYP/6-31G(d) level of theory, and the presented values are in kcal/mol.

# Chapter 4

# High-Throughput Approach for Minimum Energy Pathway Search Using the Nudged Elastic Band Method with Efficient Data Handling and Parallel Computing

*The content and figures in this chapter are reprinted or adapted from the author's previously published article: Park, H.; Pritchard, B. P.; Wang, L.-P., J. Chem. Theory Comput. 2025, 21 (23), 12048–12063.* *Copyright 2025 American Chemical Society.*

## 4.1. Introduction

The application of transition state theory to model reaction rates requires finding molecular geometries (i.e. nuclear configurations) located at energy minima and first-order saddle points on the potential energy surface (PES). [47, 48, 49] In computational quantum chemistry, the process of obtaining these critical points begins with computing the energy of the initial geometry, then approximating the shape of the PES in its local neighborhood using nuclear gradients and Hessians (first and second derivatives). The geometry is then iteratively optimized toward the desired critical point. There are various algorithms that can reliably carry out energy minimizations, [231] but locating the transition states (i.e. saddle points) is much more challenging due to the high sensitivity of the optimization method to the initial geometry. [198, 232, 233, 234, 235] A high-quality initial guess transition state (TS) structure is required to successfully locate the first-order saddle point.

The lowest energy path on the PES connecting reactants and products is known as the minimum energy pathway (MEP). The most commonly accepted representation of the MEP is a geodesic, or metric-minimizing curve, called the intrinsic reaction coordinate (IRC) pathway, defined in mass-weighted Cartesian coordinates with a metric reflecting the physical dynamics of the molecule system. [66, 67, 195, 196] The IRC pathway is determined by

following the instantaneous accelerations along the imaginary mode, starting from a first-order saddle point. [67, 68, 236] The MEP is the path with the least classical action for the transition from reactants to products compared to all other possible pathways, making it the most probable route under the Boltzmann distribution of all paths due to its lowest energy barrier. [48, 49]

When the precise location of the first-order saddle point is unknown, other techniques like Chain-of-States (CoS) methods [237] can be employed to find an approximate MEP. In these methods, the variable being optimized is a "chain" consisting of several molecular structures ("images") connecting two energy basins on a PES, and successive iterations result in better approximations to the exact MEP. Among these, the nudged elastic band (NEB) method [199, 200, 201, 202, 203, 204, 205] and string methods (SM) [206, 207, 208, 209, 210, 211, 238, 239, 240, 241, 242, 243] are the most commonly utilized approaches. These methods can search for the MEP quite effectively by minimizing forces perpendicular to the path while keeping the distances between the images constant through various means. To locate the TS precisely, the highest-energy image of the optimized chain can be used as a high-quality initial guess for local TS optimization methods. [232, 233, 244] The optimized TS structure should then be verified using the IRC method to confirm that the imaginary mode leads to both the reactant and the product. Therefore, CoS methods can be used either on their own to produce an approximate MEP, or as part of a workflow to find an optimized TS structure. Many of the concepts discussed here are also relevant to string methods, though this work focuses on NEB.

The initial chain used as input for NEB can be obtained through interpolation of the reactant and product coordinates. Cartesian interpolation is trivial, but more advanced methods such as the image-dependent pair potential [215] or geodesic interpolation [216] can be employed to avoid high-energy images that are physically not feasible. The NEB algorithm updates the chain iteratively using the gradient of the PES for each image as well

as harmonic spring forces applied to neighboring image pairs. Since the PES energy/gradient calculations for different images are mutually independent, they can be carried out in parallel employing a high-throughput distributed quantum calculation workflow such as FireWorks [245], BigChem [246], QMFlows, [247] and QCArchive. [71]

The distinguishing feature of NEB is the modification of force components that retains only the perpendicular component of the PES gradients and the parallel component of spring forces with respect to the chain, which produces better approximations of the TS structure and improved convergence properties. A number of improvements to the NEB method have been proposed in the literature. The climbing images NEB method modifies the update to the highest-energy image by forcing it to ascend the PES independently of the path direction. [202] Efforts to improve the resolution of NEB images around the high-energy region have resulted in the energy-weighted NEB [248] and adaptive spring length approaches. [249] The well-known problem of kink formation, which can prevent convergence in NEB calculations, has been recently addressed with the NEB stiffness method [250] and curvature-weighted NEB. [251] Extensions of NEB have also been developed to enable its applications in large biomolecular systems, [252] solids, [205] and reactions with multiple spin states. [253]

Data-driven approaches have gained traction in computational chemistry to address complex challenges, including molecular representation, [254, 255] protein structures prediction, [256,257] and drug discovery. [258,259,260] The success of these studies relies on two key factors: access to large volumes of high-quality data and their effective utilization. To meet the first requirement, publicly available quantum chemistry databases, such as PubChemQC [261] and Alchemy, [262] provide molecular structures and ground-state electronic properties of millions of small molecules, computed using density functional theory (DFT). Similarly, datasets like Small-molecule/Protein Interaction Chemical Energies (SPICE) offer energetic interaction data between drug-like small molecules and proteins. [263,264] To address the second need, high-throughput data selection and processing workflows, such as

The Million-Molecule Challenge [265] and NOMAD, [266] have been developed to enhance data usability.

Given this trend, there is a clear need for infrastructure to systematically organize and store chain optimization calculation results in a database, while also enabling the concurrent execution of the calculations. To our knowledge, no such publicly available platform is capable of carrying a large number of chain optimization calculations while organizing and storing the results. A database created from these could be shared and leveraged to train various ML models. For instance, Mathias and his team developed a dataset named Transition1x, which contains 9.6 million DFT calculations derived from NEB calculations of over 10,000 various organic reactions. [267] They also used this dataset to train models, employing them as potentials for the NEB method and achieving a mean absolute error of 0.23 eV and a root mean squared error of 0.52 eV on energy barrier heights. [268] Furthermore, such infrastructure could serve as a platform for benchmarking the performance of chain optimization protocols. In this study, we addressed this need by developing a new workflow that combines two open-source software packages, geomeTRIC [65] and Quantum Chemistry Archive (QCArchive). [71] By integrating these tools, we enabled parallel chain optimization calculations across distributed computing resources, with results automatically organized and stored in a database for efficient retrieval as part of data-driven research projects.

The geomeTRIC software package was originally developed as a proof-of-concept for the *translational-rotational internal coordinate* (TRIC) system, a delocalized internal coordinate system that employs primitive coordinates for translation and rotation of molecules and fragments; it is capable of multiple types of geometry optimization calculations, and interfaces with quantum chemistry software packages by passing Cartesian coordinates and receiving the energy and Cartesian nuclear gradients. [65] Our previous study demonstrated the improved performance of TRIC for geometry optimization of molecular systems dominated by

intermolecular degrees of freedom. [65] In recent years, new features, including TS optimization, conical intersection optimization, and the intrinsic reaction coordinate method have been added. The implementation of the NEB method into geomeTRIC is described in this work.

The other open-source software package used in this study is QCArchive. [71] The QCArchive infrastructure offers solutions for managing numerous calculations in parallel and storing the calculation results efficiently. [71] It comprises several components, of which QCFractal and QCPortal are described here; details about the other components and the overall structure are available in the original paper. [71] QCFractal provides a server that stores quantum chemistry calculation data in a Structured Query Language (SQL) database, [269] along with a web Application Programming Interface (API) for accessing the data. Users can remotely connect to the server via QCPortal, a Python-based front-end, to submit calculations or query results. QCFractal is also capable of distributing a very large number of submitted tasks across multiple computing resources provided by users.

This paper presents the details of the NEB method and explains how QCFractal and geomeTRIC efficiently perform a large number of chain optimization calculations and store their results in a database. We first apply this infrastructure to perform a sensitivity analysis of the chain optimization parameters (i.e. image number, spring constant, etc.) for four reactions from the RGD1 organic reaction dataset, [109] using hundreds of different protocols. This result in recommended ranges for the parameter values. [270] The chain optimizations may be conducted using various approaches to electron structure theory, including DFT and semiempirical models. For the DFT calculations, the B3LYP/6-31G(d) [28,34,224,225] and B3LYP-D3/def2-TZVP [28, 34, 39, 271, 272] levels of theory via an open-source quantum chemistry package Psi4 [70] are used, while the PM7 model [273, 274] from the Molecular Orbital PACkage (MOPAC) [275] is employed. This is followed by a performance benchmarking study where 118 reactions from seven different reaction barrier height (BH)

datasets [223] were optimized at the B3LYP/6-31G(d) level of theory (as implemented in the Q-Chem quantum chemistry package) [226] and the PM7 semi-empirical model. Lastly, three elementary reactions involving larger molecules from previous mechanistic studies on molecular transition metal catalysts are reproduced here. [276,277] All successful chain optimization calculations, specifically the highest-energy images of the final chains, were further optimized to locate the exact first-order saddle points.

The rest of this paper is organized as follows: First, we will discuss the details of the reactions used to test the algorithm in this study. Next, we will discuss how the chain optimization method, implemented in geomeTRIC, guides the initial chain toward the MEP, and how it interacts with QCFractal to manage a large number of NEB calculations concurrently. Finally, we will discuss the results from the implementation, focusing on the optimization convergence rates of the chains and the quality of the TS estimates.

## 4.2. Methods

### 4.2.1. Reactions

This study involves optimizing reactions in three different groups, each with a different purpose. First, four linearly interpolated chains generated from reactions in the RGD1 dataset were optimized using various parameter combinations and levels of theory to investigate the method's sensitivity to the parameters. Two sets of parameters that performed well (optimizing chains with fewer iterations while providing high-quality initial guesses for TS optimizations) were then selected to optimize 118 chemically diverse reactions from the BH dataset. As a supplement to the reactions in both the RGD1 and BH datasets, which involve only small molecules and simple reactions, three transition metal-catalyzed energy conversion elementary reactions were optimized to reproduce PES from previous publications. [276,277] Details of these selected reactions will be discussed in this section.

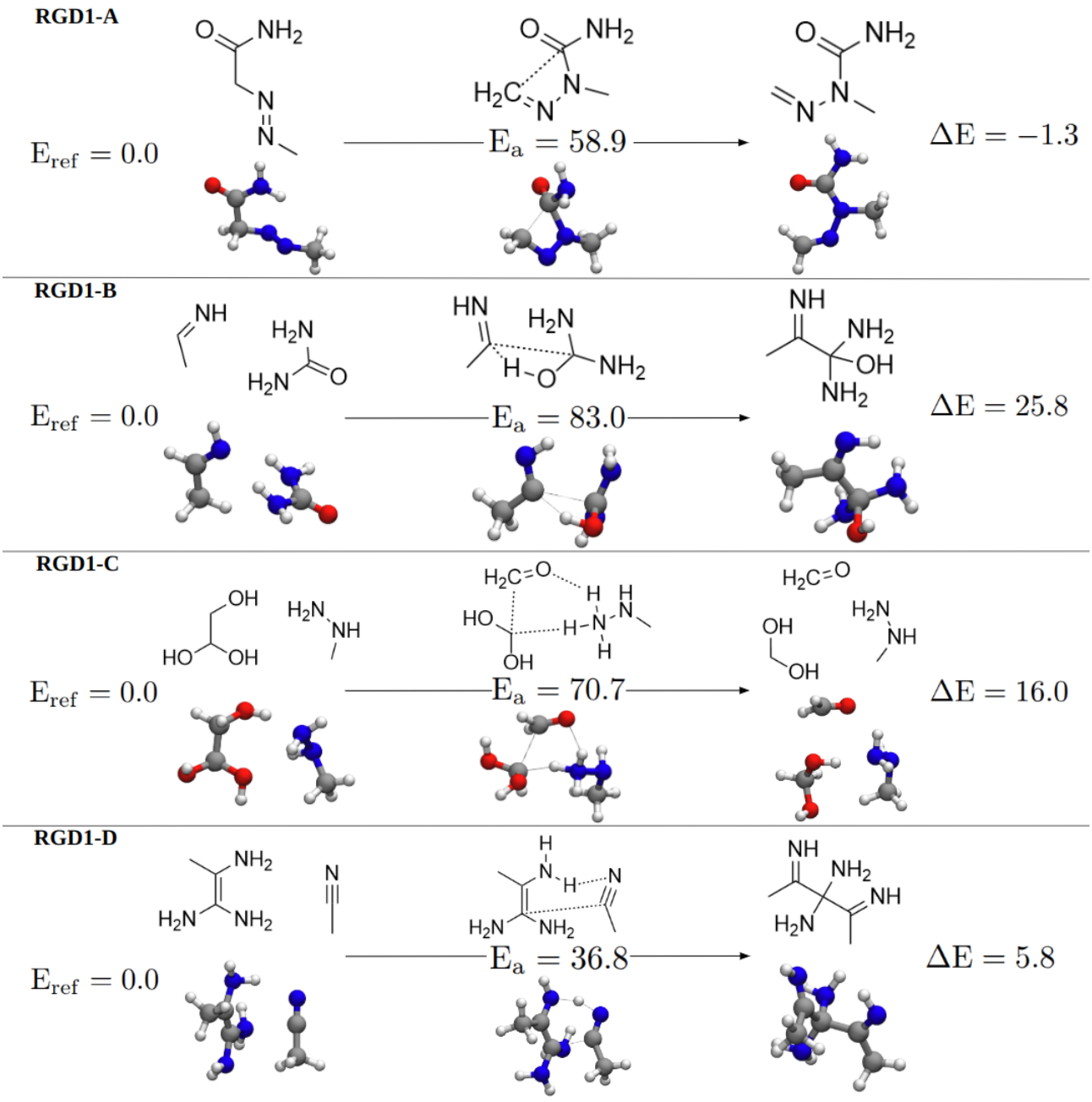



FIGURE 4.1. The endpoint and TS geometries of the four selected reactions from the RGD1 dataset. The energies reported here are in kcal/mol, optimized using the B3LYP-D3/def2-TZVP level of theory. Atoms are colored as follows: nitrogen (blue), oxygen (red), carbon (gray), and hydrogen (white).

TABLE 4.1. The chain optimization parameters and their values are listed. All possible combinations of the parameters and three band type variations were used to optimize initial chains for the RGD1 reactions. The unit of the spring constant is kcal/mol/Å$^2$.

| Parameters | Values |
|---|---|
| Spring constant | 0.1, 0.5, 1.0, 5.0, 10.0, 50.0, 100.0, 500.0, 1000.0 |
| Number of images | 7, 11, 15, 21 |
| Align | True, False |

#### 4.2.1.1. *RGD1 reactions*

The RGD1 dataset consists of over 170,000 singlet, neutral organic reactions involving up to 10 heavy atoms. [109] The four reactions selected from the RGD1 dataset are MR_100987_0, MR_421485_0, MR_547833_1, and MR_557723_0. [109] We will refer to these reactions as RGD1-A, RGD1-B, RGD1-C, and RGD1-D, respectively. The molecular structures of the reaction endpoints from the dataset, which were further optimized and visualized using VMD [134] and ChemDraw, [278] are shown in Figure 4.1. The first three (A, B, and C) reactions were selected because they are unimolecular, bimolecular, and termolecular reactions, respectively. The fourth was included due to its much lower energy barrier compared to the other three, making it more relevant to experiments. The initial input chains, consisting of 101 images, were generated through linear interpolation of the Cartesian coordinates between the two endpoints, which are raw structures from the database. The four interpolated RGD1 reaction trajectories were optimized using a total of 72 combinations of chain optimization parameters, three different band types, two different levels of DFT (B3LYP/6-31G(d) [28, 34, 224, 225] and B3LYP-D3/def2-TZVP [28, 34, 39, 271, 272], as implemented in Psi4), and the PM7 semiempirical model from MOPAC. [273, 274, 275] The parameters and their values are listed in Table 4.1, with further discussion in the following section. The TS structures from the dataset were optimized with the two DFT levels of theory to verify the final TS structures. The xyz files for the optimized reactant, product, and TS structures, along with the xyz files for the two endpoints used in the linear interpolations, are provided in

the Supplementary Information (SI). The endpoints of the chains serve as anchors for other images throughout the chain optimization procedure. To ensure they are properly placed in energy basins, they were optimized using the corresponding level of theory before the chain optimizations. The highest-energy image of the final chain is taken as the approximated TS structure and optimized to locate the first-order saddle point precisely. The TS structures are considered optimized when the following five criteria are met: the energy difference between the current structure and the previous iteration is less than $1.00 \times 10^{-6}$ Hartree; the root-mean-square displacement and maximum displacement, taken over all atoms, are less than $1.20 \times 10^{-3}$ and $1.80 \times 10^{-3}$ Å, respectively; and the root-mean-square gradient and max gradient of the current structure are less than $3.00 \times 10^{-4}$ and $4.50 \times 10^{-4}$ Hartree/Bohr, respectively. The TS optimizations were performed at the same level of theory as the chain optimizations for the DFT approach, while both B3LYP/6-31G(d) and B3LYP-D3/def2-TZVP were used to optimize the guessed TS structures from the chains optimized with the PM7 semiempirical model. The normal modes, energies, and geometries of the resulting TS structures were compared with those of the optimized TS structures from the dataset to calculate the overall success rate. A total of 3,456 chain and TS optimizations are carried with the four RGD1 dataset reactions.

#### 4.2.1.2. *BH reactions*

We carried out preliminary processing of the 118 selected BH reactions to make the data suitable for our study using geomeTRIC [65] and Q-Chem [226]. The BH subsets are described in Table 4.2. For example, in many of the reactions, the product structures were not available, so they needed to be found using IRC calculations. [67] The source data came from multiple publications; therefore, their TS structures were not all optimized using the same level of theory or convergence criteria. Details of the processing procedure, along with the optimized structure of reactants, TSs, and products for the final 118 reactions, are provided in the SI (Section 2: Preliminary Processing of the BH Dataset). Figure 4.2 presents

| Subset | Description | Number of reactions |
|---|---|---|
| BH76 | Atom transfer, nucleophilic substitution, unimolecular and association reactions | 17 |
| BHPERI | Pericyclic reactions | 26 |
| BHDIV10 | Diverse reactions | 7 |
| INV24 | Inversion/racemization processes | 18 |
| BHROT27 | Single bond rotations | 19 |
| PX13 | Proton-exchanges | 13 |
| WCPT18 | Proton-transfer in uncatalyzed and water-catalyzed reactions | 18 |
| | Total | 118 |

TABLE 4.2. Descriptions of the seven BH subsets and the number of reactions used in this study. [223]

the optimized 3D structures of the reactants, TSs, and products for seven representative BH reactions.

The 118 linearly interpolated trajectories were optimized with two different sets of recommended chain optimization parameters from the sensitivity study: one with a spring constant of 1.0 kcal/mol/Å$^2$, 21 images, NEB band type, and alignment enabled, and the other set with the band type varied to hybrid. The chain optimizations were carried out using B3LYP and PM7, resulting in a total of 472 chain optimizations. The approximate TS structures, derived from completed chain optimizations, were optimized with B3LYP. The overall performance is evaluated based on the convergence rates of chain and TS structure optimization. Q-Chem is used for performing DFT calculations, whereas MOPAC is employed for the PM7 semiempirical model.

#### 4.2.1.3. *Selected reactions*

Modern applications of DFT often involve larger molecules or more complex geometries than the systems in the BH datasets, which are small by necessity due to the need to compute benchmark-quality energies. To address this deficiency, we include three calculated elementary steps from published studies on transition metal-catalyzed energy conversion

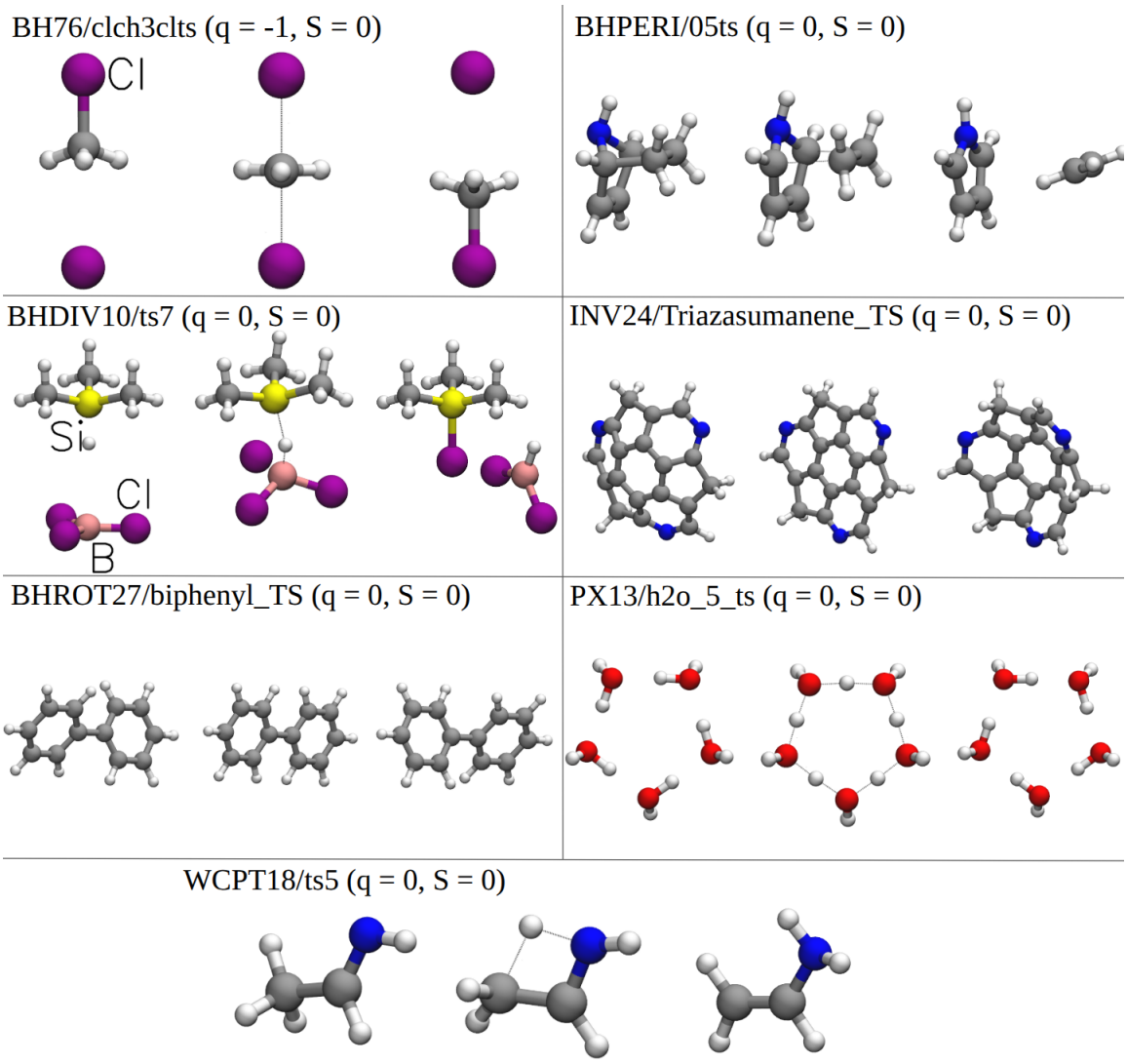



FIGURE 4.2. Seven example BH reactions, one from each subset, are presented. The structures, optimized using B3LYP/6-31G(d), depict the reactant, TS, and product from left to right. The charges and total spin quantum numbers are indicated in parentheses next to their names. Atoms are colored as follows: nitrogen (blue), oxygen (red), carbon (gray), hydrogen (white), and other elements are labeled in the figure. Their barrier heights and reaction energies, along with those of the other BH reactions, are listed in Table S1.

reactions. [276, 277] The 2D and 3D structures of the reactants, TSs, and products for these reactions are shown in Figure 4.3. One reaction involves the displacement of $O_2$ by $H_2O$ from a Ru(bpy)(tpy) coordination complex ($O_2$-displacement). [276] This reaction occurs at

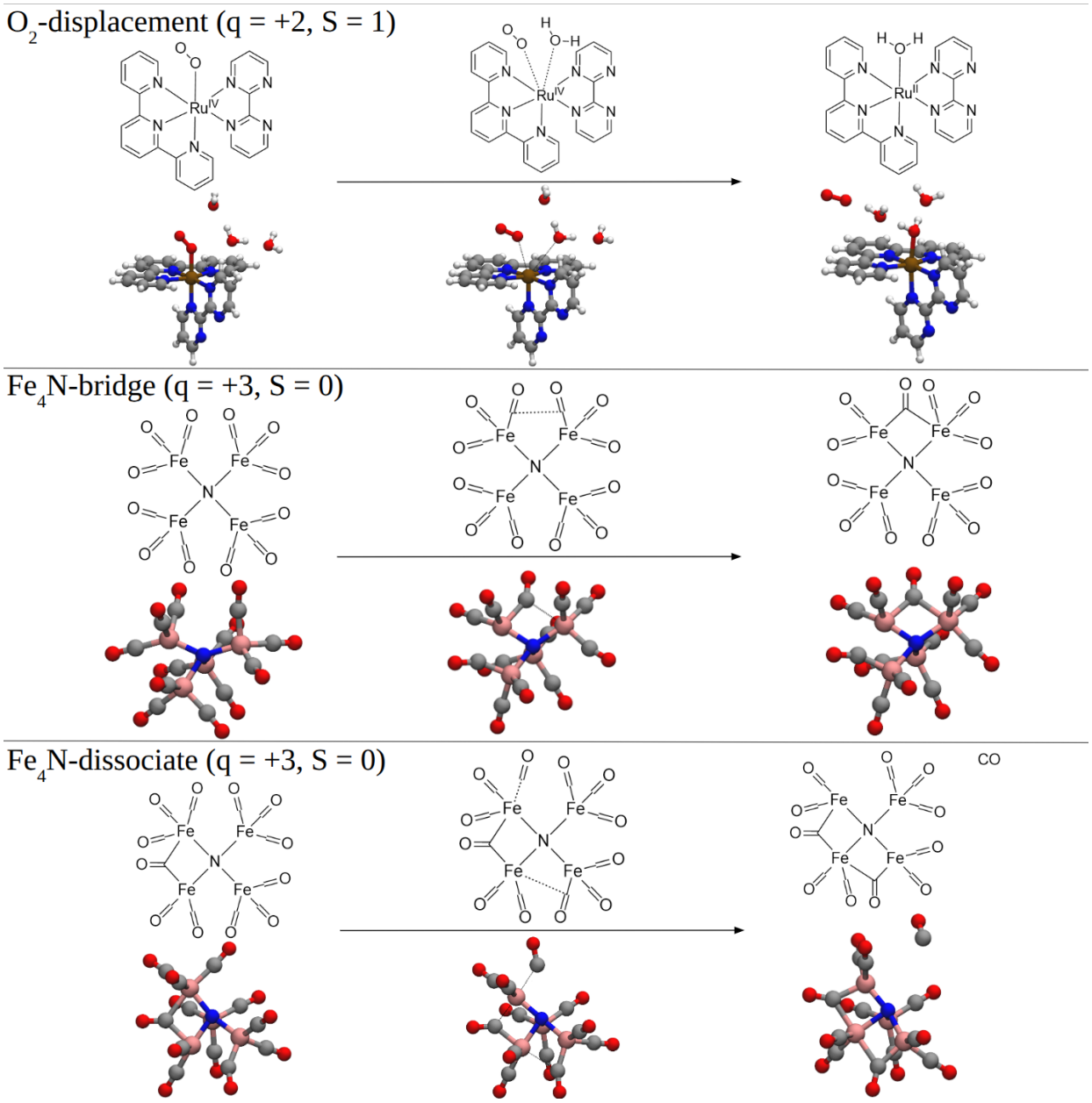



FIGURE 4.3. The three elementary reactions selected for reproduction in this study are presented. The visualized structures are from the previous results. The charges and total spin quantum numbers are indicated in parentheses next to the reaction names. Atoms are colored as follows: nitrogen (blue), oxygen (red), carbon (gray), hydrogen (white), iron (pink) and ruthenium (brown).

the end of a catalytic cycle of water oxidation in which the coordination complex is oxidized a total of four times and two water molecules are converted to one molecule of dioxygen. The energies of the metal cluster were calculated using the B3LYP/6-31G(d) level of theory, with the Ru atom represented by the LANL2DZ effective core potential (ECP). [276, 279] Two other reactions involve the geometric rearrangements of a $Fe_4N(CO)_{12}$ transition metal complex that catalyzes the electrochemical reduction of $CO_2$ to formate. [277] Our previous studies showed that the PES of this molecule contained many local minima characterized by geometric distortions of the $Fe_4N$ core, rearrangements of the ligands, and even CO ligand dissociation. In the first reaction, one CO ligand moves from a typical coordination geometry to a bridging geometry in between two Fe atoms ($Fe_4N$-bridge), whereas in the second reaction a CO ligand dissociates from the complex ($Fe_4N$-dissociate). The energies were calculated using the BP86 gradient corrected semi-local functional [27, 280], with the LANL2DZ ECP for the Fe atoms and the 6-31G(d) basis set for the rest of the atoms. These three elementary reactions represent the complex changes in geometry that often occur in transition metal catalysis and together present a more difficult challenge for the implementation. The initial input chains for the three reactions were generated by linearly interpolating between reactant and product structures. The input chains for these three reactions were optimized using the following set of parameters: a spring constant of 1.0 kcal/mol/$Å^2$, 21 images, hybrid band, and alignment enabled.

### 4.2.2. Nudged Elastic Band Method

The initial input chains were prepared by selecting a number of images with evenly spaced intervals among the 101 images including the reactant and product. The interpolation endpoints for RGD1 reactions are raw structures from the dataset, whereas we carried out energy minimizations to locate the interpolation endpoints of reactions from the other two groups. After the calculation is initiated, the two endpoints of the input chains were energy-minimized at the same level of theory as the chain optimization and are not adjusted further.

After optimizing the endpoints, the entire chain can be aligned to the first optimized frame before initiating the chain optimization. Our study includes comparing the performance of the chain optimization calculations whether or not the alignment step is used.

The improved tangent and climbing image method from Ref. 201,202 were implemented into geomeTRIC. The forces on each image include contributions from the PES ($\mathbf{F}_{\text{PES}}$) and harmonic-like spring forces to adjacent images ($\mathbf{F}_{\text{spring}}$). Depending on the components of these forces, geomeTRIC can optimize three different types of bands: plain band, hybrid band, and NEB. The plain band force components can be calculated as follows:

$$\mathbf{F}_i^{\text{Plain}} = \mathbf{F}_{\text{PES}}(\mathbf{r}_i) + \mathbf{F}_{\text{spring}}(\mathbf{r}_i) \tag{4.1}$$

$$\mathbf{F}_{\text{PES}}(\mathbf{r}_i) = -\nabla \text{E}(\mathbf{r}_i) \tag{4.2}$$

$$\mathbf{F}_{\text{spring}}(\mathbf{r}_i) = k[(\mathbf{r}_{i+1} - \mathbf{r}_i) - (\mathbf{r}_i - \mathbf{r}_{i-1})] \tag{4.3}$$

where $\mathbf{r}_i$ is the geometry of $i$-th image, $k$ is the spring constant. The force in the plain elastic band, as given by Eq. (4.1), comprises the full forces from the PES and the spring, which are detailed in Eq. (4.2) and Eq. (4.3), respectively. The NEB (Eq. (4.4)) and hybrid band (Eq. (4.5)) force components can then be defined as follows:

$$\mathbf{F}_i^{\text{NEB}} = \mathbf{F}_{\text{PES}}^{\perp}(\mathbf{r}_i) + \mathbf{F}_{\text{spring}}^{\parallel}(\mathbf{r}_i) \tag{4.4}$$

$$\mathbf{F}_i^{\text{Hybrid}} = \mathbf{F}_{\text{PES}}^{\perp}(\mathbf{r}_i) + \mathbf{F}_{\text{spring}}(\mathbf{r}_i) \tag{4.5}$$

$$\mathbf{F}_{\text{PES}}^{\perp}(\mathbf{r}_i) = \mathbf{F}_{\text{PES}} + (\nabla \text{E}(\mathbf{r}_i) \cdot \hat{\boldsymbol{\tau}}_i)\hat{\boldsymbol{\tau}}_i \tag{4.6}$$

$$\mathbf{F}_{\text{spring}}^{\parallel}(\mathbf{r}_i) = k(|\mathbf{r}_{i+1} - \mathbf{r}_i| - |\mathbf{r}_i - \mathbf{r}_{i-1}|)\hat{\boldsymbol{\tau}}_i \tag{4.7}$$

where $^{\perp}$ and $^{\parallel}$ indicate that each force contribution is decomposed into perpendicular and parallel components to the path. Equations (4.6) and (4.7) describe the projected force components, where portions of the forces are projected out using the unit vector $\hat{\boldsymbol{\tau}}_i$ along the

reaction path. This unit vector is defined based on the relative energies of adjacent images as follows: [201]

$$\boldsymbol{\tau}_i = \begin{cases} \boldsymbol{\tau}_i^+ = \mathbf{r}_{i+1} - \mathbf{r}_i & \text{if } \mathrm{E}_{i+1} > \mathrm{E}_i > \mathrm{E}_{i-1} \\ \boldsymbol{\tau}_i^- = \mathbf{r}_i - \mathbf{r}_{i-1} & \text{if } \mathrm{E}_{i+1} < \mathrm{E}_i < \mathrm{E}_{i-1} \end{cases} \tag{4.8}$$

$$\hat{\boldsymbol{\tau}}_i = \frac{\boldsymbol{\tau}_i}{|\boldsymbol{\tau}_i|}$$

where $\mathrm{E}_i$ is the energy of $i$-th image. Therefore, for images that are not local energy maxima or minima along the chain, the tangent vector is always defined using the higher-energy neighbor. For images identified as local extrema, a different unit vector $\hat{\boldsymbol{\tau}}_{\text{ext}}$ is applied instead of the one defined in Eq. (4.8).

$$\boldsymbol{\tau}_{\text{ext}} = \begin{cases} \boldsymbol{\tau}_i^+ \Delta \mathrm{E}_i^{\max} + \boldsymbol{\tau}_i^- \Delta \mathrm{E}_i^{\min} & \text{if } \mathrm{E}_{i+1} > \mathrm{E}_{i-1} \\ \boldsymbol{\tau}_i^+ \Delta \mathrm{E}_i^{\min} + \boldsymbol{\tau}_i^- \Delta \mathrm{E}_i^{\max} & \text{if } \mathrm{E}_{i+1} < \mathrm{E}_{i-1} \end{cases}$$

$$\Delta \mathrm{E}_i^{\max} = \max(|\mathrm{E}_{i+1} - \mathrm{E}_i|, |\mathrm{E}_{i-1} - \mathrm{E}_i|) \tag{4.9}$$

$$\Delta \mathrm{E}_i^{\min} = \min(|\mathrm{E}_{i+1} - \mathrm{E}_i|, |\mathrm{E}_{i-1} - \mathrm{E}_i|)$$

$$\hat{\boldsymbol{\tau}}_{\text{ext}} = \frac{\boldsymbol{\tau}_{\text{ext}}}{|\boldsymbol{\tau}_{\text{ext}}|}$$

Among the images located at extrema, the highest-energy image, known as the climbing image (CI), climbs up toward the first-order saddle point. The CI is guided by the following force defined in Eq. (4.10) employing the $\hat{\boldsymbol{\tau}}_{\text{ext}}$ from Eq. (4.9): [202]

$$\mathbf{F}(\mathbf{r}_{\text{CI}}) = -\nabla \mathrm{E}(\mathbf{r}_{\text{CI}}) + 2(\nabla \mathrm{E}(\mathbf{r}_{\text{CI}}) \cdot \hat{\boldsymbol{\tau}}_{\text{ext}})\hat{\boldsymbol{\tau}}_{\text{ext}} \tag{4.10}$$

Reintroducing part of the perpendicular spring force, $\mathbf{F}^{\perp}_{\text{spring}}$, into the NEB force could improve convergence by preventing chain kinks, as previously suggested. [200] Indeed, the doubly nudged elastic band method that retains part of $\mathbf{F}^{\perp}_{\text{spring}}$ that does not interact with $\mathbf{F}^{\perp}_{\text{PES}}$ performs more efficiently compared to the NEB method. [204] Additionally, adjusting $\mathbf{F}^{\perp}_{\text{spring}}$ based on the curvatures of the PES has shown promising performance enhancements. [249, 281] However, no studies have explored the impact of retaining the full, $\mathbf{F}_{\text{spring}}$, in the overall force when using the improved tangent. [201] Therefore, we include the hybrid band method here to benchmark it against the standard NEB.

Once the force is obtained, an optimization step based on the trust radius quasi-Newton method takes place, following the equation:

$$\boldsymbol{\delta} = -(\mathbf{H} + \lambda I)^{-1}\mathbf{g} \tag{4.11}$$

where $\boldsymbol{\delta}$, $\mathbf{H}$, and $\mathbf{g}$ represent the displacement in Cartesian coordinate, Hessian and gradient, respectively. The initial Hessian is approximated as a diagonal matrix with eigenvalues of 0.05 and is updated using the BFGS method. [51, 52, 53, 54] The regularization parameter $\lambda$ is used to control the step size. It is initially set to zero if the smallest eigenvalue of the Hessian ($\Lambda_{\text{min}}$) is bigger than or equal to an adjustable parameter $\varepsilon$ (with a default value of $1 \times 10^{-5}$). If $\Lambda_{\text{min}}$ is less than $\varepsilon$, $\lambda$ is calculated as $\varepsilon - \Lambda_{\text{min}}$. The root-mean-squared-difference (RMSD) in Cartesian coordinates before and after the step with $\boldsymbol{\delta}$ is calculated using Eq. (5) in the SI to evaluate the step size. If the RMSD is less than 110% of the given trust radius, $\mathbf{r}_{\text{trust}}$ (default is 0.1 Bohr), the algorithm proceeds to take the step and begins the next iteration. Otherwise $\lambda$ is optimized to produce a step length approximately equal to the trust radius as follows:

$$\lambda_{n+1} - \lambda_n = \left(1 - \frac{|\boldsymbol{\delta}|}{\mathbf{r}_{\text{trust}}}\right) \cdot \frac{|\boldsymbol{\delta}|}{\frac{d|\boldsymbol{\delta}|}{d\lambda_n}} = \left(\frac{|\boldsymbol{\delta}|}{\mathbf{r}_{\text{trust}}} - 1\right) \cdot \frac{|\boldsymbol{\delta}|^2}{\boldsymbol{\delta} \cdot (\mathbf{H} + \lambda_n \mathbf{I})^{-1} \cdot \boldsymbol{\delta}} \tag{4.12}$$

The iterations stop when $|\boldsymbol{\delta}|$ and $\mathbf{r}_{\text{trust}}$ agree to within 10%. The optimized $\lambda$ is then used to update the coordinates for the next step, and $\mathbf{r}_{\text{trust}}$ is adjusted based on the step quality, which is evaluated by changes in band energy and gradients, as provided in the SI. The algorithm iteratively updates the chain using step sizes computed from Eqs. (4.11) and (4.12) until the average and maximum RMS Cartesian gradients of all the images fall below $4.86 \times 10^{-4}$ Hartree/Bohr (0.025 eV/Å) and $9.72 \times 10^{-4}$ Hartree/Bohr (0.05 eV/Å), respectively.

### 4.2.3. GeomeTRIC and QCArchive Project

GeomeTRIC is an open-source software initially designed to optimize molecular geometries using translation-rotation-internal coordinates. [65] It interfaces with other quantum chemistry software packages to request SP energy and gradient calculations, then adjust the geometries to approach energy minima. The TS optimization, NEB, and IRC method have been implemented in recent years.

The QCArchive project was launched in 2019 by the Molecular Sciences Software Institute (MolSSI) to accelerate data-driven studies in computational chemistry. [71] QCFractal is the central component of the QCArchive project. [71, 282] It is a server that organizes and stores the input and output of quantum calculations as records in an SQL database. [71, 269, 282] It also supports task distribution and parallel computation through the Python library Parsl [283]. When a SP calculation is submitted through the Python API of QCPortal, input specifications include the name of the quantum chemistry software performing the calculation, the type of calculation (`driver`, such as energy, gradient, or Hessian), and method/basis set employed, all of which are stored as a `QCSpecification`. This specification, along with the initial molecule (`Molecule`), are stored in a `SinglepointRecord`. Once the calculation is completed, the result of the calculation is saved in the same record.

The input specifications for an `OptimizationRecord` (`OptimizationSpecification`), include the name of the program handling the molecular geometry optimization, as well as the `QCSpecification` for gradient calculations. A successfully completed `OptimizationRecord`

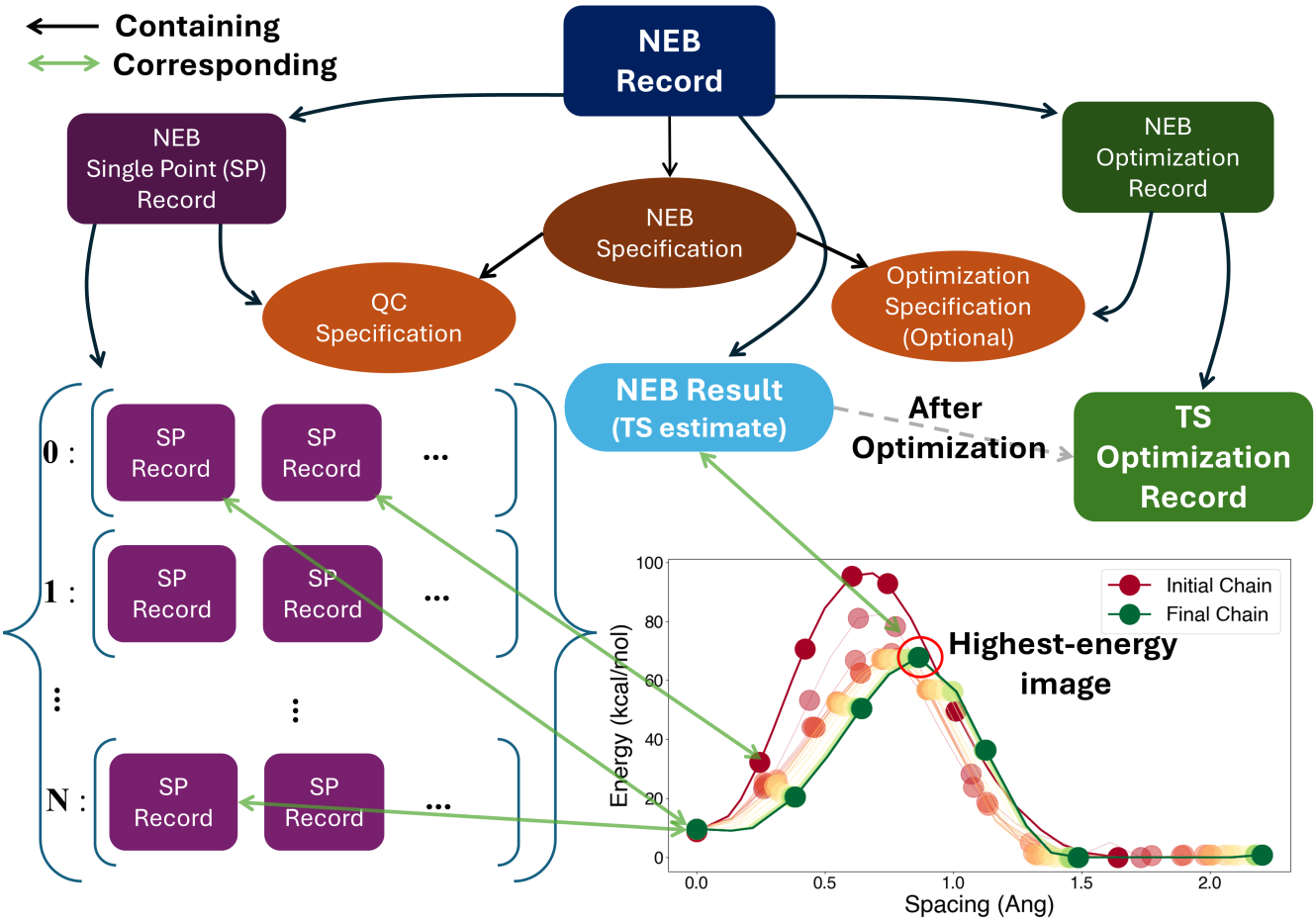



FIGURE 4.4. A schematic plot illustrating the relations between the `NEBRecord` and other associated information. If an `OptimizationSpecification` is not provided, the `QCSpecification` is used as a default to carry the TS optimizations. Additionally, two other `OptimizationRecord` entries, corresponding to the optimizations of the initial chain's endpoints prior to the chain optimization, are available through the `NEBOptimizationRecord`. The energy-versus-spacing plot is the result of a chain optimization for the H-atom transfer reaction from HCN to CNH, calculated with HF/3-21G level of theory.

contains the initial and optimized `Molecule` records, along with the full optimization trajectory (a list of `SinglepointRecord`, each representing a gradient calculation for a specific iteration).

When an NEB calculation is carried out using QCFractal and geomeTRIC, the two endpoints of the initial chain can be optimized first. All the images are optionally aligned to the first image by minimizing the RMSD. After the alignment, energies and gradients of the images are calculated in parallel and the results are sent to geomeTRIC along with their geometries. Using this information, geomeTRIC takes a step to minimize the overall forces, following the procedure described above. The Cartesian coordinates of the new chain are then passed to QCFractal to perform SP calculations for the next iteration. Both the quantum calculations and the NEB step in geomeTRIC are carried out by compute nodes to limit the server's workload.

The various results obtained from a completed NEB calculation, accessible through a `NEBRecord`, are depicted in Figure 4.4. The name of the software used for the chain optimization —currently limited to geomeTRIC— is stored in the `NEBSpecification` along with a `QCSpecification` and an optional `OptimizationSpecification`. The `QCSpecification` is utilized for optimizing the initial chain's endpoints (if enabled) and for performing energy and gradient calculations during the chain optimization. If TS optimization is enabled and the optional `OptimizationSpecification` is provided, the highest-energy image of the final chain is optimized according to the `OptimizationSpecification`. Otherwise, the `QCSpecification` from the chain optimization is reused for TS optimization. This highest-energy image is stored as the `NEBResult`, which is a `Molecule` record.

The `NEBSinglepointRecord` is a dictionary where the keys represent iterations, starting at zero, and the values are lists of `SinglepointRecord` entries for each images. A list of `Molecule` records for the initial chain and a list of `SinglepointRecord` entries for the final chain are also available from a completed `NEBRecord`.

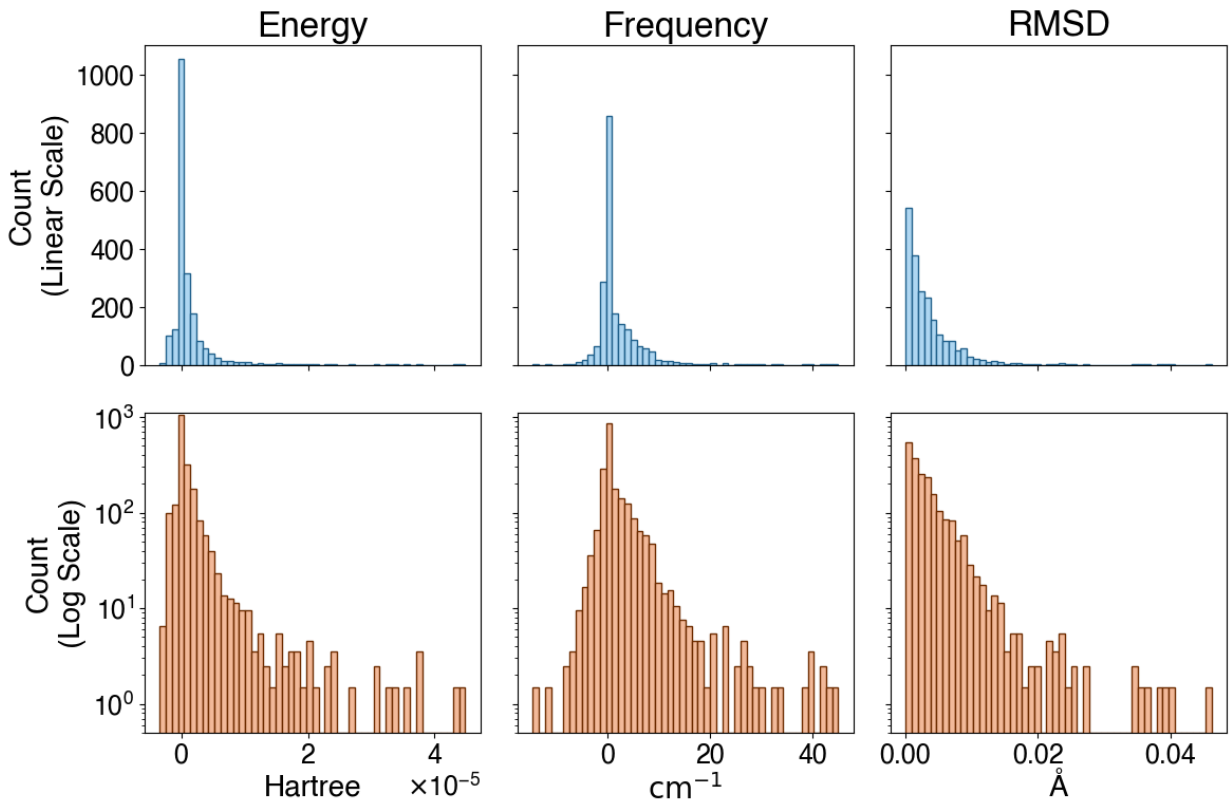



FIGURE 4.5. Histograms showing the distributions of energies, imaginary mode frequencies, and RMSDs (relative to the optimized reference TS structure) for the TS structures considered converged of the four RGD1 reactions. The first row is plotted on a linear scale, whereas the second row uses a logarithmic scale with a +0.5 shift applied to the y-values to make counts of one visible. No peaks other than those corresponding to the reference structure are observed.

## 4.3. Results and discussion

### 4.3.1. RGD1 dataset reactions

The convergence rates of both chain and TS optimizations, along with the number of SP calculations carried out for the chain optimizations, are presented here. Summaries of the results for each reaction are displayed in Figures S1 to S4 in the SI. The averages and standard deviations for the number of SP calculations were computed using only those converged within 300 iterations. The highest energy images of the final chain, regardless of the chain

TABLE 4.3. The chain and TS convergence rates, reported as percentages, for the RGD1 reactions are computed over selected parameter ranges: spring constants from 0.1 to 100 kcal/mol/Å$^2$, with 11, 15, and 21 number of images, and with NEB and alignment enabled. The values in parentheses are from the hybrid band while keeping the other parameters the same.

| | RGD1 reactions | B3LYP (%) | B3LYP-D3 (%) | PM7//B3LYP (%) | PM7//B3LYP-D3 (%) |
|---|---|---|---|---|---|
| Chain | RGD1-A | 57.1 (95.2) | 81.0 (95.2) | 85.7 (95.2) | 85.7 (95.2) |
| | RGD1-B | 19.0 (52.4) | 52.4 (71.4) | 76.2 (90.5) | 76.2 (90.5) |
| | RGD1-C | 100.0 (100.0) | 100.0 (95.2) | 100.0 (95.2) | 100.0 (100.0) |
| | RGD1-D | 100.0 (100.0) | 100.0 (100.0) | 33.3 (42.9) | 33.3 (38.1) |
| TS | RGD1-A | 19.0 (19.0) | 61.9 (57.1) | 90.5 (100.0) | 90.5 (100.0) |
| | RGD1-B | 33.3 (61.9) | 47.6 (66.7) | 81.0 (100.0) | 85.7 (100.0) |
| | RGD1-C | 100.0 (100.0) | 71.4 (81.0) | 100.0 (100.0) | 100.0 (100.0) |
| | RGD1-D | 100.0 (100.0) | 100.0 (100.0) | 90.5 (81.0) | 90.5 (81.0) |

optimization convergence state, were optimized to locate the first-order saddle points. The final TS structures are considered successfully converged if they have only one imaginary mode frequency, and if their energies, imaginary mode frequencies, and geometries (RMSD) agree within 0.001 Hartree, 100 cm$^{-1}$, and 0.1 Bohr (0.0529 Å), respectively, with those of the optimized reference TS structures. The energies and imaginary mode frequencies of the reference TS structures for the RGD1 reactions are presented in Table S2 in the SI. To verify the tightness of the convergence criteria and prevent misidentification of distinct first-order saddle points as the same converged structure, the distributions of energies, imaginary mode frequencies, and RMSDs for the converged structures were examined. Figure 4.5 displays these distributions. All three distributions are concentrated around the values of the reference structures, with only a few outliers. Given the small magnitudes of the energies, frequencies, and RMSDs for these outliers, the convergence criteria are sufficiently tight to avoid such misidentification.

Herein, we focus primarily on the NEB and hybrid band types that were aligned, as alignment has little impact on performance, while NEB and hybrid band performed significantly better than the plain band. We further screened the other parameters as follows to calculate

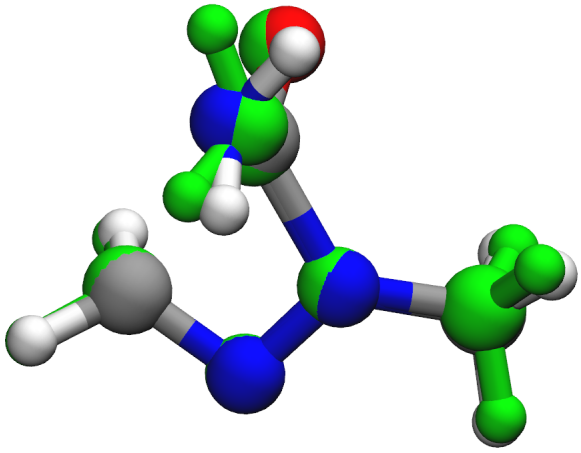


FIGURE 4.6. The molecule highlighted in green represents the optimized TS structure of RGD1-A, while the other structure corresponds to a newly detected TS. The new structure has an imaginary vibrational frequency of -725.51 $cm^{-1}$ and an energy 2.09 kcal/mol higher than that of the reference. These values were calculated using the B3LYP/6-31G(d) level of theory. The RMSD between the two structures is 0.2053 Bohr (0.1086 Å).

the overall convergence rates: spring constants ranging from 0.1 to 100 kcal/mol/$Å^2$, and the number of images set to 11, 15, and 21. The calculated chain and TS convergence rates are summarized in Table 4.3.

Among the chain convergence results, the RGD1-C reaction exhibits the highest convergence rates across all methods, while the RGD1-B shows the lowest. The relatively low TS convergence rates of RGD1-B stem from its low chain convergence rates. Generating the initial chain for the RGD1-B reaction by interpolating between optimized endpoints rather than between the raw structures may improve these results. The very low TS convergence rate of 19% of the reaction RGD1-A, when DFT was used to optimize the chain, are due to the final TS structures converging to another first-order saddle point. The structure of the newly found TS, along with its relative energy and imaginary vibrational frequency, is shown in Figure 4.6. When the TS convergence RMSD and the energy criterion are relaxed from 0.1 to 0.21 Bohr and from 0.001 to 0.01 Hartree, respectively, while keeping the other

unchanged, B3LYP TS convergence achieves 81.0% (17/21) and B3LYP-D3 achieves 90.5% (19/21) for the RGD1-A reaction using NEB, while both achieve 100.0% (21/21) with the hybrid band.

The IRC method was performed for the final TS structures that did not match with the reference structures. All IRC calculations, starting from the different TS structure presented in Figure 4.6, resulted in alternate conformers of the RGD1-A reactant and product. Some of the other IRC results led to different reactants and products, as shown in Figures S5 to S7, or were not true TSs of chemical reactions. These alternative reactions, along with all other non-converged cases, are described in detail in the SI (Section 4: Non-converged Cases and Newly Discovered Reactions). The reaction RGD1-C had the highest convergence rates, in spite of having the largest number of molecules involved. Both the NEB and hybrid band struggled to converge the chain for the RGD1-D reaction when using PM7. Interestingly, chain optimizations using PM7 achieved higher chain convergence rates with a lower number of images (Figure S4). This is most likely due to a combination of the reaction's topology, which involves a small RMSD between the reactant and product, and the shape of the PM7 PES. We hypothesize that very closely spaced images could lead to kink formation and non-convergence because the tangent direction becomes highly sensitive to small displacements and the spring forces are very small; approximations in the shape of the PES from the semiempirical model could exacerbate this problem.

The chain convergence rates for PM7//B3LYP and PM7//B3LYP-D3 should be identical, as they share the same inputs for the chain optimization. However, minor discrepancies are observed in the RGD1-C and RGD1-D results. This could be caused by floating-point non-associativity, which prevents the precise reproduction of the same calculations. [284, 285] This issue is more apparent when coupled with asynchronous parallel computations, [285] which is how the chain optimizations in this study were conducted, thereby magnifying the

discrepancies. Overall, the hybrid band performed better than NEB, and Figures S1 through S4 also support this result.

Figure 4.7 illustrates the NEB method's performance across the four RGD1 reactions with selected parameters, while Figure 4.8 presents the same results using the hybrid band. The number of SP calculations shown here are averages and standard deviations computed across all calculations that used the indicated parameter (e.g. for the bar group titled "11 images", there are 36 calculations per bar corresponding to 9 spring constant $\times$ 4 reactions). The same number of data points are used to calculate the convergence rates. Both band types exhibit similar trends, but the hybrid band outperforms the NEB based on overall chain and TS convergence rates. The number of SP calculation performed during chain optimization and the chain convergence rates do not vary significantly across spring force constants up to 100 kcal/mol/Å$^2$. However, spring constants above 100 kcal/mol/Å$^2$ result in more SP calculations and lower chain convergence rates. This is likely because high spring constants lead to higher spring forces and shrink the basin of convergence as a result. The number of SP calculations increases with the number of images, but also improves TS convergence rates by providing better resolutions near first-order saddle points. While the highest convergence rate was achieved using 21 images, the number of SP calculations was also the highest. Lowering the number of images to 15 or 11 decreases the success rate of the NEB calculations but also lowers the number of SP calculations for successful NEB runs; therefore, it may be employed as a compromise if computational resources are limited. Also, using fewer images for reactions involving small RMSDs between the reactant and product may perform better than using 21 images. However, in general, using only 7 images dramatically decreases the convergence rate and is not recommended.

Based on the overall performance, a spring constant of 1.0 kcal/mol/Å$^2$ and 21 images was chosen as a "recommended" parameter set, and employed in the next section to optimize 118 reactions from the BH dataset with both band types. Table 4.4 presents the average number

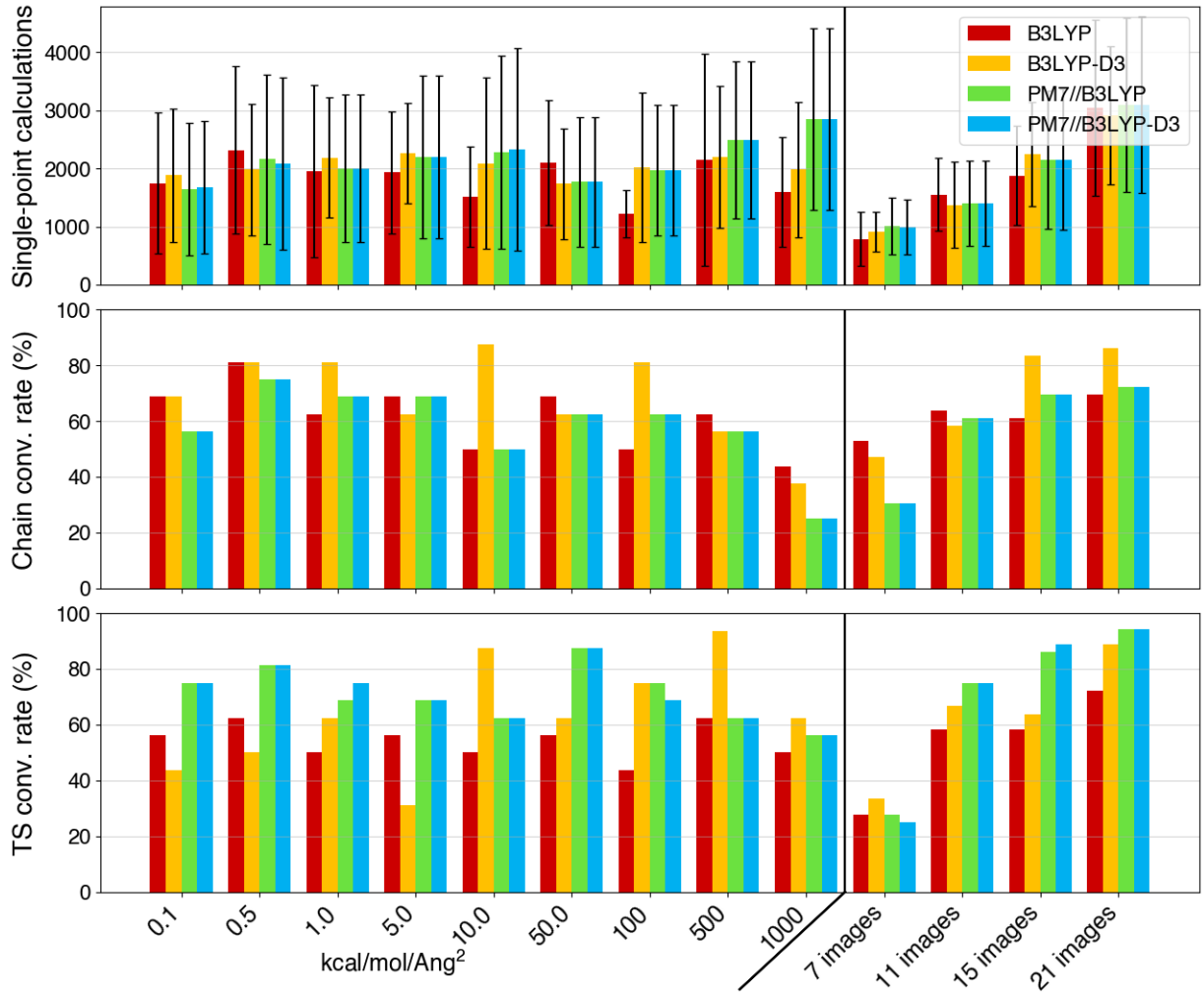



FIGURE 4.7. Averages and standard deviations of the number of single-point calculations performed during chain optimizations, excluding those that did not converge within 300 iterations, along with convergence rates for chain and TS optimizations for the four RGD1 reactions. These results are from the NEB spring type with alignment enabled, spring constants from 0.1 to 1,000 kcal/mol/Å$^2$, and image counts ranging from 7 to 21. Each bar for spring constants represents result from 16 data points (4 image counts × 4 reactions), while each bar for number of images is from 36 data points (9 spring constants × 4 reactions). Approximated TS structures from final chains based on DFT calculations were optimized at the same levels of theory (red and yellow bars), while those from the PM7 semiempirical model were optimized at both DFT levels (green and blue bars).

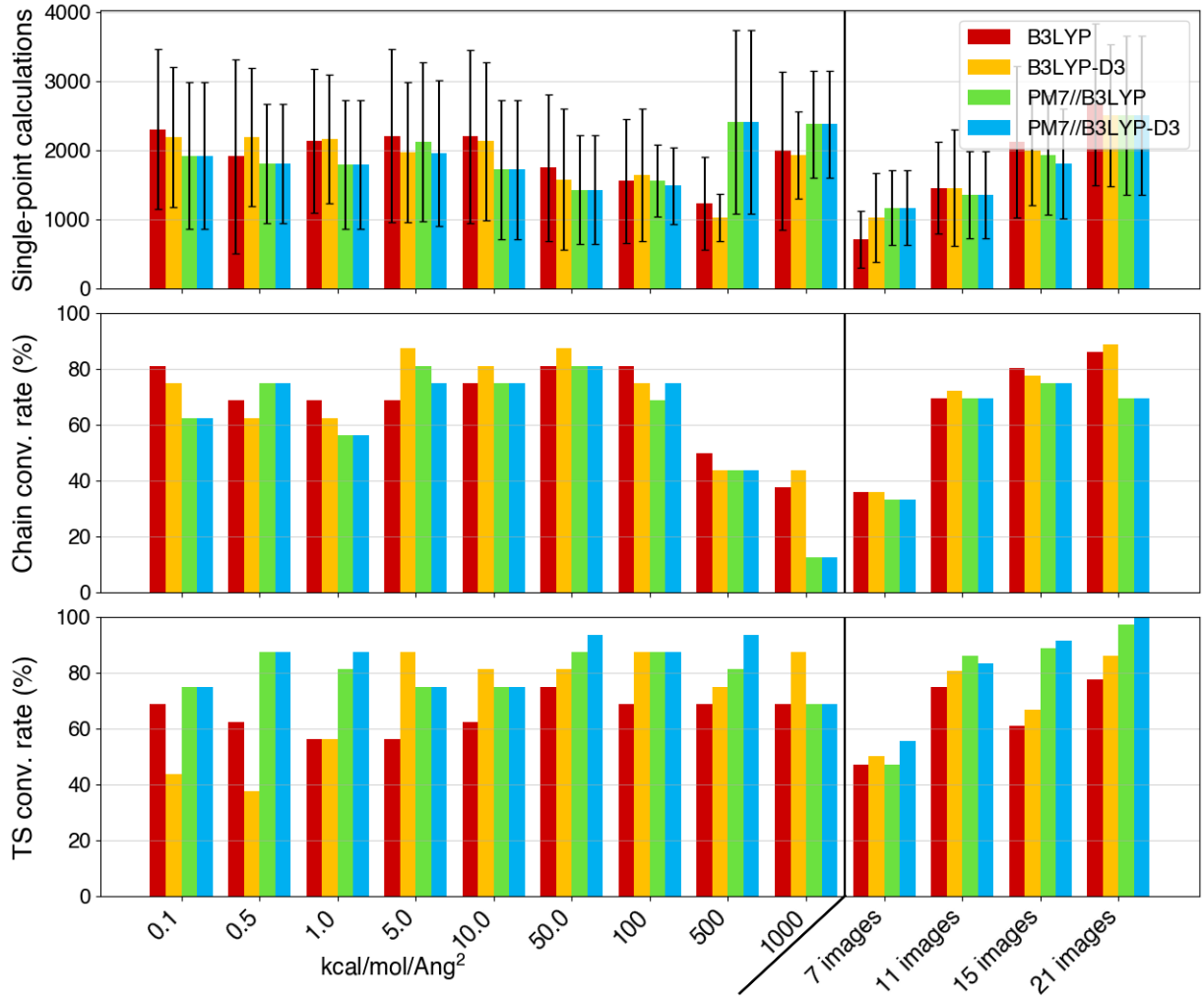



FIGURE 4.8. Averages and standard deviations of the number of single-point calculations performed during chain optimizations along with convergence rates for chain and TS optimizations for the four RGD1 reactions. These results are from the hybrid band spring type with alignment enabled, spring constants from 0.1 to 1,000 kcal/mol/Å$^2$, and image counts ranging from 7 to 21. See Figure 4.7 for other details.

of chain optimization iterations (excluding those that did not converge within the maximum 300 iterations) and the number of converged calculations for both chain and TS optimizations across the four sets of levels of theory for this recommended parameter set. Although the iteration counts varied significantly depending on the reaction type, the majority of chain and TS optimizations converged successfully. The three non-converged TS structures from

TABLE 4.4. The average number of chain optimization iterations and the number of converged results out of four calculations at different levels of theory, using a spring force constant of 1.00 kcal/mol/Å$^2$, 21 images, and aligned initial chains, are shown.

| | NEB | | | Hybrid band | | |
|---|---|---|---|---|---|---|
| | Iteration | Chain conv. | TS conv. | Iteration | Chain conv. | TS conv. |
| RGD1-A | 199.3 | 4/4 | 3/4 | 198.3 | 4/4 | 2/4 |
| RGD1-B | 133.3 | 3/4 | 4/4 | 100.5 | 4/4 | 4/4 |
| RGD1-C | 56.0 | 4/4 | 4/4 | 57.3 | 4/4 | 4/4 |
| RGD1-D | 112.5 | 2/4 | 4/4 | 99.0 | 2/4 | 4/4 |

the RGD1-A cases (one from NEB and two from the hybrid band) correspond to the alternate TS structure shown in Figure 4.6. All four non-converged chain optimizations in RGD1-D came from PM7, while the single failed case in RGD1-B came from B3LYP.

### 4.3.2. BH dataset reactions

The selected sets of chain optimization parameters were used to optimize 118 linearly interpolated BH reactions using two approaches: the B3LYP/6-31G(d) (B3LYP) level of theory for both chain and TS optimizations, and a combination of the PM7 semiempirical model for chain optimizations with B3LYP/6-31G(d) for TS optimizations (PM7//B3LYP). This resulted in 236 calculations for NEB and 236 for hybrid band. Details regarding the datasets and the number of reactions from each dataset are provided in Table 4.2. The final optimized TS structures were evaluated for convergence based on the same criteria applied to the RGD1 reactions. Table 4.5 presents the NEB and TS convergence rates. Calculations that used B3LYP for both the NEB and TS steps had a higher convergence rate compared to the PM7//B3LYP hybrid calculations that used different methods in the two steps. Of the 37 unsuccessful B3LYP results, 15 converged to alternative saddle points. Similarly, 18 of the 40 failed PM7//B3LYP calculations converged to different TS structures. Other convergence failures were attributed to SCF iteration issues in Q-Chem, MOPAC errors due to atoms being too close, failures to converge within the maximum number of TS optimization

TABLE 4.5. The overall chain optimization and TS convergence rates for the 118 reactions in the BH dataset using NEB and hybrid band methods are shown. Results in parentheses indicate hybrid band outcomes that differ from the NEB results.

| | B3LYP | | PM7//B3LYP | |
|---|---|---|---|---|
| | Chain(%) | TS(%) | Chain(%) | TS(%) |
| BH76 | 88.2 | 82.4 (88.2) | 58.8 (70.6) | 41.2 (29.4) |
| BHPERI | 100.0 | 100.0 | 96.15 | 96.15 (100.0) |
| BHDIV10 | 85.7 | 71.4 | 85.7 (100.0) | 42.9 (57.1) |
| INV24 | 100.0 | 83.3 | 100.0 | 83.3 (88.9) |
| BHROT27 | 94.7 | 78.9 | 73.7 (68.4) | 52.6 (57.9) |
| PX13 | 100.0 | 100.0 | 100.0 | 92.3 |
| WCPT18 | 100.0 | 100.0 | 100.0 | 100.0 |
| Total | 96.6 | 89.8 (90.7) | 88.1 (89.8) | 76.3 (78.0) |

iterations, or final structures exhibiting no or excess imaginary modes. Overall, both the NEB and hybrid band achieved high convergence rates, with the hybrid band performing slightly better. Adopting more advanced interpolation methods beyond linear interpolation could further increase these rates.

### 4.3.3. Selected reactions

The overall reaction coordinates, including relative energies obtained in this study and comparisons with previous work, are presented in Figure 4.9. The 3D model structures shown in the figure are the results from this study. All the xyz files of the optimized reactants, transition states, and products used for plotting, from both the previous and this study, are provided in the SI. The first-order saddle points for all three elementary reactions were successfully identified and verified based on their imaginary vibrational modes. The relative energies of the three reaction from this work and the previous work agree within $\sim$ 1 kcal/mol. The geometries of $O_2$-displacement products show the largest discrepancies (RMSD) among the critical point geometries, as detailed in Table 4.6. The RMSD values for the TS structures of the $Fe_4N$-bridge and $Fe_4N$-dissociate reactions are higher than those of the reactant and product structures. These differences are from the use of different geometry

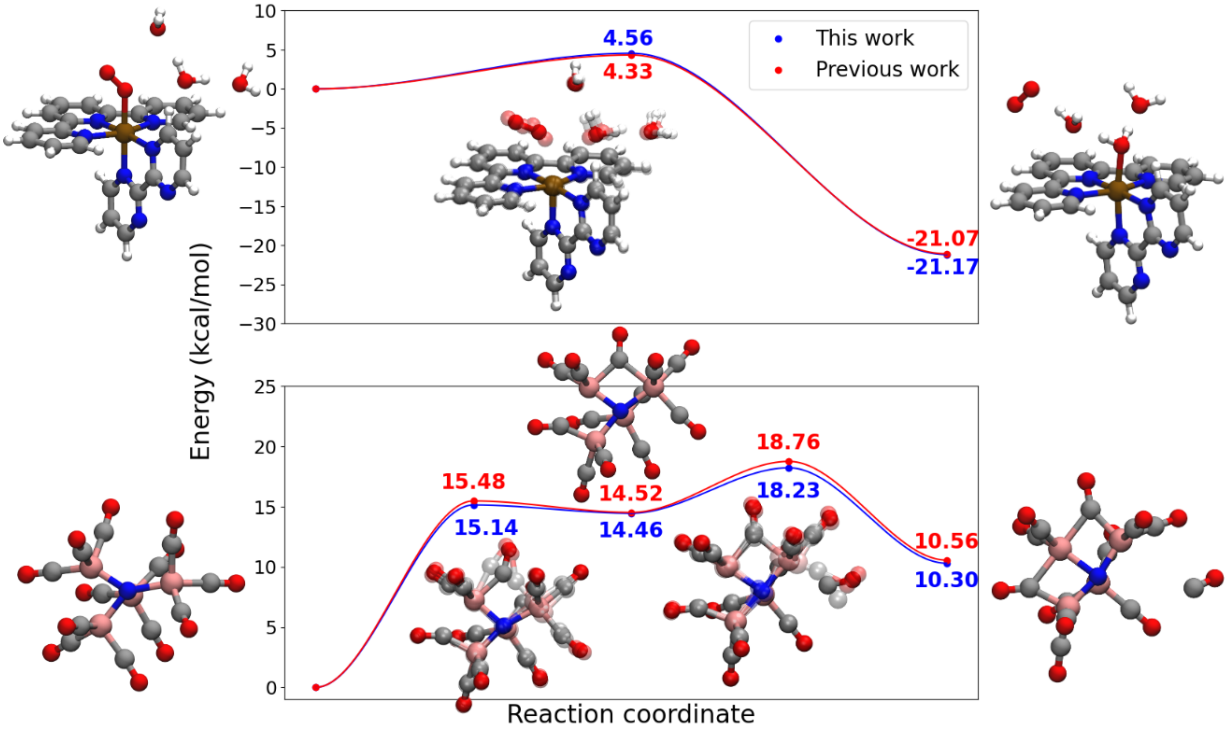



FIGURE 4.9. The endpoint 3D geometry models are positioned at the edges of the plot, while the other structures, including the TS structures with their imaginary vibrational modes, are displayed in the center. The curves do not represent the actual shapes of the PES but are fitted lines. Atoms are colored as follows: nitrogen (blue), oxygen (red), carbon (gray), hydrogen (white), iron (pink) and ruthenium (brown). The results for the $O_2$-displacement reaction (top) in this work agree within 0.3 kcal/mol with those from the previous study. The other two reactions, $Fe_4N$-bridge and $Fe_4N$-dissociate (bottom), exhibit larger differences between the two sets of results. The greatest discrepancy is from the TS structures of the $Fe_4N$-dissociate reaction, but it is no larger than 0.6 kcal/mol. Overall, the results from the two studies agree well, confirming the reliability of the chain optimization algorithm, as the previous study located the TS structures using constrained optimizations and modeling.

optimization convergence criteria in two software packages, Q-Chem and geomeTRIC. Since energies can be particularly sensitive to geometries near first-order saddle points, an RMSD value of $\sim 0.08$Å corresponded to an energy difference of $\sim 0.3$ kcal/mol for the $Fe_4N$-bridge reaction, and a $\sim 0.5$ kcal/mol difference for the $Fe_4N$-dissociate reaction. Despite these

| Reactions | Reactant | Transition State | Product |
|---|---|---|---|
| $O_2$-displacement | 0.02632 | 0.04482 | 0.08219 |
| $Fe_4N$-bridge | 0.00016 | 0.08183 | 0.01432 |
| $Fe_4N$-dissociate | 0.00036 | 0.07705 | 0.05063 |

TABLE 4.6. The RMSD values between the optimized geometries from the previous and present work is reported, with all numbers in Å.

minor discrepancies, the results from this study and the previous work show strong overall agreement.

## 4.4. Conclusion

In this study, we introduced a high-throughput approach capable of extensive parallel chain and TS optimizations, with all submitted and completed calculations stored in a database. Key features include the automation of TS optimization upon chain optimization completion and the use of different levels of theory for chain and TS optimizations. We tested this approach on four reactions from the RGD1 dataset and 72 combinations of chain optimization parameters —spring constant, number of images, and chain alignment— and three different band types resulting in 3,456 chain optimization calculations using B3LYP/6-31G(d), B3LYP-D3/def2-TZVP, and the PM7 semiempirical model. All the initial chains processed in this study were generated through linear interpolations between reactants and products. The TS structures approximated with the PM7 model were subsequently optimized with the two DFT methods, resulting in 3,456 TS optimizations. Energies, geometries, and imaginary modes of the final TS structures were compared with the optimized geometries from RGD1 dataset to assess convergence rates. Overall, both chain and TS optimizations demonstrated high convergence rates, except for TS convergence rate of RGD1-A, where many of the final TS structures converged to alternative first-order saddle points.

Building on these results, we selected two sets of chain optimization parameters to optimize 118 reactions from the BH dataset. Initial chains were optimized using B3LYP/6-31G(d) and the PM7. The approximated TS structures from both approaches were further

refined using B3LYP/6-31G(d). The TS convergence rate of the NEB method was 89.8% (90.7% for hybrid band) when B3LYP was used for both chain and TS optimizations, and 76.3% (78.0% for hybrid band) when PM7 was used for chain optimization and B3LYP/6-31G(d) for TS optimizations. The combined results from the RGD1 and BH dataset reactions show that the hybrid band surpasses NEB in performance. Additional calculations on three transition metal-catalyzed energy conversion reactions, involving larger systems, were carried out, and the results from this study align well with previous findings. Based on the overall results of this study, we recommend using this implementation with 21 images and the hybrid band, while keeping the spring constant at or below 100 kcal/mol/Å$^2$. When computational resources are limited, a smaller number of images (no fewer than 11) and a semiempirical model can be used to optimize the chains, followed by TS optimization at a higher level of theory.

This approach, implemented using the open-source software packages geomeTRIC and QCFractal, is readily available for broader use. It enables systematic studies of how chain optimization parameters influence performance, as demonstrated here, and supports the simultaneous search for MEPs across numerous reactions by leveraging parallel computing and efficient data storage. Future developments will focus on enabling internal coordinate steps for chain optimization, and incorporating advanced features of the NEB method. Furthermore, the approach will be integrated into an automated workflow capable of refining and identifying reaction pathways from first-principles molecular dynamics simulation trajectories.

## 4.5. Supplementary Information

Supplementary information is available through the journal article.

# Chapter 5

# Summary and Future Work

This dissertation discusses computational chemistry, specifically the development of reliable and efficient methods to investigate and predict molecular behaviors in chemical reactions. In Chapter 1, a broad range of foundational concepts is introduced. First, various approximations to the time-independent Schrödinger equation are described that calculate energies and forces of a given molecule. Second, the transition from single-point energy and gradient calculations to full ab initio molecular (AIMD) simulation trajectories is explained using the Velocity Verlet method. Lastly, details are provided on how high-energy simulation trajectories can be refined to extract the minimum energy pathways (MEPs) via an automated refinement workflow.

In Chapter 2, AIMD simulation and the automated trajectory refinement workflow are applied to explore astrochemical reactions. Three barrierless/low-barrier nitrogen-containing heterocycle (N-heterocycle) formation pathways were discovered and refined using the approach. N-heterocycles are considered prebiotic molecules as they are building blocks of genetic materials. The simulation was leveraged in a retrosynthetic manner, starting from the desired products and bombarding them with He atoms to discover potential high-energy reactants. Two pathways are ring-expansion reactions in which a carbon chain molecule forms a bond with a ring molecule to insert an atom into the ring. The third involves two carbon chains undergoing a ring-formation reaction. The results demonstrate the usefulness of the retrosynthetic nanoreactor simulation in discovering astrochemically relevant prebiotic molecules.

Chapter 3 presents a novel interpolation method (TRICS) that linearly interpolates internal coordinates (ICs) between reactants and products. A key challenge is that a set of ICs defined at the starting point can be reliably mapped back to Cartesian coordinates within

its original configuration space, but the conversion becomes unreliable near the endpoint in a different configuration space. This occurs because a well-defined, non-redundant set of ICs that is valid in one configuration space does not necessarily map accurately to Cartesian coordinates in another (for example, when a proper angle defined at the starting point becomes linear at the destination). The core idea of TRICS is to leverage multiple sets of ICs defined over different segments of the path, where each set remains non-redundant and well-defined. It interpolates a trajectory for each segment, and the disconnection points are smoothed iteratively by exchanging target structures with the adjacent trajectory and re-interpolating (or splicing) the segments together. Its performance is equivalent to, if not superior to, other existing approaches.

Chapter 4 introduces a high-throughput nudged elastic band (NEB) approach. The implementation integrates two open-source software projects, geomeTRIC and QCArchive. QCArchive is a cloud-based platform whose component QCFractal runs a server that distributes large numbers of quantum chemistry calculations to remotely connected compute nodes while saving the data. The stored data can then be queried seamlessly by other researchers using a Python script and an internet connection. The implementation was tested with over 3,000 NEB calculations, and the results confirm its reliability and correctness. This platform can be leveraged to generate and store a large amount of training data for either machine learning models or algorithm development.

Future work will focus on three main directions. First, additional prebiotic formation pathways, such as glycolamide and ribose, will be explored, along with further expansion of N-heterocycle formation pathways in astrochemical environments. Second, TRICS will be refined based on benchmark results against other widely used interpolation approaches, with particular attention to handling cases where atoms approach too closely or drift too far apart. Finally, advanced NEB algorithms will be implemented, including energy-weighted springs,

internal-coordinate steps, and Hessian approximations. Each improvement will be rigorously validated within the high-throughput framework. Once integrated, these algorithms will finalize the automated refinement workflow built on QCArchive, enabling massive parallelization and seamless storage and sharing of the calculation results. This workflow will accelerate the systematic discovery of molecular reactivity across astrochemistry, prebiotic chemistry, and beyond.

# Bibliography


[1] E. Schrödinger, "Quantisierung als Eigenwertproblem," *Annalen der Physik*, vol. 384, no. 4, pp. 361–376, 1926. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/andp.19263840404

[2] M. Born and P. Jordan, "Zur quantenmechanik," *Zeitschrift für Physik*, vol. 34, pp. 858–888, 1925.

[3] M. Born, W. Heisenberg, and P. Jordan, "Zur quantenmechanik ii," *Zeitschrift für Physik*, vol. 35, pp. 557–615, 1926.

[4] E. Schrödinger, "An Undulatory Theory of the Mechanics of Atoms and Molecules," *Physical Review*, vol. 28, no. 6, pp. 1049–1070, December 1926. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRev.28.1049

[5] C. Eckart, "Operator Calculus and the Solution of the Equations of Quantum Dynamics," *Physical Review*, vol. 28, no. 4, pp. 711–726, October 1926. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRev.28.711

[6] M. Born and R. Oppenheimer, "Zur Quantentheorie der Molekeln," *Annalen der Physik*, vol. 389, no. 20, pp. 457–484, 1927. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/andp.19273892002

[7] A. Szabo and N. S. Ostlund, *Modern Quantum Chemistry: Introduction to Advanced Electronic Structure Theory*. Mineola, New York: Dover Publications, 1996, pp. 45—70.

[8] W. Gerlach and O. Stern, "Der experimentelle Nachweis der Richtungsquantelung im Magnetfeld," *Zeitschrift für Physik*, vol. 9, no. 1, pp. 349–352, December 1922. [Online]. Available: https://doi.org/10.1007/BF01326983

[9] W. Pauli, "Über den zusammenhang des abschlusses der elektronengruppen im atom mit der komplexstruktur der spektren," *Zeitschrift für Physik*, vol. 31, pp. 765–783, 1925.

[10] T. Helgaker, P. Jørgensen, and J. Olsen, *Molecular Electronic-Structure Theory*. Chichester: Wiley, 2000, pp. 450—451.

[11] B. O. Roos, P. R. Taylor, and P. E. M. Siegbahn, "A complete active space scf method (casscf) using a density matrix formulated super-ci approach," *Chemical Physics*, vol. 48, no. 2, pp. 157–173, 1980.

[12] C. Møller and M. S. Plesset, "Note on an approximation treatment for many-electron systems," *Physical Review*, vol. 46, no. 7, pp. 618–622, 1934.

[13] J. Čížek, "On the correlation problem in atomic and molecular systems," *The Journal of Chemical Physics*, vol. 45, pp. 4256–4266, 1966.

[14] K. Raghavachari, G. W. Trucks, J. A. Pople, and M. Head-Gordon, "A fifth-order perturbation comparison of electron correlation theories," *Chemical Physics Letters*, vol. 157, pp. 479–483, 1989.

[15] M. Challacombe and E. Schwegler, "Linear scaling computation of the fock matrix," *J. Chem. Phys.*, vol. 106, no. 13, pp. 5526–5536, 1997.

[16] C. Ochsenfeld, C. A. White, and M. Head-Gordon, "Linear and sublinear scaling formation of hartree–fock-type exchange matrices," *J. Chem. Phys.*, vol. 109, no. 5, pp. 1663–1669, 1998.

[17] A. M. N. Niklasson, C. J. Tymczak, and M. Challacombe, "Trace resetting density matrix purification in o(n) self-consistent-field theory," *J. Chem. Phys.*, vol. 118, no. 19, pp. 8611–8620, 2003.

[18] L. H. Thomas, "The calculation of atomic fields," *Mathematical Proceedings of the Cambridge Philosophical Society*, vol. 23, no. 5, pp. 542–548, 1927.

[19] E. Fermi, "Un metodo statistico per la determinazione di alcune proprietà dell'atomo," *Rendiconti della Accademia dei Lincei*, vol. 6, pp. 602–607, 1927.

[20] P. A. M. Dirac, "Note on exchange phenomena in the Thomas atom," *Mathematical Proceedings of the Cambridge Philosophical Society*, vol. 26, no. 3, pp. 376–385, 1930. [Online]. Available: https://doi.org/10.1017/S0305004100016108

[21] A. D. Becke, "Perspective: Fifty years of density-functional theory in chemical physics," *The Journal of Chemical Physics*, vol. 140, no. 18, p. 18A301, 2014.

[22] P. Hohenberg and W. Kohn, "Inhomogeneous electron gas," *Physical Review*, vol. 136, no. 3B, pp. B864–B871, nov 1964. [Online]. Available: https://doi.org/10.1103/PhysRev.136.B864

[23] W. Kohn, "Electronic structure of matter—wave functions and density functionals," *Reviews of Modern Physics*, vol. 71, no. 5, pp. 1253–1266, 1999, nobel Lecture.

[24] R. G. Parr and W. Yang, *Density-Functional Theory of Atoms and Molecules*. New York: Oxford University Press, 1989.

[25] J. P. Perdew and K. Schmidt, "Jacob's ladder of density functional approximations for the exchange-correlation energy," in *Density Functional Theory and Its Application to Materials*, ser. AIP Conference Proceedings, V. Van Doren, C. Van Alsenoy, and P. Geerlings, Eds., vol. 577. Melville, NY: American Institute of Physics, 2001, pp. 1–20.

[26] J. P. Perdew, K. Burke, and M. Ernzerhof, "Generalized Gradient Approximation Made Simple," *Physical Review Letters*, vol. 77, no. 18, pp. 3865–3868, 1996. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevLett.77.3865

[27] A. D. Becke, "Density-functional exchange-energy approximation with correct asymptotic behavior," *Physical Review. A, General Physics*, vol. 38, no. 6, pp. 3098–3100, September 1988.

[28] C. Lee, W. Yang, and R. G. Parr, "Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density," *Physical Review B*, vol. 37, no. 2, pp. 785–789, January 1988. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevB.37.785

[29] J. P. Perdew, J. A. Chevary, S. H. Vosko, K. A. Jackson, M. R. Pederson, D. J. Singh, and C. Fiolhais, "Atoms, molecules, solids, and surfaces: Applications of the generalized gradient approximation for exchange and correlation," *Physical Review B*, vol. 46, no. 11, p. 6671, 1992.

[30] J. Tao, J. P. Perdew, V. N. Staroverov, and G. E. Scuseria, "Climbing the density functional ladder: Nonempirical meta–generalized gradient approximation designed for molecules and solids," *Physical Review Letters*, vol. 91, no. 14, p. 146401, 2003.

[31] J. P. Perdew, S. Kurth, A. Zupan, and P. Blaha, "Accurate density functional with correct formal properties: A step beyond the generalized gradient approximation," *Physical Review Letters*, vol. 82, no. 12, pp. 2544–2547, 1999.

[32] J. P. Perdew and L. A. Constantin, "Laplacian-level density functionals for the kinetic energy density and exchange-correlation energy," *Physical Review B*, vol. 75, p. 155109, 2007.

[33] C. Adamo and V. Barone, "Toward reliable density-functional methods without adjustable parameters: The PBE0 model," *The Journal of Chemical Physics*, vol. 110, no. 13, pp. 6158–6170, April 1999. [Online]. Available: https://doi.org/10.1063/1.478522

[34] A. D. Becke, "Density-functional thermochemistry. III. The role of exact exchange," *The Journal of Chemical Physics*, vol. 98, no. 7, pp. 5648–5652, April 1993. [Online]. Available: https://doi.org/10.1063/1.464913

[35] S. H. Vosko, L. Wilk, and M. Nusair, "Accurate spin-dependent electron liquid correlation energies for local spin density calculations: a critical analysis," *Canadian Journal of Physics*, vol. 58, no. 8, pp. 1200–1211, 1980.

[36] S. Grimme, "Semiempirical hybrid density functional with perturbative second-order correlation," *The Journal of Chemical Physics*, vol. 124, no. 3, p. 034108, 2006.

[37] J.-D. Chai and M. Head-Gordon, "Systematic optimization of long-range corrected hybrid density functionals," *The Journal of Chemical Physics*, vol. 128, no. 8, p. 084106, 2008.

[38] O. A. Vydrov and T. Van Voorhis, "Nonlocal van der waals density functional: The simpler the better," *The Journal of Chemical Physics*, vol. 133, no. 24, p. 244103, 2010.

[39] S. Grimme, S. Ehrlich, and L. Goerigk, "Effect of the damping function in dispersion corrected density functional theory," *Journal of Computational Chemistry*, vol. 32, no. 7, pp. 1456–1465, 2011.

[40] Y. Zhao and D. G. Truhlar, "The m06 suite of density functionals for main group thermochemistry, thermochemical kinetics, noncovalent interactions, excited states, and transition elements: two new functionals and systematic testing of four m06-class functionals and 12 other functionals," *Theoretical Chemistry Accounts*, vol. 120, no. 1-3, pp. 215–241, 2008.

[41] H. S. Yu, X. He, S. L. Li, and D. G. Truhlar, "Mn15: A kohn–sham global-hybrid exchange–correlation density functional with broad accuracy for multi-reference and single-reference systems and noncovalent interactions," *Chemical Science*, vol. 7, no. 8, pp. 5032–5051, 2016.

[42] L. Verlet, "Computer "experiments" on classical fluids. I. thermodynamical properties of lennard-jones molecules," *Physical Review*, vol. 159, no. 1, pp. 98–103, 1967.

[43] W. C. Swope, H. C. Andersen, P. H. Berens, and K. R. Wilson, "A computer simulation method for the calculation of equilibrium constants for the formation of physical clusters of molecules: Application to small water clusters," *The Journal of Chemical Physics*, vol. 76, no. 1, pp. 637–649, 1982.

[44] M. Tuckerman, B. J. Berne, and G. J. Martyna, "Reversible multiple time scale molecular dynamics," *The Journal of Chemical Physics*, vol. 97, no. 3, pp. 1990–2001, 1992.

[45] M. E. Tuckerman, *Statistical Mechanics: Theory and Molecular Simulation*. Oxford: Oxford University Press, 2010, chapter 15.

[46] L.-P. Wang, A. Titov, R. McGibbon, F. Liu, V. S. Pande, and T. J. Martínez, "Discovering chemistry with an ab initio nanoreactor," *Nature Chemistry*, vol. 6, no. 12, pp. 1044–1048, 2014.

[47] E. Wigner, "The transition state method," *Transactions of the Faraday Society*, vol. 34, no. 0, pp. 29–41, 1938.

[48] H. Eyring, "The Activated Complex in Chemical Reactions," *The Journal of Chemical Physics*, vol. 3, no. 2, pp. 107–115, 1935.

[49] R. L. Jaffe, J. M. Henry, and J. B. Anderson, "Variational theory of reaction rates: Application to F+H2↔HF+H," *The Journal of Chemical Physics*, vol. 59, no. 3, pp. 1128–1141, 1973.

[50] H. B. Schlegel, "Estimating the hessian for gradient-type geometry optimizations," *Theoretica Chimica Acta*, vol. 66, no. 5, pp. 333–340, 1984.

[51] C. G. Broyden, "The convergence of a class of double-rank minimization algorithms: 1. general considerations," *Journal of the Institute of Mathematics and its Applications*, vol. 6, no. 1, pp. 76–90, 1970.

[52] R. Fletcher, "A new approach to variable metric algorithms," *The Computer Journal*, vol. 13, no. 3, pp. 317–322, 1970.

[53] D. Goldfarb, "A family of variable-metric methods derived by variational means," *Mathematics of Computation*, vol. 24, no. 109, pp. 23–26, 1970.

[54] D. F. Shanno, "Conditioning of quasi-newton methods for function minimization," *Mathematics of Computation*, vol. 24, no. 111, pp. 647–656, 1970.

[55] E. B. Wilson, J. C. Decius, and P. C. Cross, *Molecular Vibrations: The Theory of Infrared and Raman Vibrational Spectra*. Dover, 1980.

[56] H. P. Hratchian and H. B. Schlegel, "Finding minima, transition states, and following reaction pathways on ab initio potential energy surfaces," in *Theory and Applications of Computational Chemistry: The First Forty Years*, C. E. Dykstra, G. Frenking, K. S. Kim, and G. E. Scuseria, Eds. Amsterdam: Elsevier, 2005, ch. 10, pp. 195–249.

[57] H. B. Schlegel, "Optimization of equilibrium geometries and transition structures," *Journal of Computational Chemistry*, vol. 3, no. 2, pp. 214–218, 1982.

[58] P. Pulay, G. Fogarasi, F. Pang, and J. E. Boggs, "Systematic ab initio gradient calculation of molecular geometries, force constants, and dipole moment derivatives," *Journal of the American Chemical Society*, vol. 101, no. 10, pp. 2550–2560, May 1979. [Online]. Available: https://doi.org/10.1021/ja00504a009

[59] G. Fogarasi, X. Zhou, P. W. Taylor, and P. Pulay, "The calculation of ab initio molecular geometries: efficient optimization by natural internal coordinates and empirical correction by offset forces," *Journal of the American Chemical Society*, vol. 114, no. 21, pp. 8191–8201, 1992.

[60] P. Pulay and G. Fogarasi, "Geometry optimization in redundant internal coordinates," *The Journal of Chemical Physics*, vol. 96, no. 4, pp. 2856–2860, 1992.

[61] C. Peng, P. Y. Ayala, H. B. Schlegel, and M. J. Frisch, "Using redundant internal coordinates to optimize equilibrium geometries and transition states," *Journal of Computational Chemistry*, vol. 17, no. 1, pp. 49–56, 1996.

[62] J. Baker, A. Kessi, and B. Delley, "The generation and use of delocalized internal coordinates in geometry optimization," *The Journal of Chemical Physics*, vol. 105, no. 1, pp. 192–212, 1996.

[63] J. Baker, "Constrained optimization in delocalized internal coordinates," *Journal of Computational Chemistry*, vol. 18, no. 8, pp. 1079–1095, 1997.

[64] J. Baker, D. Kinghorn, and P. Pulay, "Geometry optimization in delocalized internal coordinates: An efficient quadratically scaling algorithm for large molecules," *The Journal of Chemical Physics*, vol. 110, no. 11, pp. 4986–4991, 1999. [Online]. Available: https://doi.org/10.1063/1.478397

[65] L.-P. Wang and C. Song, "Geometry optimization made simple with translation and rotation coordinates," *The Journal of Chemical Physics*, vol. 144, no. 21, p. 214108, 2016.

[66] K. Fukui, "Formulation of the reaction coordinate," *The Journal of Physical Chemistry*, vol. 74, no. 23, pp. 4161–4163, 1970.

[67] ——, "The path of chemical reactions - the IRC approach," *Accounts of Chemical Research*, vol. 14, no. 12, pp. 363–368, 1981.

[68] C. Gonzalez and H. B. Schlegel, "Reaction path following in mass-weighted internal coordinates," *The Journal of Physical Chemistry*, vol. 94, no. 14, pp. 5523–5527, 1990.

[69] S. L. Johansen, H. Park, L.-P. Wang, and K. N. Crabtree, "Reactant discovery with an ab initio nanoreactor: Exploration of astrophysical n-heterocycle precursors and formation pathways," *ACS Earth and Space Chemistry*, vol. 8, no. 9, pp. 1771–1783, 2024, published online August 9, 2024.

[70] R. M. Parrish, L. A. Burns, D. G. A. Smith, A. C. Simmonett, A. E. I. DePrince, E. G. Hohenstein, U. Bozkaya, A. Y. Sokolov, R. Di Remigio, R. M. Richard, J. F. Gonthier, A. M. James, H. R. McAlexander, A. Kumar, M. Saitow, X. Wang, B. P. Pritchard, P. Verma, H. F. I. Schaefer, K. Patkowski, R. A. King, E. F. Valeev, F. A. Evangelista, J. M. Turney, T. D. Crawford, and C. D. Sherrill, "Psi4 1.1: An Open-Source Electronic Structure Program Emphasizing Automation, Advanced Libraries, and Interoperability," *Journal of Chemical Theory and Computation*, vol. 13, no. 7, pp. 3185–3197, 2017.

[71] D. G. A. Smith, D. Altarawy, L. A. Burns, M. Welborn, L. N. Naden, L. Ward, S. Ellis, B. P. Pritchard, and T. D. Crawford, "The molssi qcarchive project: An open-source platform to compute, organize, and share quantum chemistry data," *WIREs Computational Molecular Science*, vol. 11, no. 2, p. e1491, 2021.

[72] H. Park, B. P. Pritchard, and L. Wang, "High-throughput approach for minimum energy pathway search using the nudged elastic band method with efficient data handling and parallel computing," *Journal of Chemical Theory and Computation*, vol. 21, no. 23, pp. 12 048–12 063, 2025.

[73] F. Wöhler, "Ueber künstliche Bildung des Harnstoffs," *Ann. Phys.*, vol. 88, no. 2, pp. 253–256, 1828.

[74] A. Eschenmoser and M. Volkan Kisakürek, "Chemistry and the Origin of Life," *Helv. Chim. Acta*, vol. 79, no. 5, pp. 1249–1259, 1996.

[75] I. Hirao, M. Kimoto, and R. Yamashige, "Natural versus Artificial Creation of Base Pairs in DNA: Origin of Nucleobases from the Perspectives of Unnatural Base Pair Studies," *Acc. Chem. Res.*, vol. 45, no. 12, pp. 2055–2065, 2012.

[76] A. C. Rios and Y. Tor, "On the Origin of the Canonical Nucleobases: An Assessment of Selection Pressures across Chemical and Early Biological Evolution," *Isr. J. Chem.*, vol. 53, no. 6-7, pp. 469–483, 2013.

[77] Z. Peeters, O. Botta, S. B. Charnley, R. Ruiterkamp, and P. Ehrenfreund, "The Astrobiology of Nucleobases," *Astrophys. J.*, vol. 593, no. 2, p. L129, 2003.

[78] Z. Martins, "The nitrogen heterocycle content of meteorites and their significance for the origin of life," *Life*, vol. 8, no. 3, p. 28, July 2018.

[79] D. M. Hudgins, C. W. Bauschlicher, Jr., and L. J. Allamandola, "Variations in the peak position of the 6.2 $\mu$m interstellar emission feature: A tracer of N in the interstellar polycyclic aromatic hydrocarbon population," *Astrophys. J.*, vol. 632, no. 1, pp. 316–332, October 2005.

[80] C. P. Endres, S. Schlemmer, P. Schilke, J. Stutzki, and H. S. Müller, "The cologne database for molecular spectroscopy, cdms, in the virtual atomic and molecular data centre, vamdc," *Journal of Molecular Spectroscopy*, vol. 327, pp. 95–104, 2016. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0022285216300340

[81] S. Kwok, "Complex organics in space from Solar System to distant galaxies," *Astron. Astrophys. Rev.*, vol. 24, no. 1, p. 8, December 2016.

[82] B. A. McGuire, A. M. Burkhardt, S. Kalenskii, C. N. Shingledecker, A. J. Remijan, E. Herbst, and M. C. McCarthy, "Detection of the aromatic molecule benzonitrile (c-C6H5CN) in the interstellar medium," *Science*, vol. 359, no. 6372, pp. 202–205, 2018.

[83] T. J. Barnum, M. A. Siebert, K. L. K. Lee, R. A. Loomis, P. B. Changala, S. B. Charnley, M. L. Sita, C. Xue, A. J. Remijan, A. M. Burkhardt, B. A. McGuire, and I. R. Cooke, "A search for heterocycles in gotham observations of tmc-1," *Journal of Physical Chemistry A*, vol. 126, pp. 2716–2728, 5 2022.

[84] M. C. Mccarthy and B. A. Mcguire, "Aromatics and cyclic molecules in molecular clouds: A new dimension of interstellar organic chemistry," *J. Phys. Chem. A*, vol. 125, p. 3243, 2021.

[85] I. R. Cooke, D. Gupta, J. P. Messinger, and I. R. Sims, "Benzonitrile as a proxy for benzene in the cold ISM: Low-temperature rate coefficients for CN + C 6 H 6," *Astrophys. J.*, vol. 891, no. 2, p. L41, March 2020.

[86] J. Cernicharo, M. Agündez, C. Cabezas, B. Tercero, N. Marcelino, J. R. Pardo, and P. de Vicente, "Pure hydrocarbon cycles in tmc-1: Discovery of ethynyl cyclopropenylidene, cyclopentadiene, and indene," *Astron. Astrophys.*, vol. 649, pp. –, 2021. [Online]. Available: https://doi.org/10.1051/0004-6361/202141156

[87] B. A. McGuire, R. A. Loomis, A. M. Burkhardt, K. L. K. Lee, C. N. Shingledecker, S. B. Charnley, I. R. Cooke, M. A. Cordiner, E. Herbst, S. Kalenskii, M. A. Siebert, E. R. Willis, C. Xue, A. J. Remijan, and M. C. McCarthy, "Detection of two interstellar polycyclic aromatic hydrocarbons via spectral matched filtering," *Science*, vol. 371, no. 6535, pp. 1265–1269, 2021. [Online]. Available: https://www.science.org/doi/abs/10.1126/science.abb7535

[88] V. Vuitton, R. Yelle, and M. McEwan, "Ion chemistry and N-containing molecules in Titan's upper atmosphere," *Icarus*, vol. 191, no. 2, pp. 722–742, November 2007.

[89] C. A. Nixon, A. E. Thelen, M. A. Cordiner, Z. Kisiel, S. B. Charnley, E. M. Molter, J. Serigano, P. G. J. Irwin, N. A. Teanby, and Y.-J. Kuan, "Detection of cyclopropenylidene on titan with ALMA," *Astron. J.*, vol. 160, no. 5, p. 205, October 2020.

[90] M. Y. Palmer, M. A. Cordiner, C. A. Nixon, S. B. Charnley, N. A. Teanby, Z. Kisiel, P. G. J. Irwin, and M. J. Mumma, "ALMA detection and astrobiological potential of vinyl cyanide on Titan," *Sci. Adv.*, vol. 3, no. 7, p. e1700022, July 2017.

[91] S. Hörst, R. Yelle, A. Buch, N. Carrasco, G. Cernogora, O. Dutuit, E. Quirico, E. Sciamma-O'Brien, M. Smith, Á. Somogyi, C. Szopa, R. Thissen, and V. Vuitton, "Formation of amino acids and nucleotide bases in a titan atmosphere simulation experiment," *Astrobiology*, vol. 12, no. 9, pp. 809–817, September 2012.

[92] M. G. Trainer, A. A. Pavlov, H. L. DeWitt, J. L. Jimenez, C. P. McKay, O. B. Toon, and M. A. Tolbert, "Organic haze on Titan and the early Earth," *Proc. Natl. Acad. Sci.*, vol. 103, no. 48, pp. 18 035–18 042, November 2006.

[93] A. Coustenis, "Titans atmosphere and surface - parallels and differences with the primitive earth," *Earth Moon Planets*, vol. 67, no. 1-3, pp. 95–100, 1995.

[94] P. M. Zimmerman, "Navigating molecular space for reaction mechanisms: An efficient, automated procedure," *Mol. Simul.*, vol. 41, no. 1-3, pp. 43–54, February 2015.

[95] M. Bergeler, G. N. Simm, J. Proppe, and M. Reiher, "Heuristics-Guided Exploration of Reaction Mechanisms," *J. Chem. Theory Comput.*, vol. 11, no. 12, pp. 5712–5722, December 2015.

[96] Y. V. Suleimanov and W. H. Green, "Automated Discovery of Elementary Chemical Reaction Steps Using Freezing String and Berny Optimization Methods," *J. Chem. Theory Comput.*, vol. 11, no. 9, pp. 4248–4259, September 2015.

[97] E. Martínez-Núñez, "An automated method to find transition states using chemical dynamics simulations," *J. Comput. Chem.*, vol. 36, no. 4, pp. 222–234, 2015.

[98] ——, "An automated transition state search using classical trajectories initialized at multiple minima," *Phys. Chem. Chem. Phys.*, vol. 17, no. 22, pp. 14 912–14 921, 2015.

[99] C. W. Gao, J. W. Allen, W. H. Green, and R. H. West, "Reaction Mechanism Generator: Automatic construction of chemical kinetic mechanisms," *Comput. Phys. Commun.*, vol. 203, pp. 212–225, 2016.

[100] I. M. Pendleton, M. H. Pérez-Temprano, M. S. Sanford, and P. M. Zimmerman, "Experimental and Computational Assessment of Reactivity and Mechanism in C(sp3)–N Bond-Forming Reductive Elimination from Palladium(IV)," *J. Am. Chem. Soc.*, vol. 138, no. 18, pp. 6049–6060, May 2016.

[101] S. Maeda, Y. Harabuchi, M. Takagi, T. Taketsugu, and K. Morokuma, "Artificial Force Induced Reaction (AFIR) Method for Exploring Quantum Chemical Potential Energy Surfaces," *Chem. Rec.*, vol. 16, no. 5, pp. 2232–2248, 2016.

[102] A. L. Dewyer and P. M. Zimmerman, "Finding reaction mechanisms, intuitive or otherwise," *Org. Biomol. Chem.*, vol. 15, no. 3, pp. 501–504, January 2017.

[103] S. Kopec, E. Martínez-Núñez, J. Soto, and D. Peláez, "vdW-TSSCDS—An automated and global procedure for the computation of stationary points on intermolecular potential energy surfaces," *Int. J. Quantum Chem.*, vol. 119, no. 21, p. e26008, 2019.

[104] J. P. Unsleber and M. Reiher, "The Exploration of Chemical Reaction Networks," *Annu. Rev. Phys. Chem.*, vol. 71, no. 1, pp. 121–142, 2020.

[105] E. Martínez-Núñez, G. L. Barnes, D. R. Glowacki, S. Kopec, D. Peláez, A. Rodríguez, R. Rodríguez-Fernández, R. J. Shannon, J. J. P. Stewart, P. G. Tahoces, and S. A. Vazquez, "AutoMeKin2021: An open-source program for automated reaction discovery," *J. Comput. Chem.*, vol. 42, no. 28, pp. 2036–2048, 2021.

[106] Q. Zhao and B. M. Savoie, "Algorithmic Explorations of Unimolecular and Bimolecular Reaction Spaces**," *Angew. Chem. Int. Ed.*, vol. 61, no. 46, November 2022.

[107] I. Ismail, R. Chantreau Majerus, and S. Habershon, "Graph-Driven Reaction Discovery: Progress, Challenges, and Future Opportunities," *J. Phys. Chem. A*, vol. 126, no. 40, pp. 7051–7069, October 2022.

[108] G. Zhang, J. Li, X. Liang, and Z. Liu, "Automated reaction mechanisms and kinetics with the nudged elastic band method-based AMK_Mountain and its description of the preliminary alkaline hydrolysis of nitrocellulose monomer," *J. Comput. Chem.*, vol. 43, no. 22, pp. 1513–1523, 2022.

[109] Q. Zhao, S. M. Vaddadi, M. Woulfe, L. A. Ogunfowora, S. S. Garimella, O. Isayev, and B. M. Savoie, "Comprehensive exploration of graphically defined reaction spaces," *Scientific Data*, vol. 10, no. 1, p. 145, 2023.

[110] F. Pietrucci and W. Andreoni, "Graph Theory Meets Ab Initio Molecular Dynamics: Atomic Structures and Transformations at the Nanoscale," *Phys. Rev. Lett.*, vol. 107, no. 8, p. 085504, August 2011.

[111] L.-P. Wang, R. T. McGibbon, V. S. Pande, and T. J. Martínez, "Automated discovery and refinement of reactive molecular dynamics pathways," *Journal of Chemical Theory and Computation*, vol. 12, no. 2, pp. 638–649, 2016.

[112] L. Xie, Q. Zhao, K. F. Jensen, and H. J. Kulik, "Direct Observation of Early-Stage Quantum Dot Growth Mechanisms with High-Temperature Ab Initio Molecular Dynamics," *J. Phys. Chem. C*, vol. 120, no. 4, pp. 2472–2483, February 2016.

[113] R. A. Jara-Toro, G. A. Pino, D. R. Glowacki, R. J. Shannon, and E. Martínez-Núñez, "Enhancing Automated Reaction Discovery with Boxed Molecular Dynamics in Energy Space," *ChemSystemsChem*, vol. 2, no. 1, p. e1900024, 2020.

[114] R. J. Shannon, E. Martínez-Núñez, D. V. Shalashilin, and D. R. Glowacki, "ChemDyME: Kinetically Steered, Automated Mechanism Generation through Combined Molecular Dynamics and Master Equation Calculations," *J. Chem. Theory Comput.*, vol. 17, no. 8, pp. 4901–4912, August 2021.

[115] Q. Cui, J. Peng, C. Xu, and Z. Lan, "Automatic Approach to Explore the Multireaction Mechanism for Medium-Sized Bimolecular Reactions via Collision Dynamics Simulations and Transition State Searches," *J. Chem. Theory Comput.*, vol. 18, no. 2, pp. 910–924, February 2022.

[116] H. Hirai and R. Jinnouchi, "Discovering surface reaction pathways using accelerated molecular dynamics and network analysis tools," *RSC Adv.*, vol. 12, no. 36, pp. 23 274–23 283, 2022.

[117] M. H. Rasmussen and J. H. Jensen, "Fast and automated identification of reactions with low barriers using meta-MD simulations," *PeerJ Phys. Chem.*, vol. 4, p. e22, March 2022.

[118] A. Stan, B. V. D. Esch, and C. Ochsenfeld, "Fully Automated Generation of Prebiotically Relevant Reaction Networks from Optimized Nanoreactor Simulations," *J. Chem. Theory Comput.*, vol. 18, no. 11, pp. 6700–6712, November 2022.

[119] Y. Sumiya, Y. Harabuchi, Y. Nagata, and S. Maeda, "Quantum Chemical Calculations to Trace Back Reaction Paths for the Prediction of Reactants," *JACS Au*, vol. 2, no. 5, pp. 1181–1188, May 2022.

[120] C. Robertson, R. Hyland, A. J. D. Lacey, S. Havens, and S. Habershon, "Identifying Barrierless Mechanisms for Benzene Formation in the Interstellar Medium Using Permutationally Invariant Reaction Discovery," *J. Chem. Theory Comput.*, vol. 17, no. 4, pp. 2307–2322, April 2021.

[121] J. Zádor, C. Martí, R. Van De Vijver, S. L. Johansen, Y. Yang, H. A. Michelsen, and H. N. Najm, "Automated Reaction Kinetics of Gas-Phase Organic Species over Multiwell Potential Energy Surfaces," *J. Phys. Chem. A*, vol. 127, no. 3, pp. 565–588, January 2023.

[122] S. Habershon, "Automated Prediction of Catalytic Mechanism and Rate Law Using Graph-Based Reaction Path Sampling," *J. Chem. Theory Comput.*, vol. 12, no. 4, pp. 1786–1798, April 2016.

[123] J. A. Varela, S. A. Vázquez, and E. Martínez-Núñez, "An automated method to find reaction mechanisms and solve the kinetics in organometallic catalysis," *Chem. Sci.*, vol. 8, no. 5, pp. 3843–3851, 2017.

[124] M. Gao, A. Lyalin, S. Maeda, and T. Taketsugu, "Application of Automated Reaction Path Search Methods to a Systematic Search of Single-Bond Activation Pathways Catalyzed by Small Metal Clusters: A Case Study on H–H Activation by Gold," *J. Chem. Theory Comput.*, vol. 10, no. 4, pp. 1623–1630, April 2014.

[125] M. Jafari and P. M. Zimmerman, "Uncovering reaction sequences on surfaces through graphical methods," *Phys. Chem. Chem. Phys.*, vol. 20, no. 11, pp. 7721–7729, 2018.

[126] E. Pieri, D. Lahana, A. M. Chang, C. R. Aldaz, K. C. Thompson, and T. J. Martínez, "The non-adiabatic nanoreactor: Towards the automated discovery of photochemistry," *Chem. Sci.*, vol. 12, no. 21, pp. 7294–7307, 2021.

[127] N. Goldman, E. J. Reed, L. E. Fried, I.-F. William Kuo, and A. Maiti, "Synthesis of glycine-containing complexes in impacts of comets on early Earth," *Nat. Chem.*, vol. 2, no. 11, pp. 949–954, November 2010.

[128] A. M. Saitta and F. Saija, "Miller experiments in atomistic computer simulations," *Proc. Natl. Acad. Sci.*, vol. 111, no. 38, pp. 13 768–13 773, September 2014.

[129] D. Y. Zubarev, D. Rappoport, and A. Aspuru-Guzik, "Uncertainty of Prebiotic Scenarios: The Case of the Non-Enzymatic Reverse Tricarboxylic Acid Cycle," *Sci. Rep.*, vol. 5, no. 1, p. 8009, January 2015.

[130] Q. Zhao, S. S. Garimella, and B. M. Savoie, "Thermally Accessible Prebiotic Pathways for Forming Ribonucleic Acid and Protein Precursors from Aqueous Hydrogen Cyanide," *J. Am. Chem. Soc.*, vol. 145, no. 11, pp. 6135–6143, March 2023.

[131] D. Garay-Ruiz, M. Álvarez-Moreno, C. Bo, and E. Martínez-Núñez, "New Tools for Taming Complex Reaction Networks: The Unimolecular Decomposition of Indole Revisited," *ACS Phys. Chem. Au*, vol. 2, no. 3, pp. 225–236, May 2022.

[132] L. Martínez, R. Andrade, E. G. Birgin, and J. M. Martínez, "PACKMOL: A package for building initial configurations for molecular dynamics simulations," *J. Comput. Chem.*, vol. 30, no. 13, pp. 2157–2164, October 2009.

[133] I. S. Ufimtsev and T. J. Martinez, "Quantum chemistry on graphical processing units. 3. Analytical energy gradients, geometry optimization, and first principles molecular dynamics," *J. Chem. Theory Comput.*, vol. 5, no. 10, pp. 2619–2628, October 2009.

[134] W. Humphrey, A. Dalke, and K. Schulten, "VMD – Visual Molecular Dynamics," *Journal of Molecular Graphics*, vol. 14, pp. 33–38, 1996.

[135] T. P. Snow and B. J. McCall, "Diffuse atomic and molecular clouds," *Annu. Rev. Astron. Astrophys.*, vol. 44, pp. 367–414, 2006.

[136] Y.-S. Lin, G.-D. Li, S.-P. Mao, and J.-D. Chai, "Long-range corrected hybrid density functionals with improved dispersion corrections," *J. Chem. Theory Comput.*, vol. 9, no. 1, pp. 263–272, January 2013.

[137] Y. Shao, Z. Gan, E. Epifanovsky, A. T. Gilbert, M. Wormit, J. Kussmann, A. W. Lange, A. Behn, J. Deng, X. Feng, D. Ghosh, M. Goldey, P. R. Horn, L. D. Jacobson, I. Kaliman, R. Z. Khaliullin, T. Kuš, A. Landau, J. Liu, E. I. Proynov, Y. M. Rhee, R. M. Richard, M. A. Rohrdanz, R. P. Steele, E. J. Sundstrom, H. L. Woodcock, P. M. Zimmerman, D. Zuev, B. Albrecht, E. Alguire, B. Austin, G. J. Beran, Y. A. Bernard, E. Berquist, K. Brandhorst, K. B. Bravaya, S. T. Brown, D. Casanova, C. M. Chang, Y. Chen, S. H. Chien, K. D. Closser, D. L. Crittenden, M. Diedenhofen, R. A. Distasio, H. Do, A. D. Dutoi, R. G. Edgar, S. Fatehi, L. Fusti-Molnar, A. Ghysels, A. Golubeva-Zadorozhnaya, J. Gomes, M. W. Hanson-Heine, P. H. Harbach, A. W. Hauser, E. G. Hohenstein, Z. C. Holden, T. C.

Jagau, H. Ji, B. Kaduk, K. Khistyaev, J. Kim, J. Kim, R. A. King, P. Klunzinger, D. Kosenkov, T. Kowalczyk, C. M. Krauter, K. U. Lao, A. D. Laurent, K. V. Lawler, S. V. Levchenko, C. Y. Lin, F. Liu, E. Livshits, R. C. Lochan, A. Luenser, P. Manohar, S. F. Manzer, S. P. Mao, N. Mardirossian, A. V. Marenich, S. A. Maurer, N. J. Mayhall, E. Neuscamman, C. M. Oana, R. Olivares-Amaya, D. P. Oneill, J. A. Parkhill, T. M. Perrine, R. Peverati, A. Prociuk, D. R. Rehn, E. Rosta, N. J. Russ, S. M. Sharada, S. Sharma, D. W. Small, A. Sodt, T. Stein, D. Stück, Y. C. Su, A. J. Thom, T. Tsuchimochi, V. Vanovschi, L. Vogt, O. Vydrov, T. Wang, M. A. Watson, J. Wenzel, A. White, C. F. Williams, J. Yang, S. Yeganeh, S. R. Yost, Z. Q. You, I. Y. Zhang, X. Zhang, Y. Zhao, B. R. Brooks, G. K. Chan, D. M. Chipman, C. J. Cramer, W. A. Goddard, M. S. Gordon, W. J. Hehre, A. Klamt, H. F. Schaefer, M. W. Schmidt, C. D. Sherrill, D. G. Truhlar, A. Warshel, X. Xu, A. Aspuru-Guzik, R. Baer, A. T. Bell, N. A. Besley, J. D. Chai, A. Dreuw, B. D. Dunietz, T. R. Furlani, S. R. Gwaltney, C. P. Hsu, Y. Jung, J. Kong, D. S. Lambrecht, W. Liang, C. Ochsenfeld, V. A. Rassolov, L. V. Slipchenko, J. E. Subotnik, T. Van Voorhis, J. M. Herbert, A. I. Krylov, P. M. Gill, and M. Head-Gordon, "Advances in molecular quantum chemistry contained in the Q-Chem 4 program package," *Mol. Phys.*, vol. 113, no. 2, pp. 184–215, January 2015.

[138] K. L. Kelvin Lee, R. A. Loomis, A. M. Burkhardt, I. R. Cooke, C. Xue, M. A. Siebert, C. N. Shingledecker, A. Remijan, S. B. Charnley, M. C. McCarthy, and B. A. McGuire, "Discovery of interstellar trans -cyanovinylacetylene (HCCCHCHCN) and vinylcyanoacetylene (H 2 CCHC 3 N) in GOTHAM observations of TMC-1," *Astrophys. J.*, vol. 908, no. 1, p. L11, February 2021.

[139] M. A. Zdanovskaia, P. M. Dorman, V. L. Orr, A. N. Owen, S. M. Kougias, B. J. Esselman, R. C. Woods, and R. J. McMahon, "Rotational spectra of three cyanobutadiene isomers (C5H5N) of relevance to astrochemistry and other harsh reaction environments," *J. Am. Chem. Soc.*, vol. 143, no. 25, pp. 9551–9564, 2021.

[140] B. A. McGuire, A. M. Burkhardt, R. A. Loomis, C. N. Shingledecker, K. L. K. Lee, S. B. Charnley, M. A. Cordiner, E. Herbst, S. Kalenskii, E. Momjian, E. R. Willis, C. Xue, A. J. Remijan, and M. C. McCarthy, "Early science from GOTHAM: Project overview, methods, and the detection of interstellar propargyl cyanide (HCCCH2CN) in TMC-1," *Astrophys. J. Lett.*, vol. 900, no. 1, p. L10, September 2020.

[141] J. M. Vrtilek, C. A. Gottlieb, and P. Thaddeus, "LABORATORY AND ASTRONOMICAL SPECTROSCOPY OF C 3 H 2 , THE FIRST INTERSTELLAR ORGANIC RING," Tech. Rep., 1987.

[142] J. M. Hollis, A. J. Remijan, P. R. Jewell, and F. J. Lovas, "Cyclopropenone (c-H2C3O): A New Interstellar Ring Molecule," *Astrophys. J.*, vol. 642, no. 2, p. 933, 2006.

[143] I. R. Cooke, C. Xue, P. B. Changala, H. T. Shay, A. N. Byrne, Q. Y. Tang, Z. T. P. Fried, K. L. Kelvin Lee, R. A. Loomis, T. Lamberts, A. Remijan, A. M. Burkhardt, E. Herbst, M. C. McCarthy, and B. A. McGuire, "Detection of Interstellar E-1-cyano-1,3-butadiene in GOTHAM Observations of TMC-1," *Astrophys. J.*, vol. 948, p. 133, May 2023.

[144] H. I. Ewen and E. M. Purcell, "Observation of a line in the galactic radio spectrum: Radiation from galactic hydrogen at 1,420 Mc./sec." *Nature*, vol. 168, no. 4270, pp. 356–356, September 1951.

[145] S. T. RIDGWAY, D. N. B. HALL, S. G. KLEINMANN, D. A. WEINBERGER, and R. S. WOJSLAW, "Circumstellar acetylene in the infrared spectrum of IRC +10° 216," *Nature*, vol. 264, no. 5584, pp. 345–346, November 1976.

[146] A. L. Betz, "Ethylene in IRC +10216," *Astrophys. J.*, vol. 244, no. 2, p. L103, March 1981.

[147] J. H. Lacy, II. Evans, Neal J., J. M. Achtermann, D. E. Bruce, J. F. Arens, and J. S. Carr, "Discovery of interstellar acetylene," *Astrophys. J.*, vol. 342, p. L43, July 1989.

[148] F. F. Gardner and G. Winnewisser, "The detection of interstellar vinyl cyanide (acrylonitrile)," *Astrophys. J.*, vol. 195, p. L127, February 1975.

[149] M. Agúndez, N. Marcelino, and J. Cernicharo, "Discovery of interstellar isocyanogen (CNCN): Further evidence that dicyanopolyynes are abundant in space," *Astrophys. J.*, vol. 861, no. 2, p. L22, 2018.

[150] D. P. Zaleski, N. A. Seifert, A. L. Steber, M. T. Muckle, R. A. Loomis, J. F. Corby, O. Martinez, K. N. Crabtree, P. R. Jewell, J. M. Hollis, F. J. Lovas, D. Vasquez, J. Nyiramahirwe, N. Sciortino, K. Johnson, M. C. McCarthy, A. J. Remijan, and B. H. Pate, "DETECTION OF E-CYANOMETHANIMINE TOWARD SAGITTARIUS B2(N) IN THE GREEN BANK TELESCOPE PRIMOS SURVEY," *Astrophys. J. Lett.*, vol. 765, no. 1, p. L10, February 2013.

[151] V. M. Rivilla, J. Martín-Pintado, I. Jiménez-Serra, S. Zeng, S. Martín, J. Armijos-Abendaño, M. A. Requena-Torres, R. Aladro, and D. Riquelme, "Abundant Z-cyanomethanimine in the interstellar medium: Paving the way to the synthesis of adenine," *Mon. Not. R. Astron. Soc. Lett.*, vol. 483, no. 1, pp. L114–L119, 2019.

[152] A. Belloche, K. M. Menten, C. Comito, H. S. P. Müller, P. Schilke, J. Ott, S. Thorwirth, and C. Hieret, "Detection of amino acetonitrile in sgr B2(N)," *Astron. Astrophys.*, vol. 492, no. 3, pp. 769–773, 2008.

[153] M. Agúndez, J. Cernicharo, P. De Vicente, N. Marcelino, E. Roueff, A. Fuente, M. Gerin, M. Guélin, C. Albo, A. Barcia, L. Barbas, R. Bolaño, F. Colomer, M. C. Diez, J. D. Gallego, J. Gómez-González,

I. López-Fernández, J. A. López-Fernández, J. A. López-Pérez, I. Malo, J. M. Serna, and F. Tercero, "Probing non-polar interstellar molecules through their protonated form: Detection of protonated cyanogen (NCCNH +)," *Astron. Astrophys.*, vol. 579, pp. 1–4, 2015.

[154] B. E. Turner, H. S. Liszt, N. Kaifu, and A. G. Kisliakov, "Microwave detection of interstellar cyanamide," *Astrophys. J.*, vol. 201, p. L149, November 1975.

[155] B. A. McGuire, R. A. Loomis, C. M. Charness, J. F. Corby, G. A. Blake, J. M. Hollis, F. J. Lovas, P. R. Jewell, and A. J. Remijan, "Interstellar carbodiimide (HNCNH): A new astronomical detection from the gbt PRIMOS survey via maser emission features," *Astrophys. J. Lett.*, vol. 758, no. 2, 2012.

[156] N. Jiang, M. Melosso, L. Bizzocchi, S. Alessandrini, J.-C. Guillemin, L. Dore, and C. Puzzarini, "Spectroscopic and computational characterization of 2-Aza-1,3-butadiene, a molecule of astrochemical significance," *J. Phys. Chem. A*, vol. 126, no. 11, pp. 1881–1888, 2022.

[157] M. Guelin and P. Thaddeus, "Tentative detection of the C3N radical," *Astrophys. J. Lett.*, vol. 212, p. L81, March 1977.

[158] P. Friberg, A. Hjalmarson, M. Guelin, and W. M. Irvine, "Interstellar C3N - detection in taurus dark clouds," *Astrophys. J. Lett.*, vol. 241, pp. L99–L103, October 1980.

[159] J. Cernicharo, M. Agúndez, C. Cabezas, N. Marcelino, B. Tercero, J. R. Pardo, J. D. Gallego, F. Tercero, J. A. López-Pérez, and P. de Vicente, "Discovery of $CH_2CHCCH$ and detection of HCCN, $HC_4N$, $CH_3CH_2CN$, and, tentatively, $CH_3CH_2CCH$ in TMC-1," *Astron. Astrophys.*, vol. 647, no. L2, p. L2, March 2021.

[160] J. Cernicharo, C. Cabezas, M. Agúndez, B. Tercero, N. Marcelino, J. R. Pardo, F. Tercero, J. D. Gallego, J. A. López-Pérez, and P. DeVicente, "Discovery of allenyl acetylene, H 2 CCCHCCH, in TMC-1," *Astron. Astrophys.*, vol. 647, p. L3, March 2021.

[161] R. Fuentetaja, C. Cabezas, M. Agúndez, B. Tercero, N. Marcelino, J. R. Pardo, P. de Vicente, and J. Cernicharo, "Discovery of $CH_2CCHC_4H$ and a rigorous detection of $CH_2CCHC_3N$ in TMC-1 with the QUIJOTE line survey," *Astron. Astrophys.*, vol. 663, no. L3, p. L3, July 2022.

[162] J. Heitkämper, S. Suchaneck, J. García de la Concepción, J. Kästner, and G. Molpeceres, "The reactivity of pyridine in cold interstellar environments: The reaction of pyridine with the CN radical," *Front. Astron. Space Sci.*, vol. 9, 2022.

[163] D. S. Parker, R. I. Kaiser, O. Kostko, T. P. Troy, M. Ahmed, A. M. Mebel, and A. G. Tielens, "Gas phase synthesis of (ISO)quinoline and its role in the formation of nucleobases in the interstellar medium," *Astrophys. J.*, vol. 803, no. 2, p. 53, April 2015.

[164] B. J. Esselman, S. M. Kougias, M. A. Zdanovskaia, R. C. Woods, and R. J. McMahon, "Synthesis, purification, and rotational spectroscopy of (Cyanomethylene)Cyclopropane—An isomer of pyridine," *J. Phys. Chem. A*, vol. 125, no. 25, pp. 5601–5614, July 2021.

[165] R. A. Loomis, D. P. Zaleski, A. L. Steber, J. L. Neill, M. T. Muckle, B. J. Harris, J. M. Hollis, P. R. Jewell, V. Lattanzi, F. J. Lovas, Jr. Martinez, Oscar, M. C. McCarthy, A. J. Remijan, B. H. Pate, and J. F. Corby, "The detection of interstellar ethanimine ($CH_3CHNH$) from observations taken during the GBT PRIMOS survey," *Astrophys. J. Lett.*, vol. 765, no. L9, p. L9, March 2013.

[166] P. D. Godfrey, R. D. Brown, B. J. Robinson, and M. W. Sinclair, "Discovery of interstellar methanimine (formaldimine)," *Astrophys. Lett.*, vol. 13, p. 119, February 1973.

[167] S. Zeng, I. Jiménez-Serra, V. M. Rivilla, J. Martín-Pintado, L. F. Rodríguez-Almeida, B. Tercero, P. de Vicente, F. Rico-Villas, L. Colzi, S. Martín, and M. A. Requena-Torres, "Probing the chemical complexity of amines in the ISM: Detection of vinylamine ($C_2H_3NH_2$) and tentative detection of ethylamine ($C_2H_5NH_2$)," *Astrophys. J. Lett.*, vol. 920, no. L27, p. L27, October 2021.

[168] R. A. Motiyenko, L. Margulès, and J.-C. Guillemin, "Millimeter- and submillimeter-wave spectrum of methyleneaminoacetonitrile," *Astron. Astrophys.*, vol. 559, p. A44, November 2013.

[169] K. Luková, L. Kolesniková, J. Koucký, K. Vávra, P. Kania, J.-C. Guillemin, and Š. Urban, "Decoding millimetre-wave spectra of 2-iminopropanenitrile, a candidate for astronomical observations," *Astron. Astrophys.*, vol. 665, p. A9, September 2022.

[170] V. Bonačić-Koutecký, K. Schöffel, and J. Michl, "Electronic states of cyclobutadiene heteroanalogues. Critical biradicaloids," *J. Am. Chem. Soc.*, vol. 111, no. 16, pp. 6140–6146, 1989.

[171] A. G. Császár, J. Demaison, and H. D. Rudolph, "Equilibrium structures of three-, four-, five-, six-, and seven-membered unsaturated n-containing heterocycles," *The Journal of Physical Chemistry A*, vol. 119, no. 9, pp. 1731–1746, 2015.

[172] W. M. Irvine, P. Friberg, A. Hjalmarson, S. Ishikawa, N. Kaifu, K. Kawaguchi, S. C. Madden, H. E. Matthews, M. Ohishi, S. Saito, H. Suzuki, P. Thaddeus, B. E. Turner, S. Yamamoto, and L. M. Ziurys, "Identification of the interstellar cyanomethyl radical (CH2CN) in the molecular clouds TMC-1 and Sagittarius B2," *Astrophys. J.*, vol. 334, p. L107, November 1988.

[173] J. H. Waite, H. Niemann, R. V. Yelle, W. T. Kasprzak, T. E. Cravens, J. G. Luhmann, R. L. McNutt, W. H. Ip, D. Gell, V. De La Haye, I. Müller-Wordag, B. Magee, N. Borggren, S. Ledvina, G. Fletcher, E. Walter, R. Miller, S. Scherer, R. Thorpe, J. Xu, B. Block, and K. Arnett, "Ion Neutral Mass Spectrometer results from the first flyby of Titan," *Science*, vol. 308, no. 5724, pp. 982–986, May 2005.

[174] J. H. Waite, D. T. Young, T. E. Cravens, A. J. Coates, F. J. Crary, B. Magee, and J. Westlake, "The process of tholin formation in titan's upper atmosphere," *Science*, vol. 316, no. 5826, pp. 870–875, May 2007.

[175] V. Vuitton, R. Yelle, S. Klippenstein, S. Hörst, and P. Lavvas, "Simulating the density of organic species in the atmosphere of Titan with a coupled ion-neutral photochemical model," *Icarus*, vol. 324, pp. 120–197, May 2019.

[176] P. P. Lavvas, A. Coustenis, and I. M. Vardavas, "Coupling photochemistry with haze formation in Titan's atmosphere, Part II: Results and validation with Cassini/Huygens data," *Planet. Space Sci.*, vol. 56, no. 1, pp. 67–99, 2008.

[177] G. Maier and U. Schäfer, "Kleine ringe, 27. Versuche zur darstellung von azacyclobutadienen," *Lieb. Ann. Chem.*, vol. 1980, no. 5, pp. 798–813, 1980.

[178] M. D. Hanwell, D. E. Curtis, D. C. Lonie, T. Vandermeersch, E. Zurek, and G. R. Hutchison, "Avogadro: An advanced semantic chemical editor, visualization, and analysis platform," *J. Cheminformatics*, vol. 4, no. 1, p. 17, December 2012.

[179] D. S. N. Parker, R. I. Kaiser, O. Kostko, T. P. Troy, M. Ahmed, B.-J. Sun, S.-H. Chen, and A. H. H. Chang, "On the formation of pyridine in the interstellar medium," *Phys. Chem. Chem. Phys.*, vol. 17, no. 47, pp. 32 000–32 008, 2015.

[180] F. J. Lovas, J. M. Hollis, A. J. Remijan, and P. R. Jewell, "Detection of ketenimine ($CH_2CNH$) in sagittarius B2(N) hot cores," *Astrophys. J.*, vol. 645, no. 2, pp. L137–L140, July 2006.

[181] N. Balucani, D. Skouteris, F. Leonori, R. Petrucci, M. Hamberg, W. D. Geppert, P. Casavecchia, and M. Rosi, "Combined crossed beam and theoretical studies of the N( 2 D) + C 2 H 4 reaction and implications for atmospheric models of titan," *J. Phys. Chem. A*, vol. 116, no. 43, pp. 10 467–10 479, November 2012.

[182] H. E. Matthews and T. J. Sears, "The detection of vinyl cyanide in TMC-1," *Astrophys. J.*, vol. 272, p. 149, September 1983.

[183] S. K. Mondal, W. Iqbal, P. Gorai, B. Bhat, V. Wakelam, and A. Das, "Investigating the hot molecular core, G10.47+0.03: A pit of nitrogen-bearing complex organic molecules," *Astron. Astrophys.*, vol. 669, p. A71, January 2023.

[184] E. A. Bergin and M. Tafalla, "Cold Dark Clouds: The Initial Conditions for Star Formation," *Annu. Rev. Astron. Astrophys.*, vol. 45, no. 1, pp. 339–396, September 2007.

[185] R. T. Garrod and S. L. W. Weaver, "Simulations of Hot-Core Chemistry," *Chem. Rev.*, vol. 113, no. 12, pp. 8939–8960, December 2013.

[186] S. L. Johansen, Z. Xu, J. H. Westerfield, A. C. Wannenmacher, and K. N. Crabtree, "Coupled cluster characterization of 1-, 2-, and 3-Pyrrolyl: Parameters for vibrational and rotational spectroscopy," *J. Phys. Chem. A*, vol. 125, no. 5, pp. 1257–1268, February 2021.

[187] F. X. Brown, S. Saito, and S. Yamamoto, "Microwave spectroscopy of isotopically substituted HCCN and its molecular structure," *J. Mol. Spectrosc.*, vol. 143, no. 2, pp. 203–208, 1990.

[188] R. A. Bernheim, R. J. Kempf, J. V. Gramas, and P. S. Skell, "Electron paramagnetic resonance of triplet alternant methylenes. Propargylene and homologs," *J. Chem. Phys.*, vol. 43, no. 1, pp. 196–200, 1965.

[189] B. Z. Abbott, P. R. Hoobler, and H. F. Schaefer, "Relatives of cyanomethylene: Replacement of the divalent carbon by B-, N+, al-, si, P+, ga-, ge, and As+," *Phys. Chem. Chem. Phys.*, vol. 21, no. 48, pp. 26 438–26 452, 2019.

[190] S. Soorkia, C. a Taatjes, D. L. Osborn, T. M. Selby, A. J. Trevitt, K. R. Wilson, and S. R. Leone, "Direct detection of pyridine formation by the reaction of CH (CD) with pyrrole: A ring expansion reaction." *Phys. Chem. Chem. Phys. PCCP*, vol. 12, no. 31, pp. 8750–8, 2010.

[191] F. Holzmeier, I. Wagner, I. Fischer, A. Bodi, and P. Hemberger, "Pyrolysis of 3-Methoxypyridine. Detection and characterization of the pyrrolyl radical by threshold photoelectron spectroscopy," *J. Phys. Chem. A*, vol. 120, no. 27, pp. 4702–4710, July 2016.

[192] M. Guelin and J. Cernicharo, "Astronomical detection of the HCCN radical. Toward a new family of carbon-chain molecules?" *Astron. Astrophys.*, pp. L21–L24, January 1991.

[193] P. Lavvas, M. Sander, M. Kraft, and H. Imanaka, "Surface Chemistry and Particle Shape: Process for the Evolution Aerosols in Titan's Atmosphere," *Astrophys. J.*, vol. 728, no. 2, p. 80, February 2011.

[194] V. A. Krasnopolsky, "A photochemical model of Titan's atmosphere and ionosphere," *Icarus*, vol. 201, no. 1, pp. 226–256, May 2009.

[195] A. Tachibana and K. Fukui, "Differential geometry of chemically reacting systems," *Theoretica chimica acta*, vol. 49, no. 4, pp. 321–347, 1978.

[196] ——, "Intrinsic field theory of chemical reactions," *Theoretica chimica acta*, vol. 51, no. 4, pp. 275–296, 1979.

[197] J. L. Bao and D. G. Truhlar, "Variational transition state theory: theoretical framework and recent developments," *Chem. Soc. Rev.*, vol. 46, pp. 7548–7596, 2017.

[198] A. B. Birkholz and H. B. Schlegel, "Using bonding to guide transition state optimization," *Journal of Computational Chemistry*, vol. 36, no. 15, pp. 1157–1166, 2015. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/jcc.23910

[199] G. Mills, H. Jónsson, and G. K. Schenter, "Reversible work transition state theory: application to dissociative adsorption of hydrogen," *Surface Science*, vol. 324, no. 2, pp. 305–337, 1995.

[200] H. Jónsson, G. Mills, and K. W. Jacobsen, "Nudged elastic band method for finding minimum energy paths of transitions," in *Classical and Quantum Dynamics in Condensed Phase Simulations*. WORLD SCIENTIFIC, 1998, pp. 385–404. [Online]. Available: https://www.worldscientific.com/doi/abs/10.1142/9789812839664_0016

[201] G. Henkelman and H. Jónsson, "Improved tangent estimate in the nudged elastic band method for finding minimum energy paths and saddle points," *The Journal of Chemical Physics*, vol. 113, no. 22, pp. 9978–9985, 2000.

[202] G. Henkelman, B. P. Uberuaga, and H. Jónsson, "A climbing image nudged elastic band method for finding saddle points and minimum energy paths," *The Journal of Chemical Physics*, vol. 113, no. 22, pp. 9901–9904, 2000.

[203] J.-W. Chu, B. L. Trout, and B. R. Brooks, "A super-linear minimization scheme for the nudged elastic band method," *The Journal of Chemical Physics*, vol. 119, no. 24, pp. 12 708–12 717, 2003. [Online]. Available: https://doi.org/10.1063/1.1627754

[204] S. A. Trygubenko and D. J. Wales, "A doubly nudged elastic band method for finding transition states," *The Journal of Chemical Physics*, vol. 120, no. 5, pp. 2082–2094, 2004. [Online]. Available: https://doi.org/10.1063/1.1636455

[205] D. Sheppard, P. Xiao, W. Chemelewski, D. D. Johnson, and G. Henkelman, "A generalized solid-state nudged elastic band method," *The Journal of Chemical Physics*, vol. 136, no. 7, p. 074103, 2012. [Online]. Available: https://doi.org/10.1063/1.3684549

[206] W. E, W. Ren, and E. Vanden-Eijnden, "String method for the study of rare events," *Physical Review B*, vol. 66, no. 5, p. 052301, 2002. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevB.66.052301

[207] ——, "Finite Temperature String Method for the Study of Rare Events," *The Journal of Physical Chemistry B*, vol. 109, no. 14, pp. 6688–6693, 2005. [Online]. Available: https://doi.org/10.1021/jp0455430

[208] ——, "Simplified and improved string method for computing the minimum energy paths in barrier-crossing events," *The Journal of Chemical Physics*, vol. 126, no. 16, p. 164103, 2007. [Online]. Available: https://doi.org/10.1063/1.2720838

[209] D. Sheppard, R. Terrell, and G. Henkelman, "Optimization methods for finding minimum energy paths," *The Journal of Chemical Physics*, vol. 128, no. 13, p. 134106, 2008. [Online]. Available: https://doi.org/10.1063/1.2841941

[210] S. K. Burger and W. Yang, "Quadratic string method for determining the minimum-energy path based on multiobjective optimization," *The Journal of Chemical Physics*, vol. 124, no. 5, p. 054109, 2006. [Online]. Available: https://doi.org/10.1063/1.2163875

[211] ——, "Sequential quadratic programming method for determining the minimum energy path," *The Journal of Chemical Physics*, vol. 127, no. 16, p. 164107, 2007. [Online]. Available: https://doi.org/10.1063/1.2780147

[212] J. Steinmetzer, S. Kupfer, and S. Gräfe, "pysisyphus: Exploring potential energy surfaces in ground and excited states," *International Journal of Quantum Chemistry*, vol. 121, no. 3, p. e26390, 2021.

[213] J. Marks and J. Gomes, "Incorporation of internal coordinates interpolation into the freezing string method," *Journal of Chemical Theory and Computation*, vol. 21, no. 23, pp. 12 110–12 120, 2025.

[214] T. A. Halgren and W. N. Lipscomb, "The synchronous-transit method for determining reaction pathways and locating molecular transition states," *Chemical Physics Letters*, vol. 49, no. 2, pp. 225–232, 1977.

[215] S. Smidstrup, A. Pedersen, K. Stokbro, and H. Jónsson, "Improved initial guess for minimum energy path calculations," *The Journal of Chemical Physics*, vol. 140, no. 21, p. 214106, 2014.

[216] X. Zhu, K. C. Thompson, and T. J. Martínez, "Geodesic interpolation for reaction pathways," *The Journal of Chemical Physics*, vol. 150, no. 16, p. 164103, 2019.

[217] S. Kim, J. Woo, and W. Y. Kim, "Diffusion-based generative AI for exploring transition states from 2D molecular graphs," *Nature Communications*, vol. 15, p. 341, 2024.

[218] Y. Si, Y. Ma, T. Yu, Y. Liu, Y. Wu, W. Lai, Z. Zhang, J. Shi, L. Guo, O. V. Prezhdo, and M. Liu, "Transition state structure detection with machine learning," *npj Computational Materials*, vol. 11, p. 199, 2025.

[219] S. Choi, "Prediction of transition state structures of gas-phase chemical reactions via machine learning," *Nature Communications*, vol. 14, p. 1168, 2023.

[220] C. Duan, G.-H. Liu, Y. Du, T. Chen, Q. Zhao, H. Jia, C. P. Gomes, E. A. Theodorou, and H. J. Kulik, "Optimal transport for generating transition states in chemical reactions," *Nature Machine Intelligence*, vol. 7, pp. 615–626, 2025.

[221] S. P. Neville, M. Chergui, A. Stolow, and M. S. Schuurman, "Ultrafast X-ray spectroscopy of conical intersections," *Phys. Rev. Lett.*, vol. 120, p. 243001, 2018.

[222] Y. Boeije and M. Olivucci, "From a one-mode to a multi-mode understanding of conical intersection mediated ultrafast organic photochemical reactions," *Chem. Soc. Rev.*, vol. 52, pp. 2643–2687, 2023.

[223] L. Goerigk, A. Hansen, C. Bauer, S. Ehrlich, A. Najibi, and S. Grimme, "A look at the density functional theory zoo with the advanced GMTKN55 database for general main group thermochemistry, kinetics and noncovalent interactions," *Physical Chemistry Chemical Physics*, vol. 19, no. 48, pp. 32 184–32 215, 2017. [Online]. Available: https://pubs.rsc.org/en/content/articlelanding/2017/cp/c7cp04913g

[224] W. J. Hehre, R. Ditchfield, and J. A. Pople, "Self—Consistent Molecular Orbital Methods. XII. Further Extensions of Gaussian—Type Basis Sets for Use in Molecular Orbital Studies of Organic Molecules," *The Journal of Chemical Physics*, vol. 56, no. 5, pp. 2257–2261, 1972. [Online]. Available: https://doi.org/10.1063/1.1677527

[225] P. C. Hariharan and J. A. Pople, "The influence of polarization functions on molecular orbital hydrogenation energies," *Theoretica chimica acta*, vol. 28, no. 3, pp. 213–222, 1973.

[226] E. Epifanovsky, A. T. B. Gilbert, X. Feng, J. Lee, Y. Mao, N. Mardirossian, P. Pokhilko, A. F. White, M. P. Coons, A. L. Dempwolff, Z. Gan, D. Hait, P. R. Horn, L. D. Jacobson, I. Kaliman, J. Kussmann, A. W. Lange, K. U. Lao, D. S. Levine, J. Liu, S. C. McKenzie, A. F. Morrison, K. D. Nanda, F. Plasser, D. R. Rehn, M. L. Vidal, Z.-Q. You, Y. Zhu, B. Alam, B. J. Albrecht, A. Aldossary, E. Alguire, J. H. Andersen, V. Athavale, D. Barton, K. Begam, A. Behn, N. Bellonzi, Y. A. Bernard, E. J. Berquist, H. G. A. Burton, A. Carreras, K. Carter-Fenk, R. Chakraborty, A. D. Chien, K. D. Closser, V. Cofer-Shabica, S. Dasgupta, M. de Wergifosse, J. Deng, M. Diedenhofen, H. Do, S. Ehlert, P.-T. Fang, S. Fatehi, Q. Feng, T. Friedhoff, J. Gayvert, Q. Ge, G. Gidofalvi, M. Goldey, J. Gomes, C. E. González-Espinoza, S. Gulania, A. O. Gunina, M. W. D. Hanson-Heine, P. H. P. Harbach, A. Hauser, M. F. Herbst, M. Hernández Vera, M. Hodecker, Z. C. Holden, S. Houck, X. Huang, K. Hui, B. C. Huynh, M. Ivanov, Á. Jász, H. Ji, H. Jiang, B. Kaduk, S. Kähler, K. Khistyaev, J. Kim, G. Kis, P. Klunzinger, Z. Koczor-Benda, J. H. Koh, D. Kosenkov, L. Koulias, T. Kowalczyk, C. M. Krauter, K. Kue, A. Kunitsa, T. Kus, I. Ladjánszki, A. Landau, K. V. Lawler,

D. Lefrancois, S. Lehtola, R. R. Li, Y.-P. Li, J. Liang, M. Liebenthal, H.-H. Lin, Y.-S. Lin, F. Liu, K.-Y. Liu, M. Loipersberger, A. Luenser, A. Manjanath, P. Manohar, E. Mansoor, S. F. Manzer, S.-P. Mao, A. V. Marenich, T. Markovich, S. Mason, S. A. Maurer, P. F. McLaughlin, M. F. S. J. Menger, J.-M. Mewes, S. A. Mewes, P. Morgante, J. W. Mullinax, K. J. Oosterbaan, G. Paran, A. C. Paul, S. K. Paul, F. Pavošević, Z. Pei, S. Prager, E. I. Proynov, Á. Rák, E. Ramos-Cordoba, B. Rana, A. E. Rask, A. Rettig, R. M. Richard, F. Rob, E. Rossomme, T. Scheele, M. Scheurer, M. Schneider, N. Sergueev, S. M. Sharada, W. Skomorowski, D. W. Small, C. J. Stein, Y.-C. Su, E. J. Sundstrom, Z. Tao, J. Thirman, G. J. Tornai, T. Tsuchimochi, N. M. Tubman, S. P. Veccham, O. Vydrov, J. Wenzel, J. Witte, A. Yamada, K. Yao, S. Yeganeh, S. R. Yost, A. Zech, I. Y. Zhang, X. Zhang, Y. Zhang, D. Zuev, A. Aspuru-Guzik, A. T. Bell, N. A. Besley, K. B. Bravaya, B. R. Brooks, D. Casanova, J.-D. Chai, S. Coriani, C. J. Cramer, G. Cserey, A. E. I. DePrince, R. A. J. DiStasio, A. Dreuw, B. D. Dunietz, T. R. Furlani, W. A. I. Goddard, S. Hammes-Schiffer, T. Head-Gordon, W. J. Hehre, C.-P. Hsu, T.-C. Jagau, Y. Jung, A. Klamt, J. Kong, D. S. Lambrecht, W. Liang, N. J. Mayhall, C. W. McCurdy, J. B. Neaton, C. Ochsenfeld, J. A. Parkhill, R. Peverati, V. A. Rassolov, Y. Shao, L. V. Slipchenko, T. Stauch, R. P. Steele, J. E. Subotnik, A. J. W. Thom, A. Tkatchenko, D. G. Truhlar, T. Van Voorhis, T. A. Wesolowski, K. B. Whaley, H. L. I. Woodcock, P. M. Zimmerman, S. Faraji, P. M. W. Gill, M. Head-Gordon, J. M. Herbert, and A. I. Krylov, "Software for the frontiers of quantum chemistry: An overview of developments in the Q-Chem 5 package," *The Journal of Chemical Physics*, vol. 155, no. 8, p. 084801, 2021. [Online]. Available: https://doi.org/10.1063/5.0055522

[227] M. Castiñeira Reis, E. Martínez Núñez, and A. Fernández Ramos, "Comprehensive computational automated search of barrierless reactions leading to the formation of benzene and other c6-membered rings," *Sci. Adv.*, vol. 10, p. eadq4077, 2024.

[228] Y. Mori, Y. Sakaguchi, and H. Hayashi, "Magnetic field effects on chemical reactions of biradical radical ion pairs in homogeneous fluid solvents," *J. Phys. Chem. A*, vol. 104, no. 21, pp. 4896–4905, 2000.

[229] C. Song, "New physical insights into the supporting subspace factorization of xms-caspt2 and generalization to multiple spin states via spin-free formulation," *J. Chem. Phys.*, vol. 160, p. 124106, 2024.

[230] A. V. Marenich, E. N. Brothers, H. P. Hratchian, and M. J. Frisch, "Generalized internal coordinates for creative exploration of interatomic geometries," *J. Chem. Theory Comput.*, vol. 21, no. 21, pp. 10 930–10 944, November 2025.

[231] H. B. Schlegel, "Geometry optimization," *WIREs Computational Molecular Science*, vol. 1, no. 5, pp. 790–809, 2011, comprehensive review of geometry optimization methods, including Newton-Raphson and quasi-Newton (often called pseudo-Newton-Raphson in this context) on potential energy surfaces.

[232] A. Banerjee, N. Adams, J. Simons, and R. Shepard, "Search for stationary points on surfaces," *The Journal of Physical Chemistry*, vol. 89, no. 1, pp. 52–57, 1985.

[233] J. Baker, "An algorithm for the location of transition states," *Journal of Computational Chemistry*, vol. 7, no. 4, pp. 385–395, 1986.

[234] T. Helgaker, "Transition-state optimizations by trust-region image minimization," *Chemical Physics Letters*, vol. 182, no. 5, pp. 503–510, 1991. [Online]. Available: https://www.sciencedirect.com/science/article/pii/000926149190115P

[235] A. Heyden, A. T. Bell, and F. J. Keil, "Efficient methods for finding transition states in chemical reactions: Comparison of improved dimer method and partitioned rational function optimization method," *The Journal of Chemical Physics*, vol. 123, no. 22, p. 224101, 2005.

[236] C. Gonzalez and H. B. Schlegel, "An improved algorithm for reaction path following," *The Journal of Chemical Physics*, vol. 90, no. 4, pp. 2154–2161, 1989. [Online]. Available: https://doi.org/10.1063/1.456010

[237] F. Jenson, *Introduction to Computational Chemistry*. John Wiley & Sons, 1999.

[238] B. Peters, A. Heyden, A. T. Bell, and A. Chakraborty, "A growing string method for determining transition states: Comparison to the nudged elastic band and string methods," *The Journal of Chemical Physics*, vol. 120, no. 17, pp. 7877–7886, 2004.

[239] W. Quapp, "A growing string method for the reaction pathway defined by a Newton trajectory," *The Journal of Chemical Physics*, vol. 122, no. 17, p. 174106, 2005. [Online]. Available: https://doi.org/10.1063/1.1885467

[240] A. Goodrow, A. T. Bell, and M. Head-Gordon, "Development and application of a hybrid method involving interpolation and ab initio calculations for the determination of transition states," *The Journal of Chemical Physics*, vol. 129, no. 17, p. 174109, 2008. [Online]. Available: https://doi.org/10.1063/1.2992618

[241] ——, "Transition state-finding strategies for use with the growing string method," *The Journal of Chemical Physics*, vol. 130, no. 24, p. 244108, 2009.

[242] ——, "A strategy for obtaining a more accurate transition state estimate using the growing string method," *Chemical Physics Letters*, vol. 484, no. 4, pp. 392–398, 2010.

[243] A. Behn, P. M. Zimmerman, A. T. Bell, and M. Head-Gordon, "Incorporating Linear Synchronous Transit Interpolation into the Growing String Method: Algorithm and Applications," *Journal of Chemical Theory and Computation*, vol. 7, no. 12, pp. 4019–4025, 2011.

[244] C. J. Cerjan and W. H. Miller, "On finding transition states," *The Journal of Chemical Physics*, vol. 75, no. 6, pp. 2800–2806, 1981.

[245] A. Jain, S. P. Ong, W. Chen, B. Medasani, X. Qu, M. Kocher, M. Brafman, G. Petretto, G.-M. Rignanese, G. Hautier, D. Gunter, and K. A. Persson, "FireWorks: a dynamic workflow system designed for high-throughput applications," *Concurrency and Computation: Practice and Experience*, vol. 27, no. 17, pp. 5037–5059, 2015.

[246] C. B. Hicks and T. J. Martinez, "Massively scalable workflows for quantum chemistry: BigChem and ChemCloud," *The Journal of Chemical Physics*, vol. 160, no. 14, p. 142501, 2024. [Online]. Available: https://doi.org/10.1063/5.0190834

[247] F. Zapata, L. Ridder, J. Hidding, C. R. Jacob, I. Infante, and L. Visscher, "QMflows: A Tool Kit for Interoperable Parallel Workflows in Quantum Chemistry," *Journal of Chemical Information and Modeling*, vol. 59, no. 7, pp. 3191–3197, 2019. [Online]. Available: https://doi.org/10.1021/acs.jcim.9b00384

[248] V. Ásgeirsson, B. O. Birgisson, R. Bjornsson, U. Becker, F. Neese, C. Riplinger, and H. Jónsson, "Nudged Elastic Band Method for Molecular Reactions Using Energy-Weighted Springs Combined with Eigenvector Following," *Journal of Chemical Theory and Computation*, vol. 17, no. 8, pp. 4929–4945, 2021.

[249] D. Mandelli and M. Parrinello, "A modified nudged elastic band algorithm with adaptive spring lengths," *The Journal of Chemical Physics*, vol. 155, no. 7, p. 074103, 2021. [Online]. Available: https://doi.org/10.1063/5.0059593

[250] Y. Mitsuta and T. Asada, "Nudged elastic stiffness band method: A method to solve kinks problems of reaction paths," *Journal of Computational Chemistry*, vol. 44, no. 23, pp. 1884–1897, 2023. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/jcc.27169

[251] ——, "Curvature-weighted nudged elastic band method using the Riemann curvature," *Journal of Computational Chemistry*, vol. 44, no. 5, pp. 662–669, 2023. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/jcc.27030

[252] C. Bergonzo, A. J. Campbell, R. C. Walker, and C. Simmerling, "A Partial Nudged Elastic Band Implementation for Use with Large or Explicitly Solvated Systems," *International Journal of Quantum Chemistry*, vol. 109, no. 15, p. 3781, 2009.

[253] L. Zhao, K.-j. Watanabe, N. Nakatani, A. Nakayama, X. Xu, and J.-y. Hasegawa, "Extending nudged elastic band method to reaction pathways involving multiple spin states," *The Journal of Chemical Physics*, vol. 153, no. 13, p. 134114, 2020. [Online]. Available: https://doi.org/10.1063/5.0021923

[254] K. Atz, F. Grisoni, and G. Schneider, "Geometric deep learning on molecular representations," *Nature Machine Intelligence*, vol. 3, no. 12, pp. 1023–1032, 2021.

[255] S. Raghunathan and U. D. Priyakumar, "Molecular representations for machine learning applications in chemistry," *International Journal of Quantum Chemistry*, vol. 122, no. 7, p. e26870, 2022.

[256] M. Baek, F. DiMaio, I. Anishchenko, J. Dauparas, S. Ovchinnikov, G. R. Lee, J. Wang, Q. Cong, L. N. Kinch, R. D. Schaeffer, C. Millán, H. Park, C. Adams, C. R. Glassman, A. DeGiovanni, J. H. Pereira, A. V. Rodrigues, A. A. van Dijk, A. C. Ebrecht, D. J. Opperman, T. Sagmeister, C. Buhlheller, T. Pavkov-Keller, M. K. Rathinaswamy, U. Dalwadi, C. K. Yip, J. E. Burke, K. C. Garcia, N. V. Grishin, P. D. Adams, R. J. Read, and D. Baker, "Accurate prediction of protein structures and interactions using a three-track neural network," *Science (New York, N.Y.)*, vol. 373, no. 6557, pp. 871–876, 2021.

[257] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli, and D. Hassabis, "Highly accurate protein structure prediction with AlphaFold," *Nature*, vol. 596, no. 7873, pp. 583–589, 2021. [Online]. Available: https://www.nature.com/articles/s41586-021-03819-2

[258] E. Gawehn, J. A. Hiss, and G. Schneider, "Deep Learning in Drug Discovery," *Molecular Informatics*, vol. 35, no. 1, pp. 3–14, 2016.

[259] J. Jiménez-Luna, F. Grisoni, N. Weskamp, and G. Schneider, "Artificial intelligence in drug discovery: recent advances and future perspectives," *Expert Opinion on Drug Discovery*, vol. 16, no. 9, pp. 949–959, 2021.

[260] M. Korshunova, B. Ginsburg, A. Tropsha, and O. Isayev, "OpenChem: A Deep Learning Toolkit for Computational Chemistry and Drug Design," *Journal of Chemical Information and Modeling*, vol. 61, no. 1, pp. 7–13, 2021.

[261] M. Nakata and T. Shimazaki, "PubChemQC Project: A Large-Scale First-Principles Electronic Structure Database for Data-Driven Chemistry," *Journal of Chemical Information and Modeling*, vol. 57, no. 6, pp. 1300–1308, 2017.

[262] G. Chen, P. Chen, C.-Y. Hsieh, C.-K. Lee, B. Liao, R. Liao, W. Liu, J. Qiu, Q. Sun, J. Tang, R. Zemel, and S. Zhang, "Alchemy: A quantum chemistry dataset for benchmarking ai models," 2019. [Online]. Available: https://arxiv.org/abs/1906.09427

[263] P. Eastman, P. K. Behara, D. L. Dotson, R. Galvelis, J. E. Herr, J. T. Horton, Y. Mao, J. D. Chodera, B. P. Pritchard, Y. Wang, G. De Fabritiis, and T. E. Markland, "SPICE, A Dataset of Drug-like Molecules and Peptides for Training Machine Learning Potentials," *Scientific Data*, vol. 10, no. 1, p. 11, 2023.

[264] P. Eastman, B. P. Pritchard, J. D. Chodera, and T. E. Markland, "Nutmeg and spice: Models and data for biomolecular machine learning," 2024. [Online]. Available: https://arxiv.org/abs/2406.13112

[265] M. B. Lee, B. Blue, M. Muir, and M. Kaeberlein, "The million-molecule challenge: a moonshot project to rapidly advance longevity intervention discovery," *GeroScience*, vol. 45, no. 6, pp. 3103–3113, 2023. [Online]. Available: https://doi.org/10.1007/s11357-023-00867-6

[266] C. Draxl and M. Scheffler, "NOMAD: The FAIR concept for big data-driven materials science," *MRS Bulletin*, vol. 43, no. 9, pp. 676–682, 2018.

[267] M. Schreiner, A. Bhowmik, T. Vegge, J. Busk, and O. Winther, "Transition1x - a dataset for building generalizable reactive machine learning potentials," *Scientific Data*, vol. 9, no. 1, p. 779, 2022.

[268] M. Schreiner, A. Bhowmik, T. Vegge, P. B. Jørgensen, and O. Winther, "Neuralneb—neural networks can find reaction paths fast," *Machine Learning: Science and Technology*, vol. 3, no. 4, p. 045022, 2022.

[269] D. C. Jamison, "Structured Query Language (SQL) Fundamentals," *Current Protocols in Bioinformatics*, vol. 00, no. 1, pp. 9.2.1–9.2.29, 2003.

[270] A. W. Ruttinger, D. Sharma, and P. Clancy, "Protocol for Directing Nudged Elastic Band Calculations to the Minimum Energy Pathway: Nurturing Errant Calculations Back to Convergence," *Journal of Chemical Theory and Computation*, vol. 18, no. 5, pp. 2993–3005, 2022.

[271] S. Grimme, "Semiempirical GGA-type density functional constructed with a long-range dispersion correction," *Journal of Computational Chemistry*, vol. 27, no. 15, pp. 1787–1799, 2006.

[272] S. Grimme, J. Antony, S. Ehrlich, and H. Krieg, "A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu," *The Journal of Chemical Physics*, vol. 132, no. 15, p. 154104, 2010.

[273] J. J. P. Stewart, "Optimization of parameters for semiempirical methods V: Modification of NDDO approximations and application to 70 elements," *Journal of Molecular Modeling*, vol. 13, no. 12, pp. 1173–1213, 2007.

[274] ——, "Optimization of parameters for semiempirical methods VI: more modifications to the NDDO approximations and re-optimization of parameters," *Journal of Molecular Modeling*, vol. 19, no. 1, pp. 1–32, 2013.

[275] ——, "Mopac2016," http://openmopac.net/, 2016.

[276] L.-P. Wang, Q. Wu, and T. Van Voorhis, "Acid-base mechanism for ruthenium water oxidation catalysts," *Inorganic Chemistry*, vol. 49, no. 10, pp. 4543–4553, 2010.

[277] H. Jang, Y. Qiu, M. E. Hutchings, M. Nguyen, L. A. Berben, and L.-P. Wang, "Quantum chemical studies of redox properties and conformational changes of a four-center iron CO $_2$ reduction electrocatalyst," *Chemical Science*, vol. 9, no. 10, pp. 2645–2654, 2018.

[278] Revvity Signals Software, "Chemdraw," 2025, version 25.0. [Online]. Available: https://revvitysignals.com/products/research/chemdraw

[279] P. J. Hay and W. R. Wadt, "Ab initio effective core potentials for molecular calculations. Potentials for K to Au including the outermost core orbitals," *The Journal of Chemical Physics*, vol. 82, no. 1, pp. 299–310, 1985.

[280] J. P. Perdew, "Density-functional approximation for the correlation energy of the inhomogeneous electron gas," *Physical Review B*, vol. 33, no. 12, pp. 8822–8824, 1986.

[281] E. Maras, O. Trushin, A. Stukowski, T. Ala-Nissila, and H. Jónsson, "Global transition path search for dislocation formation in Ge on Si(001)," *Computer Physics Communications*, vol. 205, pp. 13–21, 2016.

[282] D. G. A. Smith, A. T. Lolinco, Z. L. Glick, J. Lee, A. Alenaizan, T. A. Barnes, C. H. Borca, R. Di Remigio, D. L. Dotson, S. Ehlert, A. G. Heide, M. F. Herbst, J. Hermann, C. B. Hicks, J. T. Horton, A. G. Hurtado, P. Kraus, H. Kruse, S. J. R. Lee, J. P. Misiewicz, L. N. Naden, F. Ramezanghorbani, M. Scheurer, J. B. Schriber, A. C. Simmonett, J. Steinmetzer, J. R. Wagner, L. Ward, M. Welborn,

D. Altarawy, J. Anwar, J. D. Chodera, A. Dreuw, H. J. Kulik, F. Liu, T. J. Martínez, D. A. Matthews, H. F. Schaefer, J. Šponer, J. M. Turney, L.-P. Wang, N. De Silva, R. A. King, J. F. Stanton, M. S. Gordon, T. L. Windus, C. D. Sherrill, and L. A. Burns, "Quantum Chemistry Common Driver and Databases (QCDB) and Quantum Chemistry engine (QCEngine): Automation and interoperability among computational chemistry programs," *The Journal of Chemical Physics*, vol. 155, no. 20, p. 204801, 2021.

[283] Y. Babuji, A. Woodard, Z. Li, D. S. Katz, B. Clifford, R. Kumar, L. Lacinski, R. Chard, J. M. Wozniak, I. Foster, M. Wilde, and K. Chard, "Parsl: Pervasive Parallel Programming in Python," in *Proceedings of the 28th International Symposium on High-Performance Parallel and Distributed Computing*, ser. HPDC '19. New York, NY, USA: Association for Computing Machinery, 2019, pp. 25–36.

[284] J. Demmel and H. D. Nguyen, "Parallel reproducible summation," *IEEE Transactions on Computers*, vol. 64, no. 7, pp. 2060–2070, jul 2015.

[285] S. Shanmugavelu, M. Taillefumier, C. Culver, O. Hernandez, M. Coletti, and A. Sedova, "Impacts of floating-point non-associativity on reproducibility for hpc and deep learning applications," 2024. [Online]. Available: https://arxiv.org/abs/2408.05148