

MACHINE LEARNING ESTIMATION OF REACTION ENERGY BARRIERS AND ITS APPLICATIONS IN ASTROCHEMISTRY

HONGCHEN JI

A DISSERTATION SUBMITTED TO
THE FACULTY OF GRADUATE STUDIES
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

GRADUATE PROGRAM IN CHEMISTRY
YORK UNIVERSITY
TORONTO, ONTARIO

December 2023

© Hongchen Ji, 2023

Abstract

We developed a machine learning model for fast estimating reaction energy barriers. The model was trained on data for 11,730 elementary reactions and barriers computed with an estimated accuracy of 2.3 kcal/mol by Grambow and coworkers. Although it was trained, and then applied, only for reactions involving atoms of H, C, N and O, our model can readily be generalized to molecules and reactions involving other elements. We designed and tested many molecular representations. Our best model has 363 features calculated from the chemical composition, structure, and energy of products and reactants that can all be obtained at a small computational cost.

A Kernel Ridge Regression with Laplacian kernel was found to give the best fit to the data. It makes predictions with a mean absolute error of 4.1 kcal/mol for barriers smaller than 40 kcal/mol. We used this machine learning model to estimate the barriers of 136,081 hypothetical association reactions between molecules known to exist in the circumstellar envelopes and interstellar medium, where temperatures range between 10 and 150 K. A screening procedure identified reactions likely to occur (those with near zero barriers) that could play a role in the formation of relatively complex organic molecules (those with at least one double bond and containing at least one atom each of the elements H, C, N and O). Reactions identified as the most promising were investigated with density functional theory and coupled cluster quantum chemical methods to obtain reaction pathways and energies of reactants, intermediates, transition states and products with an accuracy of roughly 1 kcal/mol. We found no barrierless reaction but found two reactions with low

barriers leading to the formation of *N*-methyleneformamide (CH_2NCHO) and imine acetaldehyde (NHCHCHO). These molecules have not yet been observed in space. The low barriers that we calculated, and the possibility that these reactions would be facilitated by adsorption on ice covered dust grains in the interstellar medium, suggest that these two molecules may exist in space and could be detected by determinations of rotational constants from spectroscopic lines in the microwave region.

Acknowledgments

I would like to express my deepest and sincerest appreciation to my supervisor, Professor René Fournier, for all his thoughtful instructions, knowledgeable teaching, considerable guidance, encouraging support, unconditional patience, friendly concerns, and eternal goodwill. This endeavour would not have been possible without him. In the past five years, every time I met him, his passion for scientific research and encouragement provided for me always filled me with positive hopes for my research and gratitude for everything. Words cannot express my appreciation to you for bringing me the best growth in my academic abilities and for professionally supporting me in this journey. Thank you very much and very sincerely for all your precious time and all your help.

I would like to extend my sincerest thanks to my family members, Yanmei Li, for your great love and endless kindness, and for being always there forever no matter who I am, and Peiju Yan, Fengqing Ji, Hongtian Ji, and Hongfu Ji, for your unconditional love, best support, and all warm moments that you bring to me in my life. Without all of you, I would never be able to become the current me. I truly and deeply appreciate all your love and appreciate your being my family, from the bottom of my heart.

Many thanks to my friends, Dr. Anita Rágyanszki, for your sincere help and infinite patience in my academic achievements, and for warmly and unconditionally helping me when I encounter sad things, Tong Ye Wang, for your good faith, generous help, and for listening with best understanding and sincere trust, and Anirudh Krishnadas, for your

constructive advice and kind help. Thank you for your accompaniment and your golden friendships.

I am also very grateful to my supervisory committee members, Professor William J. Pietro and Professor Tao Zeng, for their precious time, generous suggestions, professional expertise, kind guidance, and warm help offered to me during this journey.

I would also like to acknowledge many professors, friends, graduate program assistants, administrative assistants, and laboratory technicians, from York University, who provided me with warm-hearted help throughout this journey. Thank you for all your warmth. I also gratefully acknowledge Chengfei Xie from Zhejiang Sci-Tech University for partial technical help. Thank you for your kind help.

We acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC). This research was enabled in part by support provided by Sharcnet (sharcnet.ca) and the Digital Research Alliance of Canada (alliancecan.ca).

Table of Contents

<i>Abstract</i>	<i>ii</i>
<i>Acknowledgments</i>	<i>iv</i>
<i>Table of Contents</i>	<i>vi</i>
<i>List of Tables</i>	<i>ix</i>
<i>List of Figures</i>	<i>x</i>
<i>List of Abbreviations</i>	<i>xiii</i>
Chapter 1 Introduction	1
1.1 Contributions by Myself and Others to This Work	3
1.2 Background	4
1.3 Quantum Chemical Methods for Estimating Reaction Energy Barrier	6
1.4 Introduction to Machine Learning	12
1.4.1 Elements of a Machine Learning Model	12
1.5 Interstellar Medium (ISM) and Circumstellar Envelopes (CSE)	14
1.6 Outline of the Dissertation	16
1.6.1 Development of a Machine Learning Model.....	16
1.6.2 Computational Search for Prebiotic Chemistry in the ISM and CSE.....	17
Chapter 2 Computational Techniques	19
2.1 Quantum Chemistry and the Gaussian Software	19
2.1.1 Molecular Geometry Optimization.....	19
2.1.2 Molecular Frequency and Energy Calculation	20
2.1.3 Transition States	21
2.1.4 Intrinsic Reaction Coordinate (IRC) Calculation	23
2.1.5 Molecular Single-Point Energy Calculation.....	24
2.1.6 Potential Energy Surface Scan.....	25
2.2 Machine Learning (ML) Techniques	26
2.2.1 Methodology of Machine Learning	26
2.2.2 Data Processing of Machine Learning.....	27
2.2.3 Ordinary Least Squares (OLS)	27
2.2.4 Ridge Regression (RR)	28
2.2.5 Weighted Least Squares (WLS)	29
2.2.6 Feed-Forward Neural Network (FFNN).....	30
2.2.7 Kernel Ridge Regression (KRR)	31
2.3 Applications of ML to Chemistry	33
Chapter 3 Kernel Ridge Regression Model for Predicting Reaction Energy Barrier	37
3.1 Introduction	37
3.2 Dataset of Chemical Reactions	38

3.2.1 Initial Reaction Dataset.....	38
3.2.2 Dataset of Grambow, Pattanaik and Green.....	39
3.3 Features of Chemical Reactions	40
3.3.1 Tests and Features Selection.....	40
3.3.2 Features Used in the ML Models.....	41
3.3.3 Features Set 1.....	46
3.3.4 Features Set 2.....	52
3.3.5 Different Approaches for Calculating Features.....	56
3.4 KRR Model for Energy Barriers.....	57
3.5 Performance of Various Machine Learning Models	62
3.6 Discussion and Conclusions	74
<i>Chapter 4 Screening Reactions for Relevance to Prebiotic Chemistry in the ISM and CSE</i>	78
4.1 ISM and CSE Reaction Dataset.....	78
4.1.1 Reactant Side	80
4.1.2 Product Side.....	80
4.1.3 Enriched ISM and CSE Reaction Dataset	80
4.2 Procedure of Screening	81
4.2.1 Screening based on Energy Barrier	82
4.2.2 Screening by Biological Relevance	82
4.3 Discussion and Conclusions	85
<i>Chapter 5 The Reactions of Methylene with HCNO and HNCO.....</i>	86
5.1 Introduction.....	86
5.2 Computational Details.....	88
5.3 $^1\text{CH}_2/^3\text{CH}_2 + \text{HCNO}$	90
5.3.1 Singlet State Energy Profile for $\text{CH}_2 + \text{HCNO}$	91
5.3.2 Triplet State Energy Profile for $\text{CH}_2 + \text{HCNO}$	94
5.3.3 Reaction Paths for $\text{CH}_2 + \text{HCNO}$	98
5.4 Reactions of CH_2 with HNCO	100
5.4.1 Energy Profile of Singlet $\text{CH}_2 + \text{HNCO}$	100
5.4.2 Energy Profile of Triplet $\text{CH}_2 + \text{HNCO}$	102
5.4.3 Reaction Paths for $\text{CH}_2 + \text{HNCO}$	105
5.5 The Reaction of $^1\text{CH}_2$ with $\text{CH}_3\text{C}(=\text{O})\text{NH}_2$.....	107
5.5.1 A Precursor of Alanine May Exist in the ISM or CSE: Alaninal	107
5.5.2 Energy Profile of $^1\text{CH}_2 + \text{CH}_3\text{C}(=\text{O})\text{NH}_2$	107
5.6 The $\text{H}_2\text{O} + \text{NHCHCHO}$ Reaction.....	111
5.6.1 A Reaction Route Producing Glycine.....	111
5.6.2 Energy Profile for the Formation of Glycine.....	112
5.7 Discussion and Conclusions	114
<i>Chapter 6 Energy, Structure, and Spectroscopic Study of Conformers of N-methyleneformamide and Imine Acetaldehyde</i>	116

6.1 <i>N</i>-methyleneformamide.....	116
6.2 Imine Acetaldehyde	117
6.3 Computational Details.....	117
6.4 Results	119
6.4.1 Geometries and Relative Energies of Conformers	119
6.4.2 Spectroscopic Constants of Conformers.....	121
<i>Chapter 7 Conclusion.....</i>	123
<i>Bibliography</i>	126
<i>Appendix</i>	171
Appendix A: Molecular Structure Information of All Studied Species in Chapter 5.....	171
Appendix B: Total Energies Including Zero-Point Corrections (E, in a.u.), and Relative Energies (RE, in kcal/mol) for Reactants, Reactant Complexes, Transition States, Reaction Intermediates, and Products for Reactions in Chapter 5	209
Appendix C: Structure Information of All Conformers of <i>N</i>-methyleneformamide and Imine Acetaldehyde in Chapter 6.....	211

List of Tables

Table 3-1. Selection procedure of best hyperparameters by <i>GridSearchCV</i> for KRR.....	61
Table 3-2. Performance statistics (MAE, RMSE, R^2 , R^2_{adj}) of six ML models using the same 117 features calculated in the X,DFT//DFT approach. The mean and standard deviation over ten testing sets generated by 10-time Monte Carlo cross-validation are shown for each statistic	63
Table 3-3. Performance statistics (MAE, RMSE, R^2 , R^2_{adj}) of KRR models using features calculated in seven approaches. The mean and standard deviation over ten testing sets generated by 10-time Monte Carlo cross-validation are shown for each statistic.....	65
Table 3-4. Statistical description of two datasets	72
Table 3-5. Performance statistics (RMSE, R^2 , MAE) of KRR models for sets D ₁ and D ₂ (the entire set minus the outliers) using the features calculated in six approaches. The mean and standard deviation over ten testing sets generated by 10-time Monte Carlo cross-validation are shown for each statistic	73
Table 4-1. List of 72 molecules known, or believed, to be present in the ISM and CSE [200–260] that are in our molecule database [20,39] and their SMILES representation [163]	79
Table 4-2. Standard amino acids, their degree of unsaturation n_u , number of non-hydrogen atoms n_x , and ratio n_u/n_x	84
Table 4-3. Representative mono and disaccharides, their degree of unsaturation n_u , number of non-hydrogen atoms n_x , and ratio n_u/n_x	85
Table 6-1. Total energies including zero-point corrections (E, a.u.) and relative energies (RE, kcal/mol) calculated with DFT (in parentheses) and CCSD(T) for the conformers of <i>N</i> -methyleneformamide and imine acetaldehyde	120
Table 6-2. Structural constants including bond lengths (in angstroms) and bond angles (in degrees) optimized with ω B97XD/def2-TZVP of the conformers of <i>N</i> -methyleneformamide and imine acetaldehyde	120
Table 6-3. Spectroscopic equilibrium rotational constants (A_e , B_e , C_e , in GHz) with moments of inertia in parentheses (I_A , I_B , I_C , in amu Å ²), ground vibrational state rotational constants (A_0 , B_0 , C_0 , in GHz) with and without centrifugal distortion (in parentheses), and dipole moment (μ , in Debye) of conformers of <i>N</i> -methyleneformamide and imine acetaldehyde	121

List of Figures

Figure 1-1. Workflow of overall process presented in this dissertation with pointing out applications to astrochemistry, especially to ISM and CSE.....	16
Figure 1-2. Workflow for acquiring the most possible and biologically meaningful reactions in the ISM and CSE.....	18
Figure 2-1. Gaussian input file of one simple optimization example (geometry optimization of ethanol at ω B97XD/def2-TZVP level of theory).....	20
Figure 2-2. Gaussian input file of one simple example of frequency and energy calculation (of ethanol at ω B97XD/def2-TZVP level of theory).	21
Figure 2-3. Gaussian input file of one simple example of TS optimization (for a TS between <i>anti</i> and <i>gauche</i> conformers of ethanol at ω B97XD/def2-TZVP level of theory).	22
Figure 2-4. Gaussian input file of one simple example of IRC calculation (starting from an assumed TS between <i>anti</i> and <i>gauche</i> conformers of ethanol at ω B97XD/def2-TZVP level of theory).	23
Figure 2-5. Gaussian input file of a calculation based on the geometry of ethanol at CCSD(T)/Aug-cc-pVQZ level of theory.....	24
Figure 2-6. Gaussian input file of one simple example of PES scan of CH ₂ at ω B97XD/def2-TZVP level of theory.....	25
Figure 2-7. Flowchart of our best FFNN-3HL (produced by MATLAB Online).....	30
Figure 3-1. Coulomb matrix for ethanol at its equilibrium geometry.	47
Figure 3-2. Procedure of how to rearrange data by energy barrier and how to build the testing set where the number of total bins is M_0	59
Figure 3-3. The 5-fold cross-validation process achieved in <i>GridSearchCV</i> . The blue "S" is a testing set in 5-fold cross-validation occupying 1/5 of the original Training Set, and pink "T" is a training set in 5-fold cross-validation occupying 4/5 of the original Training Set.	60
Figure 3-4. Trend in MAE and RMSE for the training and testing sets as a function of γ with a fixed $\alpha = 0.05$. The grey rectangle indicates the range of acceptable γ values.....	61
Figure 3-5. A: In-sample correlation of predicted and reference barriers for the KRR model Babel,PM7//PM7. B: Out-of-sample correlation (same method as A).	68

Figure 3-5. C: In-sample correlation of predicted and reference barriers for the KRR model Babel,DFT//DFT. D: Out-of-sample correlation (same method as C).	69
Figure 3-6. E: In-sample correlation of predicted and reference barriers for the KRR model PM7,PM7//PM7. F: Out-of-sample correlation (same method as E).	70
Figure 3-6. G: In-sample correlation of predicted and reference barriers for the KRR model DFT,DFT//DFT. H: Out-of-sample correlation (same method as G).	71
Figure 5-1. Singlet potential energy diagrams showing the energies of reactant complexes (RCs), intermediates of reaction (IMs), transition states (TSs) and products (Ps) of reactions between $^1\text{CH}_2$ and HCNO. The relative energies calculated by CCSD(T) and DFT are shown, with DFT in parentheses.	91
Figure 5-1. (Continued).	92
Figure 5-2. Optimized geometries of singlet reactants, RCs, TSs, IMs, and products of reactions (1), (2), and (3) at $\omega\text{B97XD/def2-TZVP}$ level of theory. Cartesian coordinates of all studied species shown here are listed in Appendix A.	93
Figure 5-2. (Continued).	94
Figure 5-3. Triplet potential energy diagrams showing the energies of reactant complexes (RCs), intermediates of reaction (IMs), transition states (TSs) and products (Ps) of reactions between $^3\text{CH}_2$ and HCNO. The relative energies calculated by CCSD(T) and DFT are shown, with DFT in parentheses.	94
Figure 5-3. (Continued).	95
Figure 5-4. Optimized geometries of triplet state molecular species of reactions (1), (2), and (3) at $\omega\text{B97XD/def2-TZVP}$ level of theory. Cartesian coordinates of all studied species shown here are listed in Appendix A.	96
Figure 5-4. (Continued).	97
Figure 5-5. Singlet potential energy diagrams of reactions between $^1\text{CH}_2$ and HNCO. The relative energies calculated by CCSD(T) and DFT are shown, with DFT in parentheses.	101
Figure 5-6. Optimized geometries of singlet reactants, RCs ($^1\text{RC}_2$ and $^1\text{RC}_3$ could not be fully optimized), TSs, IMs, and products of reactions (4) and (5) at $\omega\text{B97XD/def2-TZVP}$ level of theory. Cartesian coordinates of all studied species shown here are listed in Appendix A.	102

Figure 5-7. Triplet potential energy diagrams of reactions between $^3\text{CH}_2$ and HNCO. The relative energies calculated by CCSD(T) and DFT are shown, with DFT in parentheses.	103
Figure 5-8. Optimized geometries of triplet reactants, RCs ($^3\text{RC}_3$ could not be fully optimized), TSs, IMs, and products of reactions (4) and (5) at $\omega\text{B97XD/def2-TZVP}$ level of theory. Cartesian coordinates of all studied species shown here are listed in Appendix A.	104
Figure 5-8. (Continued).	105
Figure 5-9. Singlet potential energy diagram of the reaction between $^1\text{CH}_2$ and $\text{CH}_3\text{C(=O)NH}_2$. The relative energies calculated by DFT (in parentheses) are shown. ..	108
Figure 5-10. Optimized geometries of singlet reactants, RC ($^1\text{RC}_4$ could not be fully optimized), TSs, IMs, and products of reaction (6) at $\omega\text{B97XD/def2-TZVP}$ level of theory. Cartesian coordinates of all studied species shown here are listed in Appendix A.	110
Figure 5-10. (Continued).	111
Figure 5-11. Singlet potential energy diagram of the reaction between H_2O and NHCHCHO . The relative energies calculated by CCSD(T) and DFT are shown, with DFT in parentheses.	112
Figure 5-12. Optimized geometries of singlet reactants, RC, TSs, IM, and product of reaction (7) at $\omega\text{B97XD/def2-TZVP}$ level of theory. Cartesian coordinates of all studied species shown here are listed in Appendix A.	113
Figure 6-1. 3D structures of conformers (from left to right: <i>gauche</i> , <i>cis</i> , and <i>trans</i>) of <i>N</i> -methyleneformamide at $\omega\text{B97XD/def2-TZVP}$ level of theory. Cartesian coordinates of all studied species shown here are listed in Appendix C.	119
Figure 6-2. 3D structures of conformers (the first row from left to right: <i>out-ap</i> , <i>in-ap</i> ; the second row from left to right: <i>out-sp</i> , <i>in-sp</i>) of imine acetaldehyde at $\omega\text{B97XD/def2-TZVP}$ level of theory. Cartesian coordinates of all studied species shown here are listed in Appendix C.	119

List of Abbreviations

AI	Artificial intelligence
CC	Coupled Cluster
CSE	Circumstellar envelopes
DFT	Density functional theory
DU	Degree of unsaturation
FFNN	Feedforward neural network
FFNN-3HL	Feedforward neural network with three hidden layers
FVU	Fraction of variance that is unexplained
GK	Gaussian kernel
HOMO	Highest occupied molecular orbital
IM	Intermediate of reaction
IRC	Intrinsic reaction coordinate
ISM	Interstellar medium
KRR	Kernel ridge regression
LK	Laplacian kernel
LUMO	Lowest unoccupied molecular orbital
MAE	Mean absolute error

ML	Machine learning
OLS	Ordinary least squares
PED	Potential energy diagram
PES	Potential energy surface
PPS	Possible product side
PRS	Possible reactant side
QC	Quantum chemistry
RC	Reactant complex
RDUNH	Ratio of the degree of unsaturation to the number of non-Hydrogen atom(s)
RE	Relative energy
RMSE	Root-mean square error
RR	Ridge regression
SE	Sorted eigenvalues
SEM	Semi-Empirical Method
SMILES	Simplified Molecular Input Line Entry System
SPE	Single-point energy
TS	Transition State
WLS	Weighted least squares

Chapter 1 Introduction

In quantum chemistry, predicting transition states (TS) and energy barriers of chemical reactions is of great significance for determining the chemical reactivities of molecules [1], analyzing rate constants of reactions, and depicting potential energy surfaces along reaction coordinates. In this dissertation, TS stands for ‘transition states’, and the symbol E is used to represent energy barriers in mathematical formulas. Reaction energy barriers are essential for calculating a reaction rate constant. It implies whether a reaction easily occurs or not.

The common methods for findings TS and energy barriers are quantum chemistry and electronic structure theory methods, for instance, Hartree-Fock (HF), Density Functional Theory (DFT), and Coupled Cluster (CC) theory [1–14]. Among these methods, several correlated electronic structure methods (e.g., coupled cluster with single and double excitations and with perturbative triple excitations, CCSD(T) [15–19]) can estimate reaction barrier with an error of only ~ 1 kcal/mol [8,9], but they require large computational costs for big molecule systems [20]. For example, a single-point energy calculation (based on a fixed geometry) with CCSD(T) for one molecule with 12 atoms (five non-hydrogen atoms and seven hydrogen atoms) needs 5 to 7 days to complete. Molecular geometry optimization, frequency calculation, and TS-finding with CCSD(T) take much longer. Usually, decreasing computational costs leads to a reduction in accuracy.

Hence, a more approximate method for small or middle-size systems is favoured, such as DFT. For instance, the DFT method ω B97XD [21–24], which gives a root-mean-square error (RMSE) of 2.34 kcal/mol for estimating 206 reaction energy barriers [5], takes roughly two to five days to generate the energy curve of a reaction pathway involving seven atoms and passing through two transition states with a basis set def2-TZVP [25–27]. This is much faster than applying CCSD(T) to the same problem.

However, these quantum chemical techniques are not suitable for very large and complex systems such as oceans, atmospheres, cryosphere, or the interstellar medium (ISM) and circumstellar envelopes (CSE), which may require thousands or even millions of TS, molecular geometries, and energy barrier estimates [20]. This large order of magnitude of reactions is practically impossible to handle with current quantum chemistry (QC).

The anticipated significance of the work is twofold. First, it shows a strategy for estimating energy barriers by ML that could be applied, in the future, to the study of large numbers of chemical reactions in complex systems. Second, it shows how one can identify reactions and molecules of interest for chemical evolution in the ISM and CSE [20,28,39] with estimates of energy barriers followed and applying additional criteria related to the composition and structure of products.

The next sections in this chapter, and Chapter 2, give an introduction, background knowledge, relevant concepts of our study, and describe the previous work conducted by others.

These research articles and manuscripts form the basis for chapters 2 to 6:

Ji, H.; Rágyanszki, A.; Fournier, R. Machine Learning Estimation of Reaction Energy Barriers. *Comput. Theor. Chem.* **2023**, *1229*, 114332. <https://doi.org/10.1016/j.comptc.2023.114332>. (published) (chapters 2 and 3)

Ji, H.; Rágyanszki, A.; Fournier, R. Computational Study of the Reactions of Interstellar Molecules: CH₂ Reacting with HCNO and HNCO. *ChemRxiv* **2023**. <https://doi.org/10.26434/chemrxiv-2023-kpm5r>. This content is a preprint and has not been peer-reviewed. (submitted to a peer-reviewed journal) (chapters 4 and 5)

Ji, H.; Rágyanszki, A.; Fournier, R. Computational Structural and Spectroscopic Study on Conformers of Potential Interstellar Molecules. (draft manuscript) (chapter 6)

Ji, H.; Rágyanszki, A.; Fournier, R. Computational Study of Potential Interstellar Reactions: CH₂ Reacting with CH₃C(=O)NH₂ and H₂O Reacting with NHCHCHO. (draft manuscript) (chapters 4 and 5)

1.1 Contributions by Myself and Others to This Work

Chapters 3 to 6 describe original work by myself and co-workers Anita Rágyanszki and René Fournier. I was the main contributor and the first author on research articles and manuscripts based on the work described in those chapters. To be specific: (1) I collected and pre-processed the data used in training and testing ML models; (2) I wrote or developed Python and MATLAB codes to build different ML models, to optimize parameters and hyperparameters of different ML models, and to perform model validation; (3) I ran a great number of tests to choose and optimize the features; (4) I performed the work that takes a list of molecules observed in the ISM and CSE as input and generates a list of association

reactions (this created 136,081 reactions), and I performed features calculations for those 136,081 reactions; (5) I performed all the quantum chemistry calculations that locate minima and transition states for characterizing reaction pathways shown in this dissertation; (6) The figures, tables, reaction schemes, and first drafts of manuscripts were done by me.

The following parts of the work reported in this dissertation were done collaboratively with significant and roughly equal contributions from myself, Anita Rágyanszki and René Fournier: overall strategy adopted for this project; search of literature; design and choice of features; choice of quantum chemistry and machine learning methods; methodology and criteria to select reactions of astrochemical interest for study by quantum chemistry.

Anita Rágyanszki first brought our attention to astrochemistry, guided our initial efforts in that direction, and performed the calculations for ground vibrational state rotational constants. René Fournier designed and wrote the computer code for reactant-product alignment and calculation of features designated as “Set 2” in section 3.3.4. Discussions with Dylan James and Jennifer van Wijngaarden clarified to us the relation between optimized molecular structures and rotational spectroscopic constants often used for identification of molecules in the ISM.

1.2 Background

The processes of chemical and biospheric evolution existing in space are of significant interest to scientists due to the linkage between the discovery of molecules in the universe and the origin of life [29–31]. Detection and study of possible molecules, gases, and solid matter in space contribute to exploring suitable and favourable circumstances for the

emergence of life. The list of molecules that have been identified in the ISM has roughly 250 entries and is growing [32]. More than one thousand organic molecular species were found in the Murchison meteorite [33]. As more extraterrestrial molecules are discovered, the likelihood for the presence in space of relatively big biologically important molecules increases and this matters for theories about the interaction of prebiotic molecules and early Earth [33–35].

The prerequisite for investigating the formation of life-relevant substances in a specific cosmic area is that some known small molecules must be detected. In the ISM and CSE, more than 250 molecular species have been detected through astronomical observations or are suspected to be present [32,36–39]. The increasing number of small organic molecules discovered in the ISM or CSE [32,36] implies that prebiotic molecules could be formed somewhere in space and then reach our early Earth [33–35]. This could be significant for the emergence of early-stage life [33–35,39]. Based on these known molecules, exploring more detailed ISM and CSE constituents is encouraging.

These ISM and CSE molecules could react with each other in a myriad of ways. The detailed interaction of these molecules and the mechanism of producing prebiotic organic molecules in the ISM and CSE are currently unclear due to remote distance and limited instruments. Roughly 60% of the molecules in the ISM and CSE [36] are composed of C, H, O and N atoms, which are the four most abundant elements in the universe [29–31] and are obviously important in studies of the origins of life. The chemical evolution involving the molecules composed of these elements spans a broad range of possibilities. If one has

M different molecular species and tries to generate association reactions with two reactants from those M molecules, $[(M \times (M + 1)) / 2] \times 10 = N$ association reactions can be generated if one composition of one reactant side corresponds to 10 possible products [20]. If M equals 100, this makes 50,500 possible reactions. However, the conditions in the ISM [40] and CSE [41] are such that only barrierless reactions are likely [20,39].

It is currently not possible to do accurate quantum chemistry calculations for all hypothetical reactions in the ISM and CSE that are relevant to prebiotic chemistry [20]. Therefore, estimating reaction energy barriers among ISM and CSE molecules with a feasible approach is critical for figuring out potential chemical evolution routes in the ISM and CSE.

1.3 Quantum Chemical Methods for Estimating Reaction Energy Barrier

There are various quantum chemistry methods available for optimizing molecular geometries, locating TS, calculating molecular energies, and calculating reaction energy barriers. These methods are based on quantum mechanics [42]. Common classes of methods are *ab initio* methods, density functional theory (DFT) methods, and semi-empirical methods (SEMs) [42–44].

The Hamiltonian operator ($\hat{H}$) of the Time-Independent Schrödinger equation

$$\hat{H}\psi = E\psi \tag{1}$$

for N electrons moving in the potential created by M nuclei corresponds to the sum of kinetic energies and potential energies of all the particles in the system and is described as [44a–44e]

$$\hat{H} = - \sum_{i=1}^N \frac{1}{2} \nabla_i^2 - \sum_{A=1}^M \frac{1}{2M_A} \nabla_A^2 - \sum_{i=1}^N \sum_{A=1}^M \frac{Z_A}{r_{iA}} + \sum_{i=1}^N \sum_{j>i}^N \frac{1}{r_{ij}} + \sum_{A=1}^M \sum_{B>A}^M \frac{Z_A Z_B}{R_{AB}} \quad (2)$$

where the first term is the operator corresponding to kinetic energy of N electrons (notated as $\hat{T}_e$ in Equation (3) with “ T ” representing “kinetic energy” and subscript “ e ” representing “electrons”), the second term is the operator corresponding to kinetic energy of M nuclei (notated as $\hat{T}_n$ in Equation (3) with “ T ” representing “kinetic energy” and the subscript “ n ” representing “nuclei”), the third term is the operator corresponding to Coulomb attraction of electrons-nuclei (notated as $\hat{V}_{en}$ in Equation (3) with “ V ” representing “potential energy”, and the subscript “ en ” representing attraction between electrons and nuclei), the fourth term is the operator corresponding to repulsion between electrons (notated as $\hat{V}_{ee}$ in Equation (3) with “ V ” representing “potential energy”, and the subscript “ ee ” representing repulsion between electrons), and the last term is the operator corresponding to repulsion between nuclei (notated as $\hat{V}_{nn}$ in Equation (3) with “ V ” representing “potential energy”, and the subscript “ nn ” representing repulsion between nuclei). Hence, Equation (2) can be written as

$$\hat{H} = \hat{T}_e + \hat{T}_n + \hat{V}_{en} + \hat{V}_{ee} + \hat{V}_{nn} \quad (3)$$

Nowadays, there are many methods and computation programs available to obtain approximate solutions to the Schrödinger equation (see Equations (1) to (3)) for the ground

state. Solving Schrödinger equation and obtaining the ground electronic state energy for many different sets of nuclear positions ($X_1, X_2, \dots, X_M$) generates points of the potential energy surface (PES), $U(X_1, X_2, \dots, X_M)$ (X_j is the shorthand for the three cartesian coordinates of nucleus j). This procedure is justified by the Born-Oppenheimer approximation which says, essentially, that the motion of electrons and nuclei occur on very different time scales. Chemical reactions are commonly studied by finding and calculating the energies at a few critical points of the PES associated with the nuclei in that reaction. The minima of the PES correspond to stable species during the process of that reaction, i.e., various conformers of the reactants, products, and intermediates of reaction. The saddle points of the PES correspond to TS. Every pair of minima of the PES can be connected by an intrinsic reaction coordinate (IRC) that goes through one TS (and every TS leads to two minima of the PES). Once the critical points of a PES and their energies are known, one can calculate the energy of reaction: $\Delta U_{\text{rxn}} = U(X_{\text{products}}) - U(X_{\text{reactants}})$ and the energy barriers separating pairs of connected minima 'a' and 'b': $E_{ab} = U(X_{ab}) - U(X_a)$, where X_{products} , $X_{\text{reactants}}$, X_{ab} , and X_a represent the set of nuclear positions for the products, the reactants, the transition state 'ab', and the minimum (or the molecular species) 'a'.

Ab initio methods find approximate solutions to the many-electron Schrödinger equation [42,43,45]. This is usually done by applying the variational theorem to optimize a trial wavefunction and calculating the energy from the known Hamiltonian and the approximate wavefunction [42,43,45]. Common *ab initio* methods for estimating reaction energy barriers include Hartree-Fock (HF), Møller–Plesset Perturbation theory (MP_n), Coupled Cluster theory (CC), and Quadratic Configuration Interaction (QCI).

The main differences between Kohn-Sham Density Functional Theory (KS-DFT) and *ab initio* methods are that: (1) in KS-DFT, a model Hamiltonian is used to set up a Schrödinger equation instead of using the true Hamiltonian; (2) electron correlation is treated approximately with a careful choice of a fictitious exchange-correlation potential $v_{xc}(r)$, where r denotes a position in 3D space. The great computational advantage of KS-DFT comes from the choice of $v_{xc}(r)$ which makes it unnecessary to calculate a wavefunction [46–48]. Instead, KS-DFT only requires finding $N/2$ occupied orbitals by finding $N/2$ solutions to a single one-electron Schrödinger equation which is far easier than solving the N -electron Schrödinger equation. Remarkably, it has been shown that there exists a $v_{xc}(r)$ that produces orbitals which, when squared and summed over, produce exact (nonrelativistic) electron densities [47]. And also, that there exists a functional E_{xc} of the electron density that produces exact energies. Unfortunately, v_{xc} and E_{xc} are not known exactly and there are many (roughly 200) approximate versions of KS-DFT, some of them free of any empirical parameter, some of them heavily parametrized.

Semi-empirical methods, such as PM7, proceed like *ab initio* methods with three important approximations. First, these methods neglect electronic correlation, except indirectly by fine-tuning some empirical parameters. Second, many molecular integrals that naturally arise in *ab initio* methods are neglected, they are simply set to zero [43,49,50]. Third, the elements of the Hamiltonian matrix, which are loosely related to atomic orbital energies and energy couplings, are calculated using a mix of empirical data and ‘hard calculations’ (the empirical data are mostly spectroscopic data).

In this research project, a semi-empirical method, PM7 [51–53], is utilized for optimizing molecular geometry and calculating its energy based on that geometry. We used PM7 to save a lot of computational cost when a costlier way, DFT, is no longer feasible for thousands of, or even millions of molecules. Even though SEMs are faster than standard HF or DFT approaches [50], they are less reliable. For approximating reaction energy barriers, PM7-TS [54] outputs a MAE of 22.5 kcal/mol [20,55]. For a data set of 206 reactions, 178 DFT methods can estimate reaction energy barriers with accuracies of less than 10 kcal/mol [5]. One of the best of these 178 DFT methods, ω B97XD, reproduced known energy barriers with a mean absolute error of 2.3 kcal/mol [5]. This is the method we used.

Of the three categories of methods, SEM is the fastest and least accurate, with errors of 25 kcal/mol or more [42,43,56]. Some *ab initio* methods give much better accuracy (e.g., 1 kcal/mol for estimating reaction energy barriers) [8,9] but are computationally very demanding [57]. Some DFT methods can give barriers with mean absolute errors of 2 to 5 kcal/mol, but require computing time on the order of hours per reaction which would make the study of reactions in the ISM very challenging. Furthermore, the calculation of TS and barriers by DFT, or by any other quantum chemical method, is prone to failure and requires frequent human intervention. This would be problematic whenever the number of reactions exceeds a few hundred.

Theoretically, ISM and CSE molecules could combine with each other to produce thousands of possible products due to a variety of compositions of elements, leading to

thousands of, even millions of, possible reactions. The computational cost of studying that order of magnitude of reactions with sophisticated quantum chemistry methods is tremendous. The reaction energy barrier will be acquired only after the TS is found, which is always a tough task and may require different algorithms for different kinds of reactions. Software like Gaussian allows one to search for TS. However, these searches often fail. In order to show that a critical point of the PES is a TS, normal modes are calculated: if one, and only one, of them is associated to an imaginary frequency, the point is a TS. There is another difficulty with finding TS. If one searches for the TS that connects minima A and B, and finds a TS, that TS may not actually connect A and B. To solve that problem, a calculation of the intrinsic reaction coordinate (IRC) [58,59] starting from the TS is done. An IRC always leads to two minima. If the minima are A and B, we have the correct TS for the process $A \rightarrow B$. Otherwise, the search for TS may have to be repeated with a different algorithm or different geometry guess. The bonds formation and breaking near transition states of a reaction need to be well known for approximating energies very well, which is a hard problem for molecular electronic structure calculations [20].

To solve the above issues, the goal of the first part of this research project is to establish a machine learning model for fast (less than one second) and moderately accurate (10 kcal/mol or better) estimates of energy barriers [20].

1.4 Introduction to Machine Learning

Machine learning (ML), a subfield of artificial intelligence (AI), has been successfully applied in different fields of research including chemistry, energy economics, neuroscience, and health care [60–66].

ML has shown its potential for handling large-scale chemical problems [67,68]. The two most common categories of problems that ML algorithms can tackle are regression and classification. The most natural formulation of energy barrier prediction is a regression. Suppose that we know, for many cases $j = 1$ to N , a set of input variables ($x_{j1}, x_{j2}, \dots, x_{jD}$) and one output energy barrier E_j . The integer D is the dimensionality of the representation. The D input variables for the j^{th} data are conveniently denoted by a vector x_j . A ML model can be trained to learn the relation between inputs x_j and outputs E_j . After training, the ML model is a fit function $f(x)$ that can act on any input x such that $f(x_j) = E_j + d_j$, where the d_j 's are small errors (small in a root-mean square sense). Different forms of the function give rise to different ML models, and different techniques may be used to train ML models. These are explained in section 2.2.

1.4.1 Elements of a Machine Learning Model

The first key to building a ML model is to have a large database of reactions with accurately known reaction energy barriers. Since it learns from data [72], the number and the quality of data is essential for creating a good ML model.

The second key is to develop appropriate features of reactions. Ideally, features should encode chemical reactions in a way that is computationally efficient and captures all

important aspects for barrier prediction. In our problem, it is essential that the computation of features of a reaction be many orders of magnitude faster than a DFT determination of TS and barrier. It is also good to have relatively few features, and features that are not too strongly correlated to one another. The ideal number of features is highly problem dependent. It can be less than 10 for some very simple problems, or millions in very complex problems. The design and selection of features often requires (as it did in our study) many tests and some form of optimization.

In addition to the above keys, ML requires computationally efficient ways to process data. A common difficulty in ML, which did arise in our work, is to recognize and deal with corrupt data. In our study, a few unphysical geometries were generated by the computer code that translates Simplified Molecular Input Line Entry System (SMILES) representations of molecules to 3D molecular structures. The solution we applied involved standardization: for every feature ' y ', the mean ' m ' and standard deviation ' s ' of y_j values are calculated, and the feature is redefined as $y_{std} = (y-m)/s$. After standardization, any instance where the absolute value of one of the features is larger than a tolerance T is rejected. For our data, a very conservative tolerance, $T = 18$, was adequate to reject corrupt data [20].

The last key to a successful ML is to seek the most accurate ML model for solving a specific problem. We used these statistics to assess the quality of different ML models: the root-mean-square error (RMSE), the mean absolute error (MAE), the mean signed error (MSE), the fraction of variance unexplained (FVU) by the model, the coefficient of

determination R^2 , and the adjusted coefficient of determination R_{adj}^2 . We give special importance to R_{adj}^2 in comparing models and the MAE to express typical expected errors in future use of a model. We point out that a proper assessment of a model must be based on two sets of data, a training set used for optimizing the parameters of the model (e.g., the coefficients in a linear regression), and an independent validation set that has no data in common with the training set, for assessing the predictive performance with RMSE, MAE, etc. Moreover, a model validation procedure is of great significance for avoiding overfitting or underfitting issues of the chosen model.

1.5 Interstellar Medium (ISM) and Circumstellar Envelopes (CSE)

The major application of our study is the chemical reactivity and the exploration of potential chemical evolution in the ISM and CSE [20,39]. The interstellar medium only surrounds a small portion of the Galaxy's total mass, however, the chemical, biochemical, and biophysical evolution occurring in those regions is essential to our understanding of space, the possible origins of life, and the galactic ecosystem [40].

Some molecules and dust in the ISM (roughly 80% of the total mass) were initially generated in the CSE and blown from the CSE into the ISM [41]. The rapid and advanced growth of telescope and satellite technology in the past 60-70 years has enabled more and more interstellar molecules to be identified, which furthers our understanding of space and brings renewed interest by the research community [73a].

In 1994, there were 107 gas-phase interstellar molecules reported to be present, five of which were uncertain [73b]. In 2004, there were 127 interstellar molecular species reported

to exist in space [73a]. As of 2023, more than 250 molecular species have been detected or are suspected to be present in the ISM and CSE [32,36–39]. The chemistry of the ISM or CSE must be governed to a large extent by barrierless reactions because temperatures are very low, roughly between 15 and 100 K [40,41]. However, observations and investigations of molecular clouds of ISM achieved by scientists imply that interstellar shocks may exist and lead to the generation of radiation [73a,74], which could increase the maximum temperature to 150 K for a specific area. Even with temperatures of 15 ~ 150 K, the gas-phase chemical reactions between two interstellar molecules can hardly proceed except those with zero barriers and no endothermicity [20,39,73a]. The reaction mechanisms in the ISM and CSE are not well-known. Gas phase collisions, reactions at the surface of carbonaceous or icy grains, and reactions initiated by absorption of cosmic rays are all possible [39,73a,73c,74]. It seems plausible, however, that barrierless reactions play a big role. The reactions can hardly occur if reactants need extra energy (especially a large energy) to produce a new compound [20]. We focus on association reactions among two reactants since the collision situation is rare in the ISM and CSE and it is unlikely to have one collision caused by more than two molecular species at once [73a,73b]. In the ISM, stabilization of the products of association reactions is an essential step. We did not attempt to model this. It has been suggested that newly formed products in the ISM could be stabilized by emitting a photon for a reaction in the gas phase, or by giving off energy to a dust grain where the reaction occurs.

1.6 Outline of the Dissertation

The work described in this dissertation is divided into three parts: development of a ML model for predicting energy barriers; quantum chemical study of possible reactions in the ISM or CSE, $\text{CH}_2 + \text{HCNO}/\text{HNCO}/\text{CH}_3\text{C}(=\text{O})\text{NH}_2$; calculation of equilibrium geometries and rotational constants of conformers of molecules suspected to be present in the ISM. The overall process can be depicted in Figure 1-1, which has been published as a graphical abstract in the paper:

- **Ji, H.;** Rágyanszki, A.; Fournier, R. A. Machine Learning Estimation of Reaction Energy Barriers. *Comput. Theor. Chem.* **2023**, 1229, 114332. <https://doi.org/10.1016/j.comptc.2023.114332>.

1.6.1 Development of a Machine Learning Model

The first major part of our research project is to propose a fast way of estimating reaction energy barriers. Such a fast way can be achieved by establishing ML approaches. Features can represent the information relevant for estimating energy barriers. In order to achieve

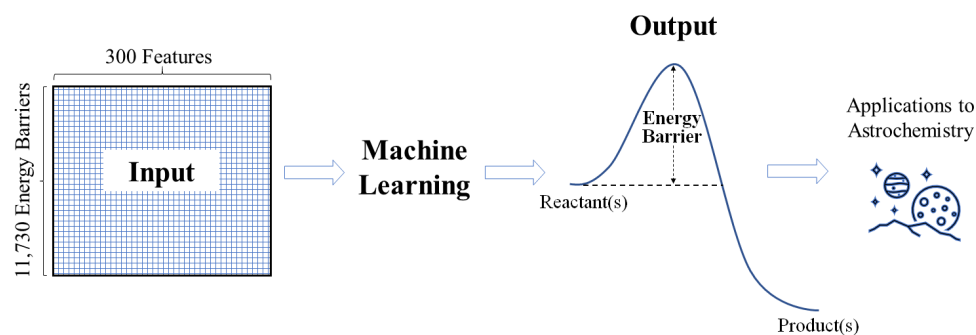


Figure 1-1. Workflow of overall process presented in this dissertation with pointing out applications to astrochemistry, especially to ISM and CSE.

good accuracy, we tested many different kinds of features combined in many different ways.

We tested five regression methods for energy barrier estimation [20]: Ordinary Least Squares (OLS) [75,76], Ridge Regression (RR) [77,78], Weighted Least Squares (WLS) [79,80], Kernel Ridge Regression (KRR) [81–84], and Feed Forward Neural Network (FFNN) [85–88]. After the ML method with the best performance is found, different features sets and many types of approaches to calculating features are applied [20].

1.6.2 Computational Search for Prebiotic Chemistry in the ISM and CSE

After the best ML model is developed, the next major part is to create feasible, efficient, and biochemistry-important screening procedures. The screening procedures presented in Chapter 5 have been completed [39] and can be depicted in Figure 1-2. The new screening proposal and corresponding plans related to the energies of conformers of critical molecules will be discussed in Chapter 7.

As shown in Figure 1-2, a dataset of 136,081 reactions was created by making combinations of reactants A and B, where A and B are chosen among the molecules known to exist in the ISM or CSE [39]. The barriers to reactions were predicted using our best ML method [20] for those 136,081 reactions and every reaction with an energy barrier larger than 6 kcal/mol was rejected [39]. After this step, approximately 2% of the reactions remained. Next, we applied a set of rules to select reactions that appear most relevant for prebiotic chemistry.

The few remaining reactions were then examined one by one, and we made educated guesses as to which are most important. Those were then studied with accurate quantum chemistry (QC) techniques to establish reaction pathways, including transition states and energy barriers, and to compute equilibrium geometries and rotational constants of conformers of the products of reactions. The details of those calculations are discussed in chapters 2, 5, and 6.

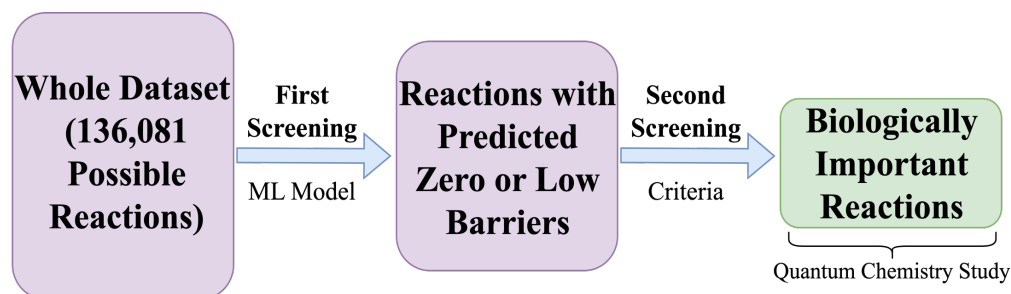


Figure 1-2. Workflow for acquiring the most possible and biologically meaningful reactions in the ISM and CSE.

Chapter 2 Computational Techniques

Here we describe the quantum chemistry and machine learning methods and software that we used. The content of this chapter closely follows our article on energy barrier estimation (Hongchen Ji *et al.* [20]).

2.1 Quantum Chemistry and the Gaussian Software

We used Gaussian16 software [89] to perform all quantum chemistry techniques in our research project, as well as GaussView [90] and Avogadro [91,92] as interfaces. All Gaussian calculations were carried out on Digital Research Alliance of Canada computers.

2.1.1 Molecular Geometry Optimization

For every reactant, reactant complex (RC), intermediate (IM), transition state (TS), and product (P) on the PES of one reaction, optimization of geometry is required. Different optimization algorithms can be utilized by specifying a keyword in the Gaussian input file [89,93a,94]. Other information that is supposed to be contained in the Gaussian input file includes an assigned name for a molecule, electric charge and the spin multiplicity of that molecule, and element symbol and cartesian coordinates of an initial approximate geometry for each atom in that molecule [95]. One simple optimization example (geometry optimization of ethanol at ω B97XD/def2-TZVP level of theory [21–27]) of the Gaussian input file is shown in Figure 2-1. The input geometry of ethanol used here and used in the next sections was obtained from the PubChem Compound Summary [93b].

```

# wB97XD/Def2TZVP Opt

Ethanol

0 1
O          -1.17120000    0.29970000    0.00000000
C          -0.04630000   -0.56650000    0.00000000
C           1.21750000    0.26680000    0.00000000
H          -0.09580000   -1.21200000    0.88190000
H          -0.09520000   -1.19380000   -0.89460000
H           2.10500000   -0.37200000   -0.01770000
H           1.24260000    0.93070000   -0.87040000
H           1.26160000    0.90520000    0.88860000
H          -1.12910000    0.83640000    0.80990000

```

Figure 2-1. Gaussian input file of one simple optimization example (geometry optimization of ethanol at ω B97XD/def2-TZVP level of theory).

2.1.2 Molecular Frequency and Energy Calculation

In our work, molecular geometry optimization was always followed by a calculation of normal mode frequencies achieved with the keyword Freq in the Gaussian input [96,97]. The Freq keyword also triggers the calculation of the zero-point energy (ZPE) and ZPE corrected energy [96,97]. Imaginary frequencies show up as negative numbers in the Gaussian output. The number ‘n’ of imaginary frequencies is used to characterize stationary points of the PES: $n = 0$ indicates a minimum, $n = 1$ a transition state, and $n > 1$ a higher-order critical point of no interest to us.

One simple example of the Gaussian input file of frequency and energy calculation (of ethanol at ω B97XD/def2-TZVP [21–27] level of theory) is shown in Figure 2-2.

```
# wB97XD/Def2TZVP Freq

Ethanol

0 1
O          -1.17120000    0.29970000    0.00000000
C          -0.04630000   -0.56650000    0.00000000
C           1.21750000    0.26680000    0.00000000
H          -0.09580000   -1.21200000    0.88190000
H          -0.09520000   -1.19380000   -0.89460000
H           2.10500000   -0.37200000   -0.01770000
H           1.24260000    0.93070000   -0.87040000
H           1.26160000    0.90520000    0.88860000
H          -1.12910000    0.83640000    0.80990000
```

Figure 2-2. Gaussian input file of one simple example of frequency and energy calculation (of ethanol at ω B97XD/def2-TZVP level of theory).

2.1.3 Transition States

The equilibrium geometries of molecules are minima of the PES. Finding the minima of the PES, a function of many variables, is an easy mathematical problem and the numerical methods to do it are fast and robust. In contrast, a TS is a saddle point, i.e., a minimum along all coordinates except one (call it ' q ') and a maximum along q . TS are difficult to locate because q is not known a priori. Numerical methods that search for TS go through many steps and often fail. Some TS searches end up at a minimum, others at a higher-order critical point, and others find a TS that is different from the one connecting the reactants and products.

There are three common keywords used for locating TS with Gaussian16: QST2, QST3, and TS [93a,94,96,98,99]. The former two approaches are both based on the Synchronous Transit-Guided Quasi-Newton (STQN) method [99,100]. Preparing input files for QST2 and QST3 is not simple since the atom order in the product(s) should be the same as in reactant(s) (for QST2), and the atom order in an initial guessing TS should be also the same

as in reactant(s) (for QST3) [93a]. Moreover, they fail often for complex TS. However, one can give the testing priority to QST2 or QST3 if there is only one small change between two conformers of a molecule, e.g., only one angle between two conformers is different. The difference between QST2 and QST3 is the input information needed: QST2 requires molecular information of reactant(s) and product(s) while QST3 requires molecular information of reactant(s), product(s), and an initial TS guess [99]. For most cases in our work, the keyword TS using the Berny algorithm [93a, 100–106] works well. Along with this approach, the computation of force constants is sometimes needed for better TS optimization, which can be achieved with the keyword calcall, specifying that, at every point, the force constants will be calculated [93a,107]. One simple example of a Gaussian input file for TS optimization with the Berny algorithm and force constants computation for a TS between *anti* and *gauche* conformers of ethanol [108] at ω B97XD/def2-TZVP [21–27] level of theory is shown in Figure 2-3.

```
# wB97XD/Def2TZVP Opt=(ts,calcall) geom=connectivity
TS between Ethanol conformers

0 1
O          -1.17120000    0.29970000    0.00000000
C          -0.04630000   -0.56650000    0.00000000
C          1.21750000    0.26680000    0.00000000
H          -0.09580000   -1.21200000    0.88190000
H          -0.09520000   -1.19380000   -0.89460000
H          2.10500000   -0.37200000   -0.01770000
H          1.24260000    0.93070000   -0.87040000
H          1.26160000    0.90520000    0.88860000
H          -1.95939917    0.31793564    0.56935651

1 2 1.0 9 1.0
2 3 1.0 4 1.0 5 1.0
3 6 1.0 7 1.0 8 1.0
4
5
6
7
8
9
```

Figure 2-3. Gaussian input file of one simple example of TS optimization (for a TS between *anti* and *gauche* conformers of ethanol at ω B97XD/def2-TZVP level of theory).

2.1.4 Intrinsic Reaction Coordinate (IRC) Calculation

TS located with the above procedure should be characterized by normal mode analysis. A TS must have only one imaginary frequency. The next step is to find which two minima are connected by the transition states. This is done by doing a small displacement of the nuclei along the imaginary frequency mode, in the positive and negative directions, and following the intrinsic reaction coordinate (IRC) [58,59] to the two minima connected by the TS [109]. This is done with the IRC keyword in Gaussian. Generating the IRC is much more costly in computing time than a molecular geometry optimization and is typically two to eight times more costly than a TS optimization. The IRC does not always end at the expected, or desired, reactants and products. When it does not, additional TS, and maybe intermediates of reactions, must be discovered to map out the reaction path. The entire process for finding the relevant transition state(s) of a reaction can take several days.

```
# IRC=(calcall) wB97XD/Def2TZVP geom=connectivity
IRC from a TS between Ethanol conformers

0 1
O          -1.17120000    0.29970000    0.00000000
C          -0.04630000   -0.56650000    0.00000000
C          1.21750000    0.26680000    0.00000000
H          -0.09580000   -1.21200000    0.88190000
H          -0.09520000   -1.19380000   -0.89460000
H          2.10500000   -0.37200000   -0.01770000
H          1.24260000    0.93070000   -0.87040000
H          1.26160000    0.90520000    0.88860000
H          -1.95939917    0.31793564    0.56935651

1 2 1.0 9 1.0
2 3 1.0 4 1.0 5 1.0
3 6 1.0 7 1.0 8 1.0
4
5
6
7
8
9
```

Figure 2-4. Gaussian input file of one simple example of IRC calculation (starting from an assumed TS between *anti* and *gauche* conformers of ethanol at ω B97XD/def2-TZVP level of theory).

One simple example of the Gaussian input file of IRC calculation starting from an assumed TS between *anti* and *gauche* conformers of ethanol at ω B97XD/def2-TZVP [21–27] level of theory is shown in Figure 2-4.

2.1.5 Molecular Single-Point Energy Calculation

Single-point energy (SPE) calculations are normally done when a more accurate and more expensive method is used, such as CCSD(T). When studying potential energy surface or one reaction with quantum chemistry, it is common to choose a DFT method for molecular geometry optimization, TS locating, and reaction pathway following with IRC, and then employ a sophisticated Coupled Cluster (CC) method for SPE calculation, such as CCSD or CCSD(T). As mentioned earlier, the computational cost is enormous if the whole process (molecular geometry optimization, TS locating, and IRC calculation) is done with CCSD(T), it could take many months with 16 processors even for a simple reaction. Applying a combination of DFT methods with CC methods to investigate the reaction mechanism is much cheaper and gives nearly the same accuracy as doing everything by CCSD(T). SPE calculations can be easily achieved with the keyword SP in the Gaussian input file [110]. One simple example of a Gaussian input file of SPE calculation based on

```
# CCSD-T/Aug-cc-pVQZ SP
Ethanol
0 1
O      -1.17120000    0.29970000    0.00000000
C      -0.04630000   -0.56650000    0.00000000
C       1.21750000    0.26680000    0.00000000
H      -0.09580000   -1.21200000    0.88190000
H      -0.09520000   -1.19380000   -0.89460000
H       2.10500000   -0.37200000   -0.01770000
H       1.24260000    0.93070000   -0.87040000
H       1.26160000    0.90520000    0.88860000
H      -1.12910000    0.83640000    0.80990000
```

Figure 2-5. Gaussian input file of a calculation based on the geometry of ethanol at CCSD(T)/Aug-cc-pVQZ level of theory.

the geometry of ethanol at CCSD(T)/Aug-cc-pVQZ [15–19,111–113] level of theory is shown in Figure 2-5.

2.1.6 Potential Energy Surface Scan

In this research project, several reactions mechanisms are studied on both singlet and triplet potential energy surfaces. A related work of finding the triplet-to-singlet surface crossing is proposed [39] because it can discover additional reaction pathways leading to ground state products.

To figure out this crossing, one feasible way is to make use of a potential energy surface (PES) scan of singlet and triplet reactant complexes based on continuously changing geometries [39]. The keyword “scan” used in the Gaussian input file can allow us to perform a lot of SPE calculations based on different combinations of bond lengths and angles with only input once [114]. The bond lengths or angle change for each step and the total number of changing steps can be easily pointed out in that Gaussian input file. One

```
# wB97XD/Def2TZVP scan geom=connectivity
CH2 CH2
0 1
C
H          1          B1
H          1          B2    2          A1

B1          1.08024155 10 0.002953999999999999
B2          1.08024155 10 0.002953999999999999
A1          134.29612176 10 -3.219419

1 2 1.0 3 1.0
2
3
```

Figure 2-6. Gaussian input file of one simple example of PES scan of CH₂ at ω B97XD/def2-TZVP level of theory.

simple example of the Gaussian input file of the PES scan of CH₂ at ω B97XD/def2-TZVP [21–27] level of theory is shown in Figure 2-6.

2.2 Machine Learning (ML) Techniques

2.2.1 Methodology of Machine Learning

Estimation of reaction energy barriers (E) with machine learning can be cast in the form of a regression problem. Assume we have a set of M chemical reactions labelled with ‘ m ’, N features (or variables) $x_1, x_2, \dots, x_n, \dots, x_N$ that vary from one reaction to the next reaction, and M known energy barriers $E_1, E_2, \dots, E_M$. We are seeking a function $f(x_1, x_2, \dots, x_N)$ that approximates reaction energy barriers.

In the Ordinary Least Squares (OLS) linear regression method, the function f is a linear combination of the features: $f = c_1x_1 + c_2x_2 + \dots + c_Nx_N$. If the features for reaction m are denoted $X_{m1}, X_{m2}, \dots, X_{mN}$, we seek a function that minimizes a loss function such as

$$L_1 = \sum_m |f(X_{m1}, X_{m2}, \dots, X_{mN}) - E_m| \quad (4)$$

or

$$L_2 = \sum_m (f(X_{m1}, X_{m2}, \dots, X_{mN}) - E_m)^2 \quad (5)$$

In the OLS method, L_2 is minimized by varying the coefficients c_n . Other ML methods use more complicated forms for f , and sometimes use different loss functions. In principle, the prediction of reaction energy barrier simply amounts to a regression (fit). In practice, however, it is a tough and complex problem with four or five key requirements.

As mentioned in section 1.4.1, a large dataset of accurately known reaction energy barriers is needed first. Second, we need to create or develop useful features, preferably chemically meaningful ones, that must be simple to calculate and interpret and should correlate with reaction energy barriers or describe significant changes between reactants and products, such as significant bond changes. Third, we need to figure out computationally efficient ways for handling all input data, including outliers and corrupt data. Fourth, we need to develop robust numerical techniques for optimizing functions f that are flexible enough to fit data accurately. As for additional problems that arise when fine tuning ML models, they will be discussed in section 3.4.

2.2.2 Data Processing of Machine Learning

When pre-processing data for ML, we handle outliers as follows. We standardize every feature by subtracting the mean and then dividing by the standard deviation of that feature. The i^{th} data will be considered as an outlier and will be rejected if any of the standardized features $\bar{X}_{ij}$ is outside of the range $[-C, +C]$. After eliminating outliers, we will recompute the mean and standard deviation of each feature, standardize current features again, and apply the outlier criterion once more. This will result in the removal of some more data. We applied this procedure with $C = 18$, which led to the removal of only one to three percent of data.

2.2.3 Ordinary Least Squares (OLS)

The OLS model is the simplest among ML techniques [75,76] and is widely used. If we have a training set with M reaction energy barriers E_i and M features vectors x_i , each of

length N , a M -by- N design matrix $\mathbf{X}$ is established: X_{ij} is the j^{th} feature of the i^{th} instance of the data ($i = 1, \dots, M, j = 1, \dots, N$). As usual for the OLS technique, we add to the set of features a constant feature (which is always one), which allows fits with nonzero intercepts in OLS. The predicted reaction energy barriers are

$$f_i = \sum_j X_{ij} \beta_j \quad (6)$$

and the parameters of the model, i.e., the regression coefficients β_j , are obtained as described

$$\boldsymbol{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{E} \quad (7)$$

In equation (7) and the rest of this dissertation, bold lowercase or Greek letters represent vectors and bold upper letters represent matrices. A superscript T denotes the transpose of a matrix, and a superscript -1 denotes the inverse of a matrix. This gives β_j s that minimize L_2 . Our OLS model, and RR, WLS and KRR models, were implemented using Python with PyCharm [115].

2.2.4 Ridge Regression (RR)

The Ridge Regression [77,78] model is similar to OLS but with a non-zero regularization parameter, λ .

$$\boldsymbol{\beta} = (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^T \mathbf{E} \quad (8)$$

With β_j s obtained from Equation (8), f minimizes a modified form of the loss L_2

$$\min \sum_{i=1}^M (f_i - E_i)^2 + \lambda \sum_{j=1}^N \beta_j^2 \quad (9)$$

which helps prevent the overfitting problem.

2.2.5 Weighted Least Squares (WLS)

The WLS model [79,80] is another variant of OLS but with weights assigned to data. The working equation for the coefficients is described as

$$\boldsymbol{\beta} = (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W} \mathbf{E} \quad (10)$$

The matrix $\mathbf{W}$ in Equation (10) is diagonal and contains weights W_{ii} that are chosen and assigned to emphasize some data, with greater W_{ii} for data deemed more accurate or more critical. As mentioned earlier, only those reactions with zero or small energy barriers matter for chemistry in the ISM and CSE. Thus, we chose to emphasize the data (reactions) with small barriers. After testing different forms of weight matrices, eventually we set

$$W_{ii} = \frac{10}{\sqrt{E_i + 10}} \quad (11)$$

due to its good performance. In Equation (11), E_i is in kcal/mol. The WLS technique solves the following minimization problem [79,80]

$$\min \sum_{i=1}^M W_{ii} \left| E_i - \sum_{j=1}^N X_{ij} \beta_j \right|^2 \quad (12)$$

2.2.6 Feed-Forward Neural Network (FFNN)

FFNN accomplishes a nonlinear regression. We tested FFNN with different numbers of hidden layers (e.g., 2, 3, 4, etc.) with different numbers of neurons for each hidden layer (e.g., 20, 30, 40, etc.) and with Bayesian Regularization backpropagation algorithm [85–88] by MATLAB Online [116]. The best FFNN with our features set 1 is the one with three hidden layers (FFNN-3HL) with 20, 10, 5 neurons respectively. One flowchart example of this FFNN-3HL is shown in Figure 2-7, which is produced by MATLAB Online [116].

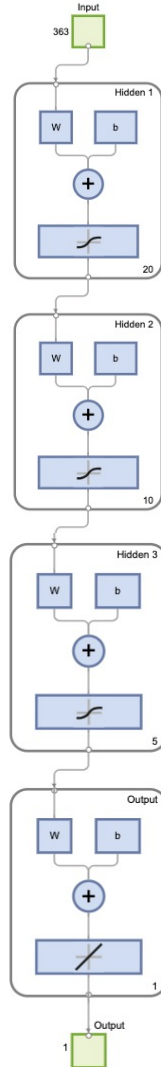


Figure 2-7. Flowchart of our best FFNN-3HL (produced by MATLAB Online).

2.2.7 Kernel Ridge Regression (KRR)

Unlike OLS, RR, and WLS, but like FFNN, Kernel Ridge Regression [81–84] does a nonlinear regression. It has been successfully used for predicting atomization energies [83] and orbital energies [84]. A predicted barrier “ f ” in OLS, RR and WLS can be written as a scalar product of two vectors, $f = \boldsymbol{\beta}^T \mathbf{x}'$, where $\mathbf{x}'$ is a features vector not in the training set. In the RR method, $\boldsymbol{\beta} = (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^T \mathbf{E}$ and $f = \mathbf{x}'^T ((\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^T \mathbf{E})$ where $\mathbf{E}$ is the M -long vector of energy barriers from the training set, and $\mathbf{X}$ is the matrix of features in the training set. It can be shown [81–84] that f can be

$$f = \boldsymbol{\beta}^T \mathbf{x}' = (\mathbf{E}^T (\mathbf{X} \mathbf{X}^T + \lambda \mathbf{I})^{-1} \mathbf{X}) \mathbf{x}' \quad (13)$$

After defining $K_{ij} = \mathbf{x}_i^T \mathbf{x}_j$ and $k_i = \mathbf{x}_i^T \mathbf{x}'$ ($i = 1, \dots, M$) and changing the symbol λ to α , Equation (13) becomes

$$f = \mathbf{E}^T (\mathbf{K} + \alpha \mathbf{I})^{-1} \mathbf{k} = \boldsymbol{\beta}^T \mathbf{k} \quad (14)$$

The K_{ij} s and k_i s are dot products between pairs of features vectors. These dot products are large when two sets of features are similar, and are small when they are dissimilar. In KRR, the dot products are replaced by adjustable “similarity functions”, or “kernels”. This amounts to an *implicit* nonlinear transformation of the features, which changes RR into KRR, a nonlinear regression to the data. The different choices of the kernel may allow more accurate fits. Two choices of the kernel that are commonly used are the Laplacian kernel and the Gaussian kernel [81–84]

$$k_L(\mathbf{x}, \mathbf{x}') = \exp(-\gamma \sum_{j=1}^N |x_j - x'_j|) \quad (15)$$

$$k_G(\mathbf{x}, \mathbf{x}') = \exp(-\gamma \sum_{j=1}^N |x_j - x'_j|^2) \quad (16)$$

where γ has different expressions of kernel width for Laplacian kernel and Gaussian kernel (it is a hyperparameter in KRR, and it is the reciprocal of kernel width σ for Laplacian kernel (i.e. $1/\sigma$), while it is in the form of $1/(2\sigma^2)$ for Gaussian kernel [84]), x_j is the j^{th} feature of the features vector $\mathbf{x}$, and x'_j is the j^{th} feature of the features vector $\mathbf{x}'$.

The hyperparameter γ can be varied to achieve the best fit. This will be discussed in section 3.4. KRR achieves this minimization [84]

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^M (f(\mathbf{x}_i) - E_i)^2 + \alpha \boldsymbol{\beta}^T \mathbf{K} \boldsymbol{\beta} \quad (17)$$

where α is the regularization hyperparameter which, like λ in RR, helps prevent the overfitting problem. The kernel $\mathbf{K}$ in Equation (17) has matrix elements given by

$$K_{ij} = k_L(\mathbf{x}_i, \mathbf{x}_j) \quad \text{or} \quad k_G(\mathbf{x}_i, \mathbf{x}_j) \quad (18)$$

where $\mathbf{x}_i$ and $\mathbf{x}_j$ are two features vectors from the training set. The coefficients β_j are given by

$$\boldsymbol{\beta} = (\mathbf{K} + \alpha \mathbf{I})^{-1} \mathbf{E} \quad (19)$$

For a given new vector $\mathbf{x}$, a prediction is made with

$$f(\mathbf{x}) = \sum_{j=1}^M \beta_j k_L(\mathbf{x}, \mathbf{x}_j) \quad \text{or} \quad \sum_{j=1}^M \beta_j k_G(\mathbf{x}, \mathbf{x}_j) \quad (20)$$

In the KRR model, there are M parameters β_j , one for each data (each reaction). These β_j s are interpreted as weights that are assigned automatically to data in the fitting process. The performance of Laplacian kernel and Gaussian kernel for our research project and the optimization of the hyperparameters α and γ will be discussed in Chapter 3.

2.3 Applications of ML to Chemistry

ML has shown great performance in predicting some molecular properties [20,60–63,83], such as atomization energies [60,83] and intrinsic bond energies [61], in chemical synthesis [69], in drug discovery [70,71], and in investigating chemical compound and reaction space [304–306].

Predictions of chemical properties can be obtained much faster with ML than with quantum chemical calculations. Meyer *et al.* [307] reported an application of ML combined with the utilization of volcano plots for fast screening of the most potential and attractive catalyst species for C-C cross-coupling reactions. This work used geometry estimates of catalysts that were quickly obtained for their ML model and predicted the energy of oxidative addition reactions with MAE of 2.6–3.1 kcal/mol using different features [307]. It avoided time-consuming optimization of molecular geometry and calculation of molecular energy with quantum chemistry methods. The ML-predicted reaction energy then was correlated with the catalysts' thermodynamic performance for optimal selection of desired catalysts [307].

The choice of molecular representation and features is essential for ML applications in chemistry [308]. Huang and von Lilienfeld [308] showed the importance of target similarity and features uniqueness in achieving fast, accurate, and transferable ML models. For example, their BAML model, whose features are based on the universal force-field [309] and can be regarded as a way to describe the molecular approximate energy, offers an amazing accuracy (a MAE of roughly 1 kcal/mol) for predicting atomization energy [308].

Many artificial neural network (ANN) models were devised to predict the outcome of chemical reactions [310–315]. The R2B model of Heinen *et al.* [310] was trained on the data of von Rudorff *et al.* [316]. It combines predictions of reaction outcomes with a modeling of kinetics and thermodynamics and could predict energy barriers with MAE as low as 2.5 kcal/mol. The R2B model can also predict the most likely mechanism among competing reaction paths and can help plan synthesis.

To accelerate the process of exploring chemical compound space, which is similar to our goal (i.e., exploring potential interstellar molecules among millions of reactions for complex ISM reaction space), a fast estimation of molecular properties and molecular energetics is needed [39,83,306]. Various ML methods, such as linear regression, *k*-nearest neighbours, KRR, and FFNN, have been applied to the prediction of molecular atomization energy [83]. In reference [83] it is shown that ML can be used successfully even when features are not directly based on physical laws. Interestingly, the KRR can be described in a way that takes the similarity between different molecules or reactions into consideration. Moreover, reference [83] emphasized the importance of a large number of training data,

appropriate model selection, hyperparameter selection, feature optimization, model validation, as well as the benefits of avoiding the sparsity of observations in some splits [83]. In this dissertation, we tested distinct ML techniques starting from simple baseline OLS, RR and successively changing to more complex WLS, FFNN and KRR, which is critical and desirable as suggested in the literature [83].

To avoid the high cost of molecular geometry optimization with common quantum chemistry methods or the difficult cases with empirical approaches, a ML model, G2S, has been constructed for generating reasonable 3D structures of molecules [317]. This ML offers a good accuracy, $MAE < 0.2 \text{ \AA}$ for interatomic distances [317]. The G2S molecular structures can serve as inputs to *ab initio* geometry optimization methods and to ML models used for property prediction [60,83,317].

A ML model, OQML [318], was developed to predict the structure of transition states. It combines ML with the L-BFGS optimization algorithm [319] and achieves a RMSE of $0.14 \text{ \AA}$ for regenerating the geometry of minimum [318a].

The success of the R2B and OQML models [310,318a] is due in part to the efficient representation with FCHL19 features [320]. These features produce efficient estimation of binding energies of water clusters [320]: this could be very helpful for simulating reactions that may take place on ice covered dust grains in the ISM [39]. Better molecular representations and features for ML are still being developed. Khan *et al.* [321] recently reported a fairly small and efficient set of features based on many-body distribution functionals.

In this work, ML can predict the reaction energy barrier directly without finding a TS structure avoiding the computational cost and problems associated with the calculation of normal mode frequencies and IRC [20]. The ML model that we established in this project can predict reaction energy barriers with a speed of more than 100 reactions per second, which is roughly ten million times faster than sophisticated DFT methods. In addition, the features that we developed for our ML satisfy two important criteria: uniqueness and target similarity [308]. Other features or ML methodologies that we can try for barrier prediction in the future can be inspired and derived from the literature [322,323]. Hence, the use of ML is imperative for the energy barrier prediction of thousands of reactions.

Chapter 3 Kernel Ridge Regression Model for Predicting Reaction Energy Barrier

3.1 Introduction

In this chapter, ML estimation of reaction energy barrier is described in detail. The ML models that we developed are fitted to a dataset of 11,730 reaction energy barriers, which were computed by Grambow, Pattanaik, and Green [117a,117b] at ω B97X-D3/def2-TZVP [21–27] level of theory. This level of theory is expected to offer a RMSE of 2.3 kcal mol⁻¹ [5]. We use a common and accurate method, KRR (presented in Chapter 2), along with features that describe geometric changes between products and reactants, and a few features derived from molecular energies that can be obtained from DFT calculations, or PM7 [51–53] semiempirical calculations, or experimental data if available. This computational approach can allow us to predict reaction E in less than a second and could be applied to quickly screen chemical reactions from all possible combinations of molecular species observed or suspected to be present in the ISM and CSE. After that quick screening process, the remaining reactions (those with zero or very low barriers) can be sorted according to their predicted energy barriers and according to the criteria that could assess the relevance of the products of reaction for the origins of life. Such criteria are developed in Chapter 4. This methodology would help prioritize quantum chemistry calculations of transition states, energy barriers, energies of potential ISM/CSE molecules, structural parameters of those molecules, reaction pathways, etc.

3.2 Dataset of Chemical Reactions

3.2.1 Initial Reaction Dataset

From September 2018 to May 2020, we initially created a small-size reaction dataset of 63 reactions and tested some features with OLS and RR models. It was gradually expanded to include 208 reactions. This was assembled painstakingly, one reaction at a time, from the literature [118–161] reporting accurate experimental results or correlated electronic structure calculations. The energy barriers in the dataset of 208 reactions change from 0.0 kcal/mol to 50.0 kcal/mol. For this dataset, we trained a RR model with four different combinations of subsets, such as the hydrogen-transfer reaction subset and its complementary subset, and with roughly 100 combinations of features derived from atomic electronegativity, hardness, and interatomic distances (see section 3.3). Among these many trials, the KRR model stood out as the one giving the lowest MAE, RMSE, and highest adjusted R -squared coefficient of determination (i.e., R_{adj}^2).

As we experimented with models, it became clear that a dataset of 208 reactions is much too small for developing a ML model of sufficient accuracy. Fortunately, Grambow *et al.* [117a,117b] published in 2020 transition states and energy barriers calculated at a good level of theory (estimated RMSE of 2.28 kcal/mol) for 12,000 elementary chemical reactions. From this point onward, we used their dataset for developing and testing ML models.

3.2.2 Dataset of Grambow, Pattanaik and Green

Grambow, Pattanaik, and Green [117a,117b] (GPG) published a very large dataset of reactions (12,000 reactions) with their energy barriers calculated at the ω B97X-D3/def2-TZVP [21–27] level of theory. The DFT method, ω B97X-D3, is a hybrid generalized gradient approximation method (GGA) [21,22–24]. The D3 descriptor indicates that empirical dispersion corrections designed by Grimme [162] are added to the energy functional [21]. Using such a DFT method to estimate 12,000 reaction E is excellent but quite costly both in human and computer time. The molecules in the GPG reaction dataset were taken from the GDB-7 dataset of molecules [117a,117c]. This dataset is our primary source of data for training our ML models. The molecules in the reaction dataset contain only four types of atoms: H, C, N, and O. The molecules have at most 23 atoms and at most seven non-hydrogen atoms.

Due to the limitations of the number of experimental data and the need for detailed atomic properties or electronic structural properties in order to build a feasible ML model, datasets containing a large number of molecules and data obtained from accurate quantum chemistry methods are extremely valuable. The dataset in the literature [316], for instance, was created at a cost of approximately 2.8 million computer core hours. The large-size reaction dataset with known energy barriers calculated at the same level of theory is significant for training a good ML model as the consistency is emphasized in the literature [316].

We re-arranged all reactions and wrote them in the exoergic direction in order to make the data more uniform and facilitate learning by the ML model. We are primarily interested in reactions where two molecules, A and B, combine to form a third one, P, i.e., association

reactions. They are reactions through which chemical evolution could happen in space. Hence, reactions with more than two reactant molecules or two product molecules are excluded. The literature presents reactants and products in ways of Simplified Molecular Input Line Entry System (SMILES) [163] representations. In the case of a few reactions, we were unable to convert molecules to 3D structures by Open Babel [164,165], Avogadro [91,92], and the online tool of NovoPro [166]. We excluded those cases. After implementing the above procedures, our dataset is a subset consisting of 11,730 reactions from reference [117a,117b]. This is the dataset we used in all of our ML work. We will refer to it as the “GPG dataset”.

The GPG dataset has energy barriers that range from 0 to 138.85 kcal/mol. It contains 10,306 molecules, many of which appear in more than one reaction. A Python program was written to calculate and compare the sorted eigenvalues of the Coulomb matrix [60,83] of molecules so that we could create the set of 10,306 distinct molecules. With the aid of Open Babel, it is simple and quick to generate a great number of molecular geometry information based on their known SMILESs.

Note that, even though KRR worked best in tests using the small initial dataset, we redeveloped and tested all ML models using the much larger GPG dataset.

3.3 Features of Chemical Reactions

3.3.1 Tests and Features Selection

In the past five years we tested hundreds of features, and combinations of features. Some of the features we tested failed because they did not improve the accuracy of the model,

others because they could not be calculated in a fast and reliable way. Among the features that we tested and rejected are those based on electrophilicity, HOMO-LUMO gap, bond orders, atomic charges, and mechanical softness (they require costly electronic structure calculations and do not help much), and the number of degrees of freedom in molecules (it does not improve the accuracy). The features that can be calculated reliably and do bring accuracy are presented in sections 3.3.3 and 3.3.4. They fall into two groups, “Set 1” and “Set 2”. The main difference between these two is that features of Set 2 require alignment of the reactant(s) supermolecule and the product(s) supermolecule, including reordering of atoms of the product(s) in a way consistent with that of the reactant(s).

3.3.2 Features Used in the ML Models

Many types of molecular features have been devised for ML estimation of PES [167], atomization energies [60,83,168], bond energies [61], and other properties [62,63]. Among those features are Coulomb matrices (CMs) [60,83], descriptors of atom-centered neighborhood [168], bags of bonds [169], and sum over bonds [63]. In addition to those, more complex features have been utilized for ML prediction of energy barriers [55,170–176], such as topological indices [55,170–172], molecular Morgan fingerprints [172,174], and molecular features from RDKit [55,175,176]. These complex features are difficult to interpret and involve many parameters. Complex features can provide better accuracy. But they have two drawbacks. First, if the number of features is very large, for instance as large as the number of data, the meaning of statistical tests of overfitting becomes unclear and making predictions on data that are not very similar to those in the training and validating sets becomes hazardous. Second, as more features get added, assessing their importance for

energy barrier prediction, and interpreting the meaning of predictions, becomes harder. These problems are generally worse with the most flexible, most accurate, ML methods, such as artificial neural networks (ANN).

In this research project, we choose to use relatively few features with simple interpretability, that can be generalized to bigger molecules and to molecules containing more elements than the four (H, C, N, O) of the GPG dataset.

An important distinction can be made between three ways for creating a features vector $x(\text{reaction})$ of length D for a chemical reaction $A + B \rightarrow P + Q$. The first way, (i), is to calculate D -long features vectors for the infinitely separated molecules and define $x(\text{reaction}) = x(P) + x(Q) - x(A) - x(B)$, denoted as the “separated” case. The second way, (ii), is to first arrange A , B , P , and Q in space in a way that resembles both reactants and products, that is, aligning reactant(s) and product(s). This can be done by putting one molecule at the origin in an arbitrary orientation, and then placing the other molecules’ center of mass at optimum positions, and orienting those molecules in an optimum way. Qualitatively speaking, the optimum alignment can be defined as having every atom of the supermolecule AB near an atom of the same element of supermolecule PQ ; and having no atom of A too close to an atom of B , and no atom of P too close to an atom of Q . This way is called rigid alignment. The detailed algorithm we used is described below. Once that is done, one can calculate the features of the reaction like this: $x(\text{reaction}) = x(PQ) - x(AB)$. A third way, (iii), for calculating features is to perform a flexible alignment of AB and PQ : this is an alignment where the degrees of freedom that get optimized are the center of mass

of the molecules, their orientation, and torsion angles around every single bond. In order to do rigid alignment, we wrote a Python program. The algorithm (pseudo steps) in this program is shown in Scheme 3-1.

Input: Geometry information of single reactant or product Output: Geometry information of oriented supermolecule Step 1. For a reaction $A + B \rightarrow P$, put the center of molecule P at the origin (0, 0, 0) and update its xyz coordinates; Step 2. Put the center of molecule A at the origin as well and update its xyz coordinates; Step 3. Generate three random Euler Angles for A: $\alpha_A, \beta_A, \gamma_A$; Step 4. Use the above three Euler Angles to create a rotation matrix, $\mathbf{R}_A$; Step 5. Generate a random position $\mathbf{P}_A = (X_A, Y_A, Z_A)$ for the center of A; Step 6. For every atom j in A with position vector $\mathbf{p}_j = (x_j, y_j, z_j)$, create the new position $\mathbf{p}'_j = (x'_j, y'_j, z'_j)$ by rotation followed by translation, i.e., $\mathbf{p}' = \mathbf{R}_A \mathbf{p} + \mathbf{P}_A$; Steps 7–12. Repeat steps 2 to 7 for the molecule B instead of A, then obtain $\alpha_B, \beta_B, \gamma_B, \mathbf{R}_B, X_B, Y_B, Z_B$, and latest xyz coordinates of B; Step 13. Randomly orientate A and B to minimize this objective function: $F(y) = \sum_i d_{im(i)} + F_{\text{penalty}}$ where $d_{im(i)}$ is the distance between atoms i and $m(i)$, i is an atom in supermolecule "AB" and $m(i)$ is the atom of the same element in P molecule that is nearest to i, and the penalty function we set up here is to prevent two atoms (one from A and the other from B) from being too close. The penalty is implemented as follows: (1) If $d_{ij} \geq \frac{3}{2}(R_i + R_j)$, $F_{\text{penalty},ij} = 0$; (2) If not, $F_{\text{penalty},ij} = (1.5(R_i + R_j) - d_{ij})^2$; (3) $F_{\text{penalty}} = \sum_{ij} F_{\text{penalty},ij}$. where i, j label atoms of supermolecule "AB", and d_{ij} is the distance between atoms i and j, and R_i, R_j are respectively the atomic radius of atom i and atom j. Since every distance depends on the above 12 variables ($\alpha_A, \beta_A, \gamma_A, X_A, Y_A, Z_A, \alpha_B, \beta_B, \gamma_B, X_B, Y_B, Z_B$), F_{obj} is a function of those same 12 variables. Step 14. Minimize F_{obj} with respect to the 12 variables, and save the nuclear coordinates of the supermolecule AB.
--

Scheme 3-1. Pseudo steps for acquiring the geometry of a reasonably oriented supermolecule.

For a more general reaction $A + B \rightarrow P + Q$, we can perform the same idea by placing A in an arbitrary orientation with geometric center at the origin, and by varying 18 parameters that specify the position and orientation of B, P, and Q. These parameters are the geometric center (X_M, Y_M, Z_M) and three Euler angles ($\alpha_M, \beta_M, \gamma_M$) with $M = B, P, Q$. Denoting a set of 18 parameters by 'y', we minimize the loss function F with respect to y:

$$F(y) = \sum_i d_{im(i)} + \sum_{i < j} [\text{Max}(0, 1.5(R_i + R_j) - d_{ij})]^2 + \sum_{m < n} [\text{Max}(0, 1.5(R_m + R_n) - d_{mn})]^2 \quad (21)$$

where atoms i and j are from supermolecule AB; atoms m and n are from supermolecule PQ; d represents interatomic distances; $m(i)$ is the atom of the same element in PQ that is nearest to i ; the first sum is over all atoms in AB; the second sum is over all atom pairs in AB; the third sum is over all atom pairs in PQ; and R_i, R_j, R_m, R_n are covalent atomic radii for the corresponding element. The second and third terms (representing the penalty term) simulate steric repulsion to prevent A and B, and P and Q, from getting too close. The first term is a measure of how dissimilar supermolecules AB and PQ are. The Python program that we developed here was to minimize this dissimilarity with steric repulsion constraints. This process could serve as a reasonable orientation model for how AB and PQ would approach the transition state.

Later, a second program was written by R. Fournier for doing flexible alignment of molecules and calculating features of 'Set 2'. This has been described in reference [20].

We now briefly describe the flexible alignment algorithm. The cartesian coordinates of A and B are merged into a supermolecule AB. In the flexible alignment, individual

molecules A and B are “remembered” by use of an indicator variable: $m_i = 1$ for atom i that came from molecule A, or $m_i = 2$ for atom i that came from molecule B. Atoms are put in decreasing order of atomic number in AB and in P. Now, a loss function F is minimized:

$$F = \sum_i d_{iAB,iP} + 10 \sum_{i,j \in AB} (1 - \delta_{m_i,m_j}) [\text{Min}(0, r_{ij} - 1.2)]^2 \quad (22)$$

where $d_{iAB,iP}$ is the atomic distance between the i 'th atom of supermolecule AB and i 'th atom of molecule P, δ_{m_i,m_j} is the Kronecker delta, equal to 1 if $m_i = m_j$, equal to 0 otherwise, $r_{ij} = d_{ij}/(R_i + R_j)$, d_{ij} is the distance between atoms i and j , and R_i, R_j are covalent atomic radii of atoms i and j . The first sum is over all atoms. The second term represents a sum of penalties that are zero except when $m_i \neq m_j$ and the distance d_{ij} between the two atoms is smaller than $1.2(R_i + R_j)$. This penalty term prevents reactants A and B from overlapping (being too close). The minimization of F is achieved by simulated annealing with four types of moves: (1) rigid displacement of all atoms of A, or all atoms of B, along x, y or z axis; (2) rotation of all atoms of A, or all atoms of B, around the x, y or z axis, without changing their center; (3) rotation around a single bond of A or B, if that bond is not part of a ring; and (4) interchange of labels i and j for two atoms (with same atomic number) in P. The last type of move is essential for finding the permutation of atoms of same element that leads to the best match between supermolecule AB (reactant side) and molecule P (product side). A small F in Equation (22) (a good alignment) gives unambiguous 1-to-1 mapping of atoms between the reactant side and product side. This level of alignment will allow us to obtain helpful features for reflecting intermolecular interactions (useful for Features Set 1), and for specific bonds and atom pairs (required for Set 2 features).

Now we have 3 ways of generating a geometry for a reaction: (i) separated; (ii) rigid alignment; and (iii) flexible alignment. That gives 3 categories of features (i), (ii) and (iii). Some features can only be calculated with flexible alignment, so they are necessarily type-3 features (those are Set 2 features). Some features can be calculated with (i), (ii), or (iii): in such case, we had to make a choice for obtaining better accuracy.

3.3.3 Features Set 1

We tried many different features similar to those in previous work [83,84,169], and a few more we developed. First, by trial and error, we found a set of 117 features by minimizing the MAE, RMSE, and maximizing the adjusted *R*-squared coefficient of determination between fit and data. The 117 features in Set 1 fall into three categories: 69 features based on Coulomb matrices [60,83,177]; 32 features that depend on the atomic hardness, electronegativity, and interatomic distances of atom pairs; and 16 features that depend on global variables of molecules or reactions. All three types of features encode, in different ways, information about the chemical composition and geometry information of molecules (or supermolecules).

3.3.3.1 Features Related to Coulomb Matrix

The Coulomb Matrix **CM** [60,83], has elements $\mathbf{CM}_{ij}$ introduced by reference 60:

$$\mathbf{CM}_{ij} = \begin{cases} \frac{1}{2}Z_i^{2.4} & \text{if } i=j \\ \frac{Z_iZ_j}{d_{ij}} & \text{if } i \neq j \end{cases} \quad (23)$$

where $\mathbf{CM}_{ij}$ is the element located in the i^{th} row and j^{th} column of $\mathbf{CM}$, Z_i is the atomic number of the i^{th} atom, Z_j is the atomic number of the j^{th} atom, and d_{ij} is the interatomic distance between atoms i and j , $\frac{1}{2}Z_i^{2.4}$ is the diagonal element, and $\frac{Z_i Z_j}{d_{ij}}$ is the off-diagonal element. Inspired by the reference [83], one example of $\mathbf{CM}$ representation (corresponding $\mathbf{CM}$ for ethanol) has been shown in Figure 3-1.

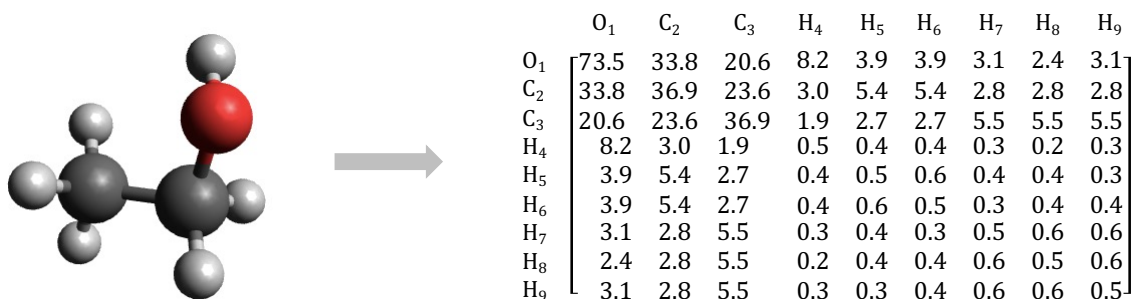


Figure 3-1. Coulomb matrix for ethanol at its equilibrium geometry.

The maximum number of atoms in any reaction of our dataset (including 11,730 reactions) is 23. All $\mathbf{CM}$ s are uniformized to 23×23 matrices by padding with zeros. When the number of atoms $N < 23$, we set $\mathbf{CM}_{ij} = 0$ whenever $i > N$ or $j > N$. This is justified by the fact that adding $(23 - N)$ atoms at infinite separation leaves the energy of the system unchanged and adds these zero matrix elements. $\mathbf{CM}$ s, in themselves, are not suitable representations of molecules because they depend on how atoms are numbered. There are two solutions to this problem mentioned in the literature [83]: take features that are invariant properties of the $\mathbf{CM}$, or follow a procedure to generate a unique ordering of atoms that is, in some sense, optimal (which is a tough procedure for the reaction that has complicated changes of bonds). We choose the first way. We calculate the 23 eigenvalues of $\mathbf{CM}$ arranged in increasing order (sorted eigenvalues) as features.

We generalize the **CM** by replacing the distance d_{ij} in off-diagonal elements (in Equation (23)) by d_{ij}^p . With $p = 1, 2, 3$, we first calculate three kinds of **CMs** (one original and two modified) for separated molecules (A, B, P, and Q) or for supermolecules (AB or PQ). After that, we compute the sorted eigenvalues (SE) of each **CM** with $p = 1$ (the original **CM**), and the SE of each modified **CM** with $p = 2$, and $p = 3$. Based on different p values, we tested two approaches (discussed in section 3.3.2) for computing **CM**-related features of reactions: $x(\text{reaction}) = x(P) + x(Q) - x(A) - x(B)$, or $x(\text{reaction}) = x(PQ) - x(AB)$, where x refers to SE.

The above process gives 69 invariant features associated with generalized **CMs**. This is a simple way of getting a finer description of molecular structure. The regression models we obtained by using all three kinds of **CMs** and 69 invariant features offer a slightly better R_{adj}^2 than models using only the 23 sorted eigenvalues from the original **CM** (with $p = 1$).

3.3.3.2 Features of distances and atoms' hardness and electronegativity

The chemical composition and its relation to geometry information are encoded in **CMs** with atomic numbers Z_i . The relation between atomic numbers and chemical bonds is, however, far from trivial. Building features that can reveal information about chemical bonding is of great importance for estimating reaction energy barriers. Electronic properties of atoms, in particular ionization energy (I) and electron affinity (A), have a much more direct relation to chemical bonding. The sum and difference of I and A are, respectively, the Mulliken's electronegativity:

$$\chi = \frac{(I + A)}{2} \quad (24)$$

and the Pearson-Parr hardness:

$$\eta = \frac{(I - A)}{2} \quad (25)$$

We use these two properties to characterize atoms. Accordingly, we define features that are sums over pairs of atoms ij of simple functions of interatomic distance d_{ij} , electronegativity difference between two atoms $|\chi_i - \chi_j|$, and mean hardness of two atoms $\frac{1}{2}(\eta_i + \eta_j)$. These features are defined by the formulas:

$$x(p, q, r) = \sum_{i>j} \left(\frac{R_i + R_j}{d_{ij}} \right)^p y_{ij,r}^q$$

$$y_{ij,1} = |\chi_i - \chi_j|$$

$$y_{ij,2} = \frac{1}{2}(\eta_i + \eta_j) \quad (26)$$

where $p = 1, 3, 5, 7$ and $q = 1, 2, 3, 4$ and $r = 1$ or 2 is a label for two categories of features; R_i is the covalent radius of atom i . The formulas with the above combination of exponentials define $(4 \times 4 \times 2) = 32$ features.

The numerical values we chose for the exponents ($p = 1, 3, 5, 7$ and $q = 1, 2, 3, 4$) are somewhat arbitrary. We tested other combinations, such as $1, 2, 3$; $1, 2, 3, 4$; $1, 3, 5, 7$; etc.

We finally chose $p = 1, 3, 5, 7$ and $q = 1, 2, 3, 4$ because it gave the highest adjusted R -square coefficient.

Like **CM**-related features, we also tested two approaches (see 3.3.2) for computing the above χ and η -related features of reactions: $x(\text{reaction}) = x(P) + x(Q) - x(A) - x(B)$, or $x(\text{reaction}) = x(PQ) - x(AB)$, where x refers to the above $x(p,q,r)$ in Equation (26) for each molecule (A, B, P, Q) or supermolecule (AB, PQ).

3.3.3.3 Global features

The energy of a reaction, $E(\text{reaction}) = E(P) + E(Q) - E(A) - E(B)$, is a desirable feature when trying to model the reaction energy barrier. To calculate the energy of a reaction, first we need to use QC methods to calculate the energy of each molecule (A, B, P, Q). In this research project, we use the enthalpy change of reaction ($\Delta_r H$), energy change of reaction ($\Delta_r E$), and Gibbs free energy change of reaction ($\Delta_r G$) as some of our features. Note that, the enthalpy of formation, $\Delta_f H$, has been determined experimentally for many small molecules, but it is not known for *all* the molecular species we may need for modelling a complex system like the ISM and CSE.

In order to calculate molecular energies, and then calculate $\Delta_r H$, $\Delta_r E$, and $\Delta_r G$, we prefer the PM7 semiempirical method [51–53] because of its low computational cost. PM7 calculations can be easily completed for the kind of molecular species we are considering, typically those with 50 atoms or fewer atoms with $Z < 19$. Generally, in a complex system, a set of "K" molecules will give rise to a far bigger number "J" of possible chemical

reactions (see section 4.1). The cost of "K" PM7 molecular energy calculations, or even "K" DFT calculations, will be a negligible fraction of the hypothetical cost of finding, by first-principles methods, the transition states and energy barriers of "J" chemical reactions. In other words, finding TS and calculating reaction energy barriers from the potential energy surfaces with DFT is much more costly than ML estimation of barrier with features from molecular energy calculations (even with DFT molecular energy calculations). For this reason, we also consider models that use DFT molecular energies as inputs. The cost of PM7 calculations is moderate even for large-size reaction datasets. For instance, PM7 calculations of optimized geometries and energies for the 26,790 molecules in our dataset from reference [117] took only 12 days using 32 cores.

We obtained $\Delta_r H$, $\Delta_r E$, and $\Delta_r G$ for the 11,730 reactions from reference [117a,117b]. Those were calculated by DFT with the ω B97X-D3 [21–24] dispersion corrected exchange-correlation functional and the triple-zeta basis set denoted def2-TZVP [25–27]. In the remainder of this chapter, the ω B97X-D3/def2-TZVP method will be referred to simply as “DFT”. In addition to making use of $\Delta_r H$, $\Delta_r E$, and $\Delta_r G$ from DFT, we also generated molecular geometries with Open Babel [164,165], and by energy minimization with the PM7 semiempirical method, and calculated $\Delta_r H$, $\Delta_r E$, and $\Delta_r G$ from PM7 energies (see section 3.3.5). We use $\Delta_r H$, $\Delta_r E$, and $\Delta_r G$ as three of the features, as well as $\sqrt{|\Delta_r H|}$ and $\Delta_r H^2$.

In addition to the above energy-related features, we use eleven more global features. They are: the number of atoms of each element (H, C, N, O); the difference between

products and reactants, in the number of molecules; and the number of single or aromatic bonds, number of double bonds, and number of triple bonds, in reactants and in products. There were many global features that we tested in the past five years: the 16 described above help improve accuracy and were used.

All the above procedures make a total of 117 features for our current features in set 1.

3.3.4 Features Set 2

In addition to the features of set 1, we considered a different kind of features that depend on a representation of bonds being broken and formed during a reaction. They complement features of section 3.3.3 and can improve the accuracy of the ML model. They are also closer to a chemist’s intuition. In addition to equilibrium nuclear positions, the inputs are standard values of average bond lengths [178] and average bond energies [179] for single, double, and triple bonds.

For acquiring Features Set 2, the first step in the calculation is a flexible alignment of the molecular geometries of reactants and products (described in section 3.3.2). This gives a 1-to-1 mapping of atoms between the supermolecule AB and the molecule P, for a reaction $A + B \rightarrow P$.

Formal bond energies B_{ij} are assigned to every atom pair ij by linear interpolation of bond energies, with interatomic distances d_{ij} as the interpolation variables, and with bond lengths and bond energies of single, double, and triple bonds as parameters. A triple bond energy

is assigned to the atom pair ij when d_{ij} is smaller than a triple bond length. When d_{ij} is larger than the bond length of a single bond, we set bond energies as:

$$B_{ij} = B_1 e^{8(1 - r_{ij})} \quad (27)$$

where $r_{ij} = d_{ij}/(R_i + R_j)$ ($r_{ij} \sim 1$ for single bonds), B_1 is the bond energy of a single bond for this pair of elements, and the parameter $a = 8$ is arbitrary. Here we still consider a reaction $A + B \rightarrow P$.

A fictitious mass $m_{ij} = |B_{ij,AB} - B_{ij,P}|$ is assigned to every pair ij of atoms: m_{ij} is big when there is a big change in bond energy (the big change can be positive or negative) from reactant to product; m_{ij} is small when there is a slight change in bond energy from reactant to product. Note that atoms pairs ij with $m_{ij} < 0.1 \text{ kJ mol}^{-1}$ are ignored from the rest of the calculation, i.e., only the features related to significant changes are added to our current features dataset.

We indicate a property of the product with a prime, e.g., B'_{ij} , and a property of the reactant with an unprimed variable, e.g., B_{ij} . Atom pairs are described with a small set of variables: an indicator of chemical composition n_{ij} (HH:1, HC:2, HN:3, HO:4, CC:5, CN:6, CO:7, NN:8, NO:9, OO:10); the bond energy change between products and reactants, $E_{ij} = B'_{ij} - B_{ij}$; the fictitious mass $m_{ij} = |E_{ij}|$; the position (cartesian coordinates) of the bond in space (details shown below), c_{ij} ; and the orientation of the bond (a unit length vector, details shown below), u_{ij} . The position of pair ij is computed as $c_{ij} = (x_i + x_j + x'_i + x'_j)/4$, where x_i

and x'_i denote atom i 's position in the reactant and product, and x_j and x'_j denote atom j 's position in the reactant and product. The unit length vector u_{ij} points in the direction from $(x_i + x'_i)/2$ to $(x_j + x'_j)/2$. The set of points with position c_{ij} and fictitious mass m_{ij} is moved, so that its center of mass is at the origin and the moments of inertia are calculated. The set of points is then rotated to a standard orientation with the principal axes along the x , y , z axes and moments of inertia $I_{xx} \leq I_{yy} \leq I_{zz}$. The distance of atom pair ij to the center is $\|c_{ij}\|$.

With the above variables (or properties), a first set of 200 features that depend on bond energy changes and distance to the center is calculated with the formulas:

$$A_{n,p,q} = \sum_{(n_{ij}=n)} E_{ij}^p \|c_{ij}\|^q \quad p = 1, 2, 3, 4; q = 0, 1, 2, 3 \quad (28)$$

$$B_{n,p,0} = \sum_{(n_{ij}=n)} m_{ij}^p \quad p = 1, 2, 3, 4 \quad (29)$$

In Equations (28) and (29), each sum is over the atom pairs of a given composition n (HH, or HC, or HN, or HO, or CC, or CN, or CO, or NN, or NO, or OO).

A second set of 36 features that depend on the relative positions of pairs of bonds are calculated with:

$$P_{p,q,r,s} = \sum_{ij} \sum_{kl} E_{ij}^p E_{kl}^q \|c_{ij} - c_{kl}\|^r (u_{ij} \cdot u_{kl})^s$$

$$p = 1, 2, 3; q = 1 \text{ to } p; r = 1, 2, 3; s = 0, 1 \quad (30)$$

where the double sum is over all pairs of atom pairs (ij, kl) , $\|c_{ij} - c_{kl}\|$ is the distance between the centers of the atom pairs (ij, kl) , and $(u_{ij} \cdot u_{kl})$ is the cosine of the angle between the orientations of the two pairs (ij, kl) .

The next four features describe the chemical composition of atom pairs with the biggest energy changes.

$$Q_k = N \sum_{ij} m_{ij} a_{ij,k} \quad k = 1, 2, 3, 4 \quad (31)$$

where the value (1, 2, 3, 4) of k indicates an element (H, C, N or O), and $a_{ij,k}$ is the number of atoms of element k in pair ij (0 for that element not appearing in the pair ij , 1 for appearing once, or 2 for appearing twice), and N is a normalization factor that makes the sum of Q_k s equal to 1. Three more features are obtained by taking the weighted average of bond orientations with weights equal to m_{ij} , and the last three are the moments of inertia.

All the above procedures make a total of 246 features for our current Features Set 2.

Sets 1 and 2 combined have 363 features. To avoid redundancy from possible similar features, we perform the feature selection [180] procedure with the minimum redundancy maximum relevance (MRMR) algorithm [181,182] using MATLAB Online [116]. The 363 features are ranked by importance, according to the MRMR algorithm, and only the 300 most important are kept. These 300 features are our latest and most helpful features. For the 63 features that are rejected by MRMR, there are 25 features from Set 1 and 38 from Set 2.

3.3.5 Different Approaches for Calculating Features

The literature [117a,117b] provided SMILES strings for reactants and products in the reaction dataset. We generated the nuclear coordinates for the molecules in our 11,730 reactions (derived from the literature [117a,117b]) by converting their SMILES strings to 3D structures using the “bunch converting” function of Open Babel [164,165]. In a few cases, this failed. In such cases, we would use the converting function of Avogadro [91,92] and the online tool of NovoPro [166]. The molecular geometries generated by the above procedure are simply denoted “Babel” in the following contents.

We tested many combinations of features and many ways (DFT, PM7, Babel) to calculate the equilibrium geometries from which they are obtained. We abandoned the use of many of these combinations and describe them briefly here only for completeness.

Strategies for calculating features are characterized by three empirical or quantum chemical methods for computing:

- a) equilibrium geometries and $\Delta_r H$, $\Delta_r E$, and $\Delta_r G$;
- b) equilibrium geometries used as input in the calculation of features of Set 1;
- c) equilibrium geometries used as input in the calculation of features of Set 2.

Using DFT for step a), and Babel for steps b) and c) is denoted Babel,Babel//DFT; using DFT for step a), PM7 for step b), and no features of Set 2 is denoted X,PM7//DFT; an approach where $\Delta_r H$, $\Delta_r E$, and $\Delta_r G$ are not used, and PM7 is used for b) and c) is denoted PM7,PM7//X; and so on. The seven approaches that we investigated are: X,Babel//X;

X,PM7//PM7; X,DFT//DFT; Babel,PM7//PM7; Babel,DFT//DFT; PM7,PM7//PM7; DFT,DFT//DFT. This will be discussed in sections 3.5 and 3.6.

3.4 KRR Model for Energy Barriers

The short-range features derived from the electronegativity and hardness of atoms ($N = 32$) are useful in describing chemical bonds but not helpful in describing intermolecular interactions. For this reason, features of *chemical reactions* ($A + B \rightarrow P + Q$) defined by Equation (26) are evaluated as $x(\text{reaction}) = x(P) + x(Q) - x(A) - x(B)$. Unlike the above case, coulomb matrices-related features ($N = 69$) are equally relevant for describing chemical bonds and nonbonding interactions. Since the supermolecules AB and PQ carry information about the general shape of transition state, it is preferable to calculate features derived from Coulomb matrices for supermolecules, i.e., $x(\text{reaction}) = x(PQ) - x(AB)$. Global features, such as $\Delta_r H$, $\Delta_r E$, and $\Delta_r G$, do not depend on the relative positions and orientations of A and B, or P and Q.

Using supermolecular geometries for all features except global ones increases the RMSE by ~ 1 kcal mol⁻¹, and using separated molecular geometries for everything increases the RMSE by ~ 2 kcal mol⁻¹. Numerical tests show that using the geometries of separated molecules for the features of equation (26) and aligned reactant and product (see 3.3.2) for all other features gives slightly more accurate fitting results.

We trained and tested all ML models described in section 2.2.3 – 2.2.7 with 10-time Monte Carlo cross-validation. Each time the entire set (minus the outliers) is divided at random into two sets: 10% of the data as the testing set, remaining as the training set. A

completely random partition of the data could sometimes lack balance, especially for a reaction dataset with energy barriers spanning a broad range. What could happen is that, for instance, a higher proportion of very large energy barriers is in one of the groups. Lacking balance might weaken ML's prediction ability. However, this problem can be eliminated, and accuracy is improved, by performing a procedure described in the reference [83] that is partly random but ensures similar distributions of reaction energy barriers in every group. The data is arranged in order of increasing energy barrier and is divided into bins of size ~ 10 . One data from each bin is then assigned to the testing set to be used in cross-validation. The procedure is shown in Figure 3-2.

ML models often give highly accurate results if they are assessed on the training sets. However, the predictive value of a ML model, and its accuracy, must be assessed on the testing set. Unless stated otherwise, the performance results reported in sections 3.5 and 3.6 are all obtained against testing sets.

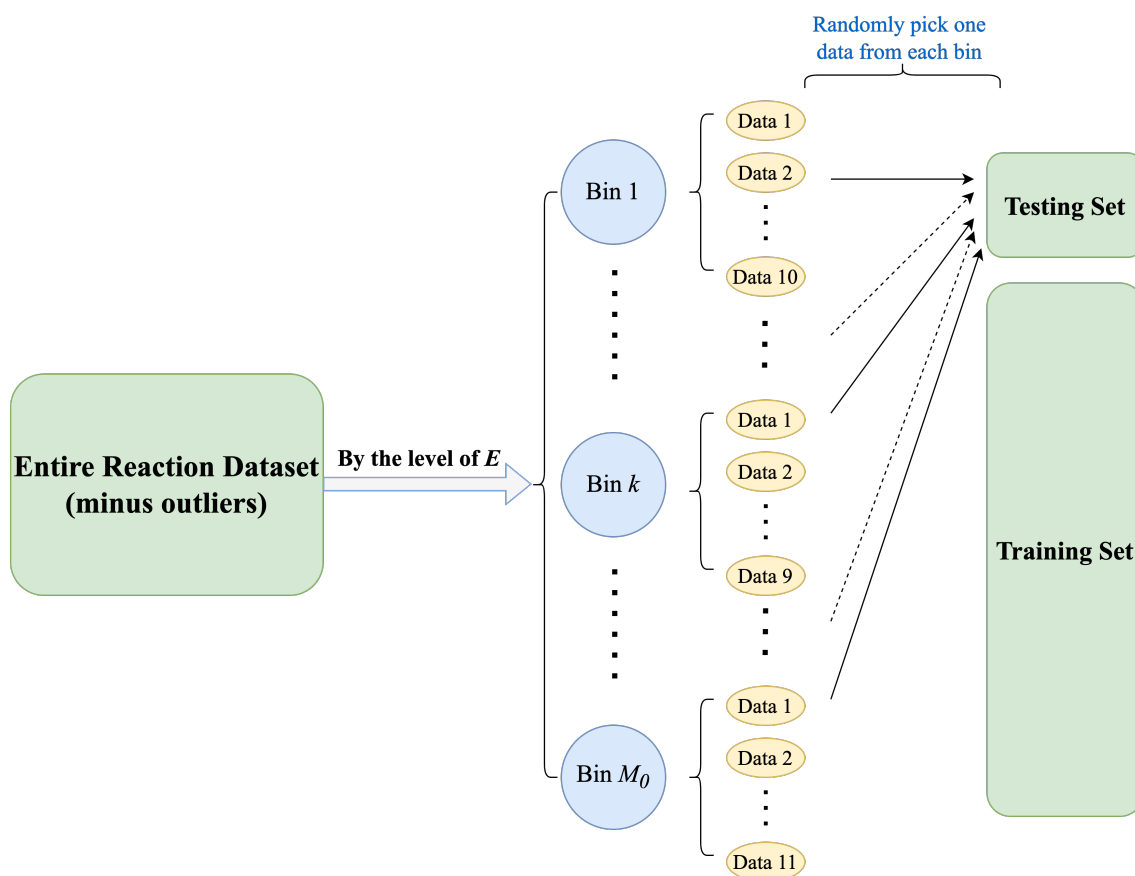


Figure 3-2. Procedure of how to rearrange data by energy barrier and how to build the testing set where the number of total bins is M_0 .

The performances of KRR models depend on the hyperparameters α and γ (introduced in section 2.2.7). We used the *GridSearchCV* function in the Python module *Scikit-learn* [68] to optimize α and γ with different choices of sets of features. The *GridSearchCV* optimization of hyperparameters was done by 5-fold cross-validation as illustrated in Figure 3-3.

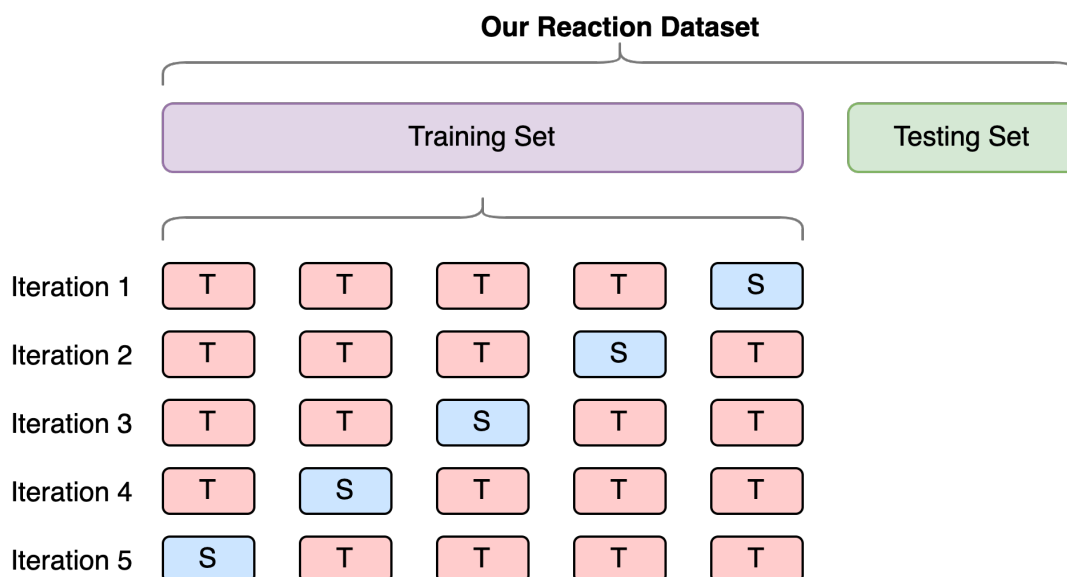


Figure 3-3. The 5-fold cross-validation process achieved in *GridSearchCV*. The blue "S" is a testing set in 5-fold cross-validation occupying 1/5 of the original Training Set, and pink "T" is a training set in 5-fold cross-validation occupying 4/5 of the original Training Set.

The input values, and optimum value, of hyperparameters optimized with *GridSearchCV* are shown in Table 3-1. The dependence of the MAE and RMSE of KRR models with different γ are shown in Figure 3-4. The optimal γ is where a minimum is found in the MAE and RMSE curves for the testing set. It is also where there is a sudden decrease in the magnitude of the slope of the MAE and RMSE for the training set. The base of the grey rectangle in Figure 3-4 indicates the range of acceptable γ values (0.005~0.02). Increasing

γ nearly always decreases the MAE and RMSE measured against the training set but can lead to higher MAE and RMSE against the testing set: this is a clear sign of overfitting. As shown in Figure 3-4, overfitting occurs at $\gamma \gtrsim 0.02$.

Table 3-1. Selection procedure of best hyperparameters by *GridSearchCV* for KRR

Hyperparameter	Input Values	Selected Value (work best)
α	First Step: $10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10$ Second Step: $5 \times 10^{-3}, 10^{-2}, 5 \times 10^{-2}, 10^{-1}, 5 \times 10^{-1}$	5×10^{-2}
γ	First Step: $10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10$ Second Step: $5 \times 10^{-3}, 10^{-2}, 5 \times 10^{-2}, 10^{-1}, 5 \times 10^{-1}$	10^{-2}

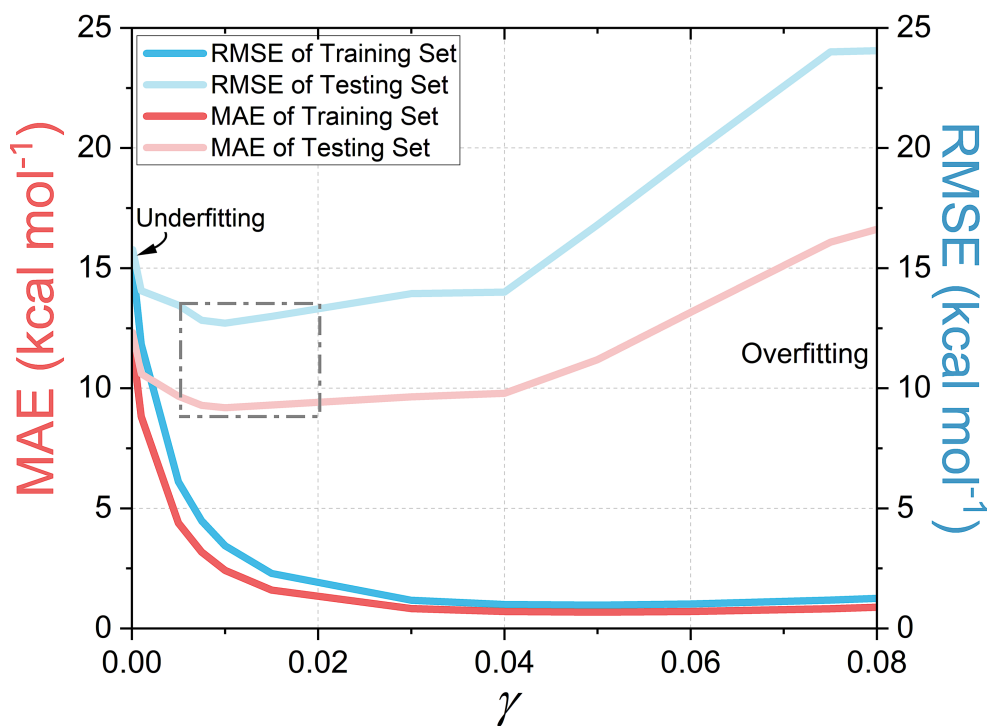


Figure 3-4. Trend in MAE and RMSE for the training and testing sets as a function of γ with a fixed $\alpha = 0.05$. The grey rectangle indicates the range of acceptable γ values.

We tested a very large number of different features, and different subsets (combinations) of features, trying to achieve the smallest MAE, RMSE, and the biggest R_{adj}^2 to arrive at the features in sections 3.3.3 and 3.3.4. The adjusted coefficient of determination R_{adj}^2 is given by

$$R_{\text{adj}}^2 = 1 - \left[\frac{(1 - R^2)(M - 1)}{M - N - 1} \right] \quad (32)$$

where M is the number of reactions in our database, and N is the number of features. We found that the 300 features described in previous sections show the best performance (i.e., $N = 300$).

3.5 Performance of Various Machine Learning Models

We first assess the performance of five regression methods (six ML models), OLS, RR, WLS, KRR with Gaussian kernel, KRR with Laplacian kernel, and FFNN-3HL using features set 1 (section 3.3.3). In each case, we use the features of set 1 derived in the X,DFT//DFT approach and a dataset of 11,624 reactions, the entire dataset of 11,730 reactions minus 106 outliers as explained in section 2.2.2.

We perform 10-time Monte Carlo cross-validation using bins of energy barriers in the procedure for splitting the entire set into training and testing sets. The optimal values of the hyperparameters are $\alpha = 0.05$ and $\gamma = 0.01$. The KRR results do not change appreciably when these hyperparameters are multiplied or divided by 2.

The statistics of the ten regressions (10-time Monte Carlo cross-validation), assessed against their respective testing sets, are listed in Table 3-2. As shown, there is no significant

difference in the performances of OLS, RR, and WLS models. In all three cases, the MAE is 13.8 or 13.9 kcal mol⁻¹ and the coefficient of determination, R^2 , is between 0.56 and 0.58. The KRR method with Gaussian kernel and FFNN-3HL are better, with a MAE of 10.8 to 10.9 kcal mol⁻¹ and R^2 is between 0.67 to 0.70. As another nonlinear method, FFNN-3HL is not the best one based on our features set 1. The KRR method with Laplacian kernel is better still, with a MAE of 9.2 kcal mol⁻¹ and $R^2 = 0.77$. The difference between the two KRR methods is significant, and this phenomenon suggests that other kernels, or other nonlinear regression methods, might improve results further.

Table 3-2. Performance statistics (MAE, RMSE, R^2 , R^2_{adj}) of six ML models using the same 117 features calculated in the X,DFT//DFT approach. The mean and standard deviation over ten testing sets generated by 10-time Monte Carlo cross-validation are shown for each statistic

ML model	MAE (kcal mol ⁻¹)	RMSE (kcal mol ⁻¹)	R^2	R^2_{adj}
OLS	13.90	17.55	0.5697	0.5215
	± 0.27	± 0.31	± 0.0147	± 0.0163
RR	13.78	17.33	0.5804	0.5334
	± 0.32	± 0.39	± 0.0181	± 0.0202
WLS	13.81	17.79	0.5582	0.5088
	± 0.18	± 0.27	± 0.0138	± 0.0153
KRR (Gaussian kernel)	10.86	15.32	0.6717	0.6349
	± 0.23	± 0.46	± 0.0193	± 0.0214
KRR (Laplacian kernel)	9.17	12.79	0.7710	0.7454
	± 0.20	± 0.23	± 0.0081	± 0.0090
FFNN-3HL	10.81	14.66	0.6998	0.6662
	± 0.25	± 0.31	± 0.0156	± 0.0174

We will not show the detailed results of numerous tests we did with different combinations of features. Instead, we will simply summarize the main conclusions of those tests. Most of the time, adding features brings only small improvements in RMSE, MAE, and R^2_{adj} . However, removing any feature from the set of 117 decreases R^2_{adj} a little. The increase of RMSE and MAE, and decrease of R^2 associated with the removal of features and method KRR (with Laplacian Kernel) is informative. Removing the features that relate to $\Delta_r H$, $\Delta_r E$, and $\Delta_r G$ increases the RMSE of the best method by 5.0 kcal mol⁻¹ and decreases R^2 by 0.21. This is big effect. Removing the 46 features from sorted eigenvalues of the modified Coulomb matrices (with $p = 2, 3$ in CM's off-diagonal elements), while keeping those of the original Coulomb matrix (with $p = 1$ in CM's off-diagonal elements), increases the RMSE by 0.2 kcal mol⁻¹ and decreases R^2 by 0.01. Removing all features based on electronegativity and hardness of atoms (from Equation (26)) increases the RMSE by 1.2 kcal mol⁻¹ and decreases R^2 by 0.05. Different choices for the set of exponents p and q in Equation (26) typically increase the RMSE by 0.01 to 1 kcal mol⁻¹ and decrease R^2 by 0.001 to 0.01. Excluding the global features other than those based on $\Delta_r H$, $\Delta_r E$, and $\Delta_r G$, increases the RMSE by 1.1 kcal mol⁻¹ and decreases R^2 by 0.04.

Tests with different hyperparameters showed that results are insensitive to α over a wide range of α . Unlike α , hyperparameter γ has a bigger effect on the performance of KRR. When $\gamma < 0.002$, KRR predictions are akin to taking averages over very many barriers of the training set: predictions are non-specific, and we have the underfitting issue. When γ is too big, say $\gamma > 0.04$, the KRR predictions are a bit like taking averages over very few

barriers of the training set, only those of reactions that are highly similar, in the sense of features, to the reaction of interest. This kind of fit is short-sighted and also not good. When the reaction for which we seek a prediction is in the training set the “prediction” appears extremely good (in the limit $\gamma \rightarrow \infty$ it is perfect), but when it is not, the prediction is very poor. That is the overfitting issue.

Table 3-3. Performance statistics (MAE, RMSE, R^2 , R^2_{adj}) of KRR models using features calculated in seven approaches. The mean and standard deviation over ten testing sets generated by 10-time Monte Carlo cross-validation are shown for each statistic

Approach of acquiring features	MAE (kcal mol ⁻¹)	RMSE (kcal mol ⁻¹)	R^2	R^2_{adj}
X,Babel//X	13.44	17.76	0.5581	0.5093
(KRR: $\alpha=0.05$, $\gamma=0.01$)	± 0.22	± 0.30	± 0.0151	± 0.0168
X,PM7//PM7	11.16	15.33	0.6705	0.6335
(KRR: $\alpha=0.05$, $\gamma=0.01$)	± 0.17	± 0.38	± 0.0160	± 0.0177
X,DFT//DFT	9.17	12.79	0.7710	0.7454
(KRR: $\alpha=0.05$, $\gamma=0.01$)	± 0.20	± 0.23	± 0.0081	± 0.0090
Babel,PM7//PM7	9.79	14.06	0.7238	0.6246
(KRR: $\alpha=0.05$, $\gamma=0.01$)	± 0.22	± 0.39	± 0.0150	± 0.0204
Babel,DFT//DFT	8.49	12.49	0.7821	0.7043
(KRR: $\alpha=0.05$, $\gamma=0.01$)	± 0.26	± 0.39	± 0.0135	± 0.0183
PM7,PM7//PM7	9.90	14.33	0.7129	0.6099
(KRR: $\alpha=0.05$, $\gamma=0.01$)	± 0.38	± 0.55	± 0.0216	± 0.0294
DFT,DFT//DFT	7.71	11.76	0.8062	0.7369
(KRR: $\alpha=0.05$, $\gamma=0.01$)	± 0.20	± 0.31	± 0.0101	± 0.0137

According to Table 3-3, the quality of the prediction power of a ML model depends on the quality of optimized molecular geometries used to calculate features and also depends on the quality of molecular energies ($\Delta_r H$, $\Delta_r E$, and $\Delta_r G$, and different QC methods for molecular energy calculation lead to different accuracy). Table 3-3 displays how different molecular geometries and energies (seven approaches to calculating features) affect the performance of KRR. As shown in Table 3-3, using KRR with Laplacian kernel in each case, the MAE goes from 13.4 kcal mol⁻¹ to 7.7 kcal mol⁻¹, and the RMSE goes from 17.8 kcal mol⁻¹ to 11.8 kcal mol⁻¹, between the simplest (X,Babel//X) and best (DFT,DFT//DFT) calculations, while R^2 goes from 0.56 to 0.81. It is very important to keep in mind that the cost of calculations varies by roughly 7 orders of magnitude as we go from a force field (X,Babel//X) to a semi-empirical quantum method like PM7, and a sophisticated but accurate DFT method like ω B97X-D3. In large-scale applications of ML estimates of reaction energy barriers involving, e.g., millions of reactions, it is possible and feasible to acquire PM7 geometries and energies for every reactant and product molecule, and the models we denote X,PM7//PM7, or Babel,PM7//PM7, or PM7,PM7//PM7 can be used. But it is harder to use an approach for generating features like DFT,DFT//DFT. On the other hand, DFT equilibrium molecular geometries (not transition states) and their energies are already available for thousands of molecules and can realistically be calculated for thousands more (it is still much cheaper than finding thousands or even millions of TS with DFT). When available, such DFT equilibrium molecular geometries and their energies should be used, and energy barrier predictions should be made with model X,DFT//DFT, Babel,DFT//DFT, or DFT,DFT//DFT. When very many energy barriers (e.g., 100,000

energy barriers) must be estimated by ML, it would be preferable to construct several regression models and predict every barrier in one or more ways. In a mixed approach like this, different energy barrier predictions will come with different uncertainty estimates.

One can use such a procedure in a future study of chemical reactions in complex systems, such as oceans, ISM, and CSE. For our application to astrochemistry, since the reactions of main interest for ISM and CSE are those with very low barriers, a hierarchy of models can be used to discard reactions above a certain threshold. One could discard, for instance, reactions with barriers bigger than 30 kcal/mol in a regression model based entirely on PM7 data; then discard from the remaining ones those with barriers bigger than 20 kcal/mol in a regression model based on DFT data; and then discard reactions with barriers bigger than 10 kcal/mol in a regression model based on CCSD(T) single-point energies and DFT optimized geometries.

Figure 3-5 shows scatter plots of energy barriers, predicted by the KRR models with 363 features from Babel,PM7//PM7 or Babel,DFT//DFT approaches vs. those of reference [117a,117b]. Figure 3-6 shows scatter plots of energy barriers, predicted by the KRR models (with 300 most important features from PM7,PM7//PM7 or DFT,DFT//DFT approaches) vs. those of reference [117a,117b]. The in-sample (training set) correlations are good (panels A, C, E, and G), but they are of little use. The out-of-sample (testing set) predictions (panels B, D, F, and H) are, unsurprisingly, less accurate. As shown in panels B, D of Figure 3-5 and F, H of Figure 3-6, a small number of points are very far from the fit line and contribute a lot to the RMSE. A comparison of RMSE to the MAE (Table 3-2)

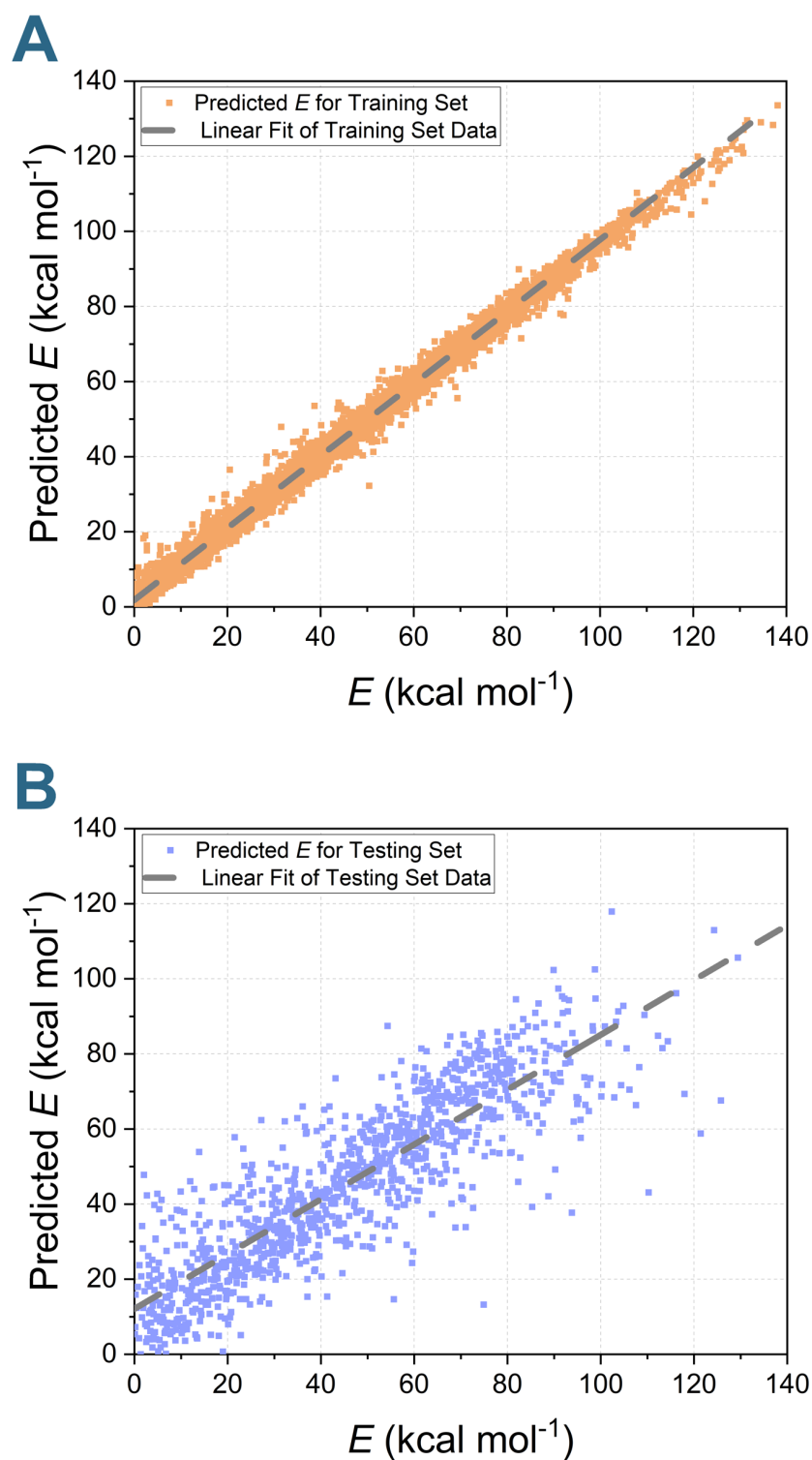


Figure 3-5. A: In-sample correlation of predicted and reference barriers for the KRR model Babel,PM7//PM7. **B:** Out-of-sample correlation (same method as A).

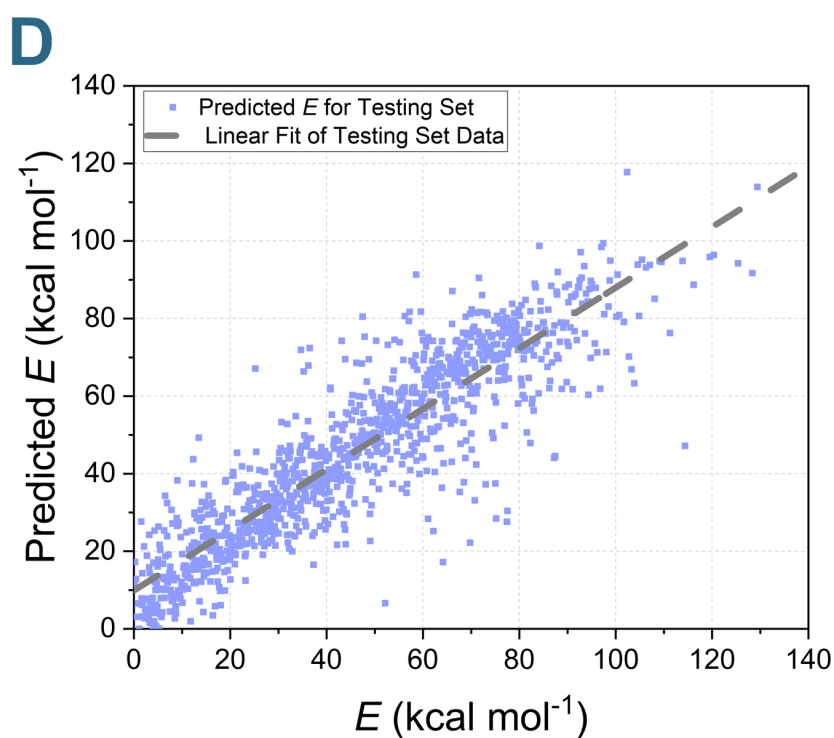
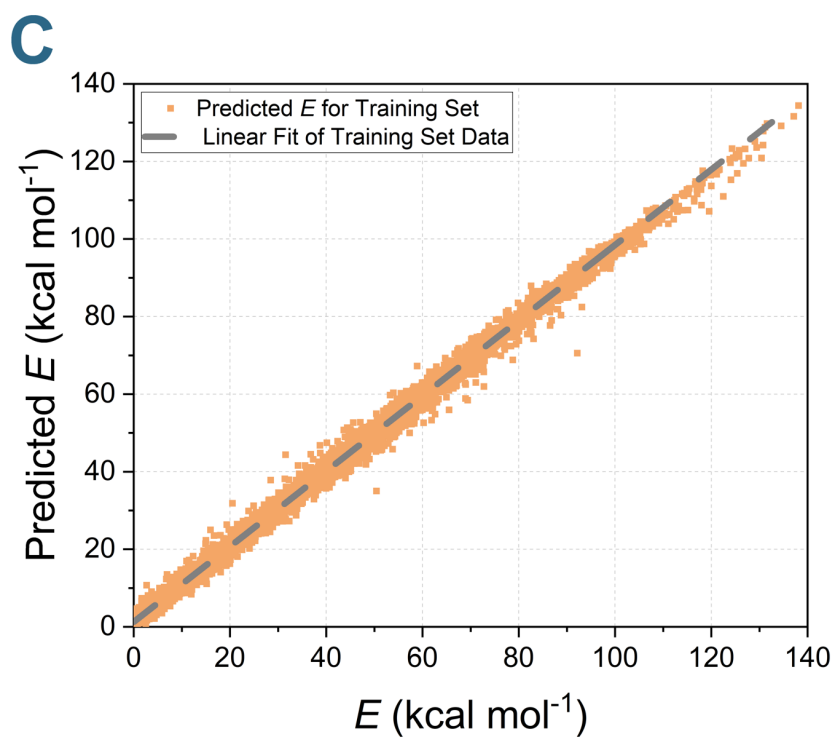


Figure 3-5. **C:** In-sample correlation of predicted and reference barriers for the KRR model Babel,DFT//DFT. **D:** Out-of-sample correlation (same method as C).

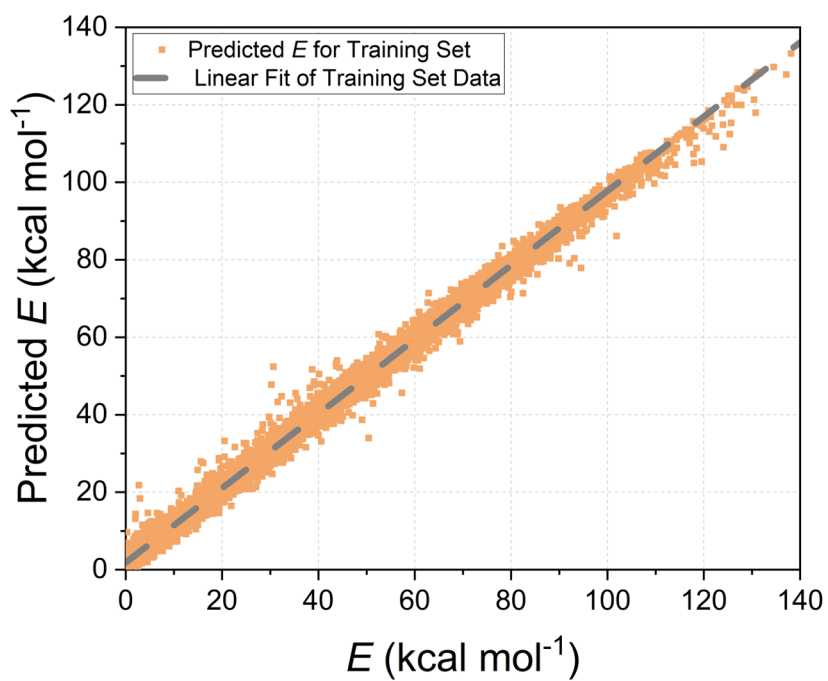
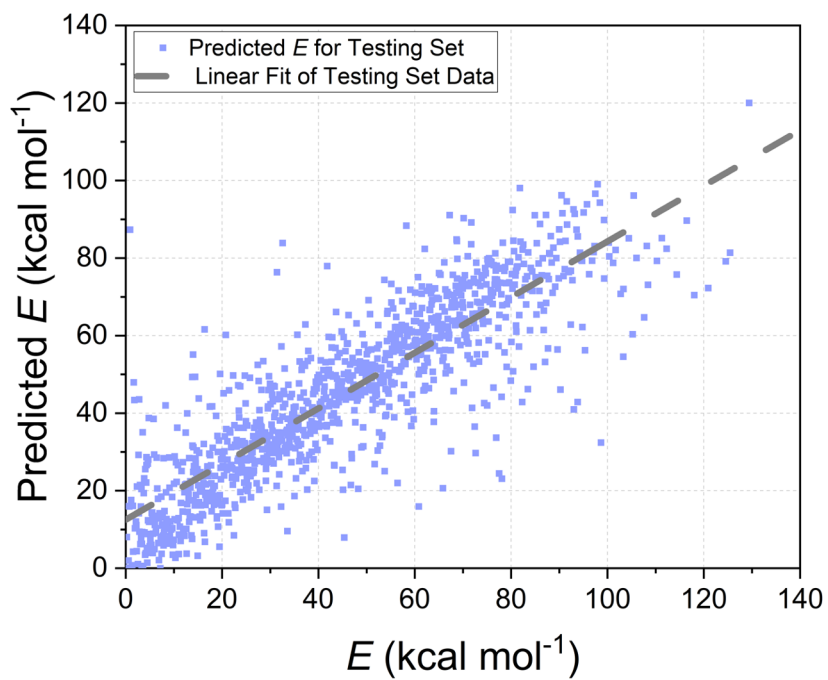
E**F**

Figure 3-6. E: In-sample correlation of predicted and reference barriers for the KRR model PM7,PM7//PM7. **F:** Out-of-sample correlation (same method as E).

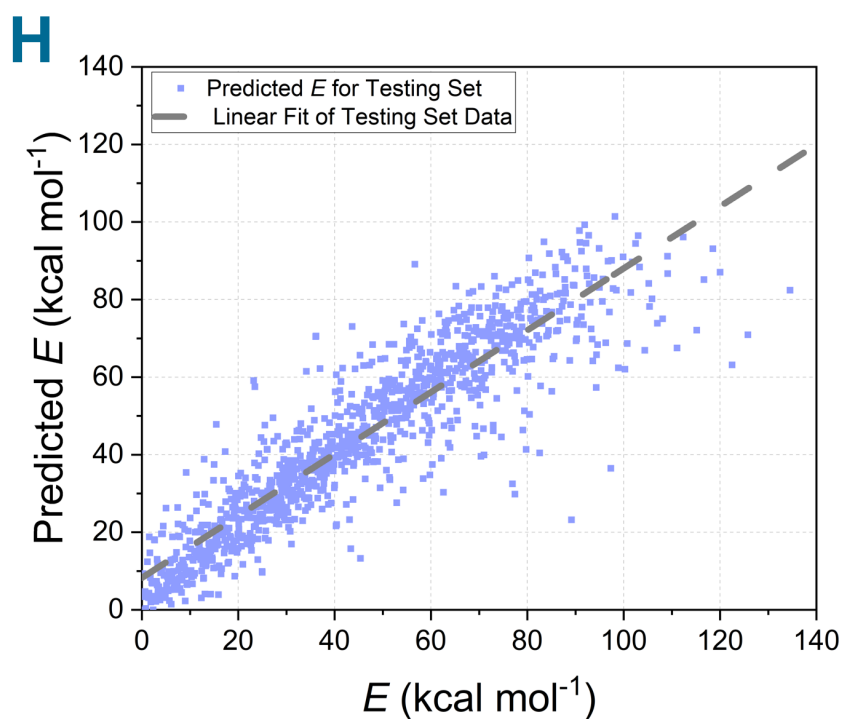
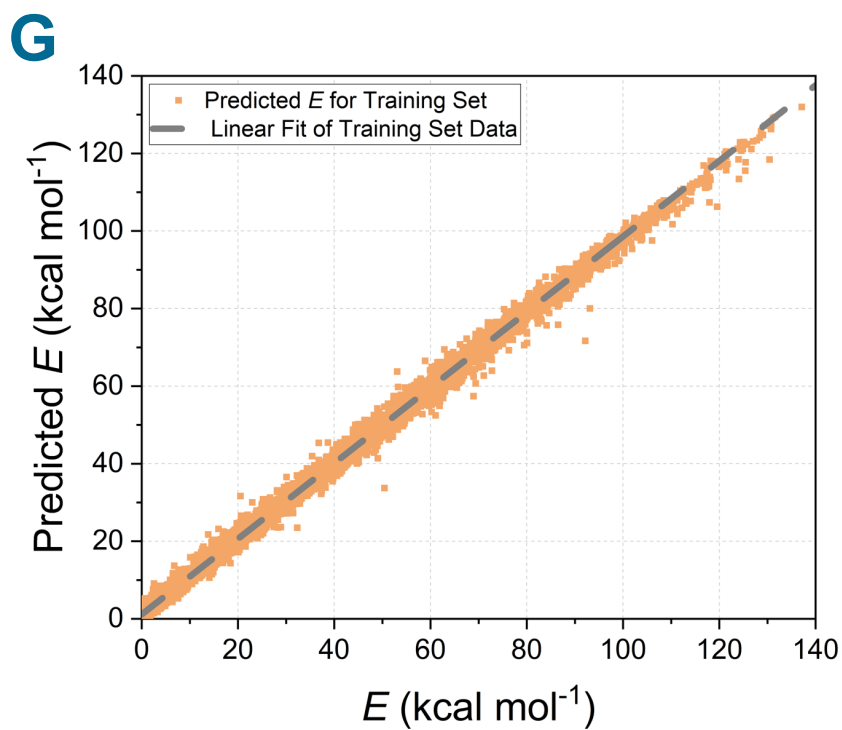


Figure 3-6. G: In-sample correlation of predicted and reference barriers for the KRR model DFT,DFT//DFT. **H:** Out-of-sample correlation (same method as G).

quantifies this observation: the MAE-to-RMSE ratio is $10.86/15.32 = 0.71$ with the Gaussian kernel and $9.17/12.79 = 0.72$ with the Laplacian kernel.

After applying an OLS model to fit the absolute error (y) between reaction energy barriers estimated by the best KRR and energy barriers from DFT of reference [117a,117b] versus the energy barriers from DFT (x), a linear function $y = 0.008x + 0.742$ was obtained. The coefficient 0.008 shows the positive correlation between the absolute error and reference reaction barriers.

As mentioned before, the most important reactions in the ISM and CSE are those with extremely low or zero barriers. Therefore, it is useful to assess the accuracy of ML models when the barrier is small. The full dataset has barriers ranging from 0 to 139 kcal mol⁻¹. We generated the subset of 5,276 reactions with barriers smaller than 40 kcal mol⁻¹ (derived from the full dataset). Some statistics about this subset (denoted D₂) and the full set (denoted D₁) are shown in Table 3-4.

Table 3-4. Statistical description of two datasets

Dataset	D ₁	D ₂
# of reactions	11,730	5,276
Smallest E (kcal mol ⁻¹)	0.00	0.00
Largest E (kcal mol ⁻¹)	138.85	40.00
Mean of E (kcal mol ⁻¹)	45.94	21.27
Standard Deviation (kcal mol ⁻¹)	26.85	11.32

Table 3-5. Performance statistics (RMSE, R^2 , MAE) of KRR models for sets D₁ and D₂ (the entire set minus the outliers) using the features calculated in six approaches. The mean and standard deviation over ten testing sets generated by 10-time Monte Carlo cross-validation are shown for each statistic

Approach of acquiring features (Hyperparameters)	D ₁ : RMSE (kcal mol ⁻¹)	D ₁ : R^2	D ₂ : RMSE (kcal mol ⁻¹)	D ₂ : MAE (kcal mol ⁻¹)	D ₂ : R^2
X,PM7//PM7 ($\alpha=0.05$, $\gamma=0.01$)	15.33 ± 0.38	0.6705 ± 0.02	7.95 ± 0.21	6.06 ± 0.22	0.5052 ± 0.03
X,DFT//DFT ($\alpha=0.05$, $\gamma=0.01$)	12.79 ± 0.23	0.7710 ± 0.01	6.59 ± 0.08	4.88 ± 0.10	0.6604 ± 0.01
Babel,PM7//PM7 ($\alpha=0.05$, $\gamma=0.01$)	14.06 ± 0.39	0.7238 ± 0.02	7.26 ± 0.25	5.28 ± 0.16	0.5872 ± 0.03
Babel,DFT//DFT ($\alpha=0.05$, $\gamma=0.01$)	12.49 ± 0.39	0.7821 ± 0.01	6.58 ± 0.25	4.68 ± 0.18	0.6604 ± 0.03
PM7,PM7//PM7 ($\alpha=0.05$, $\gamma=0.01$)	14.33 ± 0.55	0.7129 ± 0.02	7.30 ± 0.25	5.28 ± 0.19	0.5821 ± 0.03
DFT,DFT//DFT ($\alpha=0.05$, $\gamma=0.01$)	11.76 ± 0.31	0.8062 ± 0.01	6.02 ± 0.34	4.13 ± 0.19	0.7154 ± 0.03

The predictions of KRR models for the $E \leq 40$ kcal mol⁻¹ subset are shown beside those for the full dataset in Table 3-5. The RMSE on the subset is 7.3 kcal mol⁻¹ with Babel,PM7//PM7 inputs to the ML model and 6.6 kcal mol⁻¹ with Babel,DFT//DFT inputs. The corresponding MAEs are 5.3 kcal mol⁻¹ and 4.7 kcal mol⁻¹. The RMSE on the subset is 7.3 kcal mol⁻¹ with PM7,PM7//PM7 inputs to the ML model and 6.0 kcal mol⁻¹ with DFT,DFT//DFT inputs. The corresponding MAEs are 5.3 kcal mol⁻¹ and 4.1 kcal mol⁻¹. This level of accuracy will allow us to filter out the majority of hypothetical reactions that may take place in the ISM or CSE based on predicted energy barriers and a cut-off of roughly 5 kcal/mol or 10 kcal/mol.

3.6 Discussion and Conclusions

In this study, we presented ML regression models for the prediction of reaction energy barriers fitted to the DFT transition states data of Grambow *et al.* [117a,117b] for reactions involving up to 23 atoms of the elements H, C, N, O. Our best ML model, KRR, uses 300 features that fall into four categories: sorted eigenvalues of modified Coulomb matrices; features that depend on the electronegativity, hardness and interatomic distance of pairs of atoms; features that do not depend explicitly on nuclear positions, such as $\Delta_r H$, $\Delta_r E$, and $\Delta_r G$; and features that depend on bond positions and formal bond energies of relatively few atom pairs (primarily identifying broken or formed bonds during a reaction). Features of *chemical reactions* are expressed as differences between features of products P and Q and reactants A and B, (i.e., $x(\text{reaction}) = x(P) + x(Q) - x(A) - x(B)$), or features of supermolecules PQ and AB (molecules positioned and aligned in a way that prepares for the reaction, i.e., $x(\text{reaction}) = x(PQ) - x(AB)$). In the present work, features of the electronic structure (e.g., the HOMO-LUMO gap) of reactant and product molecules were not included because of their high computational cost (they are not available in the dataset of reference [117a,117b]). However, some electronic properties are readily available for 2,186 atoms and molecules [183] and databases of computed molecular properties will continue to grow. It would be interesting in the future to add features related to the HOMO-LUMO gap of molecules [184–196], atomic charges fitted to molecular electrostatic potentials [195,196], and atom-condensed Fukui functions [197]. These features might be able to improve the accuracy. The performance, on certain classes of reactions, of ML models with or without these features could bring new insight into reactivity principles like

the Hard and Soft Acids and Bases principle (HSAB) [185–189] and the Maximum Hardness Principle (MHP) [190–194].

We tried six ML methods: three variants of linear regression, two variants of KRR (with two different kernels), and FFNN (with three hidden layers). The KRR with the Laplacian kernel method worked best (please see Table 3-2 for details). It predicts out-of-sample energy barriers (testing set) with a MAE of 7.7 kcal mol⁻¹, RMSE of 11.8 kcal mol⁻¹, and a coefficient of determination $R^2 = 0.81$, for 11,409 reactions with barriers ranging from 0 to 139 kcal mol⁻¹ (even though it predicts in-sample barriers, i.e. training set, with very good accuracy, it is useless as mentioned before). The cost for training this KRR is modest (hours of computing for a data set of 11,409 reactions). The cost of using it to make one barrier prediction is less than a second (with only a two-core processor). This will allow the screening of very large sets of hypothetical chemical reactions for ISM and CSE-related applications.

The accuracy of the best method reported here is better than using only PM7-TS [198] for estimating energy barriers [55], which gave a MAE of 22.5 kcal mol⁻¹. It is comparable to that obtained by Rupp *et al.* with a similar KRR methodology for atomization energies [60]. Atomization energies cover a much bigger range than reaction barriers, which could make the former more difficult to fit. However, transition states show more structural diversity than molecules and are much harder to model. For reasons of computational cost and interpretability, we chose to not use any kind of transition state geometry estimate,

complex features or parametrization for its input. Instead, we chose to develop the features with simple interpretability and involving only several parameters.

Grambow, Pattanaik and Green [175] achieved excellent accuracy (MAE of 1.7 kcal/mol and RMSE of 3.4 kcal/mol when training a ML on a dataset of more than 20,000 training points; MAE of ~ 3.4 kcal/mol when training a ML on a dataset of roughly 10,300 training points, which is roughly equal to the number of our training points) on predicted energy barriers for a dataset similar to, and larger than, the one used here. The target barriers used for training ML were calculated with ω B97X-D3. This DFT method gives a RMSE of 2.3 kcal/mol. There are two ways to estimate the total error (RMSE) relative to the experiment: (1) the total error is equal to the sum of DFT error (2.3 kcal/mol) and ML error (3.4 kcal/mol from Grambow *et al.* [175]; 6.0 kcal/mol from us), $e_{tot} = e_{DFT} + e_{ML}$, for strongly correlated errors (i.e., DFT error is strongly correlated to ML error); (2) the total error is equal to the square root of sum of squares of each error, $e_{tot} = \sqrt{e_{DFT}^2 + e_{ML}^2}$, for uncorrelated errors (i.e., DFT error is not correlated to ML error). Hence, their actual error is between 4.1 and 5.7 kcal/mol, and ours is between 6.4 and 8.3 kcal/mol. Their Deep Learning model is based on a directed graph representation of molecules and is built on a molecular property prediction method (chemprop) [199]. It uses features of RDKit [176] including some that encode the atoms' hybridization and formal charge, and the presence of conjugated bonds, rings, aromaticity, and some fairly complicated features. By comparison, the model presented here is less accurate but simpler. We only use a three-dimensional representation of molecular geometries and relatively few parameters: atomic radii, electronegativities and hardness, and bond lengths and energies. Moreover, our model can be generalized to

molecules, clusters and solids with atoms outside the main group elements provided that a bond length-bond strength relation can be given for every pair of elements. It remains to be seen if adding more helpful features, such as HOMO or LUMO-related features, could improve the ML's accuracy without too much added complexity.

The best KRR model presented here gives energy barriers with a MAE of 9.9 kcal mol⁻¹ when using PM7 input geometries for calculating all features, 7.7 kcal mol⁻¹ when using DFT input geometries for calculating all features (see Table 3-3 for details). While encouraging, it is 13 times $k_B T$ at room temperature and too big for applications that require high accuracy. It may be adequate for exploratory or semi-quantitative work but not as a standalone predictor of chemical reactivity. The MAE of the models decreases to 5.3 kcal mol⁻¹ (PM7 inputs) and 4.1 kcal mol⁻¹ (DFT inputs) when they are fitted to reaction energy barriers smaller than 40 kcal mol⁻¹. With this level of accuracy, one can exclude many reactions unlikely to occur at low temperatures (15 to 25 kelvin, $k_B T \sim 0.04$ kcal mol⁻¹) in the ISM or CSE. Ultimately, theoretical predictions of reactivity in space will rely on highly accurate and costly quantum chemical methods, such as CCSD(T).

Chapter 4 Screening Reactions for Relevance to Prebiotic

Chemistry in the ISM and CSE

4.1 ISM and CSE Reaction Dataset

As described earlier, most molecules that have been observed in space have ten or fewer atoms, but the mechanism for the formation of bigger molecules is not well known. Chemical and biospheric evolution with a high reaction energy barrier is not likely to happen in the ISM and CSE areas given the low temperatures. Forecasting the most likely reactions in these places can be achieved by considering many combinations of ISM and CSE molecules as possible reactants and ruling out reactions with a high barrier. Bimolecular reactions between the molecular species identified in the ISM are natural targets for this kind of study. But there are thousands of possibilities and accurate calculations for all of them is currently impossible.

The strategy adopted in the present work starts by considering all reactions with reactant molecules A and B chosen among those of Table 4-1 [200–260] (see the literature [39] for the corresponding reference(s) for each molecule). This makes 2,628 combinations. The molecules in Table 4-1 [200–260] have all been observed in the ISM or CSE and contain no atom other than H, C, N and O.

Table 4-1. List of 72 molecules known, or believed, to be present in the ISM and CSE [200–260] that are in our molecule database [20,39] and their SMILES representation [163]

Molecule Formula	SMILES	Molecule Formula	SMILES
CO	[C-]#[O+]	CH ₃ NCO	O=C=NC
H ₂	[HH]	HOCH ₂ CN	OCC#N
N ₂	N#N	H ₃ CNH ₂	CN
H ₂ C	[CH2]	CH ₂ CHCN	N#CC=C
HCN	C#N	HCCCHNH	C#CC=N
CO ₂	C(=O)=O	CH ₃ CHO	O=CC
H ₂ O	O	H ₂ CHCOH	OC=C
N ₂ O	N#[N+][O-]	<i>c</i> -C ₂ H ₄ O	O1CC1
H ₂ CO	C=O	CH ₃ C ₂ H	CC#C
C ₂ H ₂	C#C	NH ₂ CH ₂ CN	NCC#N
HNCO	N=C=O	(NH ₂) ₂ CO	C(=O)(N)N
HCNO	[O-][N+]#C	CH ₂ CHCHO	O=CC=C
HCCN	N#C[CH+]	HCOOCH ₃	O=COC
CNCN	[C-]#[N+]C#N	CH ₃ COOH	CC(O)=O
HOCN	OC#N	CH ₃ CHNH	CC=N
NH ₃	N	HC ₃ H ₂ CN	C#CCC#N
CH ₄	C	H ₃ CC ₂ CN	CC#CC#N
H ₂ CNH	C=N	CH ₂ CCHCN	C=C=CC#N
H ₂ C ₂ O	C=C=O	C ₂ H ₃ NH ₂	C=CN
HNCNH	C(=N)=N	CH ₃ NHCHO	CNC=O
NH ₂ CN	N#CN	CH ₃ CONH ₂	O=C(N)C
HCC-NC	C#C[N+]#[C-]	CH ₃ OCH ₃	COC
HC(O)CN	C(=O)C#N	CH ₃ CH ₂ OH	OCC
HCOOH	C(=O)O	CH ₃ CH ₂ CN	CCC#N
HC ₃ N	C#CC#N	CH ₃ CHCH ₂	C=CC
CH ₃ OH	CO	CH ₃ C ₄ H	CC#CC#C
CH ₃ CN	CC#N	CH ₃ CHCH ₂ O	CC1CO1
C ₂ H ₄	C=C	CH ₃ CH ₂ CHO	CCC=O
HC ₂ CHO	C#CC=O	(CH ₃) ₂ CO	CC(=O)C
CH ₃ NC	[C-]#[N+]C	NH ₂ CH ₂ CH ₂ OH	NCCO
HC ₄ N	[H]C#C[C]C#N	CH ₃ COOCH ₃	O=C(OC)C
HCONH ₂	O=CN	C ₆ H ₆	c1ccccc1
<i>l</i> -H ₂ C ₄	C#CC#C	(CH ₃) ₂ CHCN	CC(C)C#N
E-HNCHCN	[H]/N=C(C#N)\[H]	C ₃ H ₇ CN	CCCC#N
CH ₂ CNH	C=C=N	C ₂ H ₅ NH ₂	CCN
<i>c</i> -H ₂ C ₃ O	C1=C(=O)=C1	C ₂ H ₅ OCH ₃	COCC

4.1.1 Reactant Side

When two ISM or CSE molecules collide, a transient intermediate may be formed first if the association process is barrierless. This intermediate may get stabilized by the emission of radiation and then may exist for a longer time [261]. In this work, our primary focus is on association reactions ($A + B \rightarrow P$) involving any two molecules among the above 72 molecules as reactants A and B. This brings us $72 \times 73 \div 2 = 2,628$ possible combinations, i.e., 2,628 possible reactant sides. In this way, we could have 26,280 conceivable reactions even if one possible reactant side matches only ten kinds of products. As described in previous chapters, with that many reactions, energy barriers cannot be obtained by quantum chemistry, they must be estimated by ML.

4.1.2 Product Side

Each possible reactant side specifies a clear composition of different atoms (C, H, O, and N). According to each composition of possible reactant side, our possible product side can be generated by matching the composition of atoms. Our matching procedure is to search for the molecules possessing the same composition as possible reactant side in our molecule database (10,306 organic small molecules in it) [20,117]. Eventually, 136,081 combinations of reactants and products were generated.

4.1.3 Enriched ISM and CSE Reaction Dataset

So far, a huge dataset of first-step possible reactions between ISM and CSE molecules, 136,081 reactions, has been created. Since our ML model was developed with the reaction

energy barrier as a function of features of reactants and products, the next step is to enrich our dataset with molecular geometry information for the following feature calculations.

In previous chapters, the geometries of all possible products have been generated by converting its SMILES representation to the 3D structure by Open Babel, Avogadro software, and the online tool of NovoPro. To ensure consistency, we first collect SMILES strings of all molecules in our possible reactant sides, i.e., 72 molecules in Table 4-1, and then transfer their SMILES strings to 3D structures with the same procedure. After molecular geometry information of the whole reaction dataset is collected, features then can be calculated, and all input data can be prepared for ML [20].

4.2 Procedure of Screening

For this project, we do two screenings: first we eliminate reactions with estimated energy barrier larger than a threshold E_0 , and then we eliminate reactions in which the product's chemical composition or functional groups are very different from those of typical small biomolecules, such as amino acids. With this procedure, we hope to discover reactions taking place in the ISM and CSE that are relevant to theories about the origins of life.

One of the possible mechanisms for reactions in the ISM is gas-phase radiative association [261]: $A + B \rightleftharpoons AB^* \rightarrow AB + h\nu$. The temperature in the ISM varies roughly between 10 and 100 K. But it can exceed 100 K in “protostellar cores”. The existence of interstellar shocks may lead to the temporary existence of radiation in the ISM and CSE [73a,74]. It then could increase the temperature to 100~150 K for areas with such shocks. Therefore, reactions with low or zero energy barriers are much more likely to occur than

those with high barriers. The formation of a relatively stable and long-lived reaction complex AB^* is also favourable [261].

Another mechanism believed to be important for ISM chemistry is the reaction of A and B at the surface of dust grains. The nature of these grains, and the mechanisms of reactions on them, are not well known and are difficult to model or simulate theoretically. For simplicity, we assume that reactions with the smallest gas-phase energy barriers are also among those most likely to occur at the surface of grains.

4.2.1 Screening based on Energy Barrier

Reaction energy barriers were estimated for 136,081 hypothetical reactions between ISM or CSE molecules. This took only twenty minutes of computing time. Although the KRR model was trained on elementary reactions, it is used here to predict the barrier of the rate limiting step of those reactions that may be elementary or complex. Despite this limitation, the KRR estimates provide a useful ranking of reactions in increasing order of estimated barriers. 98% of the reactions were ruled out on the basis of their estimated barriers.

4.2.2 Screening by Biological Relevance

Not all barrierless reactions are relevant to the biochemical or biogenic evolution even though they may play a significant role in other aspects. The primary task of this work is to discover more likely ISM and CSE molecules, especially biologically important molecules. To do so, we define the criteria for evaluating the relevance of the possible products for the emergence of life [20].

This second screening will be applied to the reactions that passed the energy barrier criterion screening. The absence of unsaturation implies an inert molecule which can hardly be a building block for larger prebiotic molecules. On the other hand, molecules with a large degree of unsaturation are often very unstable and may not exist long enough to contribute to chemical evolution in space. Moreover, the number of non-Hydrogen atoms in a molecule also matters in comparisons to the organic molecules that form the basis of life. The elements: C, H, O, and N, are four of the most abundant and commonest elements in the universe [29–31] and are fundamental in investigating life-related biomolecules.

We use two criteria: (i) there should be at least one atom of each element (H, C, N, O); (ii) the degree of unsaturation (n_u) of the product divided by the number of non-H atoms (n_x), n_u/n_x , should fall within a range of [0.05, 0.50] so that it is similar to what is found in amino acids [262–265] (see Table 4-2), mono and disaccharides [266] (see Table 4-3), alpha-hydroxy acids [267], and many drug-like molecules [268]. Only a few reactions passed both rounds of screening.

Table 4-2. Standard amino acids, their degree of unsaturation n_u , number of non-hydrogen atoms n_x , and ratio n_u/n_x

Name of Amino Acid	Abbreviation Symbol	# of Double Bond	# of Aromatic Ring	n_u	n_x	n_u/n_x
Alanine	Ala	1	0	1	6	0.17
Arginine	Arg	2	0	1.5	12	0.13
Asparagine	Asn	2	0	2	9	0.22
Aspartate	Asp	2	0	2.5	9	0.28
Cysteine	Cys	1	0	1	7	0.14
Glutamate	Glu	2	0	2.5	10	0.25
Glutamine	Gln	2	0	2	10	0.20
Glycine	Gly	1	0	1	5	0.20
Histidine	His	3	1	4	11	0.37
Isoleucine	Ile	1	0	1	9	0.11
Leucine	Leu	1	0	1	9	0.11
Lysine	Lys	1	0	0.5	10	0.05
Methionine	Met	1	0	1	9	0.11
Phenylalanine	Phe	4	1	5	12	0.42
Proline	Pro	1	0	2	8	0.25
Serine	Ser	1	0	1	7	0.14
Threonine	Thr	1	0	1	8	0.13
Tryptophan	Trp	5	1	7.5	15	0.50
Tyrosine	Tyr	4	1	5	13	0.38
Valine	Val	1	0	1	8	0.13
Range						0.05~ 0.50

Table 4-3. Representative mono and disaccharides, their degree of unsaturation n_u , number of non-hydrogen atoms n_x , and ratio n_u/n_x

Class	Name of Sugar	# of Double Bond	# of Aromatic Ring	n_u	n_x	n_u/n_x
Monosaccharide	Glucose	0	0	1	12	0.08
	Galactose	0	0	1	12	0.08
	Fructose	0	0	1	12	0.08
Disaccharide	Maltose (two molecules of glucose)	0	0	2	23	0.09
	Lactose (glucose + galactose)	0	0	2	23	0.09
	Sucrose (glucose + fructose)	0	0	2	23	0.09
Range					0.08~0.09	

4.3 Discussion and Conclusions

After our two screening procedures, five reactions stand out because they all involve the reaction of methylene (CH_2) with fulminic acid (HCNO) or its isomer isocyanic acid (HNCO). These five reactions and their computational studies with quantum chemistry calculations will be shown and discussed in the next chapter. Two more reactions are selected for quantum chemistry calculations because they have small ML-predicted barriers and are correlated with biologically important glycine and alanine somehow. One is $\text{CH}_2 + \text{CH}_3\text{C}(=\text{O})\text{NH}_2$ producing alaninal. The other one is $\text{H}_2\text{O} + \text{NHCHCHO}$ producing glycine. The seven products of reactions exhibit an interesting variety of functional groups for such a small set. Computational studies based on these seven reactions with a sophisticated level of theory are presented in Chapter 5.

Chapter 5 The Reactions of Methylene with HCNO and HNCO

5.1 Introduction

The discovery of molecules in space and the study of their chemical reactivity is of interest for their implications on the possible origins of life. The increasing number of organic molecules discovered through astronomical observations [32] suggests that prebiotic molecules may have formed in space and reached early Earth and other planets [33–35]. Until now, the spectroscopic signatures of more than 250 molecular species have been detected in the ISM and CSE [32,36–38]. It is not a simple thing to establish the evidence for a molecular species in the ISM. It is therefore plausible that many more molecules present in the ISM and CSE are yet to be discovered. Recent observation of spectral lines assigned to the amino acid tryptophan in the interstellar gas of star-forming regions [269] points to the possibilities of relatively complex and biologically important molecules. Hence, computational and theoretical studies are needed to figure out such possibility for providing directions to specify some detections.

Association reactions among known ISM and CSE molecules may take place in the gas phase or at the surface of carbonaceous grains [261]. These grains could be covered with ice and may catalyze some reactions. Considering the low temperatures in the ISM (typically between 10 and 150 K) [40,41], it is likely that only reactions with a very low (or zero) energy barrier occur with a significant rate. Computational investigations can guide the discovery of molecules and reactions in space. With quantum chemical techniques, one can locate transition states and compute energy barriers with an accuracy on the order of 1

kcal/mol [8,9]. Chemical reactions between those identified interstellar species (see Chapter 4) are targets for this part of our study.

In 2023, Panda, Chiranjibi, Awasthi, Ghosal, and Anoop [270] studied the reaction mechanisms of $\text{HCN} + \text{H}_2\text{CO}$ and $\text{HNC} + \text{H}_2\text{CO}$ by DFT and CC methods and found barriers of reactions of 15 to 41 kcal/mol giving oxiran-2-ylazanide and isocyanomethanol as the main products. Several prior experimental and theoretical studies of reactions of radicals with HCNO have been conducted [271–279]. Theoretical studies of the reaction between NH and HCNO have been performed by a combination of DFT and quadratic configuration interaction (QCI) methods giving a reaction barrier of 2.9 kcal/mol for the kinetically favoured products, HCN and HON [271]. Another study reported DFT, CC and QCI calculations predicting a zero barrier to the most feasible products, HCN and HNO [272]. An experimental kinetics study of $\text{OH} + \text{HCNO}$ reaction was conducted by Feng, Meyer, and Herschberger [275], and this reaction produces two major kinds of products $\text{CO} + \text{H}_2\text{NO}$ and $\text{HNO} + \text{HCO}$. A computational study of the same reaction by DFT and CC methods gave $\text{CO} + \text{NHOH}$ as the main products with a zero barrier [274]. The product channels of $\text{CH}_3 + \text{HCNO}$ have been theoretically studied by DFT and CC methods: the favoured route produced *Z* and *E*- $\text{HC}(\text{CH}_3)\text{NO}$ with a reaction barrier of only 4 kcal/mol [276]. There have been several experimental and theoretical studies of reactions of radicals and HNCO [280–284] as well, including reactions of $\text{C}_3\text{H}_3 + \text{HNCO}$ [280], $\text{OH} + \text{HNCO}$ [281], and $\text{CH}_2\text{CH} + \text{HNCO}$ [283]. Moreover, a theoretical study on the hydrogen transfer path of $\text{CH}_2 + \text{HNCO}$ reaction has also been reported in the literature [284].

Redondo *et al.* recently published high-level theoretical spectroscopic constants for hydroxy-azirine [285] and imine acetaldehyde [286]. These molecules were proposed as potential interstellar prebiotic molecules, but searches for their spectroscopic signature in the ISM have failed so far. Fourré *et al.* [287] suggested using the minimum energy principle (MEP) for screening potential ISM molecules. We have the same goal. Interestingly, the screening of reactions with estimated energy barriers (present work) produced some of the same molecules proposed in the reference [287], which is encouraging.

5.2 Computational Details

As mentioned in Chapter 2, all quantum chemistry calculations in our research project were completed with the Gaussian 16 software [89] using GaussView [90] and Avogadro [91,92] as interfaces. The geometries of reactants, products, and intermediates of reactions were optimized by DFT with the ω B97XD method for exchange-correlation and dispersion corrections [21–24] (this DFT gives a RMSE of 2.34 kcal/mol for estimating reaction energy barriers [5]) and a def2-TZVP basis set [25–27]. The search for transition states was also done at the ω B97XD/def2-TZVP level of theory [21–27]. All stationary points of the potential energy surface were characterized by normal mode analysis, those with one imaginary frequency being TS. The intrinsic reaction coordinate, IRC [58,59], originating from every TS was followed to find the corresponding reactant and product (confirming that the TS is the real one, i.e., the desired one for a specific reaction), and all IRC calculations are done at the same level of theory, ω B97XD/def2-TZVP. The energy profiles

in Figures 5-1, 5-3, 5-5, 5-7, 5-9, 5-11 show sequences of minima (reactant, intermediate, or product) and the TS connecting each pair of minima. Single point (fixed geometry) calculations were performed with CCSD(T) (coupled cluster theory with single and double excitations and perturbative treatment of triple excitations) with a big basis set, Aug-cc-pVQZ [15–19,111–113] at the optimized minima and TS structures to get more accurate molecular energies and then get more accurate energy barriers and energies of reactions. Unless stated otherwise, the energies reported in this chapter are differences between zero kelvin energies at the CCSD(T)/Aug-cc-pVQZ [15–19,111–113] level with zero-point energy corrections and optimized geometries by ω B97XD/def2-TZVP, a method denoted CCSD(T)/Aug-cc-pVQZ// ω B97XD/def2-TZVP.

The five reactions of Scheme 5-1 and Scheme 5-2 were examined on the triplet and singlet potential energy surfaces. The two reactions of Scheme 5-3 and Scheme 5-4 were examined only on the singlet potential energy surface.

The ground states of HCNO, HNCO, $\text{CH}_3\text{C}(=\text{O})\text{NH}_2$, H_2O , and NHCHCHO are all singlet. The ground state of CH_2 is triplet, with the lowest excited singlet state being 9.00 kcal/mol higher [288]. The CH_2 singlet excited state energy is calculated as +9.1 kcal/mol, very close to the experimental value.

For the molecular species symbols in the following contents, RC stands for reactant complex, IM for reaction intermediate, TS for transition state, P for product, the superscripts 1 and 3 represent singlet and triplet species, and subscripts are arbitrarily assigned to the structures (depicted in Figures 5-2, 5-4, 5-6, 5-8, 5-10, 5-12) in the order in which they

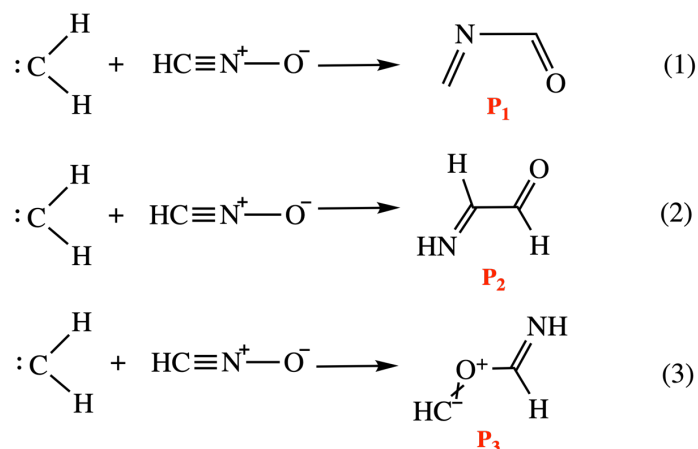
appear in Figures 5-1, 5-3, 5-5, 5-7, 5-9, 5-11. In a few cases, following the IRC from different transition states produced slightly different reaction complex geometries with nearly identical energies. For such cases, only one version of each reaction complex is shown (in Figures 5-2, 5-4, 5-6, 5-8, 5-10, 5-12).

5.3 $^1\text{CH}_2/^3\text{CH}_2 + \text{HCNO}$

Three reactions of methylene (CH_2) with fulminic acid (HCNO) are shown in Scheme 5-1, leading to three products (P_1 , P_2 , and P_3). The computational study on reactions of singlet CH_2 and HCNO is reported first (Figures 5-1, and 5-2), which is followed by the study on reactions of triplet CH_2 and HCNO (Figures 5-3, and 5-4).

The products of reactions in Scheme 5-1 are *N*-methyleneformamide, imine acetaldehyde, and ((λ^3 -methylene)- λ^3 -oxidaneyl)methanimine, respectively. None of them has been observed in the ISM or CSE [36,287].

In the next chapter, the structural and spectroscopic computational studies of conformers of *N*-methyleneformamide and imine acetaldehyde molecules will be presented and discussed. This kind of procedure of study allows us to investigate potential interstellar molecules or their most possible conformers more deeply.



Scheme 5-1. The three reactions of CH₂ with HCNO and their products.

5.3.1 Singlet State Energy Profile for CH₂ + HCNO

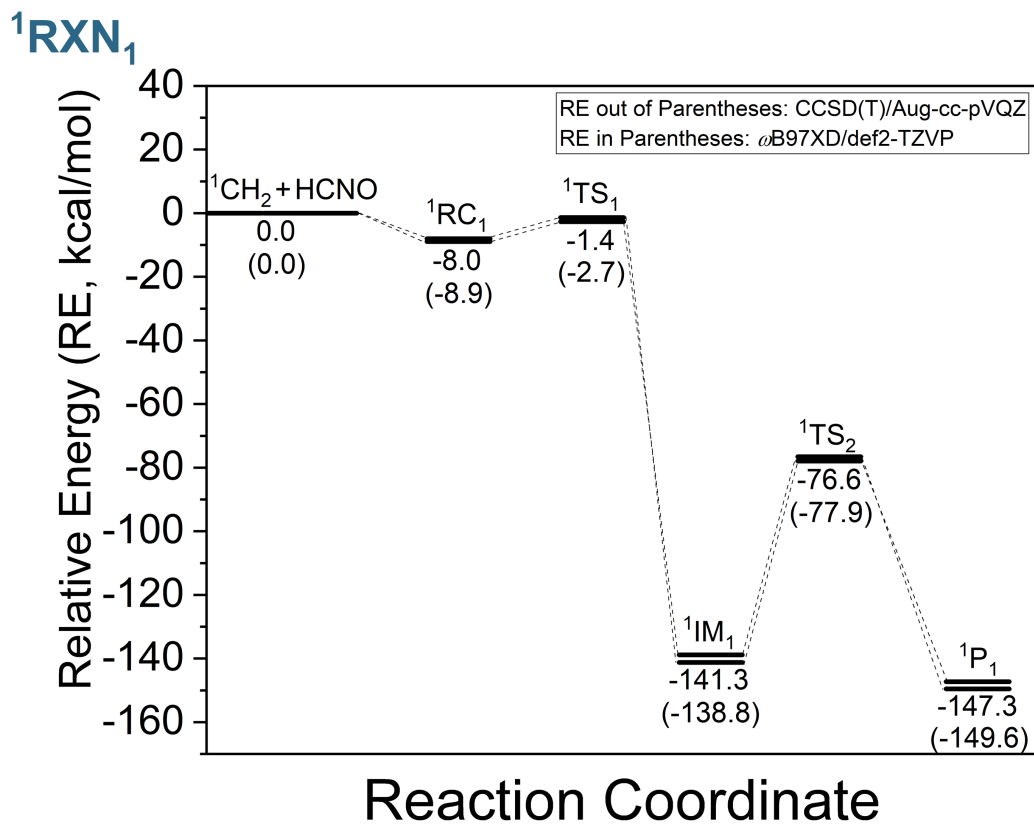


Figure 5-1. Singlet potential energy diagrams showing the energies of reactant complexes (RCs), intermediates of reaction (IMs), transition states (TSs) and products (Ps) of reactions between ¹CH₂ and HCNO. The relative energies calculated by CCSD(T) and DFT are shown, with DFT in parentheses.

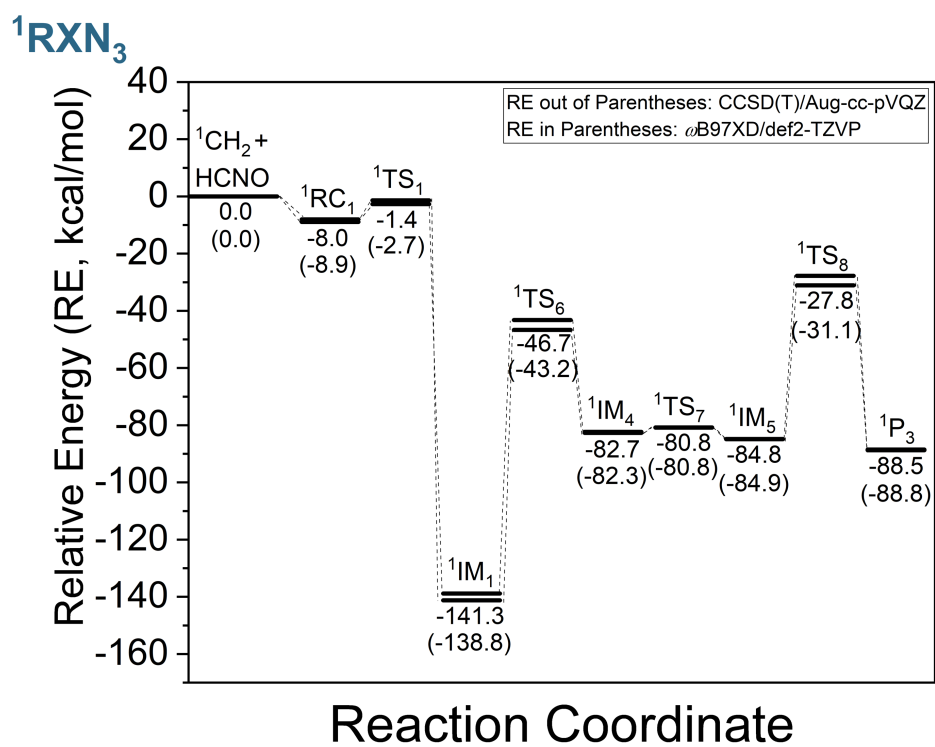
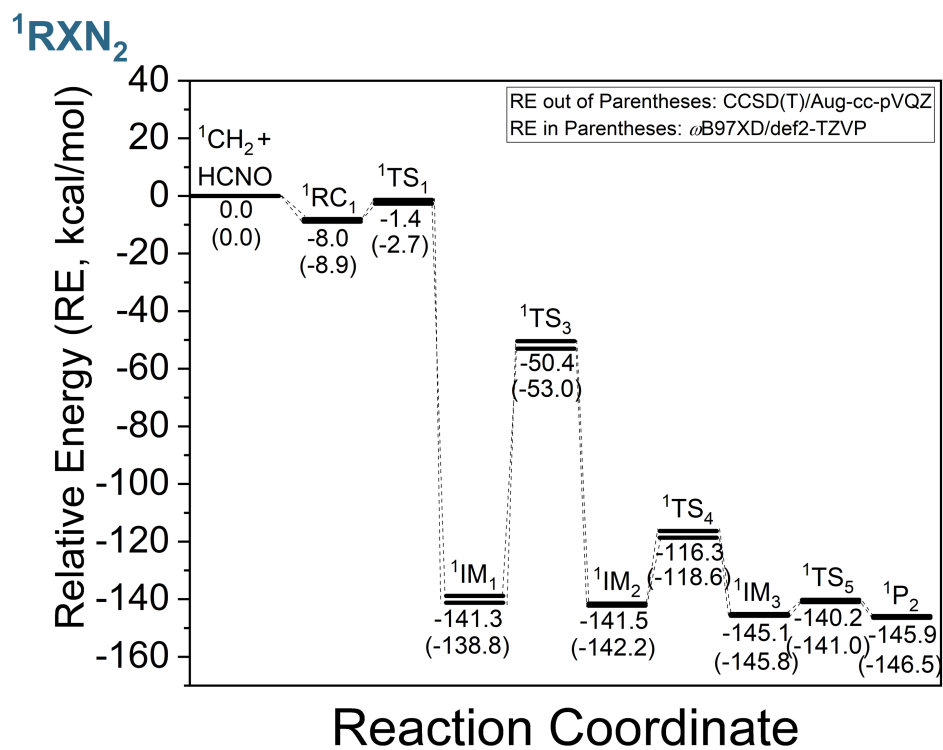


Figure 5-1. (Continued).

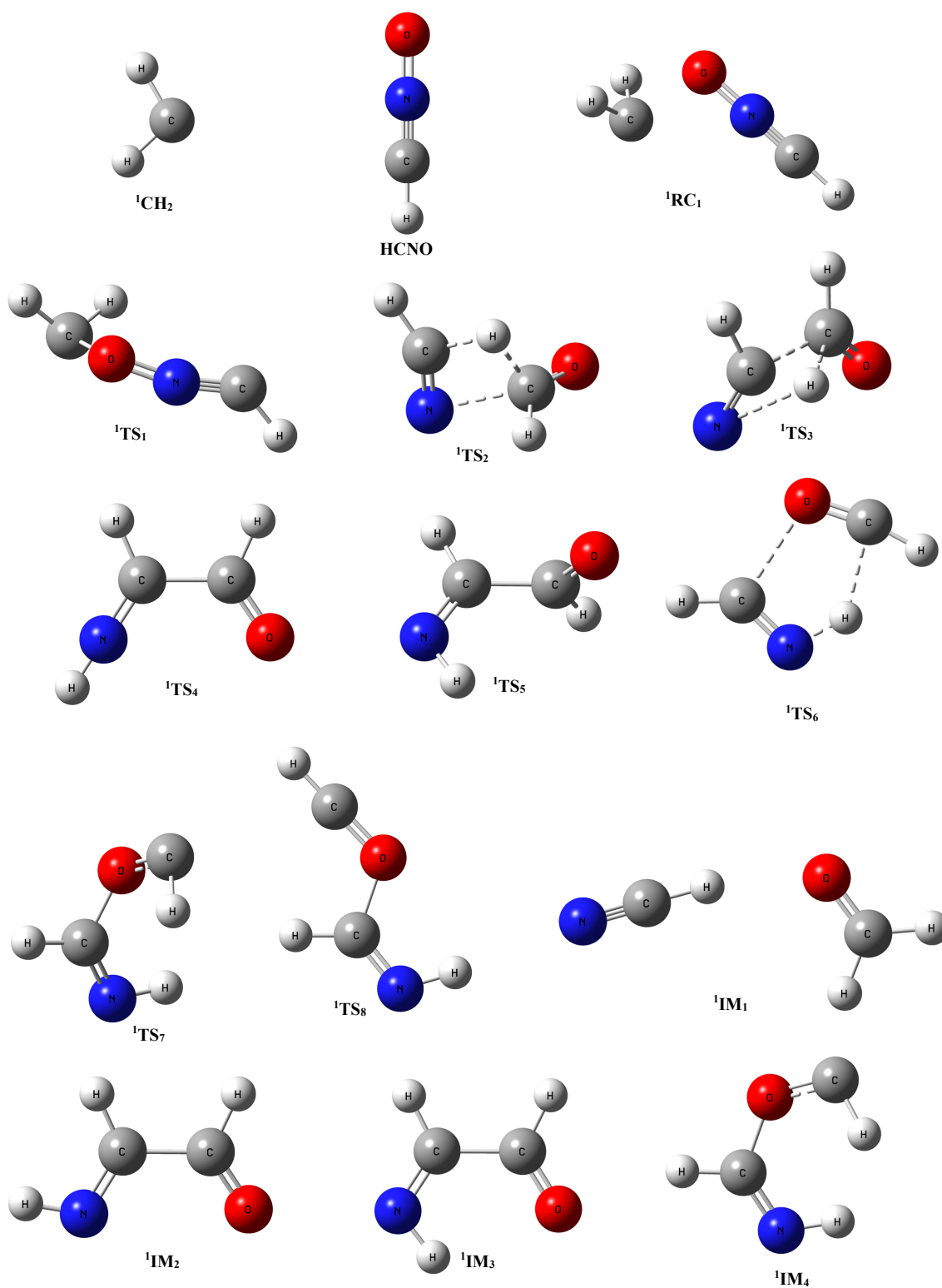


Figure 5-2. Optimized geometries of singlet reactants, RCs, TSs, IMs, and products of reactions (1), (2), and (3) at ω B97XD/def2-TZVP level of theory. Cartesian coordinates of all studied species shown here are listed in Appendix A.

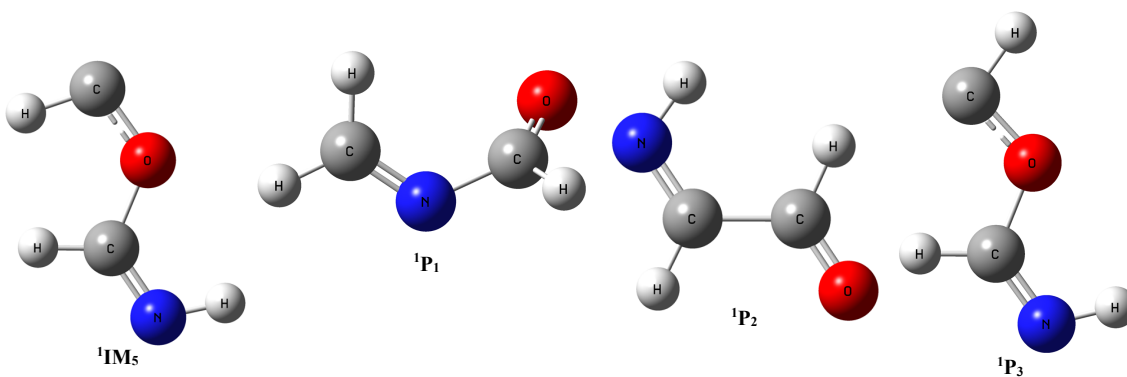


Figure 5-2. (Continued).

5.3.2 Triplet State Energy Profile for $\text{CH}_2 + \text{HCNO}$

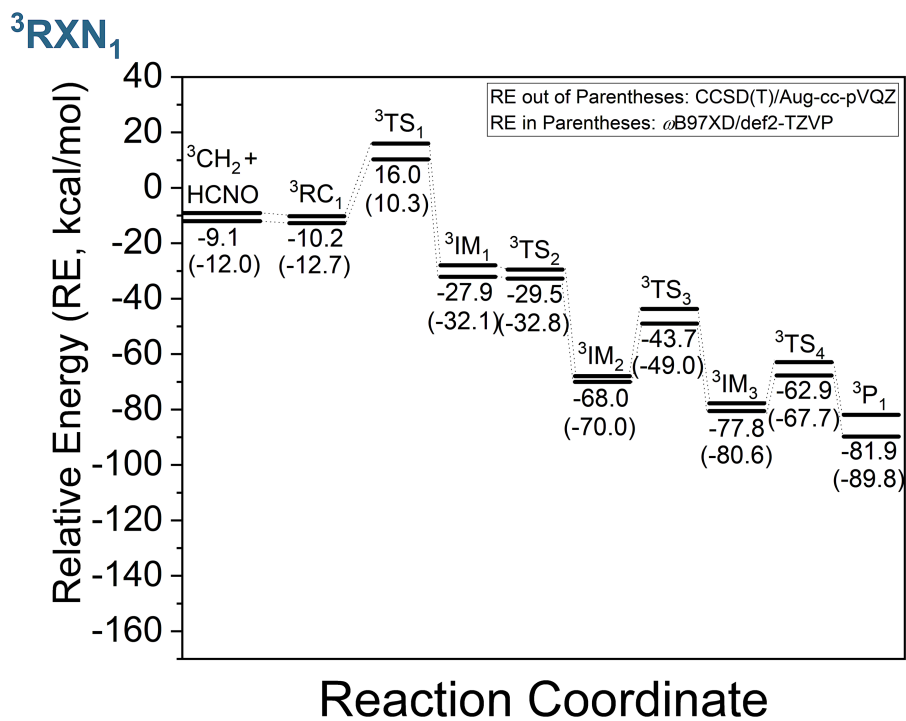


Figure 5-3. Triplet potential energy diagrams showing the energies of reactant complexes (RCs), intermediates of reaction (IMs), transition states (TSs) and products (Ps) of reactions between $^3\text{CH}_2$ and HCNO. The relative energies calculated by CCSD(T) and DFT are shown, with DFT in parentheses.

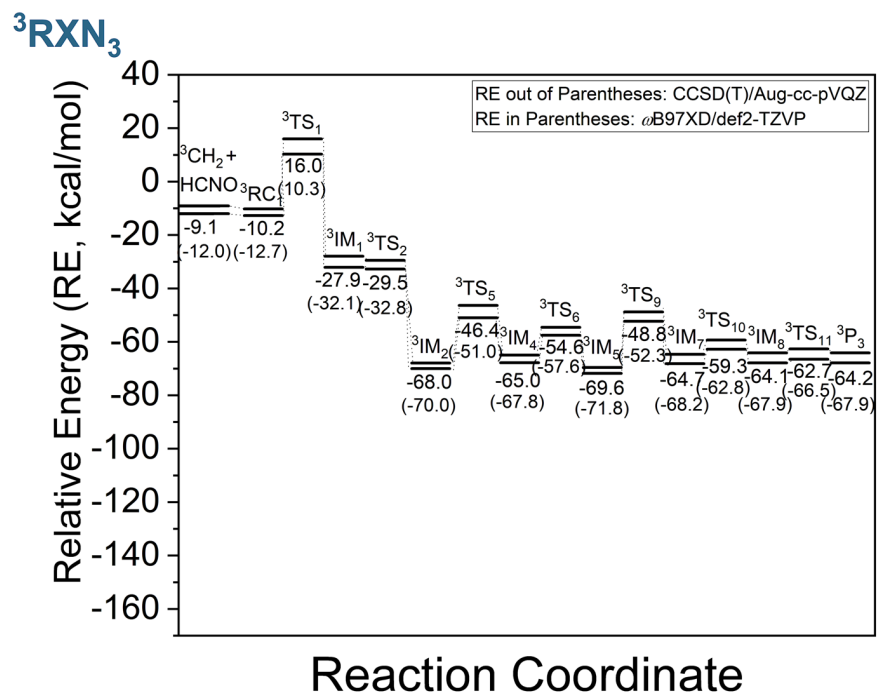
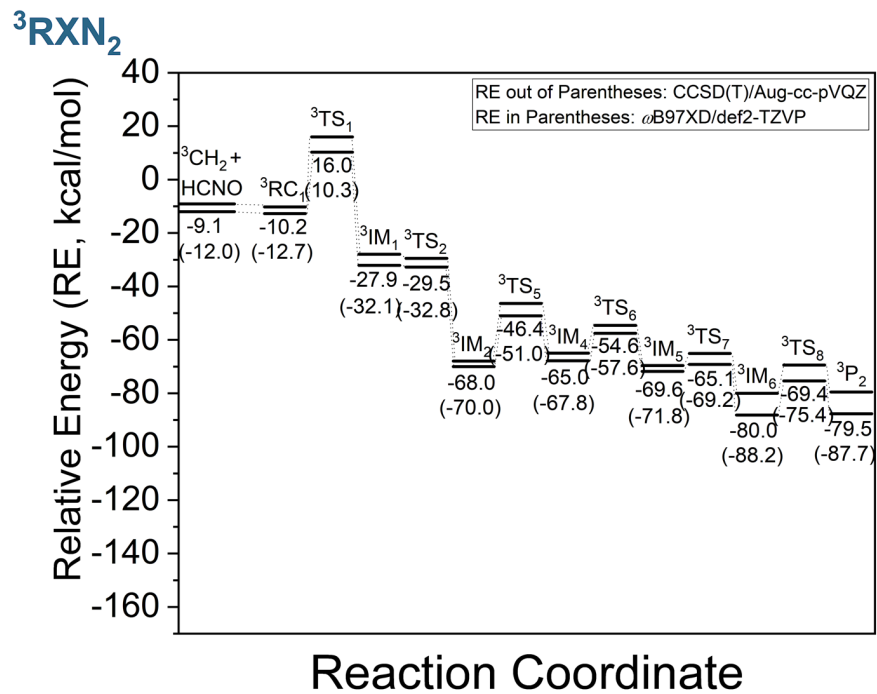


Figure 5-3. (Continued).

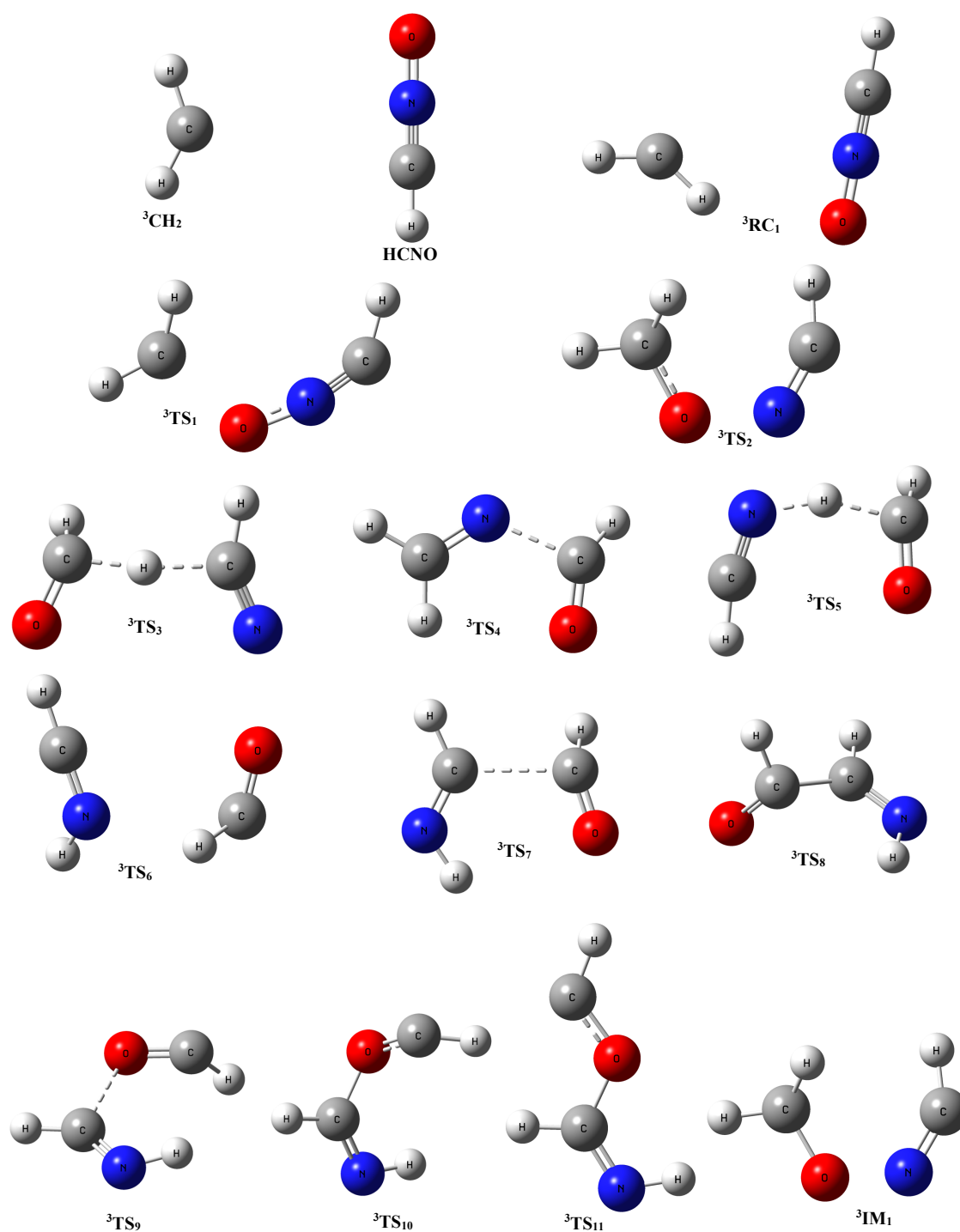


Figure 5-4. Optimized geometries of triplet state molecular species of reactions (1), (2), and (3) at ω B97XD/def2-TZVP level of theory. Cartesian coordinates of all studied species shown here are listed in Appendix A.

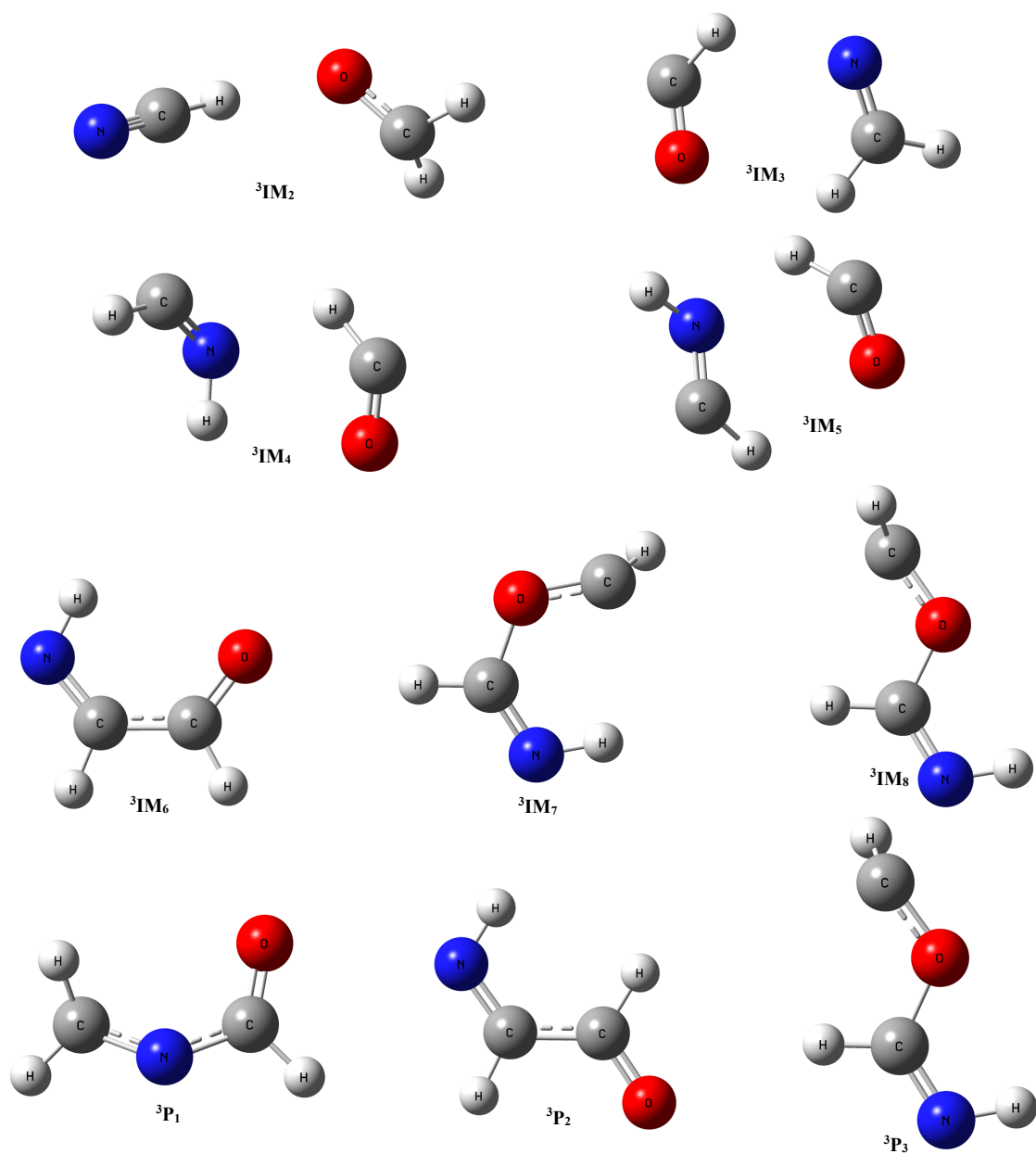


Figure 5-4. (Continued).

5.3.3 Reaction Paths for CH₂ + HCNO

Total energies including zero-point corrections (Es) for all studied molecular species are listed in Appendix B along with their relative energies where the zero of energy is assigned to singlet CH₂ plus singlet HCNO.

The singlet and triplet optimized geometries of reactants, reactant complexes (RCs), intermediates (IMs), transition states (TSs), and products (Ps) in reactions (1), (2), and (3) are shown in Figures 5-2 (in singlet) and 5-4 (in triplet) and the cartesian coordinates can be found in Appendix A. Some relative energies have numerical errors of approximately 2 kcal/mol. The energy of ³IM₁ is found 1.6 kcal/mol higher than ³TS₂ even though IRC following connects them while the energy of ³IM₁, without zero-point energy, is 0.22 kcal/mol lower than the ³TS₂. However, the zero-point correction pushes ³IM₁ 1.6 kcal/mol higher than ³TS₂.

Reactions k ($k = 1, 2, 3$) on the singlet and triplet energy surfaces are represented by ¹RXN _{k} and ³RXN _{k} . Their energy profiles are shown in Figures 5-1 (singlet) and 5-3 (triplet). In reaction ¹RXN₁, singlet CH₂ and HCNO form a stable reactant complex, then CH₂ picks up the oxygen atom of HCNO in a barrierless process (see ¹TS₁ in Figure 5-2 and ¹RXN₁ in Figure 5-1). This leads to formation of ¹IM₁, a complex of hydrogen cyanide (HCN) and formaldehyde (H₂CO) both of which were identified in the ISM or CSE [200,205,206,210]. The complex of HCN and H₂CO can go on to react in many ways, see ¹IM₁ and the transition states ¹TS₂, ¹TS₃, and ¹TS₆ in ¹RXN₁, ¹RXN₂, and ¹RXN₃, respectively. These transition states are well below the ¹(CH₂ + HCNO) separation limit but well above HCN

+ H₂CO (by 65 kcal/mol, 91 kcal/mol and 95 kcal/mol). The transition state ¹TS₁ is above triplet CH₂ plus HCNO by only 7.6 kcal/mol (9.0 – 1.4 = 7.6).

One conclusion that we draw from energy profiles ¹RXN₁, ¹RXN₂, and ¹RXN₃ is that singlet CH₂, if it is present in the ISM, can react with HCNO without barrier. For reaction (1), the thermodynamically favoured product of this reaction would be CH₂NCHO (*N*-methyleneformamide, ¹P₁), at –147.3 kcal/mol relative to reactants. Relevant computational studies of *N*-methyleneformamide have reported vibrational spectra [289] and relative energies of conformers [290]. In the literature [290], it has been suggested that the *gauche*-conformation of *N*-methyleneformamide is the most stable one. The conformer of *N*-methyleneformamide that we obtained here is the *gauche*-conformation, which has the lowest energy among all conformers [290]. The product of reaction (2), the conformer of imine acetaldehyde NHCHCHO, ¹P₂ has a slightly higher energy, –145.9 kcal/mol. Some of the intermediates of reactions are also very close in energy: the HCN-H₂CO complex ¹IM₁ (–141.3 kcal/mol), and two other conformers of imine acetaldehyde, ¹IM₂ (–141.5 kcal/mol) and ¹IM₃ (–145.1 kcal/mol).

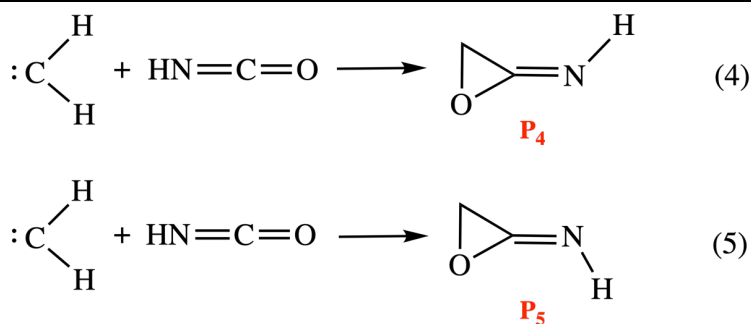
Nevertheless, it is much more likely that, in the ISM and CSE, CH₂ is in its triplet ground state. Then, the formation of CH₂NCHO (*N*-methyleneformamide, ¹P₁) would require approximately 6 to 9 kcal/mol excess energy and triplet-to-singlet surface crossing. Crossing might occur at some configurations of the CH₂ + HCNO reactant complex at an energetic cost of 9 kcal/mol (at large CH₂-HCNO separation) or maybe less, at some hypothetical favourable configuration, but we were unable to locate such a configuration.

The thermodynamically favoured triplet state product is 3P_1 (see 3RXN_1 in Figure 5-3) with an energy of -82 kcal/mol (-73 kcal/mol relative to the ground state when considering that singlet CH_2 is 9 kcal/mol higher than triplet CH_2 , i.e., $-82 + 9 = -73$). It is 65 kcal/mol higher than singlet *N*-methyleneformamide.

5.4 Reactions of CH_2 with HNCO

Two reactions of methylene (CH_2) with isocyanic acid (HNCO) are shown in Scheme 5-2, leading to two products (P_4 and P_5). The computational study on reactions of singlet CH_2 and HNCO is reported first (Figures 5-5, and 5-6), which is followed by the study on reactions of triplet CH_2 and HNCO (Figures 5-7, and 5-8).

The products of reactions in Scheme 5-2 are two conformers of oxiran-2-ylazanide. None of them has been observed in the ISM or CSE [36,287].



Scheme 5-2. The two reactions of CH_2 with HNCO and their products.

5.4.1 Energy Profile of Singlet CH_2 + HNCO

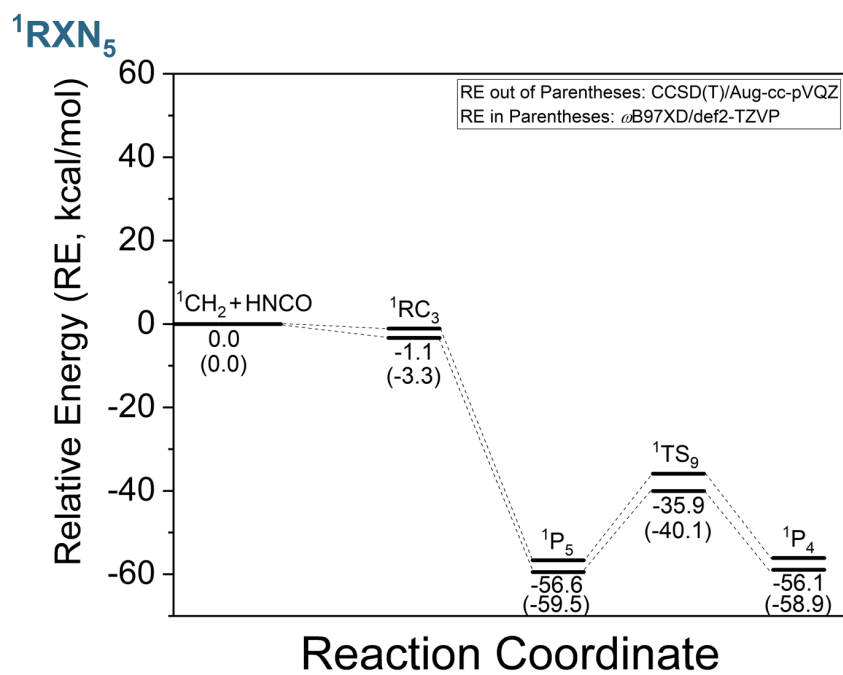
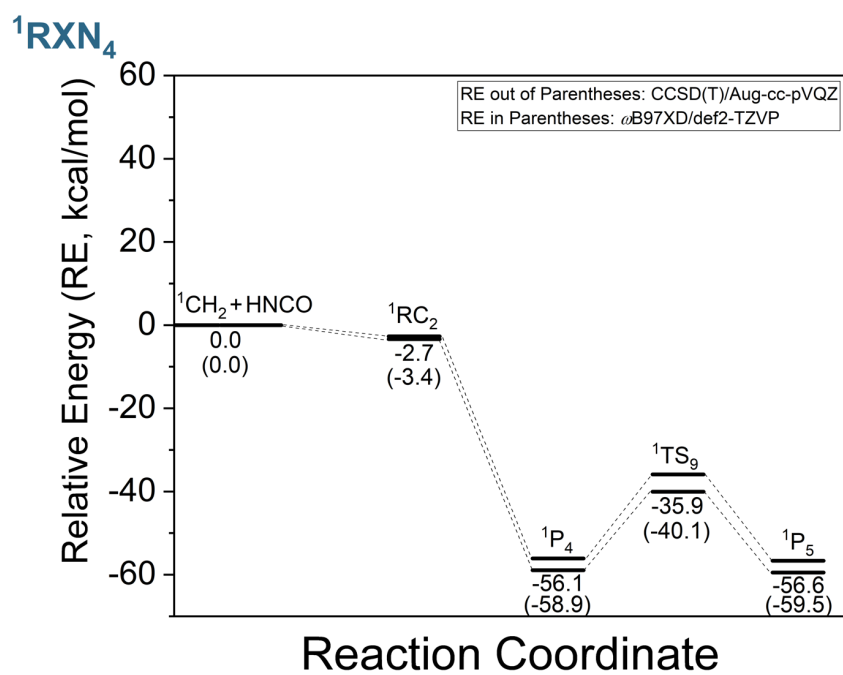


Figure 5-5. Singlet potential energy diagrams of reactions between ¹CH₂ and HNCO. The relative energies calculated by CCSD(T) and DFT are shown, with DFT in parentheses.

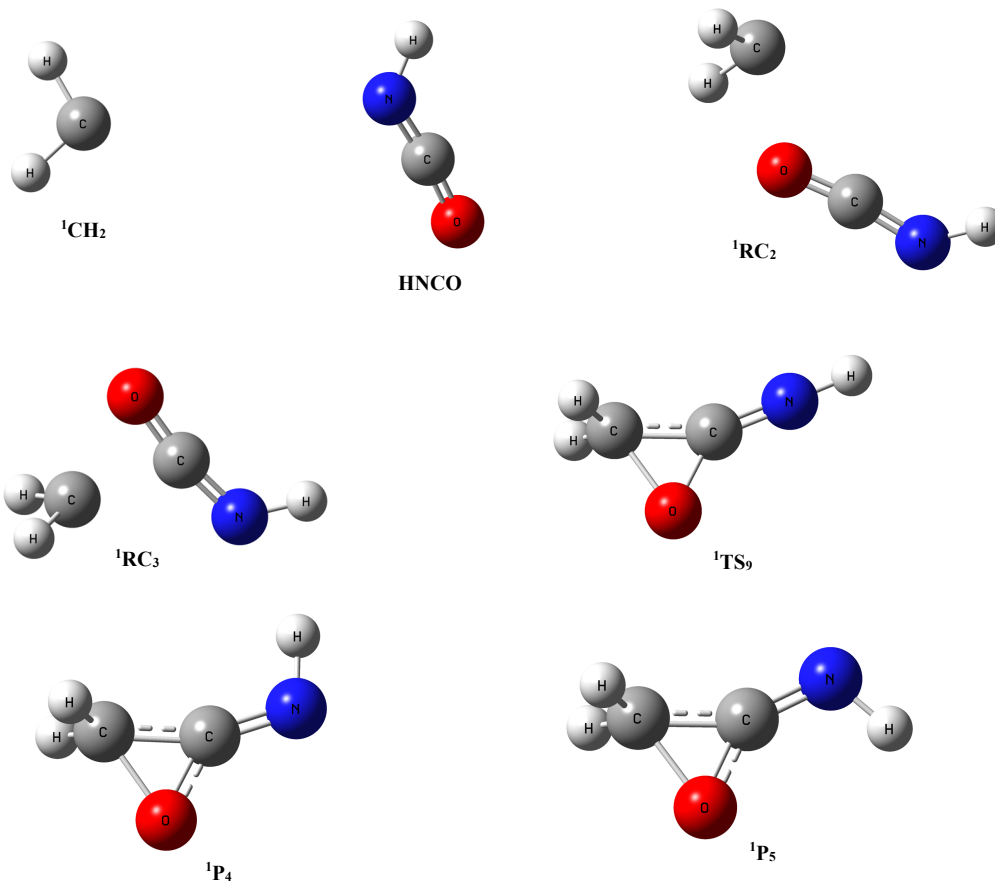


Figure 5-6. Optimized geometries of singlet reactants, RCs ($^1\text{RC}_2$ and $^1\text{RC}_3$ could not be fully optimized), TSs, IMs, and products of reactions (4) and (5) at ω B97XD/def2-TZVP level of theory. Cartesian coordinates of all studied species shown here are listed in Appendix A.

5.4.2 Energy Profile of Triplet $\text{CH}_2 + \text{HNCO}$

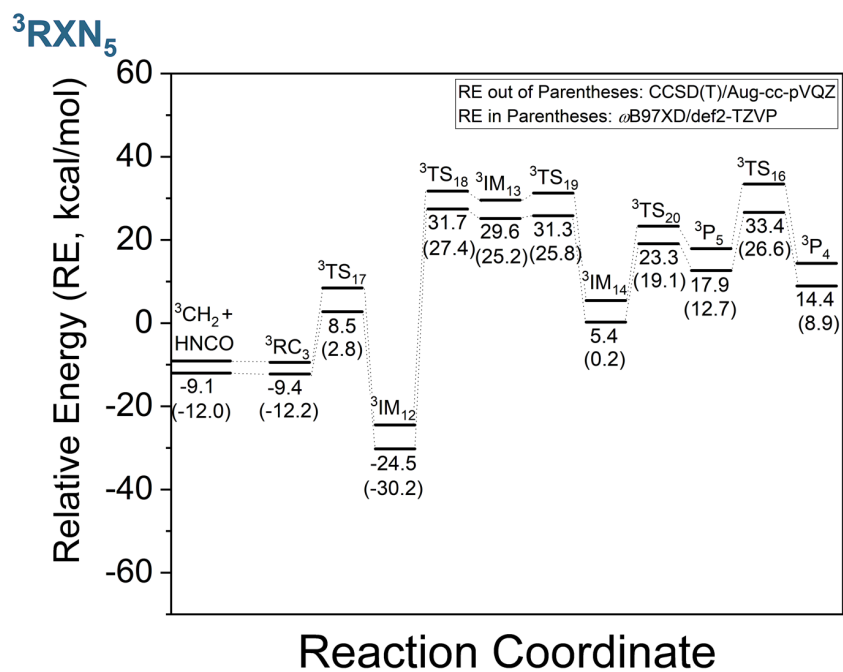
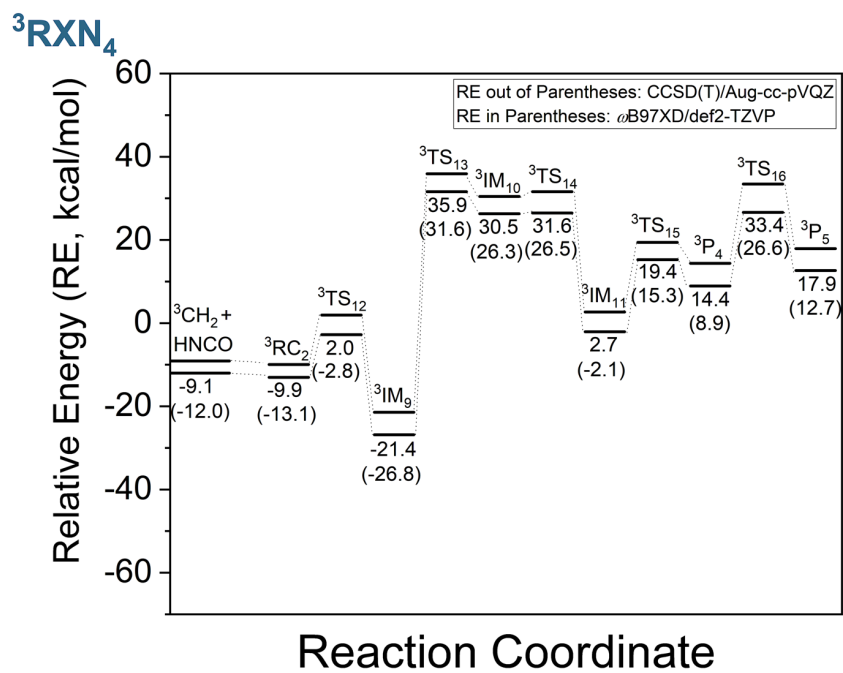


Figure 5-7. Triplet potential energy diagrams of reactions between ³CH₂ and HNCO. The relative energies calculated by CCSD(T) and DFT are shown, with DFT in parentheses.

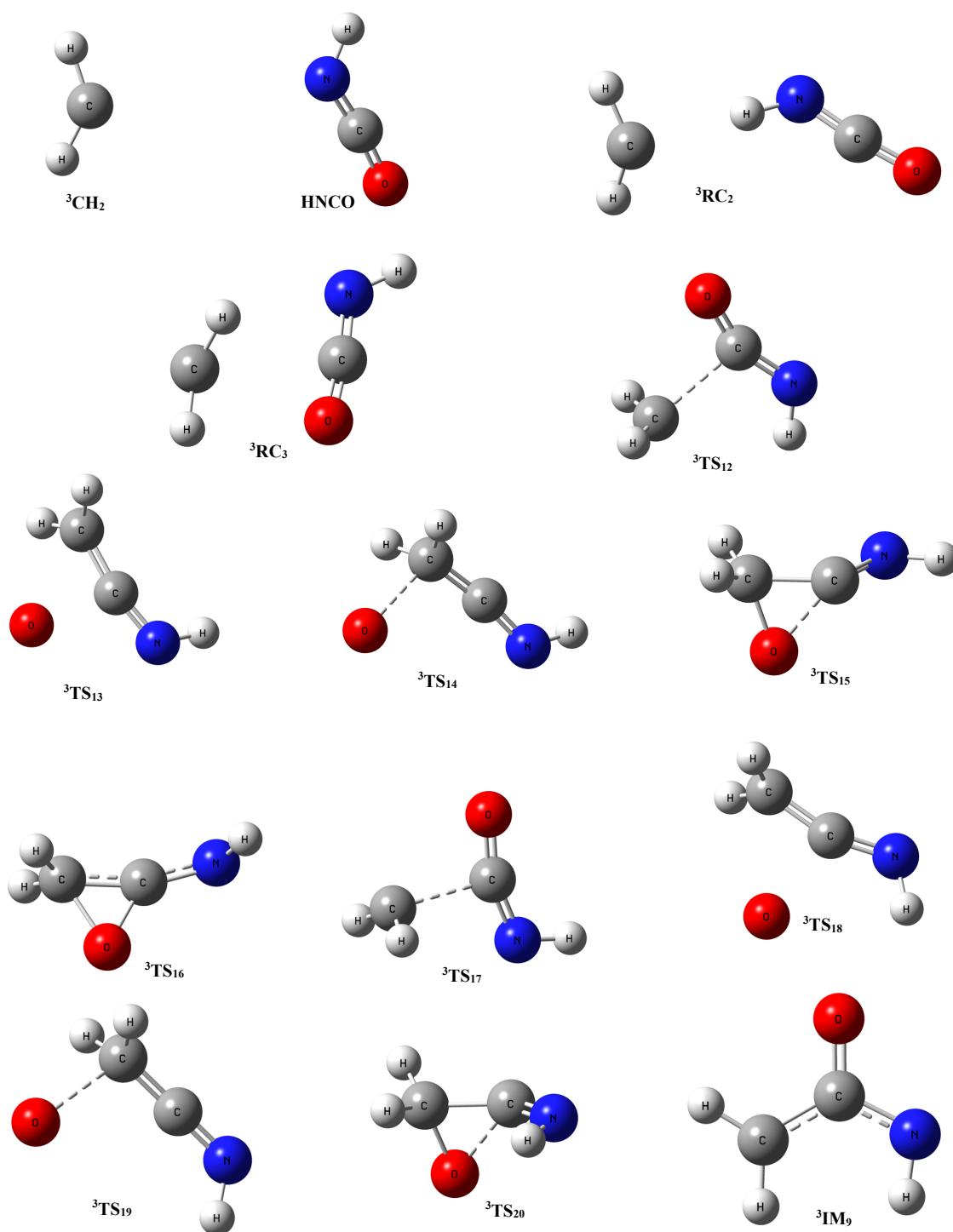


Figure 5-8. Optimized geometries of triplet reactants, RCs ($^3\text{RC}_3$ could not be fully optimized), TSs, IMs, and products of reactions (4) and (5) at ω B97XD/def2-TZVP level of theory. Cartesian coordinates of all studied species shown here are listed in Appendix A.

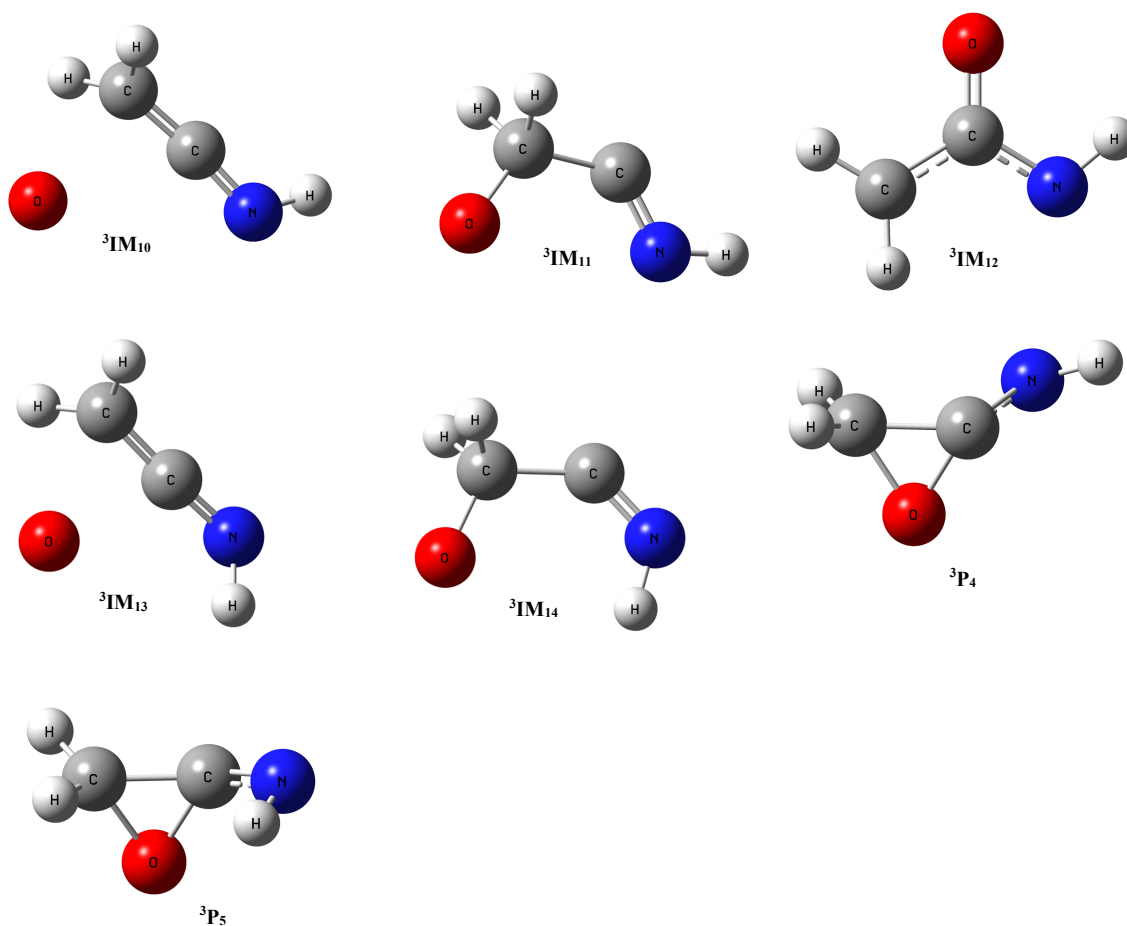


Figure 5-8. (Continued).

5.4.3 Reaction Paths for $\text{CH}_2 + \text{HNC}$

Total energies including zero-point corrections for all studied molecular species are listed in Appendix B along with their relative energies where the zero of energy is assigned to singlet CH_2 plus singlet HNC . The singlet and triplet optimized geometries of reactants, RCs, IMs, TSs, and Ps of reactions (4) and (5) are shown in Figures 5-6 and 5-8 and the cartesian coordinates can be found in Appendix A. The basis set used here is not expected to give accurate results in weakly bound complexes. In particular, the geometries of $^1\text{RC}_2$,

$^1\text{RC}_3$, and $^3\text{RC}_3$ could not be fully optimized. Results on reactant complexes RCs should be considered only in a qualitative sense. Reactions 4 and 5 on the singlet and triplet energy surfaces are represented by $^1\text{RXN}_4$, $^1\text{RXN}_5$, $^3\text{RXN}_4$, and $^3\text{RXN}_5$, respectively. Their singlet and triplet energy profiles are shown in Figures 5-5 (singlet) and 5-7 (triplet).

Recently, Panda, Chiranjibi, Awasthi, Ghosal, and Anoop [270] investigated in detail many reactions on the singlet energy surface of $\text{C}_2\text{H}_3\text{NO}$ by DFT and coupled cluster methods. They found many reaction pathways of $\text{HCN} + \text{H}_2\text{CO}$, but none of them with a barrier smaller than 42 kcal/mol [270]. For their studied reaction pathways of H_2CO with HNC (another molecule observed in the ISM [36,291,292]), the lowest reaction energy barrier found was 15 kcal/mol [270], leading to the formation of a conformer of oxiran-2-ylazanide, $^1\text{P}_5$ (shown in Figure 5-6).

On the singlet surface, reactions (4) and (5) are barrierless and lead to the formation of two conformers of oxiran-2-ylazanide ($^1\text{P}_4$ and $^1\text{P}_5$). On the triplet surface, reactions 4 and 5 have rate limiting barriers of 11.1 and 17.6 kcal/mol, respectively.

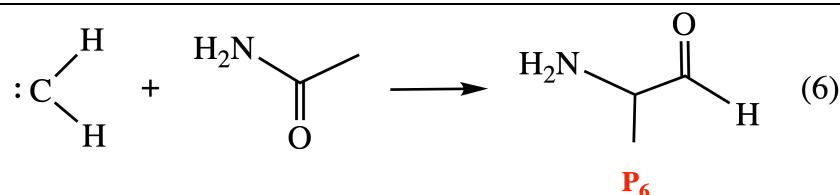
The kinetically favoured triplet product among all triplet potential surfaces is $^3\text{IM}_9$ (see Figure 5-7, panel $^3\text{RXN}_4$). The barrier to forming $^3\text{IM}_9$, is 11.1 kcal/mol, making this reaction practically impossible in the gas phase in the ISM (especially the areas with no radiation and with a very low temperature), but maybe possible in hot corinos regions or at the surface of dust grains. Molecules $^3\text{P}_4$ and $^3\text{P}_5$ are 27 and 23 kcal/mol higher than the reactants and are not likely products of these two triplet reactions (even though molecules $^1\text{P}_4$ and $^1\text{P}_5$ are products of singlet reactions (4) and (5)).

5.5 The Reaction of $^1\text{CH}_2$ with $\text{CH}_3\text{C}(=\text{O})\text{NH}_2$

5.5.1 A Precursor of Alanine May Exist in the ISM or CSE: Alaninal

The reaction (6) of methylene (CH_2) with acetamide ($\text{CH}_3\text{C}(=\text{O})\text{NH}_2$) is shown in Scheme 5-3, leading to one product (P_6). Only the computational study on the reaction of singlet CH_2 and $\text{CH}_3\text{C}(=\text{O})\text{NH}_2$ is reported now (Figures 5-9 and 5-10).

The product of this reaction is alaninal, which is a precursor of alanine that has not been observed in the ISM or CSE [36,287]. The IMs in this reaction are also interesting. The difference between alaninal and alanine is only one oxygen atom. The reaction (6) shown here is a zero-barrier reaction, which might lead to the existence of alaninal in the ISM or CSE, or even alanine when an O atom is available.



Scheme 5-3. The reaction of CH_2 with $\text{CH}_3\text{C}(=\text{O})\text{NH}_2$ and their product.

5.5.2 Energy Profile of $^1\text{CH}_2 + \text{CH}_3\text{C}(=\text{O})\text{NH}_2$

Single point (fixed geometry) calculations with CCSD(T)/Aug-cc-pVQZ are not fully conducted for this reaction for this moment because some calculations based on IMs or TSs need to take more than 7 days even with 32 cores (the maximum time for running a single calculation with Graham is 7 days [293]). Now we are planning to run our SP energy calculations with other ways. Thus, the energies reported here are differences between 0 K

energies at the ω B97XD/def2-TZVP level with zero-point energy corrections and optimized geometries also by ω B97XD/def2-TZVP.

Total energies including zero-point corrections, E , for all studied molecular species reported here are shown in Appendix B along with their relative energies where the zero of energy is assigned to singlet CH_2 plus singlet $\text{CH}_3\text{C}(=\text{O})\text{NH}_2$. The singlet optimized geometries of reactants, RCs, IMs, TSs, and Ps are shown in Figure 5-10 and the cartesian coordinates can be found in Appendix A. As stated before, the basis set used here is not expected to give accurate results in weakly bound complexes. The geometry of $^1\text{RC}_4$ could not be fully optimized. Reaction (6) on the singlet energy profile is represented by $^1\text{RXN}_6$. The energy profile is shown in Figure 5-9.

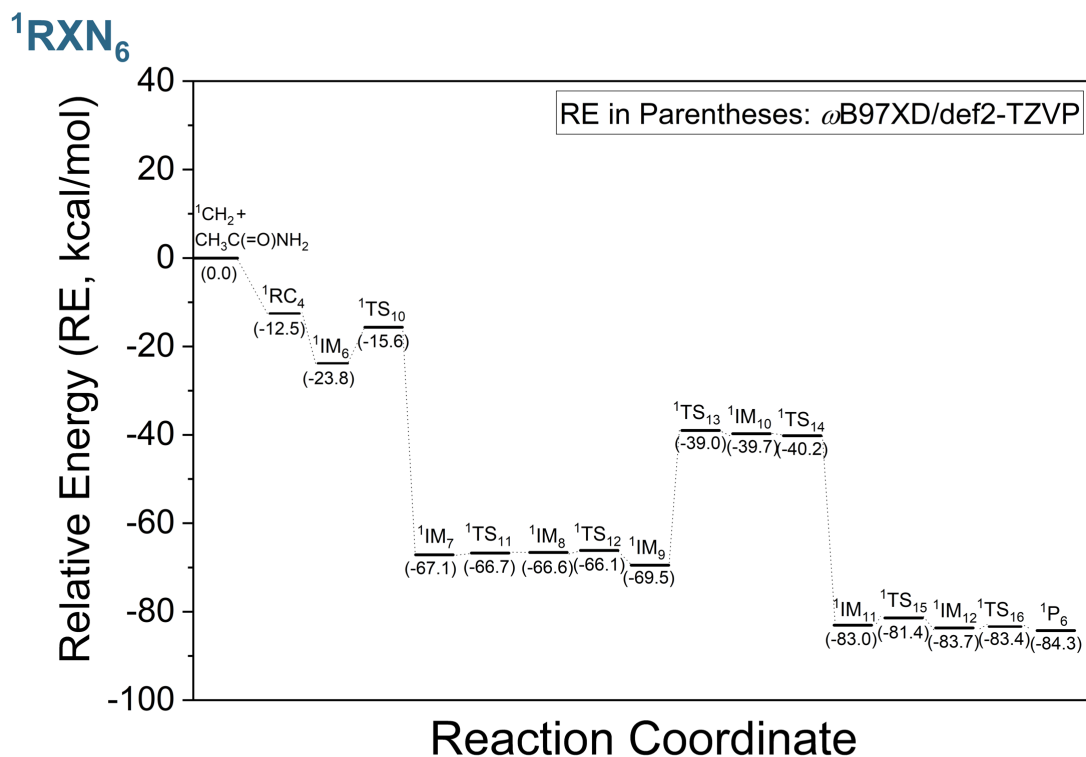


Figure 5-9. Singlet potential energy diagram of the reaction between $^1\text{CH}_2$ and $\text{CH}_3\text{C}(=\text{O})\text{NH}_2$. The relative energies calculated by DFT (in parentheses) are shown.

As shown in Figure 5-9, the singlet reaction of methylene and acetamide (producing alaninal) is barrierless. Two TSs ($^1\text{TS}_{11}$, $^1\text{TS}_{14}$) show a slightly smaller energy than one of their neighbouring IMs, which implies an error of 0.1~0.7 kcal/mol with the current quantum chemistry method. $^1\text{TS}_{11}$ is 0.18 kcal/mol higher than $^1\text{IM}_8$ before adding zero-point corrections to their energies. The product, alaninal, has an energy of -84.3 kcal/mol, which is the lowest energy along this potential energy surface.

As described before, triplet CH_2 needs 9 kcal/mol to reach the singlet excited state. If the triplet-to-singlet crossing occurs, then the triplet reaction with reactants ($^3\text{CH}_2 + \text{CH}_3\text{C}(=\text{O})\text{NH}_2$) might only need to overcome a barrier of less than 9 kcal/mol to “skip” to the singlet energy surface. However, the triplet study of this reaction will be performed in the near future to investigate the detailed mechanisms, which may help us learn more about the potential interstellar molecules.

Alaninal has a very similar structure to alanine except lacking an oxygen atom. It has been reported that O atoms are present and common in the ISM [294,295]. The studies of interstellar O atoms have listed potential reactions with H atoms [296] and alkenes [297]. Pearce *et al.* [298] reported a potential chemical mechanism in the ISM: the O atom could be inserted into the C-H bond of methane, leading to the formation of methanol. Hence, the potential of the interstellar reaction between alaninal and O atoms is worth studying. It brings an interesting possibility, if alanine can be formed by the insertion of the interstellar O atom into the C-H bond of the functional group $-\text{CHO}$ of potential interstellar alaninal.

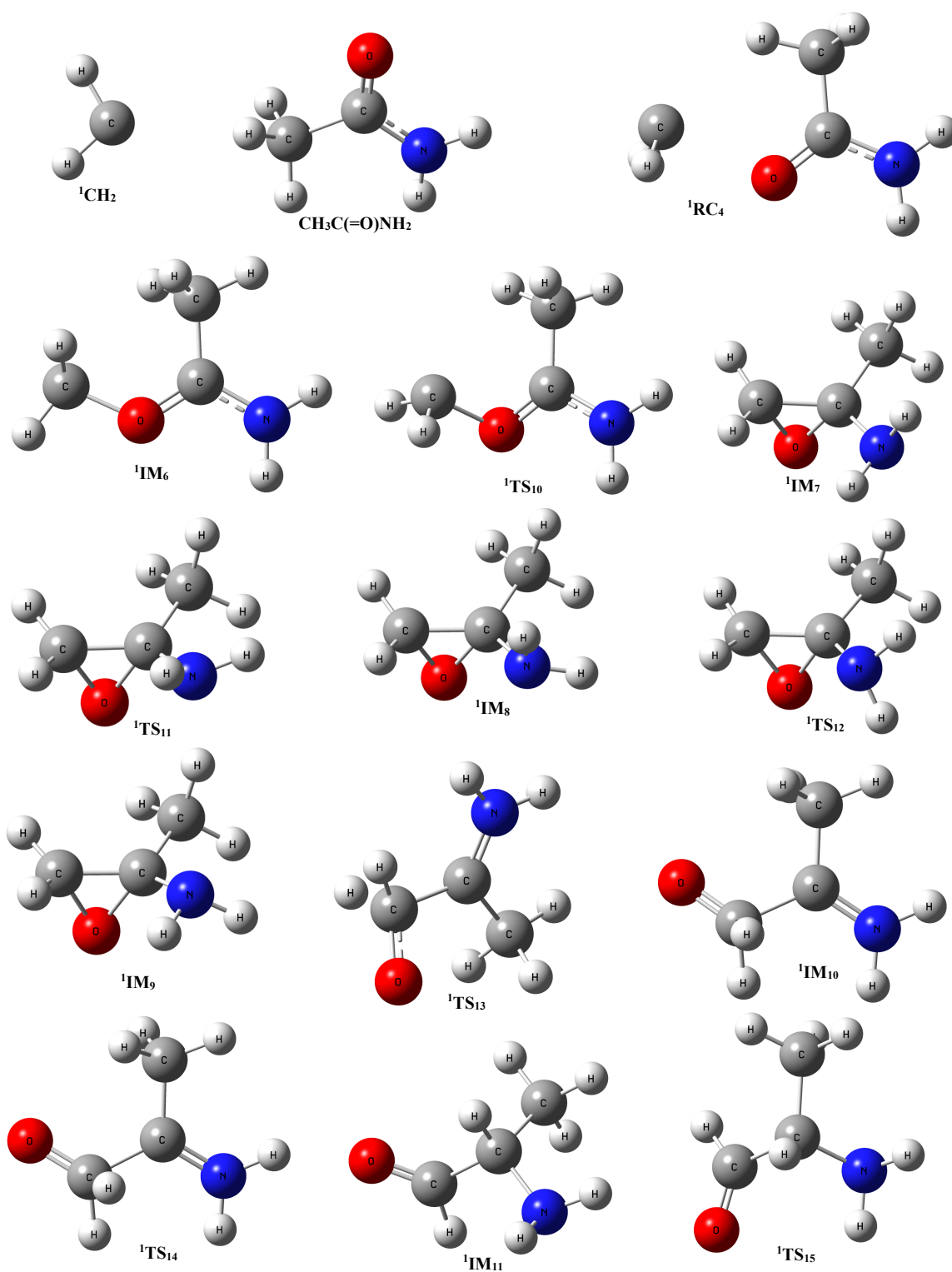


Figure 5-10. Optimized geometries of singlet reactants, RC ($^1\text{RC}_4$ could not be fully optimized), TSs, IMs, and products of reaction (6) at $\omega\text{B97XD/def2-TZVP}$ level of theory. Cartesian coordinates of all studied species shown here are listed in Appendix A.

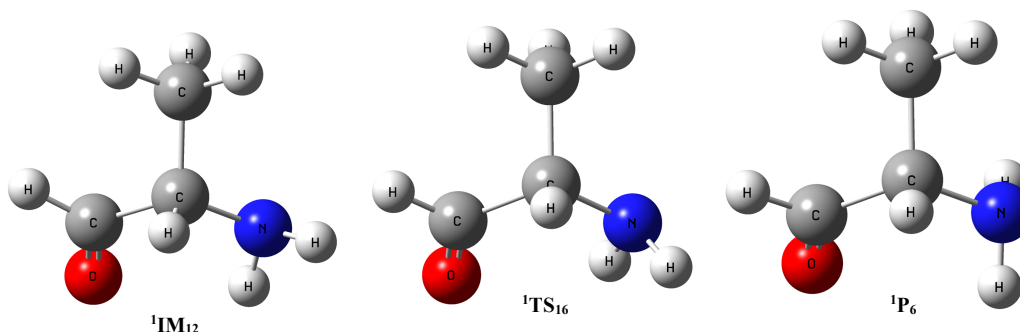
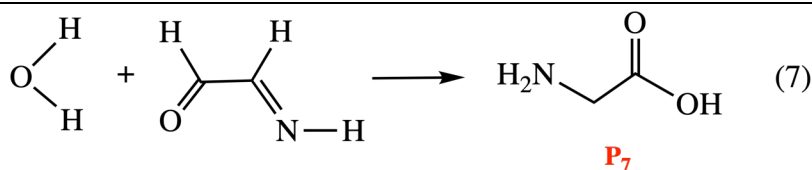


Figure 5-10. (Continued).

5.6 The H₂O + NHCHCHO Reaction

5.6.1 A Reaction Route Producing Glycine

The reaction (7) of water (H₂O) with one conformer of imine acetaldehyde (NHCHCHO) [286] (which is an IM in our reaction (2)) is shown in Scheme 5-4, leading to one product (P₇). Since their ground states are both singlet, the computational study on the reaction of H₂O + NHCHCHO is only reported in singlet (Figures 5-11, and 5-12). The product of this reaction is glycine, which is tagged as an “unconfirmed” molecule in the ISM or CSE [36] and remains the subject of study.



Scheme 5-4. The reaction of H₂O with NHCHCHO and their product.

5.6.2 Energy Profile for the Formation of Glycine

Total energies including zero-point corrections (E) are shown in Appendix B for all minima and transition states along with relative energies where the zero of energy is assigned to singlet H₂O plus singlet NHCHCHO. The optimized geometries of reactants, RCs, IMs, TSs, and Ps are shown in Figure 5-12 and the cartesian coordinates can be found in Appendix A. Reaction (7) on the singlet potential energy surface is represented by ¹RXN₇. The energy profile is shown in Figure 5-11.

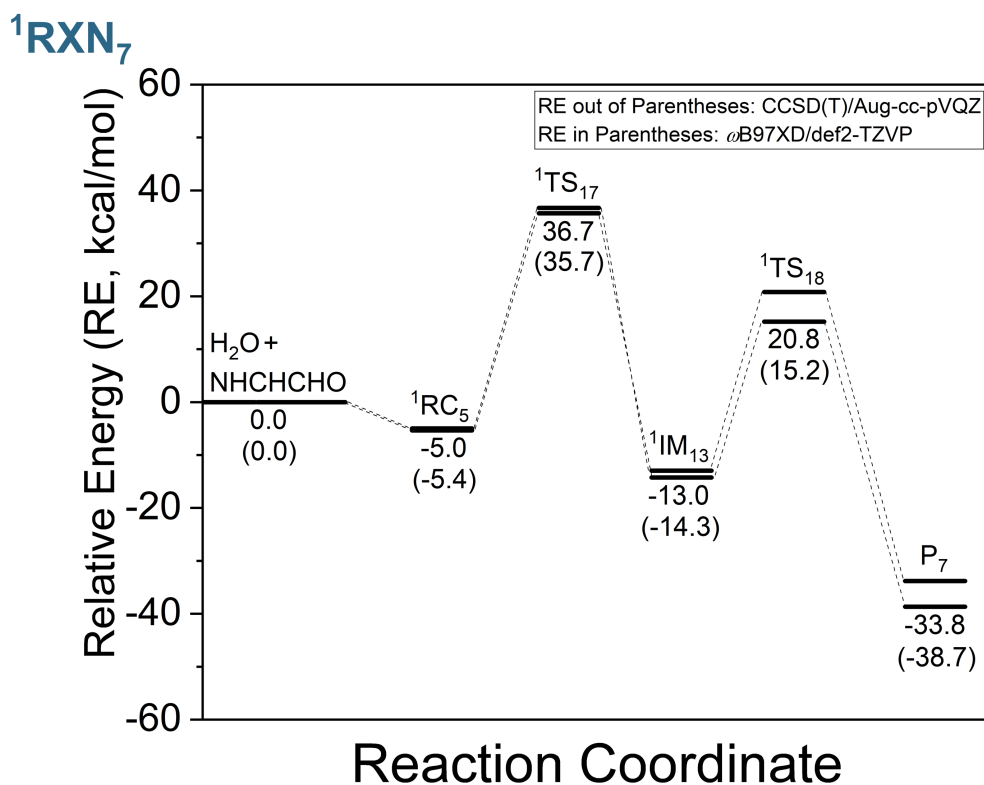


Figure 5-11. Singlet potential energy diagram of the reaction between H₂O and NHCHCHO. The relative energies calculated by CCSD(T) and DFT are shown, with DFT in parentheses.

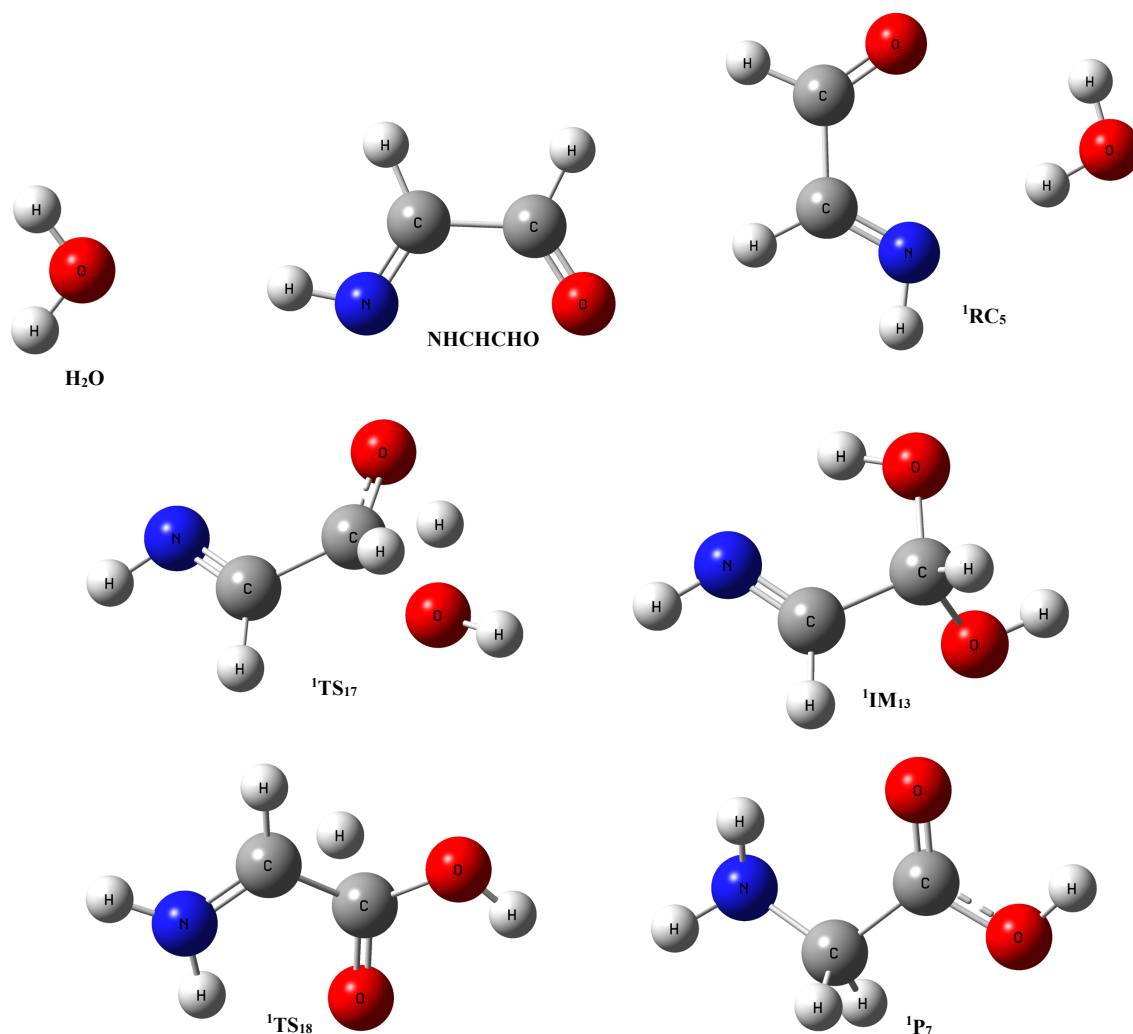


Figure 5-12. Optimized geometries of singlet reactants, RC, TSs, IM, and product of reaction (7) at ω B97XD/def2-TZVP level of theory. Cartesian coordinates of all studied species shown here are listed in Appendix A.

This reaction starts from a possible interstellar molecule, imine acetaldehyde, and a well-known interstellar molecule, water. Even though it shows a barrier of 36.7 kcal/mol, the reaction between those two molecules is interesting, especially since it leads to glycine as the final product. There are some theoretical studies [299–301] that have been trying to

search for interstellar glycine in different ways, which could help us understand better the possible emergence of amino acids in the ISM and CSE [299].

However, the barrier of reaction (7) reported here is approximately 37 kcal/mol, which makes the reaction impossible in the ISM or CSE areas with very low temperatures.

5.7 Discussion and Conclusions

The $\text{CH}_2 + \text{HCNO}$, $\text{CH}_2 + \text{HNCO}$, and $\text{CH}_2 + \text{CH}_3\text{C}(=\text{O})\text{NH}_2$ reactions were identified as possible sources of as yet undetected interstellar prebiotic molecules by means of a screening procedure that combines energy barriers estimated by machine learning (described in Chapter 3) and empirical rules about chemical composition or some others (described in Chapter 4). The reactions of CH_2 with HCNO/HNCO were studied on the singlet and triplet energy surfaces by DFT at the $\omega\text{B97XD}/\text{def2-TZVP}$ level for geometry optimization, TS finding, IRC, and normal mode frequencies and at the $\text{CCSD(T)}/\text{Aug-cc-pVQZ}/\omega\text{B97XD}/\text{def2-TZVP}$ level for computing energy differences such as barriers to reaction ($\text{CH}_2 + \text{CH}_3\text{C}(=\text{O})\text{NH}_2$ reaction was reported only on the singlet energy surface and only at the $\omega\text{B97XD}/\text{def2-TZVP}$ level so far). Singlet CH_2 reacts with HCNO without a barrier to produce *N*-methyleneformamide ($^1\text{P}_1$) and Imine Acetaldehyde ($^1\text{P}_2$), with the former one being thermodynamically favoured by only ~ 1 kcal/mol (or we can say, these two products are both thermodynamically favoured). However, singlet CH_2 is 9.00 kcal/mol higher than ground state triplet CH_2 [288]. Therefore, reactions on the singlet energy surface would require triplet-to-singlet crossing and collision energies of ~ 9 kcal/mol, or less than 9 kcal/mol if triplet-to-singlet surface crossing occurs at a favorable reactant complex

geometry. As shown in $^3\text{RXN}_4$, triplet CH_2 can react with HNCO with a barrier of 11.1 kcal/mol to produce $^3\text{IM}_9$, $\text{CH}_2\text{C}(=\text{O})\text{NH}$. These results do not give very strong evidence for the formation of prebiotic molecules from reactions of triplet CH_2 with HCNO or HNCO in the ISM or CSE. However, the barriers to reaction relative to ground state reactants are found to be small for several reactions: ~ 9 kcal/mol for reactions 1 to 6 if triplet-to-singlet surface crossing occurs at large intermolecular distances, possibly less than 9 kcal/mol if crossing occurs at shorter intermolecular distances; and on the triplet surface, approximately 11 kcal/mol for reaction (4), and approximately 18 kcal/mol for reaction (5).

The above results suggest three areas for future research: (1) a more precise characterization of the triplet and singlet energy surfaces of weakly bound $\text{CH}_2\text{-HCNO}$ or $\text{CH}_2\text{-HNCO}$ or $\text{CH}_2\text{-CH}_3\text{C}(=\text{O})\text{NH}_2$ complexes, with a search for possible triplet-singlet crossing (especially those leading to a reaction $E < 9$ kcal/mol); (2) the modelling of possible reactions of CH_2 with ice-covered dust grains without and with coadsorbed HCNO or HNCO or $\text{CH}_3\text{C}(=\text{O})\text{NH}_2$; and (3) a search for the spectroscopic signature of *N*-methyleneformamide, imine acetaldehyde, and other potential biomolecules in the ISM and CSE.

Chapter 6 Energy, Structure, and Spectroscopic Study of

Conformers of *N*-methyleneformamide and Imine

Acetaldehyde

6.1 *N*-methyleneformamide

The main source of information about ISM molecules are spectral lines in the microwave region in light collected by telescopes. These lines correspond to transitions between rotational states and give precise information about the moments of inertia of a molecule. Quantum chemical calculations of equilibrium geometries provide moments of inertia and spectroscopic constants A, B and C that can be compared to lines in the microwave region for molecule identification. When several conformers are possible, relative energies can be calculated and the spectroscopic constants of the most stable conformer are used for identifying molecules. This kind of computational study has been done for imine acetaldehyde [286]. Here we describe a similar study we did for *N*-methyleneformamide. Theoretical studies available on *N*-methyleneformamide are few. There have been reports of vibrational spectra [289] and of relative energies of conformers [39,290]. The conformer we found in reaction (1) is the *gauche* conformer which was found to be the most stable one in earlier work [39,290]. *N*-methyleneformamide is our thermodynamically favoured product of reaction (1) shown in section 5.3, with a relative energy of -147.3 kcal/mol relative to reactants. It is reported as a potential ISM or CSE molecule in Chapter 5.

Therefore, a computational study on its conformers' structural constants and spectroscopic parameters will give useful information for possible identification by astronomers.

6.2 Imine Acetaldehyde

In 2018, Redondo *et al.* [286] reported four conformers of imine acetaldehyde and the relative energies calculated at different levels of theory, and the equilibrium geometries of the conformers. They proposed imine acetaldehyde as a likely interstellar molecule [286]. Interestingly, we also came to the conclusion that imine acetaldehyde is an interesting possible ISM molecule [39] but for different reasons from those of reference [286]. Our study of reaction (2) in Chapter 5 leads to finding the different conformations of imine acetaldehyde with the antiperiplanar conformer as the most stable. The names of four conformations (*out-ap*, *in-ap*, *out-sp*, and *in-sp*) are defined in the reference [286]: “*ap*” and “*sp*” represent antiperiplanar and synperiplanar, respectively, “*out*” or “*in*” mean the H atom of NH group is toward outside (*out*) of the skeleton of imine acetaldehyde, or toward inside (*in*).

6.3 Computational Details

The same computational method we described in section 5.2 was used to study the conformers of *N*-methyleneformamide and imine acetaldehyde. All conformers were characterized by normal mode analysis, those with zero imaginary frequency being minima. Conformers (two of *N*-methyleneformamide: *gauche* and *trans*; four of imine acetaldehyde: *out-ap*, *in-ap*, *out-sp*, and *in-sp*) presented in this chapter are identified as minima. Note

that the *cis* conformation is not a minimum according to our DFT calculation, it has one imaginary frequency. This is consistent with the study in literature [289]. The seven conformers we found are shown in Figures 6-1 and 6-2.

Spectroscopic equilibrium rotational constants (A_e , B_e , C_e , in GHz) and dipole moment (μ , in Debye) were calculated at ω B97XD/def2-TZVP level of theory based on ω B97XD-optimized geometries. The moments of inertia (I_A , I_B , I_C , in amu $\text{\AA}^2$) were computed using A_e , B_e , and C_e as inputs through the converting platform of NIST [302]. Spectroscopic ground vibrational state rotational constants (A_0 , B_0 , C_0 , in GHz) without and with centrifugal distortion were both calculated at ω B97XD/def2-TZVP level of theory. These spectroscopic parameters of all studied conformers are listed in Table 6-3.

6.4 Results

6.4.1 Geometries and Relative Energies of Conformers

The optimized geometries of conformers of *N*-methyleneformamide are shown in Figure 6-1, and those of imine acetaldehyde are shown in Figure 6-2.

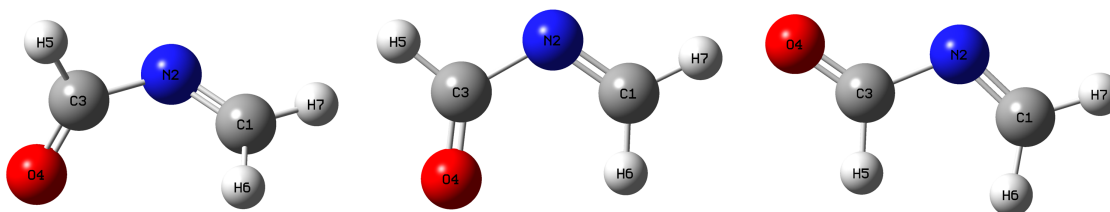


Figure 6-1. 3D structures of conformers (from left to right: *gauche*, *cis*, and *trans*) of *N*-methyleneformamide at ω B97XD/def2-TZVP level of theory. Cartesian coordinates of all studied species shown here are listed in Appendix C.

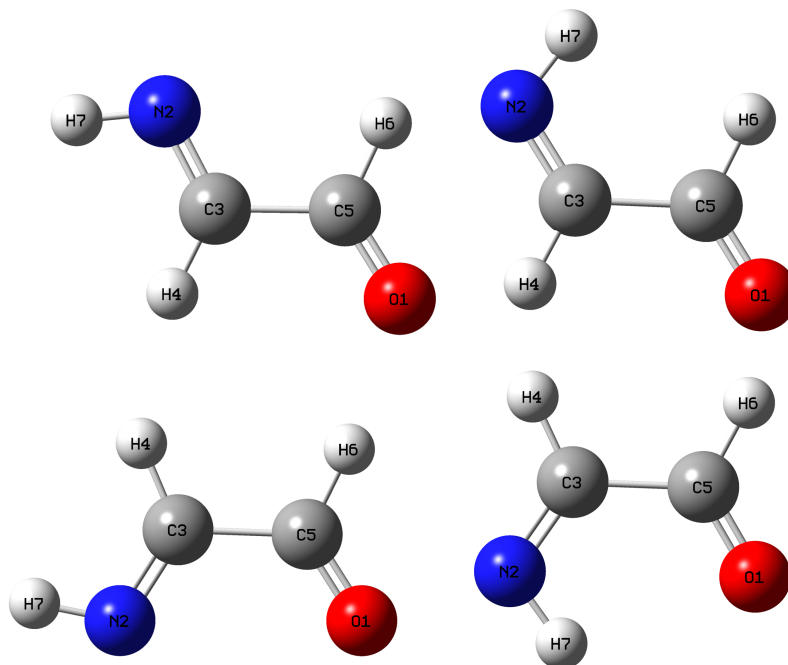


Figure 6-2. 3D structures of conformers (the first row from left to right: *out-ap*, *in-ap*; the second row from left to right: *out-sp*, *in-sp*) of imine acetaldehyde at ω B97XD/def2-TZVP level of theory. Cartesian coordinates of all studied species shown here are listed in Appendix C.

The total energy with zero-point correction and relative energy (RE) of each conformer of two molecules have been listed in Table 6-1. The bond lengths (in angstrom) and angles (in degree) of each conformer of two molecules are listed in Table 6-2.

Table 6-1. Total energies including zero-point corrections (E, in a.u.) and relative energies (RE, in kcal/mol) calculated with DFT (in parentheses) and CCSD(T) for the conformers of *N*-methyleneformamide and imine acetaldehyde

<i>N</i> -methyleneformamide	E (a.u.)	RE (kcal/mol)
<i>Gauche</i>	-207.647474 (-207.915835)	0.00 (0.00)
<i>Cis</i>	-207.6448483 (-207.911813)	1.65 (2.52)
<i>Trans</i>	-207.6453077 (-207.912370)	1.36 (2.17)
Imine Acetaldehyde	E (a.u.)	RE (kcal/mol)
<i>Out-ap</i>	-207.6471170 (-207.912993)	0.00 (0.00)
<i>In-ap</i>	-207.6451982 (-207.910901)	1.20 (1.31)
<i>Out-sp</i>	-207.6383110 (-207.904179)	5.53 (5.53)
<i>In-sp</i>	-207.6440502 (-207.909825)	1.92 (1.99)

Table 6-2. Structural constants including bond lengths (in angstroms) and angles (in degrees) optimized with ω B97XD/def2-TZVP of the conformers of *N*-methyleneformamide and imine acetaldehyde

Bonds of <i>N</i> -methyleneformamide	Bond Length			Angles of <i>N</i> -methyleneformamide	Angle		
	<i>Gauche</i>	<i>Cis</i>	<i>Trans</i>		<i>Gauche</i>	<i>Cis</i>	<i>Trans</i>
C1-H6	1.093	1.093	1.098	H6-C1-H7	117.84	118.56	117.37
C1-H7	1.088	1.087	1.087	H6-C1-N2	122.92	122.78	123.14
C1-N2	1.259	1.266	1.264	H7-C1-N2	119.23	118.66	119.49
C3-N2	1.396	1.435	1.423	C1-N2-C3	120.79	115.32	114.26
C3-H5	1.103	1.099	1.113	N2-C3-O4	124.65	126.20	123.01
C3-O4	1.198	1.196	1.189	N2-C3-H5	112.24	111.55	114.79
				O4-C3-H5	122.96	122.25	122.20

Bonds of Imine Acetaldehyde	Bond Length				Angles of Imine Acetaldehyde	Angle			
	<i>Out-ap</i>	<i>In-ap</i>	<i>Out-sp</i>	<i>In-sp</i>		<i>Out-ap</i>	<i>In-ap</i>	<i>Out-sp</i>	<i>In-sp</i>
O1-C5	1.199	1.197	1.195	1.199	O1-C5-H6	123.03	122.01	121.38	121.45
C5-H6	1.106	1.112	1.110	1.106	O1-C5-C3	122.62	123.49	124.33	122.13
C3-C5	1.496	1.499	1.511	1.511	C3-C5-H6	114.35	114.49	114.30	116.43
C3-H4	1.094	1.090	1.096	1.091	C5-C3-H4	115.88	116.22	114.99	117.79
C3-N2	1.262	1.262	1.260	1.260	H4-C3-N2	125.97	120.15	124.45	119.79
N2-H7	1.018	1.024	1.019	1.025	C5-C3-N2	118.14	123.63	120.56	123.02
					C3-N2-H7	111.79	111.03	111.20	109.30

6.4.2 Spectroscopic Constants of Conformers

The spectroscopic rotational constants, moments of inertia, dipole moment of each conformer of *N*-methyleneformamide (denoted **1**) and imine acetaldehyde (denoted **2**) are listed in Table 6-3. The spectroscopic constants (in GHz) derived from our calculations have no more than 3 or 4 significant figures.

Table 6-3. Spectroscopic equilibrium rotational constants (A_e , B_e , C_e , in GHz) with moments of inertia in parentheses (I_A , I_B , I_C , in amu Å²), ground vibrational state rotational constants (A_0 , B_0 , C_0 , in GHz) with and without centrifugal distortion (in parentheses), and dipole moment (μ , in Debye) of conformers of *N*-methyleneformamide and imine acetaldehyde

Molecule with Conformation	A_e (I_A)	B_e (I_B)	C_e (I_C)	A_0	B_0	C_0	μ
1 (<i>Gauche</i>)	30.3769 (16.6369)	6.0906 (82.9773)	5.3850 (93.8503)	31.0745 (31.0745)	5.9712 (5.9711)	5.3212 (5.3214)	2.90
1 (<i>Cis</i>)	25.3471 (19.9383)	7.0244 (71.9458)	5.5002 (91.8841)	25.2167 (25.2167)	6.9785 (6.9785)	5.4470 (5.4470)	1.58
1 (<i>Trans</i>)	56.6808 (8.9162)	5.1748 (97.6625)	4.7418 (106.5787)	55.7268 (55.7268)	5.1437 (5.1437)	4.7242 (4.7242)	4.27

2 (<i>Out-ap</i>)	54.0216 (9.3551)	4.7831 (105.6593)	4.3941 (115.0144)	53.4062 (53.4062)	4.7597 (4.7598)	4.3710 (4.3710)	2.35
2 (<i>In-ap</i>)	50.9031 (9.9283)	4.7957 (105.3828)	4.3828 (115.3109)	50.3256 (50.3256)	4.7698 (4.7698)	4.3581 (4.3581)	1.54
2 (<i>Out-sp</i>)	26.2563 (19.2480)	6.1269 (82.4858)	4.9677 (101.7338)	26.4571 (26.4571)	6.0223 (6.0223)	4.9387 (4.9387)	4.10
2 (<i>In-sp</i>)	24.3416 (20.7619)	6.5056 (77.6843)	5.1336 (98.4463)	24.2953 (24.2953)	6.4403 (6.4403)	5.0892 (5.0892)	2.23

Chapter 7 Conclusion

A fast computational method for estimating reaction energy barriers was developed. It is a machine learning model trained on 11,730 reactions with accurately known barriers to reactions [117], with features obtained after many trials to minimize the MAE and RMSE. It uses Kernel Ridge Regression with a Laplacian kernel and 300 features that we designed to describe chemical reactions [20]. These features are relatively easy to interpret. They are based on combinations of: atomic number, electronegativity and hardness of atoms; interatomic distances; and empirical bond lengths and bond energies. Such features could easily be generalized to elements other than H, C, N and O. The computational cost for estimating an energy barrier is low, one-hundredth of a second. This makes it possible to estimate barriers for millions of chemical reactions.

The computational efficiency of the KRR ML model enabled the second part of our work on hypothetical reactions in interstellar space. A set of 136,081 chemical reactions were generated by combining two reactants chosen among 72 molecules known to exist in the ISM or CSE, with the constraint that the product is a single molecule that is in the dataset we used for training the ML model [20,39]. These reactions were screened to eliminate those with a non-zero ML energy barrier estimate. A second round of screening eliminated reactions with products that lack similarity to known, biologically relevant, organic molecules. Among the top ranked reactions after these two rounds of screening were the reactions of CH_2 with HCNO and HNCO. We studied these reactions with accurate quantum chemistry methods to obtain possible reaction paths, including intermediates,

transition states, products, and intrinsic reaction coordinates, calculated with a good DFT method. More accurate energy calculations, with the CCSD(T) method, were done at the stationary points of the potential energy surfaces. We found two reactions with barriers of approximately 10 kcal/mol, maybe less if triplet-singlet crossing occurs at a favourable geometry [39]. These reactions lead to formation of *N*-methyleneformamide and imine acetaldehyde and might play a role in chemical evolution in the ISM and CSE. We identified two more possible interstellar molecules, oxiran-2-ylazanide and alaninal.

In the third part of this work, we calculated accurate energies, equilibrium geometries, and rotational constants for the possible conformers of *N*-methyleneformamide and imine acetaldehyde. We hope these results will stimulate the search for evidence of these molecules in the ISM and CSE by analysis of microwave spectral lines.

Our future work will be concerned, firstly, with a computational characterization of reaction paths on the ground state triplet energy surface of CH_2 reacting with $\text{CH}_3\text{C}(=\text{O})\text{NH}_2$. Second, we plan to create a list of all products of reaction that rank high after the two rounds of screening (those with low barrier and bio-organic characteristics), find their most stable conformers, and compute equilibrium geometries and rotational constants, as we did for *N*-methyleneformamide and imine acetaldehyde. In this study, we will also use a minimum energy principle [287,303] as another criterion for screening to find interesting candidates (possible ISM molecules). This will help the search and astronomical detection of interstellar molecules [286]. A third topic for future work is to design a procedure to search for singlet-triplet energy surface crossings for CH_2 interacting

with small closed-shell molecules. These crossings could be important for finding possible low-energy pathways of reactions of CH₂ with other ISM molecules. A fourth plan for future work is to describe the advantages of KRR related to its good interpretability, i.e., how to explain the results predicted by KRR from a point of view of chemistry. Compared to the complicated neural networks with difficult interpretability [324,325], KRR is an interpretable ML method, and its predictions can be described on the basis of similarities between chemical reactions. Every prediction by KRR is a linear combination of known energy barriers in our training dataset. If a specific reaction of the training set is quite similar to the predicted reaction, then its known energy barrier will be assigned a large weight. Based on the similarities between chemical reactions provided by KRR, another future trial could be clustering categories of reactions. In this way, more information about the mechanism of different types of reactions, such as S_N2 and E2 [310,326], might be conveyed. Another future work could be the addition of the potential interstellar molecules we reported to the previous ISM and CSE molecule database for extending our possible reaction dataset size and then for the next reaction generation loop.

Bibliography

- (1) Roston, D.; Demapan, D.; Cui, Q. Leaving Group Ability Observably Affects Transition State Structure in a Single Enzyme Active Site. *J. Am. Chem. Soc.* **2016**, *138* (23), 7386–7394. <https://doi.org/10.1021/jacs.6b03156>.
- (2) Dybala-Defratyka, A.; Paneth, P.; Pu, J.; Truhlar, D. G. Benchmark Results for Hydrogen Atom Transfer between Carbon Centers and Validation of Electronic Structure Methods for Bond Energies and Barrier Heights. *J. Phys. Chem. A* **2004**, *108* (13), 2475–2486. <https://doi.org/10.1021/jp037312j>.
- (3) Jackson, R.; Zhang, W.; Pearson, J. TSNet: Predicting Transition State Structures with Tensor Field Networks and Transfer Learning. *Chem. Sci.* **2021**, *12* (29), 10022–10040. <https://doi.org/10.1039/D1SC01206A>.
- (4) Junginger, A.; Craven, G. T.; Bartsch, T.; Revuelta, F.; Borondo, F.; Benito, R. M.; Hernandez, R. Transition State Geometry of Driven Chemical Reactions on Time-Dependent Double-Well Potentials. *Phys. Chem. Chem. Phys.* **2016**, *18* (44), 30270–30281. <https://doi.org/10.1039/C6CP02519F>.
- (5) Mardirossian, N.; Head-Gordon, M. Thirty Years of Density Functional Theory in Computational Chemistry: An Overview and Extensive Assessment of 200 Density Functionals. *Mol. Phys.* **2017**, *115* (19), 2315–2372. <https://doi.org/10.1080/00268976.2017.1333644>.
- (6) Yang, K.; Zheng, J.; Zhao, Y.; Truhlar, D. G. Tests of the RPBE, RevPBE, τ -HCTHhyb, ω B97X-D, and MOHLYP Density Functional Approximations and 29 Others

against Representative Databases for Diverse Bond Energies and Barrier Heights in Catalysis. *J. Chem. Phys.* **2010**, *132* (16), 1–10. <https://doi.org/10.1063/1.3382342>.

(7) Sousa, S. F.; Fernandes, P. A.; Ramos, M. J. General Performance of Density Functionals. *J. Phys. Chem. A* **2007**, *111*, 10439–10452. <https://doi.org/10.1021/jp0734474>.

(8) Lynch, B. J.; Truhlar, D. G. Small Representative Benchmarks for Thermochemical Calculations. *J. Phys. Chem. A* **2003**, *107* (42), 8996–8999. <https://doi.org/10.1021/jp035287b>.

(9) Zheng, J.; Zhao, Y.; Truhlar, D. G. The DBH24/08 Database and Its Use to Assess Electronic Structure Model Chemistries for Chemical Reaction Barrier Heights. *J. Chem. Theory Comput.* **2009**, *5* (4), 808–821. <https://doi.org/10.1021/ct800568m>.

(10) Lynch, B. J.; Fast, P. L.; Harris, M.; Truhlar, D. G. Adiabatic Connection for Kinetics. *J. Phys. Chem. A* **2000**, *104* (21), 4811–4815. <https://doi.org/10.1021/jp000497z>.

(11) Varandas, A. J. C.; Martínez González, M.; Montero-Cabrera, L. A.; Garcia de la Vega, J. M. Assessing How Correlated Molecular Orbital Calculations Can Perform versus Kohn–Sham DFT: Barrier Heights/Isomerizations. *Chem. Eur. J.* **2017**, *23* (38), 9122–9129. <https://doi.org/10.1002/chem.201700928>.

(12) Zhao, Y.; González-García, N.; Truhlar, D. G. Benchmark Database of Barrier Heights for Heavy Atom Transfer, Nucleophilic Substitution, Association, and Unimolecular Reactions and Its Use to Test Theoretical Methods. *J. Phys. Chem. A* **2005**, *109* (9), 2012–2018. <https://doi.org/10.1021/jp045141s>.

(13) Lynch, B. J.; Truhlar, D. G. How Well Can Hybrid Density Functional Methods

Predict Transition State Geometries and Barrier Heights? *J. Phys. Chem. A* **2001**, *105* (13), 2936–2941. <https://doi.org/10.1021/jp004262z>.

(14) Lynch, B. J.; Truhlar, D. G. What Are the Best Affordable Multi-Coefficient Strategies for Calculating Transition State Geometries and Barrier Heights? *J. Phys. Chem. A* **2002**, *106* (5), 842–846. <https://doi.org/10.1021/jp014002x>.

(15) Čížek, J. On the Use of the Cluster Expansion and the Technique of Diagrams in Calculations of Correlation Effects in Atoms and Molecules. In *Advances in Chemical Physics*; LeFebvre, R., Moser, C., Eds.; Wiley Interscience: New York, 1969; Vol. 14, pp 35–89.

(16) Purvis III, G. D.; Bartlett, R. J. A Full Coupled-Cluster Singles and Doubles Model – The Inclusion of Disconnected Triples. *J. Chem. Phys.* **1982**, *76*, 1910–1918. <https://doi.org/10.1063/1.443164>.

(17) Scuseria, G. E.; Janssen, C. L.; Schaefer III, H. F. An Efficient Reformulation of the Closed-Shell Coupled Cluster Single and Double Excitation (CCSD) Equations. *J. Chem. Phys.* **1988**, *89*, 7382–7387. <https://doi.org/10.1063/1.455269>.

(18) Scuseria, G. E.; Schaefer III, H. F. Is Coupled Cluster Singles and Doubles (CCSD) More Computationally Intensive than Quadratic Configuration-Interaction (QCISD)? *J. Chem. Phys.* **1989**, *90*, 3700–3703. <https://doi.org/10.1063/1.455827>.

(19) Pople, J. A.; Head-Gordon, M.; Raghavachari, K. Quadratic Configuration Interaction – A General Technique for Determining Electron Correlation Energies. *J. Chem. Phys.* **1987**, *87*, 5968–5975. <https://doi.org/10.1063/1.453520>.

(20) Ji, H.; Rágyanszki, A.; Fournier, R. A. Machine Learning Estimation of Reaction

Energy Barriers. *Comput. Theor. Chem.* **2023**, *1229*, 114332.
<https://doi.org/10.1016/j.comptc.2023.114332>.

(21) Lin, Y. S.; Li, G. D.; Mao, S. P.; Chai, J. D. Long-range Corrected Hybrid Density Functionals with Improved Dispersion Corrections. *J. Chem. Theory Comput.* **2013**, *9*, 263–272. <https://doi.org/10.1021/ct300715s>.

(22) Chai, J. D.; Head-Gordon, M. Long-range Corrected Hybrid Density Functionals with Damped Atom–atom Dispersion Corrections. *Phys. Chem. Chem. Phys.* **2008**, *10*, 6615–6620. <https://doi.org/10.1039/B810189B>.

(23) (a) Grimme, S. Semiempirical GGA-type Density Functional Constructed with a Long-range Dispersion Correction. *J. Comput. Chem.* **2006**, *27* (15), 1787–1799. <https://doi.org/10.1002/jcc.20495>;

(b) Grimme, S.; Antony, J.; Ehrlich, S.; Krieg, H. A Consistent and Accurate *Ab Initio* Parametrization of Density Functional Dispersion Correction (DFT-D) for the 94 Elements H–Pu. *J. Chem. Phys.* **2010**, *132* (15), 154104. <https://doi.org/10.1063/1.3382344>.

(24) Chai, J. D.; Head-Gordon, M. Systematic Optimization of Long-range Corrected Hybrid Density Functionals. *J. Chem. Phys.* **2008**, *128* (8), 084106. <https://doi.org/10.1063/1.2834918>.

(25) Schaefer, A.; Huber, C.; Ahlrichs, R. Fully Optimized Contracted Gaussian-Basis Sets of Triple Zeta Valence Quality for Atoms Li to Kr. *J. Chem. Phys.* **1994**, *100*, 5829–5835. <https://doi.org/10.1063/1.467146>.

(26) Weigend, F.; Ahlrichs, R. Balanced Basis Sets of Split Valence, Triple Zeta Valence and Quadruple Zeta Valence Quality for H to Rn: Design and Assessment of Accuracy.

Phys. Chem. Chem. Phys. **2005**, 7, 3297–3305. <https://doi.org/10.1039/B508541A>.

(27) Weigend, F. Accurate Coulomb-fitting Basis Sets for H to Rn. *Phys. Chem. Chem. Phys.* **2006**, 8, 1057–1065. <https://doi.org/10.1039/B515623H>.

(28) Rágyanszki, A.; Ji, H.; Fournier, R. Understanding the Origins of Life - the Constituents of Interstellar Medium as the Source of Life's Building Blocks. *Biophys. J.* **2020**, 118 (3, Suppl. 1), 339a–340a. <https://doi.org/10.1016/j.bpj.2019.11.1891>.

(29) Oró, J. Chemical Evolution and the Origin of Life. *Adv. Space Res.* **1983**, 3 (9), 77–94.

(30) Scattergood, T.; Des Marais, D.; Jahnke, L. Life's Origin: the Cosmic, Planetary and Biological Processes. *Planet Rep.* **1987**, 7 (6), 4–5.

(31) Trimble, V. Origin of the Biologically Important Elements. *Orig. Life Evol. Biosph.* **1997**, 27, 3–21. <https://doi.org/10.1023/A:1006561811750>.

(32) Guélin, M.; Cernicharo, J. Organic Molecules in Interstellar Space: Latest Advances. *Front. Astron. Space Sci.* **2022**, 9. <https://doi.org/10.3389/fspas.2022.787567>;

(33) (a) Martins, Z.; Botta, O.; Fogel, M. L.; Sephton, M. A.; Glavin, D. P.; Watson, J. S.; Dworkin, J. P.; Schwartz, A. W.; Ehrenfreund, P. Extraterrestrial Nucleobases in the Murchison Meteorite. *Earth Planet. Sci. Lett.* **2008**, 270, 130–136. <https://doi.org/10.1016/j.epsl.2008.03.026>.

(b) Pizzarello, S.; Shock, E. The Organic Composition of Carbonaceous Meteorites: The Evolutionary Story ahead of Biochemistry. *Cold Spring Harb Perspect. Biol.* **2010**, 2 (3), a002105. <https://doi.org/10.1101/cshperspect.a002105>.

(c) Matson, J. Meteorite That Fell in 1969 Still Revealing Secrets of the Early Solar

System. <https://www.scientificamerican.com/article/murchison-meteorite/> (accessed October 2, 2023).

(34) Chyba, C. F.; Sagan, C. Comets as a Source of Prebiotic Organic Molecules for the Early Earth. In *Comets and the Origin and Evolution of Life*; Thomas, P. J., Chyba, C. F., McKay, C. P., Eds.; Springer: New York, 1997; pp 147–173.

(35) Barone, V.; Biczysko, M.; Puzzarini, C. Quantum Chemistry Meets Spectroscopy for Astrochemistry: Increasing Complexity toward Prebiotic Molecules. *Acc. Chem. Res.* **2015**, *48* (5), 1413–1422. <https://doi.org/10.1021/ar5003285>.

(36) List of interstellar and circumstellar molecules. https://en.wikipedia.org/wiki/List_of_interstellar_and_circumstellar_molecules#Molecules (accessed April 30, 2019).

(37) Tielens, A. G. G. M.; Charnley, S. B. Circumstellar and Interstellar Synthesis of Organic Molecules. *Biophys. J.* **1997**, *27*, 23–51.

(38) Rehder, D. The Interstellar Medium. *Chemistry in Space: from Interstellar Matter to the Origin of Life*; Wiley-VCH: Weinheim, 2010; pp 45–98.

(39) Ji, H.; Rágyanszki, A.; Fournier, R. A. Computational Study of the Reactions of Interstellar Molecules: CH₂ Reacting with HCNO and HNCO. *ChemRxiv* **2023**. <https://doi.org/10.26434/chemrxiv-2023-kpm5r>. This content is a preprint and has not been peer-reviewed. [Submitted for publication].

(40) Ferrière, K. M. The Interstellar Environment of Our Galaxy. *Rev. Mod. Phys.* **2001**, *73* (4), 1031–1066. <https://doi.org/10.1103/RevModPhys.73.1031>.

(41) Ziurys, L. M. The Chemistry in Circumstellar Envelopes of Evolved Stars:

Following the Origin of the Elements to the Origin of Life. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103* (33), 12274–12279. <https://doi.org/10.1073/pnas.0602277103>.

(42) Bladon, P.; Gorton, J. E.; Hammond, R. B. Software for Molecular Modelling and Computational Chemistry. *Molecular Modelling: Computational Chemistry Demystified*; The Royal Society of Chemistry: Cambridge, 2012; pp 31–46.

(43) Goodman, J. M. Molecular Orbital Theory. *Chemical Applications of Molecular Modelling*; The Royal Society of Chemistry: Cambridge, 1998; pp 122–149.

(44) (a) Szabo, A.; Ostlund, N. S. *Modern Quantum Chemistry: Introduction to Advanced Electronic Structure Theory*; Dover Publications, Inc.: New York, 1996;

(b) Fedorov, D. A. Quantum Nuclear Dynamics: Application to Spin-Forbidden Transitions and Lifetime of Vibrational States. Ph.D. dissertation, University of Nevada, Reno, Reno, Nevada, 2017;

(c) Aliano, A.; Cicero, G. Ab Initio DFT Simulations of Nanostructures. In *Encyclopedia of Nanotechnology*; Bhushan, B., Ed.; Springer: Dordrecht, 2012; pp 11–17;

(d) Levine, I. N. *Quantum Chemistry – Volume I: Quantum Mechanics and Molecular Electronic Structure*; The Allyn and Bacon Chemistry Series; Allyn and Bacon: Boston, 1970;

(e) Hamiltonian (quantum mechanics). [https://en.wikipedia.org/wiki/Hamiltonian_\(quantum_mechanics\)](https://en.wikipedia.org/wiki/Hamiltonian_(quantum_mechanics)) (accessed May 30, 2020).

(45) Marqués, L. A.; Aboy, M.; López, P.; Santos, I.; Pelaz, L.; Fisicaro, G. Atomistic Modeling of Laser-Related Phenomena. In *Laser Annealing Processes in Semiconductor*

Technology; Cristiano, F., Magna, A. L., Eds.; Woodhead Publishing Series in Electronic and Optical Materials; Woodhead Publishing, 2021; pp 79–136.

(46) Rimola, A.; Ferrero, S.; Germain, A.; Corno, M.; Ugliengo, P. Computational Surface Modelling of Ices and Minerals of Interstellar Interest—Insights and Perspectives. *Minerals* **2021**, *11* (1), 26. <https://doi.org/10.3390/min11010026>.

(47) Parr, R. G.; Yang, W. Density-Functional Theory of the Electronic Structure of Molecules. *Annu. Rev. Phys. Chem.* **1995**, *46*, 701–728.

(48) Khalatur, P. G. Molecular Dynamics Simulations in Polymer Science: Methods and Main Results. In *Polymer Science: A Comprehensive Reference*; Matyjaszewski, K., Möller, M., Eds.; Elsevier, 2012; Vol. 1, pp 417–460.

(49) Johansson, M. P.; Kaila, V. R. I.; Sundholm, D. *Ab Initio*, Density Functional Theory, and Semi-Empirical Calculations. In *Biomolecular Simulations*; Monticelli, L., Salonen, E., Eds.; Methods in Molecular Biology; Humana Press: Totowa, 2013; Vol. 924, pp 3–27.

(50) Cui, Q.; Elstner, M. Density Functional Tight Binding: Values of Semi-Empirical Methods in an *Ab Initio* era. *Phys. Chem. Chem. Phys.* **2014**, *16* (28), 14368–14377. <https://doi.org/10.1039%2Fc4cp00908h>.

(51) Stewart, J. J. P. Optimization of Parameters for Semiempirical Methods. V. Modification of NDDO Approximations and Application to 70 Elements. *J. Mol. Model.* **2007**, *13*, 1173–1213. <https://doi.org/10.1007/s00894-007-0233-4>.

(52) Throssell, K. T. Evaluating and Improving Approximate LCAO-MO Theory with Restored Overlap and Bond Order Bond Energy Corrections. Ph.D. dissertation, Wesleyan

University, Middletown, CT, 2017.

(53) Throssel, K.; Frisch, M. J. Evaluation and Improvement of Semi-Empirical Methods I: PM7R8: A Variant of PM7 with Numerically Stable Hydrogen Bonding Corrections. In Prep (unpublished results).

(54) Stewart, J. J. P. Optimization of Parameters for Semiempirical Methods VI: More Modifications to the NDDO Approximations and Re-Optimization of Parameters. *J. Mol. Model.* **2013**, *19*, 1–32. <https://doi.org/10.1007/s00894-012-1667-x>.

(55) García-Andrade, X.; Tahoces, P. G.; Pérez-Ríos, J.; Núñez, E. M. Barrier Height Prediction by Machine Learning Correction of Semiempirical Calculations. *J. Phys. Chem. A* **2023**, *127*, 2274–2283. <https://doi.org/10.1021/acs.jpca.2c08340>.

(56) Pople, J. A. Some Deficiencies of MINDO/3 Semiempirical Theory. *J. Am. Chem. Soc.* **1975**, *97*, 5306–5308.

(57) Hu, H.; Yang, W. Development and Application of Ab Initio QM/MM Methods for Mechanistic Simulation of Reactions in Solution and in Enzymes. *Theochem (J. Mol. Struct.)* **2009**, *898*, 17–30. <https://doi.org/10.1016/j.theochem.2008.12.025>.

(58) Fukui, K. The Path of Chemical Reactions – The IRC Approach. *Acc. Chem. Res.* **1981**, *14* (12), 363–368. <https://doi.org/10.1021/ar00072a001>.

(59) Hratchian, H. P.; Schlegel, H. B. Finding Minima, Transition States, and Following Reaction Pathways on *Ab Initio* Potential Energy Surfaces. In *Theory and Applications of Computational Chemistry: The First 40 Years*; Dykstra, C. E., Frenking, G., Kim, K. S., Scuseria, G., Eds.; Elsevier: Amsterdam, 2005; pp 195–249.

(60) Rupp, M.; Tkatchenko, A.; Müller, K.; von Lilienfeld, O. A. Fast and Accurate

Modeling of Molecular Atomization Energies with Machine Learning. *Phys. Rev. Lett.* **2012**, *108* (5), 058301. <https://doi.org/10.1103/PhysRevLett.108.058301>.

(61) Yao, K.; Herr, J. E.; Brown, S. N.; Parkhill, J. Intrinsic Bond Energies from a Bonds-in-Molecules Neural Network. *J. Phys. Chem. Lett.* **2017**, *8* (12), 2689–2694. <https://doi.org/10.1021/acs.jpcclett.7b01072>.

(62) Carande, W. H.; Kazakov, A.; Muzny, C.; Frenkel, M. Quantitative Structure-Property Relationship Predictions of Critical Properties and Acentric Factors for Pure Compounds. *J. Chem. Eng. Data* **2015**, *60* (5), 1377–1387. <https://doi.org/10.1021/je501093v>.

(63) Elton, D. C.; Boukouvalas, Z.; Butrico, M. S.; Fuge, M. D.; Chung, P. W. Applying Machine Learning Techniques to Predict the Properties of Energetic Materials. *Sci. Rep.* **2018**, *8* (1), No. 9059. <https://doi.org/10.1038/s41598-018-27344-x>.

(64) Ghoddusi, H.; Creamer, G. G.; Rafizadeh, N. Machine Learning in Energy Economics and Finance: A Review. *Energy Econ.* **2019**, *81*, 709–727. <https://doi.org/10.1016/j.eneco.2019.05.006>.

(65) Richiardi, J.; Achard, S.; Bunke, H.; Van De Ville, D. Machine Learning with Brain Graphs: Predictive Modeling Approaches for Functional Imaging in Systems Neuroscience. *IEEE Signal Process. Mag.* **2013**, *30* (3), 58–70. <https://doi.org/10.1109/MSP.2012.2233865>.

(66) Asri, H.; Mousannif, H.; Moatassime, H. A.; Noel, T. Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis. *Procedia Comput. Sci.* **2016**, *83*, 1064–1069. <https://doi.org/10.1016/j.procs.2016.04.224>.

(67) Borboudakis, G.; Stergiannakos, T.; Frysali, M.; Klontzas, E.; Tsamardinos, I.; Froudakis, G. E. Chemically Intuited, Large-Scale Screening of MOFs by Machine Learning Techniques. *NPJ Comput. Mater.* **2017**, *3*, No. 40. <https://doi.org/10.1038/s41524-017-0045-8>.

(68) Jinich, A.; Sanchez-Lengeling, B.; Ren, H.; Harman, R.; Aspuru-Guzik, A. A Mixed Quantum Chemistry/Machine Learning Approach for the Fast and Accurate Prediction of Biochemical Redox Potentials and Its Large-Scale Application to 315 000 Redox Reactions. *ACS Cent. Sci.* **2019**, *5*, 1199–1210. <https://doi.org/10.1021/acscentsci.9b00297>.

(69) Haywood, A. L.; Redshaw, J.; Gaertner, T.; Taylor, A.; Mason, A. M.; Hirst, J. D. Machine Learning for Chemical Synthesis. In *Machine Learning in Chemistry: The Impact of Artificial Intelligence*; Cartwright, H. M., Ed.; The Royal Society of Chemistry: Cambridge, 2020; pp 169–194.

(70) Roberts, M. G.; Lawrence, R. MedChemInformatics: An Introduction to Machine Learning for Drug Discovery. In *Machine Learning in Chemistry: The Impact of Artificial Intelligence*; Cartwright, H. M., Ed.; The Royal Society of Chemistry: Cambridge, 2020; pp 37–75.

(71) Carracedo-Reboredo, P.; Liñares-Blanco, J.; Rodríguez-Fernández, N.; Cedrón, F.; Novoa, F. J.; Carballal, A.; Maojo, V.; Pazos, A.; Fernandez-Lozano, C. A Review on Machine Learning Approaches and Trends in Drug Discovery. *Comput. Struct. Biotechnol. J.* **2021**, *19*, 4538–4558. <https://doi.org/10.1016/j.csbj.2021.08.011>.

(72) Alpaydin, E. *Introduction to Machine Learning*, 3rd ed.; MIT Press: Cambridge, Massachusetts, 2014.

- (73) (a) Flower, D. *Molecular Collisions in the Interstellar Medium*, 2nd ed.; Cambridge University Press: Cambridge, 2007;
- (b) Herbst, E. Chemistry in the Interstellar Medium. *Annu. Rev. Phys. Chem.* **1995**, *46*, 27–53. <https://doi.org/10.1146/annurev.pc.46.100195.000331>;
- (c) Maggiolo, R.; Gibbons, A.; Cessateur, G.; De Keyser, J.; Dhooghe, F.; Gunell, H.; Loreau, J.; Mousis, O.; Vaeck, N. Effect of the Surface Roughness of Icy Grains on Molecular Oxygen Chemistry in Molecular Clouds. *Astrophys. J.* **2019**, *882* (2), 131. <https://doi.org/10.3847/1538-4357/ab3400>.
- (74) McKee, C. F.; Draine, B. T. Interstellar Shock Waves. *Science* **1991**, *252*, 397–403. <https://doi.org/10.1126/science.252.5004.397>.
- (75) Baddeley, M. C.; Barrowclough, D. V. An Introduction to Ordinary Least Squares. In *Running Regressions: A Practical Guide to Quantitative Research in Economics, Finance and Development Studies*; Cambridge University Press: Cambridge, 2009; pp 11–35.
- (76) Cantrell, C. A. Technical Note: Review of Methods for Linear Least-Squares Fitting of Data and Application to Atmospheric Chemistry Problems. *Atmos. Chem. Phys.* **2008**, *8*, 5477–5487. <https://doi.org/10.5194/acp-8-5477-2008>.
- (77) McDonald, G. C. Ridge Regression. *WIREs Comp. Stat.* **2009**, *1*, 93–100. <https://doi.org/10.1002/wics.14>.
- (78) Arashi, M., Saleh, A. K. M. E., Kibria, B. M. G., Eds.; *Theory of Ridge Regression Estimation with Applications*; John Wiley & Sons: Hoboken, 2009.
- (79) Strutz, T. *Data Fitting and Uncertainty: A Practical Introduction to Weighted Least*

Squares and Beyond, 2nd ed.; Springer Fachmedien Wiesbaden: Wiesbaden, 2015.

(80) Holland, P. W. Weighted Ridge Regression: Combining Ridge and Robust Regression Methods. *NBER Work. Pap.* **1973**, No. w0011. <https://ssrn.com/abstract=259324>.

(81) a) Kung, S. Y. Kernel Ridge Regressors and Variants. In *Kernel Methods and Machine Learning*; Cambridge University Press: New York, 2014; pp 219–336;

b) Murphy, K. R. *Machine Learning: A Probabilistic Perspective*; The MIT Press: Cambridge, Massachusetts, 2012.

(82) Hastie, T., Tibshirani, R., Friedman, J., Eds.; *The Elements of Statistical Learning: Data Mining, Inference and Prediction*, 2nd ed.; Springer: New York, 2009.

(83) Hansen, K.; Montavon, G.; Biegler, F.; Fazli, S.; Rupp, M.; Scheffler, M.; von Lilienfeld, O. A.; Tkatchenko, A.; Müller, K. -R. Assessment and Validation of Machine Learning Methods for Predicting Molecular Atomization Energies. *J. Chem. Theory Comput.* **2013**, 9 (8), 3404–3419. <https://doi.org/10.1021/ct400195d>.

(84) Stuke, A.; Todorović, M.; Rupp, M.; Kunkel, C.; Ghosh, K.; Himanen, L.; Rinke, P. Chemical Diversity in Molecular Orbital Energy Predictions with Kernel Ridge Regression. *J. Chem. Phys.* **2019**, 150 (20), 204121. <https://doi.org/10.1063/1.5086105>.

(85) Kriegeskorte, N.; Golan, T. Neural Network Models and Deep Learning. *Current Biology* **2019**, 29 (7), R231–R236. <https://doi.org/10.1016/j.cub.2019.02.034>.

(86) MacKay, D. J. C. Bayesian Interpolation. *Neural Comput.* **1992**, 4 (3), 415–447. <https://doi.org/10.1162/neco.1992.4.3.415>.

(87) Foresee, F. D.; Hagan, M. T. Gauss-Newton Approximation to Bayesian Learning.

Proceedings of the International Joint Conference on Neural Networks, June, 1997.

(88) Wu, D.; Huang, H.; Qiu, S.; Liu, Y.; Wu, Y.; Ren, Y.; Mou, J. Application of Bayesian Regularization Back Propagation Neural Network in Sensorless Measurement of Pump Operational State. *Energy Rep.* **2022**, *8*, 3041–3050. <https://doi.org/10.1016/j.egy.2022.02.072>.

(89) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Petersson, G. A.; Nakatsuji, H.; Li, X.; Caricato, M.; Marenich, A. V.; Bloino, J.; Janesko, B. G.; Gomperts, R.; Mennucci, B.; Hratchian, H. P.; Ortiz, J. V.; Izmaylov, A. F.; Sonnenberg, J. L.; Williams-Young, D.; Ding, F.; Lipparini, F.; Egidi, F.; Goings, J.; Peng, B.; Petrone, A.; Henderson, T.; Ranasinghe, D.; Zakrzewski, V. G.; Gao, J.; Rega, N.; Zheng, G.; Liang, W.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Throssell, K.; Montgomery, J. A., Jr.; Peralta, J. E.; Ogliaro, F.; Bearpark, M. J.; Heyd, J. J.; Brothers, E. N.; Kudin, K. N.; Staroverov, V. N.; Keith, T. A.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A. P.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Millam, J. M.; Klene, M.; Adamo, C.; Cammi, R.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Farkas, O.; Foresman, J. B.; Fox, D. J. *Gaussian 16*, Revision C.01; Gaussian, Inc.: Wallingford, CT, 2016.

(90) Dennington, R. Keith, T. A. Millam, J. M. *GaussView*, Version 6; Semichem Inc.: Shawnee Mission, KS, 2016.

(91) *Avogadro: An Open-source Molecular Builder and Visualization Tool*, Version 1.2.0. <http://avogadro.cc/> (accessed June 10, 2021).

- (92) Hanwell, M. D.; Curtis, D. E.; Lonie, D. C.; Vandermeersch, T.; Zurek, E.; Hutchison, G. R. Avogadro: An Advanced Semantic Chemical Editor, Visualization, and Analysis Platform. *J. Cheminform.* **2012**, *4*, 1–17. <https://doi.org/10.1186/1758-2946-4-17>.
- (93) (a) Expanding the Limits of Computational Chemistry. Optimization. <https://gaussian.com/opt/> (accessed January 9, 2023);
- (b) National Center for Biotechnology Information. PubChem Compound Summary for CID 702, Ethanol. <https://pubchem.ncbi.nlm.nih.gov/compound/Ethanol> (accessed September 17, 2023).
- (94) Foresman, J. B.; Frisch, Æ. Geometry Optimizations. In *Exploring Chemistry with Electronic Structure Methods*, 2nd ed.; Gaussian, Inc.: Pittsburgh, PA, 1996; pp 39–59.
- (95) Expanding the Limits of Computational Chemistry. Molecule Specifications. <https://gaussian.com/molspec/> (accessed October 31, 2018).
- (96) Foresman, J. B.; Frisch, Æ. Frequency Calculations. In *Exploring Chemistry with Electronic Structure Methods*, 2nd ed.; Gaussian, Inc.: Pittsburgh, PA, 1996; pp 61–90.
- (97) Expanding the Limits of Computational Chemistry. Freq. <https://gaussian.com/freq/> (accessed January 9, 2023).
- (98) Foresman, J. B.; Frisch, Æ. Studying Chemical Reactions and Reactivity. In *Exploring Chemistry with Electronic Structure Methods*, 2nd ed.; Gaussian, Inc.: Pittsburgh, PA, 1996; pp 165–211.
- (99) Expanding the Limits of Computational Chemistry. Transition State Optimizations with Opt=QST2. <https://gaussian.com/qst2/> (accessed September 17, 2023).
- (100) Peng, C.; Schlegel, H. B. Combining Synchronous Transit and Quasi-Newton

Methods for Finding Transition States. *Israel J. Chem.* **1993**, *33*, 449–454.
<https://doi.org/10.1002/ijch.199300051>.

(101) Li, X.; Frisch, M. J. Energy-Represented DIIS within a Hybrid Geometry Optimization Method. *J. Chem. Theory and Comput.* **2006**, *2*, 835–839.
<http://dx.doi.org/10.1021/ct050275a>.

(102) Pulay, P.; Fogarasi, G.; Fogarasi, F.; Fogarasi, J. E. Systematic Ab Initio Gradient Calculation of Molecular Geometries, Force Constants, and Dipole-Moment Derivatives. *J. Am. Chem. Soc.* **1979**, *101*, 2550–2560. <http://dx.doi.org/10.1021/ja00504a009>.

(103) Fogarasi, G.; Fogarasi, X.; Fogarasi, P.; Fogarasi, P. The Calculation of Ab Initio Molecular Geometries: Efficient Optimization by Natural Internal Coordinates and Empirical Correction by Offset Forces. *J. Am. Chem. Soc.* **1992**, *114*, 8191–8201.
<http://dx.doi.org/10.1021/ja00047a032>.

(104) Pulay, P.; Fogarasi, G. Geometry Optimization in Redundant Internal Coordinates. *J. Chem. Phys.* **1992**, *96*, 2856–2860. <http://dx.doi.org/10.1063/1.462844>.

(105) Baker, J. Techniques for Geometry Optimization – a Comparison of Cartesian and Natural Internal Coordinates. *J. Comp. Chem.* **1993**, *14*, 1085–1100.
<http://dx.doi.org/10.1002/jcc.540140910>.

(106) Peng, C.; Ayala, P. Y.; Schlegel, H. B.; Frisch, M. J. Using Redundant Internal Coordinates to Optimize Equilibrium Geometries and Transition States. *J. Comp. Chem.* **1996**, *17*, 49–56.

(107) The ‘Art’ of Finding Transition States Part 1.
<https://joaquinbarroso.com/2016/05/26/the-art-of-finding-transition-states-part-1/>

(accessed January 26, 2023).

(108) Scheiner, S.; Seybold, P.G. Quantum Chemical Analysis of the Energetics of the *Anti* and *Gauche* Conformers of Ethanol. *Struct. Chem.* **2009**, *20*, 43–48. <https://doi.org/10.1007/s11224-008-9395-7>.

(109) The ‘Art’ of Finding Transition States Part 2. <https://joaquinbarroso.com/2016/05/31/the-art-of-finding-transition-states-part-2/>

(accessed January 26, 2023).

(110) Expanding the Limits of Computational Chemistry. SP. <https://gaussian.com/sp/> (accessed June 3, 2023).

(111) Kendall, R. A.; Dunning Jr., T. H.; Harrison, R. J. Electron Affinities of the First-Row Atoms Revisited. Systematic Basis Sets and Wave Functions. *J. Chem. Phys.* **1992**, *96*, 6796–6806. <https://doi.org/10.1063/1.462569>.

(112) Woon, D. E.; Dunning Jr., T. H. Gaussian-Basis Sets for Use in Correlated Molecular Calculations. III. The Atoms Aluminum Through Argon. *J. Chem. Phys.* **1993**, *98*, 1358–1371. <https://doi.org/10.1063/1.464303>.

(113) Papajak, E.; Zheng, J.; Leverentz, H. R.; Truhlar, D. G. Perspectives on Basis Sets Beautiful: Seasonal Plantings of Diffuse Basis Functions. *J. Chem. Theory and Comput.* **2011**, *7*, 3027–3034. <https://doi.org/10.1021/ct200106a>.

(114) Expanding the Limits of Computational Chemistry. Scan. <https://gaussian.com/scan/> (accessed August 4, 2023).

(115) *PyCharm (Community Edition)*, Version 2020.1.2.; JetBrains, 2020 <https://www.jetbrains.com/pycharm/> (accessed June 1, 2020).

- (116) MATLAB Online. <https://www.mathworks.com/products/matlab-online.html> (accessed May 29, 2023).
- (117) (a) Grambow, C. A.; Pattanaik, L.; Green, W. H. Reactants, Products, and Transition States of Elementary Chemical Reactions Based on Quantum Chemistry. *Sci. Data* **2020**, 7 (137). <https://doi.org/10.1038/s41597-020-0460-4>;
- (b) Grambow, C. A.; Pattanaik, L.; Green, W. H. Reactants, Products, and Transition States of Elementary Chemical Reactions Based on Quantum Chemistry (1.0.1) [Data set], 2020. Zenodo. <https://doi.org/10.5281/zenodo.3715478> (accessed June 9, 2021);
- (c) Ruddigkeit, L.; Van Deursen, R.; Blum, L. C.; Reymond, J. L. Enumeration of 166 Billion Organic Small Molecules in the Chemical Universe Database GDB-17. *J. Chem. Inf. Model.* **2012**, 52 (11), 2864–2875. <https://doi.org/10.1021/ci300415d>.
- (118) Xu, Z. F.; Raghunath, P.; Lin, M. C. Ab Initio Chemical Kinetics for the $\text{CH}_3 + \text{O}(^3\text{P})$ Reaction and Related Isomerization–Decomposition of CH_3O and CH_2OH Radicals. *J. Phys. Chem. A* **2015**, 119 (28), 7404–7417.
- (119) Ruscic, B. IUPAC Critical Evaluation of Thermochemical Properties of Selected Radicals. Part I. *J. Phys. Chem. Ref. Data* **2005**, 34, 573.
- (120) Chen, H.; Zhu, R.; Chen, H.; Li, H.; Ju, S. Ab Initio Study on Mechanisms and Kinetics for Reaction of NCS with NO. *J. Phys. Chem. A* **2008**, 112 (24), 5495–5501.
- (121) Ruscic, B.; Berkowitz, J. The H–NCS Bond Energy, $\Delta H_f^\circ(\text{HNCS})$, $\Delta H_f^\circ(\text{NCS})$, and $\text{IP}(\text{NCS})$ from Photoionization Mass Spectrometric Studies of HNCS, NCS, and $(\text{NCS})_2$. *J. Chem. Phys.* **1994**, 101, 7975.

- (122) Chase, M. W.; Davies, C. A.; Downey, J. R.; Frurip, D. J.; McDonald, R. A.; Syverud, A. N. JANAF Thermochemical Tables Third Edition. *J. Phys. Chem. Ref. Data* **1985**, *14*, 1.
- (123) Troya, D. *Ab Initio* and Direct Quasiclassical-Trajectory Study of the $F + CH_4 \rightarrow HF + CH_3$ Reaction. *J. Chem. Phys.* **2005**, *123* (21).
- (124) Johnson, R. D., III. NIST Computational Chemistry Comparison and Benchmark Database, NIST Standard Reference Database Number 101, Release 21, August 2020. <http://cccbdb.nist.gov/> (accessed May 06, 2021).
- (125) Peiró-García, J.; Nebot-Gil, I.; Merchán, M. An ab Initio Study on the Mechanism of the Atmospheric Reaction $NH_2 + O_3 \rightarrow H_2NO + O_2$. *Chemphyschem* **2003**, *4*, 366-372.
- (126) Lu, X.; Shao, K.; Fu, B.; Wang, X.; Zhang, D. H. An Accurate Full-Dimensional Potential Energy Surface and Quasiclassical Trajectory Dynamics of the $H + H_2O_2$ Two-Channel Reaction. *Phys. Chem. Chem. Phys.* **2018**, *20*, 23095–23105.
- (127) Faragó, E. P.; Szori, M.; Owen, M. C.; Fittschen, C.; Viskolcz, B. Critical Evaluation of the Potential Energy Surface of the $CH_3 + HO_2$ Reaction System. *J. Chem. Phys.* **2015**, *142* (5).
- (128) Kaur, R.; Vikas. Conflict in the Mechanism and Kinetics of the Barrierless Reaction between SH and NO_2 Radicals. *J. Phys. Chem. A* **2018**, *122*, 1926–1937.
- (129) Jamal, A.; Mebel, A. M. Theoretical Investigation of the Mechanism and Product Branching Ratios of the Reactions of Cyano Radical with 1- and 2 - Butyne and 1, 2 - Butadiene. *J. Phys. Chem. A* **2013**, *117*, 741–755.

- (130) Feng, W.; Hershberger, J. F. Kinetics of the $\text{CN} + \text{CS}_2$ and $\text{CN} + \text{SO}_2$ Reactions. *J. Phys. Chem. A* **2011**, *115* (3), 286–290.
- (131) Xu, K.; Xu, Z.; Lin, M. C. Theoretical Study on the Kinetics and Mechanism for the Reaction of FCO with NO. *J. Phys. Chem. A* **2006**, *110* (21), 6718–6723.
- (132) Castillo, J. F.; Aoiz, F. J.; Martínez-Haya, B. Theoretical Study of the Dynamics of $\text{Cl} + \text{O}_3$ Reaction I. Ab Initio Potential Energy Surface and Quasiclassical Trajectory Results. *Phys. Chem. Chem. Phys.* **2011**, *13*, 8537–8548.
- (133) Hou, H.; Wang, B. Mechanistic and Kinetic Study of the $\text{O} + \text{CH}_2\text{OH}$ Reaction. *J. Phys. Chem. A* **2005**, *109* (21), 4796–4803.
- (134) Nguyen, H. M. T.; Nguyen, H. T.; Nguyen, T.; Hoang, H. V.; Vereecken, L. Theoretical Study on the Reaction of the Methylidyne Radical, $\text{CH} (\text{X}^2\Pi)$, with Formaldehyde, CH_2O . *J. Phys. Chem. A* **2014**, *118*, 8861–8871.
- (135) Wang, M.; Sun, X.; Bian, W.; Cai, Z. A Global 12-Dimensional Ab Initio Potential Energy Surface and Dynamical Studies for the $\text{SiH}_4 + \text{H} \rightarrow \text{SiH}_3 + \text{H}_2$ Reaction. *J. Chem. Phys.* **2006**, *124* (23), 234311.
- (136) Talbi, D.; Smith, I. W. M. A Theoretical Analysis of the Reaction between CN Radicals and NH_3 . *Phys. Chem. Chem. Phys.* **2009**, *11* (38), 8477–8483.
- (137) Kerkeni, B.; Clary, D. C. Ab Initio Rate Constants from Hyperspherical Quantum Scattering: Application to $\text{H} + \text{C}_2\text{H}_6$ and $\text{H} + \text{CH}_3\text{OH}$. *J. Chem. Phys.* **2004**, *121* (14), 6809–6821.
- (138) Bryukov, M. G.; Slagle, I. R.; Knyazev, V. D. Kinetics of Reactions of H Atoms

with Ethane and Chlorinated Ethanes. *J. Phys. Chem. A* **2001**, *105* (28), 6900–6909.

(139) Nguyen, T. L.; Stanton, J. F. Ab Initio Thermal Rate Calculations of $\text{HO} + \text{HO} = \text{O}(^3\text{P}) + \text{H}_2\text{O}$ Reaction and Isotopologues. *J. Phys. Chem. A* **2013**, *117* (13), 2678–2686.

(140) Ju, L.; Xie, T.; Zhang, X.; Han, K. A Modified Potential Energy Surface for the $\text{C}_2\text{H} + \text{H}_2 \rightarrow \text{C}_2\text{H}_2 + \text{H}$ Reaction and a Theoretical Study on Its Rate Constants. *Chem. Phys. Lett.* **2005**, *409*, 249–254.

(141) Dong, H.; Ding, Y.; Sun, C. $\text{C}_2\text{H} + \text{H}_2\text{CO}$: A New Route for Formaldehyde Removal. *J. Chem. Phys.* **2005**, *122* (20).

(142) Blitz, M. A.; Hughes, K. J.; Pilling, M. J.; Robertson, S. H. Combined Experimental and Master Equation Investigation of the Multiwell Reaction $\text{H} + \text{SO}_2$. *J. Phys. Chem. A* **2006**, *110* (9), 2996–3009.

(143) Goumri, A.; Rocha, J. D. R.; Laakso, D.; Smith, C. E.; Marshall, P. Characterization of Reaction Pathways on the Potential Energy Surfaces for $\text{H} + \text{SO}_2$ and $\text{HS} + \text{O}_2$. *J. Phys. Chem. A* **1999**, *103* (51), 11328–11335.

(144) Frank, A. J.; Sadílek, M.; Ferrier, J. G.; Tureček, F. Sulfur Oxyacids and Radicals in the Gas Phase. A Variable-Time Neutralization-Photoexcitation-Reionization Mass Spectrometric and Ab Initio/RRKM Study. *J. Am. Chem. Soc.* **1997**, *119* (50), 12343–12353.

(145) Alecu, I. M.; Truhlar, D. G. Computational Study of the Reactions of Methanol with the Hydroperoxyl and Methyl Radicals. 1. Accurate Thermochemistry and Barrier Heights. *J. Phys. Chem. A* **2011**, *115* (13), 2811–2829.

- (146) Huo, R.; Zhang, X.; Huang, X.; Li, J.; Sun, C. Direct Ab Initio Dynamics Study of Radical $C_4H(\tilde{X}^2\Sigma^+)$ + CH_4 Reaction. *J. Phys. Chem. A* **2011**, *115* (15), 3576–3582.
- (147) Li, Q.; Wang, C. Y. Direct Dynamic Study on the Hydrogen Abstraction Reaction $CH_3CN + OH \rightarrow CH_2CN + H_2O$. *J. Comput. Chem.* **2004**, *25* (2), 251–257.
- (148) Tachikawa, H.; Kawabata, H. Effects of a Single Water Molecule on the Reaction Barrier of Interstellar CO_2 Formation Reaction. *J. Phys. Chem. A* **2016**, *120* (33), 6596–6603.
- (149) Zanchet, A.; del Mazo, P.; Aguado, A.; Roncero, O.; Jlménez, E.; Canosa, A.; Agúndez, M.; Cernlcharo, J. Full Dimensional Potential Energy Surface and Low Temperature Dynamics of the $H_2CO + OH \rightarrow HCO + H_2O$ Reaction. *Phys. Chem. Chem. Phys.* **2018**, *20* (8), 5415–5426.
- (150) Beckett, D.; Edelman, M.; Raff, J. D.; Raghavachari, K. Hidden Complexities in the Reaction of H_2O_2 and HNO Revealed by: Ab Initio Quantum Chemical Investigations. *Phys. Chem. Chem. Phys.* **2017**, *19* (43), 29549–29560.
- (151) Berman, M. R.; Tsuchiya, T.; Gregušová, A.; Perera, S. A.; Bartlett, R. J. HNNC Radical and Its Role in the $CH + N_2$ Reaction. *J. Phys. Chem. A* **2007**, *111* (29), 6894–6899.
- (152) Louie, M. K.; Francisco, J. S.; Verdicchio, M.; Klippenstein, S. J.; Sinha, A. Hydrolysis of Ketene Catalyzed by Formic Acid: Modification of Reaction Mechanism, Energetics, and Kinetics with Organic Acid Catalysis. *J. Phys. Chem. A* **2015**, *119* (19), 4347–4357.
- (153) Feng, H.; Sun, W.; Xie, Y.; Schaefer, H. F. Is There an Entrance Complex for the

F + NH₃ Reaction? *Chem. Asian J.* **2011**, 6 (11), 3152–3156.

(154) Bacskay, G. B.; Mackie, J. C. Oxidation of CO by SO₂: A Theoretical Study. *J. Phys. Chem. A* **2005**, 109 (9), 2019–2025.

(155) Savchenkova, A. S.; Semenikhin, A. S.; Chechet, I. V.; Matveev, S. G.; Konnov, A. A.; Mebel, A. M. Mechanism and Rate Constants of the CH₂ + CH₂CO Reactions in Triplet and Singlet States: A Theoretical Study. *J. Comput. Chem.* **2019**, 40 (2), 387–399.

(156) Song, H.; Guo, H. Mode Specificity in the HCl + OH → Cl + H₂O Reaction: Polanyi's Rules vs Sudden Vector Projection Model. *J. Phys. Chem. A* **2015**, 119 (5), 826–831.

(157) Dong, H.; Ding, Y.; Sun, C. Radical - Molecule Reactions HCO/HOC + C₂H₂: Mechanistic Study. *J. Phys. Chem. A* **2005**, 109 (51), 11941–11955.

(158) Saheb, V. Quantum Chemical and Theoretical Kinetics Study of the O(³P) + CS₂ Reaction. *J. Phys. Chem. A* **2011**, 115, 4263–4269.

(159) Zhao, Z.; Liu, J.; Wang, L.; Zhang, H.; Hou, C.; Sun, C. A Study of the Hydrogen Abstraction Reactions of C₂H Radical with CH₃CN, C₂H₅CN, and C₃H₇CN by Dual-Level Generalized Transition State Theory. *J. Phys. Chem. A* **2008**, 112 (36), 8455–8463.

(160) Tan, T.; Yang, X.; Krauter, C. M.; Ju, Y.; Carter, E. A. Ab Initio Kinetics of Hydrogen Abstraction from Methyl Acetate by Hydrogen, Methyl, Oxygen, Hydroxyl, and Hydroperoxy Radicals. *J. Phys. Chem. A* **2015**, 119, 6377–6390.

(161) Westley, F. *Table of Recommended Rate Constants for Chemical Reactions Occurring in Combustion*; NSRDS-NBS 67; U.S. Department of Commerce, National

Bureau of Standards: Washington, DC, 1980; pp 8-109.

(162) Grimme, S.; Antony, J.; Ehrlich, S.; Krieg, H. A Consistent and Accurate *Ab Initio* Parametrization of Density Functional Dispersion Correction (DFT-D) for the 94 Elements H-Pu. *J. Chem. Phys.* **2010**, *132* (15), 154104. <https://doi.org/10.1063/1.3382344>.

(163) Weininger, D. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31–36. <https://doi.org/10.1021/ci00057a005>.

(164) O’Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Tim, V.; Hutchison, G. R. Open Babel: An Open Chemical Toolbox. *J. Cheminform.* **2011**, *3*, 1–14. <https://doi.org/10.1186/1758-2946-3-33>.

(165) Yoshikawa, N.; Hutchison, G. R. Fast, Efficient Fragment-based Coordinate Generation for Open Babel. *J. Cheminform.* **2019**, *11*, 1–9. <https://doi.org/10.1186/s13321-019-0372-5>.

(166) NovoPro Home Page, Convert SMILES to 3D structure. <https://www.novoprolabs.com>, 2014 (accessed June 10, 2021).

(167) Kanamori, K.; Toyoura, K.; Honda, J.; Hattori, K.; Seko, A.; Karasuyama, M.; Shitara, K.; Shiga, M.; Kuwabara, A.; Takeuchi, I. Exploring a Potential Energy Surface by Machine Learning for Characterizing Atomic Transport. *Phys. Rev. B* **2018**, *97* (12), 125124. <https://doi.org/10.1103/PhysRevB.97.125124>.

(168) Nigam, J.; Pozdnyakov, S.; Fraux, G.; Ceriotti, M. Unified Theory of Atom-Centered Representations and Message-Passing Machine-Learning Schemes. *J. Chem. Phys.* **2022**, *156* (20), 204115. <https://doi.org/10.1063/5.0087042>.

- (169) Hansen, K.; Biegler, F.; Ramakrishnan, R.; Pronobis, W.; von Lilienfeld, O. A.; Müller, K.; Tkatchenko, A. Machine Learning Predictions of Molecular Properties: Accurate Many-Body Potentials and Nonlocality in Chemical Space. *J. Phys. Chem. Lett.* **2015**, *6* (12), 2326–2331. <https://doi.org/10.1021/acs.jpclett.5b00831>.
- (170) Estrada, E. Characterization of the Folding Degree of Proteins. *Bioinformatics* **2002**, *18*, 697–704. <https://doi.org/10.1093/bioinformatics/18.5.697>.
- (171) Gutman, I.; Trinajstić, N. Graph Theory and Molecular Orbitals. Total π -Electron Energy of Alternant Hydrocarbons. *Chem. Phys. Lett.* **1972**, *17*, 535–538. [https://doi.org/10.1016/0009-2614\(72\)85099-1](https://doi.org/10.1016/0009-2614(72)85099-1).
- (172) Choi, S.; Kim, Y.; Kim, J. W.; Kim, Z.; Kim, W. Y. Feasibility of Activation Energy Prediction of Gas-Phase Reactions by Machine Learning. *Chem. Eur. J.* **2018**, *24*, 12354–12358. <https://doi.org/10.1002/chem.201800345>.
- (173) Morgan, H. L. The Generation of a Unique Machine Description for Chemical Structures - A Technique Developed at Chemical Abstracts Service. *J. Chem. Doc.* **1965**, *5*, 107–113. <https://doi.org/10.1021/c160017a018>.
- (174) Sadeghi, A.; Ghasemi, S. A.; Schaefer, B.; Mohr, S.; Lill, M. A.; Goedecker, S. Metrics for Measuring Distances in Configuration Spaces. *J. Chem. Phys.* **2013**, *139*, 184118. <https://doi.org/10.1063/1.4828704>.
- (175) Grambow, C. A.; Pattanaik, L.; Green, W. H. Deep Learning of Activation Energies. *J. Phys. Chem. Lett.* **2020**, *11*, 2992–2997. <https://dx.doi.org/10.1021/acs.jpclett.0c00500>.
- (176) Landrum, G. RDKit: Open-Source Cheminformatics, 2006. <http://rdkit.org>.

- (177) Behler, J. Perspective: Machine Learning Potentials for Atomistic Simulations. *J. Chem. Phys.* **2016**, *145*, 170901. <https://doi.org/10.1063/1.4966192>.
- (178) Rumble, J. R. *CRC Handbook of Chemistry and Physics*, 104th ed.; CRC Press: Boca Raton, FL, 2023.
- (179) Song, K.; Le, D. CHEMnetBase, University of California Santa Barbara Library. <https://www.library.ucsb.edu/> (accessed July 13, 2023).
- (180) Guyon, I.; Elisseeff, A. An Introduction to Variable and Feature Selection. *J. Mach. Learn. Res.* **2003**, *3*, 1157–1182. <https://doi.org/10.1162/153244303322753616>.
- (181) Ding, C.; Peng, H. Minimum Redundancy Feature Selection from Microarray Gene Expression Data. *J. Bioinform. Comput. Biol.* **2005**, *3* (2), 185–205. <https://doi.org/10.1142/S0219720005001004>.
- (182) Darbellay, G. A.; Vajda, I. Estimation of the Information by an Adaptive Partitioning of the Observation Space. *IEEE Trans. Inf. Theory* **1999**, *45* (4), 1315–1321.
- (183) Editor: Russell D. Johnson III. NIST Computational Chemistry Comparison and Benchmark Database, NIST Standard Reference Database Number 101. <http://cccbdb.nist.gov/>, Release 22, May 2022.
- (184) Miranda-Quintana, R. A.; Heidar-Zadeh, F.; Fias, S.; Chapman, A. E. A.; Liu, S.; Morell, C.; Gómez, T.; Cárdenas, C.; Ayers, P. W. Molecular Interactions from the Density Functional Theory for Chemical Reactivity: Interaction Chemical Potential, Hardness, and Reactivity Principles. *Front. Chem.* **2022**, *10*, 929464. <https://doi.org/10.3389/fchem.2022.929464>.
- (185) Pearson, R. G. Hard and Soft Acids and Bases, HSAB, Part 1: Fundamental

Principles. *J. Chem. Educ.* **1968**, *45*, 581–587. <https://doi.org/10.1021/ed045p581>.

(186) Pearson, R. G. Hard and Soft Acids and Bases, HSAB, Part II: Underlying Theories. *J. Chem. Educ.* **1968**, *45*, 643–648. <https://doi.org/10.1021/ed045p643>.

(187) Nalewajski, R. F. Electrostatic Effects in Interactions between Hard (Soft) Acids and Bases. *J. Am. Chem. Soc.* **1984**, *106*, 944–945. <https://doi.org/10.1021/ja00316a020>.

(188) Chattaraj, P. K.; Lee, H.; Parr, R. G. HSAB Principle. *J. Am. Chem. Soc.* **1991**, *113*, 1855–1856. <https://doi.org/10.1021/ja00005a073>.

(189) Miranda-Quintana, R. A.; Kim, T. D.; Cárdenas, C.; Ayers, P. W. The HSAB Principle from a Finite-Temperature Grand-Canonical Perspective. *Theor. Chem. Acc. B*, *136*, 135. <https://doi.org/10.1007/s00214-017-2167-y>.

(190) Pearson, R. G. Recent Advances in the Concept of Hard and Soft Acids and Bases. *J. Chem. Educ.* **1987**, *64*, 561–567. <https://doi.org/10.1021/ed064p561>.

(191) Zhou, Z.; Parr, R. G. New Measures of Aromaticity: Absolute Hardness and Relative Hardness. *J. Am. Chem. Soc.* **1989**, *111*, 7371–7379. <https://doi.org/10.1021/ja00201a014>.

(192) Parr, R. G.; Chattaraj, P. K. Principle of Maximum Hardness. *J. Am. Chem. Soc.* **1991**, *113*, 1854–1855. <https://doi.org/10.1021/ja00005a072>.

(193) Palke, R. G.; Palke, W. E. Support for a Principle of Maximum Hardness. *J. Phys. Chem.* **1992**, *96*, 3283–3285. <https://doi.org/10.1021/j100187a020>.

(194) Miranda-Quintana, R. A. Note: The Minimum Electrophilicity and the Hard/Soft Acid/Base Principles. *J. Chem. Phys.* **2017**, *146*, 046101. <https://doi.org/10.1063/1.4974987>.

- (195) Fuentealba, P.; Simón-Manso, Y.; Chattaraj, P. K. Molecular Electronic Excitations and the Minimum Polarizability Principle. *J. Phys. Chem. A* **2000**, *104*, 3185–3187. <https://doi.org/10.1021/jp992973v>.
- (196) Morell, C.; Labet, V.; Grand, A.; Chermette, H. Minimum Electrophilicity Principle: An Analysis based upon the Variation of Both Chemical Potential and Absolute Hardness. *Phys. Chem. Chem. Phys.* **2009**, *11*, 3417–3423. <https://doi.org/10.1039/b818534d>.
- (197) Oller, J.; Saez, D. A.; Vöhringer-Martinez, E. Atom-Condensed Fukui Function in Condensed Phases and Biological Systems and its Application to Enzymatic Fixation of Carbon Dioxide. *J. Phys. Chem. A* **2020**, *124* (5), 849–857. <https://doi.org/10.1021/acs.jpca.9b07012>.
- (198) Stewart, J. J. P. Optimization of Parameters for Semiempirical Methods VI: More Modifications to the NDDO Approximations and Re-Optimization of Parameters. *J. Mol. Model.* **2013**, *19*, 1–32. <https://doi.org/10.1007/s00894-012-1667-x>.
- (199) Yang, K.; Swanson, K.; Jin, W.; Coley, C.; Eiden, P.; Gao, H.; Guzman-Perez, A.; Hopper, T.; Kelley, B.; Mathea, M.; Palmer, A.; Settels, V.; Jaakkola, T.; Jensen, K.; Barzilay, R. Analyzing Learned Molecular Representations for Property Prediction. *J. Chem. Inf. Model.* **2019**, *59*, 3370–3388. <https://doi.org/10.1021/acs.jcim.9b00237>.
- (200) Ziurys, L. M. The Chemistry in Circumstellar Envelopes of Evolved Stars: Following the Origin of the Elements to the Origin of Life. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103* (33), 12274–12279.
- (201) Dent, W. R. F.; Wyatt, M. C.; Roberge, A.; Augereau, J. -C.; Casassus, S.; Corder,

S.; Greaves, J. S.; de Gregorio-Monsalvo, I.; Hales, A.; Jackson, A. P.; Meredith Hughes, A.; Lagrange, A. -M.; Matthews, B.; Wilner, D. Molecular Gas Clumps from the Destruction of Icy Bodies in the β Pictoris Debris Disk. *Science* **2014**, *343* (6178), 1490–1492.

(202) Carruthers, G. R. Rocket Observation of Interstellar Molecular Hydrogen. *Astrophys. J.* **1970**, *161*, L81–L85.

(203) Knauth, D. C.; Andersson, B. -G.; McCandliss, S. R.; Moos, H. W. The Interstellar N₂ Abundance towards HD 124314 from Far-ultraviolet Observations. *Nature* **2004**, *429*, 636–638.

(204) Hollis, J. M.; Jewell, P. R.; Lovas, F. J. Confirmation of Interstellar Methylene. *Astrophys. J.* **1995**, *438*, 259–264.

(205) Fuente, A.; García-Burillo, S.; Gerin, M.; Teyssier, D.; Usero, A.; Rizzo, J. R.; de Vicente, P. Photon-dominated Chemistry in the Nucleus of M82: Widespread HOC⁺ Emission in the Inner 650 Parsec Disk. *Astrophys. J.* **2005**, *619*, L155–L158.

(206) Snyder, L. E.; Buhl, D. Observations of Radio Emission from Interstellar Hydrogen Cyanide. *Astrophys. J.* **1971**, *163*, L47–L52.

(207) Whittet, D. C. B.; Walker, H. J. On the Occurrence of Carbon Dioxide in Interstellar Grain Mantles. *Mon. Notices Royal Astron. Soc.* **1991**, *252* (1), 63–67.

(208) European Space Agency. Europe's Space Telescope Iso Finds Water in Distant Places. https://web.archive.org/web/20061222224048/http://www.iso.vilspa.esa.es/outreach/esa_pr/in9712.htm (accessed March 7, 2023).

(209) Ziurys, L. M.; Apponi, A. J.; Hollis, J. M.; Snyder, L. E. Detection of Interstellar N₂O: A New Molecule Containing an N-O Bond. *Astrophys. J. Lett.* **1994**, *436*, L181–

L184.

(210) Snyder, L. E.; Buhl, D.; Zuckerman, B.; Palmer, P. Microwave Detection of Interstellar Formaldehyde. *Phys. Rev. Lett.* **1969**, 22 (13), 679–681.

(211) Ridgway, S. T.; Hall, D. N. B.; Kleinmann, S. G.; Weinberger, D. A.; Wojslaw, R. S. Circumstellar Acetylene in the Infrared Spectrum of IRC +10° 216. *Nature* **1976**, 264, 345–346.

(212) Johansson, L. E. B.; Andersson, C.; Elldér, J.; Friberg, P.; Hjalmarson, Å.; Höglund, B.; Irvine, W. M.; Olofsson, H.; Rydbeck, G. Spectral Scan of Orion A and IRC + 10216 from 72 to 91 GHz. *Astron. Astrophys.* **1984**, 130, 227–256.

(213) Marcelino, N.; Cernicharo, J.; Tercero, B.; Roueff, E. Discovery of Fulminic Acid, HCNO, in Dark Clouds. *Astrophys. J.* **2009**, 690, L27–L30.

(214) Guélin, M.; Cernicharo, J. Astronomical Detection of the Hccn Radical. Toward a New Family of Carbon-Chain Molecules? *Astron. Astrophys.* **1991**, 244, L21–L24.

(215) Agúndez, M.; Marcelino, N.; Cernicharo, J. Discovery of Interstellar Isocyanogen (Cncn): Further Evidence That Dicyanopolynes Are Abundant in Space. *Astrophys. J. Lett.* **2018**, 861, L22.

(216) Brünken, S.; Belloche, A.; Martín, S.; Verheyen, L.; Menten, K. M. Interstellar HOCN in the Galactic Center Region. *Astron. Astrophys.* **2010**, 516, A109.

(217) Rieu, N. Q.; Graham, D.; Bujarrabal, V. Ammonia and Cyanotriacetylene in the Envelopes of CRL 2688 and IRC + 10216. *Astron. Astrophys.* **1984**, 138, L5–L8.

(218) Lacy, J. H.; Carr, J. S.; Evans II, N. J.; Baas, F.; Achtermann, J. M.; Arens, J. F. Discovery of Interstellar Methane: Observations of Gaseous and Solid CH₄ Absorption

toward Young Stars in Molecular Clouds. *Astrophys. J.* **1991**, 376, 556–560.

(219) Dickens, J. E.; Irvine, W. M.; DeVries, C. H.; Ohishi, M. Hydrogenation of Interstellar Molecules: A Survey for Methylenimine (CH_2NH). *Astrophys. J.* **1997**, 479, 307–312.

(220) McGuire, B. A.; Loomis, R. A.; Charness, C. M.; Corby, J. F.; Blake, G. A.; Hollis, J. M.; Lovas, F. J.; Jewell, P. R.; Remijan, A. J. Interstellar Carbodiimide (HNCNH): A New Astronomical Detection From the Gbt Primos Survey via Maser Emission Features. *Astrophys. J. Lett.* **2012**, 758, L33.

(221) Turner, B. E.; Liszt, H. S.; Kaifu, N.; Kisliakov, A. G. Microwave Detection of Interstellar Cyanamide. *Astrophys. J.* **1975**, 201, L149–L152.

(222) Ligterink, N. F. W.; El-Abd, S. J.; Brogan, C. L.; Hunter, T. R.; Remijan, A. J.; Garrod, R. T.; McGuire, B. M. The Family of Amide Molecules toward NGC 6334I. *Astrophys. J.* **2020**, 901, 37.

(223) Kawaguchi, K.; Ohishi, M.; Ishikawa, S. Detection of Isocyanoacetylene HCCNC in TMC-1. *Astrophys. J.* **1992**, 386, L51–L53.

(224) Remijan, A. J.; Hollis, J. M.; Lovas, F. J.; Stork, W. D.; Jewell, P. R.; Meier, D. S. Detection of Interstellar Cyanoformaldehyde (CNCHO). *Astrophys. J.* **2008**, 675, L85–L88.

(225) Zuckerman, B.; Ball, J. A.; Gottlieb, C. A. Microwave Detection of Interstellar Formic Acid. *Astrophys. J.* **1971**, 163, L41–L45.

(226) Liu, S.; Mehringer, D. M.; Snyder, L. E. Observations of Formic Acid in Hot Molecular Cores. *Astrophys. J.* **2001**, 552, 654–663.

- (227) Walmsley, C. M.; Winnewisser, G.; Toelle, F. Cyanoacetylene and Cyanodiacetylene in Interstellar Clouds. *Astron. Astrophys.* **1980**, *81*, 245–250.
- (228) Remijan, A. J.; Hollis, J. M.; Lovas, F. J.; Plusquellic, D. F.; Jewell, P. R. Interstellar Isomers: The Importance of Bonding Energy Differences. *Astrophys. J.* **2005**, *632*, 333–339.
- (229) Betz, A. L. Ethylene in IRC +10216. *Astrophys. J.* **1981**, *244*, L103–L105.
- (230) Irvine, W. M.; Brown, R. D.; Cragg, D. M.; Friberg, P.; Godfrey, P. D.; Kaifu, N.; Matthews, H. E.; Ohishi, M.; Suzuki, H.; Takeo, H. A New Interstellar Polyatomic Molecule: Detection of Propynal in the Cold Cloud TMC-1. *Astrophys. J.* **1988**, *335*, L89–L93.
- (231) Hollis, J. M.; Lovas, F. J.; Remijan, A. J.; Jewell, P. R.; Ilyushin, V. V.; Kleiner, I. Detection of Acetamide (CH_3CONH_2): The Largest Interstellar Molecule with a Peptide Bond. *Astrophys. J.* **2006**, *643*, L25–L28.
- (232) Cernicharo, J.; Heras, A. M.; Tielens, A. G. G. M.; Pardo, J. R.; Herpin, F.; Guélin, M.; Waters, L. B. F. M. Infrared Space Observatory's Discovery of C_4H_2 , C_6H_2 , and Benzene in CRL 618. *Astrophys. J.* **2001**, *546*, L123–L126.
- (233) Zaleski, D. P.; Seifert, N. A.; Steber, A. L.; Muckle, M. T.; Loomis, R. A.; Corby, J. F.; Martinez Jr., O.; Crabtree, K. N.; Jewell, P. R.; Hollis, J. R. Detection of E-Cyanomethanimine Toward Sagittarius B2(N) In the Green Bank Telescope Primos Survey. *Astrophys. J.* **2013**, *765*, L10.
- (234) National Radio Astronomy Observatory. Researchers Use NRAO Telescope to Study Formation of Chemical Precursors to Life.

<https://www.nrao.edu/pr/2006/gbtmolecules/> (accessed March 7, 2023).

(235) Hollis, J. M.; Remijan, A. J.; Jewell, P. R.; Lovas, F. J. Cyclopropenone ($c\text{-H}_2\text{C}_3\text{O}$): A New Interstellar Ring Molecule. *Astrophys. J.* **2006**, *642*, 933–939.

(236) Halfen, D. T.; Ilyushin, V. V.; Ziurys, L. M. Interstellar Detection of Methyl Isocyanate CH_3NCO in Sgr B2(N): A Link from Molecular Clouds to Comets. *Astrophys. J. Lett.* **2015**, *812*, L5.

(237) Zeng, S.; Quénard, D.; Jiménez-Serra, I.; Martín-Pintado, J.; Rivilla, V. M.; Testi, L.; Martín-Doménech, R. First Detection of the Pre-biotic Molecule Glycolonitrile (HOCH_2CN) In the Interstellar Medium. *Mon. Not. R. Astron. Soc.: Lett.* **2019**, *484* (1), L43–L48.

(238) Kaifu, N.; Takagi, K.; Kojima, T. Excitation of Interstellar Methylamine. *Astrophys. J.* **1975**, *198*, L85–L88.

(239) Bizzocchi, L.; Prudenzeno, D.; Rivilla, V. M.; Pietropolli-Charmet, A.; Giuliano, B. M.; Caselli, P.; Martín-Pintado, J.; Jiménez-Serra, I.; Martín, S.; Requena-Torres, M. A.; Rico-Villas, F.; Zeng, S.; Guillemin, J. -C. Propargylimine in the Laboratory and in Space: Millimetre-Wave Spectroscopy and Its First Detection in the ISM. *Astron. Astrophys.* **2020**, *640*, A98.

(240) Dickens, J. E.; Irvine, W. M.; Ohishi, M.; Ikeda, M.; Ishikawa, S.; Nummelin, A.; Hjalmarson, Å. Detection of Interstellar Ethylene Oxide ($c\text{-C}_2\text{H}_4\text{O}$). *Astrophys. J.* **1997**, *489*, 753–757.

(241) National Radio Astronomy Observatory. Scientists Toast the Discovery of Vinyl Alcohol in Interstellar Space. <https://www.nrao.edu/pr/2001/vinylalco/> (accessed March 7,

2023).

(242) Belloche, A.; Menten, K. M.; Comito, C.; Müller, H. S. P.; Schilke, P.; Ott, J.; Thorwirth, S.; Hieret, C. Detection of Amino Acetonitrile in Sgr B2(N). *Astron. Astrophys.* **2008**, *482*, 179–196.

(243) Remijan, A. J.; Snyder, L. E.; McGuire, B. A.; Kuo, H.; Looney, L. W.; Friedel, D. N.; Golubiatnikov, G. Y.; Lovas, F. J.; Ilyushin, V. V.; Alekseev, E. A.; Dyubko, S. F.; McCall, B. J.; Hollis, J. M. Observational Results of a Multi-Telescope Campaign in Search of Interstellar Urea [(NH₂)₂CO]. *Astrophys. J.* **2014**, *783* (77).

(244) Mehringer, D. M.; Snyder, L. E.; Miao, Y. Detection and Confirmation of Interstellar Acetic Acid. *Astrophys. J.* **1997**, *480*, L71–L74.

(245) Loomis, R. A.; Zaleski, D. P.; Steber, A. L.; Neill, J. L.; Muckle, M. T.; Harris, B. J.; Hollis, J. M.; Jewell, P. R.; Lattanzi, V.; Lovas, F. J. The Detection of Interstellar Ethanamine (CH₃CHNH) from Observations Taken During the Gbt Primos Survey. *Astrophys. J. Lett.* **2013**, *765*, L9.

(246) McGuire, B. A.; Burkhardt, A. M.; Loomis, R. A.; Shingledecker, C. N.; Lee, K. L. K.; Charnley, S. B.; Cordiner, M. A.; Herbst, E.; Kalenskii, S.; Momjian, E. Early Science from GOTHAM: Project Overview, Methods, and the Detection of Interstellar Propargyl Cyanide (HCCCH₂CN) in TMC-1. *Astrophys. J. Lett.* **2020**, *900*, L10.

(247) Lovas, F. J.; Remijan, A. J.; Hollis, J. M.; Jewell, P. R.; Snyder, L. E. Hyperfine Structure Identification of Interstellar Cyanoallene toward TMC-1. *Astrophys. J.* **2006**, *637*, L37–L40.

(248) Zeng, S.; Jiménez-Serra, I.; Rivilla, V. M.; Martín-Pintado, J.; Rodríguez-Almeida,

L. F.; Tercero, B.; de Vicente, P.; Rico-Villas, F.; Colzi, L.; Martín, S.; Requena-Torres, M. A. Probing the Chemical Complexity of Amines in the ISM: Detection of Vinylamine ($\text{C}_2\text{H}_3\text{NH}_2$) and Tentative Detection of Ethylamine ($\text{C}_2\text{H}_5\text{NH}_2$). *Astrophys. J. Lett.* **2021**, *920*, L27.

(249) Snyder, L. E.; Buhl, D.; Schwartz, P. R.; Clark, F. O.; Johnson, D. R.; Lovas, F. J.; Giguere, P. T. Radio Detection of Interstellar Dimethyl Ether. *Astrophys. J.* **1974**, *191*, L79–L82.

(250) Zuckerman, B.; Turner, B. E.; Johnson, D. R.; Clark, F. O.; Lovas, F. J.; Fourikis, N.; Palmer, P.; Morris, M.; Lilley, A. E.; Ball, J. A.; Gottlieb, C. A.; Litvak, M. M.; Penfield, H. Detection of Interstellar Trans-Ethyl Alcohol. *Astrophys. J.* **1975**, *196*, L99–L102.

(251) Marcelino, N.; Cernicharo, J.; Agúndez, M.; Roueff, E.; Gerin, M.; Martín-Pintado, J.; Mauersberger, R.; Thum, C. Discovery of Interstellar Propylene (CH_2CHCH_3): Missing Links in Interstellar Gas-Phase Chemistry. *Astrophys. J.* **2007**, *665*, L127–L130.

(252) Remijan, A. J.; Hollis, J. M.; Snyder, L. E.; Jewell, P. R.; Lovas, F. J. Methyltriacetylene ($\text{CH}_3\text{C}_6\text{H}$) toward TMC-1: The Largest Detected Symmetric Top. *Astrophys. J.* **2006**, *643*, L37–L40.

(253) McGuire, B. A.; Carroll, P. B.; Loomis, R. A.; Finneran, I. A.; Jewell, P. R.; Remijan, A. J.; Blake, G. A. Discovery of the Interstellar Chiral Molecule Propylene Oxide ($\text{CH}_3\text{CHCH}_2\text{O}$). *Science* **2016**, *352* (6292), 1449–1452.

(254) Snyder, L. E.; Lovas, F. J.; Mehringer, D. M.; Miao, N. Y.; Kuan, Y.; Hollis, J. M.; Jewell, P. R. Confirmation of Interstellar Acetone. *Astrophys. J.* **2002**, *578*, 245–255.

- (255) Rivilla, V. M.; Jiménez-Serra, I.; Martín-Pintado, J.; Briones, C.; Rodríguez-Almeida, L. F.; Rico-Villas, F.; Tercero, B.; Zeng, S.; Colzi, L.; de Vicente, P.; Martín, S.; Requena-Torres, M. A. Discovery in Space of Ethanolamine, the Simplest Phospholipid Head Group. *Proc. Natl. Acad. Sci. U.S.A.* **2021**, *118* (22), e2101314118.
- (256) Tercero, B.; Kleiner, I.; Cernicharo, J.; Nguyen, H. V. L.; López, A.; Muñoz Caro, G. M. Discovery of Methyl Acetate and Gauche Ethyl Formate in Orion. *Astrophys. J. Lett.* **2013**, *770*, L13.
- (257) Belloche, A.; Garrod, R. T.; Müller, H. S. P.; Menten, K. M. Detection of a Branched Alkylmolecule in the Interstellar Medium: Iso-Propyl Cyanide. *Science* **2014**, *345* (6204), 1584–1587.
- (258) Belloche, A.; Garrod, R. T.; Müller, H. S. P.; Menten, K. M.; Comito, C.; Schilke, P. Increased Complexity in Interstellar Chemistry: Detection and Chemical Modeling of Ethyl Formate and N-Propyl Cyanide in Sagittarius B2(N). *Astron. Astrophys.* **2009**, *499*, 215–232.
- (259) Fuchs, G. W.; Fuchs, U.; Giesen, T. F.; Wyrowski, F. Trans-Ethyl Methyl Ether in Space: A New Look at a Complex Molecule in Selected Hot Core Regions. *Astron. Astrophys.* **2005**, *444*, 521–530.
- (260) Guélin, M.; Cernicharo, J. Organic Molecules in Interstellar Space: Latest Advances. *Front. Astron. Space Sci.* **2022**, *9*. <https://doi.org/10.3389/fspas.2022.787567>.
- (261) Herbst, E. Unusual Chemical Processes in Interstellar Chemistry: Past and Present. *Front. Astron. Space Sci.* **2021**, *8*.
- (262) MacKee, T.; MacKee, J. R. Amino Acids, Peptides, and Proteins. *Biochemistry:*

The Molecular Basis of Life, 6th ed.; Oxford University Press: New York, 2016; pp 130–135.

(263) MacKee, T.; MacKee, J. R.; Sehorn, M. G. Amino Acids, Peptides, and Proteins. *Student Study Guide/Solutions Manual for Use with Biochemistry: The Molecular Basis of Life*, 6th ed.; Oxford University Press: New York, 2016; pp 59–61.

(264) Mathews, C. K.; Van Holde, K. E.; Appling D. R.; Anthony-Cahill, S. J. Introduction to Proteins: The Primary Level of Proteins Structure. *Biochemistry*, 4th ed.; Pearson: Toronto, 2013; pp 138–139.

(265) Otaki, J. M.; Tsutsumi, M.; Gotoh, T.; Yamamoto, H. Secondary Structure Characterization Based on Amino Acid Composition and Availability in Proteins. *J. Chem. Inf. Model.* **2010**, *50* (4), 690–700.

(266) Mathews, C. K.; Van Holde, K. E.; Appling D. R.; Anthony-Cahill, S. J. Carbohydrates: Sugars, Saccharides, Glycans. *Biochemistry*, 4th ed.; Pearson: Toronto, 2013; pp 309–311.

(267) Chandru, K.; Guttenberg, N.; Giri, C.; Hongo, Y.; Butch, C.; Mamajanov, I.; Cleaves II, H. J. Simple Prebiotic Synthesis of High Diversity Dynamic Combinatorial Polyester Libraries. *Commun. Chem.* **2018**, *1* (30).

(268) Mao, F.; Ni, W.; Xu, X.; Wang, H.; Wang, J.; Ji, M.; Li, J. Chemical Structure-Related Drug-Like Criteria of Global Approved Drugs. *Molecules* **2016**, *21* (75).

(269) Iglesias-Groth, S. A Search for Tryptophan in the Gas of the IC 2348 Star Cluster of the Perseus Molecular Cloud. *Mon. Notices Royal Astron. Soc.* **2023**, *523* (2), 2876–2886. <https://doi.org/10.1093/mnras/stad1535>.

- (270) Panda, S.; Chiranjibi, A.; Awasthi, D.; Ghosal, S.; Anoop, A. Exploring the Chemical Space of C_2H_3NO Isomers and Bimolecular Reactions with Hydrogen Cyanide and Formaldehyde: Insights into the Emergence of Life. *ACS Earth Space Chem.* **2023**, *7*, 1739–1752. <https://doi.org/10.1021/acsearthspacechem.3c00113>.
- (271) Li, Y.; Liu, H.-L.; Sun, Y.-B.; Li, Z.; Huang, X.-R.; Sun, C.-C. Radical Reaction $HCNO + {}^3NH$: A Mechanistic Study. *Theor. Chem. Acc.* **2009**, *124*, 123–137. <https://doi.org/10.1007/s00214-009-0591-3>.
- (272) Gao, Y.; Jia, X.-J.; Li, S.; Yu, Y.-B.; Wang, R.-S.; Pan, X.-M. Theoretical Study on the Singlet and Triplet Potential Energy Surfaces of $NH(X^3\Sigma^-) + HCNO$ Reaction. *Theor Chem Acc* **2010**, *127*, 81–94. <https://doi.org/10.1007/s00214-009-0707-9>.
- (273) Miller, J. A.; Klippenstein, S. J.; Glarborg, P. A Kinetic Issue in Reburning: the Fate of $HCNO$. *Combust. Flame* **2003**, *135* (3), 357–362. <https://doi.org/10.1016/j.combustflame.2003.07.002>.
- (274) Nguyen, H. M. T.; Nguyen, T. N. Calculations on the Complex Mechanism of the $HCNO + OH$ Reaction. *Chem. Phys. Lett.* **2014**, *599*, 15–22. <https://doi.org/10.1016/j.cplett.2014.03.001>.
- (275) Feng, W.; Meyer, J. P.; Hershberger, J. F. Kinetics of the $OH + HCNO$ Reaction. *J. Phys. Chem. A* **2006**, *110*, 4458–4464. <https://doi.org/10.1021/jp058305t>.
- (276) Nguyen, T. N.; Nguyen, H. M. T. Prediction of the Product Channels in the Reaction of the Methyl Radical with Fulminic Acid. *Int. J. Chem. Kinet.* **2021**, *53*, 479–491. <https://doi.org/10.1002/kin.21458>.
- (277) Pang, J.-L.; Xie, H.-B.; Zhang, S.-W.; Ding, Y.-H.; Tang, A.-Q. Theoretical Study

on Reaction Mechanism of Fulminic Acid HCNO with CN Radical. *J. Phys. Chem. A* **2008**, *112* (23), 5251–5257. <https://doi.org/10.1021/jp709700u>.

(278) Mu, D.; Kong, W.-X.; Wang, S. A Density Functional Investigation of the Reaction Mechanism of $\text{H}_2\text{O} + \text{HCNO}$. *Int. J. Quantum Chem.* **2011**, *111*, 165–173. <https://doi.org/10.1002/qua.22745>.

(279) Feng, W.; Hersherberger, J. F. Theoretical Study of the $\text{O}(^3\text{P}) + \text{HCNO}$ Reaction. *Chem. Phys. Lett.* **2008**, *457*, 307–311. <https://doi.org/10.1016/j.cplett.2008.04.012>.

(280) Pham, T. V.; Tran, A. V. Gas Phase Reaction of Isocyanic Acid: Kinetics, Mechanisms, and Formation of Isopropyl Aminocarbonyl. *ACS Omega* **2021**, *6* (50), 34661–34674. <https://doi.org/10.1021/acsomega.1c05063>.

(281) Tully, F. P.; Perry, R. A.; Thorne, L. R.; Allendorf, M. D. Free-Radical Oxidation of Isocyanic Acid. *Symp. (Int.) Combust.* **1989**, *22* (1), 1101–1106. [https://doi.org/10.1016/S0082-0784\(89\)80120-1](https://doi.org/10.1016/S0082-0784(89)80120-1).

(282) Tsang, W. Chemical Kinetic Data Base for Propellant Combustion. II. Reactions Involving CN, NCO, and HNCO. *J. Phys. Chem. Ref. Data* **1992**, *21*, 753–791. <https://doi.org/10.1063/1.555914>.

(283) Li, L.-C.; Zheng, Y.; Zha, D.; Tian, A.-M.; Xu, M.-H. Theoretical Studies on the Reaction Mechanism of CH_2CH Radical with HNCO. *Int. J. Quantum Chem.* **2006**, *106*, 1672–1682. <https://doi.org/10.1002/qua.20917>.

(284) Ji, Y.-Q.; Feng, W.-L.; Xu, Z.-F.; Lei, M.; Hao, M.-R. MP2 and QCISD Study of Hydrogen Transfer Reaction Path of the Reaction HNCO with Carbon-Hydrogen Radicals CH_x ($x = 1\sim 3$). *Acta Chim. Sinica* **2002**, *60*, 1167–1172.

(285) Redondo, P.; Sanz-Novo, M.; Barrientos, C.; Largo, A. Structure, Spectroscopic Signatures, and Formation of Hydroxy-Azirine: A Potential Interstellar Prebiotic Molecule. *Mon. Notices Royal Astron. Soc.* **2023**, *520* (2), 1659–1671. <https://doi.org/10.1093/mnras/stad205>.

(286) Redondo, P.; Largo, A.; Barrientos, C. Structure and Spectroscopic Properties of Imine Acetaldehyde: A Possible Interstellar Molecule. *Mon. Notices Royal Astron. Soc.* **2018**, *478* (3), 3042–3048. <https://doi.org/10.1093/mnras/sty1175>.

(287) Fourré, I.; Matz, O.; Ellinger, Y.; Guillemin, J.-C. Relative Thermodynamic Stability of the [C,N,O] Linkages as an Indication of the Most Abundant Structures in the ISM. *Astron. Astrophys.* **2020**, *639*, A16. <https://doi.org/10.1051/0004-6361/202037839>.

(288) NIST Chemistry WebBook, NIST Standard Reference Database Number 69. <https://doi.org/10.18434/T4D303> (accessed July 17, 2023).

(289) Badawi, H. M. DFT and MP2 Vibrational Spectra and Assignments for *Gauche* *N*-Methyleneformamide $\text{CH}_2=\text{N}-\text{CHO}$. *Theochem (J. Mol. Struct.)* **2002**, *617*, 9–15. [https://doi.org/10.1016/S0166-1280\(02\)00395-0](https://doi.org/10.1016/S0166-1280(02)00395-0).

(290) Afanas'yev, D. Y.; Prosyaniuk, A. V. Why *N*-Methyleneformamide, $\text{CH}_2=\text{N}-\text{CHO}$, Prefers *Gauche*-Conformation? A DFT/NBO Study. *Theochem (J. Mol. Struct.)* **2005**, *730*, 45–49. <https://doi.org/10.1016/j.theochem.2005.05.001>.

(291) Schilke, P.; Benford, D. J.; Hunter, T. R.; Lis, D. C.; Phillips, T. G. A Line Survey of Orion-KL from 607 to 725 GHz. *Astrophys. J.* **2001**, *132*, 281–364. <https://doi.org/10.1086/318951>.

(292) Schilke, P.; Comito, C.; Thorwirth, S. First Detection of Vibrationally Excited

HNC in Space. *Astrophys. J.* **2003**, 582, L101–L104. <https://doi.org/10.1086/367628>.

(293) (a) We acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC). This research was enabled in part by support provided by Sharcnet (sharcnet.ca) and the Digital Research Alliance of Canada (alliancecan.ca).

(b) Graham. <https://docs.alliancecan.ca/wiki/Graham> (accessed October 31, 2018).

(294) Minissale, M.; Congiu, E.; Dulieu, F. Oxygen Diffusion and Reactivity at Low Temperature on Bare Amorphous Olivine-Type Silicate. *J. Chem. Phys.* **2014**, 140 (7), 074705. <https://doi.org/10.1063/1.4864657>.

(295) Wang, Z.-C.; Bierbaum, V. M. Reactions of Substituted Benzene Anions with N and O Atoms: Chemistry in Titan’s Upper Atmosphere and the Interstellar Medium. *J. Chem. Phys.* **2016**, 144, 214304. <https://doi.org/10.1063/1.4952454>.

(296) Goumans, T. P. M.; Catlow, C. R. A.; Brown, W. A.; Kästner, J.; Sherwood, P. An Embedded Cluster Study of the Formation of Water on Interstellar Dust Grains. *Phys. Chem. Chem. Phys.* **2009**, 11, 5431–5436. <https://doi.org/10.1039/B816905E>.

(297) Ward, M. D.; Price, S. D. Thermal Reactions of Oxygen Atoms With Alkenes at Low Temperatures on Interstellar Dust. *Astrophys. J.* **2011**, 741 (2), 121. <http://dx.doi.org/10.1088/0004-637X/741/2/121>.

(298) Pearce, M. P.; Bussemaker, M. J.; Cooper, P. D.; Lapere, K. M.; Wild, D. A.; McKinley, A. J. Formation of Methanol from Methane and Water in An Electrical Discharge. *Phys. Chem. Chem. Phys.* **2012**, 14, 3444–3449. <https://doi.org/10.1039/C2CP22135G>.

(299) Nhlabatsi, Z.P.; Bhasi, P.; Sitha, S. Hemiaminal Route for the Formation of

Interstellar Glycine: A Computational Study. *J. Mol. Model* **2019**, *25*, 335.
<https://doi.org/10.1007/s00894-019-4224-z>.

(300) Joshi, P.R.; Lee, Y.P. A Chemical Link between Methylamine and Methylene Imine and Implications for Interstellar Glycine Formation. *Commun. Chem.* **2022**, *5*, 62.
<https://doi.org/10.1038/s42004-022-00677-5>.

(301) Thripati, S. Computational Studies on the Possible Formation of Glycine via Open Shell Gas-Phase Chemistry in the Interstellar Medium. *Org. Biomol. Chem.* **2022**, *20*, 4189–4203. <https://doi.org/10.1039/D2OB00407K>.

(302) (a) NIST Computational Chemistry Comparison and Benchmark DataBase (CCCBDB) VII.D.3. Converting between rotational constants and moments of inertia. <https://cccbdb.nist.gov/convertmomint.asp> (accessed September 15, 2023);

(b) NIST Computational Chemistry Comparison and Benchmark Database. NIST Standard Reference Database Number 101. Release 22, May 2022, Editor: Russell D. Johnson III. <http://cccbdb.nist.gov/> (accessed September 15, 2023).

(303) Callen, H. B. *Thermodynamics and an Introduction to Thermostatistics*, 2nd ed.; John Wiley & Sons: New York, 1985.

(304) von Lilienfeld, O. A. Quantum Machine Learning in Chemical Compound Space. *Angew. Chem. Int. Ed.* **2018**, *57*, 4164. <https://doi.org/10.1002/anie.201709686>.

(305) Huang, B.; von Lilienfeld, O. A. Ab Initio Machine Learning in Chemical Compound Space. *Chem. Rev.* **2021**, *121*, 10001–10036.
<https://doi.org/10.1021/acs.chemrev.0c01303>.

(306) Heinen, S. N. Quantum Machine Learning Applied to Chemical Reaction Space.

Ph.D. dissertation, University of Basel, Basel, Switzerland, 2022.

(307) Meyer, B.; Sawatlon, B.; Heinen, S.; von Lilienfeld, O. A.; Corminboeuf, C. Machine Learning Meets Volcano Plots: Computational Discovery of Cross-Coupling Catalysts. *Chem. Sci.* **2018**, *9*, 7069–7077. <https://doi.org/10.1039/c8sc01949e>.

(308) Huang, B.; von Lilienfeld, O. A. Communication: Understanding Molecular Representations in Machine Learning: The Role of Uniqueness and Target Similarity. *J. Chem. Phys.* **2016**, *145*, 161102. <https://doi.org/10.1063/1.4964627>.

(309) Rappe, A. K.; Casewit, C. J.; Colwell, K. S.; Goddard III, W. A.; Skiff, W. M. UFF, A Full Periodic Table Force Field for Molecular Mechanics and Molecular Dynamics Simulations. *J. Am. Chem. Soc.* **1992**, *114*, 10024–10035.

(310) Heinen, S.; von Rudorff, G. F.; von Lilienfeld, O. A. Toward the Design of Chemical Reactions: Machine Learning Barriers of Competing Mechanisms in Reactant Space. *J. Chem. Phys.* **2021**, *155*, 064105. <https://doi.org/10.1063/5.0059742>.

(311) Kayala, M. A.; Azencott, C.-A.; Chen, J. H.; Baldi, P. Learning to Predict Chemical Reactions. *J. Chem. Inf. Model.* **2011**, *51* (9), 2209–2222. <https://doi.org/10.1021/ci200207y>.

(312) Wei, J. N.; Duvenaud, D.; Aspuru-Guzik, A. Neural Networks for the Prediction of Organic Chemistry Reactions. *ACS Cent. Sci.* **2016**, *2* (10), 725–732. <https://doi.org/10.1021/acscentsci.6b00219>.

(313) Sandfort, F.; Strieth-Kalthoff, F.; Kühnemund, M.; Beecks, C.; Glorius, F. A Structure-Based Platform for Predicting Chemical Reactivity. *Chem* **2020**, *6* (6), 1379–1390. <https://doi.org/10.1016/j.chempr.2020.02.017>.

- (314) Fooshee, D.; Mood, A.; Gutman, E.; Tavakoli, M.; Urban, G.; Liu, F.; Huynh, N.; Van Vranken, D.; Baldi, P. Deep Learning for Chemical Reaction Prediction. *Mol. Syst. Des. Eng.* **2018**, *3*, 442–452. <https://doi.org/10.1039/C7ME00107J>.
- (315) Schwaller, P.; Gaudin, T.; Lányi, D.; Bekas, C.; Laino, T. “Found in Translation”: Predicting Outcomes of Complex Organic Chemistry Reactions using Neural Sequence-to-Sequence Models. *Chem. Sci.* **2018**, *9*, 6091–6098. <https://doi.org/10.1039/C8SC02339E>.
- (316) von Rudorff, G. F.; Heinen, S. N.; Bragato, M.; von Lilienfeld, O. A. Thousands of Reactants and Transition States for Competing E2 and S_N2 Reactions. *Mach. Learn.: Sci. Technol.* **2020**, *1*, 045026. <https://doi.org/10.1088/2632-2153/aba822>.
- (317) Lemm, D.; von Rudorff, G. F.; von Lilienfeld, O. A. Machine Learning Based Energy-Free Structure Predictions of Molecules, Transition States, and Solids. *Nat. Commun.* **2021**, *12*, 4468. <https://doi.org/10.1038/s41467-021-24525-7>.
- (318) (a) Heinen, S.; von Rudorff, G. F.; von Lilienfeld, O. A. Transition State Search and Geometry Relaxation throughout Chemical Compound Space with Quantum Machine Learning. *J. Chem. Phys.* **2022**, *157*, 221102. <https://doi.org/10.1063/5.0112856>;
- (b) Christensen, A. S.; Faber, F. A.; von Lilienfeld, O. A. Operators in Quantum Machine Learning: Response Properties in Chemical Space. *J. Chem. Phys.* **2019**, *150* (6), 064105. <https://doi.org/10.1063/1.5053562>;
- (c) Christensen, A. S.; von Lilienfeld, O. A. On the Role of Gradients for Machine Learning of Molecular Energies and Forces. *Mach. Learn.: Sci. Technol.* **2020**, *1*, 045018. <https://doi.org/10.1088/2632-2153/abba6f>.
- (319) (a) Liu, D. C.; Nocedal, J. On the Limited Memory BFGS Method for Large Scale

Optimization. *Math. Program.* **1989**, *45* (1–3), 503–528.
<https://doi.org/10.1007/BF01589116>;

(b) Numerical Optimization: Understanding L-BFGS.
<https://aria42.com/blog/2014/12/understanding-lbfgs> (accessed December 2, 2023).

(320) Christensen, A. S.; Bratholm, L. A.; Faber, F. A.; von Lilienfeld, O. A. FCHL Revisited: Faster and More Accurate Quantum Machine Learning. *J. Chem. Phys.* **2020**, *152*, 044107. <https://doi.org/10.1063/1.5126701>.

(321) Khan, D.; Heinen, S.; von Lilienfeld, O. A. Kernel Based Quantum Machine Learning at Record Rate: Many-Body Distribution Functionals as Compact Representations. *J. Chem. Phys.* **2023**, *159*, 034106. <https://doi.org/10.1063/5.0152215>.

(322) von Lilienfeld, O. A.; Müller, K.-R.; Tkatchenko, A. Exploring Chemical Compound Space with Quantum-Based Machine Learning. *Nat. Rev. Chem.* **2020**, *4*, 347–358. <https://doi.org/10.1038/s41570-020-0189-9>.

(323) Ramakrishnan, R.; Dral, P. O.; Rupp, M.; von Lilienfeld, O. A. Big Data Meets Quantum Chemistry Approximations: The Δ -Machine Learning Approach. *J. Chem. Theory Comput.* **2015**, *11*, 2087–2096. <https://doi.org/10.1021/acs.jctc.5b00099>.

(324) Zhang, Q.-s.; Zhu, S.-c. Visual Interpretability for Deep Learning: A Survey. *Front. Inform. Technol. Electron. Eng.* **2018**, *19*, 27–39. <https://doi.org/10.1631/FITEE.1700808>.

(325) Rudin, C. Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. *Nat. Mach. Intell.* **2019**, *1*, 206–215. <https://doi.org/10.1038/s42256-019-0048-x>.

(326) Schore, N. E.; Vollhardt, K. P. C. *Organische Chemie*; John Wiley & Sons, 2011.

Appendix

Appendix A: Molecular Structure Information of All Studied Species in

Chapter 5

$^1\text{CH}_2$

Element Symbol	X	Y	Z
C	0.00000000	0.00000000	0.17441000
H	0.00000000	0.86308100	-0.52322900
H	0.00000000	-0.86308100	-0.52322900

$^3\text{CH}_2$

Element Symbol	X	Y	Z
C	0.00000000	0.00000000	0.10487800
H	0.00000000	0.99545500	-0.31463500
H	0.00000000	-0.99545500	-0.31463500

HCNO

Element Symbol	X	Y	Z
O	0.00000000	0.00000000	-1.17321700
N	0.00000000	0.00000000	0.01889800
C	0.00000000	0.00000000	1.17032500
H	0.00000000	0.00000000	2.23150100

HNCO

Element Symbol	X	Y	Z
-------------------	---	---	---

O	-0.42744800	1.12628500	0.00000000
N	0.31021800	-1.11827300	0.00000000
C	0.00000000	0.04857200	0.00000000
H	1.24805300	-1.47379900	0.00000000

CH₃C(=O)NH₂

Element Symbol	X	Y	Z
C	-1.35732100	-0.33787000	0.00080800
C	0.07588800	0.14663100	0.00837700
N	1.02620800	-0.82645900	0.00327600
H	1.99102800	-0.54971600	-0.02953700
H	0.79624200	-1.80101500	0.00038200
O	0.35982700	1.32378400	-0.00063600
H	-1.94396100	0.30187400	0.65665600
H	-1.75177500	-0.23133700	-1.01094800
H	-1.46500500	-1.37743300	0.31049400

H₂O

Element Symbol	X	Y	Z
O	0.00000000	0.00000000	0.11596100
H	0.00000000	0.76183400	-0.46384500
H	0.00000000	-0.76183400	-0.46384500

NHCHCHO

Element Symbol	X	Y	Z
O	1.43541900	-0.55758000	0.00007500
N	-1.39187200	-0.55618300	-0.00008900

C	-0.71330300	0.50562500	0.00008600
H	-1.14110400	1.51520600	0.00032600
C	0.79702400	0.45222700	-0.00008400
H	-2.39056500	-0.35522800	0.00004800
H	1.28909700	1.44683700	-0.00036100

¹RC₁

Element Symbol	X	Y	Z
O	-0.41865500	0.70680000	-0.00002700
N	0.67469300	0.17705000	-0.00002600
C	1.69707800	-0.32784500	0.00002900
H	2.65039900	-0.80017600	0.00006600
C	-1.64116600	-0.62034100	0.00009900
H	-2.17950000	-0.20256300	-0.86740600
H	-2.17997700	-0.20189800	0.86697000

¹RC₂

Element Symbol	X	Y	Z
O	-0.31880200	0.67096800	-0.00975400
H	2.36132900	-0.84100500	-0.55099800
C	-2.03623400	-0.56177500	-0.04343500
H	-2.50981300	0.20031500	-0.68962900
H	-2.27576300	-0.19948500	0.97241200
N	1.83214400	-0.29436500	0.10167900
C	0.72784300	0.15060600	-0.01748200

¹RC₃

Element	X	Y	Z
----------------	----------	----------	----------

Symbol

C	1.08464500	-0.69211100	-0.00480400
H	1.64795300	-0.94624000	0.90918800
H	1.62818200	-1.21790400	-0.80813400
C	-0.34928300	0.21493500	-0.00180000
N	-1.38637200	-0.43006600	0.00697000
H	-2.29624300	0.00336200	0.00986300
O	0.53906800	1.00428700	-0.01501000

¹RC₄

Element Symbol	X	Y	Z
C	0.41858600	1.41468200	-0.00116200
C	0.53331300	-0.07691000	-0.00484400
N	1.76798600	-0.60270300	0.01343500
H	1.87580200	-1.60178500	0.00265900
H	2.58096800	-0.01715000	0.03111900
O	-0.43107200	-0.84821300	-0.02870300
C	-2.46556900	0.06727400	0.03037000
H	-2.70694600	-0.62632400	0.85112700
H	-2.76648000	-0.48132100	-0.87696200
H	-0.64298200	1.65552300	0.07483300
H	0.97960200	1.83367400	0.83557700
H	0.83473000	1.81173000	-0.92895400

¹RC₅

Element Symbol	X	Y	Z
O	-0.34076000	1.43918700	-0.01492900

N	-0.16534900	-1.36476600	-0.00958600
C	-1.27247300	-0.76423400	0.00365400
H	-2.24521200	-1.26636000	0.01284600
C	-1.31483500	0.74684000	0.00971400
H	-2.33168200	1.18459900	0.03915200
H	1.69471100	-0.45790600	-0.00152000
O	2.53642200	0.01550000	0.00916300
H	2.27297000	0.93624900	-0.00600200
H	-0.27479200	-2.37634800	-0.01145200

³RC₁

Element Symbol	X	Y	Z
O	0.59833900	-1.20138600	-0.00019000
N	0.94585500	-0.05827500	-0.00014400
C	1.27635700	1.04341900	-0.00010200
H	1.56364500	2.06520900	-0.00001600
C	-2.53062600	0.30329500	0.00175300
H	-3.60579900	0.40259700	-0.00891200
H	-1.83992800	-0.52907500	0.00155000

³RC₂

Element Symbol	X	Y	Z
C	-2.91430700	-0.33664600	0.09889600
H	-3.03522200	-1.03571000	-0.71719100
H	-3.41527700	0.57006200	0.41114500
C	1.12013200	0.07878600	-0.00734000
N	0.12451300	0.75617500	-0.10009900

H	-0.82598900	0.41424600	-0.03751100
O	2.14624300	-0.46183300	0.06186400

³RC₃

Element Symbol	X	Y	Z
C	2.75264000	-0.08211100	0.01950200
H	2.10778800	0.73822100	-0.26140200
H	2.60391200	-1.14396700	0.15069600
C	-0.98056400	-0.00316200	0.01384800
N	-0.72444200	1.17364500	-0.08725100
H	-1.01572000	1.88637100	0.55714200
O	-1.15716700	-1.14806300	-0.00447200

¹TS₁

Element Symbol	X	Y	Z
O	0.50106800	-0.57057200	-0.03124300
N	-0.64462600	-0.17204900	0.05818300
C	-1.72364500	0.25346300	-0.14259300
H	-2.54464100	0.53181500	0.50295700
C	1.61100600	0.42427500	0.15820900
H	1.29760300	1.29002400	-0.43700400
H	2.42670800	-0.11934900	-0.31699100

¹TS₂

Element Symbol	X	Y	Z
C	-1.20116200	0.44941400	-0.06206300
N	-1.01084100	-0.73742000	-0.13427300

C	0.62553600	0.00626700	0.42321900
O	1.50886100	0.05714000	-0.35699000
H	0.48148400	-0.38615400	1.42130000
H	-0.17431400	1.10520200	0.61093500
H	-1.84842100	1.25169200	-0.40333400

¹TS₃

Element Symbol	X	Y	Z
O	1.47639300	-0.48211600	0.04660100
N	-1.42684400	-0.69919100	-0.11476200
C	-0.95061900	0.45010300	0.11359000
H	-1.27319300	1.43423900	-0.28508700
C	0.76422800	0.42820200	-0.09552300
H	-0.41473600	0.62366100	1.12366300
H	0.98304000	1.42353900	-0.51645600

¹TS₄

Element Symbol	X	Y	Z
O	1.40822900	-0.60142400	-0.00017300
N	-1.45368700	-0.42064300	0.00009900
C	-0.70806800	0.55269300	-0.00015800
H	-1.06581900	1.60184400	-0.00081300
C	0.82018000	0.43850700	0.00020900
H	-2.05454500	-1.20500600	0.00036200
H	1.35767000	1.41185700	0.00082900

¹TS₅

Element	X	Y	Z
----------------	----------	----------	----------

Symbol

O	-1.56550100	-0.32786300	-0.27500100
N	1.63708800	-0.34556000	-0.00515700
C	0.68005100	0.45795200	-0.17324900
H	0.85575200	1.38025700	-0.73266100
C	-0.71777300	0.24185300	0.35091300
H	-0.91657000	0.62968400	1.37187400
H	1.35154700	-1.16695300	0.53090400

¹TS₆

Element Symbol	X	Y	Z
N	-1.26571800	0.68305300	-0.00000200
C	-1.13556400	-0.51369400	0.00000300
O	0.98713900	-0.68733200	0.00000100
C	1.37717400	0.42164500	-0.00000800
H	1.56359100	1.47448500	0.00004400
H	-1.81792600	-1.35137300	0.00000800
H	-0.23241100	1.14647700	-0.00000700

¹TS₇

Element Symbol	X	Y	Z
N	1.54655764	0.37089899	0.02142507
C	0.61719664	-0.43007601	0.21268307
O	-0.66130936	-0.33208601	-0.38000193
C	-1.60853036	0.38674999	0.15770707
H	-1.18211036	0.82187899	1.10059607
H	0.70388964	-1.32011301	0.82882307

H	1.28890764	1.13115499	-0.60637193
---	------------	------------	-------------

¹TS₈

Element Symbol	X	Y	Z
N	1.73944300	-0.01112200	0.00028300
C	0.58201800	0.41856700	-0.00026000
O	-0.58088500	-0.46872800	-0.00034200
C	-1.72407200	0.06403300	0.00022500
H	-2.68992400	0.50572500	0.00095000
H	0.23792700	1.45614900	-0.00046600
H	1.77529800	-1.02979500	0.00048300

¹TS₉

Element Symbol	X	Y	Z
C	1.04847000	-0.58966600	0.00000600
H	1.51308200	-0.92789000	0.92032000
H	1.51304800	-0.92790300	-0.92032000
C	-0.30043600	-0.06781100	0.00002200
N	-1.50256200	-0.12102100	-0.00007400
H	-2.48626900	-0.17662500	0.00033200
O	0.68623400	0.85305300	0.00000200

¹TS₁₀

Element Symbol	X	Y	Z
C	0.59413500	1.41481200	-0.03822600
C	0.35327400	-0.05614600	-0.07089100
N	1.39040800	-0.86065200	0.16669800

H	1.25588000	-1.85579700	0.12550100
H	2.30140600	-0.49351600	0.36938100
O	-0.71600900	-0.63402400	-0.47449200
C	-1.76980300	0.07776000	0.33297200
H	-1.92076700	-0.59171200	1.18154100
H	-2.64989400	0.08736000	-0.30987900
H	-0.24513400	1.90589800	-0.52229900
H	1.54751900	1.63193600	-0.53852300
H	0.64057000	1.79403500	0.98020200

¹TS₁₁

Element Symbol	X	Y	Z
C	1.14142797	-1.10811568	-0.14042316
C	0.16473497	0.03564332	-0.07439716
N	0.75542997	1.32112732	0.02766684
H	0.43528097	1.98612032	-0.65592616
H	1.76035097	1.30143632	0.06740584
O	-0.95009303	-0.20481468	0.78692984
C	-1.18581303	-0.10169668	-0.60870216
H	-1.72597303	0.79012032	-0.92056016
H	-1.48962503	-1.01806668	-1.11016516
H	0.61519497	-2.06083168	-0.12122116
H	1.81157597	-1.07306468	0.72202284
H	1.74266697	-1.05073268	-1.04957516

¹TS₁₂

Element Symbol	X	Y	Z
---------------------------	----------	----------	----------

C	1.11283400	-1.10091000	-0.01044700
C	0.12125500	0.03216700	-0.06760800
N	0.66221000	1.36324800	-0.14421400
H	1.59780000	1.35844400	-0.52488100
H	0.69680600	1.76652100	0.78270500
O	-0.96433900	-0.06034500	0.82130200
C	-1.23486200	-0.18448700	-0.57595000
H	-1.75562400	0.65854100	-1.01684600
H	-1.55542600	-1.17343500	-0.89198800
H	0.60134800	-2.06141600	0.04016400
H	1.74851700	-0.99481600	0.87120000
H	1.75045300	-1.09444100	-0.89724600

¹TS₁₃

Element Symbol	X	Y	Z
C	-0.64478991	-1.41515494	0.08264707
C	-0.50280391	0.04951306	0.09855307
N	-1.53070691	0.81293706	-0.13125993
H	-2.46468091	0.45132006	-0.24663693
H	-1.39855591	1.81291306	-0.18345093
O	1.79267009	-0.10942294	-0.29975093
C	0.88329909	0.62089706	0.27627807
H	0.81166109	1.70635306	-0.02431193
H	0.90706509	0.68779706	1.41872707
H	-0.00099191	-1.83807294	0.85384207
H	-0.21388891	-1.75628294	-0.86194193

H	-1.67649391	-1.75508894	0.17769407
---	-------------	-------------	------------

¹TS₁₄

Element Symbol	X	Y	Z
C	0.61914794	1.43283454	0.04481041
C	0.51240994	-0.03661446	0.02021941
N	1.58629994	-0.78894746	-0.09201259
H	2.51119194	-0.42665746	0.06730441
H	1.48394294	-1.78596846	-0.20041359
O	-1.88069706	0.05609954	-0.15846759
C	-0.84723006	-0.62089546	0.15615941
H	-0.80055006	-1.71777546	-0.08336659
H	-0.57692606	-0.64228346	1.31762241
H	-0.12942706	1.83062654	0.72891641
H	0.31422094	1.78290854	-0.94675259
H	1.61952494	1.79906454	0.27647841

¹TS₁₅

Element Symbol	X	Y	Z
O	1.90913800	-0.13109641	0.09902376
N	-0.57790300	1.36046159	-0.24881924
C	-0.46745600	0.03202659	0.34518776
C	-1.66726700	-0.81945841	-0.02580224
C	0.82943500	-0.56525741	-0.18611224
H	-0.36108800	0.05774159	1.44043976
H	-2.58594600	-0.38681141	0.37358576
H	-1.57083800	-1.82849141	0.37983976

H	-1.76687000	-0.88359341	-1.11094424
H	0.22957200	1.91921659	-0.00430224
H	-1.40131000	1.83508659	0.09671776
H	0.72850200	-1.42904441	-0.87659624

¹TS₁₆

Element Symbol	X	Y	Z
O	1.78080763	-0.16432604	-0.25065145
N	-0.47717837	1.41735996	-0.11113945
C	-0.51951237	0.06442196	0.40816755
C	-1.61814637	-0.77208704	-0.25827645
C	0.80270163	-0.65361604	0.24232355
H	-0.71325937	0.06339196	1.48894555
H	-2.58650337	-0.30983504	-0.07319445
H	-1.64021337	-1.79419904	0.12750455
H	-1.46056937	-0.80428004	-1.33738245
H	0.41232563	1.57190596	-0.57202245
H	-0.56079237	2.10181596	0.62453655
H	0.81149163	-1.70558104	0.59381055

¹TS₁₇

Element Symbol	X	Y	Z
O	0.71466878	1.20918628	-0.06026899
N	-1.89860722	0.11770228	-0.31983399
C	-0.99258522	-0.51197272	0.28385801
H	-1.09400822	-1.52780272	0.68249801
C	0.34342678	0.13618328	0.54881001

H	0.68256078	-0.04124472	1.58539301
H	1.42943578	0.29768528	-0.64678199
O	1.40573778	-0.85330172	-0.37080399
H	2.18221778	-1.09185672	0.14825601
H	-2.74326522	-0.44586372	-0.39847399

¹TS₁₈

Element Symbol	X	Y	Z
O	0.41007760	-1.29566585	-0.00919418
N	-1.84154740	0.01067115	0.11594582
C	-0.77564340	0.73676215	-0.03132718
H	-0.82921840	1.80820715	-0.17470418
C	0.46553360	-0.04711285	-0.21925918
H	0.26251660	0.26725515	-1.37389618
H	-1.62505740	-0.98126885	0.24810882
O	1.60383360	0.67787115	0.16410182
H	2.29001860	0.00742915	0.22081582
H	-2.78096240	0.34342315	-0.00612418

³TS₁

Element Symbol	X	Y	Z
O	-0.36656400	-0.75659900	0.03031200
N	0.77716300	-0.30499200	0.00780800
C	1.72866400	0.38963500	-0.02255200
H	2.00983000	1.43077600	-0.05351800
C	-1.77085600	0.49163300	-0.05018900
H	-1.54204800	1.48039000	0.32157800

H	-2.72225400	-0.01103900	-0.12875600
---	-------------	-------------	-------------

³TS₂

Element Symbol	X	Y	Z
O	-0.64869700	-0.65793500	-0.06486200
N	0.87056000	-0.56454100	0.11051500
C	1.44258100	0.47459700	-0.05638700
H	1.40402700	1.51586700	-0.32913400
C	-1.28319400	0.52511200	-0.00957100
H	-0.92407000	1.28081000	0.68258200
H	-2.34062400	0.42033100	-0.21240900

³TS₃

Element Symbol	X	Y	Z
O	-1.80627542	-0.50663427	0.03428047
N	1.90415058	-0.58372727	-0.10026953
C	1.44964158	0.48480173	0.07018547
H	1.62984658	1.53741073	0.24228047
C	-1.29156842	0.60158273	0.04218447
H	-1.30199042	1.22330473	-0.88493253
H	-0.00003942	0.45505573	0.15380647

³TS₄

Element Symbol	X	Y	Z
O	-1.31681200	-0.65058300	0.02571000
N	0.75652500	0.63786500	0.02039900
C	1.51378200	-0.35234900	0.00874900

H	2.58289700	-0.18398100	-0.13290600
C	-1.04189000	0.50914300	-0.14947100
H	-1.32509200	1.35505900	0.49726700
H	1.14966100	-1.37223700	0.13149100

³TS₅

Element Symbol	X	Y	Z
O	1.38057261	-0.66801013	0.02006859
N	-1.28940239	0.64634287	0.05736359
C	-1.64025839	-0.45613913	-0.07591341
H	-1.74055839	-1.51369413	-0.20717241
C	1.32892161	0.54629587	0.05802859
H	-0.04430139	0.87905087	0.15022059
H	1.49661261	1.18245387	-0.84562941

³TS₆

Element Symbol	X	Y	Z
O	1.77121355	0.54243744	-0.07902937
N	-1.48327445	-0.57074556	-0.03780837
C	-1.83681245	0.58950244	0.08573663
H	-2.12566745	1.60799344	0.19722563
C	1.47579655	-0.58725756	0.00938563
H	-2.13161445	-1.24309856	-0.47615437
H	1.05452655	-1.07036156	0.92590063

³TS₇

Element Symbol	X	Y	Z
---------------------------	----------	----------	----------

O	1.49270026	-0.54238305	-0.00371119
N	-1.64343774	-0.49617605	-0.02109719
C	-1.08408774	0.58521995	-0.00492419
H	-1.50143174	1.57441895	0.18791181
C	1.13737426	0.58533795	0.09711681
H	1.10430026	1.32657495	-0.73837719
H	-1.09307874	-1.33657005	-0.19522519

³TS₈

Element Symbol	X	Y	Z
O	-1.52329831	-0.39077673	-0.23867802
N	1.66483169	-0.29095873	0.03946898
C	0.68078169	0.43083227	-0.22942002
H	0.76645669	1.20746127	-0.99303002
C	-0.68920031	0.30490127	0.45140498
H	-1.07268131	1.26174927	0.83378598
H	1.45055869	-0.98487073	0.75570498

³TS₉

Element Symbol	X	Y	Z
N	1.46226600	-0.57989400	-0.02394800
C	0.97203300	0.52634800	0.05008700
O	-0.94915700	0.60342800	-0.02944300
C	-1.46861400	-0.49908900	-0.11015700
H	-1.79628600	-1.11888600	0.74566900
H	1.35870300	1.53244900	0.14432300
H	0.77446400	-1.34527900	-0.12638700

³TS₁₀

Element Symbol	X	Y	Z
N	-1.51669400	0.42473200	-0.03204600
C	-0.64356100	-0.45341200	-0.15019100
O	0.64598800	-0.38425500	0.39103100
C	1.57907500	0.27671700	-0.28656700
H	1.83627500	1.32187800	-0.10898900
H	-0.82361500	-1.39865400	-0.65406600
H	-1.17679200	1.23785800	0.47967900

³TS₁₁

Element Symbol	X	Y	Z
N	1.73139000	-0.06040100	0.00000100
C	0.56973200	0.39757100	0.00000600
O	-0.55285200	-0.38927700	-0.00000300
C	-1.75524900	0.18918400	-0.00001800
H	-2.68189100	-0.36977200	0.00008600
H	0.35980300	1.46499600	0.00001300
H	1.73827600	-1.07872600	-0.00001000

³TS₁₂

Element Symbol	X	Y	Z
C	-1.64652309	0.01980852	0.00001846
H	-1.98941209	-0.27372948	0.98109446
H	-1.98952709	-0.27317948	-0.98118354
C	0.37496191	-0.00781848	-0.00000254

N	0.63110491	1.20606552	0.00000146
H	-0.10006709	1.89756252	0.00017346
O	0.65847491	-1.14997348	-0.00001854

³TS₁₃

Element Symbol	X	Y	Z
C	-0.59857551	1.22143754	0.00000116
H	-1.04826151	1.52427554	-0.93391984
H	-1.04830451	1.52421354	0.93392216
C	0.37474949	0.29264054	0.00000516
N	1.42889049	-0.31021546	-0.00000184
O	-1.09297851	-1.07425446	-0.00000384
H	2.25448149	0.28308254	0.00000916

³TS₁₄

Element Symbol	X	Y	Z
C	0.41171519	0.89537271	0.00006262
H	0.78168719	1.30819771	0.92737162
H	0.78271919	1.30752771	-0.92715838
C	-0.75973981	0.26864471	-0.00030938
N	-1.70936381	-0.47524329	0.00017562
H	-2.65046181	-0.09662729	-0.00029638
O	1.82603019	-0.61595529	0.00003662

³TS₁₅

Element Symbol	X	Y	Z
C	-0.91885108	-0.61131405	0.07143272

H	-0.94047908	-1.08671905	1.05493572
H	-1.54057008	-1.09422905	-0.68510428
C	0.41482092	-0.16882805	-0.44192928
N	1.48102292	-0.06031405	0.24097972
H	2.31556792	0.21770495	-0.27066428
O	-0.90802208	0.76179795	0.06074172

³TS₁₆

Element Symbol	X	Y	Z
C	1.06819900	-0.56853700	0.01588500
H	1.56653400	-0.90941200	-0.89065000
H	1.50767900	-0.91733400	0.95011000
C	-0.28290300	-0.11283000	-0.03027000
N	-1.63216900	-0.10756200	-0.12429500
H	-2.08137400	-0.09408400	0.80526300
O	0.71507100	0.84524500	0.01145700

³TS₁₇

Element Symbol	X	Y	Z
C	-1.56849618	-0.31240557	0.00021190
H	-1.81232418	-0.69070157	-0.98115810
H	-1.81072918	-0.69741957	0.97936190
C	0.35835182	0.07390643	0.00000490
N	0.85032882	-1.05885157	-0.00029410
H	1.86310582	-1.09834557	-0.00087110
O	0.27435182	1.26259443	0.00026290

³TS₁₈

Element Symbol	X	Y	Z
C	0.82870091	-1.10040814	0.00000007
H	1.32396491	-1.31858014	-0.93438593
H	1.32396891	-1.31857614	0.93438507
C	-0.30260209	-0.37318114	0.00000007
N	-1.48318609	-0.09735614	0.00000007
H	-1.69799109	0.89494586	-0.00000293
O	0.84393091	1.30055086	0.00000007

³TS₁₉

Element Symbol	X	Y	Z
C	0.94626500	-0.18060900	0.00000000
H	1.45809200	-0.40131900	0.92594800
H	1.45809200	-0.40131900	-0.92594800
C	0.00000000	0.75656900	0.00000000
N	-0.85712200	1.60366300	0.00000000
H	-1.83429800	1.32245800	0.00000000
O	-0.09495300	-1.90015200	0.00000000

³TS₂₀

Element Symbol	X	Y	Z
C	-0.88965900	0.59432300	0.07895400
H	-1.52142200	1.09480800	-0.65767100
H	-0.87124800	1.08087800	1.05946500
C	0.41836600	0.10968900	-0.46750300
N	1.56239300	-0.03996000	0.03443300

H	1.67278000	0.11124600	1.04003900
O	-0.92363800	-0.77891000	0.08105300

¹IM₁

Element Symbol	X	Y	Z
O	1.45471800	0.55573500	-0.00026700
N	-2.74477600	-0.19684000	-0.00048100
C	-1.64482500	0.11899000	0.00066600
H	-0.60499600	0.40133600	0.00103600
C	2.16941100	-0.40677800	0.00003000
H	1.76441500	-1.43491500	0.00018100
H	3.26875300	-0.30769000	0.00011100

¹IM₂

Element Symbol	X	Y	Z
O	1.43541900	-0.55758000	0.00007500
N	-1.39187200	-0.55618300	-0.00008900
C	-0.71330300	0.50562500	0.00008600
H	-1.14110400	1.51520600	0.00032600
C	0.79702400	0.45222700	-0.00008400
H	-2.39056500	-0.35522800	0.00004800
H	1.28909700	1.44683700	-0.00036100

¹IM₃

Element Symbol	X	Y	Z
O	1.36603800	-0.55204700	0.00004200
N	-1.46881800	-0.48817900	-0.00019200

C	-0.74372700	0.54192700	0.00023300
H	-1.20600400	1.53015200	0.00010700
C	0.76655200	0.48602200	-0.00007300
H	1.29521300	1.45778300	-0.00045100
H	-0.87273800	-1.32199600	0.00039200

¹IM₄

Element Symbol	X	Y	Z
N	1.42764400	0.47101800	-0.03178000
C	0.67517700	-0.50661900	0.07689700
O	-0.74378700	-0.53925200	-0.11529200
C	-1.52681700	0.48044000	0.04355700
H	-0.89091100	1.34497100	0.38127800
H	1.01262300	-1.51237700	0.29853400
H	0.94491900	1.34136700	-0.25773300

¹IM₅

Element Symbol	X	Y	Z
N	1.70643200	-0.01601100	0.00002800
C	0.53862300	0.39647500	-0.00005000
O	-0.58304600	-0.46665800	-0.00006200
C	-1.79911500	-0.01013700	0.00008500
H	-1.72401200	1.11228600	0.00013400
H	0.24767200	1.44835700	-0.00013400
H	1.75863800	-1.03333200	0.00008700

¹IM₆

Element	X	Y	Z
---------	---	---	---

Symbol

C	0.37469500	1.42900900	-0.02750800
C	0.43755300	-0.03969000	0.02143800
N	1.61428300	-0.70368400	0.05364400
H	1.60240400	-1.69952400	-0.08233900
H	2.46407900	-0.20991200	-0.14135700
O	-0.61285600	-0.72399300	0.00362800
C	-1.93806100	-0.11979200	-0.12644200
H	-2.04363900	0.59832300	0.69440800
H	-2.60150300	-0.96656000	0.04392600
H	-0.38848800	1.66327700	-0.78690500
H	1.33886400	1.88303100	-0.24832500
H	-0.01397600	1.83193400	0.91113600

¹IM₇

Element Symbol	X	Y	Z
C	1.20275900	-1.00791700	-0.03671800
C	0.12730400	0.04382000	-0.04070400
N	0.64968500	1.37054900	0.04084300
H	-0.08461000	2.06324300	-0.00263400
H	1.32234300	1.55046100	-0.69115900
O	-0.96730200	-0.18372900	0.81440100
C	-1.20162700	-0.24668700	-0.59403700
H	-1.79844700	0.56786100	-0.99678500
H	-1.44326800	-1.23393800	-0.97963300
H	0.76096900	-2.00290600	-0.04426700

H	1.81824700	-0.89749700	0.85656600
---	------------	-------------	------------

H	1.84478100	-0.90653900	-0.91444000
---	------------	-------------	-------------

¹IM₈

Element Symbol	X	Y	Z
C	1.08867900	-1.11433100	-0.09830700
C	0.12664100	0.04492900	-0.03677100
N	0.68283600	1.35783700	0.02449500
H	0.89387300	1.71671300	-0.89512400
H	1.52297000	1.38456100	0.58548300
O	-0.98018400	-0.17142800	0.82520900
C	-1.22861400	-0.07356000	-0.56643800
H	-1.74920000	0.82730500	-0.87837200
H	-1.53887300	-0.99375000	-1.05479800
H	0.55611900	-2.06438500	-0.12387300
H	1.73222800	-1.10782700	0.78510300
H	1.72426400	-1.03827600	-0.98246700

¹IM₉

Element Symbol	X	Y	Z
C	1.21673100	-0.98802500	-0.06714400
C	0.13082300	0.05142300	-0.06383500
N	0.62124900	1.37842700	-0.14771200
H	1.28721500	1.56820200	0.58948500
H	-0.12441500	2.05760900	-0.09091300
O	-0.95023500	-0.21304200	0.84460100
C	-1.20359500	-0.25413300	-0.55702600

H	-1.82813200	0.54934800	-0.94112900
H	-1.42303900	-1.23702400	-0.96638400
H	0.78807900	-1.98492600	0.01803700
H	1.88828100	-0.83237600	0.78074200
H	1.80138500	-0.92108100	-0.98462800

¹IM₁₀

Element Symbol	X	Y	Z
C	0.52643700	1.42230200	0.03780700
C	0.46827700	-0.04386000	0.01860400
N	1.55220000	-0.76440500	-0.09676300
H	2.47535800	-0.36732800	-0.03627000
H	1.47255600	-1.76744400	-0.17841700
O	-1.91051800	0.05237100	-0.17910200
C	-0.89536400	-0.66681300	0.14863700
H	-0.81171500	-1.71468400	-0.26123400
H	-0.74678300	-0.85725700	1.28421700
H	-0.10531700	1.78234600	0.85115000
H	0.00479800	1.75192400	-0.86698000
H	1.53374200	1.83452800	0.08741000

¹IM₁₁

Element Symbol	X	Y	Z
C	-0.95271800	1.29856000	0.07797900
C	-0.36152200	-0.04597500	-0.33697900
N	-1.22565400	-1.12444600	0.11516600
H	-2.19459400	-0.92729600	-0.09193400

H	-0.98712700	-2.00292500	-0.32357800
O	2.04094300	0.00536000	-0.18269300
C	0.98420100	-0.18919200	0.34799500
H	0.91158300	-0.48007700	1.41763200
H	-0.17827900	-0.03888900	-1.41801200
H	-0.30485600	2.11791400	-0.23296700
H	-1.08545800	1.34056700	1.16105500
H	-1.92900400	1.43858000	-0.39078000

¹IM₁₂

Element Symbol	X	Y	Z
O	1.89280300	-0.13105100	-0.20115400
N	-0.36083200	1.39997100	-0.14811300
C	-0.42990800	0.05600900	0.37627600
C	-1.59173800	-0.72993500	-0.22722900
C	0.87486800	-0.66529200	0.13987900
H	-0.53956200	0.01956600	1.47629500
H	-2.53620400	-0.24042200	0.01111200
H	-1.63054400	-1.74800400	0.16684800
H	-1.49518300	-0.76731200	-1.31286900
H	0.59010400	1.74400700	-0.11354100
H	-0.96508300	2.03355100	0.35257200
H	0.84054400	-1.75746700	0.33205400

¹IM₁₃

Element Symbol	X	Y	Z
O	0.54388500	1.32290900	0.00139400

N	-1.78299400	0.09719500	-0.23931700
C	-0.85653300	-0.61055600	0.23399000
H	-0.95053700	-1.66128100	0.51713800
C	0.51819600	0.00137300	0.39769700
H	0.81660700	-0.03469700	1.45483300
H	-0.26213200	1.48011900	-0.51145100
O	1.37061300	-0.81880400	-0.36404000
H	2.24706500	-0.42967600	-0.33788500
H	-2.65601000	-0.41257100	-0.33637500

³IM₁

Element Symbol	X	Y	Z
O	-0.60557700	-0.65401000	-0.04103800
N	0.83058200	-0.56396100	0.08432600
C	1.43541500	0.47818900	-0.04600300
H	1.31407100	1.52807400	-0.27345900
C	-1.27562000	0.51592600	-0.02212100
H	-0.91016300	1.31314600	0.61686800
H	-2.33213300	0.37389600	-0.19664400

³IM₂

Element Symbol	X	Y	Z
O	1.52926200	-0.61610800	-0.00020700
N	-2.84654900	0.17941300	-0.00007700
C	-1.72816500	-0.06118900	0.00015900
H	-0.67951800	-0.28538300	0.00035600
C	2.20135300	0.49598200	0.00016300

H	2.76551500	0.67544200	-0.92462200
---	------------	------------	-------------

H	2.76661700	0.67416100	0.92452200
---	------------	------------	------------

³IM₃

Element Symbol	X	Y	Z
O	-1.64226200	-0.61982700	0.06651000
N	1.38107800	0.72158400	0.00581700
C	1.80173400	-0.44029000	-0.01312800
H	2.87568600	-0.65820000	0.03638600
C	-1.66952300	0.52822500	-0.16361200
H	-1.31007500	1.32874100	0.53090300
H	1.11166800	-1.29062200	-0.07964600

³IM₄

Element Symbol	X	Y	Z
O	2.08714100	0.45726600	0.03923600
N	-1.25538700	0.15376900	0.05236000
C	-2.41187600	-0.21909500	-0.01509700
H	-3.36878400	0.29865700	-0.15460600
C	1.63117200	-0.60137500	-0.16801600
H	-0.99684200	1.14009300	-0.02654800
H	1.14043400	-1.25044000	0.59942300

³IM₅

Element Symbol	X	Y	Z
O	1.70410900	0.53305400	0.01538900
N	-1.48228200	-0.52502000	-0.01345200

C	-1.70168000	0.68103700	0.01195200
H	-0.94131700	1.46501100	0.09352600
C	1.49915400	-0.61100200	-0.12906400
H	-2.26912500	-1.16519100	-0.09992800
H	1.16870100	-1.30932600	0.68012600

³IM₆

Element Symbol	X	Y	Z
O	1.46432600	-0.55121200	-0.00001700
N	-1.54991700	-0.45519300	0.00000200
C	-0.70986100	0.51993000	0.00000200
H	-1.09250700	1.53794000	-0.00011800
C	0.72225800	0.47993100	0.00000100
H	1.26466300	1.43034400	0.00014100
H	-1.11171900	-1.37140600	0.00007200

³IM₇

Element Symbol	X	Y	Z
N	1.39466500	-0.55315400	0.00079100
C	0.70413700	0.48469400	0.03701300
O	-0.67560000	0.59517600	-0.02957600
C	-1.42318800	-0.49578900	-0.12060300
H	-1.99687600	-0.90826100	0.70731200
H	1.12937100	1.47855300	0.12367700
H	0.82395400	-1.39305300	-0.09837700

³IM₈

Element	X	Y	Z
----------------	----------	----------	----------

Symbol

N	-1.72470700	-0.02303000	0.03600600
C	-0.55330800	0.39874500	-0.01478000
O	0.55181100	-0.43118200	-0.02838200
C	1.76697900	0.09963100	-0.11320000
H	2.43938100	0.20485800	0.73753700
H	-0.30188000	1.45629600	-0.05389900
H	-1.76106600	-1.04073800	0.05925200

³IM₉

Element Symbol	X	Y	Z
C	-1.30139400	-0.38136000	0.00001200
H	-1.56155500	-1.43148600	-0.00067700
H	-2.08994300	0.35726800	0.00042700
C	0.05566400	0.06268700	-0.00001100
N	1.12562100	-0.79896600	-0.00010600
O	0.30808200	1.29270600	0.00000200
H	0.78187300	-1.76262700	0.00096800

³IM₁₀

Element Symbol	X	Y	Z
C	-0.29287300	1.01543000	-0.00002400
H	-0.67892800	1.39802600	-0.93282700
H	-0.67869600	1.39849900	0.93268300
C	0.75080100	0.21537800	0.00004100
N	1.62309600	-0.62132200	0.00005000
O	-1.91765600	-0.68984800	0.00000900

H	2.58963300	-0.31334400	-0.00037400
---	------------	-------------	-------------

³IM₁₁

Element Symbol	X	Y	Z
C	0.75207100	0.52447500	0.00000600
H	1.10942000	1.11001300	0.86429400
H	1.10949000	1.11009700	-0.86419500
C	-0.75301100	0.52766100	-0.00002300
N	-1.53573700	-0.41536500	0.00000800
H	-2.53380500	-0.22299500	0.00000400
O	1.38383800	-0.67529700	-0.00000800

³IM₁₂

Element Symbol	X	Y	Z
C	1.36410200	-0.27902000	-0.00497900
H	2.08473300	0.52485900	0.02897500
H	1.69197600	-1.30612500	-0.07512700
C	-0.02434400	0.03734400	-0.01200400
N	-0.93469300	-0.99076600	0.04770200
H	-1.83853800	-0.60816400	-0.23985800
O	-0.42923400	1.22185700	0.00675000

³IM₁₃

Element Symbol	X	Y	Z
C	0.28717200	1.14892500	0.00005200
H	0.66244000	1.54011000	0.93343700
H	0.66183400	1.53963700	-0.93416800

C	-0.67326600	0.25022400	0.00010500
N	-1.60953100	-0.51273100	-0.00001900
H	-1.39734000	-1.50628500	-0.00005300
O	1.70704300	-0.79740500	-0.00000300

³IM₁₄

Element Symbol	X	Y	Z
C	-0.74846200	-0.53739000	0.00000400
H	-1.13145300	-1.10270200	0.86606800
H	-1.13173700	-1.10364300	-0.86518000
C	0.75820000	-0.56818800	-0.00016700
N	1.62614300	0.28612100	0.00011000
H	1.36864100	1.27706900	-0.00023200
O	-1.31836000	0.69498700	-0.00005700

¹P₁

Element Symbol	X	Y	Z
C	1.47808900	-0.32557500	0.09280700
N	0.61166300	0.51973000	-0.25230800
C	-0.71268900	0.39983000	0.17274200
O	-1.44825600	-0.49502600	-0.13401800
H	-1.04043000	1.26873600	0.76832500
H	1.26064900	-1.15781400	0.76778300
H	2.49178300	-0.23435600	-0.29110300

¹P₂

Element Symbol	X	Y	Z
-------------------	---	---	---

O	-1.74066100	-0.12821400	-0.00089300
N	1.77017000	0.07005700	-0.00087300
C	0.61717300	-0.44268900	0.00061300
H	0.49986300	-1.52626200	0.00246100
C	-0.64848900	0.36051900	0.00085300
H	-0.49546600	1.46235400	0.00236100
H	1.71759200	1.09223600	-0.00036600

¹P₃

Element Symbol	X	Y	Z
N	1.71474100	-0.06990600	0.00004700
C	0.56976300	0.40876600	-0.00004500
O	-0.58414000	-0.39954900	-0.00006900
C	-1.71911100	0.24033000	0.00008600
H	-2.45980500	-0.58678900	-0.00001800
H	0.31911800	1.46654800	-0.00008900
H	1.70670500	-1.08859900	0.00008400

¹P₄

Element Symbol	X	Y	Z
C	1.02633000	-0.60026800	0.00002600
H	1.47542500	-0.95017100	0.92290400
H	1.47543400	-0.95022000	-0.92282900
C	-0.30349900	-0.04674900	0.00000300
N	-1.54661700	0.00002400	-0.00002100
H	-1.94241100	-0.93393200	0.00000200
O	0.68511000	0.83953200	-0.00001300

¹P₅

Element Symbol	X	Y	Z
C	-1.05608700	-0.55888900	0.00000400
H	-1.52562500	-0.87976100	-0.92288800
H	-1.52562200	-0.87975500	0.92289900
C	0.29494600	-0.09231700	0.00000000
N	1.52554700	-0.27040900	-0.00000200
H	2.02991900	0.61347400	-0.00000500
O	-0.63633200	0.86826800	-0.00000200

¹P₆

Element Symbol	X	Y	Z
O	1.86231400	-0.10787300	-0.29522300
N	-0.48809700	1.43444000	-0.00223000
C	-0.43433200	0.05290200	0.41962000
C	-1.52828300	-0.76707200	-0.26012500
C	0.90741600	-0.62760200	0.21030700
H	-0.61243900	0.01287500	1.50293300
H	-2.50323200	-0.34186100	-0.02719400
H	-1.51148700	-1.80678000	0.07048500
H	-1.39769700	-0.75375600	-1.34505200
H	-0.37107800	1.48974400	-1.00710000
H	0.28659900	1.95118600	0.39493400
H	0.95869500	-1.67888000	0.56957800

¹P₇

Element	X	Y	Z
---------	---	---	---

Symbol

H	-0.71582200	-1.15673400	1.11113800
C	-0.72570800	-0.71695400	0.10222700
H	-0.64409200	-1.55012400	-0.59946400
N	-1.88084800	0.10047400	-0.19212300
H	-2.71493000	-0.27189600	0.23537800
H	-1.73198900	1.04287700	0.14413400
C	0.53625900	0.10109700	0.01338600
O	0.58556000	1.29928300	0.04560800
O	1.62890300	-0.67277000	-0.06615200
H	2.39376100	-0.08441000	-0.07564900

 3P_1 **Element
Symbol****X****Y****Z**

O	1.37365100	-0.60920600	0.02248600
N	-0.54710200	0.54771700	-0.18241200
C	-1.53730900	-0.32639400	0.10032600
H	-2.45469600	0.08415200	0.50448400
C	0.79828500	0.45485200	0.05428800
H	1.30306200	1.42532900	0.14967100
H	-1.57371700	-1.24059600	-0.48484200

 3P_2 **Element
Symbol****X****Y****Z**

O	-1.78489100	-0.12628900	0.00005200
N	1.80092600	0.04111000	0.00003500
C	0.59218800	-0.41442700	-0.00003500

H	0.44690000	-1.49201900	-0.00038500
C	-0.60116700	0.36433900	0.00006500
H	-0.55919100	1.45656900	-0.00070000
H	1.83881500	1.05851800	0.00024300

³P₃

Element Symbol	X	Y	Z
N	1.72492800	-0.02302000	0.03554600
C	0.55341500	0.39876700	-0.01473400
O	-0.55202000	-0.43117600	-0.02755000
C	-1.76718400	0.09932500	-0.11381100
H	-2.43933800	0.20626900	0.73656400
H	0.30174900	1.45634000	-0.05271100
H	1.76187300	-1.04061200	0.05899500

³P₄

Element Symbol	X	Y	Z
C	-0.98255600	-0.58924800	0.02852800
H	-1.02857000	-1.10236200	0.98415000
H	-1.69536400	-0.88672500	-0.73340800
C	0.28802800	0.01490700	-0.41899500
N	1.47912800	-0.13831900	0.22278100
H	2.26186500	-0.04113500	-0.42465500
O	-0.71558200	0.80556300	0.11965700

³P₅

Element Symbol	X	Y	Z
---------------------------	----------	----------	----------

C	0.97277000	-0.59074100	0.03424000
H	1.72110300	-0.86458600	-0.70197700
H	1.00196800	-1.12285900	0.98194800
C	-0.27850100	0.01587300	-0.44311000
N	-1.55321300	-0.10238900	-0.00349200
H	-1.57823200	-0.30621700	1.00370700
O	0.69525500	0.80744900	0.14924800

**Appendix B: Total Energies Including Zero-Point Corrections (E, in a.u.),
and Relative Energies (RE, in kcal/mol) for Reactants, Reactant
Complexes, Transition States, Reaction Intermediates, and Products for
Reactions in Chapter 5**

Table B-1. Total energies including zero-point corrections (E, in a.u.), and relative energies (RE, in kcal/mol) for the reactants, reactant complexes (RCs), transition states (TSs), intermediates (IMs), and products (Ps) for the reactions in Chapter 5^d

Singlet	E (a.u.)	RE (kcal/mol)	Singlet	E (a.u.)	RE (kcal/mol)
¹ CH ₂ + HCNO	-207.4127626 (-207.677513)	0.00 (0.00)	¹ TS ₁₆	(-248.410565)	(-83.35)
¹ CH ₂ + HNCO	-207.5233873 (-207.789059)	0.00 (0.00)	¹ TS ₁₇	-283.9217749 (-284.263399)	36.69 (35.67)
¹ CH ₂ + CH ₃ CONH ₂	(-248.277738)	(0.00)	¹ TS ₁₈	-283.9470902 (-284.296029)	20.80 (15.20)
H ₂ O + NHCHCHO	-283.9802451 (-284.320248)	0.00 (0.00)	¹ IM ₁	-207.6378576 (-207.898754)	-141.25 (-138.83)
¹ RC ₁	-207.4255828 (-207.691713)	-8.04 (-8.91)	¹ IM ₂	-207.6383110 (-207.904179)	-141.53 (-142.24)
¹ RC ₂	-207.5276957 (-207.794436)	-2.70 (-3.37)	¹ IM ₃	-207.6440502 (-207.909825)	-145.14 (-145.78)
¹ RC ₃	-207.5251279 (-207.794367)	-1.09 (-3.33)	¹ IM ₄	-207.5445046 (-207.808732)	-82.67 (-82.34)
¹ RC ₄	(-248.297685)	(-12.52)	¹ IM ₅	-207.5479113 (-207.812760)	-84.81 (-84.87)
¹ RC ₅	-283.9881557 (-284.328769)	-4.96 (-5.35)	¹ IM ₆	(-248.315679)	(-23.81)
¹ TS ₁	-207.4149357 (-207.681737)	-1.36 (-2.65)	¹ IM ₇	(-248.384722)	(-67.13)
¹ TS ₂	-207.5347565 (-207.801660)	-76.55 (-77.90)	¹ IM ₈	(-248.383871)	(-66.60)
¹ TS ₃	-207.4930707 (-207.762031)	-50.39 (-53.04)	¹ IM ₉	(-248.388438)	(-69.47)
¹ TS ₄	-207.5981231 (-207.866564)	-116.32 (-118.63)	¹ IM ₁₀	(-248.341009)	(-39.70)
¹ TS ₅	-207.6361853 (-207.902231)	-140.20 (-141.01)	¹ IM ₁₁	(-248.410085)	(-83.05)
¹ TS ₆	-207.4871563 (-207.746413)	-46.68 (-43.24)	¹ IM ₁₂	(-248.411044)	(-83.65)
¹ TS ₇	-207.5414508 (-207.806329)	-80.75 (-80.83)	¹ IM ₁₃	-284.0009127 (-284.342970)	-12.97 (-14.26)
¹ TS ₈	-207.4570025 (-207.727097)	-27.76 (-31.11)	¹ P ₁	-207.6474740 (-207.915835)	-147.28 (-149.55)
¹ TS ₉	-207.5805994 (-207.852908)	-35.90 (-40.07)	¹ P ₂	-207.6451982 (-207.910901)	-145.86 (-146.45)
¹ TS ₁₀	(-248.302645)	(-15.63)	¹ P ₃	-207.5538158 (-207.818978)	-88.51 (-88.77)
¹ TS ₁₁	(-248.384075)	(-66.73)	¹ P ₄	-207.6127669 (-207.882985)	-56.09 (-58.94)
¹ TS ₁₂	(-248.383107)	(-66.12)	¹ P ₅	-207.6136523 (-207.883834)	-56.64 (-59.47)
¹ TS ₁₃	(-248.339856)	(-38.98)	¹ P ₆	(-248.412015)	(-84.26)
¹ TS ₁₄	(-248.341743)	(-40.16)	¹ P ₇	-284.0341295 (-284.381897)	-33.81 (-38.69)
¹ TS ₁₅	(-248.407456)	(-81.40)			
Triplet	E (a.u.)	RE (kcal/mol)	Triplet	E (a.u.)	RE (kcal/mol)
³ CH ₂ + HCNO	-207.4272761 (-207.696637)	-9.11 (-12.00)	³ TS ₁₈	-207.4728192 (-207.745356)	31.73 (27.42)

³ CH ₂ + HNCO	-207.5379008 (-207.808183)	-9.11 (-12.00)	³ TS ₁₉	-207.4735590 (-207.747913)	31.27 (25.82)
³ RC ₁	-207.4290474 (-207.697743)	-10.22 (-12.69)	³ TS ₂₀	-207.4862105 (-207.758658)	23.33 (19.08)
³ RC ₂	-207.5392324 (-207.809856)	-9.94 (-13.05)	³ IM ₁	-207.4572629 (-207.728688)	-27.92 (-32.11)
³ RC ₃	-207.5384130 (-207.808547)	-9.43 (-12.23)	³ IM ₂	-207.5210788 (-207.789096)	-67.97 (-70.02)
³ TS ₁	-207.3873336 (-207.661128)	15.96 (10.28)	³ IM ₃	-207.5366754 (-207.805904)	-77.76 (-80.57)
³ TS ₂	-207.4597489 (-207.729711)	-29.48 (-32.75)	³ IM ₄	-207.5162937 (-207.785628)	-64.97 (-67.84)
³ TS ₃	-207.4824223 (-207.755575)	-43.71 (-48.98)	³ IM ₅	-207.5236526 (-207.791974)	-69.58 (-71.83)
³ TS ₄	-207.5129872 (-207.785379)	-62.89 (-67.69)	³ IM ₆	-207.5402356 (-207.818013)	-79.99 (-88.17)
³ TS ₅	-207.4866958 (-207.758795)	-46.39 (-51.01)	³ IM ₇	-207.5158644 (-207.786206)	-64.70 (-68.21)
³ TS ₆	-207.4997159 (-207.769253)	-54.56 (-57.57)	³ IM ₈	-207.5149810 (-207.785729)	-64.14 (-67.91)
³ TS ₇	-207.5164750 (-207.787758)	-65.08 (-69.18)	³ IM ₉	-207.5575397 (-207.831826)	-21.43 (-26.84)
³ TS ₈	-207.5233604 (-207.797667)	-69.40 (-75.40)	³ IM ₁₀	-207.4748374 (-207.747182)	30.47 (26.28)
³ TS ₉	-207.4905591 (-207.760856)	-48.82 (-52.30)	³ IM ₁₁	-207.5191279 (-207.792362)	2.67 (-2.07)
³ TS ₁₀	-207.5073032 (-207.777536)	-59.33 (-62.77)	³ IM ₁₂	-207.5623994 (-207.837195)	-24.48 (-30.21)
³ TS ₁₁	-207.5126371 (-207.783471)	-62.67 (-66.49)	³ IM ₁₃	-207.4762519 (-207.748969)	29.58 (25.16)
³ TS ₁₂	-207.5202805 (-207.793478)	1.95 (-2.77)	³ IM ₁₄	-207.5147488 (-207.788697)	5.42 (0.23)
³ TS ₁₃	-207.4661591 (-207.738704)	35.91 (31.60)	³ P ₁	-207.5433281 (-207.820537)	-81.93 (-89.75)
³ TS ₁₄	-207.4730302 (-207.746845)	31.60 (26.49)	³ P ₂	-207.5395170 (-207.817229)	-79.54 (-87.67)
³ TS ₁₅	-207.4924691 (-207.764729)	19.40 (15.27)	³ P ₃	-207.5149922 (-207.785738)	-64.15 (-67.91)
³ TS ₁₆	-207.4701042 (-207.746624)	33.44 (26.63)	³ P ₄	-207.5004641 (-207.774811)	14.38 (8.94)
³ TS ₁₇	-207.5098942 (-207.784682)	8.47 (2.75)	³ P ₅	-207.4948651 (-207.768897)	17.90 (12.65)

^ATotal energies including zero-point corrections (E, in a.u.) and relative energies (RE, in kcal/mol) were calculated at the CCSD(T) (out of parentheses), and ω B97XD (in parentheses), levels of theory.

Appendix C: Structure Information of All Conformers of *N*-methyleneformamide and Imine Acetaldehyde in Chapter 6

Gauche Conformation of *N*-methyleneformamide

Element Symbol	X	Y	Z
C	1.47808900	-0.32557500	0.09280700
N	0.61166300	0.51973000	-0.25230800
C	-0.71268900	0.39983000	0.17274200
O	-1.44825600	-0.49502600	-0.13401800
H	-1.04043000	1.26873600	0.76832500
H	1.26064900	-1.15781400	0.76778300
H	2.49178300	-0.23435600	-0.29110300

Cis Conformation of *N*-methyleneformamide

Element Symbol	X	Y	Z
C	1.36592400	-0.40284800	0.00000000
N	0.67527700	0.65835700	-0.00000100
C	-0.74618900	0.46530900	0.00000000
O	-1.31626900	-0.58596400	0.00000000
H	-1.28385800	1.42423000	0.00000200
H	0.91851900	-1.40028800	0.00000200
H	2.45014100	-0.31949100	-0.00000100

Trans Conformation of *N*-methyleneformamide

Element Symbol	X	Y	Z
C	-1.63552200	0.10871500	0.00008300

N	-0.54131000	-0.52366700	-0.00011000
C	0.61448500	0.30726500	-0.00012300
O	1.72243900	-0.12415200	0.00011900
H	0.40345500	1.40035900	-0.00028800
H	-1.69523500	1.20542800	0.00027000
H	-2.57234200	-0.44277600	0.00008200

Out-ap Conformation of Imine Acetaldehyde

Element Symbol	X	Y	Z
O	-1.74516200	-0.16664600	0.00013700
N	1.69293200	0.26036900	0.00001900
C	0.61272500	-0.39237700	-0.00051100
H	0.52057600	-1.48262900	0.00082100
C	-0.67347200	0.37166200	-0.00012500
H	-0.55088300	1.47095500	0.00058400
H	2.50556400	-0.35345600	0.00118200

In-ap Conformation of Imine Acetaldehyde

Element Symbol	X	Y	Z
O	-1.74066100	-0.12821400	-0.00089300
N	1.77017000	0.07005700	-0.00087300
C	0.61717300	-0.44268900	0.00061300
H	0.49986300	-1.52626200	0.00246100
C	-0.64848900	0.36051900	0.00085300
H	-0.49546600	1.46235400	0.00236100
H	1.71759200	1.09223600	-0.00036600

Out-sp Conformation of Imine Acetaldehyde

Element Symbol	X	Y	Z
O	1.43541900	-0.55758000	0.00007500
N	-1.39187200	-0.55618300	-0.00008900
C	-0.71330300	0.50562500	0.00008600
H	-1.14110400	1.51520600	0.00032600
C	0.79702400	0.45222700	-0.00008400
H	-2.39056500	-0.35522800	0.00004800
H	1.28909700	1.44683700	-0.00036100

***In-sp* Conformation of Imine Acetaldehyde**

Element Symbol	X	Y	Z
O	1.36603800	-0.55204700	0.00004200
N	-1.46881800	-0.48817900	-0.00019200
C	-0.74372700	0.54192700	0.00023300
H	-1.20600400	1.53015200	0.00010700
C	0.76655200	0.48602200	-0.00007300
H	1.29521300	1.45778300	-0.00045100
H	-0.87273800	-1.32199600	0.00039200