

# **MACHINE LEARNING REPRESENTATION OF POTENTIAL ENERGY SURFACES AND ENERGY TRANSFER IN GAS-SURFACE SCATTERING**

By

**RUPAYAN BISWAS**

**Enrolment No CHEM11201604025**

**National Institute of Science Education and Research, Bhubaneswar**

*A thesis submitted to the*

*School of Chemical Sciences*

*(as applicable)*

*In partial fulfillment of requirements*

*for the Degree of*

**DOCTOR OF PHILOSOPHY**

*of*

**HOMI BHABHA NATIONAL INSTITUTE**



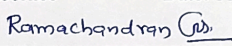
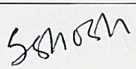
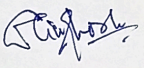
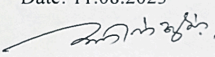


**August, 2023**

# Homi Bhabha National Institute

## Recommendations of the Viva Voce Committee

As members of the Viva Voce Committee, we certify that we have read the dissertation prepared by **Rupayan Biswas** entitled “**Machine Learning Representations of Potential Energy Surfaces and Energy Transfer in Gas-Surface Scattering**” and recommend that it may be accepted as fulfilling the thesis requirement for the award of Degree of Doctor of Philosophy.

Chairman – Prof. M. Sarkar		Date: 11.08.2023
Guide / Convener – Prof. U. Lourderaj		Date: 11.08.2023
Co-Guide –		Date:
Examiner – Prof. C. N. Ramachandran, IIT Roorkee		Date: 11.08.2023
Member 1 – Prof. Subhadip Ghosh		Date: 11.08.2023
Member 2 – Prof. Arindam Ghosh		Date: 11.08.2023
Member 3 – Prof. A. V. Anil Kumar		Date: 11.08.2023

Final approval and acceptance of this thesis is contingent upon the candidate's submission of the final copies of the thesis to HBNI.

We hereby certify that we have read this thesis prepared under my/our direction and recommend that it may be accepted as fulfilling the thesis requirement.

Date: 11.08.2023

Place: Bhubaneswar

(Co-Guide)

Prof. U. Lourderaj

(Guide)

### **STATEMENT BY AUTHOR**

This dissertation has been submitted in partial fulfillment of requirements for an advanced degree at Homi Bhabha National Institute (HBNI) and is deposited in the Library to be made available to borrowers under rules of the HBNI.

Brief quotations from this dissertation are allowable without special permission, provided that accurate acknowledgement of source is made. Requests for permission for extended quotation from or reproduction of this manuscript in whole or in part may be granted by the Competent Authority of HBNI when in his or her judgment the proposed use of the material is in the interests of scholarship. In all other instances, however, permission must be obtained from the author.

Name & Signature of the Student

Rupayan Biswas

Date:

## **DECLARATION**

I hereby declare that I am the sole author of this thesis in partial fulfillment of the requirements for a postgraduate degree from National Institute of Science Education and Research (NISER). I authorize NISER to lend this thesis to other institutions or individuals for the purpose of scholarly research.

Name & Signature of the Student

Rupayan Biswas

Date:

### List of Publications arising from the thesis

#### Journal

1. **R. Biswas**, R. Rashmi and U. Lourderaj, “Machine learning in chemical dynamics” *Reson.* **25**, 59–75 (2020)
2. K. Giri, L. González-Sánchez, **R. Biswas**, E. Yurtsever, F. A. Gianturco, N. Sathya-murthy, U. Lourderaj, and R. Wester, “HeH<sup>+</sup> collisions with H<sub>2</sub>: Rotationally inelastic cross sections and rate coefficients from quantum dynamics at interstellar temperatures” *J. Phys. Chem. A* **126** 14, 2244-2261 (2022)
3. **R. Biswas**, U. Lourderaj and N. Sathya-murthy, “Artificial neural networks and their utility in fitting potential energy curves and surfaces and related problems” *J. Chem. Sci.* **135**, 2, 22, (2023)
4. **R Biswas**, K Giri, L González-Sánchez, F A Gianturco, U Lourderaj, N Sathya-murthy, A Veselinova, E Yurtsever, R Wester, “Rotational state-changes in C<sub>5</sub>N<sup>−</sup> by collisions with He and H<sub>2</sub>”, *Mon. Not. R. Astron. Soc.* **522**, 4, 5775-5787, (2023)
5. **R. Biswas** and U. Lourderaj, “Inelastic Scattering of Formaldehyde on Au(111) surface”, *Journal of Chemical Physics*, 10.26434/chemrxiv-2023-bm5qc-v2 (Submitted 2023)
6. **R. Biswas**, F. A. Gianturco, K. Giri, L. González-Sánchez, U. Lourderaj, N. Sathya-murthy and E. Yurtsever “An improved artificial neural network fit of the *ab initio* potential energy surface points for HeH<sup>+</sup> + H<sub>2</sub> and its ensuing rigid rotors quantum dynamics” *Artificial Intelligence Chemistry* (Submitted 2023)

#### Manuscript under preparation

1. **R. Biswas** and U. Lourderaj, “On-the-fly representation of potential energy surfaces using machine learning methods”
2. **R. Biswas** and U. Lourderaj, “Rotational excitation of formaldehyde upon scattering on graphene sheet”

## Conferences

1. Presented poster at titled “On-the-fly representation of PES using Machine Learning Methods” at **iCOMET-2019**, Hefei, China
2. Presented poster at titled “On-the-fly representation of PES using GPR” **ML4SC-2019**, IIIT Hyderabad, Telengana, India
3. Presented poster at titled “Gas-Surface Dynamics: Formaldehyde on Au (111) surface” at **TCS-2021**, IACS, Kolkata, India
4. Presented poster at titled “On-the-fly representation of PES using ANN” **ML4SC-2023**, Kodaikanal, Tamil Nadu, India

Name & Signature of the Student

Rupayan Biswas

Date:

## ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my thesis supervisor Prof. U. Lourderaj for his patience, motivation and continuous support throughout the course of my Ph.D. work. I would also like to extend the gratitude to his family members, who were always very welcoming and they made sure that I never felt like I was far from home. I also thank the thesis monitoring committee members, Prof. M. Sarkar, Prof. A. Ghosh, Prof. S. Ghosh and Prof. A. V. Anil Kumar for their support and comments. I also thank all my collaborators, Prof. N. Sathyamurthy, Prof. F. A. Gianturco, Prof. L. González Sánchez, Ms. A. Veselinova, Prof. K. Giri, Prof. E. Yurtserver, Prof. R. Wester and Prof. J. Franz for giving me the opportunity to work on exciting projects. I would also like to thank Dr. N. Sharma, Ms. R. Rashmi and D. Kumar for their support in my research work.

I thank all my lab-mates/summer-students who have visited NISER throughout the years, and provided their valuable inputs and critical comments, during the group meetings we had. Also worth mentioning are all my friends, who had less faith in me than I had for myself, yet they made the journey worth it. I would also like to acknowledge the constant support provided by Bacteria. Last but not least, I would like to thank the members of my family for their unconditional love and support.

## Synopsis

Understanding chemical reactions in order to gain control has baffled chemists for decades. There have been numerous experiments designed to understand chemical reactions at the atomic level, and theoretical models such as transition state theory (TST) and Rice-Ramsperger-Kassel-Marcus (RRKM) theory which can be used to estimate reaction rates for unimolecular reactions using energetics of the system. It is now established that dynamics plays a crucial role in accurately describing chemical reactions and that, the static picture alone based on energetics and statistical theories may be insufficient. To this end, the Ph.D. work reported here is divided into two parts. The first part deals with representing potential energy surfaces (PES) using machine learning (ML) methods. The second part reports energy transfer at gas-surface interfaces. We focus on inelastic non-reactive inelastic scattering of formaldehyde scattered from two surfaces- gold and graphene.

Chapter 1 gives an introduction of the basic concepts in chemical reaction dynamics, the tools required to understand chemical reactions computationally; (i) the construction of PESs, (ii) classical trajectory simulations, (iii) sampling methods and (iv) fundamentals of two ML methods, gaussian processes for regression (GPR) and artificial neural networks (ANN) for representing PESs.

Chapter 2 reports the ML representation of the PESs of two 4-dimensional systems,  $\text{HeH}^+ + \text{H}_2$  and  $\text{C}_5\text{N}^- + \text{H}_2$  and a 5-dimensional system  $\text{C}_2^- + \text{H}_2$ , that are important molecules in the interstellar medium. For the  $\text{HeH}^+ + \text{H}_2$  system, the  $\text{HeH}^+$  and the  $\text{H}_2$  were treated as rigid rotors. The PES was represented using the variables  $R$ ,  $\alpha$ ,  $\beta$  and  $\theta$  where  $R$  is the center of mass distance between the two rotors, and  $\alpha$ ,  $\beta$ ,  $\theta$  correspond to the angles which define the orientation of the rotors with respect to each other. Using 14741 potential values computed at the CCSD(T)/CBS level of theory in the range  $-12000 \text{ cm}^{-1}$  to  $10000 \text{ cm}^{-1}$ , the PES was fitted using GPR and ANN ML algorithms and different sampling schemes. We found that the fit was better when we sampled the points randomly than by using the Latin Hypercube Sampling (LHS). The final fit had an RMSD of  $35.4 \text{ cm}^{-1}$  for all the data. The fit was appended with the corrected long range potential and was converted to a FORTRAN code, which could then be used to study inelastic rotational cross sections using MOLSCAT. In addition, a deep neural network fit was also done for this system, with two hidden layers

and 40 nodes in each layer, which resulted in an RMSD of  $4.9\text{ cm}^{-1}$ . As in the case with  $\text{HeH}^+ + \text{H}_2$ , the  $\text{C}_5\text{N}^- + \text{H}_2$  system was also 4-dimensional with rigid  $\text{C}_5\text{N}^-$  and  $\text{H}_2$ , but the well depth for this system was at most  $600\text{ cm}^{-1}$ . Using 16000 data points computed at CCSD(T)-F12b with aug-cc-PVTZ level of theory, additional data points were generated using symmetry of the system resulting in a total of 45000 points. We used shallow neural network as well as deep neural network for this system. The RMSD of the shallow neural network was  $9.27\text{ cm}^{-1}$  while for deep neural network fit it was  $0.87\text{ cm}^{-1}$ . The deep neural fit was converted to a FORTRAN code, and the asymptotically correct long range potential was appended to the code, which was then used for quantum scattering calculations. For the  $\text{C}_2 + \text{H}_2$  5-dimensional system, the total number of data points computed at CCSD(T)-F12b with aug-cc-PVTZ level of theory, was 68011. Fitting was done using deep neural network and the best fit obtained had an RMSD of  $0.82\text{ cm}^{-1}$ .

Chapter 3 reports the “on-the-fly” representation of the PES of a system during *ab initio* trajectory simulation (AICT). Studying the dynamics of large systems using AICT are often impeded by the high CPU cost to compute forces during the trajectory calculation. In order to generate statistically significant results, we need to compute many trajectories. In this work we look at the possibility of representing the PES “on-the-fly” using GPR and ANN methods. In the first scheme we store the data, the internal coordinates and the potential values at those coordinates for a few pilot trajectories. After that for any new trajectory if the trajectory is near to any part of the configuration space where we have stored prior data, we use the stored data to interpolate using GPR. In the second scheme, the PES is represented a sum of ANNs ( $V_i^{NN}$ ) that each define a different region of the PES. The individual ANNs,  $V_i^{NN}$  are trained using data obtained from different trajectories. The two schemes were implemented with 2-dimensional and 6-dimensional model potentials and the robustness of the algorithm was tested.

Chapter 4 reports the non-reactive scattering of formaldehyde on gold surface. Recent quantum state resolved molecular beam experiments[*Phys. Chem. Chem. Phys.* **19**, 19904, (2017)] have revealed that formaldehyde upon scattering from Au(111) surface, do not follow a Boltzmann distribution of rotational states, rather rotational excitation along a specific axis is preferred than the others, which in literature is known as “Rotational Rainbow”. We used classical trajectory simulations using Sutton-Chen potential for gold, Lennard-Jones

potential for gold-formaldehyde interaction and Bowman's potential for formaldehyde, to model and study the dynamics of this system. We compute trajectories for different collision energies, surface temperatures and orientations. Our trajectory results are in good accord with the experimentally observed values for axis-specific rotational energies and trapping probabilities. We found that the average rotational energy of scattered formaldehyde increases as a function of collision energy. To understand the origin of the rotational rainbow we looked into the possible mechanism of scattering. We find that the rotational rainbow is dictated by the minimum energy surface of the interaction potential. Further the rotational rainbow seems to be independent of the surface temperature.

In Chapter 5 we look at the scattering of formaldehyde on a single layer graphene surface. We model the graphene potential using AMBER force fields, the graphene-formaldehyde interaction using modified Buckingham's potential and Bowman's potential for formaldehyde. We find that the trapping probability of formaldehyde on graphene sheet decreases with increasing collision energy, however it is much greater than that from gold surface. Despite having very different interaction surface than gold-formaldehyde, our results indicate that the average rotational energy of the direct-scattered molecules from graphene surface are very similar to that of gold-formaldehyde system. However, the rotational rainbow in this system is dependent on the collision energy of formaldehyde.

# Contents

<b>List of Figures</b>	<b>i</b>
<b>List of Tables</b>	<b>vi</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Potential Energy Surface (PES)	3
1.2 Classical Trajectory Simulations	5
1.2.1 Velocity-Verlet Integration	6
1.2.2 Sampling	6
1.3 Machine Learning	8
1.3.1 Gaussian Processes for Regression	9
1.3.2 Artificial Neural Network	12
<b>2 Representation of potential energy surfaces using GPR and ANN</b>	<b>16</b>
2.1 $\text{HeH}^+ + \text{H}_2$ <i>ab initio</i> PES	17
2.1.1 Data Set	19
2.1.2 GPR	19
2.1.3 Shallow Neural Network	20
2.1.4 Extension to Long Range	26
2.1.5 Deep Neural Network	28
2.2 $\text{C}_5\text{N}^- + \text{H}_2$ <i>ab initio</i> PES	31
2.2.1 Data Set	32
2.2.2 Shallow Neural Network	35
2.2.3 Deep Neural Network	36
2.2.4 Extension to Long Range	40

2.2.5	FORTTRAN code for ANN . . . . .	40
2.3	$C_2^- + H_2$ <i>ab initio</i> PES . . . . .	40
2.3.1	Data Set . . . . .	42
2.3.2	Deep Neural Network . . . . .	42
2.4	Summary and Conclusions . . . . .	46
<b>3</b>	<b>“On-the-fly” Representation of Potential Energy Surfaces using Machine Learning Methods</b>	<b>47</b>
3.1	“On-the-fly” Representation using GPR . . . . .	48
3.1.1	Advantages and Drawbacks . . . . .	52
3.2	“On-the-fly” Representation using ANN . . . . .	55
3.2.1	Six Dimensional Non-Reactive ANN . . . . .	61
3.2.2	Six Dimensional Reactive ANN . . . . .	64
3.3	Summary and Conclusions . . . . .	66
<b>4</b>	<b>Gas-Surface dynamics: Formaldehyde on Au(111) surface</b>	<b>70</b>
4.1	Methods . . . . .	72
4.1.1	Potential energy model . . . . .	72
4.1.2	Trajectory Simulations . . . . .	73
4.1.3	Trajectory Analysis . . . . .	74
4.2	Results and discussion . . . . .	76
4.2.1	Direct vs indirect scattering . . . . .	77
4.2.2	Effect of Temperature . . . . .	80
4.2.3	Effect of orientation . . . . .	82
4.2.4	Energy Transfer . . . . .	87
4.2.5	Mechanism of Energy Transfer . . . . .	92
4.3	Summary and Conclusions . . . . .	96
<b>5</b>	<b>Gas-Surface dynamics: Formaldehyde on graphene surface</b>	<b>98</b>
5.1	Potential Energy Model . . . . .	99
5.1.1	Trajectory Simulations . . . . .	101
5.2	Results and Discussion . . . . .	101

5.2.1	Direct vs Indirect Scattering . . . . .	103
5.2.2	Effect of Orientations . . . . .	105
5.3	Energy Transfer . . . . .	109
5.3.1	Effect of Orientation on Energy Transfer . . . . .	113
5.4	Summary and Conclusions . . . . .	115
<b>6</b>	<b>Conclusions and Outlook</b>	<b>116</b>
	<b>References</b>	<b>118</b>
	<b>Appendix A</b>	<b>126</b>
	<b>Appendix B</b>	<b>129</b>

# List of Figures

1.1	Potential energy curve of $H_2$ molecule . . . . .	5
1.2	Squared exponential function correlation function . . . . .	11
1.3	An example of an ANN with two hidden layers. . . . .	12
1.4	Working of a node in neural network . . . . .	13
1.5	Plot of the sigmoid type transfer function . . . . .	14
2.1	Coordinate system used to describe the $HeH^+ + H_2$ system . . . . .	19
2.2	Neural Network design for the four dimensional PES function for the $HeH^+ + H_2$ system . . . . .	20
2.3	Comparison of residuals with respect to different sampling methods for the $HeH^+ + H_2$ ANN . . . . .	22
2.4	Scatter plot for the residuals of $HeH^+ + H_2$ ANN . . . . .	22
2.5	Slice of training data generated by random sampling . . . . .	23
2.6	Plot to show smoothness of the ANN fit where data was not available . . . . .	23
2.7	Plot of the ANN fit for $HeH^+ + H_2$ system . . . . .	25
2.8	Plot of the ANN PES for long range . . . . .	26
2.9	One dimensional slices for long range potential for $HeH^+ + H_2$ system. . . . .	27
2.10	Slice of Deep Neural Network fit for $\alpha = 0^\circ$ and $\beta = 0^\circ$ for $HeH^+ + H_2$ system . . . . .	29
2.11	Residuals of Deep Neural Network fit for $HeH^+ + H_2$ system . . . . .	29
2.12	Variables used in the representation of $C_5N^- + H_2$ PES . . . . .	32
2.13	Expanding data set for $C_5N^- + H_2$ system using cubic spline method . . . . .	34
2.14	Plot of residuals for the PES fitted using a Shallow Neural Network for $C_5N^- + H_2$ system . . . . .	35
2.15	Neural Network design for $C_5N^- + H_2$ system . . . . .	37

2.16	Plot of residuals for Deep Neural Network fit for $C_5N^- + H_2$ system . . . .	38
2.17	Plot of Deep Neural Network fit for $C_5N^- + H_2$ system for $\alpha = 0^\circ$ and $\beta = 0^\circ$	38
2.18	Plot of Deep Neural Network fit for $C_5N^- + H_2$ system for $\alpha = 60^\circ$ and $\beta = 60^\circ$ . . . . .	39
2.19	Plot of Deep Neural Network fit for $C_5N^- + H_2$ system for $\alpha = 90^\circ$ and $\beta = 90^\circ$ . . . . .	39
2.20	Variables defining the $C_2^- + H_2$ system . . . . .	42
2.21	Deep Neural Network design for $C_2^- + H_2$ system. . . . .	43
2.22	Plot for residuals for $C_2^- + H_2$ system . . . . .	44
2.23	Slice of the potential energy surface for $\theta = 40$ , $\alpha = 60$ and $\beta = 90$ for $C_2^- + H_2$ system . . . . .	45
2.24	Slice of the potential energy surface for $\theta = 40$ , $\alpha = 60$ and $\beta = 90$ for $C_2^- + H_2$ system . . . . .	45
3.1	Illustration of active data selection and prediction in the “on-the-fly” local GPR PES algorithm . . . . .	50
3.2	Algorithm for “on-the-fly” local PES using GPR . . . . .	51
3.3	Malonaldehyde system . . . . .	52
3.4	Two-dimensional PES contours for malonaldehyde system . . . . .	52
3.5	Training data collected during 40 trajectories for “on-the-fly” local PES using GPR algorithm . . . . .	53
3.6	First 100 steps of the 41 <sup>st</sup> trajectory using the “on-the-fly” local PES using GPR algorithm . . . . .	53
3.7	Illustration of the working of $V^{NN}$ and $A^{NN}$ in “on-the-fly” NN PES algo- rithm . . . . .	56
3.8	Training data for “on-the-fly” NN PES algorithm . . . . .	57
3.9	Algorithm for “on-the-fly” NN PES . . . . .	58
3.10	Working of “on-the-fly” NN PES algorithm on the 2D malonaldehyde po- tential . . . . .	59
3.11	Comparison of the performance of “on-the-fly” GPR and NN algorithms . .	60

3.12	Comparison of norms of “on-the-fly” GPR and NN algorithms with the true trajectory . . . . .	60
3.13	Residuals across the potential energy range as a function of the number of representations for “on-the-fly” NN algorithm . . . . .	62
3.14	The six internal coordinates used to represent the PES of formaldehyde . . . . .	62
3.15	Illustration of the working of “on-the-fly” NN PES algorithm for formaldehyde . . . . .	63
3.16	Percentage of points predicted during the trajectory using “on-the-fly” NN PES algorithm for formaldehyde . . . . .	64
3.17	Cut off for classification of reactive and non-reactive trajectories for formaldehyde . . . . .	65
3.18	Non-reactive ANNs prediction for the formaldehyde trajectory . . . . .	66
3.19	Reactive ANNs prediction for the formaldehyde trajectory . . . . .	67
4.1	The three axes of rotation of formaldehyde . . . . .	76
4.2	Results from Wodtke et al’s work. . . . .	77
4.3	Rotational energies and trapping probabilities as a function of collision energy for $T_s = 300$ K . . . . .	78
4.4	Population distribution of rotational energies of formaldehyde . . . . .	78
4.5	Number of bounces as a function of collision energy . . . . .	79
4.6	Average rotational energy of trapping-desorption trajectories and direct-scatter trajectories . . . . .	80
4.7	Rotational energy distribution of trapping-desorption trajectories and direct-scatter trajectories for 1.21 eV of collision energy and $T_s = 300$ K . . . . .	81
4.8	Effect of surface temperature on rotational energies and the trapping probabilities . . . . .	82
4.9	Scattering angle distribution for random initial orientation of formaldehyde for all temperatures and collision energies . . . . .	84
4.10	Different orientations used in the study for formaldehyde-gold system . . . . .	85
4.11	Trapping probabilities as a function of collision energy for different orientations . . . . .	85

4.12	Effect of orientation on rotational energies for $T_s = 300$ K . . . . .	86
4.13	Effect of orientation on rotational energies for $T_s = 200$ K . . . . .	86
4.14	Effect of orientation on rotational energies for $T_s = 400$ K . . . . .	87
4.15	Distribution of energy transfer between formaldehyde and the gold sheet for 1.21 eV of collision energy and $T_s = 300$ K . . . . .	88
4.16	Energy transfer for trajectories starting with random orientation for $T_s =$ 300 K for 1.21 eV . . . . .	90
4.17	Energy transfer for trajectories starting with OR <sub>1</sub> orientation for $T_s = 300$ K for 1.21 eV . . . . .	91
4.18	Variables used to define the interaction potential between gold and formalde- hyde . . . . .	92
4.19	Interaction potential as a function of $\beta$ , $\gamma$ and $C_z$ . . . . .	93
4.20	The initial orientations at the starting of the trajectory . . . . .	94
4.21	Orientation at the point of closest approach . . . . .	95
4.22	Dynamics after the point of closest approach . . . . .	96
5.1	The graphene surface and the formaldehyde at the starting geometry . . . . .	99
5.2	Trapping probability as a function of collision energy for formaldehyde- gold and formaldehyde-graphene systems . . . . .	102
5.3	$C_z$ distribution of trapped formaldehyde. . . . .	103
5.4	Average rotational energy of formaldehyde as a function of collision energy	104
5.5	Population distribution of rotational energies of formaldehyde for 1.21 eV of collision energy . . . . .	104
5.6	Average rotational energy of trapping-desorption and direct-scattered tra- jectories . . . . .	105
5.7	Different orientations used in the study of formaldehyde-graphene system . .	106
5.8	Trapping probabilities for different orientations of formaldehyde . . . . .	106
5.9	Average total rotational energies for different orientations of formaldehyde .	107
5.10	The components of rotational energies for different orientations . . . . .	108
5.11	Minimum potential energy surface of formaldehyde-graphene interaction as a function . . . . .	109

5.12	Point of closest approach for formaldehyde-graphene system . . . . .	110
5.13	Projection of trajectories on the minimum energy surface of the interaction potential . . . . .	111
5.14	Energy transfer for 1.21 eV of collision energy . . . . .	113
5.15	Energy transfer for 1.21 eV of collision energy for different orientations . .	114

# List of Tables

2.1	Table of different sampling methods used for ANN fit for $\text{HeH}^+ + \text{H}_2$ system	21
2.2	Comparison of fits obtained using different networks . . . . .	30
2.3	Table of different data sets generated for fitting the $\text{C}_5\text{N}^- + \text{H}_2$ PES . . . . .	34
2.4	Training and network performance for different networks when 100% data was used for $\text{C}_5\text{N}^- + \text{H}_2$ PES . . . . .	36
2.5	Training and network performance for different networks when 70% data was used for $\text{C}_5\text{N}^- + \text{H}_2$ PES . . . . .	36
2.6	Training and network performance for different networks when data was sampled using LHS scheme for $\text{C}_5\text{N}^- + \text{H}_2$ PES . . . . .	37
2.7	Table of different fits for the $\text{C}_2^- + \text{H}_2$ system . . . . .	43
4.1	Definition of L-J parameters used to represent gold-formaldehyde interaction.	73
5.1	Parameters for graphene-formaldehyde interaction. . . . .	100
A.1	Statistics of the trajectories for gold-formaldehyde system . . . . .	126
B.1	Statistics of the trajectories for graphene-formaldehyde system . . . . .	129

# Chapter 1

## Introduction

Chemistry is the study of matter at the atomistic level. How electrons are bound to the nucleus of an atom, and how these atoms interact with other atoms/molecules/light, gives rise to many types of interesting chemical phenomenon. Understanding chemical reactions and controlling them, such as the use of fire to extract metals from ores, or extracting chemicals from plants to be used as medicine, have been the cornerstone for modern civilization. Knowledge of chemistry has advanced not only the material industry, but also has led us to deepen our understanding of biological systems and the universe.

Experimental methods to study fundamental processes governing chemical reactions, like energy transfer have their limitations. Experiments are challenging to design and the observations are difficult to assign to individual events. On the other hand, statistical theories such as transition state theory<sup>1,2</sup> (TST) and Rice-Ramsperger-Kassel-Marcus theory<sup>2-5</sup> (RRKM) can give accurate rate constants for a great deal of chemical reactions<sup>6,7</sup> based on the energetics of the system, but there have been exceptions to this.<sup>8-10</sup> They fail to capture the dynamical nature of chemical reactions.

One way to overcome this issue is to use computations, starting from fundamental properties of sub-atomic particles, like the interaction between electrons and the nucleus of atoms, all the way to simulating entire chemical reactions, which give atomistic insights into the very fundamental processes. However, this method can overwhelm most current computational facilities for systems involving only a few tens of atoms. Thus, approximations are made, such as decoupling nuclear and electronic motion and using analytical representations of potential energy surfaces (PESs) to simulate reactions. However repre-

sensation of PES using analytical functions is itself quite challenging and has been limited to systems with only a few atoms (less than 10 atoms) and is not routine.

To this end, this thesis is divided into two types of work. The first two chapters discuss the use of Machine Learning (ML) methods to represent PES. The next two chapters explore inelastic non-reactive scattering of the polyatomic molecule  $\text{H}_2\text{CO}$  on surfaces.

PESs are fundamental to understand chemical reaction dynamics. Having the knowledge of the PES, one can simulate a reaction in order to get dynamical information about how the reaction proceeds. It has been observed that dynamical interpretation of chemical reactions can often differ from that predicted by statistical theories. One way to study the dynamics is by using classical mechanics i.e., solving the Newton's<sup>11</sup> equations of motion (EOM) for the system. But to do that, we need accurate information about the PES because forces acting on each atom at any given time depends on the gradient of PES. For a system of  $N$  atoms, the PES is a function of  $3N - 6$  variables which becomes a bottleneck for simulating larger systems. And in order to generate statistically significant observables from computer simulations, an ensemble of trajectories need to be computed.

*Ab initio* dynamics, also known as direct dynamics solves the issue of tackling larger systems, albeit at the expense of computational cost. This algorithm solves the classical EOMs using quantum chemical methods to generate forces at each time step. The forces are calculated by solving the electronic time-independent Schrödinger equation (TISE), for a given a set of nuclear coordinates.

In what follows, is a brief introduction of PES, the methods used for computing trajectories, the details of two machine learning (ML) methods, namely Gaussian Processes for Regression (GPR) and Artificial Neural Networks (ANN).

## 1.1 Potential Energy Surface (PES)

Given a molecule with  $N$  atoms, the non-relativistic Hamiltonian of the system in atomic units is,<sup>12</sup>

$$\hat{H} = -\frac{1}{2} \sum_{\alpha} \frac{1}{m_{\alpha}} \nabla_{\alpha}^2 - \frac{1}{2} \sum_i \nabla_i^2 + \sum_{\alpha} \sum_{\beta > \alpha}^N \frac{Z_{\alpha} Z_{\beta}}{R_{\alpha\beta}} - \sum_{\alpha} \sum_i \frac{Z_{\alpha}}{r_{i\alpha}} + \sum_j \sum_{i > j} \frac{1}{r_{ij}} \quad (1.1)$$

where  $\alpha$  and  $\beta$  are the indices of the nuclei, and  $i$  and  $j$  are the indices of the electrons.  $m$  is the mass of the nuclei,  $Z$  is the nuclear charge, and  $R$  is the distance between two nuclei,  $r_{i\alpha}$  is the distance between electron and nuclei, and  $r_{ij}$  is the distance between two electrons. The first and the second terms in Eqn. 1.1 are the kinetic energy of the nuclei and electrons respectively. The third term is the nuclear-nuclear repulsion term, fourth the electron-nuclei attraction term and the last term is the electron-electron repulsion term. The time-independent Schrödinger equation is then,

$$\hat{H}\psi(q_i, q_{\alpha}) = E\psi(q_i, q_{\alpha}) \quad (1.2)$$

where  $\psi(q_i, q_{\alpha})$  is the wavefunction of the molecule,  $q_{\alpha}$  are the nuclear coordinates, and  $q_i$  are the electronic coordinates. By invoking Born-Oppenheimer (BO) approximation, which states that due to the electronic motion being much faster than the nuclear motion, we can separate the wavefunction into nuclear part and electronic part and thus we can write the wavefunction as a product of the two. The wave function hence can be written as,

$$\psi(q_i, q_{\alpha}) = \psi_{\text{el}}(q_i; q_{\alpha}) \psi_{\text{n}}(q_{\alpha}) \quad (1.3)$$

where  $\psi_{\text{el}}$  is the electronic wavefunction and  $\psi_{\text{n}}$  is the nuclear wavefunction. The electronic wavefunction depends parametrically on the nuclear coordinates. The electronic Hamiltonian is then,

$$\hat{H}_{\text{el}} = -\frac{1}{2} \sum_i \nabla_i^2 - \sum_{\alpha} \sum_i \frac{Z_{\alpha}}{r_{i\alpha}} + \sum_j \sum_{i > j} \frac{1}{r_{ij}} \quad (1.4)$$

The electronic energies for a given set of nuclear coordinates can be computed by solving,

$$\hat{H}_{\text{el}}\psi_{\text{el}} = E_{\text{el}}\psi_{\text{el}} \quad (1.5)$$

The potential energy surface ( $V$ ) is defined as the sum of electronic energy ( $V_{\text{el}}$ ) and the nuclear-nuclear repulsion term ( $V_{\text{nn}}$ ),

$$V = E_{\text{el}} + V_{\text{nn}} \quad (1.6)$$

For an  $N$  atom system there are  $3N$  Cartesian coordinates describing the nuclear geometry of the molecule; we can remove the translations and the rotations and write the nuclear coordinates in the form of  $3N - 6$  internal coordinates ( $Q$ ). The form for the PES is then written as,

$$V(q_1, q_2, \dots, q_{3N}) \equiv V(Q_1, Q_2, \dots, Q_{3N-6}) \quad (1.7)$$

where,

$$\{q_1, q_2, \dots, q_{3N}\} \equiv \{(x_1, y_1, z_1), (x_2, y_2, z_2), \dots, (x_N, y_N, z_N)\} \quad (1.8)$$

As an example see Figure 1.1 for the potential energy curve (PEC) of  $\text{H}_2$  molecule, which is a function of only the internuclear distance  $r$ , computed using the coupled cluster singles and doubles (CCSD) method and the aug-ccpVTZ basis set.

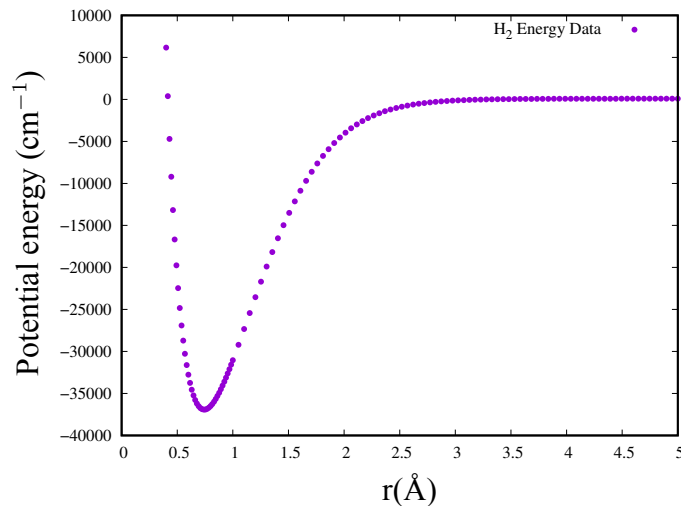


Figure 1.1: Potential energy curve of  $H_2$  molecule.

Forces on any atom can be computed by taking derivative of the PES with respect to its coordinates. Once the forces are computed, the trajectory of the entire system can be simulated by integrating the classical EOMs of the system, and the system evolves with time as illustrated in the following section.

## 1.2 Classical Trajectory Simulations

A trajectory of a molecule is defined as the set of coordinates and momenta of all atoms as a function of time. To compute trajectories we need the initial conditions, specifically the initial positions and momenta. A trajectory of the  $i^{th}$  atom in the system then can be defined as,

$$(x_i(t), y_i(t), z_i(t), p_x^i(t), p_y^i(t), p_z^i(t)) \quad (1.9)$$

where  $i$  goes from 1 to  $N$ . After an infinitesimally small amount of time  $\Delta t$ , according to the Taylor expansion of Eqn 1.9 the new position of the  $i^{th}$  atom can be written as,

$$\begin{aligned} x_i(t + \Delta t) &= x_i(t) + \dot{x}_i(t)\Delta t + \frac{1}{2}\ddot{x}_i\Delta t^2 + \dots \\ y_i(t + \Delta t) &= y_i(t) + \dot{y}_i(t)\Delta t + \frac{1}{2}\ddot{y}_i\Delta t^2 + \dots \\ z_i(t + \Delta t) &= z_i(t) + \dot{z}_i(t)\Delta t + \frac{1}{2}\ddot{z}_i\Delta t^2 + \dots \end{aligned} \quad (1.10)$$

We can truncate the series after three terms because the higher powers of  $\Delta t$  tend to zero. The first term in Eqn 1.10 is the initial position (at time  $t$ ) of the  $i^{th}$  atom, the  $\dot{x}_i(t)$  is the velocity at time  $t$ , and  $\ddot{x}_i = a_x^i$  is the acceleration felt by the  $i^{th}$  atom at time  $t$  due to the force. The acceleration of the  $i^{th}$  atom with mass  $m_i$  can be computed using  $\vec{F}_i = m_i \cdot \vec{a}_i$ , which can be written in the Cartesian coordinates as,

$$\begin{aligned} F_{x_i} &= -\frac{\partial V}{\partial Q_i} \times \frac{\partial Q_i}{\partial x_i} \\ F_{y_i} &= -\frac{\partial V}{\partial Q_i} \times \frac{\partial Q_i}{\partial y_i} \\ F_{z_i} &= -\frac{\partial V}{\partial Q_i} \times \frac{\partial Q_i}{\partial z_i} \end{aligned} \quad (1.11)$$

### 1.2.1 Velocity-Verlet Integration

Velocity Verlet integration scheme is an algorithm to compute trajectories. In this scheme the steps are,

- Calculate new positions using,  $x_i(t + \Delta t) = x_i(t) + \dot{x}_i(t)\Delta t + \ddot{x}_i(t)\Delta t^2$
- Compute  $\ddot{x}_i(t + \Delta t)$  from potential.
- Calculate new velocities using,  $\dot{x}_i(t + \Delta t) = \dot{x}_i(t) + \frac{1}{2}(\ddot{x}_i(t) + \ddot{x}_i(t + \Delta t))\Delta t$ .

### 1.2.2 Sampling

To compute trajectories, the initial conditions, i.e., the momenta and coordinates are needed. Each combination of momenta and coordinates defines a point in the  $6N$  dimensional phase

space, and each point has its corresponding total energy. The total energy of the molecule can also be expressed as,

$$E_{\text{Total}} = E_{\text{el}} + E_{\text{vibrational}} + E_{\text{rotational}} + E_{\text{translational}} \quad (1.12)$$

where  $E_{\text{el}}$  is the electronic energy,  $E_{\text{vibrational}}$  is the energy associated with the oscillatory motion of all the nucleus in the system,  $E_{\text{rotational}}$  is the kinetic energy corresponding to the rotation of the molecule and  $E_{\text{translational}}$  is the kinetic energy of translation for the molecule. Sampling is the technique used to select an ensemble of trajectories for which either the total energy is constant ( $NVE$ ), or the temperature is constant ( $NVT$ ). For different types of reactions one could do different types of sampling. For example, for a gas phase bimolecular collision reaction, the kinetic energy of approach of the two molecules can be chosen from a Boltzmann distribution of velocities corresponding to a given temperature. Other modes of kinetic energy such as rotational energy and vibrational energy can also be chosen in a similar way.

In this thesis, we have sampled rotational energies and vibrational energies for the ground electronic state of the polyatomic molecule formaldehyde, using Boltzmann distribution. The rotational energy at a temperature of  $T_{\text{rot}}$  is sampled assuming  $I_x < I_y = I_z$ , where  $x$  defines the symmetry axis. The  $x$  component ( $J_{x(z)}$ ) of the total angular momentum ( $J$ ) is sampled using the probability distribution,<sup>13</sup>

$$P(J_{x(z)}) = e^{-J_{x(z)}^2/2I_{x(z)}k_B T_{\text{rot}}} \quad 0 \leq J_{x(z)} \leq \infty \quad (1.13)$$

$$P(J) = J e^{-J^2/2I_{z(x)}k_B T_{\text{rot}}} \quad J_{x(z)} \leq J \leq \infty \quad (1.14)$$

where  $J_{x(z)}$  is sampled from  $P(J_{x(z)})$  by the rejection method.  $J$  is sampled by the cumulative distribution formula.

$$J = \left[ J_{x(z)}^2 - 2I_{z(x)}k_B T_{\text{rot}} \ln(1 - R_1) \right]^2 \quad (1.15)$$

where  $R_1$  is a freshly chosen random number uniformly distributed between 0 to 1.

Similarly the normal mode vibrational energies are sampled from a Boltzmann distribution at temperature  $T_{\text{vib}}$ , in which the quantum number  $n_i$  in the  $i^{\text{th}}$  mode (vibrational frequency  $\nu_i$ ) is sampled using the probability distribution,

$$P(n_i) = e^{\left(\frac{-n_i h \nu_i}{k_B T_{\text{vib}}}\right)} \left[1 - e^{\left(\frac{h \nu_i}{k_B T_{\text{vib}}}\right)}\right] \quad (1.16)$$

### 1.3 Machine Learning

Machine learning, as the name suggests is the study of making man-made machines, in this case computers, learn, recognize patterns and predict events before they occur.<sup>14</sup> This came into existence to meet engineering demands, such as computer codes which can handle events for which they have not been coded for, like recognizing speech or handwritten letters, and in the field of weather prediction. ML deals with representing data obtained by experiments. Traditionally, such problems have been dealt with using procedures known as “regression” methods, such as Linear Least Square method. They solve the problem by fitting the observed data to some analytical function. This method requires the nature of the function to be known beforehand, and becomes increasingly difficult for higher dimensional data.

In other fields of science including chemistry, ML has found many applications.<sup>15</sup> The two main problems that ML deals with is that of regression and classification. For regression related problems, ML has been used to represent high dimensional PESs,<sup>16–19</sup> using  $\Delta$ –machine learning methods<sup>20</sup> to bring DFT-based PESs to CCSD(T) accuracy, predicting bulk properties of materials and even coming up with new potential materials which will have a certain desired property.<sup>21</sup> On the other hand classification related problems have more diverse applications, such as predicting product distributions given the initial

state of the reactants,<sup>22</sup> classification of spectra for different materials,<sup>23</sup> identification of mechanism of reactions using kinetic data,<sup>24</sup> etc.

Two of the popular ML methods will be discussed here briefly, namely artificial neural networks<sup>25</sup> and gaussian processes for regression.<sup>26</sup>

### 1.3.1 Gaussian Processes for Regression

GPR is a kernel based statistical learning technique, also known as “kriging”. It was first developed by Daniel G. Krige in the 1950s,<sup>27</sup> for finding out mineral concentrations at unknown places in a mine, based on samples collected elsewhere. And in 1960, French mathematician Georges Matheron continued work on it and developed the field of geo-statistics.<sup>28</sup> More recently, GPR has found application in the field of representing PES.<sup>16</sup> It has been used to represent the 6-dimensional PES of  $N_4$  using only 240 points. Further, a representation of the PES using around 1500 points using GPR, whose accuracy was comparable to a traditional fit using 16000 *ab initio* points was also illustrated.

A one-dimensional Gaussian is given by,

$$f(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad (1.17)$$

where  $\mu$  is the mean, and  $\sigma^2$  is the variance. In Gaussian process, each point of training data output, is considered as a Gaussian. With  $N$  data points, one ends up with an  $N$ -dimensional Gaussian distribution.

Consider a  $q$  dimensional real valued function,  $F : \mathbb{R}^q \mapsto \mathbb{R}$ , for which the values at  $N$  points are known. This constitutes the training set,

$$Y^N = \{F(\vec{\mathbf{x}}^1), F(\vec{\mathbf{x}}^2), \dots, F(\vec{\mathbf{x}}^N)\}, \quad (1.18)$$

$$\vec{\mathbf{x}}^i = (x_1^i, x_2^i, \dots, x_q^i)^T \quad (1.19)$$

We assume that multiple outputs of GP at the given  $N$  points follow a multivariate normal distribution, i.e. they are jointly distributed as,

$$Y^N \sim MVN(\beta, \sigma^2 \mathbf{A}), \quad (1.20)$$

where  $\beta$  is the mean vector, and  $\mathbf{A}$  is an  $N \times N$  matrix, defined by,

$$\mathbf{A} = \begin{bmatrix} 1 & R(\vec{\mathbf{x}}^1, \vec{\mathbf{x}}^2) & \dots & R(\vec{\mathbf{x}}^1, \vec{\mathbf{x}}^N) \\ R(\vec{\mathbf{x}}^2, \vec{\mathbf{x}}^1) & 1 & & \vdots \\ \vdots & & \ddots & \\ R(\vec{\mathbf{x}}^N, \vec{\mathbf{x}}^1) & \dots & & 1 \end{bmatrix} \quad (1.21)$$

and,

$$R(\vec{\mathbf{x}}, \vec{\mathbf{x}}') = \left\{ \prod_{i=1}^q \exp\left(-\frac{(x_i - x_i')^2}{\omega_i^2}\right) \right\} \quad (1.22)$$

is the correlation function. For demonstration purpose we have chosen it to be a squared exponential function. Other types of correlation functions maybe used, such as Matérn correlation function. The best strategy that has been suggested involves building accurate PES with the smallest number of potential energy calculations.<sup>29</sup> The GPR model is determined by correlations between points in a multidimensional configuration space. The correlation function captures how far the inputs ( $\vec{x}_i$  and  $\vec{x}_j$ ) are from each other as shown in Figure 1.2. The closer they are, we can expect the value of the correlation function to be higher. And it can be said that the value of the function,  $F(\vec{x}_i)$  and  $F(\vec{x}_j)$  will be similar.

The log-marginal-likelihood function is given by,

$$\log \mathcal{L}(\omega \mid Y^N, X) = -\frac{1}{2}(N \log \hat{\sigma}^2 + \log(\det(\mathbf{A})) + N) \quad (1.23)$$

where,

$$\hat{\sigma}^2(\omega) = \frac{1}{N}(Y^N - \beta)^T \mathbf{A}^{-1}(Y^N - \beta), \quad (1.24)$$

$$\hat{\beta}(\omega) = (\mathbf{I}^T \mathbf{A}^{-1} \mathbf{I})^{-1} \mathbf{I}^T \mathbf{A}^{-1} Y^N \quad (1.25)$$

are the maximum likelihood estimators, and  $\mathbf{I}$  is the identity matrix. These estimators are computed by taking the derivative of the log-marginal-likelihood function with respect to

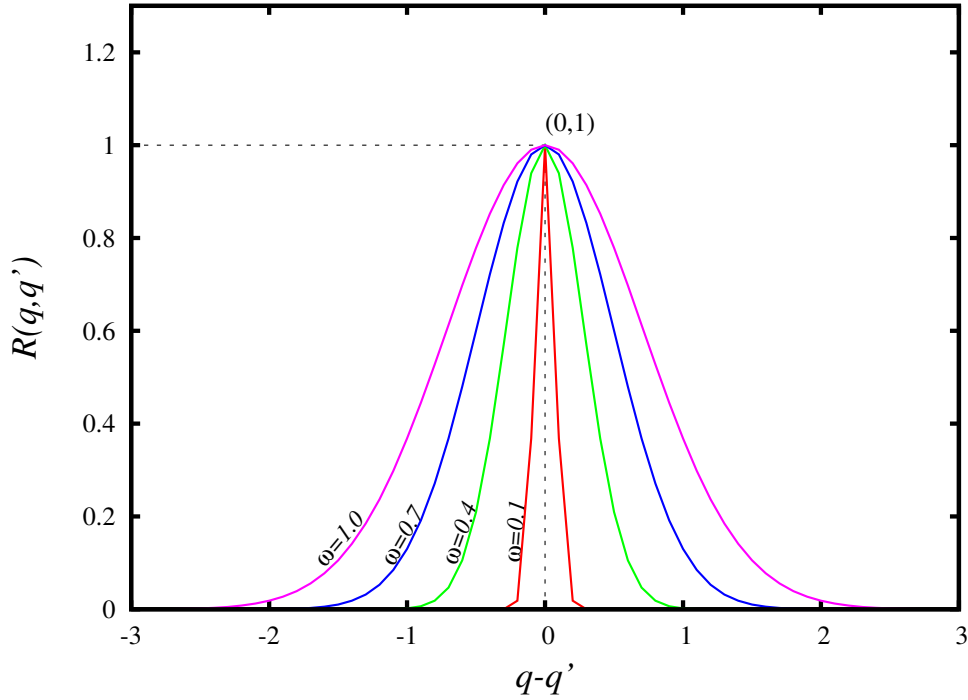


Figure 1.2: Squared exponential function correlation function.

the  $\sigma^2$  and  $\beta$ , and setting them to zero. Once the maximum of log-likelihood function is iteratively determined, with respect the  $\omega_i$ s, we call the system to have been trained, and prediction at an untried point  $\vec{x} = \vec{x}_0$  can be made, by considering the point  $Y_0 = F(\vec{x}_0)$  to be jointly distributed as,

$$\begin{pmatrix} Y_0 \\ Y^N \end{pmatrix} \sim MVN \left[ \begin{pmatrix} 1 \\ \mathbf{I} \end{pmatrix}, \sigma^2 \begin{pmatrix} 1 & \mathbf{A}_0^T \\ \mathbf{A}_0 & \mathbf{A} \end{pmatrix} \right] \quad (1.26)$$

where  $\mathbf{A}_0 = (R(\vec{\mathbf{x}}^0, \vec{\mathbf{x}}^1), R(\vec{\mathbf{x}}^0, \vec{\mathbf{x}}^2), \dots, R(\vec{\mathbf{x}}^0, \vec{\mathbf{x}}^N))^T$  is a column vector determining the correlation of the new point to all other points. The mean of the prediction is now given by,

$$\tilde{\mu}(\vec{\mathbf{x}}^0) = \beta + \mathbf{A}_0^T \mathbf{A}^{-1} (\mathbf{Y}^N - \beta) \quad (1.27)$$

and the conditional variance as,

$$\tilde{\sigma}^2(\vec{\mathbf{x}}^0) = \sigma^2 (1 - \mathbf{A}_0^T \mathbf{A}^{-1} \mathbf{A}_0) \quad (1.28)$$

GPR by construction, does not have the problem of overfitting, and are guaranteed to become more accurate when trained using more points. As can be seen from Eqn 1.24 and Eqn 1.25, training for GPR consists of inverting a square matrix of size  $N \times N$  iteratively, which becomes a bottleneck as the number of training points ( $N$ ) increases.

### 1.3.2 Artificial Neural Network

ANNs, also known as “Multilayer Perceptron”, try to mimic the neurons of mammalian brains, in order to solve problems of similar nature described in the previous section. It consists of interconnected nodes, which pass information between each other. Every interconnection between nodes, has a direction, and a weight associated with it. How these nodes are connected to each other gives the ability to the neural network to form complex relationships.

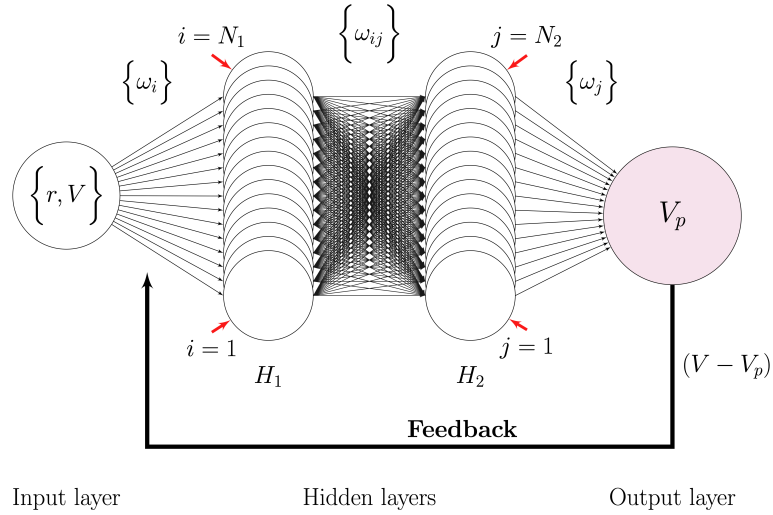


Figure 1.3: An example of an ANN with two hidden layers, consisting of  $N_1$  and  $N_2$  nodes in each layer.

ANNs usually consists of three types of nodes, which can be categorized into input layer, hidden layer/layers, and output layer, and information flow occurs according to this given

order, see Figure 1.3. Each node, other than the input layer nodes, receives multiple inputs, which are then summed, see Figure 1.4. Inputs to each node are modified by an associated weight. And the output of each node is the summation of the weighted inputs, which then is passed through a transfer function. Example of a transfer function is shown in Figure 1.5. The only restriction is that the output of any node from a layer, has to move forward to the next layer of nodes, hence they are also referred to as “Feed Forward Network”.

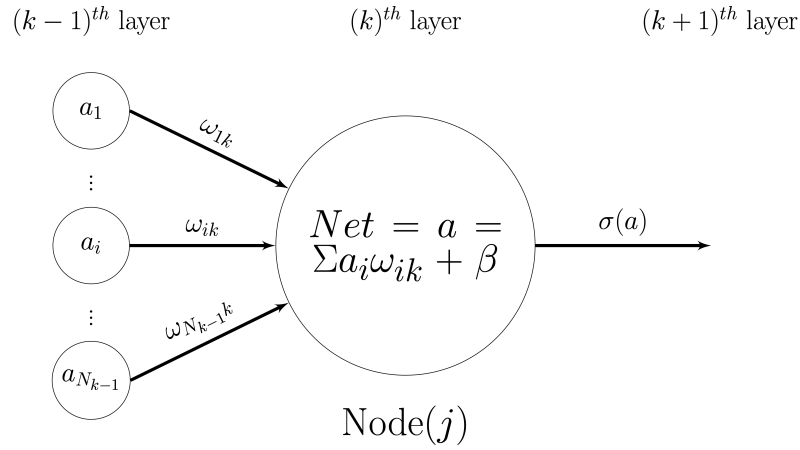
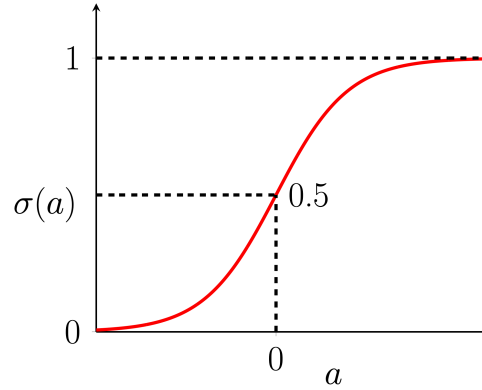


Figure 1.4: Schematics of the input to a node in the neural network and the output through a transfer function.

The size of the neural network (number of hidden layers and the number of nodes in each layer) depends on the complexity of the function one is trying to fit. Training involves dividing the entire data set into two sets, training set and test set. Training set is used to train the network, whereas the test set is used to validate the fitting, and making sure that network has not been over trained (over-fitting). This division is often done by randomly dividing the set into a ratio of 80:20 (training set:test set) or 70:30. In our case, we have also done fitting with 98% of data in the training set.



$$\text{Transfer function } \sigma(a) = \frac{1}{1+e^{-a}}$$

Figure 1.5: Plot of the sigmoid type transfer function.

The weights and biases (parameters) of the network are selected randomly. The process of training involves optimizing the parameters in such a way that the output of the network becomes closer to the real value of the function. This is done by first dividing the training set into equal sized batches. Each batch is fed to the network, and the errors are computed. The weights are modified in such a way that the error for the entire batch is reduced. First the parameters of connections between the penultimate layer and the output layer are modified. And then the connections in the layer before that are modified, this is done till the weights in the input layer and first hidden layer are modified. This process is known as “back propagation of errors”.

This is done iteratively for all the batches, and the process of exposing the entire training data to the neural network once, is known as an “epoch”. Training involves multiple epochs, and is stopped when the desired level of accuracy is achieved. It should be noted that for function fitting, multiple starting conditions for the parameters need to be tested. If a given size of network is not able to achieve the desired level of accuracy, a larger network is

recommended. This is done on a hit and trail basis. Nevertheless, this method is preferred over traditional curve-fitting methods because of its simplicity and that, a functional form need not be defined.

There is no straightforward way to select the network architecture and network size. In principle, one starts of with a small network size, and keeps increasing the network size (the number of layer/number of nodes in each layer) till the level of desired accuracy is achieved. Having more number of layers have been shown to be effective in classification tasks, whereas increasing the depth of the network seems to impede the quality of predictions for regression tasks.<sup>30</sup> Hence it is suggested to use a wider network for PES representations. It should be pointed out that having a network which is too large can result in overfitting. Appropriate cross-validation checks need to be in place so that overfitting is avoided. The correct network size for a given problem can be estimated in this way, where the network is large enough to represent the data, but not so large that the predictions at test data (data not used for training) have large errors. Having similar residuals for both training data and test data is recommended.

The learning rate is a parameter, which decides how much the weights are modified in during each iteration (epoch). Having a low learning rate ensures that while optimizing the weights, the network avoids going into a local minima. Which is essential during any optimization task. The training time depends on a few factors, such as the network size, the size of the training data, the complexity of the function one is trying to fit and also the level of accuracy that is required. For fitting potential energy surfaces, we also need to make sure that the function behaves smoothly where no data is available.

## Chapter 2

# Representation of potential energy surfaces using GPR and ANN

Astrochemistry is the interdisciplinary science of studying the formation and destruction of molecules in the astrophysical environment, which occur under the umbrella of gas-phase reactions or reactions at surfaces (dust grains). Understanding fundamental reactions in the interstellar medium (ISM) helps us to build a network of reactions, and studying their underlying energy transfer processes gives us answers to the possible origin of stars. Concentration of different light elements/ions in interstellar clouds greatly impact the future of the cloud. The ISM is very different from the typical environment observed on earth. The molecular densities are of the order of  $10^{14}$  molecules/cm<sup>3</sup> compared to about  $10^{21}$  molecules/cm<sup>3</sup> found on earth. Also the temperature ranges from 5 K to about a 100 K in the ISM, which makes studying these processes in detail in a laboratory environment challenging. On the other hand, direct observation of abundances in the ISM are made using sub-millimeter radio telescopes to observe line spectra of different molecules. But due to the presence of multiple types of molecular species in the ISM, it becomes difficult to categorize them.

Astrophysical detection relies heavily on the spectroscopic investigations of such species in the laboratory and on matching sighted lines with those observed in the laboratory. Experiments are regularly conducted to compute rate coefficients for formation and destruction of different molecules in laboratory mimicking as closely as possible, the conditions in ISM, which then helps in deciphering the observed data from telescopes. Another way to overcome this problem is to study these chemical reactions using computer simulations,

either using quantum dynamics or classical dynamics to compute rate coefficient, which in turn can give an estimate of the abundances of such chemicals in the ISM and helps us to understand the evolution of these chemicals. In such types of studies, the fitting of PES is important to carry out further calculations. To this end, two ML methods were used to fit the PESs of three molecular systems found in the ISM. Here we fitted the PES of three systems.  $\text{HeH}^+ + \text{H}_2$ ,  $\text{C}_5\text{N}^- + \text{H}_2$  and  $\text{C}_2^- + \text{H}_2$ , using ANN and GPR methods. The fitted PESs were then used to compute rate coefficients for rotational excitation processes. The rate coefficients were computed by our collaborators<sup>†</sup> and hence not reported here.

## 2.1 $\text{HeH}^+ + \text{H}_2$ *ab initio* PES

When the temperature of the universe cooled to about 4000 K, hundred thousand years after the big bang, ions of lighter elements started recombining in the reverse order of their ionization potential.<sup>31</sup> So  $\text{He}^{++}$  formed before  $\text{He}^+$ , followed by  $\text{H}^+$ .  $\text{HeH}^+$  was long believed to be the first compound in the universe formed via radiative association of neutral helium and  $\text{H}^+$ ,



which would lead to the formation of  $\text{H}_2^+$  by reaction with hydrogen atoms.



The  $\text{H}_2^+$  can then undergo charge transfer reaction to form  $\text{H}_2$  molecule,



<sup>†</sup> Electronic structure calculations were done by E. Yurtsever, *Department of Chemistry, Koc University Rumelifeneriyolu, TR 34450 Istanbul, Turkey*; Quantum scattering calculations were done by K. Giri, *Department of Computational Sciences, Central University of Punjab, Bathinda, Punjab 151401, India*; and L. González-Sánchez, *Departamento de Química Física, University of Salamanca, 37008 Salamanca, Spain*; and analysis was done by F. A. Gianturco, *Institut für Ionenphysik und Angewandte Physik, Universität Innsbruck, A-6020 Innsbruck, Austria*; and N. Sathyamurthy, *Indian Institute of Science Education and Research Mohali, SAS Nagar, Punjab 140306, India*. For details see *J. Phys. Chem. A* (2022), **126**, 14, 2244–2261.

But this compound ( $\text{HeH}^+$ ) had not been observed until recently in 2019, by Gusten et al.,<sup>32</sup> analyzing the data from Stratospheric Observatory for Infrared Astronomy (SOFIA) in the NGC 7027 planetary nebula. So it becomes important to study the dynamics of  $\text{HeH}^+$  when colliding with other partners. Interaction of  $\text{HeH}^+$  with  $\text{H}_2$  is known to be reactive giving the following reaction,



Adams et al.<sup>33</sup> estimated the rate coefficient of this reaction to be greater than  $3.5 \times 10^{-11} \text{ cm}^3 \text{ molecule}^{-1} \text{ s}^{-1}$  at 200 K ( $\approx 0.017 \text{ eV}$ ) using flowing afterglow method. Ryan and Graham<sup>34</sup> measured it to be  $(1.4 \pm 0.2) \times 10^{-9} \text{ cm}^3 \text{ molecule}^{-1} \text{ s}^{-1}$  at a mean ion energy of 0.1 eV using an ion trap in an ion source mass spectrometer. By investigating crossed ion beam-neutral gas collisions, Rutherford and Vroom<sup>35</sup> estimated the rate coefficient value to be  $2.3 \times 10^{-9} \text{ cm}^3 \text{ molecule}^{-1} \text{ s}^{-1}$  at a mean energy of 0.3 eV. They reported the reaction cross section to be  $38 \text{ \AA}^2$  for a relative translational energy ( $E_{\text{trans}}$ ) of 0.3 eV, which decreases to  $\approx 1 \text{ \AA}^2$  around  $E_{\text{trans}} = 6 \text{ eV}$ . They also found the decay in the reaction cross section to be inversely proportional to the relative velocity of the reactants for  $E_{\text{trans}}$  in the range 0.4–2 eV, as predicted by Gioumousis and Stevenson.<sup>36</sup>

Here we report the ANN representation of the PES for the  $\text{HeH}^+ + \text{H}_2$  system.

### 2.1.1 Data Set

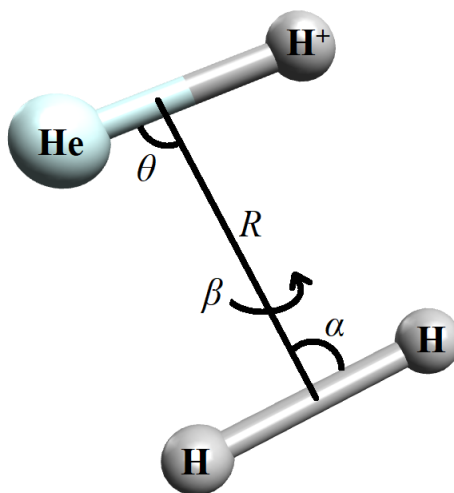


Figure 2.1: Coordinate system used to describe the  $\text{HeH}^+ + \text{H}_2$  system.

The coordinate system used to describe the  $\text{HeH}^+ + \text{H}_2$  system is shown in Figure 2.1. *Ab initio* calculations were done at the CCSD(T)/CBS level of theory to obtain the energies of different conformations, by varying the four variables described in Figure 2.1. The distance  $R$  was varied from 1.0 Å to 12 Å,  $\theta$  from  $0^\circ$  to  $180^\circ$ ,  $\alpha$  from  $0^\circ$  to  $180^\circ$ , and  $\beta$  from  $0^\circ$  to  $180^\circ$ . A total number of 15928 points were calculated. The electronic energies were calculated by our collaborator Ersin Yurtsever.

### 2.1.2 GPR

In this work, from the low energy points (less than  $3000 \text{ cm}^{-1}$ ), we used the Latin Hypercube Sampling<sup>37</sup> (LHS) to choose a data set of 1732 points to use in the GPR. LHS previously has been shown to perform better for representing multidimensional systems.<sup>16</sup>

The GPR representation of the PES using the 1732 data points gave a root mean square deviation (RMSD) of  $1415.43 \text{ cm}^{-1}$  for the entire data set. To improve the fitting, we explored other methods to fit the data.

### 2.1.3 Shallow Neural Network

The four dimensional PES for  $\text{HeH}^+ + \text{H}_2$  was represented using ANN. The MATLAB<sup>38</sup> software was used to train the ANN, and also to generate the subroutine for the given network. We used a shallow network consisting of one hidden layer with 60 nodes shown in Figure 2.2. A modified logistic sigmoid function of the form,

$$\sigma(a) = \frac{2}{1 + e^{-2a}} - 1 \quad (2.5)$$

where  $a$  is the sum of inputs to any given node, and  $\sigma(a)$  is the output of the node, was used as the transfer function. Training was done till the RMSD reached acceptable levels, after which the fits were tested for overfitting using the data points not included in the training set.

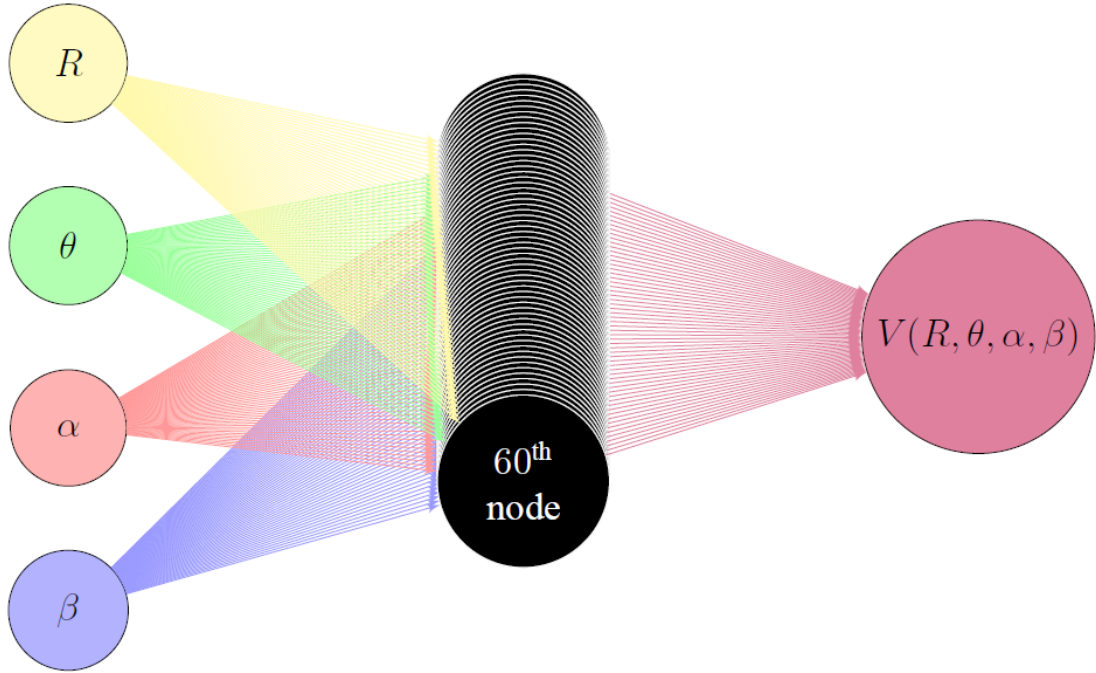


Figure 2.2: Neural Network design for the four dimensional PES function.

First low energy data points (total 14741 points) were used for training. Five types

of fitting were performed with different choices of data sets. In data sets 1 and 2, the *ab initio* points were sampled randomly using 70% and 98% of the data respectively. In set 3, the points were chosen using a uniform grid considering 50% of the data and including boundary data points. Sets 4 and 5 were chosen using LHS scheme of different sizes.

Table 2.1: Table of different sampling methods and their corresponding RMSD values of the trained ANN. “+” symbol for Set 6 signifies that it included higher energy data points.

Data Set	Sampling	No. of Training points	Maximum error (cm <sup>-1</sup> )	RMSD (cm <sup>-1</sup> )
1	Random(70%)	10318	208	11.10
2	Random(98%)	14446	-158	8.70
3	Grid	9026	9026	10.11
4	LHS1	2124	5462	107.10
5	LHS2	4055	-16027	244.00
6	Random(98%)+	15610	-934	35.35

The accuracy of the different ANNs produced are compared in Table 2.1 and Figures 2.3 and 2.4. All though the RMSD values reported here are reasonable, all fits (other than Set 6) suffered from over-fitting for different combinations of  $\alpha$  and  $\beta$ . Over-fitting occurs when the RMSD of a fit is low, yet the ANN fit does not vary smoothly over all the variables, yielding erroneous values where the values of the function are not known. It can be seen that the LHS performed poorly for the current system. This can be attributed to the fact that the number of points for the LHS were much lower compared to other schemes of sampling. LHS is done by iteratively finding points in  $n$ -dimensional space, in this case four dimensions. Every new point added to the set of generated LHS points has to be optimized with respect to all the points that have already been generated. Hence it becomes difficult to generate LHS points for higher dimensions when the number of points needed is large.

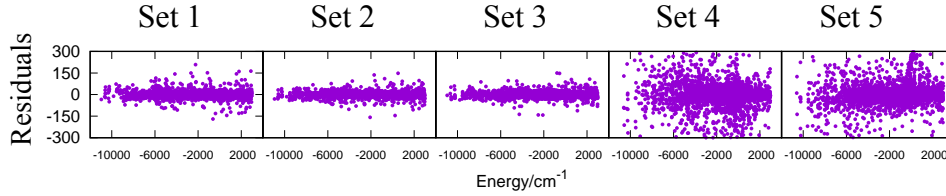


Figure 2.3: Comparison of residuals ( $\text{cm}^{-1}$ ) with respect to different sampling methods.

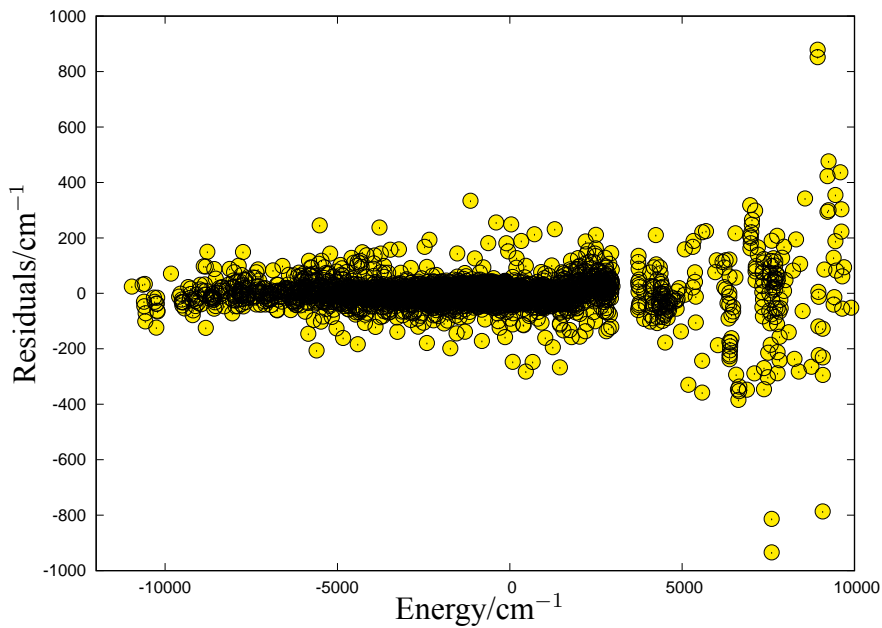


Figure 2.4: Scatter plot for the residuals of ANN (Set 6) compared to the *ab initio* data across the energy range of the PES. Darker regions shows higher concentration of points.

For Set 6 we added higher energy points to the data set, taking the total number of points to 15928 and did another ANN fit by randomly choosing 98% of the data for training. The data sampled for the fit is shown in Figure 2.5 for  $\alpha = 0^\circ$  and  $\beta = 0^\circ$ . We can see that the data covers the full region of the PES. This fit gave an overall RMSD of  $35.35 \text{ cm}^{-1}$  and a maximum error of  $-934 \text{ cm}^{-1}$  at  $\theta = 30^\circ$ ,  $\alpha = 180^\circ$ ,  $\beta = 0^\circ$  and  $R = 1.6 \text{ \AA}$  for  $V = 7597.80 \text{ cm}^{-1}$ . The RMSD for the test set was  $43.7 \text{ cm}^{-1}$  and that for the training set was  $35.2 \text{ cm}^{-1}$ . The residuals for all the data points obtained from the ANN fit are shown in Figure 2.4. The residuals near the minimum energy region of the PES is less than 200

$\text{cm}^{-1}$ , which is optimal for using the ANN PES for scattering calculations. This fit also avoided over-fitting, i.e, the value of  $V$  varied smoothly over  $\alpha$  and  $\beta$  variables as shown in Figure 2.6.

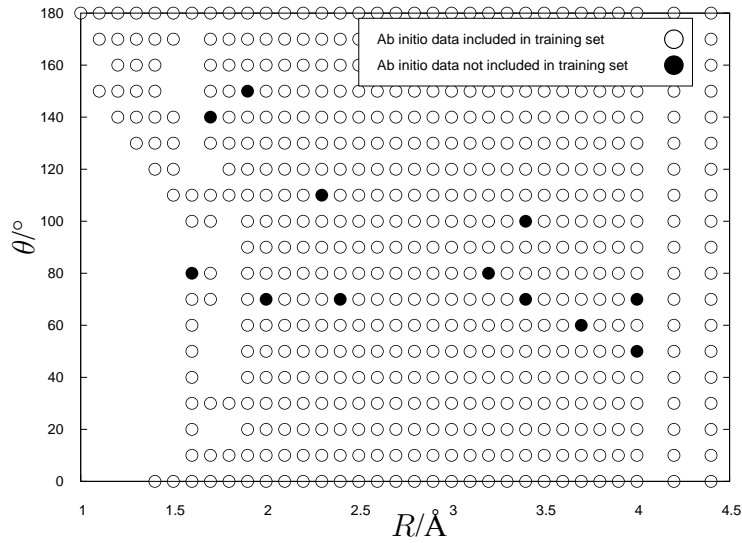


Figure 2.5: Slice of training data for  $\alpha = 0^\circ$  and  $\beta = 0^\circ$  generated by sampling 98% of the *ab initio* data randomly, which was used to train the ANN for Set 6.

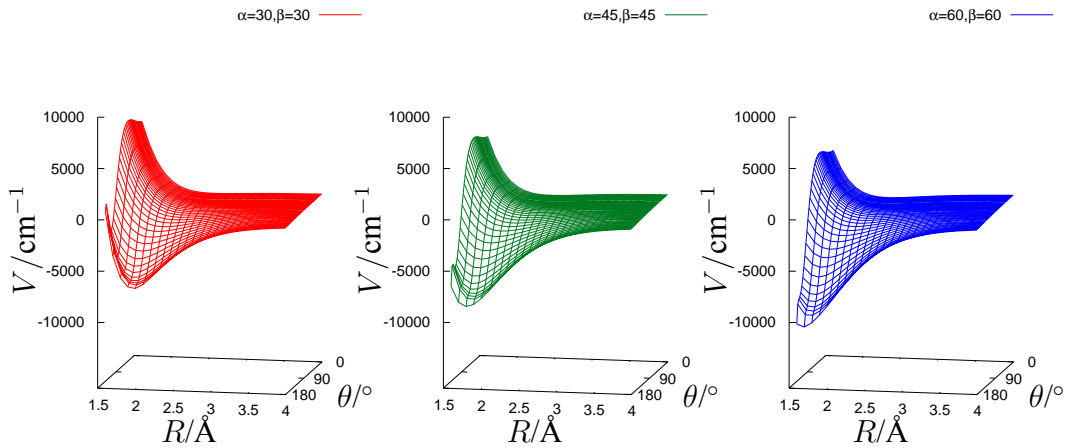


Figure 2.6:  $V$  as a function of  $R$  and  $\theta$  for different combinations of  $\alpha$  and  $\beta$ . The red plot (left) and the blue plot (right) are for values of  $\alpha$  and  $\beta$  where data is available, and the green plot (center) is for  $\alpha = 45^\circ$  and  $\beta = 45^\circ$ , where no data was available.

The PES and their corresponding contours for three different sets of fixed values of  $\alpha$  and  $\beta$  are plotted as a function of  $R$  and  $\theta$  in Figure 2.7. We can see that the ANN potential represents the *ab initio* data accurately without overfitting.

## 2 Representation of potential energy surfaces using GPR and ANN

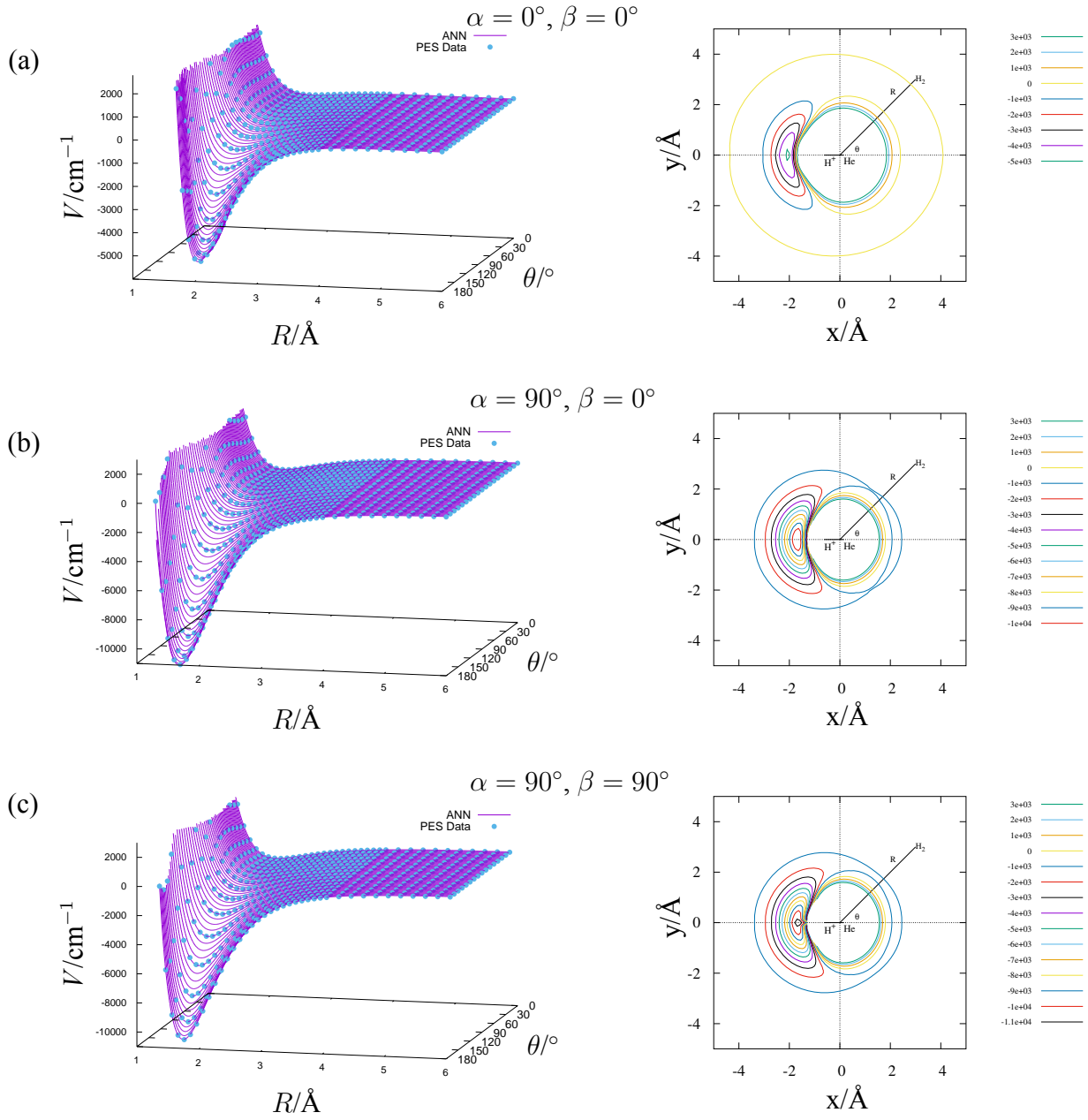


Figure 2.7: Plot of the PES (left) obtained from ANN predictions and ab initio data (blue dots) plotted against  $R, \theta$ . The respective potential energy contour plots (right), for different sets of  $\alpha$  and  $\beta$ .

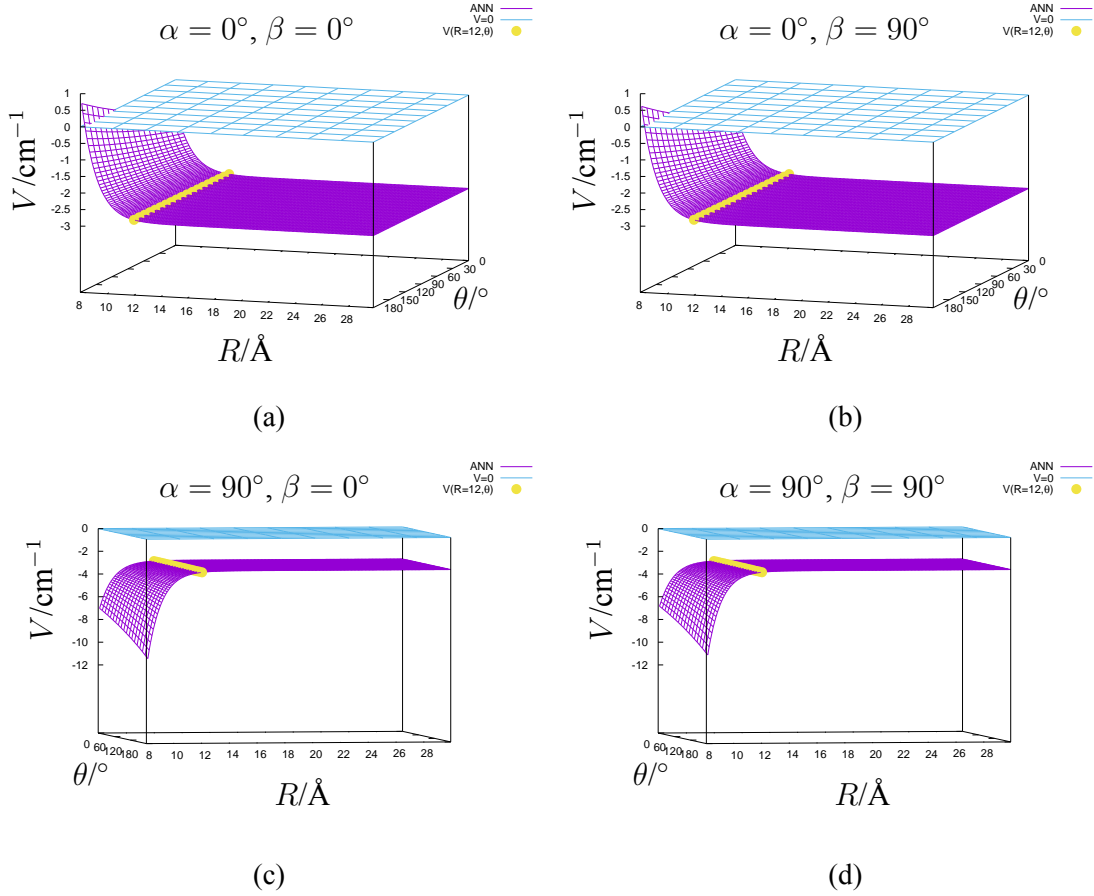


Figure 2.8: Plot of the PES obtained from ANN predictions plotted against  $R, \theta$  for long range ( $R$  up to 30 Å). The yellow points show  $V(R)$  for the upper limit of  $R$  up to which data is available ( $R \leq 12$  Å).

### 2.1.4 Extension to Long Range

The fit was tested if it could be used to extrapolate at the long range (upto 30 Å). The long range PES plotted against  $R$  and  $\theta$  for different values of  $\alpha$  and  $\beta$  are shown in Figure 2.8.

As Neural networks are known to be poor at extrapolating beyond the range of the training data, we can see that the ANN fit dies off to a value of  $-2.822 \text{ cm}^{-1}$  at large  $R$  values, hence another approach is required to include the long range part of the PES. Since the input data for the 4D PES was available for  $R \leq 12$  Å, we used the asymptotically correct long range potential,<sup>39</sup>  $V_L$  for larger  $R$  values:

$$V_L = -\alpha_0/(2R^4) + (2\alpha_0\mu/R^5) \cos \theta, \quad (2.6)$$

The first term arises due to the charge of  $\text{HeH}^+$  interacting with the polarizability of  $\text{H}_2$  molecule, whereas the second term arises due to the dipole-polarizability interaction term, where  $\alpha_0 = (\alpha_{\parallel} + 2\alpha_{\perp})/3$  and  $\alpha_2 = \alpha_{\parallel} - \alpha_{\perp}$ . Using the results of  $\alpha_{\parallel} = 6.38049$  and  $\alpha_{\perp} = 4.57769$  au for  $r = 1.4$  au ( $0.7408 \text{ \AA}$ ) of  $\text{H}_2$  reported by Kolos et al,<sup>40</sup>  $\alpha_0 = 5.1786$  au. It is worth re-iterating that  $\theta = 0$  corresponds to the He end of  $\text{HeH}^+$  and  $\theta = 180^\circ$  to the H end.  $r(\text{HeH}^+) = 0.774 \text{ \AA}$ ,  $r(\text{H-H}) = 0.744155 \text{ \AA}$ . The final long range behavior of the PES is shown in Figure 2.9.

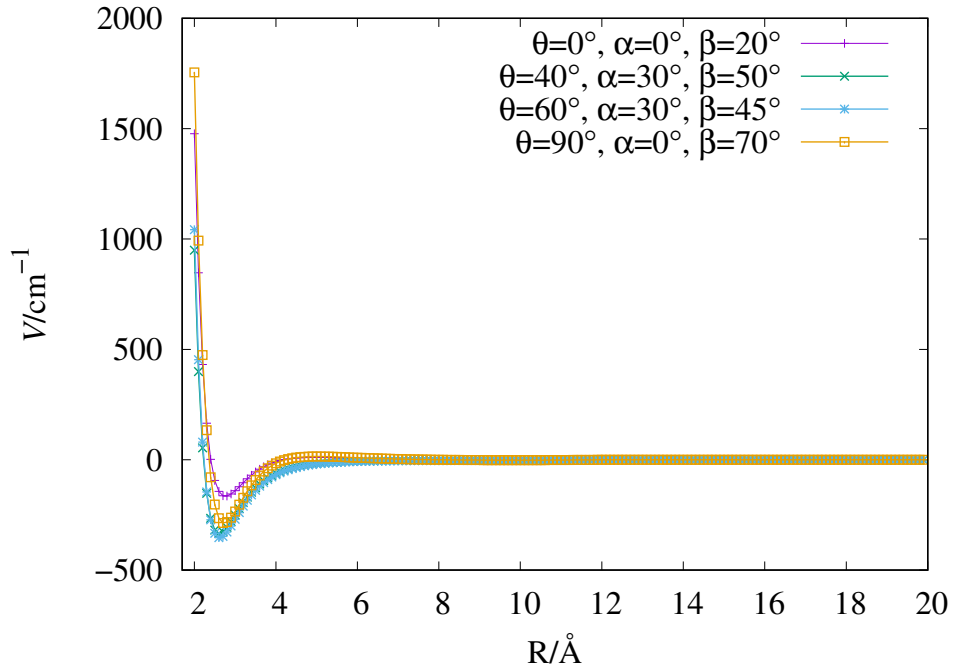


Figure 2.9: One dimensional slices of the PES for long range potential for different combinations of  $\theta$ ,  $\alpha$  and  $\beta$ .

MATLAB was used to produce a standalone C++ code for the ANN with Random (98%) sampling (Set 6), which contained all the optimized weights and biases. It was further converted to FORTRAN programming language, so that it can be used with quantum scattering packages, such as MOLSCAT.<sup>41</sup> The long range part was also added to the FORTRAN code with a switching function.

### 2.1.5 Deep Neural Network

The results of the different fits undertaken are summarized in Table 2.2 and compared with that of the single-layer 60 node fit. We can see that expanding the network architecture resulted in better fits. For the test data, the RMSD of the (20, 20) network was  $82 \text{ cm}^{-1}$  which improved to  $7.7 \text{ cm}^{-1}$  for the (30, 30) neuronal fit. A further increase in the network size to (35, 35) resulted in a similar fit with an RMSD of  $8.0 \text{ cm}^{-1}$ . The (40, 40) network gave an improved fit with an RMSD of  $6.4 \text{ cm}^{-1}$  for the test data and  $4.95 \text{ cm}^{-1}$  for the entire data set. Increasing the nodes to 50 in each layer gave a low RMSD of  $1.0 \text{ cm}^{-1}$  for the training data, however, a large RMSD of  $582 \text{ cm}^{-1}$  for the test data indicating over-fitting. To check if the fit can further be improved, we chose a three-hidden layer network. From Table ??, we can see that increasing the number of hidden layers did not improve the fit with large RMSD values for the test data again indicating over-fitting. Thus, for the  $\text{HeH}^+\text{-H}_2$  system, the (40, 40) network size seems to be optimal and have been chosen as the best fit to represent the PES. The RMSD for this fit was  $4.9 \text{ cm}^{-1}$ . A slice of the PES is shown in Figure 2.10, and the residual for this fit is shown in Figure 2.11. The maximum error for the fit was found to be  $164.843 \text{ cm}^{-1}$  at  $\theta = 170^\circ$ ,  $\alpha = 120^\circ$ ,  $\beta = 60^\circ$ , and  $R = 1.4 \text{ \AA}$  for  $V = 2799.66 \text{ cm}^{-1}$ .

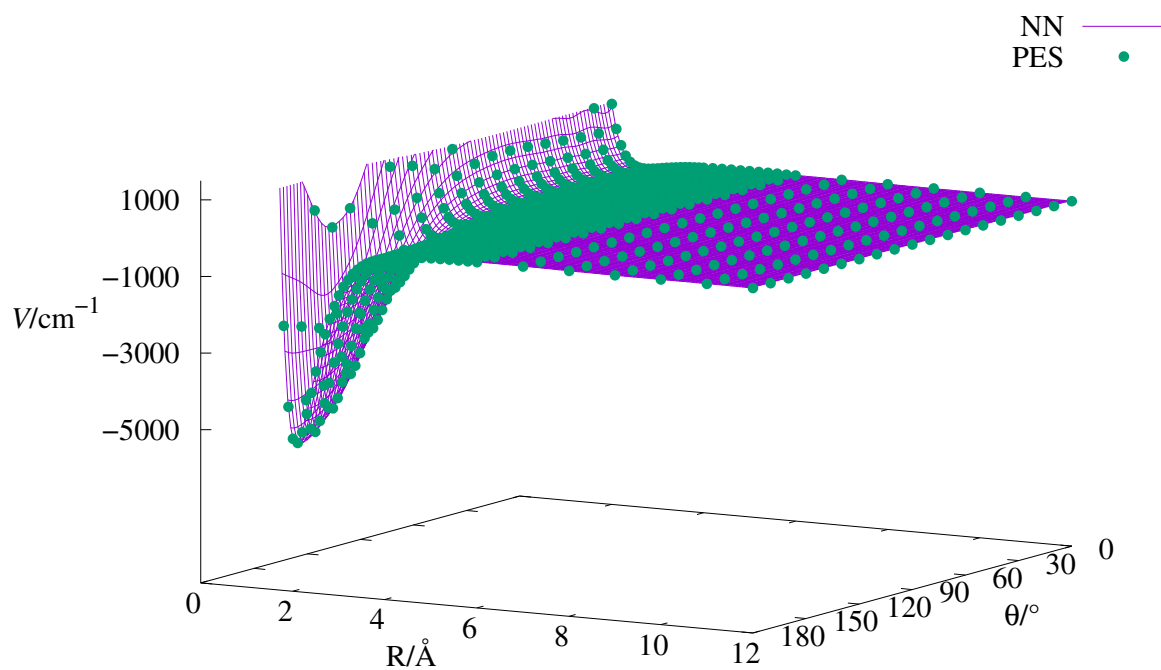


Figure 2.10: Slice of the PES for  $\alpha = 0^\circ$  and  $\beta = 0^\circ$ .

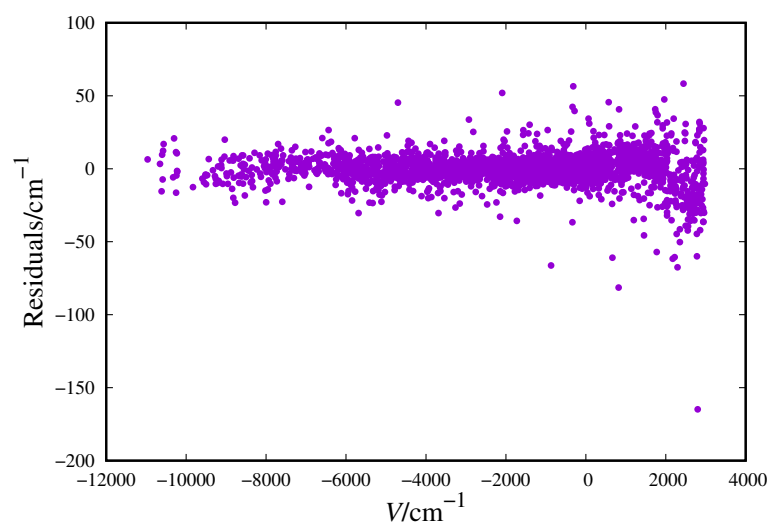


Figure 2.11: Residuals for deep neural network representation of the PES.

Table 2.2: Comparison of fits obtained using different networks

Network	NTP <sup>a</sup>	RMSD (cm <sup>-1</sup> )		ME <sup>b</sup> (cm <sup>-1</sup> )
		Training	Test	
60	15610	35.2	43.7	-934
(20, 20)	14335	18.8	81.9	1478
(30, 30)	14335	7.6	7.7	166
(35, 35)	14335	6.1	8.0	170
(40, 40)	14335	4.8	6.4	-165
(50, 50)	14335	1.0	581.8	-19 488
(30, 30, 10)	14335	22.0	40.9	1386
(30, 30, 20)	14335	23.6	92.0	656
(30, 30, 30)	14335	19.3	56.4	1199
(30, 40, 30)	14335	6.6	421.0	1113

<sup>a</sup> Number of training points <sup>b</sup> Maximum error

As can be seen from the fits and the previous section that a deep neural network gives a much more accurate PES for this system. However, further detailed checks show that salient features of the excitation function for different inelastic processes remain unaltered in going from shallow to deep neural network, although there are quantitative differences between the state-to-state integral cross section data on the two fitted surfaces (not reported here).

## 2.2 $\text{C}_5\text{N}^- + \text{H}_2$ *ab initio* PES

Dalgarno and McCray<sup>42</sup> first pointed out the importance of negative ion chemistry in the ISM. They hypothesised that in interstellar clouds which have a source of ionization, negative ions may play an important role in forming complex molecules. In recent years many C-bearing and (C,N)-bearing chains of molecular anions have been observed at various sites in the ISM.<sup>43–47</sup> The molecular anions detected so far in the interstellar and circumstellar gas are all fairly heavy linear carbon chains consisting of three or more carbon atoms, and with neutral counterparts with large electron affinities:  $\text{C}_4\text{H}^-$ ,<sup>45</sup>  $\text{C}_6\text{H}^-$ ,<sup>47</sup>  $\text{C}_8\text{H}^-$ ,<sup>46</sup>  $\text{C}_3\text{N}^-$ ,<sup>43</sup> and  $\text{C}_5\text{N}^-$ .<sup>43</sup> Cernicharo et al first detected  $\text{C}_3\text{N}^-$  and  $\text{C}_5\text{N}^-$  nitrile anions towards the cold dark core TMC-1 (Taurus molecular cloud), using the Yebes 40 m telescope. The Taurus region is a stellar nursery 430 ly away from earth, containing hundreds of newly formed stars. The abundance of these anions relative to the neutral counterparts increases with both size and the electron affinity of the neutral molecule, as expected for formation by Radiative Electron Attachment (REA).<sup>48</sup> However, this process fails to explain the abundance of the shortest observed anions. While a more direct chemical route by reaction of the  $\text{HC}_5\text{N}$  with  $\text{H}^-$  has also been put forward<sup>49</sup> since the calculated rates were found to be large enough to be relevant within the chemical evolution producing these anions.

In this work, we present the *ab initio* potential energy data, generated by our collaborator Ersin Yurtsever, for the surface describing the interaction of the linear anion ( $\text{C}_5\text{N}^- +$ ) with  $\text{H}_2$ . We employ a deep ANN fit for representing the PES. The fitted PES, was further used to carry out quantum scattering calculations, by our collaborators, Kaushik Giri, Lola-González Sánchez and Anzhela Veselinova Marinova, to yield rotationally inelastic cross sections for collisions with  $\text{H}_2$  and obtain the corresponding inelastic rate coefficients as a function of temperature.

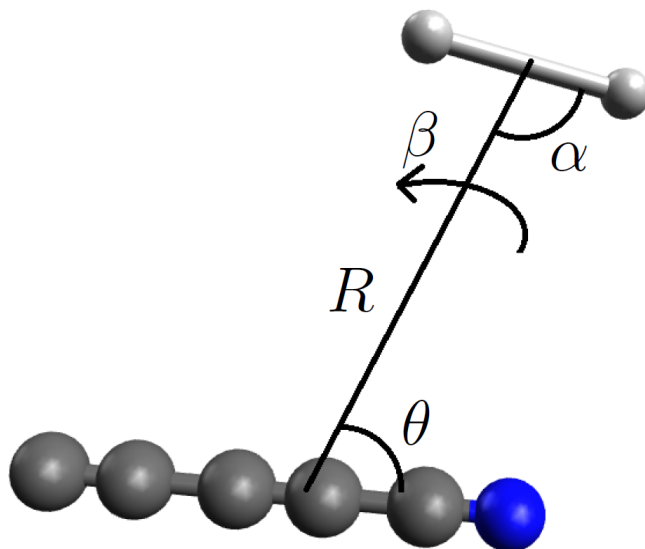


Figure 2.12: The four variables used to generate the PES for the  $C_5N^- + H_2$  system.

### 2.2.1 Data Set

The four dimensional coordinate system is given in Figure 2.12. Both  $C_5N^-$  anion and  $H_2$  considered to be rigid. MOLPRO<sup>50</sup> was used to generate the ab initio data points at the CCSD(T)-F12b/aug-cc-pVTZ level of theory. Furthermore, they are corrected for basis set superposition error (BSSE). 16025 points were generated by varying the four variables  $\theta = 0^\circ$  to  $\theta = 180^\circ$ ,  $\alpha = 0^\circ$  to  $\alpha = 150^\circ$ ,  $\beta = 0^\circ$  to  $\beta = 90^\circ$  and  $R = 2 \text{ \AA}$  to  $R = 25 \text{ \AA}$ . The electronic energies were calculated by our collaborator Ersin Yurtsever.

The ANN representation of the PES requires smooth data over the range of variables. In order to generate smooth data of the PES, the number of data points were increased (see Table 2.3).

For uniform distribution of data across various values of  $\alpha$  and  $\beta$ , additional data points

---

This is a collaborative work, electronic structure calculations were done by E. Yurtsever. R. Biswas and U. Lourderaj were responsible for ANN fitting and the FORTRAN subroutine, K. Giri, L. Gonzalez and A. Veselinova were responsible for quantum scattering calculations, and analysis was done by F. A. Gianturco and N. Sathyamurthy.

(around 4000) were added. For  $\alpha = 180^\circ$  and  $\beta = \{30^\circ, 60^\circ, 90^\circ\}$ ,  $V(\beta = 0^\circ, \alpha = 0^\circ)$  energy values were used. Using the symmetry of the system, additional data was added considering the parameter  $\beta$ . For example potential for  $\beta = 60^\circ$  is the same as the potential for  $\beta = 120^\circ$ . The density of points at high  $R$  values ( $R > 10\text{\AA}$ ) was less and this region is prone to overfitting. For this reason we used one dimensional cubic spline method to interpolate potentials at larger  $R$ s (see Figure 2.13) and added around 15000 points. We had a total of 761 different combinations of  $\theta$ ,  $\alpha$  and  $\beta$ , and for each combination we select 7 points which has  $R$  value greater than or equal to  $8\text{\AA}$ . We define a cubic polynomial between any two consecutive points (6 polynomials in total), defined by 4 coefficients each. Each of these polynomials are further subjected to 4 conditions, the first being the requirement that the polynomials should pass through the known points. Secondly we impose that the tangents at the known points of any two consecutive polynomial, should be the same. Additional two conditions are then put, which are that at the beginning and the end of the data set the slope goes to zero. With these conditions in place, it is possible to find out all the coefficients for the 6 polynomials. In this scheme, we have three points before  $R = 11\text{\AA}$ , ensuring that the interpolation between  $R = 10\text{\AA}$  and  $R = 12\text{\AA}$ , are correct. All the cubic splines were computed using MATLAB.<sup>38</sup> Lastly, we added data points beyond the limits of  $\theta$  variable, for example  $V(\theta = 200^\circ)$  is equal to  $V(\theta = 160^\circ)$ . This was done so that the ANN fit behaves smoothly at the limits of  $\theta$ . These points will not be used in predictions, they will only be used for training the ANN. The domain of  $\theta$  remains between  $0^\circ$  to  $180^\circ$ .

Table 2.3: This table summarizes the different data sets generated and used to do the fitting.

Data Set	Description	Number of points
1	Original set	16025
1a	Set 1 + 4000	20000
2	Symmetry $\beta$	29897
3	Set 2 + cubic splines	46064
4	Set 3 + Boundary	58448
5	Set 4 ( $V < 2000/\text{cm}^{-1}$ )	52734
6	Set 5 (w.o. boundary)	44915

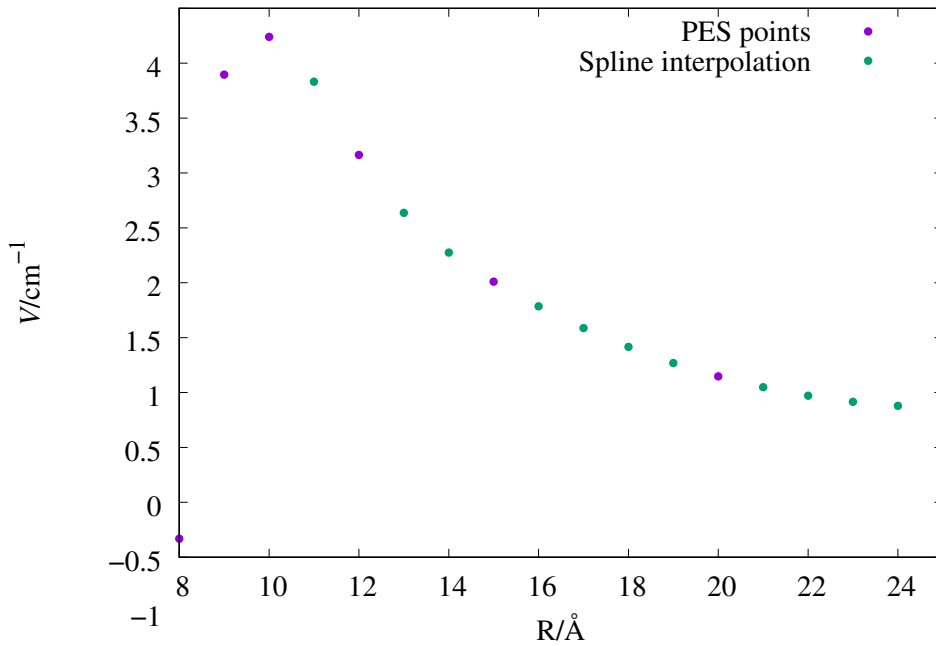


Figure 2.13: Example of cubic spline interpolation for  $\theta = 10^\circ$ ,  $\alpha = 90^\circ$  and  $\beta = 90^\circ$ .

The data Set 5 was used fit the PES using ANN for potential energies less than  $2000 \text{ cm}^{-1}$ . Set 6 was used to test the performance of the neural networks. For the transfer function we use a modified version of the logistic sigmoid function given by Eqn 2.5. The Bayesian Regularization<sup>51</sup> optimizer is used to train the network as implemented in MATLAB.<sup>38</sup>

### 2.2.2 Shallow Neural Network

We started with a shallow neural network with 60 neurons and increased the number of neurons in the hidden layer upto 300. We did multiple fits and checked the performance of the fit. The best fit we obtained used a network consisting of 300 neurons in one layer, and had an RMSD of  $9.27 \text{ cm}^{-1}$  with maximum error of  $176 \text{ cm}^{-1}$  at  $\theta = 160^\circ$ ,  $\alpha = 150^\circ$ ,  $\beta = 90^\circ$  and  $R = 5.1 \text{ \AA}$  for  $V = 1966.88 \text{ cm}^{-1}$  as shown in Figure 2.14. The maximum well depth was  $600 \text{ cm}^{-1}$  compared to  $12000 \text{ cm}^{-1}$  for the  $\text{HeH}^+ + \text{H}_2$  system described in the previous section, coupled with the fact that the rotational constant of  $\text{C}_5\text{N}^-$  is fairly small ( $0.04633 \text{ cm}^{-1}$ ), we can say that our fit quality was poor. So we switched to a deep neural network.

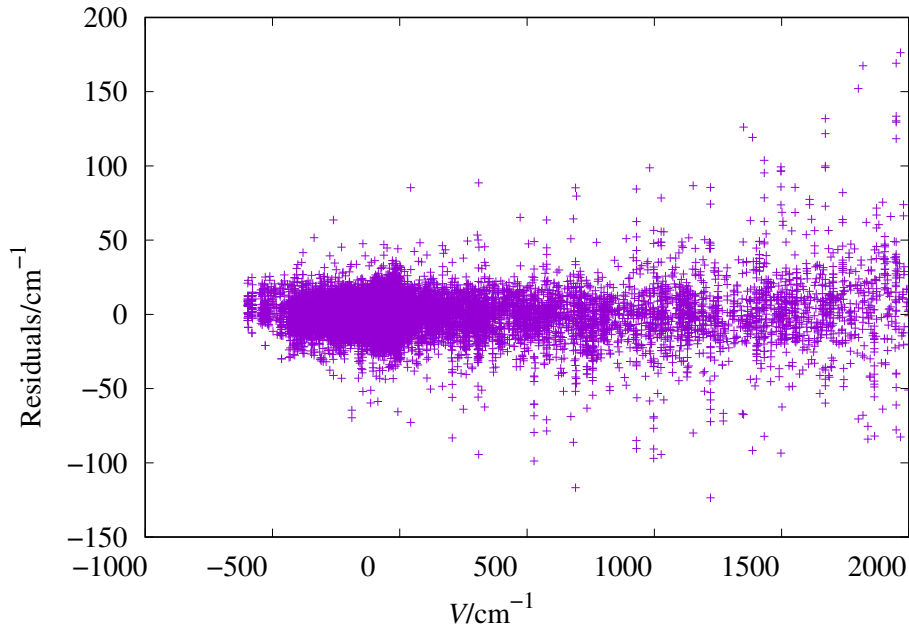


Figure 2.14: Plot of residuals for the PES fitted using a shallow neural network for  $\text{C}_5\text{N}^- + \text{H}_2$ .

Table 2.4: Training and network performance for different networks when 100% data was used.

Fit	$h_1$	$h_2$	RMSD/cm <sup>-1</sup>	Max Error/cm <sup>-1</sup>
1	30	30	5.09	101.8 at $V = 1950$
2	30	30	2.99	52.42 at $V = 1671$
3	30	30	1.03	35.3 at $V = 778$
4	30	30	0.87	38 at $V = 778$
5	35	35	0.68	35.7 at $V = 778$

Table 2.5: Training and network performance for different networks when 70% data was used.

Fit	$h_1$	$h_2$	RMSD	RMSD <sub>30</sub>	Max <sub>All</sub>	Max <sub>30</sub>
6	35	35	2.4	2.59	42 at $V = 778$	-36 at $V = 1687$
7	35	35	1.51	1.61	38 at $V = 778$	-23 at $V = 73$
8	30	30	1.5	1.56	45 at $V = 778$	-24 at $V = 1980$

### 2.2.3 Deep Neural Network

We used networks consisting of two hidden layers ( $h_1$  and  $h_2$ ) and varied the number of nodes in each hidden layer. Figure 2.15 shows the schematics of the network. First we use 100% of the data to train the network. The performance of the network is given in Table 2.4. Then we tried using only 70% of the data (from Set 5) and the performances of which are summarized in Table 2.5. RMSD<sub>30</sub> is the root mean square deviation of the data points (from Set 6) which are not included in the training data, similarly Max<sub>30</sub> is the maximum deviation for the same. After this we also tried using LHS<sup>37</sup> in the four input dimensions to train the network using less number of data points. The results are summarized in Table 2.6. LHS did not yield a good fit as the maximum error was -624.1 cm<sup>-1</sup>, so we chose to work with Fit 4 which had an RMSD of 0.87 cm<sup>-1</sup> and was better in terms of fitting (no overfitting) in comparison to Fit 5. The schematics of the ANN for Fit 4 are given in Figure 2.15 and the residual plot is given in Figure 2.16.

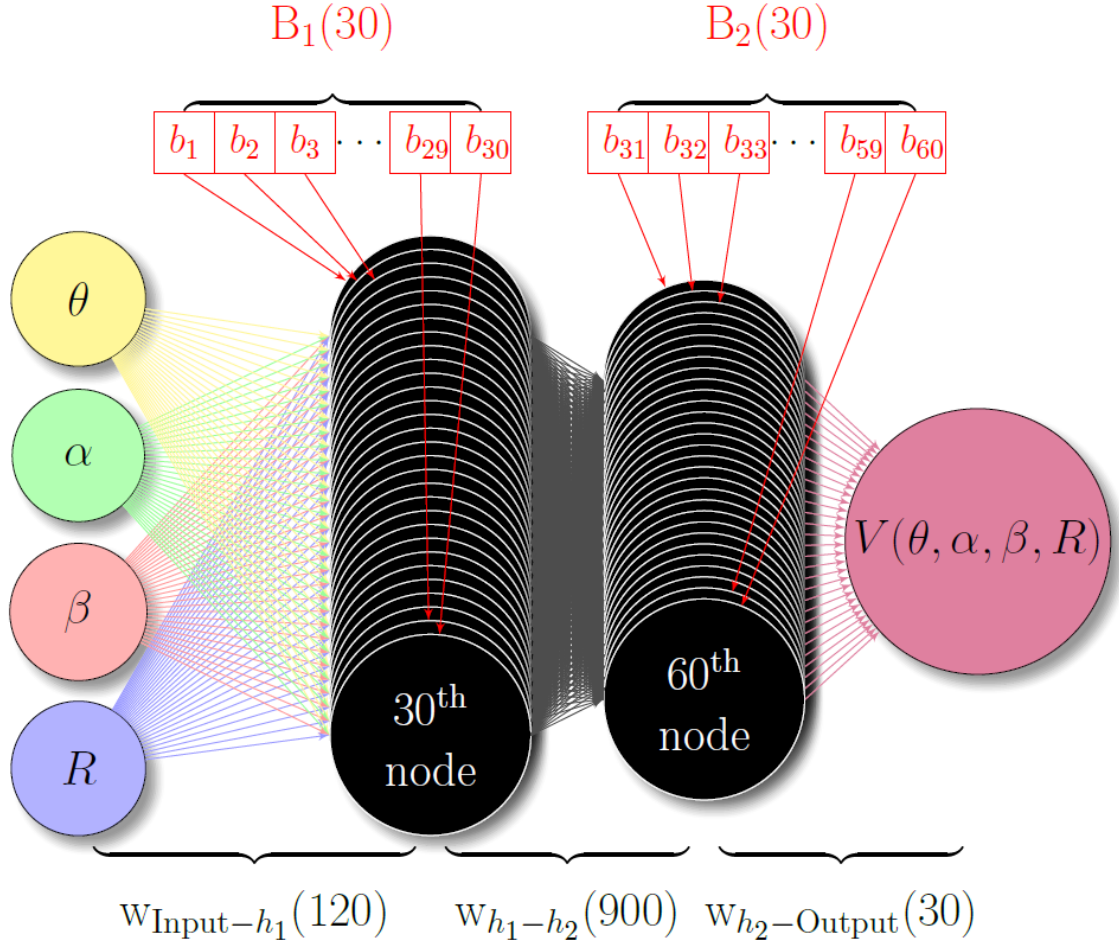


Figure 2.15: Neural Network used to train  $C_5N^- + H_2$  PES. Weights are in black, and biases in red. Numbers inside the bracket are the corresponding number of parameters.

Table 2.6: Training and network performance for different networks for LHS sampling. Brackets show % of data used for training of these networks.

Fit	Number of points (%)	$h_1$	$h_2$	RMSD/cm $^{-1}$	Max Error/cm $^{-1}$
LHS1	1000(1.9)	35	35	148.367	1159.91 at $V = -299.793$
LHS2	1334(2.5)	35	35	129.317	-1705.48 at $V = 260.675$
LHS3	1667(3.2)	35	35	116.17	1321.86 at $V = -545.714$
LHS4	2000(3.8)	35	35	58.2475	706.121 at $V = -120.537$
LHS5	3334(6.3)	35	35	86.0675	1145.16 at $V = -578.757$
LHS6	7717(14.6)	30	30	15.4669	536.445 at $V = 1400.92$
LHS7	7717(14.6)	35	35	11.54	-624.175 at $V = 1950.47$

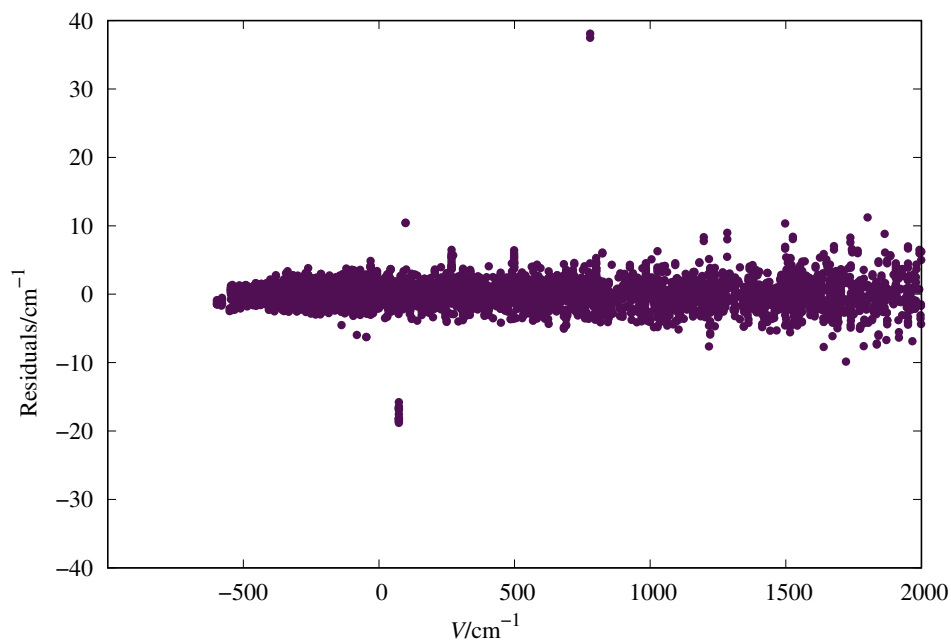


Figure 2.16: Plot for residuals for Fit 4.

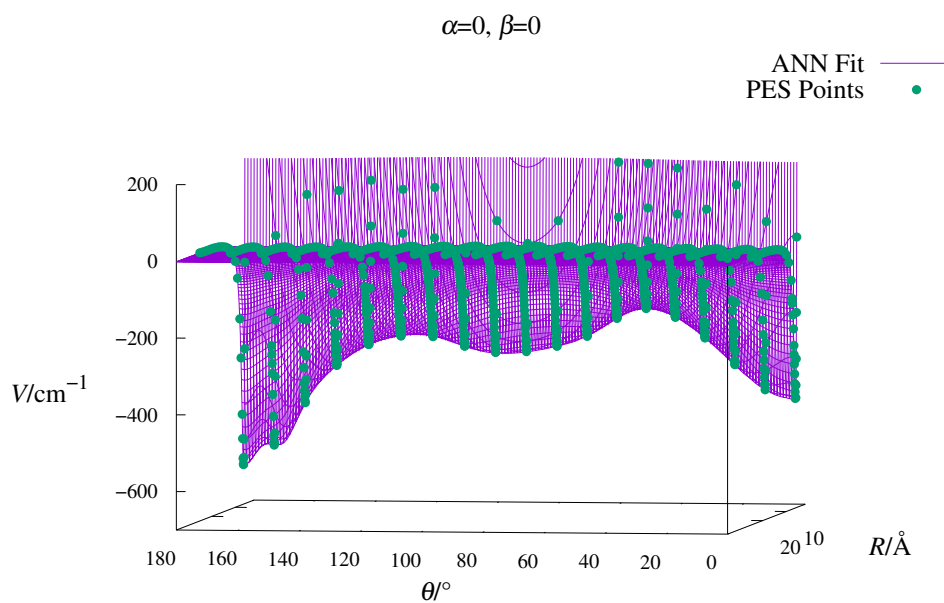


Figure 2.17: Plot for Fit 4 compared to data points, for  $\alpha = 0^\circ$  and  $\beta = 0^\circ$ .

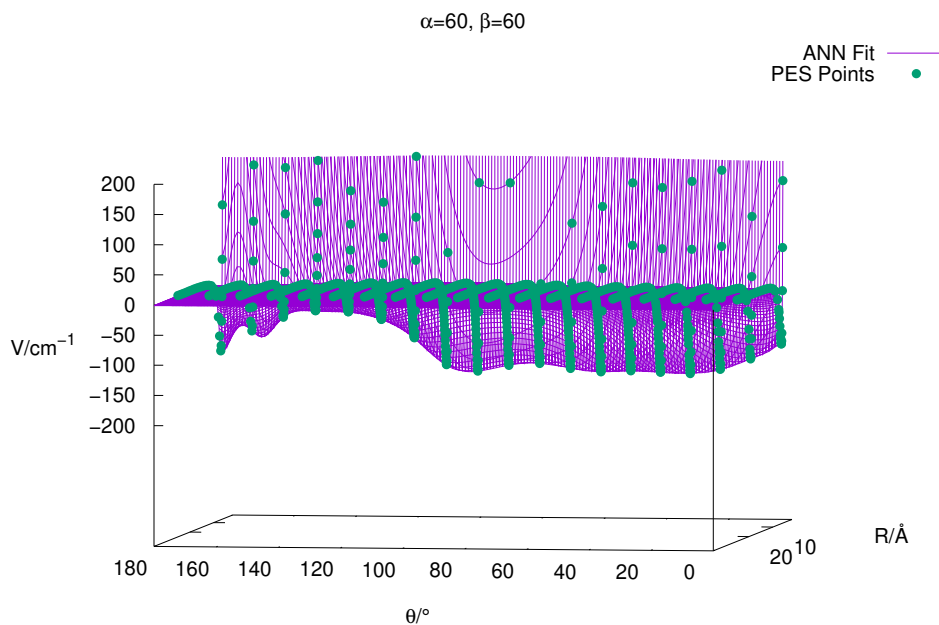


Figure 2.18: Plot for Fit 4 compared to data points, for  $\alpha = 60^\circ$  and  $\beta = 60^\circ$ .

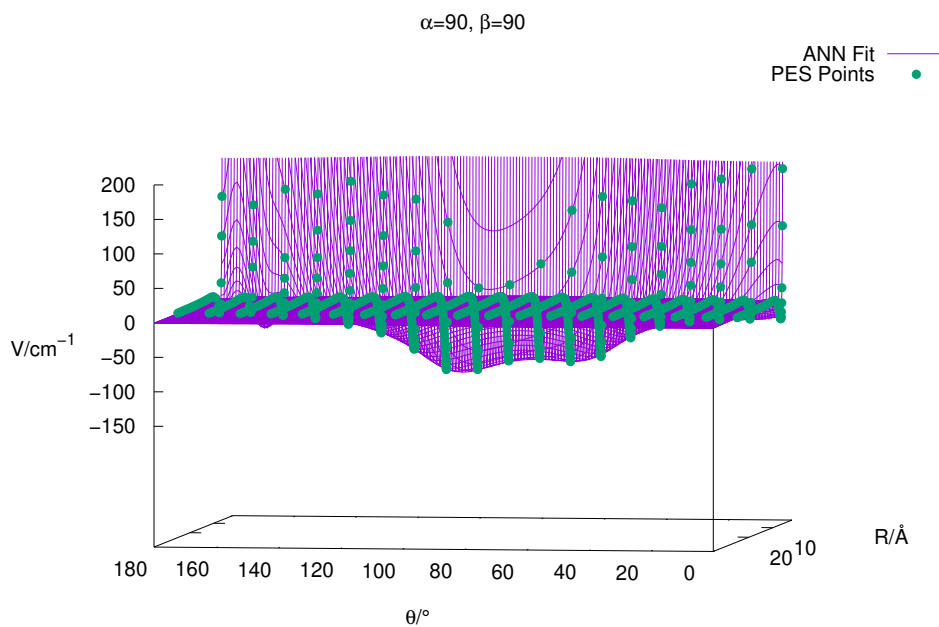


Figure 2.19: Plot for Fit 4 compared to data points, for  $\alpha = 90^\circ$  and  $\beta = 90^\circ$ .

### 2.2.4 Extension to Long Range

The asymptotically correct long range potential is given by,

$$V_L(\theta) = -\frac{a_0}{2R^4} + \frac{2a_0\mu}{R^5} \cos(\theta) \quad (2.7)$$

where  $a_0 = (a_\perp + 2a_\parallel)/3$  and  $a_2 = a_\parallel - a_\perp$ . Using the results of  $a_\parallel = 6.38049$  au and  $a_\perp = 4.57769$  au for  $r = 1.4$  au (0.7408 Å) of H<sub>2</sub> reported by Kolos and Wolniewicz,<sup>40</sup>  $a_0 = 5.7186$  au. The relaxed dipole moment of C<sub>5</sub>N anion is  $\mu = 3.31370$  au.

The final form of the potential using the switching function is given by,

$$V_f = f_s V_{\text{ANN}} + (1 - f_s) V_L \quad (2.8)$$

where the switching function is,

$$f_s(R) = \frac{1}{e^{\frac{(R-R_0)}{\Delta R}} + 1} \quad (2.9)$$

and  $R_0 = 22$  Å and  $\Delta R = 0.5$  Å.

### 2.2.5 FORTRAN code for ANN

The final fit was converted to C++ code using MATLAB coder, and was further converted to a FORTRAN subroutine. The asymptotically correct long range potential with a switching function was also added to the FORTRAN code.

## 2.3 C<sub>2</sub><sup>-</sup> + H<sub>2</sub> *ab initio* PES

Controlling quantum states of atoms and molecules have potential applications in many fields of science, such as quantum information, state specific chemistry or even studying important processes found in the ISM. Precise control of molecules started with the invention of supersonic molecular beams for cooling<sup>52</sup> and coherent control for manipulation of internal states.<sup>53</sup> However, cooling molecular gases where quantum phenomena can be

observed is relatively recent development.<sup>54</sup> Cooling and controlling quantum states of molecular ions<sup>55–57</sup> has opened up a large field of research on precise studies of elastic and inelastic collision processes,<sup>58–60</sup> controlled chemical reactions,<sup>61,62</sup> as well as on precision spectroscopy and measurements of fundamental constants.<sup>63–65</sup>

Research of negative ions gained interest due to their proposed application to sympathetically cool antiprotons,<sup>66</sup> hence providing a new method of generating cold antihydrogen<sup>67</sup> in large abundance. However, the loosely bound excess electron in an anion makes it a more fragile system in which multiple electronic states are rarely encountered unless in the form of a dipole bound state close to the detachment threshold. Only few atomic negative ions are known to possess valence excited electronic states. To prepare for negative ion laser cooling these anions have been explored in high-resolution spectroscopy.<sup>68–70</sup> An interesting alternative are small molecular anions, in particular the carbon dimer anion  $C_2^-$ .<sup>71</sup>

Recent experiments conducted by Wester et al.,<sup>72</sup> demonstrate optical pumping of the  $C_2^-$  anion into its first vibrational level  $v = 1$  with high efficiency. They also measured the absolute rate coefficient for collisional quenching back into the ground vibrational level  $v = 0$  in collisions with molecular hydrogen at a temperature of 18 K. Further, the measured rate is compared with theoretical calculations. The theoretical curves obtained by reduced-dimensional quantum scattering calculations carried out using the coupled channel method for scattering of an atom with a diatomic molecule as tests of the *ab initio* calculations have shown that they are not expected to change significantly with a larger basis set or an explicitly correlated method of computation. The experimentally observed quenching rate is an order of magnitude smaller than those obtained by theory, which is attributed to the fact that the six internal degrees of freedom in the  $C_2^- + H_2$  system had to be reduced to three to perform the calculation with ASPIN.<sup>73</sup>

To this end, in the work reported here we represent the five dimensional PES of  $C_2^- + H_2$ , where the  $H_2$  is fixed at its equilibrium geometry, using ANN.

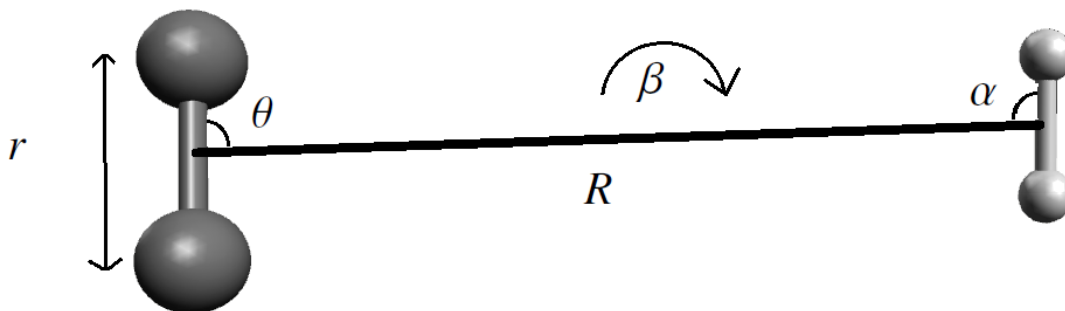


Figure 2.20: The five variables defining the  $\text{C}_2^- + \text{H}_2$  system.

### 2.3.1 Data Set

The five dimensional coordinate system is given in Figure 2.20.  $\text{H}_2$  molecule is considered to be rigid. MOLPRO<sup>50</sup> was used to generate the *ab initio* data points at the CCSD(T)-F12b/aug-cc-pVTZ level of theory. 68011 points were generated by varying the five variables  $\theta = 0^\circ$  to  $\theta = 90^\circ$ ,  $\alpha = 0^\circ$  to  $\alpha = 180^\circ$ ,  $\beta = 0^\circ$  to  $\beta = 90^\circ$ ,  $R = 3.5 \text{ \AA}$  to  $R = 25 \text{ \AA}$  and  $r = 1.18 \text{ \AA}$  to  $r = 1.4 \text{ \AA}$ . The data was generated by our collaborator Jan Franz.

### 2.3.2 Deep Neural Network

We used a deep neural network consisting of two layers as shown in Figure 2.21. We kept the number of nodes in the first layer to be 30, and varied the number of nodes in the second hidden layer from 30 to 60. For training we used 98% of the data, which was selected randomly and the rest 2% was used for testing. The performance of the fits are given in Table 2.7. The best fit (Fit 5) obtained (shown in red in Table 2.7) had an RMSD of  $0.829 \text{ cm}^{-1}$ , and the maximum error of  $36.20 \text{ cm}^{-1}$  at  $R = 3.6 \text{ \AA}$ ,  $r = 1.4 \text{ \AA}$ ,  $\theta = 0^\circ$ ,  $\alpha = 150^\circ$ ,  $\beta = 30^\circ$  for  $V=4123.57 \text{ cm}^{-1}$ . The residuals for this fit are shown in Figure 2.22 and the quality of fit are shown for two slices in Figures 2.23 and 2.24.

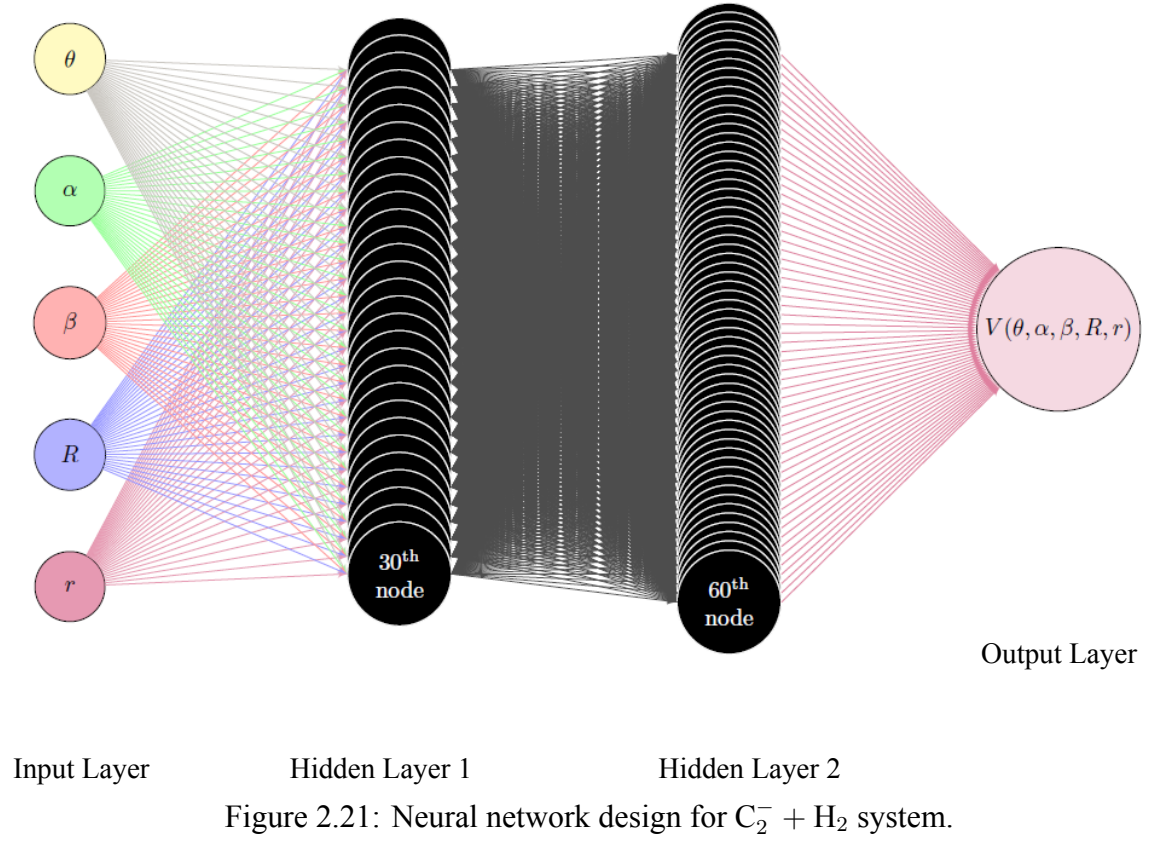


Table 2.7: Table of different fits for the  $C_2^- + H_2$  system. Best fit is given in red. The numbers under the column “Network” indicate the number of nodes in each hidden layer.

Fit	Network	$RMSD_{\text{training}} \text{ (cm}^{-1}\text{)}$	$RMSD_{\text{test}} \text{ (cm}^{-1}\text{)}$	$RMSD_{\text{total}} \text{ (cm}^{-1}\text{)}$
1	30,40	0.98	1.2	0.99
2	30,50	1.13	1.3	1.14
3	30,60	0.96	1.2	0.96
4	30,60	0.86	1.19	0.87
5	30,60	0.82	1.04	0.83

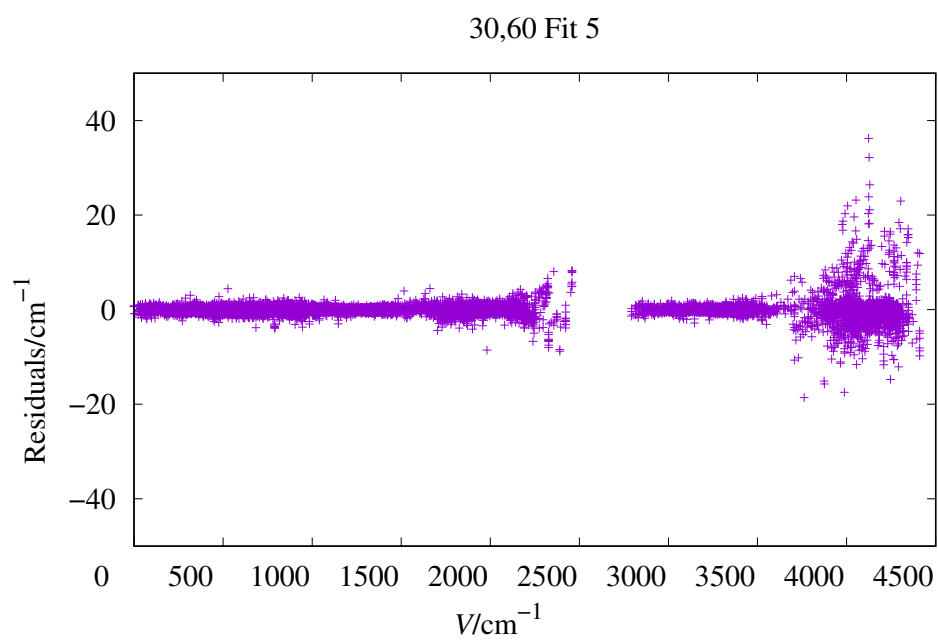


Figure 2.22: Plot for residuals for the best fit, Fit 5

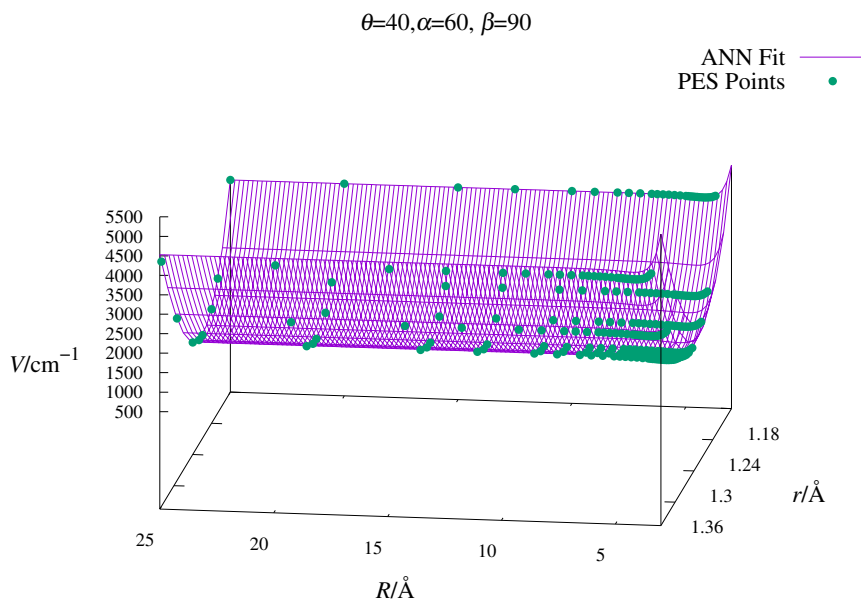


Figure 2.23: Slice of the potential energy surface for  $\theta = 40$ ,  $\alpha = 60$  and  $\beta = 90$ .

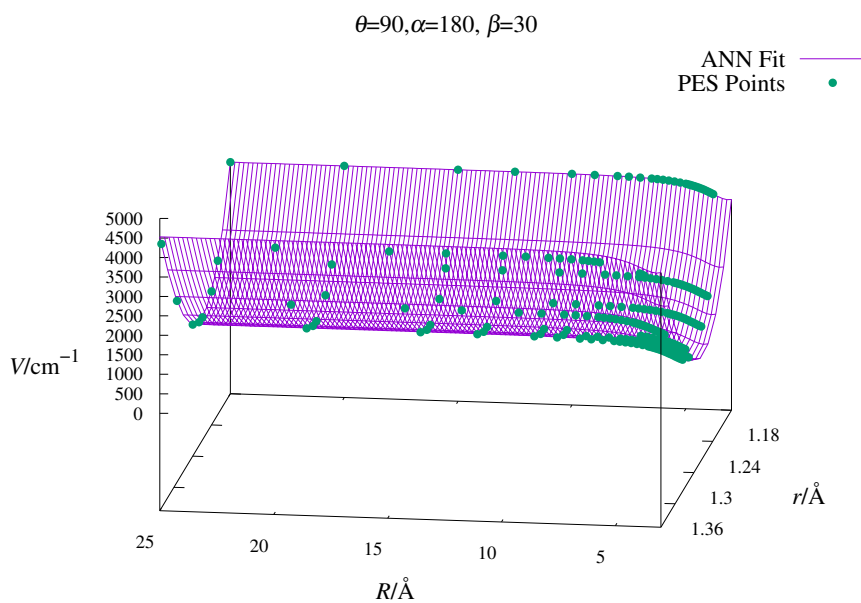


Figure 2.24: Slice of the potential energy surface for  $\theta = 90$ ,  $\alpha = 180$  and  $\beta = 30$ .

## **2.4 Summary and Conclusions**

We have successfully represented the PESs of two 4-dimensional systems,  $\text{HeH}^+ + \text{H}_2$  and  $\text{C}_5\text{N}^- + \text{H}_2$  and a 5-dimensional system  $\text{C}_2^- + \text{H}_2$ , that are important molecules in the interstellar medium. The representation was done using ML methods, GPR and ANN. We found that for low number of points GPR tends to work better, however it becomes difficult to represent the PES using GPR when higher number of data points are used. On the contrary, ANN works better with higher number of points. Moreover, we found that deep neural networks perform better over shallow neural networks. All the representations of the PESs are converted to FORTRAN code and can be merged with quantum dynamics software like MOLSCAT.

In conclusion we find that ML method, ANN, works well for four and five dimensional systems. Yet the correct network size for any given data set needs to be optimized. From our work we find that it is better to start with a smaller network (in terms of number of nodes per layer), and increase the size incrementally. Smaller networks will not be able to represent the data set, yielding larger RMSDs. We increase the network size to the point where the network overfits the data. The optimal network size is often the network used before the overfitting happens.

## Chapter 3

# “On-the-fly” Representation of Potential Energy Surfaces using Machine Learning Methods

Potential energy surfaces are fundamental to understand the dynamics of chemical reactions. In direct dynamics, the classical trajectories are calculated using the forces computed “on-the-fly” at a given level of theory.<sup>74–84</sup> This step becomes increasingly CPU intensive with increasing system size. The aim of this project is to store the force and potential energy data generated during *ab initio* dynamics and use them to predict forces for future trajectories. Using the data obtained from the trajectories, the PES is represented using ML methods. The PES is then improved with more data available from new trajectories making it an “on-the-fly” PES.<sup>85–88</sup> The GPR,<sup>88</sup> ANN<sup>89</sup> and interpolating moving least square methods<sup>85,90</sup> have been used to represent the PES “on-the-fly”. GPR has been shown to work very well with less number of data points to represent high dimensional data,<sup>16,91,92</sup> and it has been established that the GPR method becomes more accurate with the increase in the number of data points. However, the training time and prediction time scales poorly with respect to the number of training points. This is due to the fact that the training of the GPR model, for a data set of  $N$  points the design matrix of size  $N \times N$ , has to be inverted iteratively until convergence.<sup>26,92</sup> So, one has to invoke the concept of “Active Data Selection”, where during the trajectory of a system only few crucial data points are stored. In the context of PESs, the term “crucial” has been defined in multiple ways over the years,<sup>85,90,93,94</sup> with the common aim of improving the PES representation. To this end, sampling methods based on trajectories and distance between data points for interpolating moving least-squares<sup>90</sup> and

ANN<sup>89</sup> have been proposed. Selecting new *ab initio* energy points with large deviations between different NN fits among all the points generated by running trajectories on one of the NN PESs has also been explored.<sup>95</sup> In other fields, Ramakrishnan et al<sup>96</sup> and Krause et al<sup>97</sup> have explored the idea of selecting data points based on entropy criterion and mutual information, so that the final quality of predictions improves. However, these methods become challenging for problems beyond two dimensions.

Here, we propose two “proof of concept” methods, one using gaussian processes for regression and the other using artificial neural networks, to represent the PES “on-the-fly” during direct dynamics. For the purposes of developing algorithms, *ab initio* data that are required in our methods are replaced by model potentials which have an analytical form. In what follows we present the two algorithms and discuss the results obtained on two model potential energy functions: 2D malonaldehyde potential and 6D potential of formaldehyde.

### 3.1 “On-the-fly” Representation using GPR

The first algorithm uses GPR and tries to interpolate potential and forces locally. This algorithm is divided into two parts. The first step is “active data selection” phase, which is then followed by “prediction” phase. Active data selection involves the storage of potential energy and/or force data obtained from trajectory calculated using a suitable electronic structure method. To start, the first point  $Q_1$  and the associated energy  $V_1$  of the first trajectory is stored in the set  $P$ . Then, any point ( $Q_i$ ) generated during the trajectory is stored based on a distance criteria. The euclidean norm of the all the points (in set  $P$ ) with respect to  $Q_i$  is calculated. If distance between  $Q_i$  and the rest of the points is greater than the predefined trust radius ( $s_0$ ), the point  $Q_i$  and its energy  $V_i$  is added in to the set  $P$ . This is continued for a set of pilot trajectories. After the pilot trajectories, we execute the prediction phase where we run trajectories and start predicting the potentials and forces at the

point  $Q(t_n)$  based on how far it is from all the stored data within a new trust radius ( $p_0$ ) for the prediction step. If the number of data points are high we do a local GPR to predict the potential and forces using data points within  $p_0$ . Thus, the current algorithm defines a local PES for the system at a given point during the trajectory, in contrast to “on-the-fly” PES developed by others.<sup>88</sup> The working of the algorithm is illustrated in Figure 3.1 and the algorithm is presented in Figure 3.2. In Figure 3.1 panels a) and b) show the active data selection phase whereas panels c) and d) show prediction phase. Further, we demonstrate the working of the algorithm using a 2D system.

For the two dimensional case, we use malonaldehyde potential,<sup>98</sup> which is a function of two coordinates  $(x, y)$  given by,

$$V(x, y) = 2y + y^2 + (y + 0.4x^2)x^2 \quad (3.1)$$

and the physical meaning of  $x, y$  are shown in Figure 3.3.<sup>98</sup> The  $x$  and  $y$  coordinates correspond to the coordinates of the proton involved in the transfer, with the origin corresponding to the transition state. We run trajectories for a particle of unit mass subjected to the 2D potential given by Eqn. 3.1. The initial conditions are selected such that the total energy of the particle is 30 arbitrary units, and the time step is taken to be 0.1 units. We use velocity-Verlet to integrate the trajectories. The first step is to collect data which is generated during *ab initio* dynamics. For this we run 40 trajectories of 1000 time steps, and collect data using distance criteria, i.e. we only select data which are sufficiently far from each other. When the trajectory starts, we store the first point, i.e. the internal coordinates and the potential at that point ( $\vec{r} = (x_{t=0}, y_{t=0}), V(\vec{r})$ ) into the data set. For every new point  $\vec{r}_j$  (where  $\vec{r}_j = (x_{t_n}, y_{t_n})$ ) during the trajectory, we calculate the distance of  $\vec{r}_j$  from  $\vec{r}_i$  ( $i$  index runs over all stored data) as  $d_{ij}$  given by,

$$d_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2} \quad (3.2)$$

If  $d_{ij}$ s for all elements in the stored data are greater than some trust radius (in the 2D case

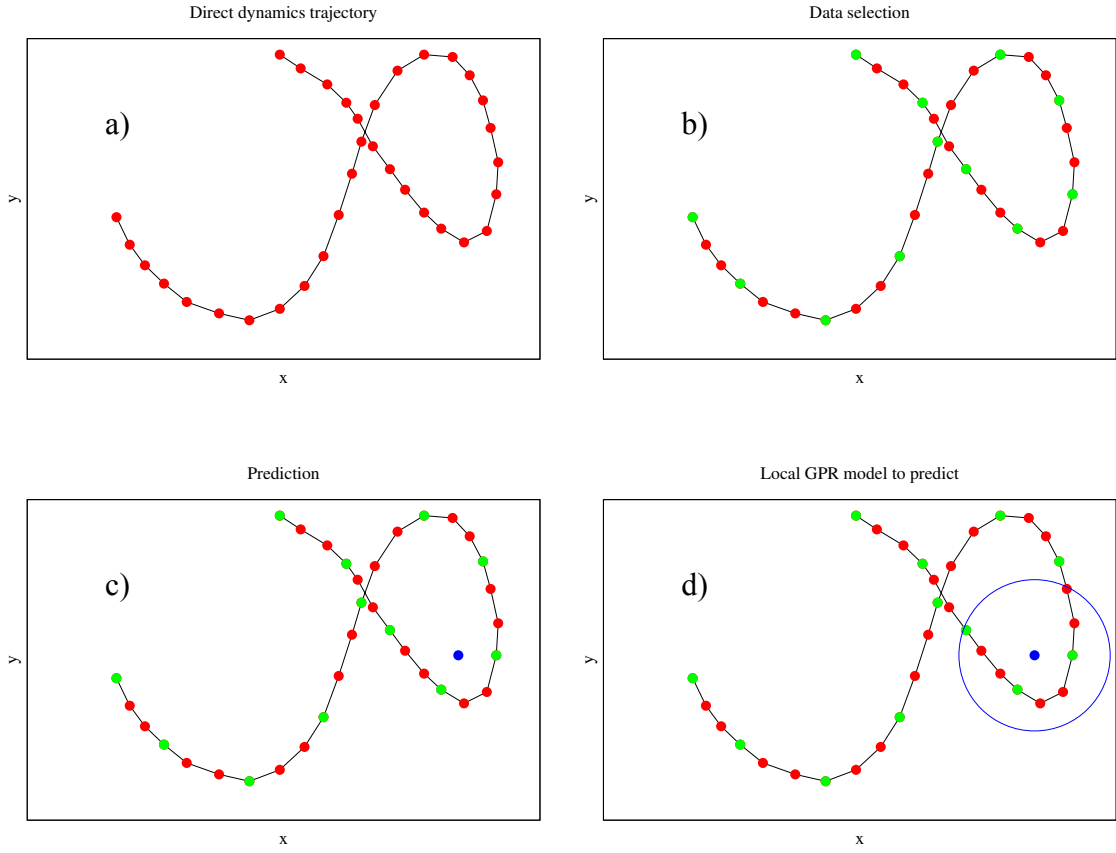


Figure 3.1: Illustration of active data selection and prediction in the “on-the-fly” local GPR PES algorithm (a) A 2D representative direct dynamics trajectory, Red dots show points where the forces and potential energy were computed and the black line shows the trajectory. (b) Green dots show the data points that have been stored using the distance criteria mentioned in the text. (c) Blue dot shows the first point of the next trajectory. (d) The blue circle is based on the trust radius for prediction ( $p_0$ ), within which three stored points lie, which will be used to predict the forces and potentials at the coordinate given by the blue dot.

presented here the trust radius was chosen to be 0.8 arbitrary units), we store the new point. If not, which signifies, we have at least one point already stored, in the vicinity of the new point so as to predict the potential energy, and hence we do not store this new point. The data collected is known as the training data. The second step is the prediction step, which starts from 41<sup>st</sup> trajectory. We calculate the number of points in the training data which are

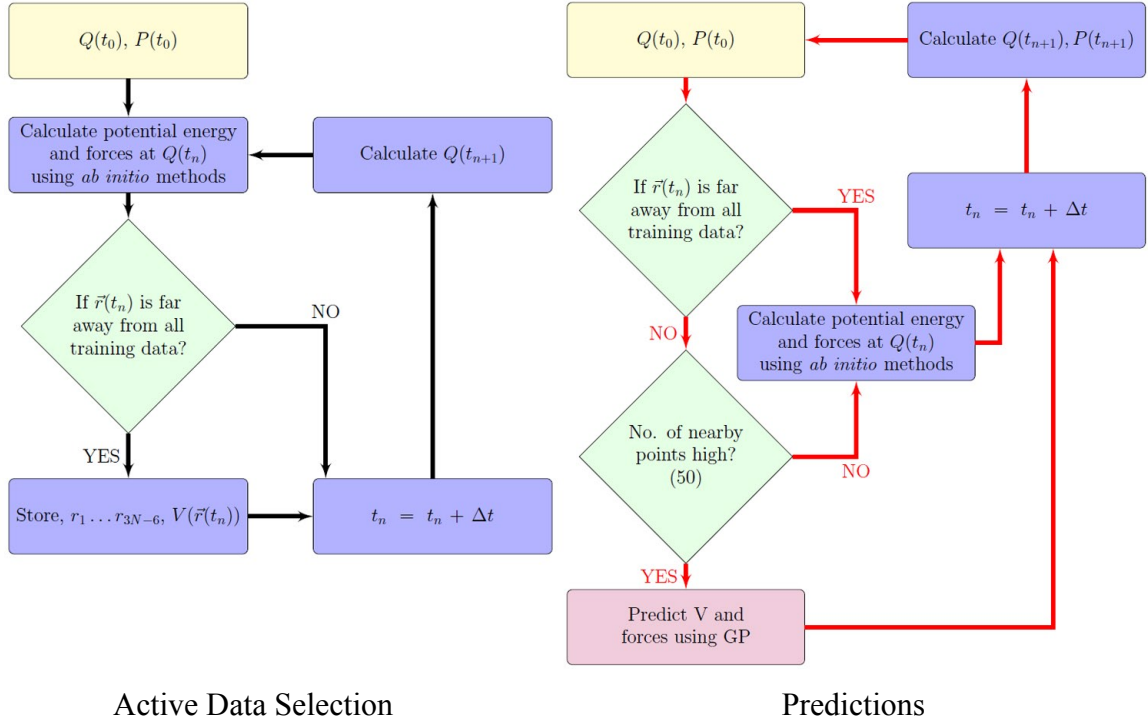


Figure 3.2: Algorithm for “on-the-fly” local PES using GPR.

nearby (defined by the prediction trust radius  $p_0$ , for the current case we used 1.0) to the current position. If the number of points is high (in this case, more than 50 data points), we use GPR to predict the forces and potential, else we use *ab initio* method. For this step we use all the nearby points which fall within the trust radius, to train a local GPR model.

After 40 trajectories, the number of points selected were 4821, see Figure 3.5. We can see that the data points are equally spaced in terms of  $x$  and  $y$ .

For the 41<sup>st</sup> trajectory we run two trajectories simultaneously, with the same initial conditions, one using the GPR predictions and the second using the potential given in equation 3.1. Figure 3.6 shows the 41<sup>st</sup> trajectory, green is for the trajectory in which forces are predicted using GPR, and red is the trajectory where forces are computed using 3.1. It can be seen that both the trajectories deviate after a few steps, although the total energy is conserved.

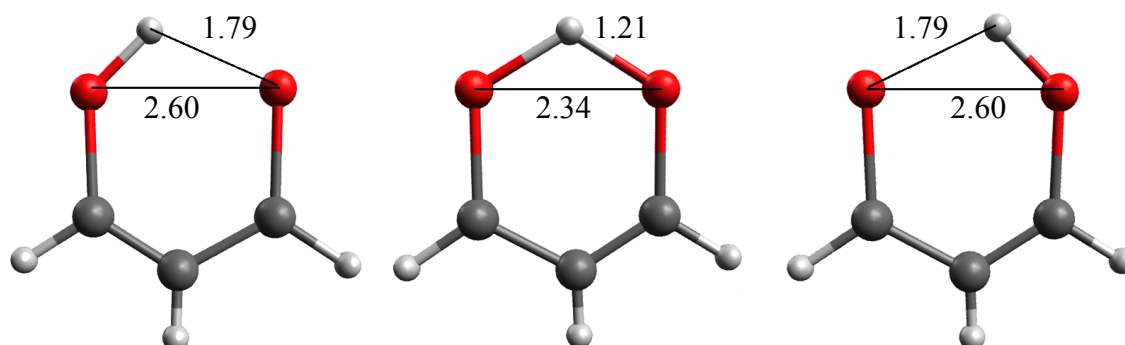


Figure 3.3: The left and the right figures show the reactant and product of malonaldehyde, whilst the one in between shows the transition state for the proton transfer reaction.

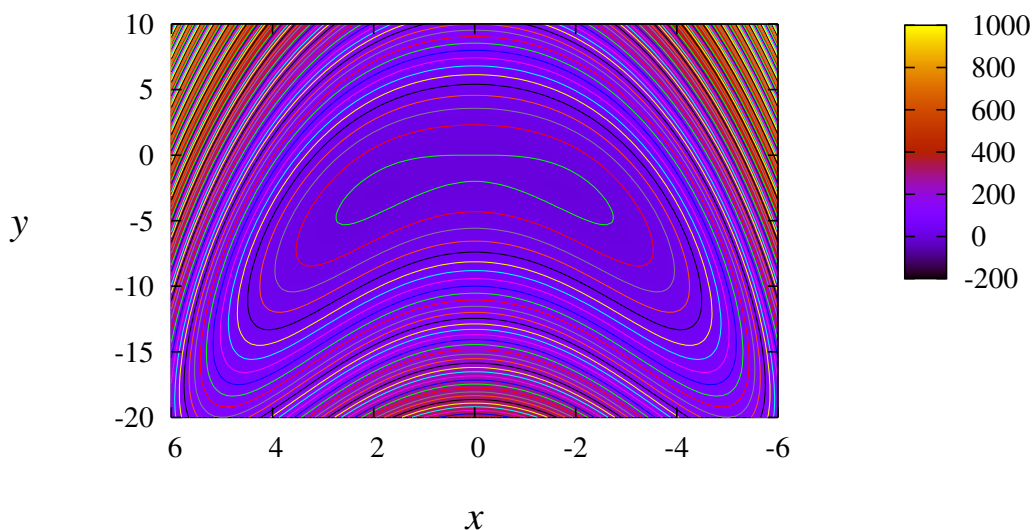


Figure 3.4: Two-dimensional PES contours for malonaldehyde system.<sup>98</sup> There is a transition state at  $x = 0$ ,  $y = 0$ . To the left and right of  $(0,0)$ , there are two minimas which correspond to the reactant and product regions.

### 3.1.1 Advantages and Drawbacks

This algorithm is relatively easy to implement, and because we are interpolating locally, the GPR training time is low with high accuracy. However, as we can see from the previous section the number of points collected for a two dimensional system is 4821. In order to increase GPR prediction accuracy, it is imperative that we store more data. This number becomes a bottleneck in the algorithm, and tends to increase rapidly as we go for higher

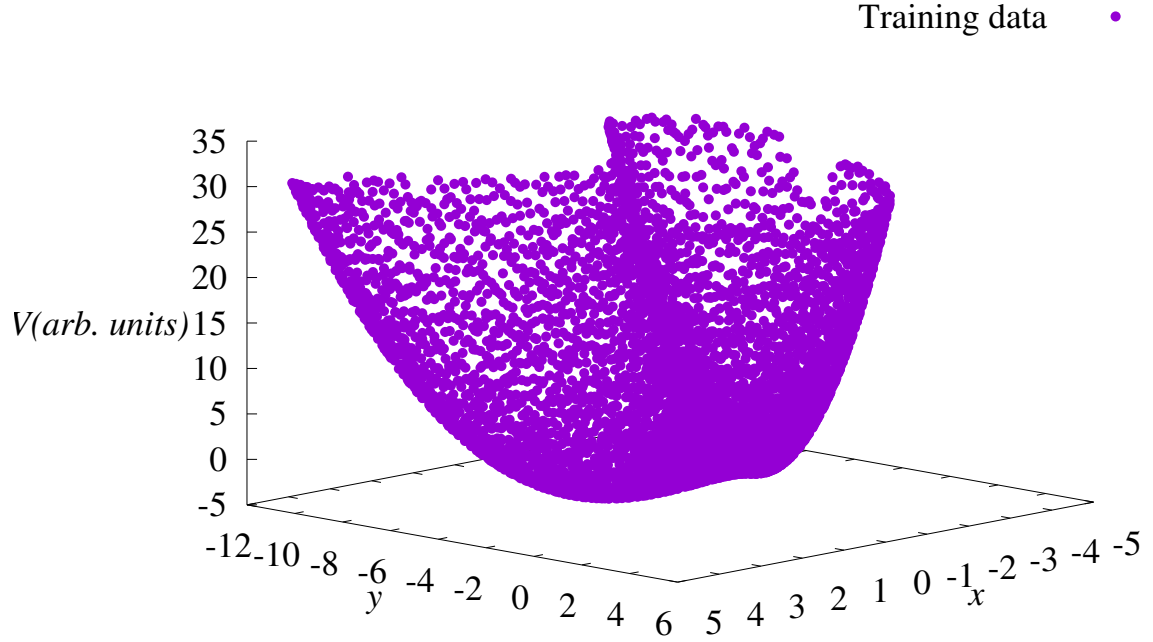


Figure 3.5: Training data collected during 40 trajectories. Number of points here are 4821.

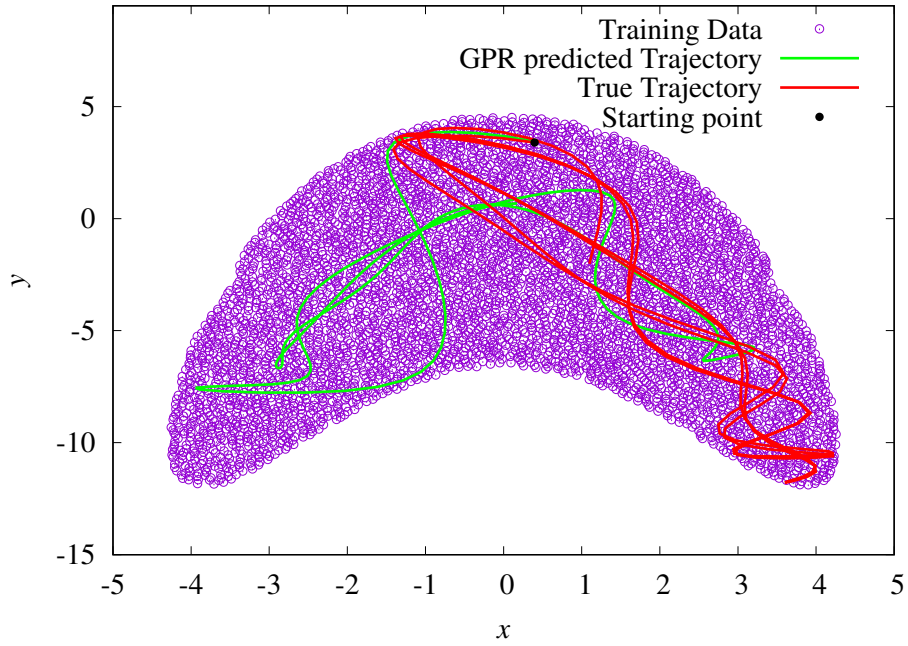


Figure 3.6: First 100 steps of the 41<sup>st</sup> trajectory. Red line is the true trajectory and green line is the trajectory obtained using local GPR PES.

dimensions. That means the for every  $n^{\text{th}}$  time step, we have to search through all the stored data, which can become very CPU intensive. One way to overcome this issue is

to store less data, which comes at the cost of poorer accuracy. Another problem that we encounter is the problem of defining a boundary (for a given total energy). For example, in Figure 3.6, let's say we want to predict at  $x = -4$  and  $y = -12$ . Using the distance criteria mentioned earlier we will find that the number of stored data points around this point is high, yet if we use these points to predict the potential, it will be highly erroneous because the data is distributed unevenly, and the GPR model will be extrapolating. We see this in the 2D case, that for potential energy prediction at the edges where data has been collected, the errors in the prediction are higher. For systems with higher dimensions the boundary becomes more difficult to visualize. One work around is to use some kind of measure which will ensure that the data is evenly distributed around the point of interest. We can divide the configuration space into grids with dimensions close to that of the trust radius used for prediction, and then check for if the point of interest (where we want to predict) lies in any one of the grids, which is populated evenly. For that, after the active data collection phase, all the grid points need to be classified as either “evenly distributed” or “unevenly distributed”, and we predict only in those grids which are “evenly distributed”. However, the minimum number checks required for each time step is  $2^2$  i.e., 4 checks. For example we can divide the entire configuration space into an  $m \times n$  grid, of grid size  $g_x$  and  $g_y$  in the  $x$  and  $y$  directions respectively. For any new point  $(x, y)$ , we need to check iteratively if,

$$M * g_x > x \geq (M - 1) * g_x \text{ and,} \quad (3.3)$$

$$N * g_y > y \geq (N - 1) * g_y \quad (3.4)$$

where the index  $M$  runs from 1 to  $m$  and the index  $N$  runs from 1 to  $n$ . For a system of six dimensions, the number of checks goes to  $2^6$ , which becomes infeasible. Other methods for the check of uniform distribution have been implemented.<sup>99</sup>

### 3.2 “On-the-fly” Representation using ANN

To overcome the problems presented in the previous section, we tried using ANN to represent the PES. ANNs are known to be poor in extrapolating.<sup>100</sup> In this algorithm, the potential energy data obtained from each trajectory is represented by a NN (local representations,  $V^{NN}$ ) and stored. In addition, the coordinates (domain) corresponding to each trajectory is represented by another NN ( $A^{NN}$ ). Thus, the  $V^{NN}$  learns the energy information while the  $A^{NN}$  learns the coordinate information. The global PES of the system is then represented by the linear combination of the local  $V_i^{NN}$  s weighted by the coefficients ( $C_i$ ) given by

$$V_p = \sum_{i=1}^n V_i^{NN} \times C_i \quad (3.5)$$

where,

$$C_i = \frac{A_i^{NN}}{S_n} \quad (3.6)$$

and,

$$S_n = \sum_{i=1}^n A_i^{NN} \quad (3.7)$$

Thus, the coefficients are related to the domain  $A_i^{NN}$  s by Eqn. 3.6.  $V_i^{NN}$  is a neural network which is trained on the data with coordinates as input and the potential energy as output.  $A_i^{NN}$  is a neural network which is trained on the data with coordinates as input and the outputs 0 or 1. The index  $i$  runs over all the trajectories. The training data for  $A^{NN}$  is obtained as follows. From the trajectory, we obtain coordinates where energy information is available and is assigned a value of 1 as output. In addition, random coordinate points where energy data is not available are generated and assigned a value of 0 as output. Together, these two sets of data form the training set for  $A^{NN}$ . The  $A^{NN}$  data with 0 outputs are at least half the number of points from the trajectory data. This ensures that  $C_i$  smoothly go to 0 where energy data is not available. The scheme is illustrated using two sample trajectories

on a one dimensional chemisorption potential well in Figure 3.7. If  $C_i$  is close to 0, the prediction will be poor, however if it is close to 1, we know that a prior trajectory has gone near this point and a prediction maybe valid. In the case when two trajectories have gone through the same region in the configuration space, we have multiple ANNs representing the same region of the PES. Eqn. 3.5 ensures that the final prediction is a weighted average over all the representations. Thus, given enough representations, the final prediction ( $V_p$ ) is expected to tend towards the true potential. One of the main advantages of this design is that we do not store any data for the energies (we can discard potential data after training a network).

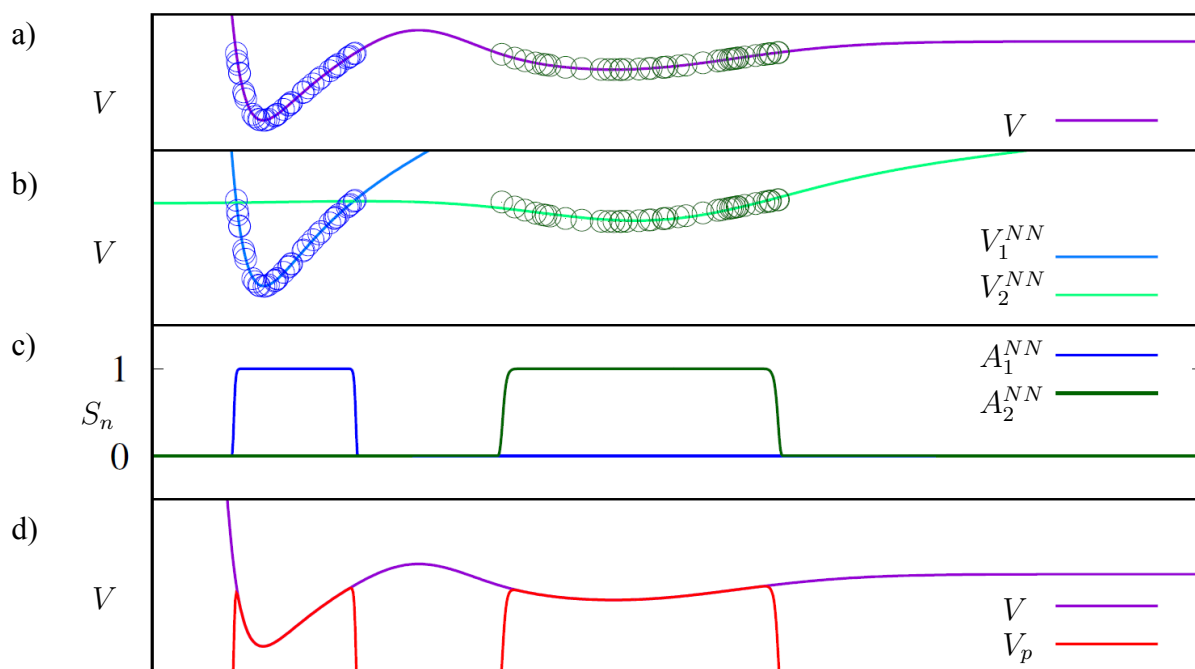


Figure 3.7: Illustration of the working of  $V^{NN}$  and  $A^{NN}$  in “on-the-fly” NN PES. (a) Representative chemisorption potential energy function given in purple, and data collected during two trajectories (blue and green). (b) Two corresponding ANNs produced ( $V_1^{NN}$  and  $V_2^{NN}$ ) by training on the trajectory data. (c) Two domain ANNs ( $A_1^{NN}$  and  $A_2^{NN}$ ) which were trained to learn the coordinate information and (d) Final prediction (in red) using the scheme ( $V_p$ ).

We implemented this algorithm on the 2D malonaldehyde potential. For this, we ran 40 trajectories, each for thousand time steps and train them and start predicting from the 41<sup>st</sup> trajectory. In the training process each of the first 40 trajectories is trained individually to  $V_i^{NN}$  and  $A_i^{NN}$ , for  $i=1$  to 40. For  $V^{NN}$  the training data is the set containing the set of triplets  $(x^i(t), y^i(t), V^i(t))$  for  $t=1$  to 1000. For  $A^{NN}$  the input is the set of all the points that the trajectory has visited, that is  $(x^i(t), y^i(t))$  for  $t=1$  to 1000, and the output for those input points are selected to be 1. We add another randomly generated 500 points  $(x^j, y^j)$  to this set that are far away from the trajectory points (greater than  $r_0$ ) with an output value of 0. The parameter  $r_0$  controls how fast  $A^{NN}$  should die off to zero. For the 2D case, we have set  $r_0 = 1$ . A sample trajectory is given in Figure 3.8. For each  $V^{NN}$  and  $A^{NN}$ , we select a

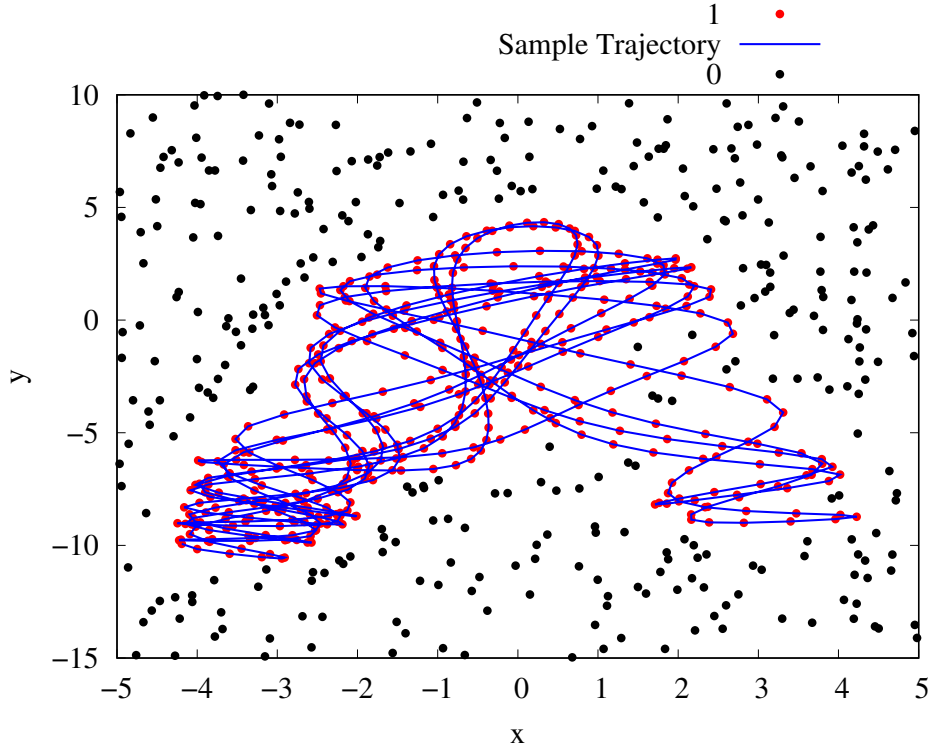


Figure 3.8: Blue line shows a sample trajectory, black dots show the random points selected where the output of  $A^{NN}$  is expected to be 0, and red dots represent where  $A^{NN}$  is 1.

neural network consisting of two hidden layers, with 10 nodes in each layer. The amount of data generated for each trajectory in terms of memory is  $3 \times 1000$  double precision variables. On the other hand, each NN after being trained consists of 151 (130 weights and 21 biases) double precision variables, and because there are two NNs for each trajectory, we store 302 variables. As with the GPR algorithm, we start predicting from the 41<sup>st</sup> trajectory, and for comparison we have taken the initial conditions of this trajectory to be the same.

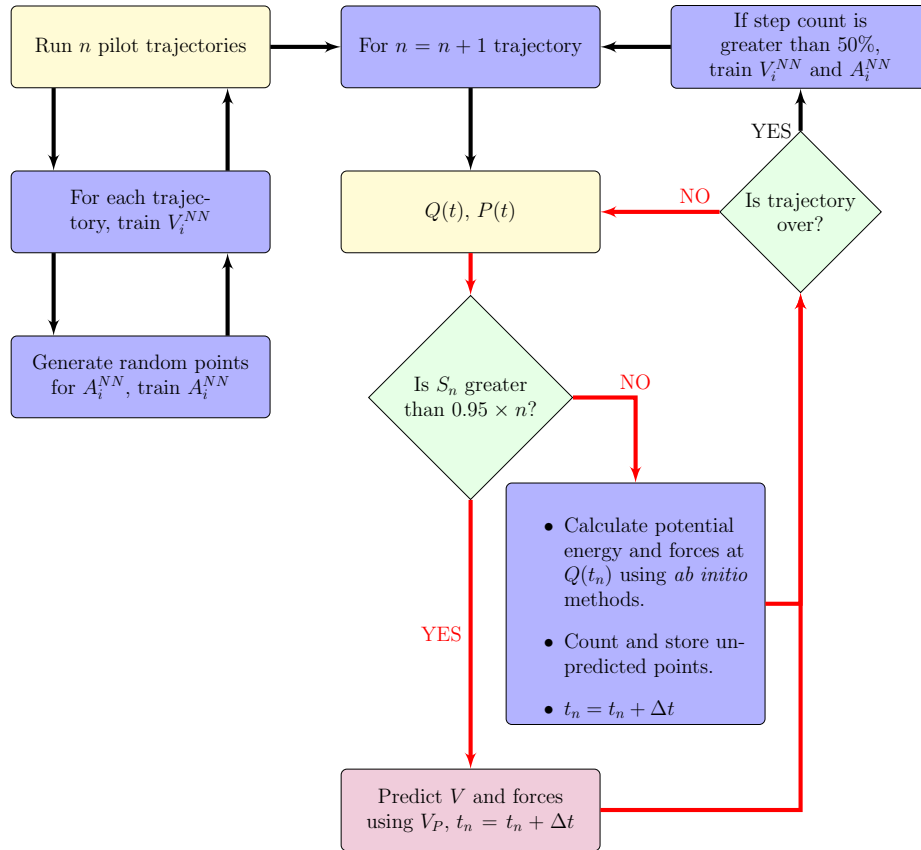


Figure 3.9: Outline of the algorithm used for on the fly representation of PES using ANNs. The prediction part is shown with red arrows.

The algorithm is shown in Figure 3.9 and the 41<sup>st</sup> trajectory using this algorithm is shown in Figure 3.10. During prediction phase, for every timestep we compute the denominator

of Eqn. 3.5 for the current  $x$  and  $y$ . If it is greater than 0.95 times the total number of trajectories run, we use the “on-the-fly” NN PES, if not, we use *ab initio* forces. Eqn. 3.7 is computed before Eqn. 3.5 so that we avoid dividing by zero, if no representations are present at that point. Also, after the completion of the set of pilot trajectories, for each new trajectory we count the number of points that have been predicted by the algorithm. If less than 50% of points are predicted, we append the set of NNs with the representation of this new trajectory.

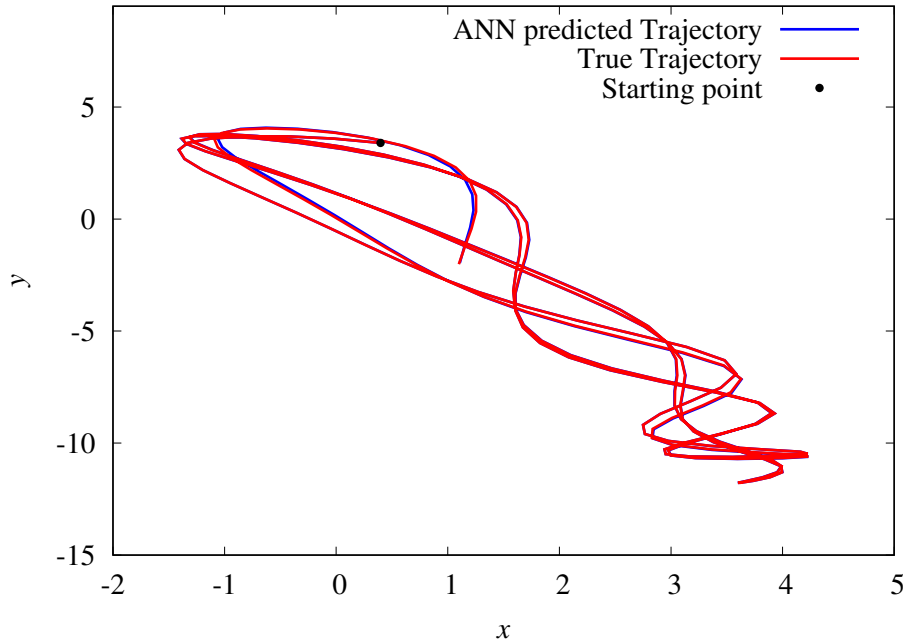


Figure 3.10: Blue line the ANN predicted trajectory, red line is the true trajectory. This shows 200 time steps.

To compare both the methods, we use the euclidean norm of predicted trajectories with respect to true trajectory for each time step of integration. This tells us how much and how fast a predicted trajectory is deviating from the true trajectory. Given enough time, it is expected that the predicted trajectory and the true trajectory will deviate into two different trajectories because of accumulation of errors. See Figures 3.11 and 3.12 for comparison of the methods.

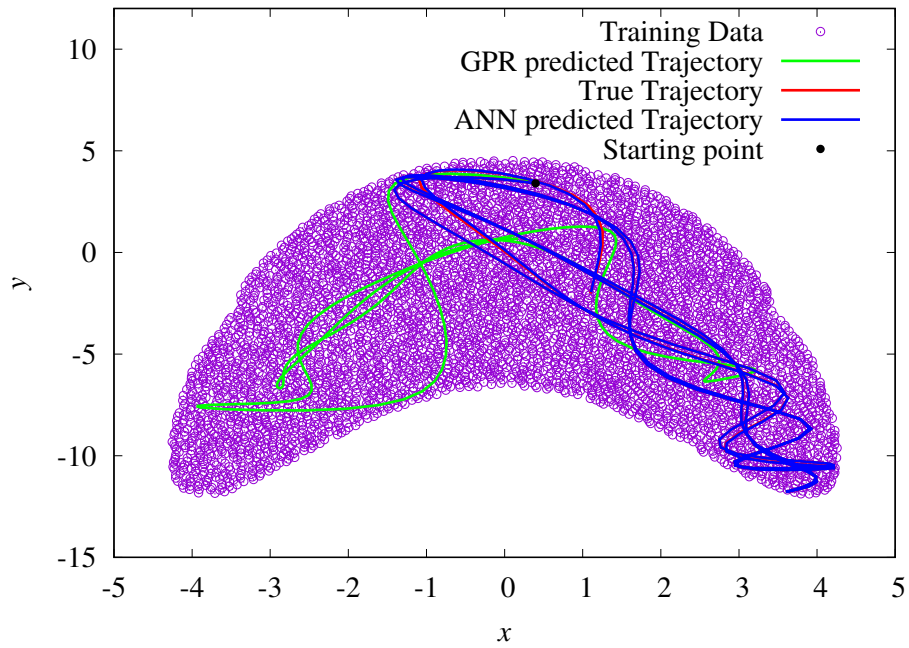


Figure 3.11: Comparison of the performance of “on-the-fly” GPR and NN algorithms for 200 integration steps.

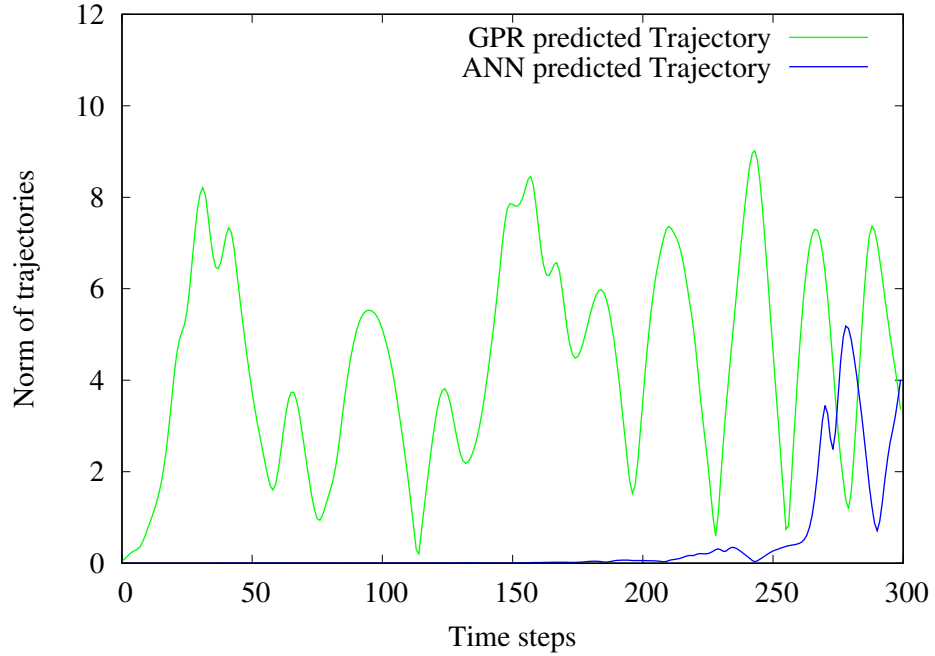


Figure 3.12: Comparison of norms of “on-the-fly” GPR and NN algorithms with the true trajectory.

We also test for the consistency of this method by generating grid of points on the PES, where the potential energies are less than or equal to 30 units. This grid contained 2417 points and was used to test how the algorithm performs as we add more representations by checking the residuals at these points using 10, 20, 30, 40 and 50 representations. The performance of the algorithm is given in Figure 3.13. It can be seen from this figure that as we add more representation, the RMSD over the grid of points tends to decrease. Further the maximum error also decreases.

We can see that the ANN method works better for the two dimensional case, and also that no data had to be stored. It is worth pointing out that the potential used here is non-reactive. If there were a region in the configuration space where the reactant can go through to perform a reaction, results of the predictions will be very different as we will show later on in the 6 dimensional case of formaldehyde.

### 3.2.1 Six Dimensional Non-Reactive ANN

For the 6-D case we used Bowman’s<sup>101</sup> potential for formaldehyde. We use 6 internuclear distances to represent the potential, i.e.  $V(R_1, R_2, R_3, R_4, R_5, R_6)$ . The six internal coordinates are given in Figure 3.14.

We use the ANN approach to fit the potential, given by equation 3.5. With this we set  $n = 10$ , that is we ran 10 trajectories of 0.8 ps, with a time step of 0.2 fs to generate data for training. The neural networks used here consists of two hidden layers containing 25 nodes each, for both  $V^{NN}$ s and  $A^{NN}$ s. The  $r_0$  value used in this system is 0.05. The total energy was set to be 102 kcal/mol, which is below the bond breaking threshold. From the 11<sup>th</sup> trajectory onwards we started predicting. The prediction criteria was set as  $S_n \geq 0.95 \times n$ . Note that in the 6-D case, if less than 90% of points are predicted, we append the set of NNs with the representation of the new trajectory. As in the case of 2D, here also we run another trajectory simultaneously which uses the *ab initio* potentials. The results are summarized

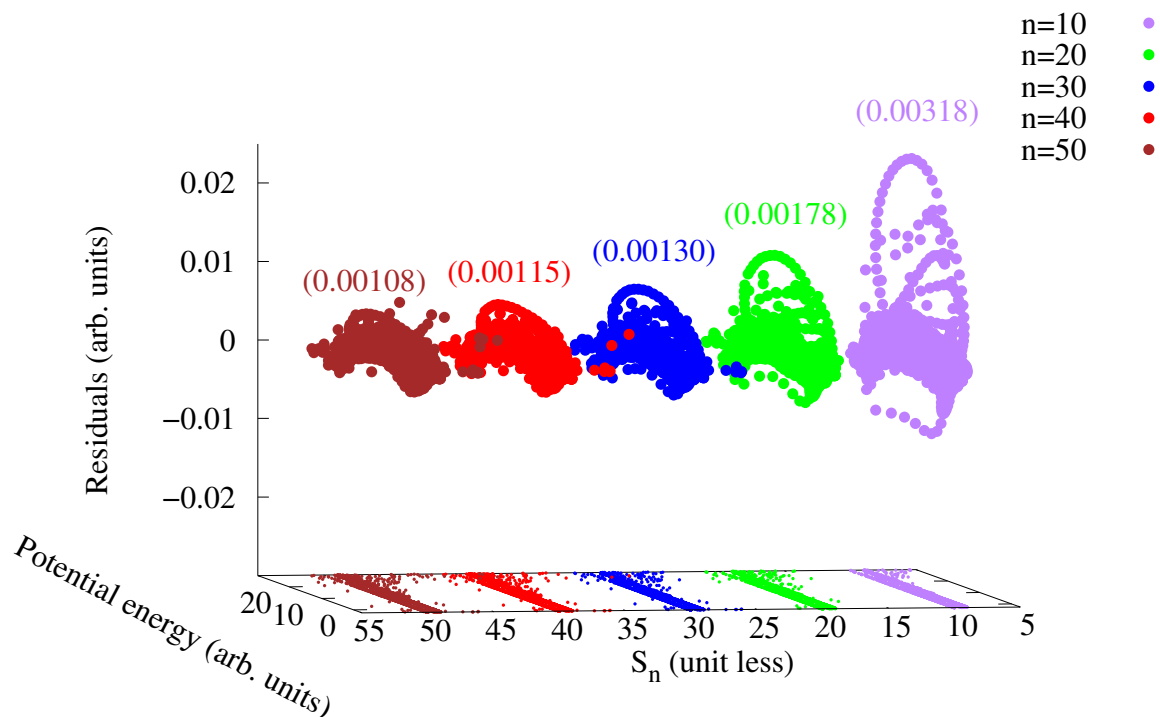


Figure 3.13: Residuals across the potential energy range as a function of the number of representations (in this figure, given by  $S_n$ ). The numbers in the bracket represent the RMSD in arbitrary units.

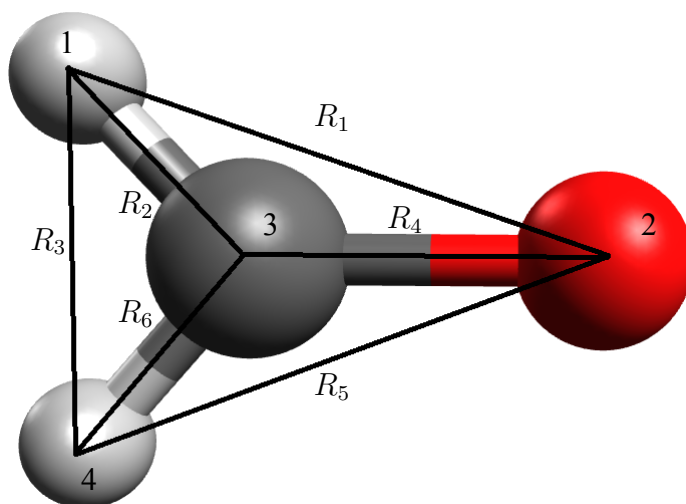


Figure 3.14: The six internal coordinates used to represent the PES of formaldehyde.

in Figure 3.15.

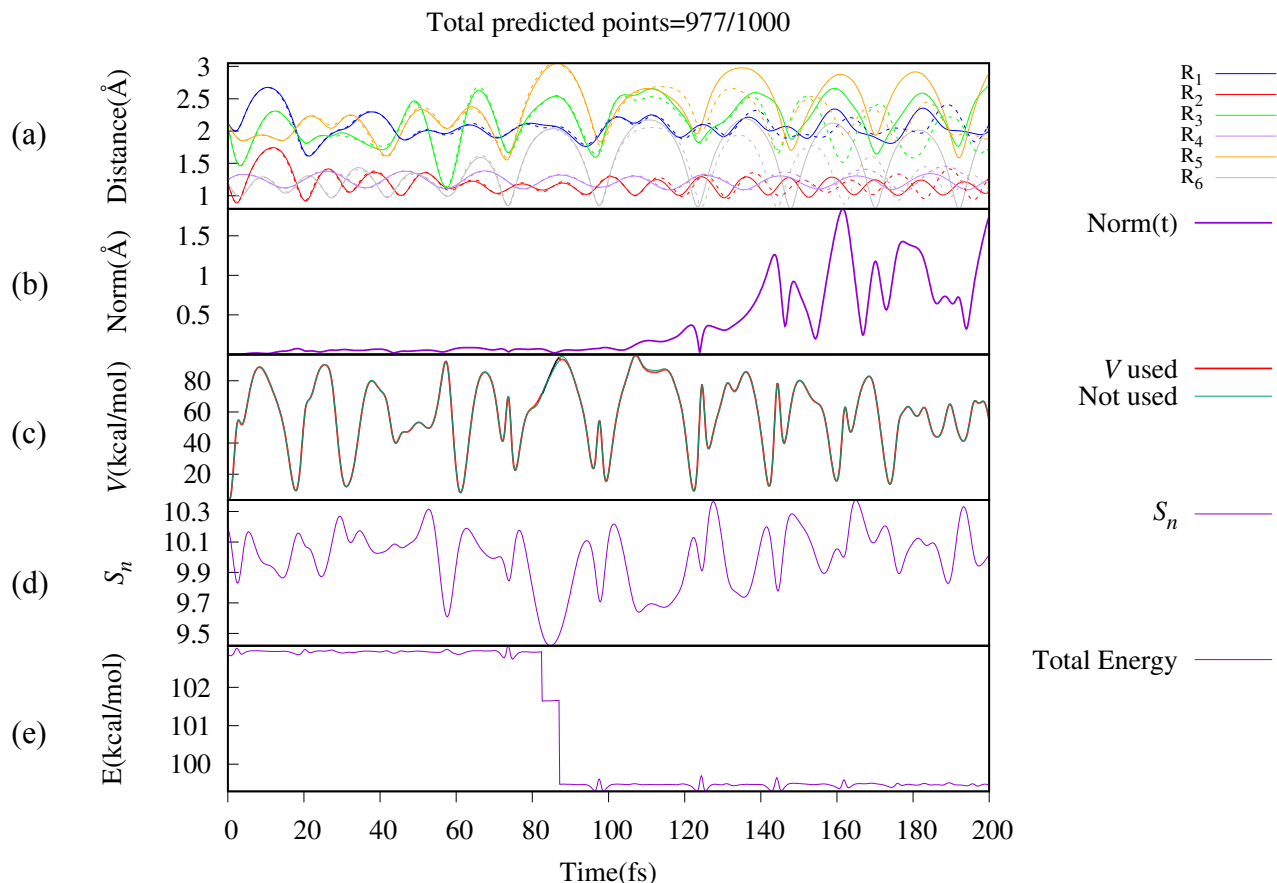


Figure 3.15: Illustration of the working of “on-the-fly” NN PES algorithm for formaldehyde for the 11<sup>th</sup> trajectory. (a) Internal coordinates as a function of time (dotted lines for predicted, and solid lines for true trajectory), (b) euclidean norm of the internal coordinates compared to the true trajectory, (c) potential as a function of time, red is the potential used for integration, and black shows where the algorithm reverted to using *ab initio* points, (d)  $S_n$  values, and (e) shows the total energy as a function of time.

In Figure 3.15 (b) we see that the amount of time required for the predicted trajectory to deviate from the true trajectory was about 100 fs. Other trajectories showed similar results. Figure 3.15 only shows 1000 steps of the trajectory, it can be seen that around 97% of time, the representation could predict the forces, however for the entire length of the trajectory only 53% of the points were predicted. Figure 3.16 shows how the algorithm performs for the six dimensional case. It can be seen that after around 25 trajectories, almost 100% of

all data points could be predicted.

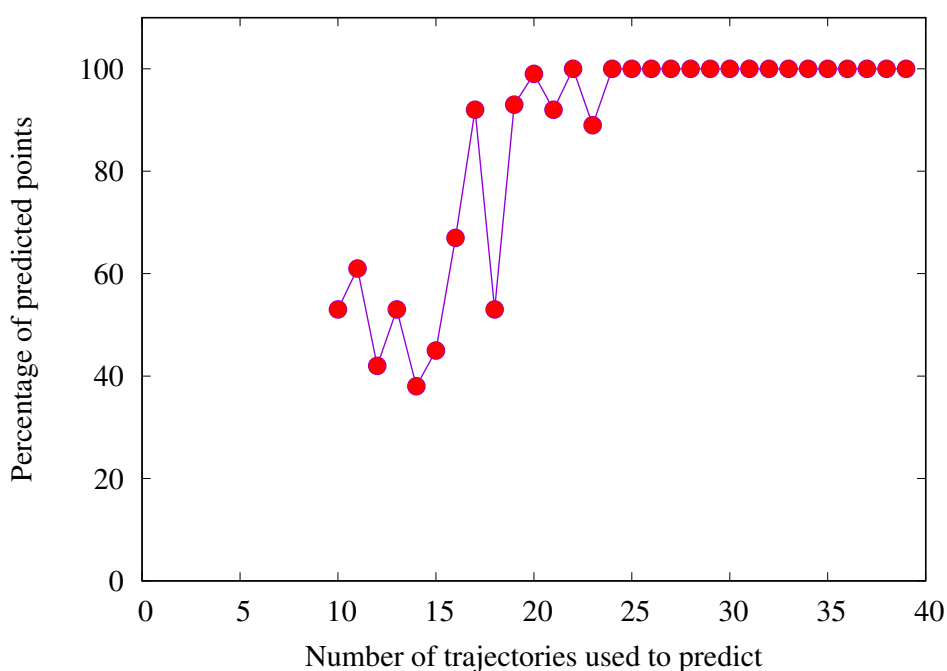


Figure 3.16: Percentage of points predicted during the trajectory as we add more and more representations.

### 3.2.2 Six Dimensional Reactive ANN

For the case of reactive trajectories, we set the total energy at 110 kcal/mol. We ran 10 reactive trajectories with a time step of 0.25 fs, for a maximum trajectory length of 1 ps. We further divided the trajectory into two parts, reactive and non-reactive. We trained each part as individual neural network (Figure 3.17). If any internal coordinate goes above 3.5 Å, we call the reaction to have occurred.

For the trajectory shown in Figure 3.17 we computed the potentials using the reactive ANNs as well as the non-reactive ANNs, see Figures 3.18 and 3.19. Here we did not use

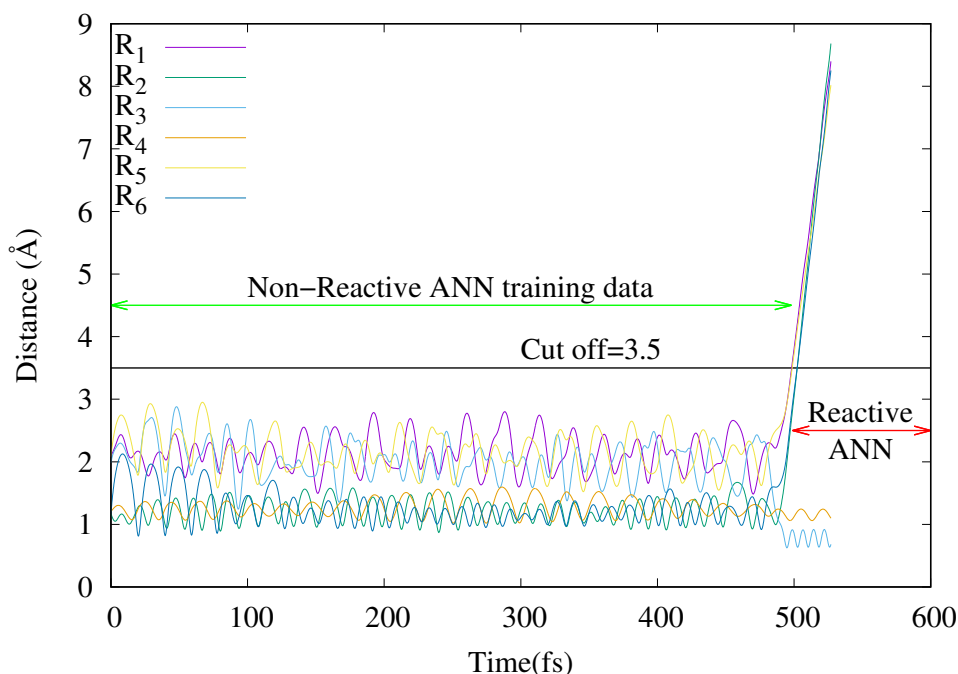


Figure 3.17: Internal coordinates as a function of time for a representative reactive trajectory, where the products formed are  $\text{H}_2$  and  $\text{CO}$ .

“on-the-fly” forces to compute the trajectory, rather we are only comparing “on-the-fly” potential energy values computed during the 11<sup>th</sup> trajectory with *ab initio* potentials. We find that the non-reactive ANNs can predict the non-reactive region well. However the performance of these are not good when we only trained the ANNs on the reactant well region, as shown in the previous section. Figure 3.19 shows the performance of reactive ANNs, when the trajectory is in the reactant well, as well as when the bond-breaking occurs. It can be seen that the performance of these set of ANNs is not as good in the reactant region, which is expected because they were not trained on any data from this region. However, the set of reactive ANNs were also performing poorly during bond breaking. This can be attributed to the fact that the amount of data generated in the reactive region is less. This type

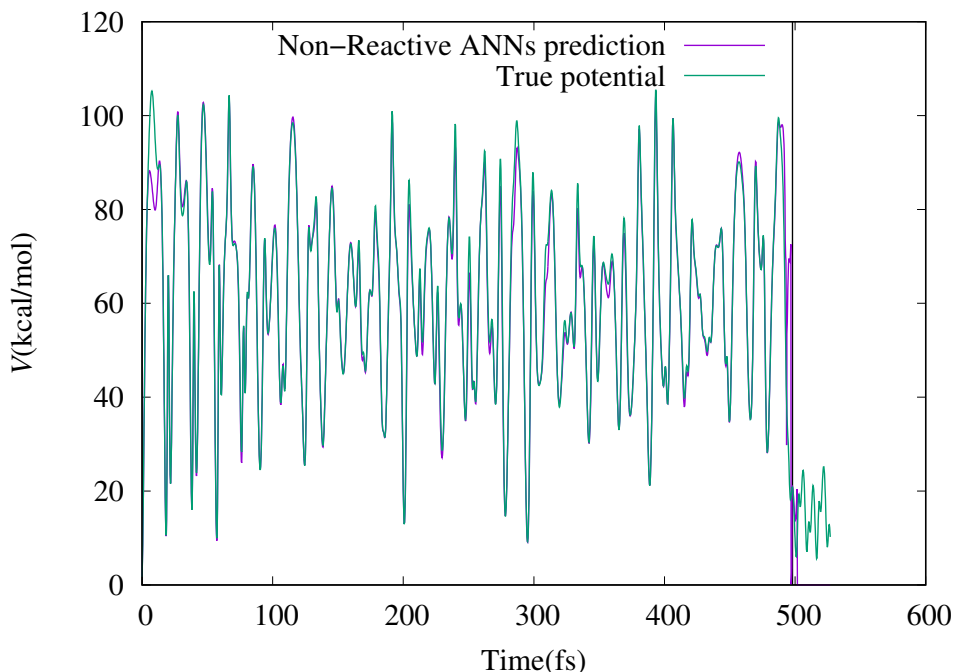


Figure 3.18: Non-reactive ANNs prediction for the entire trajectory. The black line at around 500 fs shows where the reaction has occurred.

of behavior is expected in the reactive region, and a better algorithm needs to be developed.

### 3.3 Summary and Conclusions

The problem of representing PESs “on-the-fly” is different from traditional PES fitting. The available data comes as a time series, and with respect to predictions, it is difficult to say which data point will be important. To this end, we developed two algorithms to tackle the problem of “on-the-fly” representation of PESs, one using GPR, and the other using ANNs. The local GPR PES works well for low-dimensional systems. On the other hand, “on-the-fly” NN PES works better for 2D and works reasonably well for 6D.

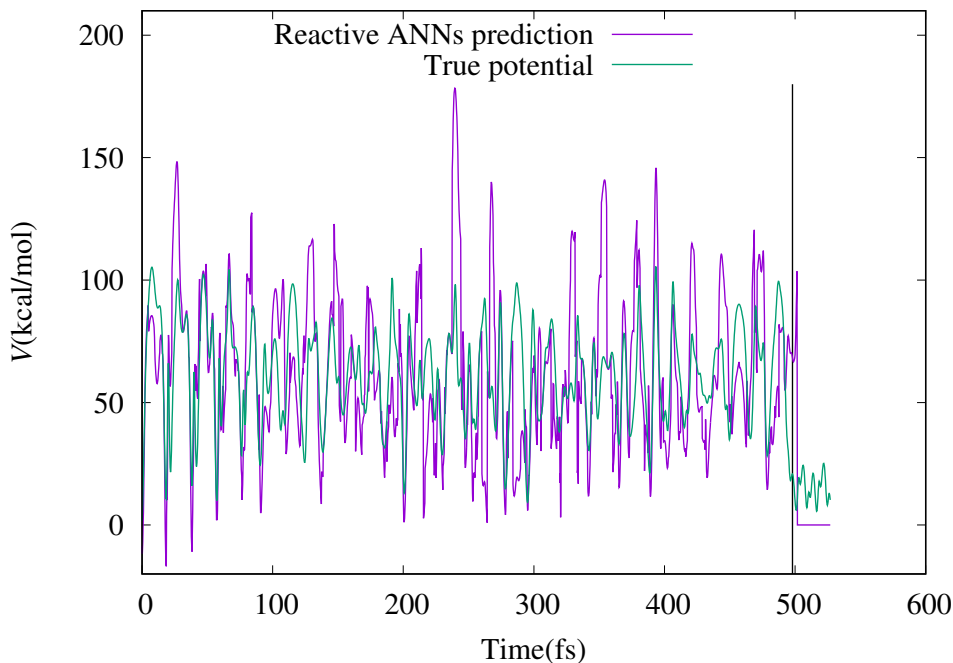


Figure 3.19: Reactive ANNs prediction for the entire trajectory. The black line at around 500 fs shows where the reaction has occurred.

For the case of ANNs, another approach would be to use all data generated during the dynamics from one of the trajectories to train an ANN and for that we can not say with any certainty that, the ANN will be able to predict energy and forces for the next trajectory. While one can always retrain the ANN, but then the network will not work for the first trajectory. In order to make it work for both the trajectories, all the prior information (trajectories one and two) needs to be stored, and then the network needs to be retrained on the combined data. In this respect “on-the-fly” NN PES, avoids the problem of retraining by storing individual representations ( $V_i^{NN}$ ), rather than the data. Further the problem of ensuring that an individual ANN will always be interpolating, a parallel ANN ( $A_i^{NN}$ ) learns the domain information.

Further, in traditional ANN fitting from our experience we find that usual network size for a six dimensional system is expected to be around at least of the order of 100 neurons per hidden layer, with two hidden layers, which yields a total of 10500 weights and biases. However, if we split the representation into smaller ones, we can use smaller neural networks as shown in this chapter. Yet, the cost (in terms of memory) of storing multiple neural networks is higher ( $\approx 37500$  weights and biases for 25 representations). Multiple modifications can be made to the “on-the-fly” NN PES algorithm to improve it, for example discarding representations which have a poor performance. The algorithm is in principle, completely parallelizable, that is the output for each NN can be computed by a different CPU. After a bunch of *ab initio* trajectories have been computed, one ends up with multiple representations, which can be used elsewhere. One major problem here is that there is no quality check for the predictions because no data is being stored to compare it with. Further, the training time for each  $V_i^{NN}$  for the 2D case was around 1 minute each, whereas for the 6D case it varied from 20 minutes to 1 hour and in general the training time of  $A_i^{NN}$  was faster, around 5 minutes each for the 6D case. These CPU times are much larger than typical *ab initio* force calculation CPU times. However, if the ANN fits are already available, in a region of the PES, the computation time for the potential energy and force calculations (numerical derivatives for 12 cartesian coordinates) with 25 representation was around 3 to 4 seconds for each time step, which is faster than the usual *ab initio* calculations.

In conclusion, we find from our study that ANNs perform well at learning the PES “on-the-fly” for reactant region, however more work needs to be done to make it work for reactive regions of the PES. In most molecular systems, the amount of time the reactant spends in the reactant or the product well is large. To this end, our “On-the-fly” ANN approach can greatly reduce computational cost of computing forces/potentials in these regions once many trajectories have sampled those regions. On the other hand, reactive regions often are less likely to be sampled during trajectories. Hence it is reasonable to assume that the data

collected in these regions would be less. It would be beneficial to store points at the reactive region from multiple reactive trajectories till the concentration of the points in this region is high enough. After that a single ANN can be trained to represent the reactive region. Consequently, when going from reactant well to reactive region, the algorithm can shift to the reactive PES as and when required.

## Chapter 4

# Gas-Surface dynamics: Formaldehyde on Au(111) surface

Inelastic scattering between gas molecules and surfaces is a fundamental process that has been investigated extensively.<sup>102–112</sup> During this process, energy exchange between the gas and the surface results in changes in the internal and translational degrees of freedom of the scattered gas molecules. However, the mechanism of energy transfer between the molecule and surface is still not well understood. In recent gas-surface scattering experiments<sup>113</sup> on formaldehyde scattering off the gold surface, a high propensity to excite twirling motion about the C-O bond axis was observed for the scattered formaldehyde molecules.

In the work presented here, we used classical trajectory simulation to understand energy transfer in formaldehyde-surface collisions, to probe the mechanism of interconversion of translational energy to rotational energy. Our simulations capture the trend of the experiments for the formaldehyde-Au scattering, both in terms of rotational energy distribution, as well as trapping probabilities.

The interaction of molecules with surfaces result in the physisorption and/or chemisorption of the adsorbate which are generally the initial steps in catalytic processes. Hence, the interaction of atoms or molecules with solid surfaces has been a topic of research for almost a century due to its importance in heterogeneous catalysis. In particular, several studies have focused on investigation of the energy transfer between the adsorbate and the surface since it directly influences the rate of adsorption and the ‘stickiness’ of the adsorbate on the surface.<sup>104, 106, 114–116</sup> The amount of time the atoms or molecules spend at the gas-surface interface determines the extent to which energy transfer takes place.<sup>117</sup> Also, the interac-

tion between the adsorbate and substrate opens up channels of energy transfer between the rotational, vibrational, and translational degrees of freedom of the species.<sup>105, 118–121</sup>

Rainbow in atom-atom scattering has been observed in experiments,<sup>122, 123</sup> where there lies a singularity in scattering intensity with respect to the scattering angle, which can be mapped as a function of the impact parameter between the two colliding atoms.<sup>111, 124</sup> Similarly, rainbow scattering has been observed in scattering of atoms from a surface, in which the scattering intensity also turns out to be a function of impact parameter for the surface.<sup>107–110, 125, 126</sup> Furthermore, rotational rainbow where one or both of the scattered molecules is rotationally hot, has been observed for molecule-molecule collisions,<sup>122, 123, 127</sup> and in diatomic molecules scattering from surfaces.<sup>126</sup> In the latter case it has been pointed out that the amount of rotational excitation of the departing molecule depends on the incident orientation with respect to the surface. Experiments where such singularities are observed, either in scattering intensities as a function of scattering angle or rotational excitations with respect to orientation angle, in principle make it possible to deduce the interaction potential between the two by looking at the scattered products.

There have been a few experimental studies looking at rotational rainbow of polyatomic molecules from collision on surfaces,<sup>113, 128, 129</sup> CH<sub>4</sub> from LiF(100), NH<sub>3</sub> from Au(111) and H<sub>2</sub>CO on Au(111). The study involving NH<sub>3</sub> scattering from Au(111) surface, was supported by classical trajectory simulations,<sup>130</sup> where they find a high propensity for rotationally cold NH<sub>3</sub> molecules to be scattered into low- $k$  states (tumbling rather than twirling). The rotational rainbow was attributed to dynamical steering effect, where the rotationally cold NH<sub>3</sub> molecules had enough time to reorient itself in a particular orientation due to the nature of interaction potential, before the collision.

Recent series of quantum state-resolved REMPI experiments have shown high propensity to excite  $a$ -axis in collisions of formaldehyde with Au(111) surface.<sup>113, 117</sup> They find that direct scattered formaldehyde molecules follow a non-maxwellian distribution in terms

of rotational states.

In this study, we propose mechanisms which could explain what happens at molecule-surface scattering, before and after the collision. The *raison d'être* for the rotational rainbow turns out to be a consequence of dynamics on the interaction potential between the formaldehyde and Au(111) surface. In what follows, we first reproduce the experimental trends. Then we move on to the effect of surface temperature and orientation and finally we look at energy transfer.

## 4.1 Methods

### 4.1.1 Potential energy model

The inelastic scattering of formaldehyde on gold surface was studied using all-atom classical trajectory simulations. The potential energy of the system is written as

$$V = V_{\text{gold}} + V_{\text{form}} + V_{\text{int}} \quad (4.1)$$

where,  $V_{\text{gold}}$ ,  $V_{\text{form}}$ , and  $V_{\text{int}}$  are potential energies of gold surface, formaldehyde, and the interaction between gold surface and formaldehyde respectively. We used a rectangular-shaped HCP lattice of gold (111) surface which consisted of three layers of Au atoms, with a total ( $n_g$ ) of 1045 atoms. The bottom layer comprised of 342 gold atoms, while the middle and top layers contained 361 and 342 atoms each. The four corner atoms of the first two layers and all the atoms of the bottom layer were kept static through out the simulations. The many-body semi-empirical Sutton–Chen (SC) interaction potential<sup>131</sup> was used to represent the bulk of gold atoms,

$$V_{\text{gold}} = \epsilon \sum_i^{n_g} \left[ \frac{1}{2} \sum_{j \neq i}^{n_g} \left( \frac{a}{r_{ij}} \right)^n - c \sqrt{\rho_i} \right], \quad (4.2)$$

$$\rho_i = \sum_{j \neq i} \left( \frac{a}{r_{ij}} \right)^m. \quad (4.3)$$

where,  $\epsilon = 1.2793 \times 10^{-2}$  eV,  $a = 4.08$  Å,  $m = 8$ ,  $n = 10$ , and  $c$  is a dimensionless quantity with a value of 34.408.

The formaldehyde was represented using Bowman's 6D potential ( $V_{\text{form}}$ ),<sup>101</sup> which is direct product of multinomials in Morse variables for six internuclear distances,  $r_1, \dots, r_6$ . The interaction between gold and formaldehyde was represented using the 12-6 Lennard-Jones (L-J) potential,

$$V_{\text{int}} = \sum_X \sum_Y V_{XY} \quad (4.4)$$

where, the Xs are the Au atoms of the surface and Ys are the C, H, and O atoms of formaldehyde. The parameters used in the three types of L-J potential are given in Table 4.1.

Table 4.1: Definition of L-J parameters used to represent gold-formaldehyde interaction.

X	Y	$\epsilon_{XY}$ (kcal/mol)	$\sigma_{XY}$ (Å)
Au	C	0.295	2.99 <sup>132</sup>
Au	H	0.002	3.95 <sup>133</sup>
Au	O	0.863	3.18 <sup>133</sup>

### 4.1.2 Trajectory Simulations

All trajectory calculations were done with velocity Verlet algorithm using VENUS chemical dynamics software package.<sup>134</sup> The initial conditions used in the simulations were chosen to model the experiments by Wodtke and coworkers.<sup>113</sup> First, the gold surface was equilibrated, for 40 ps with velocities re-scaled at every 60 fs to a desired temperature of  $T_s$  K. After this, the surface was annealed for another 40 ps. The average temperature during annealing was monitored and was found to be  $T_s$  K. This set of momenta and coordinates of the gold atoms obtained at the end of the annealing were used as the initial conditions for the gold surface for that trajectory.

For formaldehyde, the rotational temperature was set to 15 K, and a vibrational temperature of 15 K was sampled from a Boltzmann distribution. The angle between the formalde-

hyde beam and the surface normal was kept at  $0^\circ$ . Trajectories were initiated keeping the distance between the center-of-mass (COM) of formaldehyde and the aiming point on the surface at  $20 \text{ \AA}$  and the system was time evolved for a maximum of 10 ps. Trajectories were stopped if the distance between the COM of formaldehyde and the surface attained a distance of  $22 \text{ \AA}$  or if the COM of formaldehyde was close to the edge of the surface.

More than 8000 trajectories were carried out to study the formaldehyde-gold scattering for five incident kinetic energies (collision energies): 0.33, 0.39, 0.47, 0.88 and 1.21 eV at a surface temperature of 300 K which correspond to the experimental conditions.<sup>113</sup>

To study the effect of surface temperature on the dynamics at gas-surface scattering, more than 2500 trajectories were calculated for each collision energy, for five different set of surface temperatures, which are  $T_s = 140, 200, 300, 400$  and  $480 \text{ K}$ . The temperature range was chosen to be such based on the experimental setup of Wodtke and coworkers.<sup>113</sup>

Furthermore, to understand the effect of orientation around 1000 trajectories were initiated for each collision energy for a given set of five orientation, with the initial rotational temperature set at 0 K. Three surface temperatures were simulated, which are  $T_s = 200, 300, \text{ and } 400 \text{ K}$ . Table A.1 gives the list of all computed trajectories.

### 4.1.3 Trajectory Analysis

To understand trapping probability, each set of trajectories were divided into two categories, either trapping, or scattered. Any trajectory in which the formaldehyde was scattered away by a distance of  $22 \text{ \AA}$  after integration time of 10 ps, were classified as scattered. For the rest of the trajectories, if the formaldehyde was still close to the surface after 10 ps, we calculated formaldehyde's kinetic energy in the positive  $z$ -direction along the surface normal and compared it to the interaction potential. If the kinetic energy was greater than the interaction potential, we considered the trajectory as scattered. The rest of the trajectories were classified as trapped. The ratio of the number of trapped trajectories with respect to

the total trajectories defined the trapping probability.

For all other analysis presented in this work, like average rotational energies, angular distribution, we considered only those trajectories which were away from the surface by a distance of 22 Å after 10 ps of integration time. Further, out of those trajectories we only consider the ones in which the final vibrational energy of formaldehyde was greater than the zero point energy.

The rotational energy of the direct scattered formaldehyde was studied in the three axis of rotation, which are *a*-axis (twirling), *b*-axis (tumbling) and *c*-axis (cartwheeling) as shown in the Figure 4.1. The blue arrow indicates the principle axis(*x*) of rotation. The total rotational energy of formaldehyde is given by,

$$E_{\text{rot}} = E_a + E_b + E_c \quad (4.5)$$

$$E_{\text{rot}} = \frac{L_x^2}{2I_x} + \frac{L_y^2}{2I_y} + \frac{L_z^2}{2I_z} \quad (4.6)$$

where  $L_x$ ,  $L_y$  and  $L_z$  are the rotational angular momenta of formaldehyde,  $I_x$ ,  $I_y$  and  $I_z$  are the moment of inertia tensor elements and  $x$ ,  $y$  and  $z$  are the orthogonal directions in the molecular frame of formaldehyde shown by blue, green and red arrows respectively in Figure 4.1. The three molecular axis are shown as red, green and yellow. The rotation with the blue axis as the axis of rotation denotes *a*-axis rotation. Similarly the rotation with rotation with the green axis as the axis of rotation denotes *b*-axis rotation and the rotation with the red axis as the axis of rotation denotes *c*-axis rotation is the After the end of each scattered trajectory, we calculate the rotational energy in each axis averaged over another 300 time steps to get the final rotational energy.

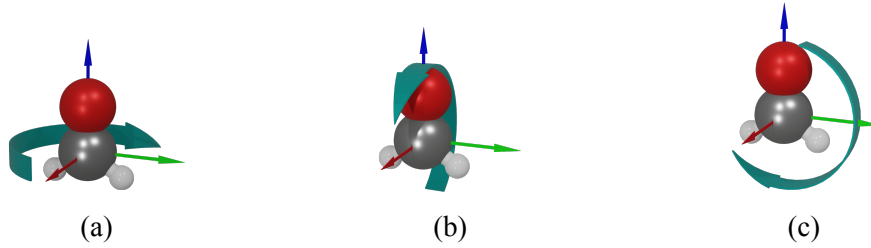


Figure 4.1: The three axes of rotation of formaldehyde. (a) Shows  $a$ -axis rotation, similarly (b) and (c) show  $b$ -axis rotation and  $c$ -axis rotation respectively.

## 4.2 Results and discussion

For  $T_s = 300$  K, a total of 8000 trajectories are analysed for each incident kinetic energy. Figure 4.2 shows the average rotational energies of formaldehyde and trapping probabilities for different collision energies obtained from the experiments.<sup>113</sup> It also shows the rotational energy distributions along the three axes of rotation obtained from simulations by Wodtke et al.<sup>113</sup> From our simulations, as shown in Figure 4.3(a), it can be seen that the average total rotational energy ( $E_{\text{rot}}$ ) increases almost linearly with increase in collision energy, and the magnitude of the rotational energies are comparable to experimentally observed values, see Figure 4.2.<sup>113</sup> And for all collision energies other than 0.33 eV, rotational energy in the  $a$ -axis( $E_a$ ) dominates. The slight dip in the  $a$ -axis rotational energy when going from 0.88 to 1.21 eV is also observed in the experiments, but it is less pronounced than in experiments. Figure 4.3(b) shows the trapping probabilities as a function of collision energy, compared to the experimentally observed values. At low collision energies formaldehyde has enough time to interact with the gold sheet, and it is expected that the trapping probability will tend towards unity. Figure 4.4 shows the distribution of  $E_{\text{rot}}$  for collision energy of 0.33 and 1.21 eV. The insets show the distribution in the individual axis of rotation. These distributions are very different in terms of magnitude to the previously reported data (see panels (c), (d) and (e) in Figure 4.2).<sup>113</sup> We can see that, for 0.33 eV, the  $E_{\text{rot}}$  distribution is exponential in nature, while for 1.21 eV, it is not. It has signatures of a Gaussian centered around 0.5 eV.

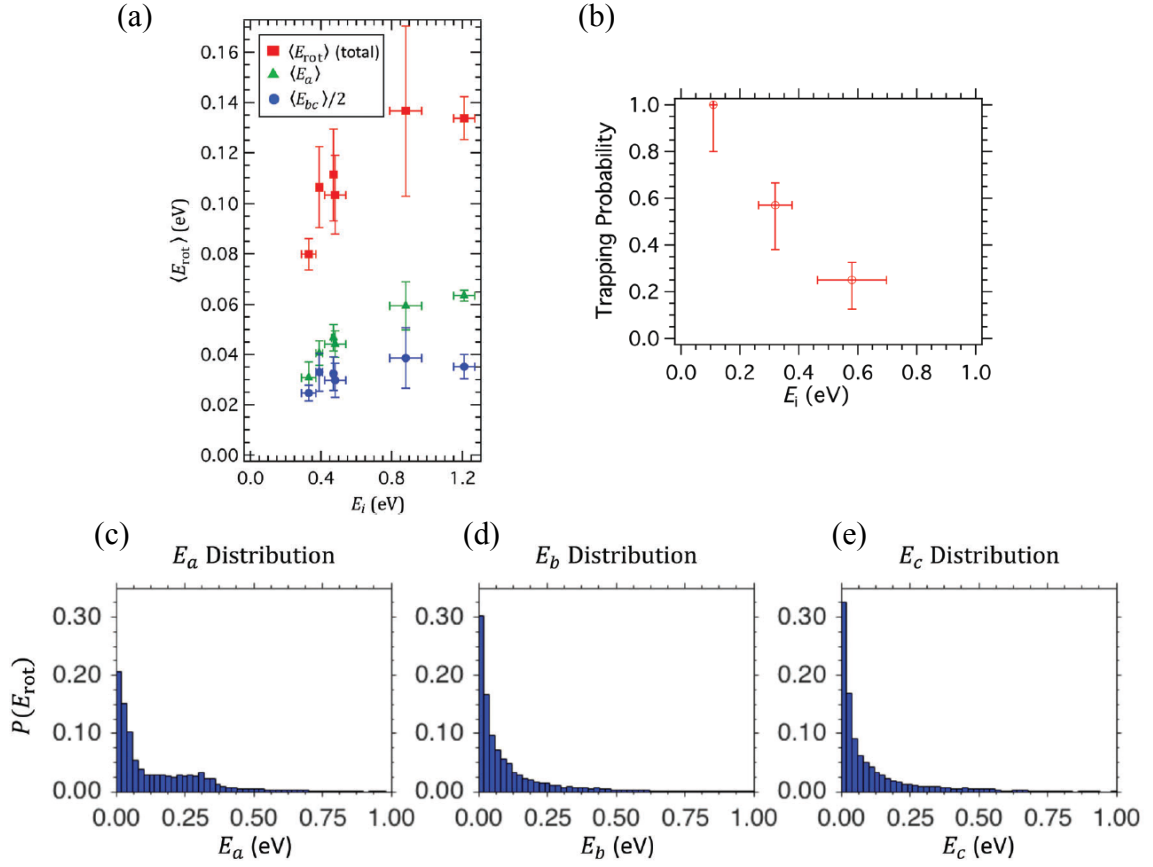


Figure 4.2: This figure shows results from Wodtke et al's work.<sup>113</sup> Panels (a) and (b) shows the average rotational energies and the trapping probabilities obtained from experiments. Panels (c), (d) and (e) show the distribution of rotational energies for the three axis of rotation computed from simulations.

### 4.2.1 Direct vs indirect scattering

Some trajectories fall under the category of trapping-desorption, where the formaldehyde molecule bounces multiple times on the surface before leaving. Such trajectories have a different distribution of final states. Figure 4.5 shows percentage of trajectories showing trapping desorption as a function of collision energy for  $T_s = 300$  K. It can be seen that regardless of collision energy most of the trajectories are single bounce (direct scattered). For all collision energies the percentage of direct scattered trajectories is more than 90%.

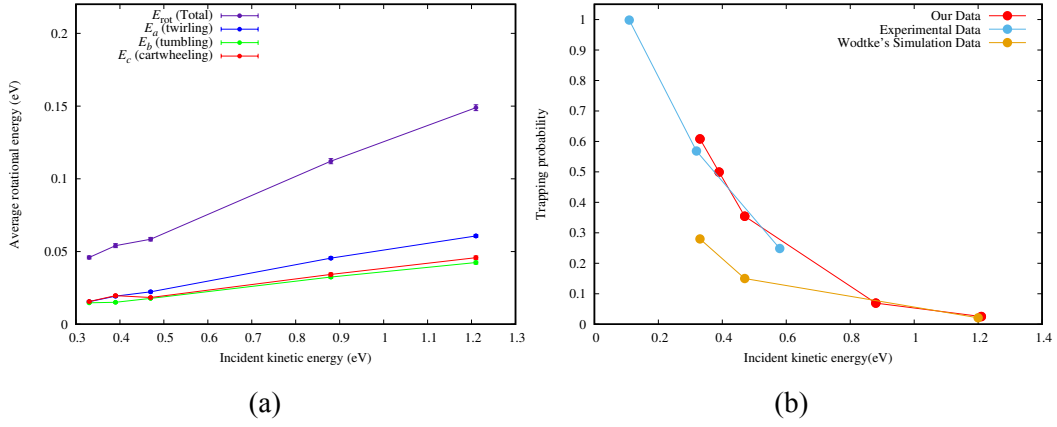


Figure 4.3: (a) Rotational energies of formaldehyde as a function of collision energy and (b) Trapping probabilities as a function of collision energy for  $T_s = 300$  K. The error associated with the trapping probabilities is maximum for 0.39 eV at 1.04 %.

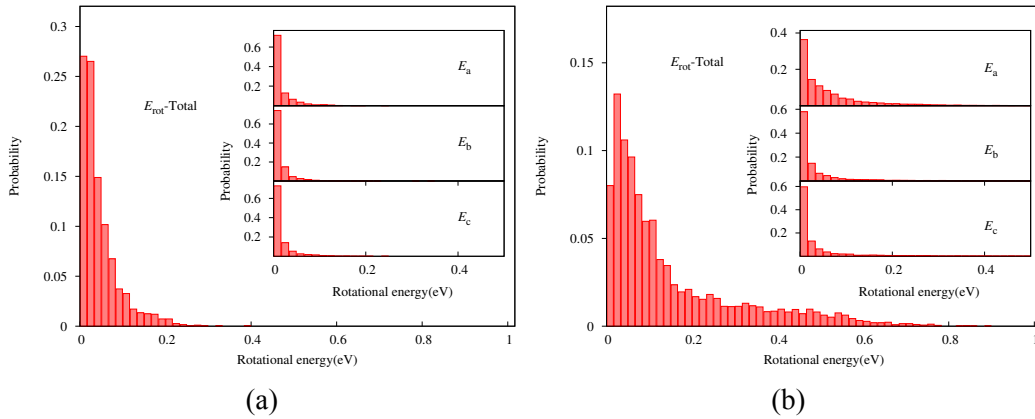


Figure 4.4: Population distribution of rotational energies of formaldehyde for (a) 0.33 eV and (b) 1.21 eV of collision energies and  $T_s = 300$  K.

Further, the percentage of direct scattered trajectories increases as a function of collision energy. Trajectories having more than one bounce all have been clumped together in Figure 4.6 to show the average total rotational energy.

From Figure 4.6 it is clear that the more amount of time formaldehyde spends on the surface, the more energy it can gain from the surface. Also the amount of rotational excitation seems to be higher for higher collision energies which is known as “memory effect”.<sup>135</sup> It states that the projectile’s final distributions in internal and external degrees of freedom

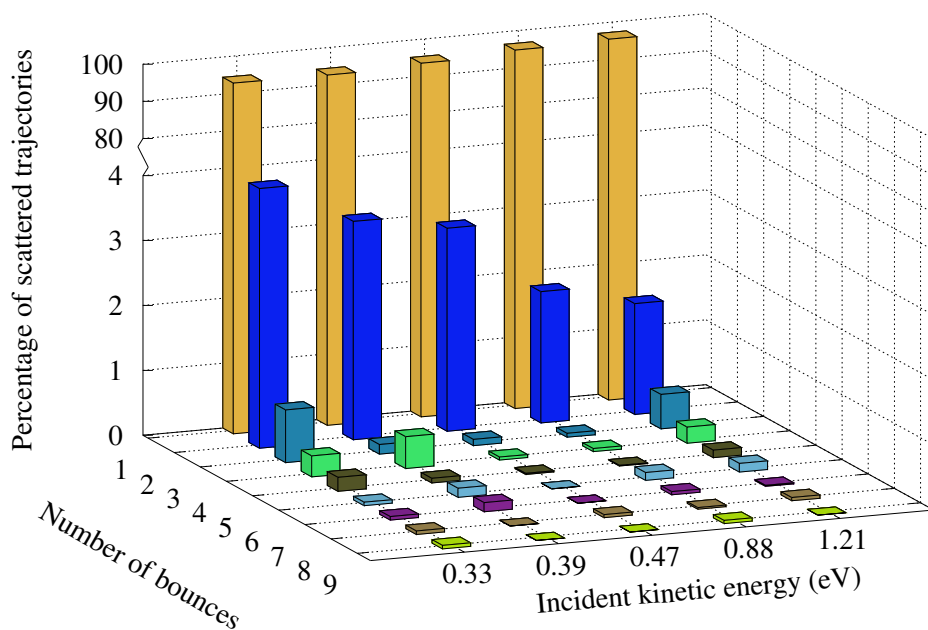


Figure 4.5: Number of bounces as a function of collision energy.

depend on incidence parameters.<sup>117</sup> And yet the contribution of trapping-desorption trajectories to total rotational energy distribution of formaldehyde is hardly significant as the number of trajectories is very low.

Figure 4.7 shows the distribution of direct and indirect scattered formaldehyde. It is evident from the figure that the trajectories exhibiting multiple bounces tend to leave the surface with much higher rotational energies.

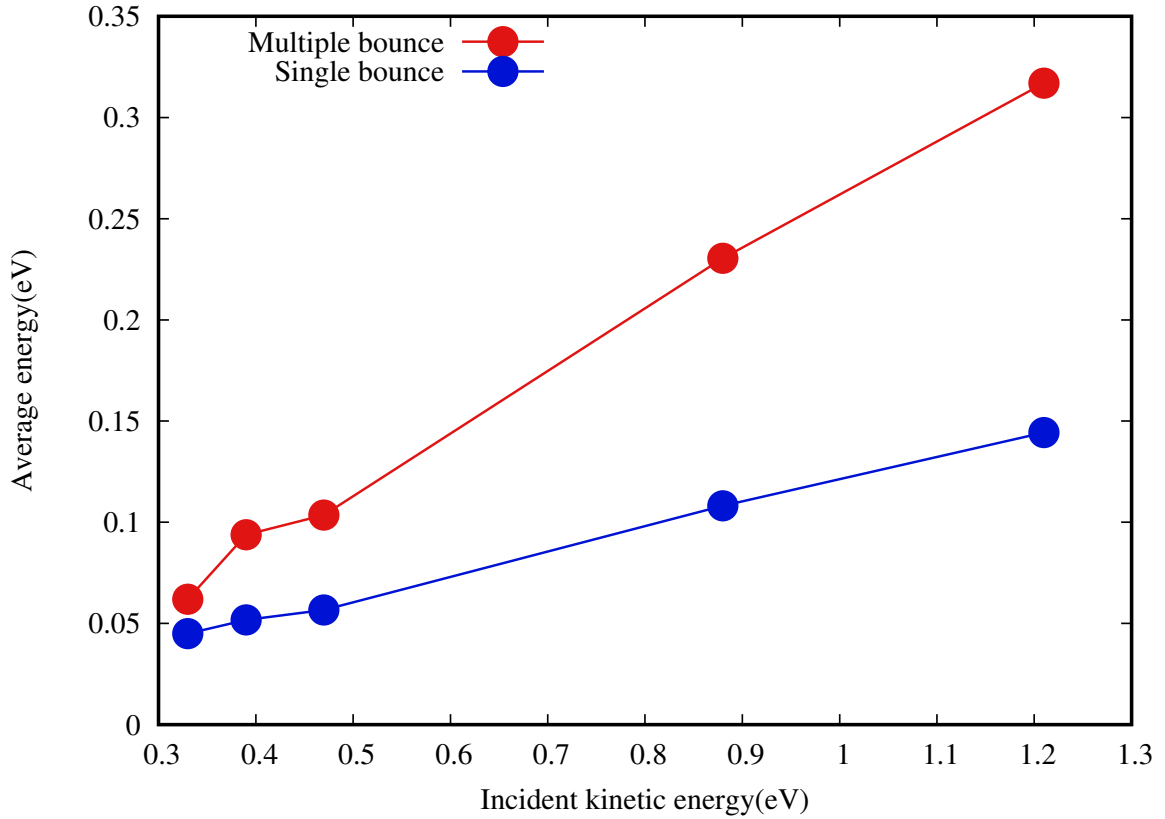


Figure 4.6: Average rotational energy of trapping-desorption (multiple bounce) trajectories are shown in red and direct-scatter trajectories (single bounce) are shown in blue.

### 4.2.2 Effect of Temperature

To understand the role of the surface temperature ( $T_s$ ) on the dynamics of rotational excitation of formaldehyde, we simulated trajectories for different surface temperatures.

Figure 4.8(a) shows the  $E_{\text{rot}}$  as a function of collision energy for different surface temperatures. For low collision energy ( $T_s$  of 140 K) the average rotational energy is the lowest and increases almost linearly with the temperature. This clear distinction washes away as we increase the collision energy. Figure 4.8(b) shows the average of individual components of rotational energy. It can be clearly seen that  $\alpha$ -axis rotation dominates through out the surface temperature range sampled in this study. Figure 4.8(c) shows that the trapping probability is mostly independent of the surface temperature. The slight rise in trapping

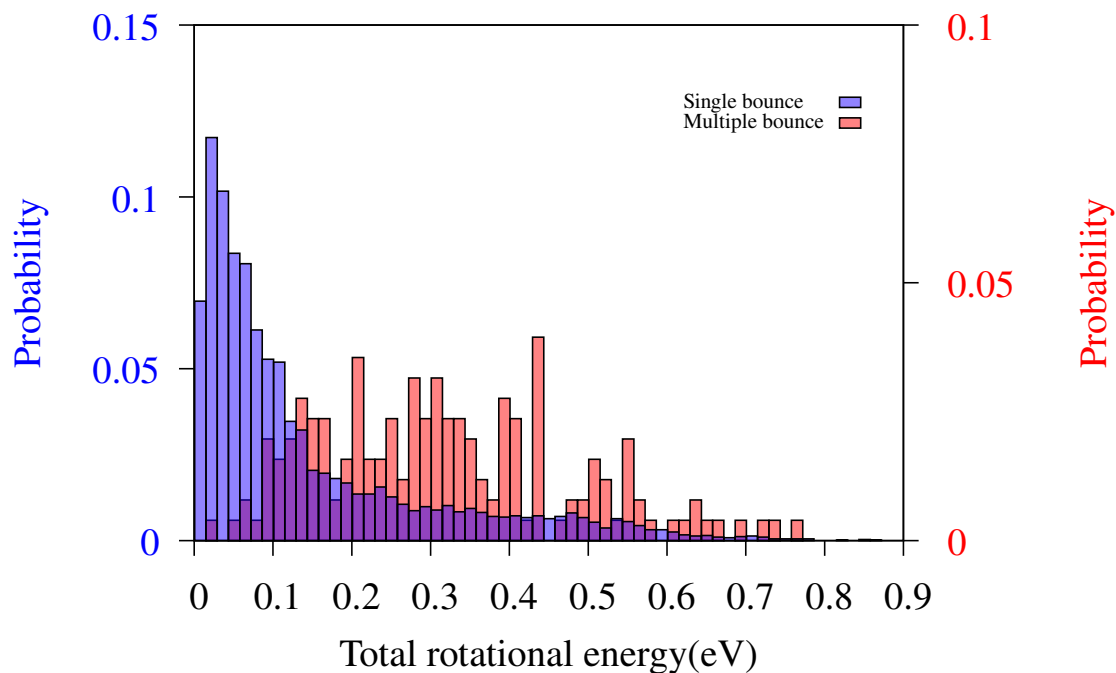


Figure 4.7: Rotational energy distribution of trapping-desorption (multiple bounce) trajectories (shown in red) and direct-scatter (single bounce) trajectories (shown in blue) for 1.21 eV of collision energy and  $T_s = 300$  K. Please note that the populations are not additive and each distribution is normalized independently for clarity.

probability seen for  $T_s = 400$  K at 1.21 eV of collision energy was checked by doubling the number of trajectories, and it still remained. Population distribution of rotational energies of formaldehyde are also very similar for different surface temperatures (not shown).

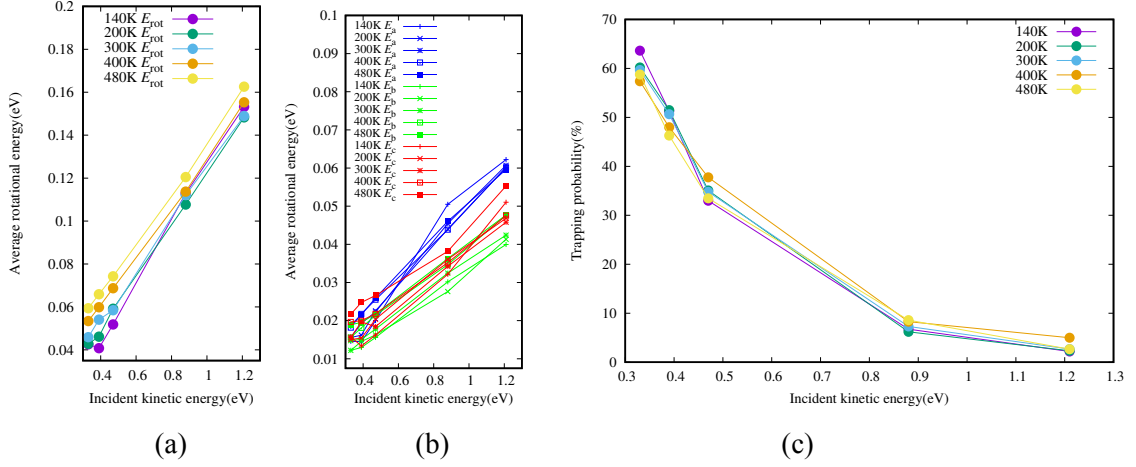


Figure 4.8: Effect of surface temperature on (a) average rotational energy, (b)  $a$ -axis(blue),  $b$ -axis(green) and  $c$ -axis(red) rotational energies, and (c) the trapping probabilities. The maximum error associated with trapping probabilities is 2.1% for  $T_s = 200$  K and collision energy of 0.39 eV.

### Scattering Angle

Scattering angle is defined as the angle subtended by the ejected molecule with respect to the surface normal ( $\vec{n}_1$ ). Figure 4.9 shows the scattering angle for all surface temperatures and collision energies considered in this study. It can be seen that the scattering angle for high collision energies are lower than that of low collision energies. The mean scattering angle for 0.33 eV collisions, regardless of surface temperature, is around  $30^\circ$ , whereas the mean scattering angle for 1.21 eV is around  $15^\circ$ . Interestingly, low collision energies have a higher variation of scattering angle.

#### 4.2.3 Effect of orientation

Five different orientations used in this study to see what could be the source of the propensity for  $a$ -axis rotation is shown in Figure 4.10. The orientations named OR<sub>1</sub> to OR<sub>5</sub> are chosen in such a way that different orientations are expected to lead to different dynamics. 1024 trajectories were initiated for each collision energy, for each orientation. Also three  $T_s$

were sampled, which are 200, 300 and 400 K. For these set of trajectories the rotational temperature was chosen to be 0 K.

Figure 4.11(a) shows the trapping probabilities of different orientations. It can be seen that orientation plays an important role in the dynamics with respect to trapping probabilities. At lower collision energies these effects are more pronounced. Figure 4.11(b) shows the average trapping probabilities for the five orientations compared to the set of trajectories for random orientation,  $T_s = 300$  K, which shows a similar trend.

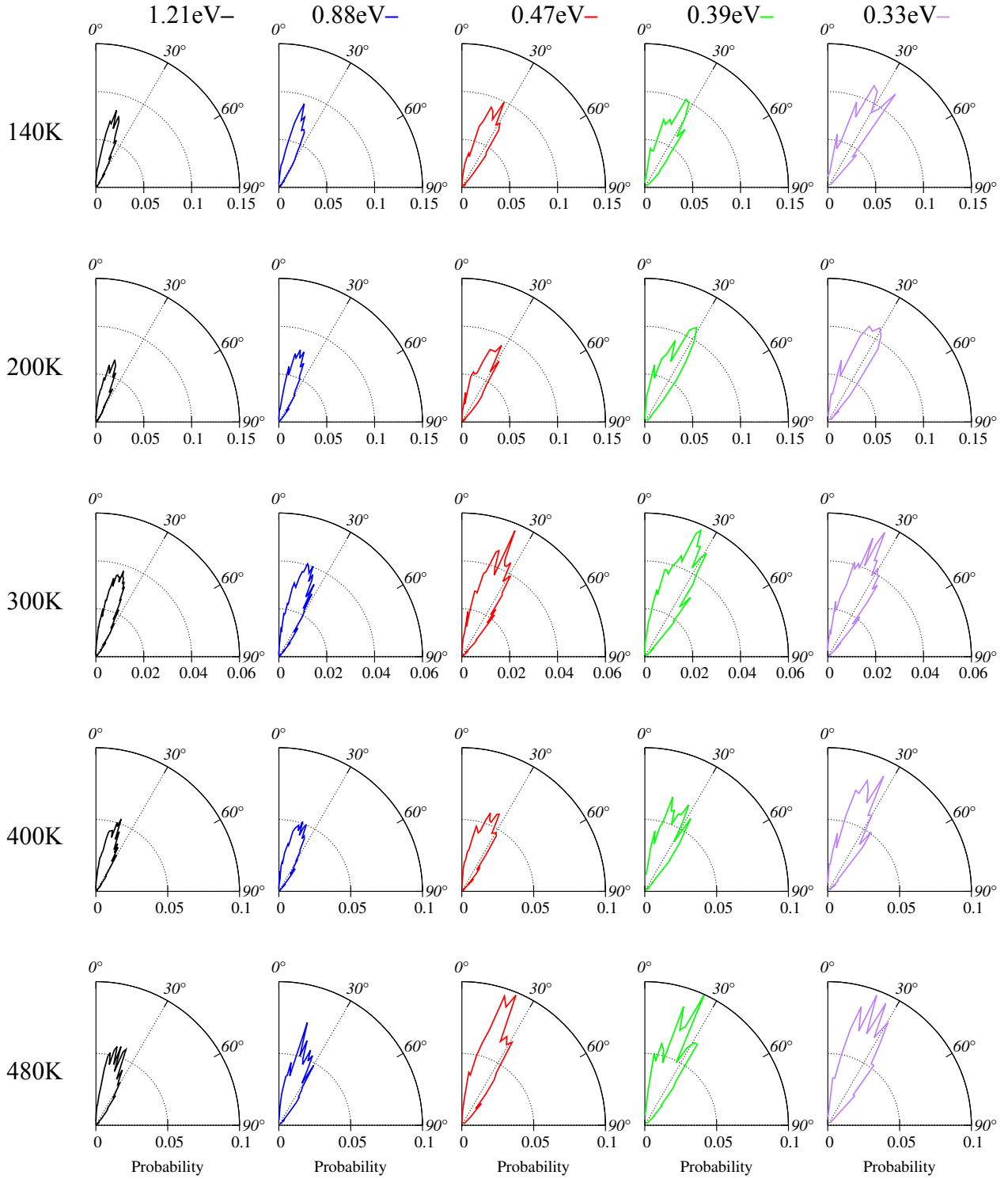


Figure 4.9: Scattering angle distribution for random initial orientation of formaldehyde for all temperatures and collision energies.

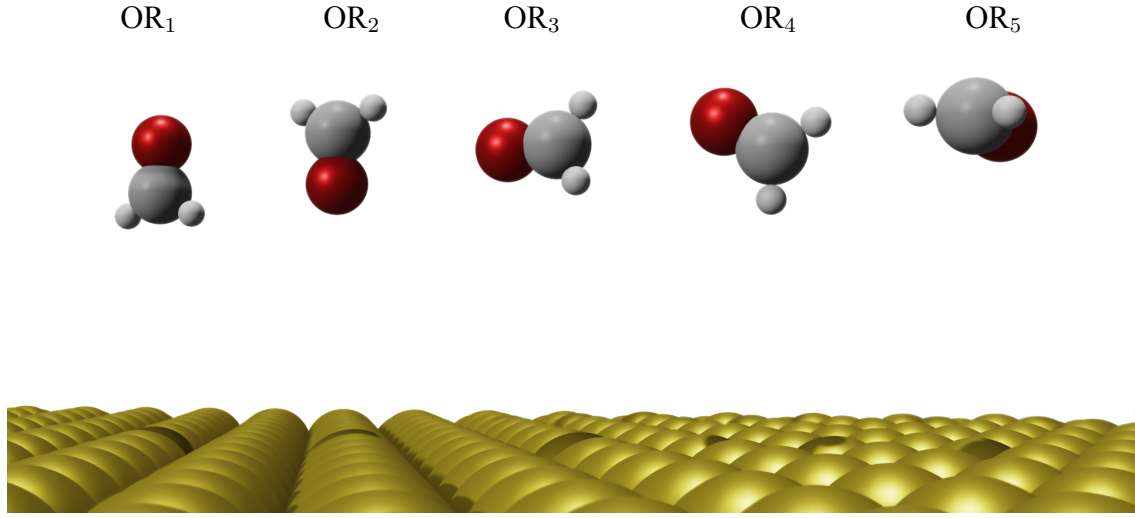


Figure 4.10: Different orientations used in this study.

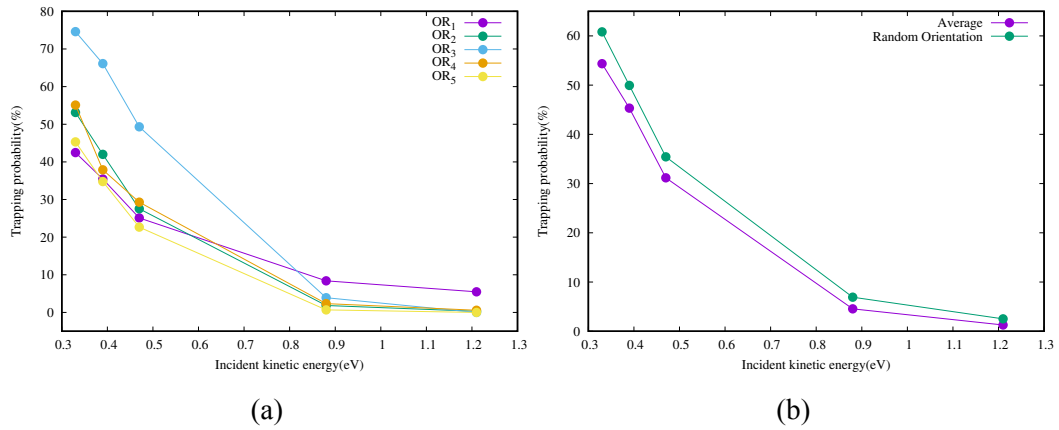


Figure 4.11: (a) Trapping probabilities as a function of collision energy for different orientations and (b) trapping probabilities as a function of collision energy averaged over different orientations compared to that of random orientation for  $T_s = 300$  K.

Figure 4.12 (a) shows the average rotational energy for different orientations as a function of collision energy.  $OR_1$ , in which oxygen atom is facing away from the gold surface has the highest average rotational energy at high collision energies. This is attributed to the high rotational excitation of  $b$ -axis and  $c$ -axis, as shown in Figure 4.12(b). For  $OR_5$ , where the formaldehyde molecule is parallel to the surface, the total average rotational energy decreases at higher collision energies. From Figure 4.12, it is evident that two orientations,

OR<sub>3</sub> and OR<sub>4</sub> contribute to *a*-axis rotation and OR<sub>1</sub> is responsible for high energy tail for total average rotational energy.

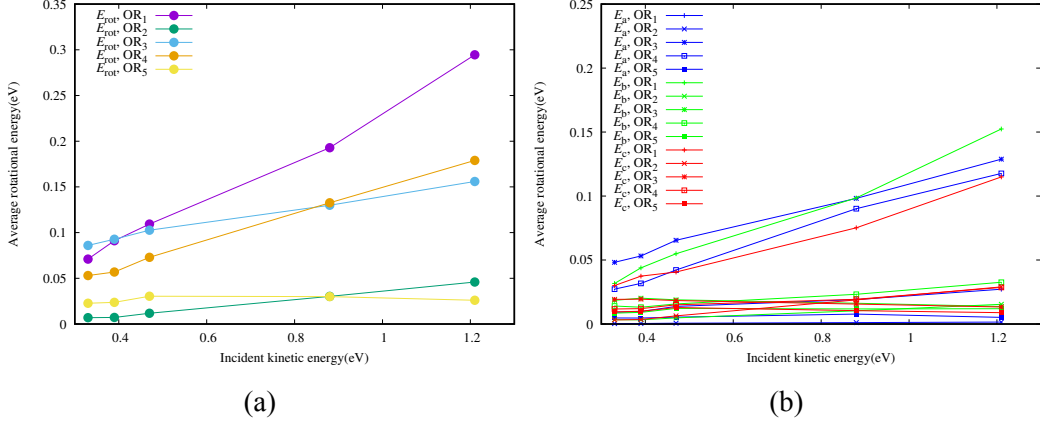


Figure 4.12: Effect of orientation on (a) total average rotational energy, (b) *a*-axis, *b*-axis and *c*-axis rotational energies, for  $T_s = 300$  K.

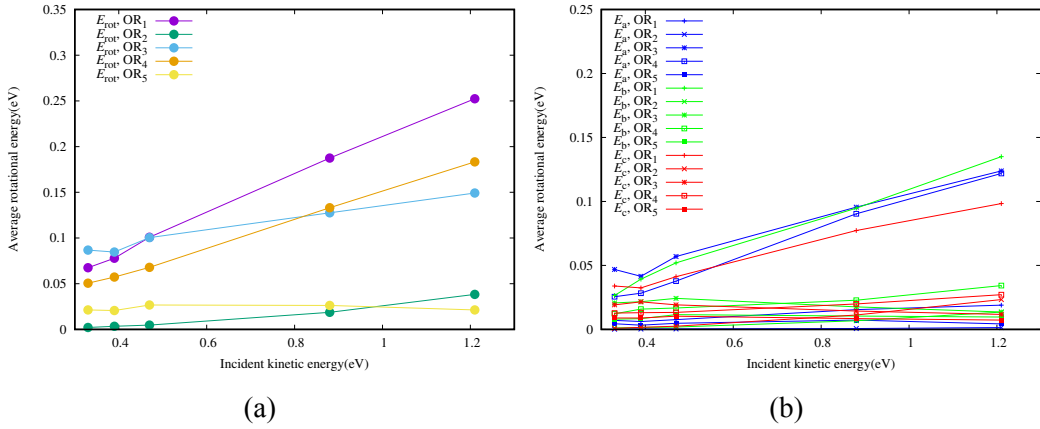


Figure 4.13: Effect of orientation on (a) total average rotational energy, (b) *a*-axis, *b*-axis and *c*-axis rotational energies for  $T_s = 200$  K.

We carried out trajectories to understand orientation effect for  $T_s = 200$  K and  $T_s = 400$  K also (see Figures 4.13 and 4.14). Irrespective of the surface temperature we see that average rotational energies for different orientations, change as a function of collision energy. At higher collision energies, OR<sub>1</sub> gives rise to high rotational excitation, while for low collision energies, OR<sub>3</sub> seems to dominate.

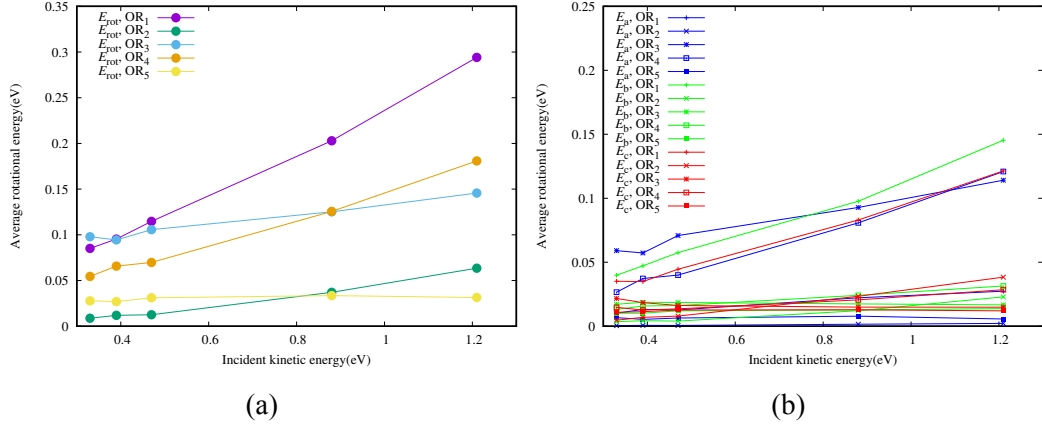


Figure 4.14: Effect of orientation on (a) total average rotational energy, (b)  $a$ -axis,  $b$ -axis and  $c$ -axis rotational energies for  $T_s = 400$  K.

#### 4.2.4 Energy Transfer

To understand the source of the rotational energy excitations, we looked at the energy transfer of formaldehyde to the gold surface. When the formaldehyde is far away from the gold surface, their total energies are independent of each other. The total energy of formaldehyde is given as,

$$E_{\text{Total}} = V(r_1, \dots, r_6) + K \quad (4.7)$$

$$E_{\text{Total}} = V(r_1, \dots, r_6) + E_{\text{rot}} + E_{\text{trans}} + E_{\text{vib}}$$

where  $V(r_1, \dots, r_6)$  is the potential energy of formaldehyde,  $E_{\text{rot}}$  is the total rotational energy for formaldehyde,  $E_{\text{trans}}$  is the kinetic energy of the center of mass of formaldehyde and  $E_{\text{vib}}$  is the vibrational energy. We calculated all the components of the total energy before the collision and compared them with the components after the collision, when formaldehyde is not interacting with the surface anymore. The difference among the individual components gives insight to the source of rotational excitation. Figure 4.15 shows the different distributions for energy transfer for  $T_s = 300$  K and 1.21 eV collision energy, for formaldehyde initiated with random orientation.

We can see that formaldehyde always loses vibrational energy (green) to the gold sur-

face. Interestingly, the blue curve which depicts the loss in translational energy of formaldehyde matches in magnitude to the gain in rotational energy. For different orientations, i.e., OR<sub>1</sub> to OR<sub>5</sub>, we see very similar trends (not shown). The change in vibrational energy of formaldehyde was found to be independent for all orientations, other than OR<sub>5</sub>, where we see that the scattered formaldehyde molecules are somewhat vibrationally cold.

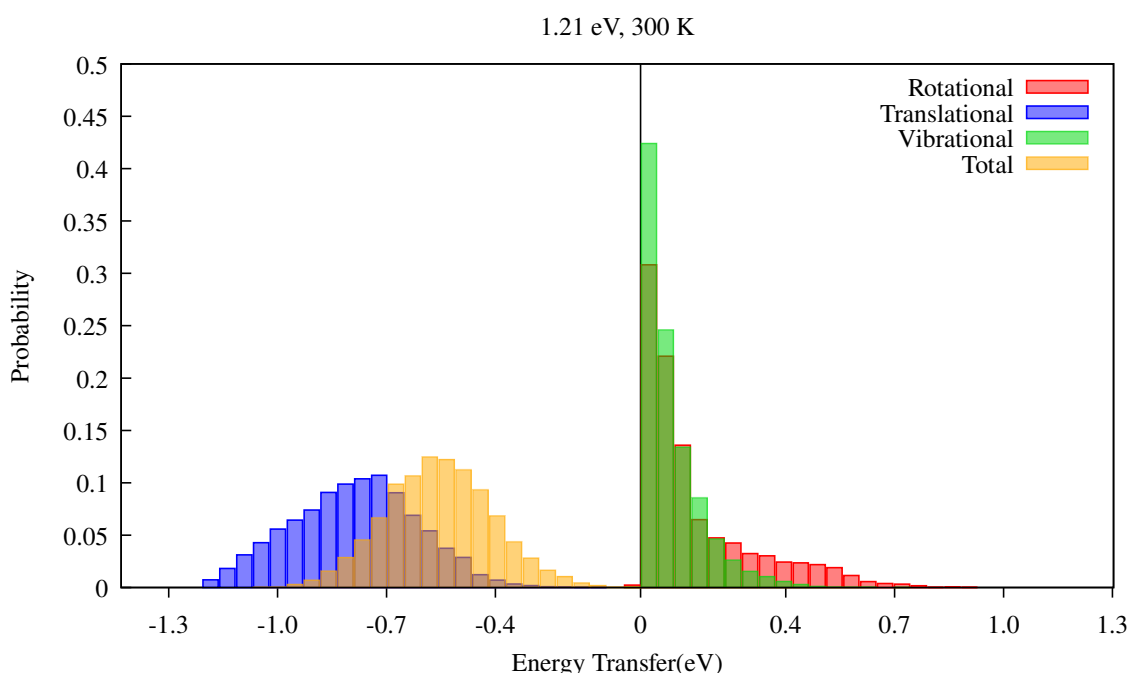
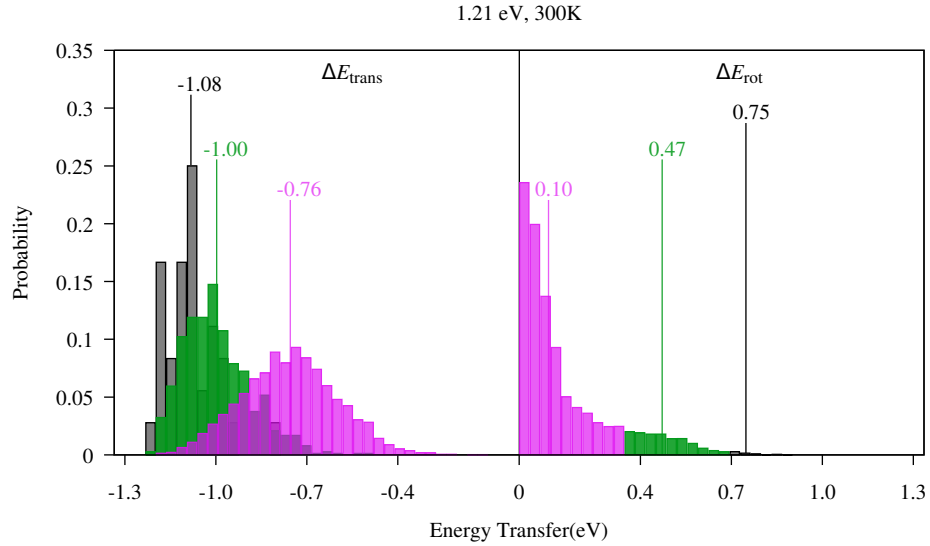


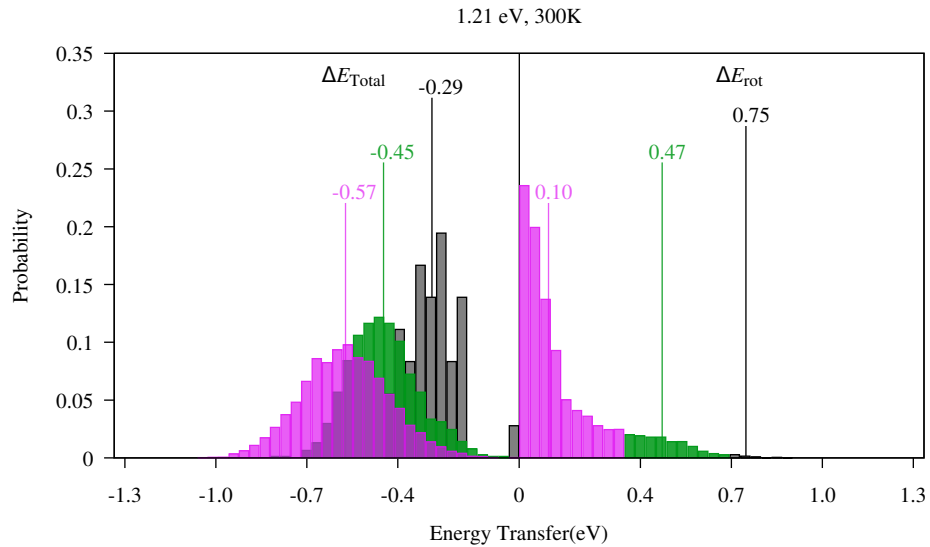
Figure 4.15: Distribution of energy transfer between formaldehyde and the gold sheet for 1.21 eV of collision energy and  $T_s = 300$  K, calculated from trajectories, starting with random orientation. Black line in the middle corresponds to 0 eV transfer.

To confirm that the exact source of the rotational energy excitation was in fact the initial kinetic energy of the incident formaldehyde molecule, we looked at the correlation between  $E_{\text{trans}}$  and  $E_{\text{rot}}$  as shown in Figure 4.16. We divided the rotational distribution into three groups, with average rotational energy gain of 0.1 eV, 0.48 eV and 0.76 eV. For these three groups we plotted the gain in rotational energy with loss in translational energy (Figure 4.16(a)), and loss in total energy (Figure 4.16(b)). It becomes evident that as the formalde-

hyde rotational excitation increases, it is more likely that the formaldehyde would have lost translational kinetic energy. It should be pointed out that these results are independent of the grouping considered here. Figure 4.17 shows the same analysis for OR<sub>1</sub> and similar trends are observed.



(a)



(b)

Figure 4.16: (a) shows gain in rotational energy of formaldehyde compared to the loss in translational energy and (b) shows gain in rotational energy of formaldehyde compared to the loss in total energy for  $T_s = 300$  K. The gain in rotational energy is divided into 3 regions marked by the color pink, yellow and black. For the same trajectories the graph on the left shows the corresponding loss in translational energy(a), and (b) Total energy.(Random orientation).

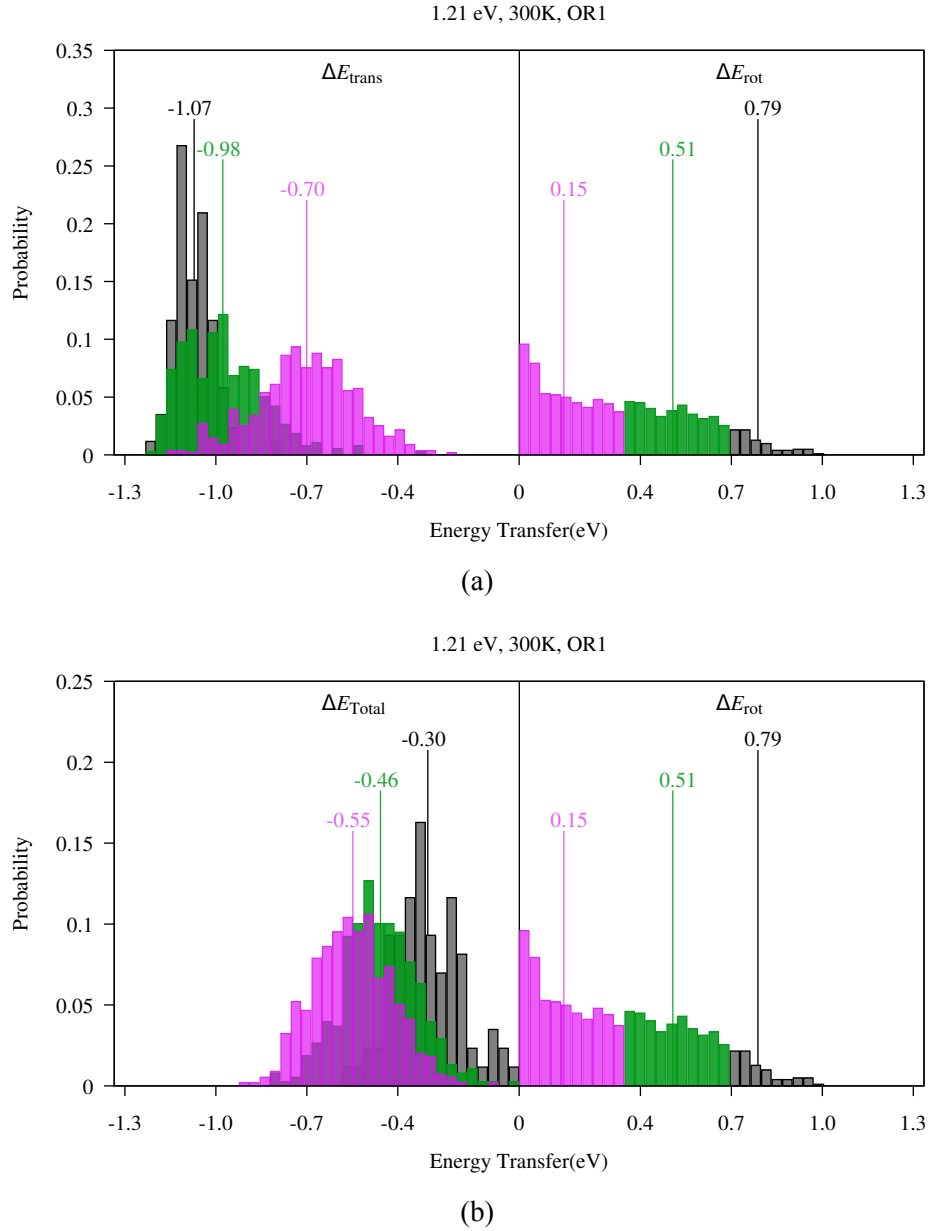


Figure 4.17: (a) shows gain in rotational energy of formaldehyde compared to the loss in translational energy and (b) shows gain in rotational energy of formaldehyde compared to the loss in total energy for  $T_s = 300$  K. The gain in rotational energy is divided into 3 regions marked by the color pink, yellow and black. For the same trajectories the graph on the left shows the corresponding loss in translational energy(a), and (b) Total energy.(OR<sub>1</sub>).

### 4.2.5 Mechanism of Energy Transfer

To understand the importance of the orientation on rotational excitation and to deduce the mechanism of energy transfer, we generated the interaction potential of formaldehyde with that of the gold surface (for the equilibrium geometries of formaldehyde and gold). The coordinate system used for the generation of the interaction potential is given in Figure 4.18. Here,  $\beta$  is the angle between the normal of the plane containing CH<sub>2</sub> group ( $\vec{n}_2$ ) and surface normal ( $\vec{n}_1$ ),  $\gamma$  is the angle between vector  $\vec{CO}$  and surface normal ( $\vec{n}_1$ ) and  $C_z$  the distance of center of mass of formaldehyde from the surface. We assume that the geometries of gold and formaldehyde do not change much from their equilibrium structures during the dynamics.

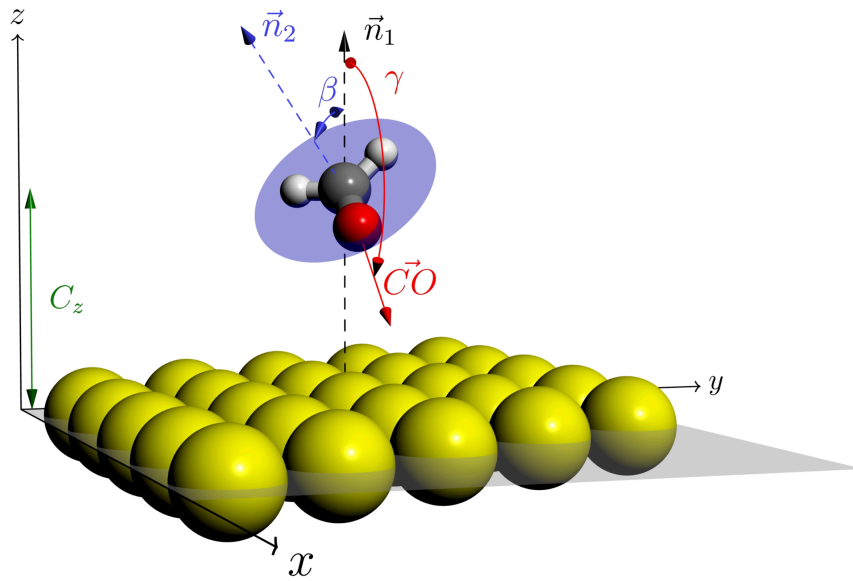


Figure 4.18: Illustration of  $\beta$ ,  $\gamma$  angles and  $C_z$  distance.  $\vec{n}_1$  is the gold surface normal,  $\vec{n}_2$  is the normal of the surface containing CH<sub>2</sub> (denoted by blue).  $\beta$ (blue) is the angle between  $\vec{n}_1$  and  $\vec{n}_2$ ,  $\gamma$ (red) is the angle between  $\vec{n}_1$  and  $\vec{CO}$  and  $C_z$ (green) is the distance of center of mass from the gold surface.

The interaction potential  $V_{\text{int}}(\beta, \gamma, C_z)$  is shown in Figure 4.19 (a). Green colour rep-

resents negative  $V_{\text{int}}$  values, whereas red colour represents positive values, and black corresponds to  $V_{\text{int}} = 0$  eV. When the distance of formaldehyde from the surface is large,  $V_{\text{int}}$  goes to 0 eV. But as it decreases, we can see that the different orientations feel different interaction energies.

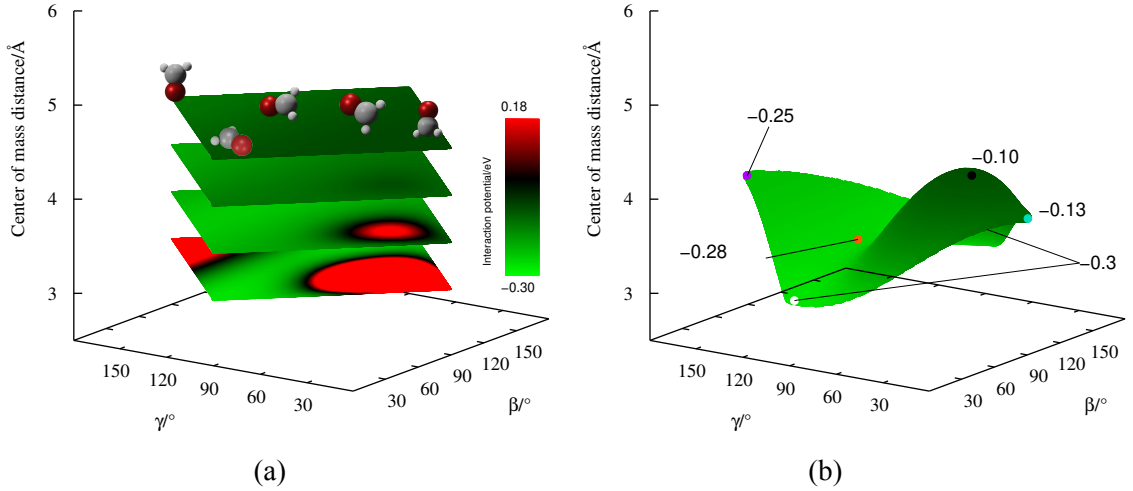


Figure 4.19: Interaction potential as a function of  $\beta$ ,  $\gamma$  and  $C_z$ . (a) Four contour plots, starting from the bottom, of the interaction potential at  $C_z$  values of 3.7, 4.2, 4.7 and 5.2 Å are shown. (b) The minimum energy surface for all possible orientations.

Figure 4.19 (b) shows the minimum energy surface for all possible orientations. The energy spanned by the minimum energy surface goes from  $-0.1$  eV ( $\beta = 90^\circ$ ,  $\gamma = 36^\circ$ ,  $C_z = 4.3$  Å) to  $-0.3$  eV ( $\beta = 178.8^\circ$ ,  $\gamma = 88^\circ$ ,  $C_z = 3.19$  Å). The iso-energetic surface corresponding to 1.21 eV, which is the “repulsive wall”, happens to be almost identical in shape to the minimum energy surface, but about 1.5 Å lower than it (not shown).

To investigate the mechanism of rotational rainbow, we divided the scattered trajectories for  $T_s = 300$  K, starting with random orientations into two sets. Set 1 has trajectories in which  $E_a \geq (E_b + E_c)$ , and rest of the trajectories are in Set 2. Figure 4.20 shows the starting orientation of the two sets. Set 1 has 47% of scattered trajectories, whereas Set 2 has 53%. We then plotted the orientations at the point of closest approach, of these two sets of trajectories as shown in Figure 4.21. We can clearly see that as the formaldehyde comes

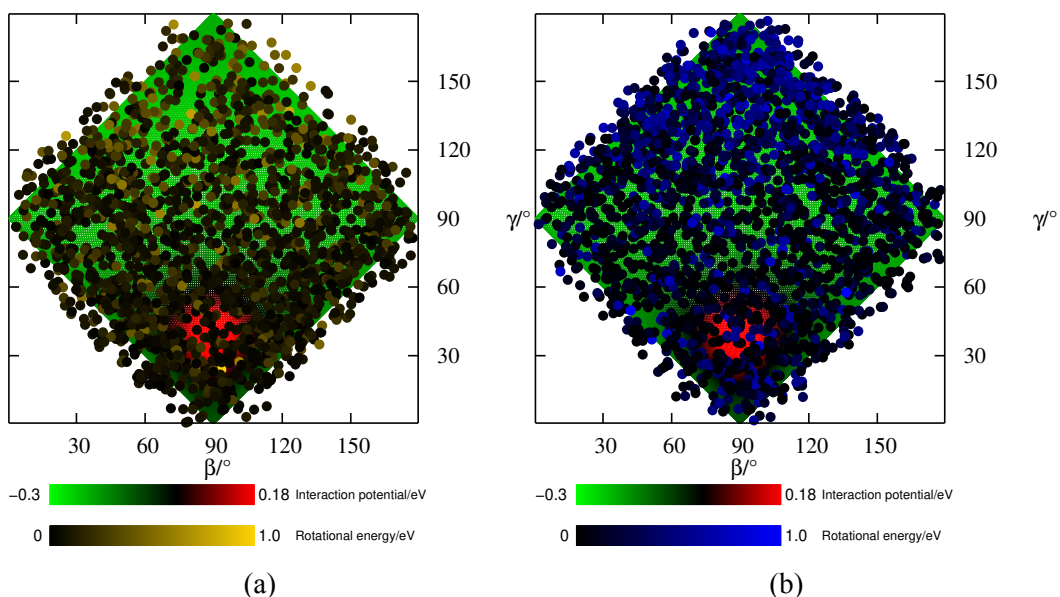


Figure 4.20: The initial orientations at the starting of the trajectory (dots) projected on the surface for interaction potential of formaldehyde with the gold surface at a distance of 3.7 Å for trajectories where (a)  $E_a \geq E_b + E_c$  and (b)  $E_a < E_b + E_c$ .

closer to the gold surface, only some orientations are selected, which can be attributed to the shape of the minimum potential energy surface in Figure 4.19(b). Additionally, the colour of the dots represent  $E_{\text{rot}}$ , which gives us information about the underlying mechanism of rotational excitation. We can see that most of the trajectories with low rotational excitation, approach the minimum energy region with an orientation similar to that of OR<sub>5</sub>. For the rest of the trajectories the orientation in which the oxygen atom is pointing towards the gold surface (OR<sub>2</sub>) is almost always selected.

To understand why  $a$ -axis rotation is higher, we looked at the dynamics, after the molecule has reached the point of closest approach. It should be pointed out that oscillations in  $\beta$  corresponds to  $a$ -axis rotation, and oscillations in  $\gamma$  angle or oscillations in a combination of  $\beta$  and  $\gamma$  angles correspond to  $b$ -axis or  $c$ -axis rotations. There are some orientations for which the  $(\beta, \gamma)$  combination will not work. For example if the formaldehyde is undergoing  $a$ -axis rotation but the  $\vec{\text{CO}}$  is exactly parallel or anti-parallel to the surface normal ( $\vec{n}_1$ ), then the  $\beta$

angle will remain constant as a function of time.

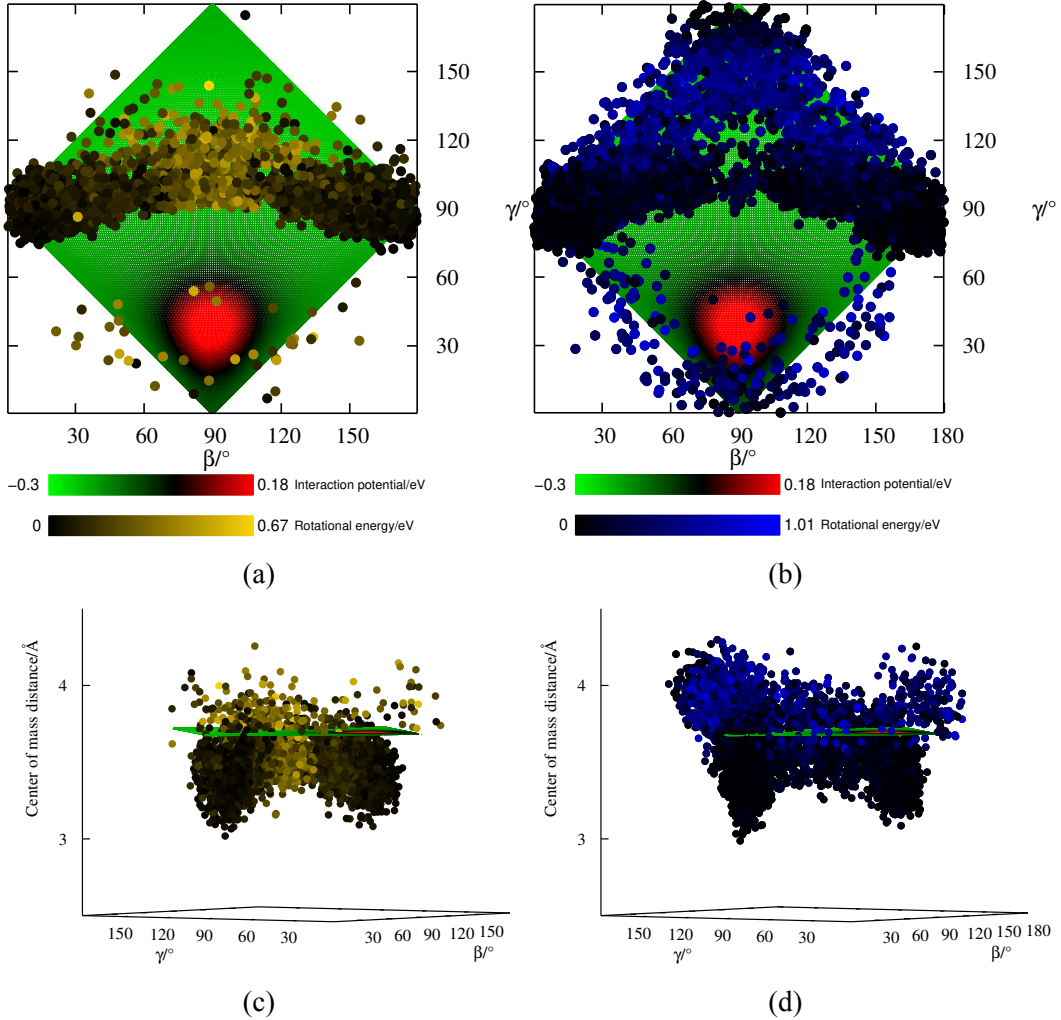


Figure 4.21: The dots represent the orientation at the point of closest approach, and the surface is the interaction potential of formaldehyde with the gold surface at a distance of 3.7 Å. (a) and (c) are for trajectories where  $E_a \geq E_b + E_c$  and (b) and (d) are for trajectories where  $E_a < E_b + E_c$ .

Figure 4.22 shows a few steps after the molecule has reached the point of closest approach, and is en route to scatter, for those trajectories in which the amount of rotational excitation is greater than 0.3 eV. Part (a) of Figure 4.22 is for cases where  $a$ -axis rotation is more, and we see that the orientation shifts towards lower  $\gamma$  value, which corresponds to  $c$ -axis rotation, but before the molecule can exit, it is affected by the barrier that lies at

$\gamma = 36^\circ$  ( $\approx 0.18$  eV).

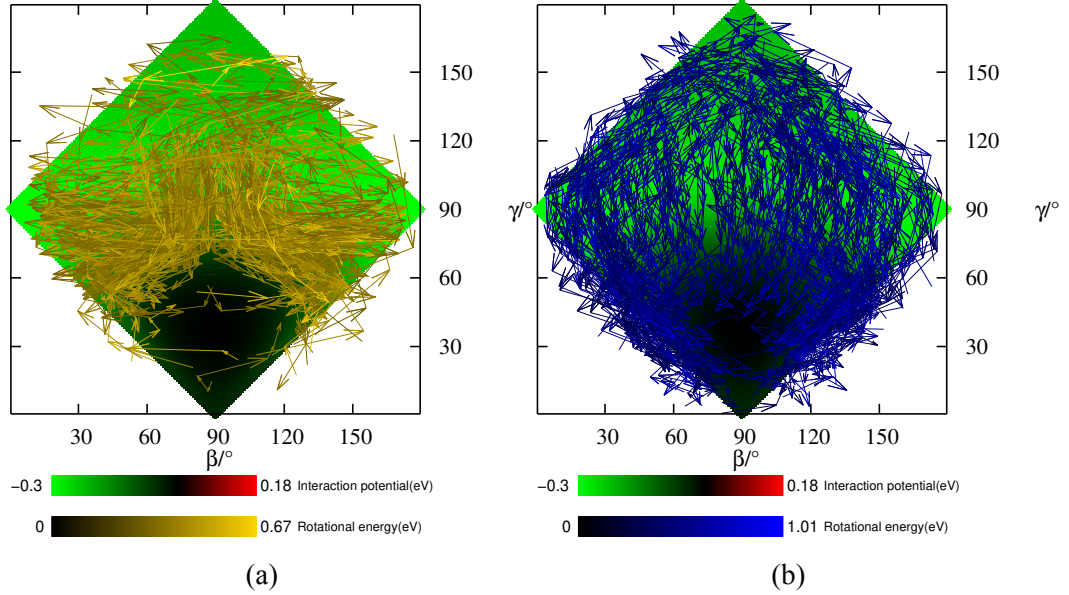


Figure 4.22: Dynamics after the point of closest approach for the two sets (a) Set 1 (b) Set 2, projected onto the minimum energy surface. All the trajectories shown here have  $E_{\text{rot}} > 3$  eV. Only 200 trajectories for each set are plotted for clarity. Each arrow signifies a timestep of 2 fs, and 10 fs worth of trajectories have been plotted after the formaldehyde reaches the point of closest approach.

The barrier induces a torque on formaldehyde along the CO-axis, converting the  $c$ -axis angular momentum to  $a$ -axis rotation (oscillations along  $\beta$ ). In part (b) of Figure 4.22, we see that  $\beta$  and  $\gamma$  start changing at the same time, and gives rise to  $b$ -axis rotation, or the oscillations in  $\gamma$  take over, but by then the molecules are already far enough from the surface to be affected by the barrier. This explains the near 50% probability of getting  $a$ -axis rotation.

### 4.3 Summary and Conclusions

In this study, we model and simulate the scattering of formaldehyde from gold(111) surface using classical mechanics. We could reproduce the trends seen in experiment,<sup>113</sup> both in terms of rotational energies as well as trapping probabilities. Moreover the translational

energies of scattered formaldehyde also matched with the experiments.<sup>117</sup> We classified trajectories based on the number of bounces as direct scattered or indirect scattered and we found that indirect scattered trajectories had higher rotational energies. After that we looked in to the effects of surface temperature, and we found that the final product state distributions were independent of the surface temperature. We also looked at the orientation effect, and it turns out to play an important role in deciding the final state distributions. Further, we looked at the energy transfer of formaldehyde to the surface, and it was found that the source of the rotational excitation was the initial kinetic energy of formaldehyde. Finally, we looked at the rotational dynamics and we could identify a mechanism which shows why the scattered formaldehyde molecules shows rotational rainbow in *a*-axis. It turns out that most of the scattered trajectories at the point of closest approach, have oxygen atom pointed towards the gold surface. Around 50% of these trajectories move towards a barrier in the interaction potential and then are deflected by the barrier, giving rise to *a*-axis rotation. Rest of the trajectories end up having either *b*-axis rotation or *c*-axis rotation as the dominant rotational mode.

In conclusion, we could understand the mechanism of rotational rainbow arising from the interaction of polyatomic gas molecule with a metal surface at the atomistic level. The interaction of the gas molecule with the minimum energy surface seems to give rise to the rotational rainbow. During collision, the formaldehyde reorients to access certain regions of the PES following the minimum energy path. Such orientational steering has been reported in earlier work for CO<sup>126</sup> and NH<sub>3</sub>.<sup>130</sup> Furthermore, the transfer of energy from translational to rotational mode, may help in trapping of such molecules when gold is used as a catalyst and the surface involved is coarse or irregular.

## Chapter 5

# Gas-Surface dynamics: Formaldehyde on graphene surface

To continue our investigation of gas-surface interaction, we looked into the rotational excitation of formaldehyde upon scattering from a single layer graphene (SLG) sheet. Graphene has been used as a surface in a number of surface scattering experiments. Inelastic scattering of hydrogen atoms from graphene sheet has been studied by Wodtke et al,<sup>136</sup> and they find that the dynamics of adsorption of hydrogen atoms to the graphene surface involves atleast four carbon atoms. Hydrogen atom colliding with graphene sheet with defects has also been modeled computationally, by Lischka et al.<sup>137</sup> They reported the collision of hydrogen atoms around an already existing  $sp^3$  CH defect on a single layer graphene sheet. They found that despite the energy surface being globally repulsive, local channels can be created because of the defect on the surface, through which reactants (other hydrogen atoms) can move and react without a barrier. Recently, the scattering of NO from graphene sheet has been explored both experimentally<sup>138, 139</sup> as well as theoretically<sup>140</sup> by Koehler et al, and it was seen that the scattered NO exhibited rotational rainbow.

Graphene as surface is very different from gold, in terms of rigidity and molecular motion. Also, the interaction of formaldehyde atoms with graphene quite different from that of gold surface. In gold-formaldehyde, the interaction of oxygen atom with gold atoms was attractive, whereas for the graphene surface, the hydrogen atom facing the C-atoms is the most attractive.<sup>141</sup> Thus, the dynamics of formaldehyde on this surface is expected to be different.

In this chapter, the study of formaldehyde scattering from SLG sheet using classical

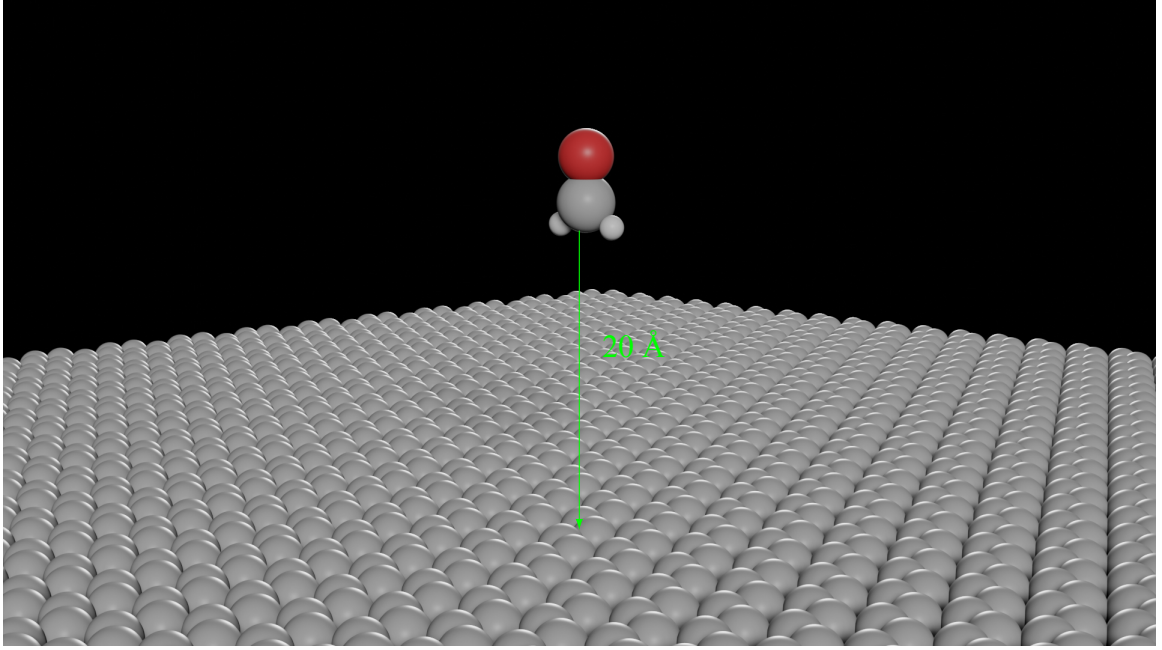


Figure 5.1: The graphene surface and the formaldehyde at the starting geometry. All atoms in this image are at 0 K, but for trajectories we equilibrate the surface to 300 K.

trajectory simulations is reported.

## 5.1 Potential Energy Model

The graphene was modelled using a single layer of size  $100 \times 100$  Å consisting of 3986  $sp^2$ -hybridized carbon atoms. Four atoms at the edges of the sheet were fixed during the course of the simulations. The system is shown in Figure 5.1. The potential of the system is given by,

$$V = V_{\text{Formaldehyde}} + V_{\text{Graphene}} + V_{\text{interaction}} \quad (5.1)$$

where  $V_{\text{Formaldehyde}}$  is the 6D potential of formaldehyde given by Bowman,<sup>101</sup>  $V_{\text{Graphene}}$  is the potential of graphene surface modelled using AMBER force fields and  $V_{\text{interaction}}$  is the interaction potential between graphene surface and formaldehyde. The graphene-formaldehyde interaction potential<sup>141</sup> was expressed as the sum of two-body interactions between the

atoms  $X$  of graphene sheet and the atoms  $Y$  of formaldehyde, given by,

$$V_{\text{interaction}} = \sum_X \sum_Y V_{X\dots Y} \quad (5.2)$$

where,

$$V_{X\dots Y} = Ae^{-Br} + f(r) \left[ \frac{C}{r^2} + \frac{D}{r^4} + \frac{E}{r^6} + \frac{F}{r^8} \right] \quad (5.3)$$

and,

$$f(r) = \begin{cases} e^{-\left(\frac{R}{r}-1\right)^2}, & r < R \\ 1, & r \geq R \end{cases} \quad (5.4)$$

The parameters used in the interaction potential are given in Table 5.1.

Table 5.1: Parameters for graphene-formaldehyde interaction.

$X\dots Y$	A	B	C	D	E	F	R
$C_p\dots C$	8305	2.87	1.00	-55.40	-145	-2184	2.4
$C_p\dots O$	42624	3.36	0.81	84.80	-1442	-3720	2.3
$C_p\dots H$	9719	3.50	-1.00	1.85	-222	-334	2.3

The SLG sheet represented using the AMBER<sup>142</sup> force field is given by,

$$V_{\text{Graphene}} = V_{\text{stretch}} + V_{\text{bend}} + V_{\text{dihedral}} \quad (5.5)$$

where  $V_{\text{stretch}}$  is the two-body potential given by,

$$V_{\text{stretch}} = \sum_{i=1}^{N_{\text{stretch}}} \frac{1}{2} f_r (r_i - r_0)^2 \quad (5.6)$$

$V_{\text{bend}}$  is the three-body potential given by,

$$V_{\text{bend}} = \sum_{i=1}^{N_{\text{bend}}} \frac{1}{2} f_{\theta} (\theta_i - \theta_0)^2 \quad (5.7)$$

and  $V_{\text{dihedral}}$  are the four-body potentials given by,

$$V_{\text{dihedral}} = \sum_{i=1}^{N_{\text{dihedral}}} \frac{k_d^n}{2} \left[ 1 + \cos(n\phi - \gamma_n) \right] \quad (5.8)$$

### 5.1.1 Trajectory Simulations

All trajectory calculations were done with velocity-Verlet algorithm using VENUS chemical dynamics software package.<sup>134</sup> The surface was equilibrated for 2 ps, with velocities rescaled every one fs, to a temperature of 300 K. Further, the surface was annealed for another 60 ps. The list of trajectories calculated for different conditions is given in Table B.1. We start the trajectories with formaldehyde randomly oriented, at a distance of 20 Å above the surface. The vibrational temperature and the rotational temperature of formaldehyde is set at 15 K sampled from a Boltzmann distribution which correspond to typical chamber conditions in molecular beam experiments for formaldehyde-gold system.<sup>113</sup> As in the case of formaldehyde-gold system, we sample five collision energies ( $E_{\text{coll}}$ ) i.e., 0.33, 0.39, 0.47, 0.88 and 1.21 eV. For each collision energy we run,  $\approx 2500$  trajectories for 10 ps. Further, to study the effect of orientations, three orientations were sampled which are OR<sub>1</sub>, OR<sub>2</sub>, OR<sub>5</sub> (shown below). In this case, the rotational temperature of formaldehyde was set to be 0 K and the vibrational temperature was set at 15 K, sampled from a Boltzmann distribution. For each of these orientations, we sampled the five collision energies of formaldehyde and for each combination, around 6000 trajectories were computed. In all the analysis reported here, if after 10 ps, the formaldehyde was less than 20 Å away from the surface we labelled the trajectories as trapped, else the trajectories were classified as scattered. The scattered trajectories were analyzed and their internal energies ( $E_{\text{trans}}$ ,  $E_{\text{rot}}$  and  $E_{\text{vib}}$ ) of formaldehyde were calculated.

## 5.2 Results and Discussion

The calculated trapping probability of graphene-formaldehyde system are compared to that of gold-formaldehyde system in Figure 5.2. From the figure it can be seen that the trapping probabilities are much higher than for the gold-formaldehyde system. Further, the decline

in trapping probability with increase in collision energy is nearly linear for this system. The large trapping probability of formaldehyde seen in the formaldehyde-graphene system is possibly due to strong interaction between formaldehyde and graphene and their higher propensity for attractive interaction due to the two hydrogen atoms of formaldehyde. Figure 5.3 shows the distribution of  $C_z$  for trapped formaldehyde. It can be seen that more than 96% of trapped formaldehyde are below 10 Å after 10 ps of trajectory simulation.

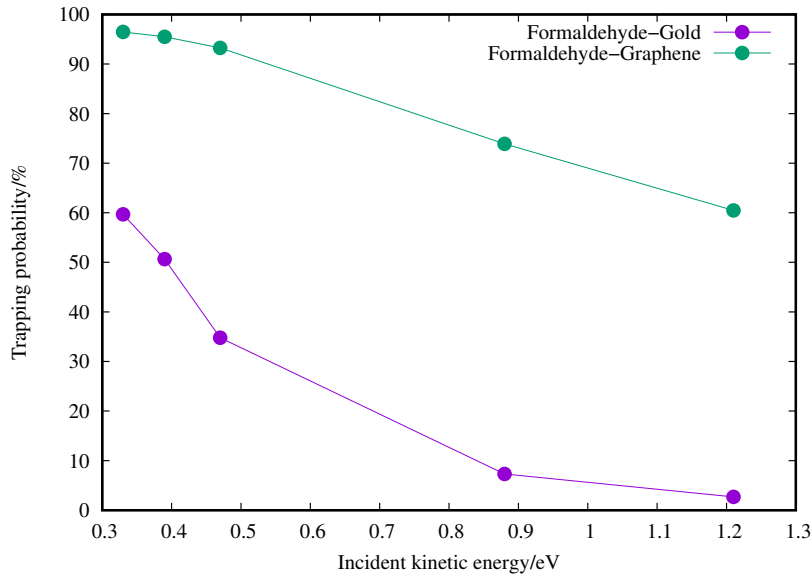


Figure 5.2: Trapping probability as a function of collision energy for formaldehyde-gold (purple) and formaldehyde-graphene (green) systems.

The average total rotational energies ( $\langle E_{\text{rot}} \rangle$ ), of scattered formaldehyde for different collision energies are shown in Figure 5.4. Despite having very different trapping probabilities,  $\langle E_{\text{rot}} \rangle$  follow a similar trend to that seen in the gold-formaldehyde system. However, the formaldehyde are rotationally hotter in the formaldehyde-graphene system and the contribution from each rotational differ for each collision energy. At 0.33 eV collision energy we see that  $a$ -axis rotation dominates and yet, above 0.88 eV the rotational rainbow can be seen in the  $b$  and  $c$  axis. The population distribution of rotational states for collision energy

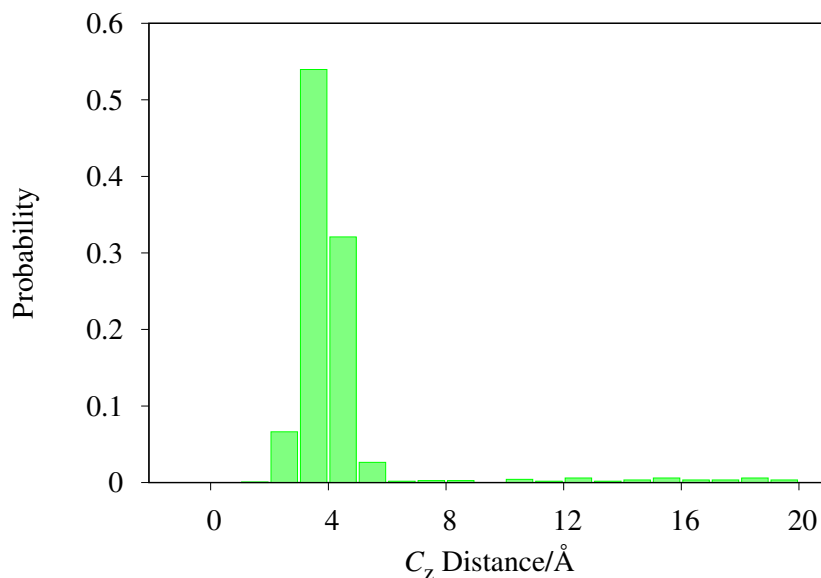


Figure 5.3: The distribution of  $C_z$  for the trapped formaldehyde.

of 1.21 eV, decreases as an exponential function, unlike in the case of gold-formaldehyde system where a gaussian hump could be seen in the total rotational energy distribution, centered at around 0.5 eV, see Figures 4.4 and 5.5. In the inset of Figure 5.5, population distribution of rotational states are shown for different axis of rotation. The  $a$ -axis rotation has more population with low energies, whereas  $b$ -axis and  $c$ -axis have longer tails with high rotational energies, which results in higher average rotational energies in those modes of rotation.

### 5.2.1 Direct vs Indirect Scattering

Similar to previous chapter, we classify trajectories as either trapping-desorption, when the formaldehyde bounces multiple times on the graphene sheet before leaving the surface, or as direct-scattered, where the formaldehyde bounces only once. Figure 5.6 shows the

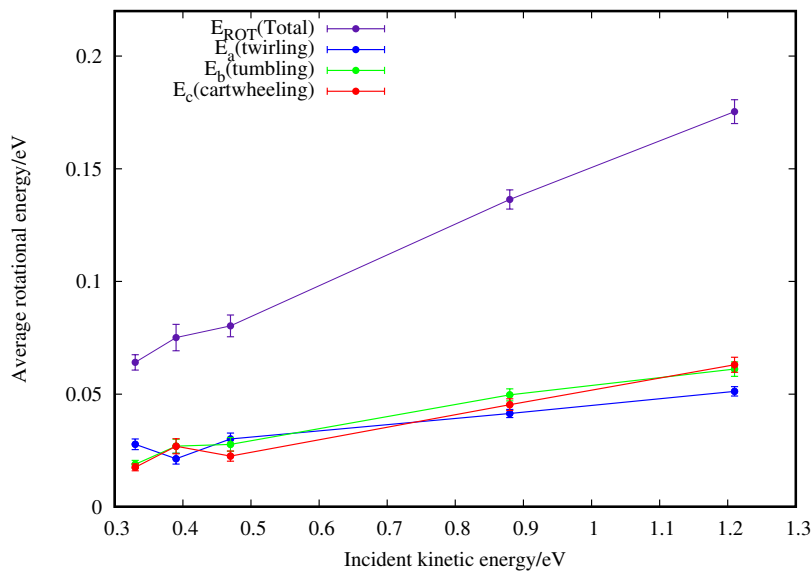


Figure 5.4: Average rotational energy of formaldehyde as a function of collision energy.

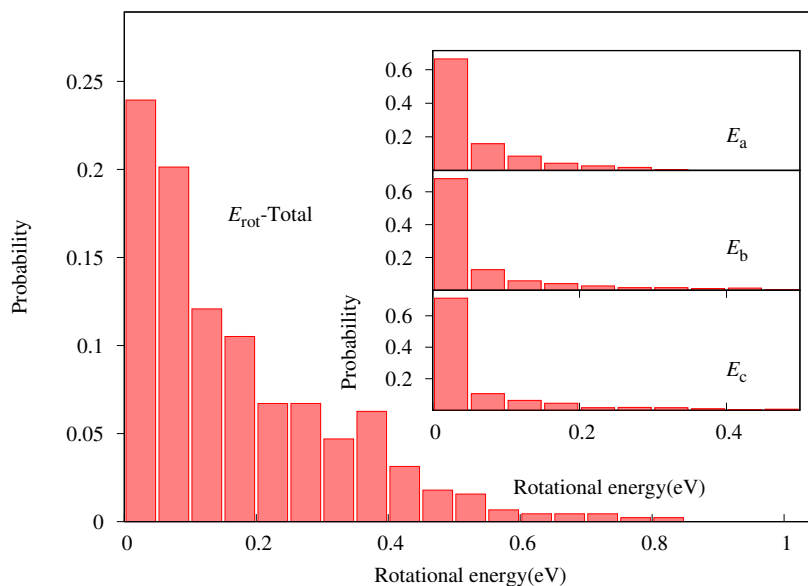


Figure 5.5: Population distribution of rotational energies of formaldehyde for 1.21 eV of collision energy. The insets show the population distribution of rotational states in each axis of rotation.

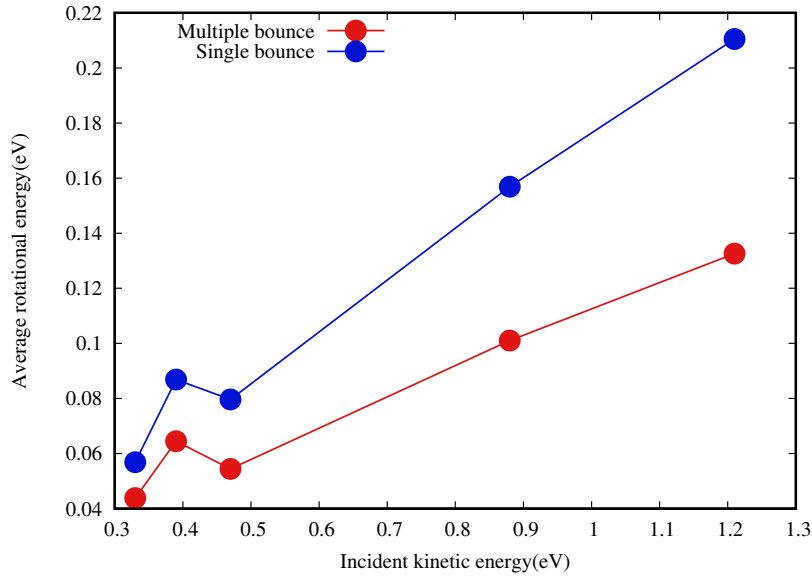


Figure 5.6: Average rotational energy of trapping-desorption (multiple bounce) trajectories shown in red, and direct-scattered trajectories (single bounce) are shown in blue.

average rotational excitation of formaldehyde for direct-scattered and trapping-desorption trajectories. Interestingly, for graphene-formaldehyde we see that the direct-scattered trajectories lead to higher rotational excitation of scattered formaldehyde. In contrast, in the gold-formaldehyde system we observed trapping-desorption trajectories to have higher rotational excitation.

### 5.2.2 Effect of Orientations

We simulated trajectories to understand the role of orientation of formaldehyde on the rotational excitation. The three orientations used are shown in Figure 5.7, and the corresponding trapping probabilities for these orientations are shown in Figure 5.8.

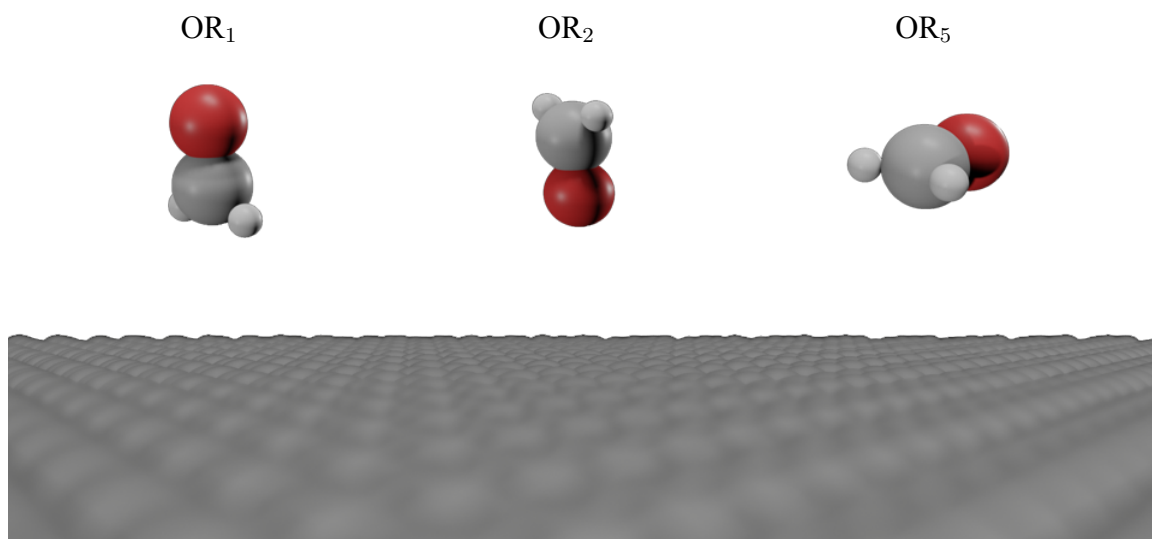


Figure 5.7: Different orientations used in this study.

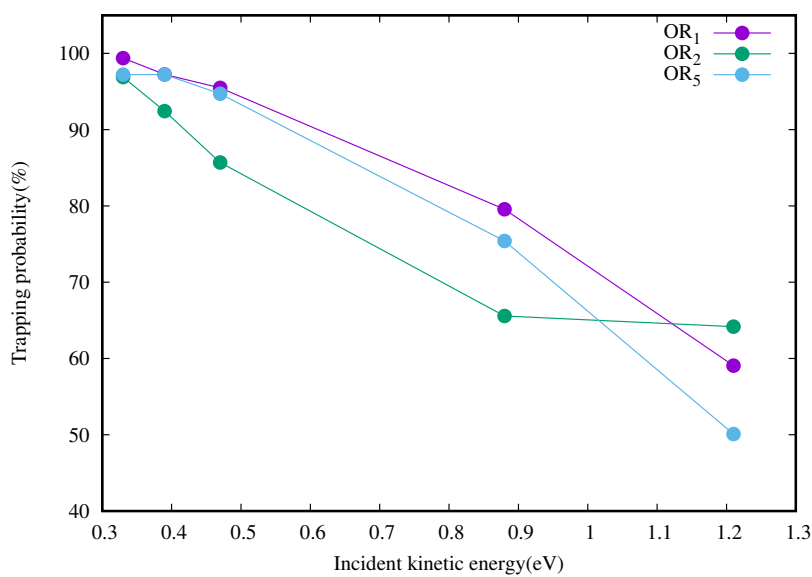


Figure 5.8: Trapping probabilities for different orientations.

We can see that OR<sub>1</sub> and OR<sub>5</sub> follow similar trend, while for OR<sub>2</sub>, the trapping probability shows a sharp decrease before levelling off. This is consistent with the attractive interaction felt in OR<sub>1</sub> and OR<sub>5</sub>. The average total rotational energies for these three ori-

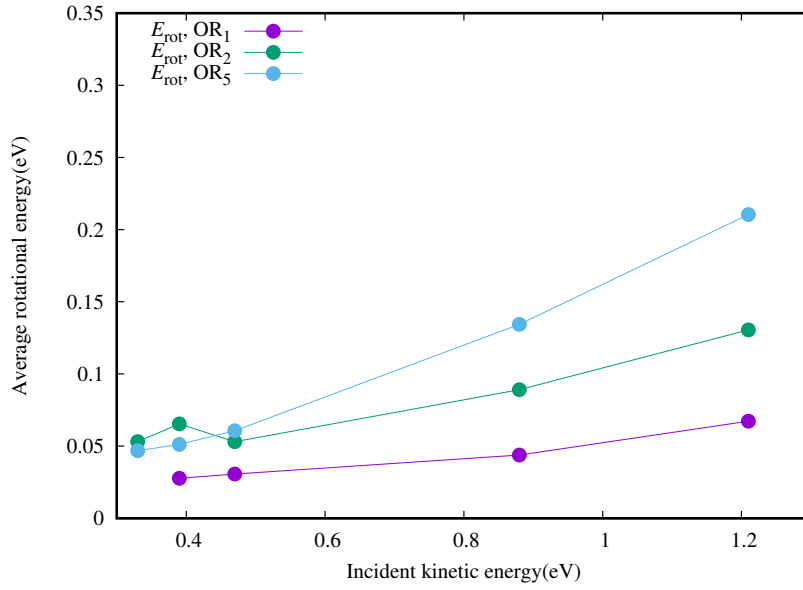


Figure 5.9: Average total rotational energies for different orientations.

entations are shown in Figure 5.9. OR<sub>1</sub> has the lowest rotational excitation throughout the collision energy range covered in this study, and at collision energies beyond 0.39 eV, OR<sub>5</sub> orientation has the highest rotational energies. Figure 5.10 shows the components of rotational energy along the three axis of rotation for the three orientations. We can see that, for OR<sub>1</sub> and OR<sub>2</sub>, the  $E_a$  is nearly constant increase in  $E_{\text{coll}}$ . However, for OR<sub>5</sub>,  $E_a$  increases with  $E_{\text{coll}}$ . In addition, for OR<sub>1</sub>, we see that  $E_a$ ,  $E_b$  and  $E_c$  are nearly independent of the collision energies indicating complimentary  $E_{\text{vib}}/E_{\text{trans}}$  excitation. Interestingly of OR<sub>2</sub> and OR<sub>5</sub>,  $E_b$  and  $E_c$  are higher than  $E_a$ .

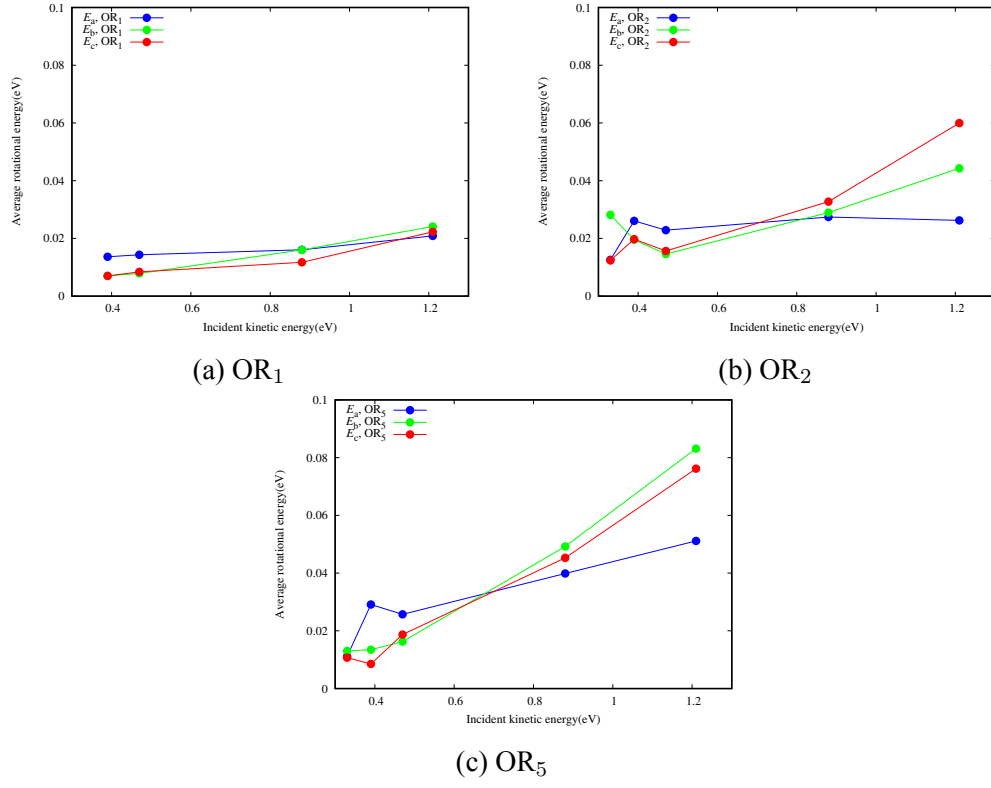


Figure 5.10: The components of rotational energies for different orientations. (a)  $OR_1$  (b)  $OR_2$  and (c)  $OR_5$ .

### 5.3 Energy Transfer

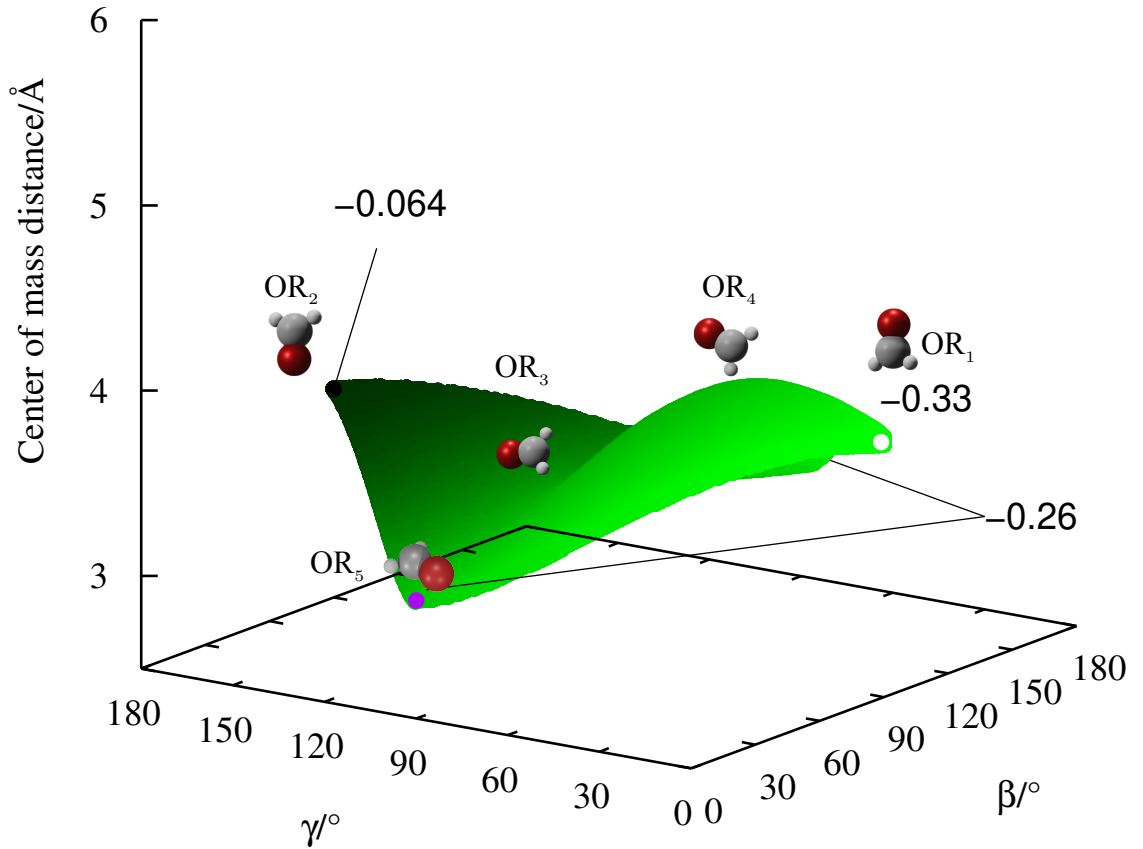


Figure 5.11: Minimum potential energy surface of formaldehyde-graphene interaction as a function of angle  $\beta$ ,  $\gamma$  and the center of mass distance of formaldehyde from the surface. Different orientations are shown, with OR<sub>1</sub> having the least potential well depth (white dot). All energies are in eV.

To investigate the energy transfer process that happens during the scattering process, we looked into the interaction potential during the approach of the projectile. The interaction potential is defined as a function of the orientation of formaldehyde (equilibrium geometry), by the angles  $\beta$  and  $\gamma$ , with respect to the plane defined by the rigid atoms of the graphene sheet and the distance of the center of mass of formaldehyde to the plane of rigid atoms of

graphene sheet, denoted as  $C_z$ . Figure 5.11 shows the minimum potential energy surface of formaldehyde and graphene surface.

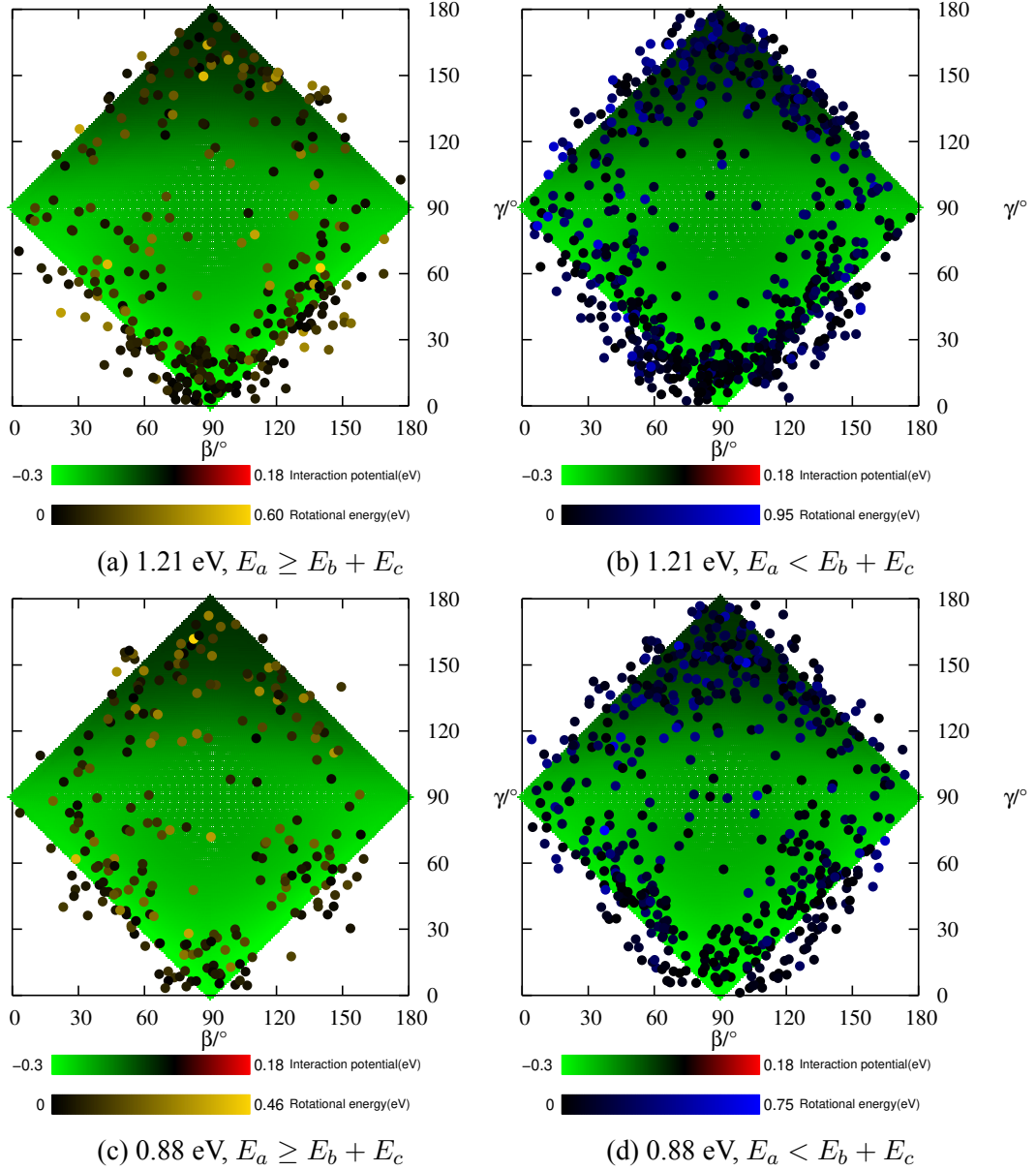


Figure 5.12: The dots represent the orientation at the point of closest approach, and the surface is the minimum interaction potential of formaldehyde with graphene surface. (a) and (c) are for trajectories where  $E_a \geq E_b + E_c$  and (b) and (d) are for trajectories where  $E_a < E_b + E_c$ , (a) and (b) are for 1.21 eV collision energy, and (c) and (d) are for 0.88 eV collision energy.

We then identified the point of closest approach by following the  $C_z$  (i.e. the lowest  $C_z$ ). The point of closest approach for each trajectory is projected onto the minimum energy surface in Figure 5.12 for two collision energies, 0.88 eV and 1.21 eV for trajectories with random initial orientation. The color indicates the rotational excitation. It is evident from the figure that at the point of closest approach only  $OR_1$ ,  $OR_2$  and  $OR_5$  are preferred. The orientations where the plane of formaldehyde is perpendicular to the surface and the  $\vec{CO}$  is parallel to the surface (region near  $\beta=90^\circ$  and  $\gamma=90^\circ$ ) are almost always avoided.

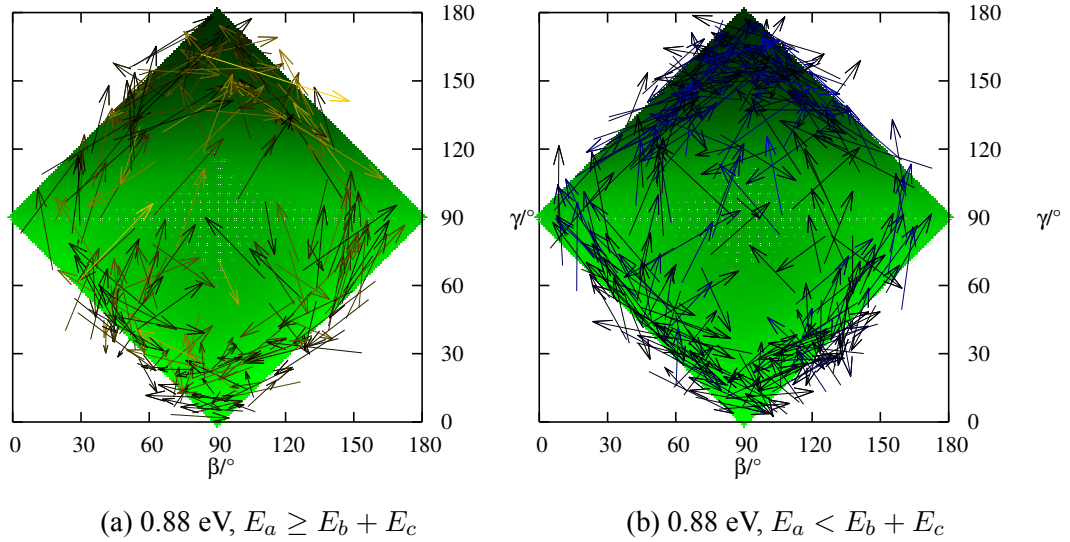


Figure 5.13: The arrows represents 100 time steps or 20 fs worth of trajectory. (a) is for trajectories where  $E_a \geq E_b + E_c$  and (b) is for trajectories where  $E_a < E_b + E_c$ , both (a) and (b) are for 0.88 eV collision energy.

We then followed the direction of trajectories on the  $(\beta, \gamma)$  plane for 20 fs from the point of closest approach and are shown in Figure 5.13. From Figure 5.13, it is not very clear how the mechanism proceeds. The only thing that we are confident about is that only a few selected configurations take part in rotational excitation. All the rest of the collision energies showed similar trajectories as shown in Figure 5.13. This can be attributed to the fact that graphene sheet during simulations show large amplitude (in the  $z$  – direction,

perpendicular to the plane of SLG) undulations moving across the surface, unlike in the case of a hard surface like gold. It can be said that the projection of trajectories on the minimum energy surface may be a good discription to identify the energy transfer mechanism for soft surfaces like graphene.

Figure 5.14, gives the energy transfer computed as the difference between the initial and final energies of formaldehyde in the trajectories for  $E_{\text{coll}} = 1.21$  eV. It can be seen that almost all of the trajectories lose most of its translational kinetic energy. Further, the average loss in total energy of formaldehyde when scattering off of graphene surface larger (0.87 eV), and than that for gold surface (0.55 eV). We also looked at the number of bounces for 1.21 eV. Around 50% trajectories bounce multiple times on the surface before getting scattered. Interestingly, we found that the average rotational excitation for trajectories with multiple bounces in formaldehyde-graphene case was 0.13 eV, compared to single bounce trajectories which had an average rotational energy of 0.21 eV. Also, for all the trajectories it can be seen that formaldehyde always loses total energy to the graphene sheet upon collision.

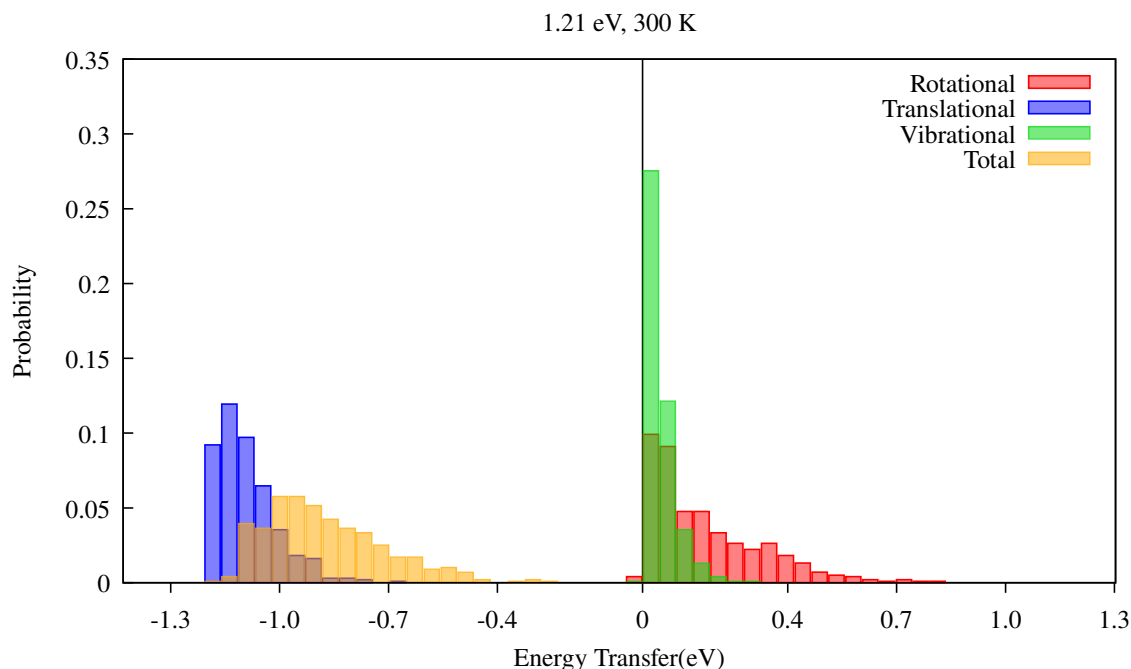
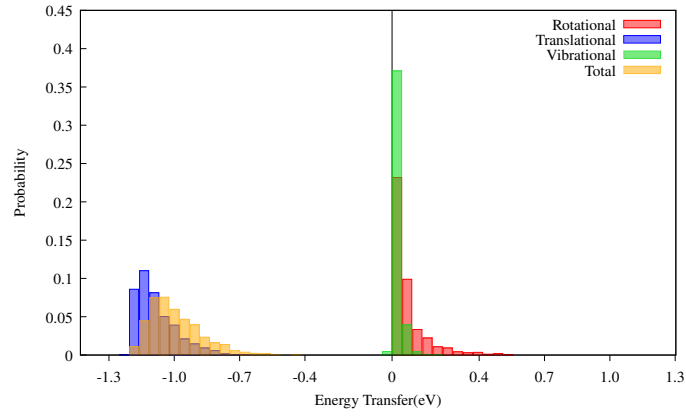


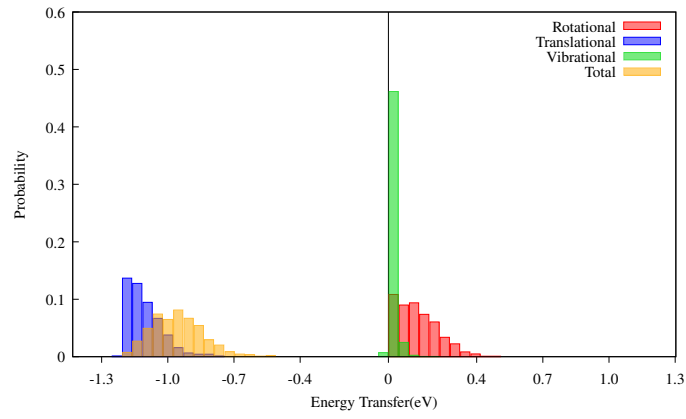
Figure 5.14: Energy transfer for 1.21 eV of collision energy.

### 5.3.1 Effect of Orientation on Energy Transfer

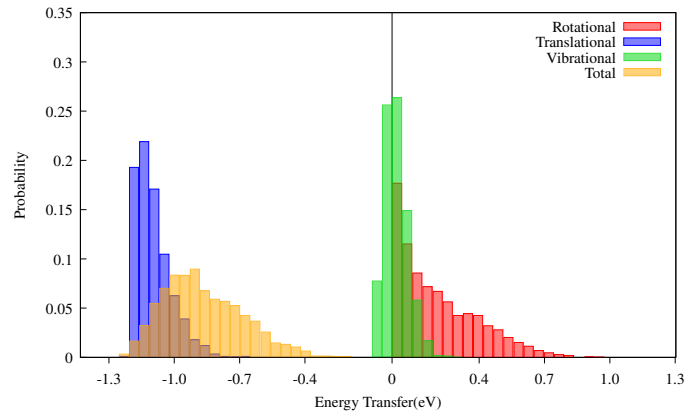
We also looked into the effect of orientation on energy transfer. The results are compiled in Figure 5.15 for  $E_{\text{coll}} = 1.21$  eV. Here also we find that in all the orientations sampled in this study, yielded scattered formaldehyde molecules with less energy than that before collision with the SLG. From Figure 5.15, we can see that OR<sub>1</sub> is the orientation in which the least amount of energy is transferred into the rotational modes (0.07 eV), followed by OR<sub>2</sub> (0.13 eV) and OR<sub>5</sub> (0.21 eV). Further, for the OR<sub>5</sub> orientation, the formaldehyde loses most of its translational kinetic energy (1.1 eV), and also it loses its vibrational energy (0.02 eV), which are transferred to formaldehyde's rotational modes.



(a) OR<sub>1</sub>



(b) OR<sub>2</sub>



(c) OR<sub>5</sub>

Figure 5.15: Energy transfer for 1.21 eV of collision energy for different orientations. (a) OR<sub>1</sub>, (b) OR<sub>2</sub> and (c) OR<sub>5</sub>.

## 5.4 Summary and Conclusions

After studying two surfaces for scattering of formaldehyde, we can say that the surface plays a vital role in final product state distribution and trapping probabilities. In the case of gold and formaldehyde, we saw a rotational rainbow (both experimentally and computationally) which was independent of the collision energy. Whereas for the formaldehyde-graphene system, the rotational rainbow depends on the collision energy. Further, the rotational rainbow seen in graphene and formaldehyde is mostly in the  $b$  and  $c$  axis, unlike  $a$  axis rotational rainbow in gold-formaldehyde system. The mechanism of such rotational rainbow also seems to be different. Despite the different mechanisms, the average rotational energy of scattered formaldehyde for graphene-formaldehyde system is comparable to that of gold-formaldehyde system. In addition, the energy transfer between the surface and formaldehyde have distinctly different signatures. For the gold-formaldehyde system we saw that the more time the formaldehyde spends on the surface (multiple bounce trajectories), the more energetic (internal states) the scattered formaldehyde was. On the contrary, graphene surface seems to reverse this effect, where the longer time the formaldehyde spends on the surface the more energy is lost. We also looked at three orientations, and found that certain orientations ( $OR_5$ ) give rise to scattered formaldehyde with higher average rotational energies. It must be stressed that these results for formaldehyde-graphene have no experimental counterpart, however from preliminary calculations it is evident that the gas-surface interaction decides the final internal states of scattered gas molecules.

## Chapter 6

### Conclusions and Outlook

In this thesis, research on the use of machine learning methods to represent potential energy surfaces and energy transfer in gas-surface scattering are reported. We summarize below the important findings of the research work reported in the thesis.

- The potential energy surface for three systems that are important molecules in the interstellar medium, namely  $\text{HeH}^+ + \text{H}_2$ ,  $\text{C}_5\text{N}^- + \text{H}_2$  and  $\text{C}_2^- + \text{H}_2$  were investigated. We used machine learning methods, viz Gaussian process for regression and artificial neural networks to represent highly accurate potential energy surfaces of each of these systems. All the potential energy surfaces were generated as FORTRAN subroutines that can be used for dynamical simulations.
- The possibility of representing the potential energy surface “on-the-fly” using machine learning methods, during direct dynamics simulations have been explored. Two algorithms were developed and tested, using Gaussian processes for regression and neural networks. The Gaussian process for regression algorithm works well for low dimension, whereas the neural network scheme was found to work upto 6 dimensions.
- The origin of the rotational rainbow observed in experiments<sup>113</sup> during scattering of formaldehyde from gold surface was investigated using classical trajectory simulations. We found that the orientational steering driven by the minimum energy path causes the rotational rainbow during the scattering process.
- The influence of the nature of the surface on the energy transfer in gas-surface scattering was investigated by taking formaldehyde-graphene as a model system. The

results of the classical simulations point out that the rotational rainbow in this system is different from that of formaldehyde-gold system and depends on the collision energy of formaldehyde.

Machine learning has completely replaced the daunting task involved with fitting analytic functions to represent the potential energy surface for higher dimensional systems. It has emerged as a successful alternative to the traditional process of “curve fitting” and has been extended to systems as large as 21 dimensions. They generally involve huge data in orders of millions of points for fitting. The representation of potential energy surfaces by machine learning using minimum number of points is still a bottleneck. To this end, the “on-the-fly” representation of potential energy surfaces provides an alternative. However, efficient schemes for practical implementation need to be developed.

For the gas-surface scattering it would be interesting to investigate the dependence of rotational rainbow on the incident angle. In addition, it is worth exploring the influence of (i) the nature of the surface using different surfaces and (ii) different gas colliders.

# References

- <sup>1</sup> H. Eyring. *J. Chem. Phys.*, **3**(2):107–115, (1935).
- <sup>2</sup> J. I. Steinfeld, J. S. Francisco, and W. L. Hase. Prentice Hall Upper Saddle River, NJ, (1999).
- <sup>3</sup> L. S. Kassel. *J. Phys. Chem.*, **32**(2):225–242, (1928).
- <sup>4</sup> O. K. Rice and H. C. Ramsperger. *J. Am. Chem. Soc.*, **49**(7):1617–1629, (1927).
- <sup>5</sup> R. A. Marcus. *J. Chem. Phys.*, **20**(3):359–364, (1952).
- <sup>6</sup> W. Forst. *J. Chim. Phys.*, **87**:715–741, (1990).
- <sup>7</sup> R. A. Marcus, W. L. Hase, and K. Swamy. *J. Phys. Chem.*, **88**(26):6717–6720, (1984).
- <sup>8</sup> William L Hase. *Science*, **266**(5187):998–1002, (1994).
- <sup>9</sup> D. L. Bunker and W. L. Hase. *J. Chem. Phys.*, **59**(9):4621–4632, (1973).
- <sup>10</sup> B. Jayee and W. L. Hase. *Annu. Rev. Phys. Chem.*, **71**:289–313, (2020).
- <sup>11</sup> H. Goldstein. *Classical mechanics*. Pearson Education India, 2011.
- <sup>12</sup> I. N. Levine, D. H. Busch, and H. Shull. *Quantum chemistry*, volume **6**. (2009).
- <sup>13</sup> D. L. Bunker and E. A. Goring-Simpson. *Faraday Discuss.*, **55**:93–99, (1973).
- <sup>14</sup> C. M. Bishop and N. M. Nasrabadi. *Pattern recognition and machine learning*, volume **4**. (2006).
- <sup>15</sup> A. Karthikeyan and U. D. Priyakumar. *J. Chem. Sci.*, **134**:1–20, (2022).
- <sup>16</sup> J. Cui and R. V. Krems. *J. Phys. B.*, **49**(22):224001, (2016).
- <sup>17</sup> C. Qu, Qi. Yu, and J. M. Bowman. *Annu. Rev. Phys. Chem.*, **69**:151–175, (2018).
- <sup>18</sup> C. M. Handley and P. L. A Popelier. *J. Phys. Chem. A*, **114**(10):3371–3383, (2010).
- <sup>19</sup> Jörg Behler. *J. Phys. Condens. Matter*, **26**(18):183001, (2014).
- <sup>20</sup> M. Ruth, D. Gerbig, and P. R. Schreiner. *J. Chem. Theory Comput.*, **18**(8):4846–4855, (2022).

- <sup>21</sup> P. Pattnaik *et al.* *J. Phys. Chem. A*, **124**(34):6954–6967, (2020).
- <sup>22</sup> B. K. Carpenter, G. S. Ezra, S. C. Farantos, Z. C. Kramer, and S. Wiggins. *J. Phys. Chem. B*, **122**(13):3230–3241, (2017).
- <sup>23</sup> C. A. M. Ramirez, M. Greenop, L. Ashton, and I. U. Rehman. *Appl. Spectrosc. Rev.*, **56**(8-10):733–763, (2021).
- <sup>24</sup> H. Prats and K. Chan. *Phys. Chem. Chem. Phys*, **23**(48):27150–27158, (2021).
- <sup>25</sup> R. Biswas, U. Lourderaj, and N. Sathyamurthy. *J. Chem. Sci.*, **135**(2):22, (2023).
- <sup>26</sup> R. Biswas, R. Rashmi, and U. Lourderaj. *Resonance*, **25**(1):59–75, (2020).
- <sup>27</sup> D.G. Krige. *J. South. Afr. Inst. Min. Metall.*, **52**(6):119–139, (1951).
- <sup>28</sup> G. Matheron. *Econ. Geol.*, **58**(8):1246–1266, (1963).
- <sup>29</sup> J. Dai and R. V. Krems. *J. Chem. Theory Comput.*, **16**(3):1386–1395, (2020).
- <sup>30</sup> A. Radhakrishnan, M. Belkin, and C. Uhler. *Proc. Natl. Acad. Sci. U.S.A.*, **120**(14):e2208779120, 2023.
- <sup>31</sup> G. Daniele and P. Francesco. *Annual Review of Astronomy and Astrophysics*, **51**:163–206, (2013).
- <sup>32</sup> G. Rolf *et al.* *Nature*, **568**(7752):357–359, (2019).
- <sup>33</sup> E. E. Ferguson N. G. Adams, D. K. Bohme. *J. Chem. Phys.*, **52**(10):5101–5105, (1970).
- <sup>34</sup> K. R. Ryan and I. G. Graham. *J. Chem. Phys.*, **59**(8):4260–4271, (1973).
- <sup>35</sup> J. A. Rutherford and D. A. Vroom. *J. Chem. Phys.*, **59**(8):4561–4562, (1973).
- <sup>36</sup> G. Gioumousis and D. P. Stevenson. *J. Chem. Phys.*, **29**(2):294–299, (1958).
- <sup>37</sup> M. Stein. *Technometrics*, **29**(2):143–151, (1987).
- <sup>38</sup> MATLAB 2018b. *The MathWorks Inc.: Natick, MA*, (2018).
- <sup>39</sup> F.A. Gianturco, K. Giri, L. Gonzalez-Sanchez, E. Yurtsever, N. Sathyamurthy, and R. Wester. *The Journal of Chemical Physics*, **154**(5), 2021.
- <sup>40</sup> W. Kolos and L. Wolniewicz. *J. Chem. Phys.*, **46**(4):1426–1432, (1967).
- <sup>41</sup> J. M. Hutson and C. R. Le Sueur. *Comput. Phys. Commun.*, **241**:9–18, (2019).
- <sup>42</sup> A. Dalgarno and R. A. McCray. *Astrophys. J.*, **181**:95–100, (1973).

- <sup>43</sup> M. Agúndez *et al.* *Astron. Astrophys.*, **478**(1):L19–L22, (2008).
- <sup>44</sup> M. Agúndez *et al.* *Astron. Astrophys.*, **517**:L2, (2010).
- <sup>45</sup> J. Cernicharo *et al.* *Astrophys. J.*, **688**(2):L83, (2008).
- <sup>46</sup> H. Gupta, S. Brünken, F. Tamassia, C. A. Gottlieb, M. C. McCarthy, and P. Thaddeus. *Astrophys. J.*, **655**(1):L57, (2007).
- <sup>47</sup> M. C. McCarthy, C. A. Gottlieb, H. Gupta, and P. Thaddeus. *Astrophys. J.*, **652**(2):L141, (2006).
- <sup>48</sup> E. Herbst and Y. Osamura. *Astrophys. J.*, **679**(2):1670, (2008).
- <sup>49</sup> M. Satta, F. A. Gianturco, F. Carelli, and R. Wester. *Astrophys. J.*, **799**(2):228, (2015).
- <sup>50</sup> H. J. Werner, P. J. Knowles, G. Knizia, F. R. Manby, and M. Schütz. *Wiley Interdiscip. Rev. Comput. Mol. Sci.*, **2**(2):242–253, (2012).
- <sup>51</sup> D. J. C. MacKay. *Neural computation*, **4**(3):448–472, (1992).
- <sup>52</sup> G. Scoles, D. Bassi, U. Buck, D. Laine, and C. Braun. *Appl. Opt.*, **28**(16):3258, (1989).
- <sup>53</sup> M. Shapiro and P. Brumer. *Quantum control of molecular processes*. John Wiley & Sons, (2012).
- <sup>54</sup> K. K. Ni *et al.* *Science*, **322**(5899):231–235, (2008).
- <sup>55</sup> T. Schneider, B. Roth, H. Duncker, I. Ernsting, and S. Schiller. *Nat. Phys.*, **6**(4):275–278, (2010).
- <sup>56</sup> P. F. Sta anum, K. Højbjerg, P. S. Skyt, K. A. Hansen, and M. Drewsen. *Nat. Phys.*, **6**(4):271–274, (2010).
- <sup>57</sup> P. R. Stollenwerk, M. G. Kokish, A. G. S. de Oliveira-Filho, F. R. Ornellas, and B. C. Odom. *Atoms*, **6**(3):53, (2018).
- <sup>58</sup> W. G. Rellergert *et al.* *Nature*, **495**(7442):490–494, (2013).
- <sup>59</sup> A. K. Hansen *et al.* *Nature*, **508**(7494):76–79, (2014).
- <sup>60</sup> D. Hauser *et al.* *Nat. Phys.*, **11**(6):467–470, (2015).
- <sup>61</sup> P. Puri *et al.* *Nat. Chem.*, **11**(7):615–621, (2019).
- <sup>62</sup> A. D. Dörfler *et al.* *arXiv preprint arXiv:1906.12285*, (2019).
- <sup>63</sup> M. Germann, X. Tong, and S. Willitsch. *Nat. Phys.*, **10**(11):820–824, (2014).

- <sup>64</sup> F. Wolf *et al.* *Nature*, **530**(7591):457–460, (2016).
- <sup>65</sup> S. Patra *et al.* *Science*, **369**(6508):1238–1241, (2020).
- <sup>66</sup> G. Cerchiari, A. Kellerbauer, M. S. Safronova, U. I. Safronova, and P. Yzombard. *Phys. Rev. Lett.*, **120**(13):133205, (2018).
- <sup>67</sup> C. J. Baker *et al.* *Nature*, **592**(7852):35–42, (2021).
- <sup>68</sup> U. Warring *et al.* *Phys. Rev. Lett.*, **102**(4):043001, (2009).
- <sup>69</sup> E. Jordan, G. Cerchiari, S. Fritzsche, and A. Kellerbauer. *Phys. Rev. Lett.*, **115**(11):113001, (2015).
- <sup>70</sup> R. Tang *et al.* *Phys. Rev. Lett.*, **123**(20):203002, (2019).
- <sup>71</sup> P. Yzombard, M. Hamamda, S. Gerber, M. Doser, and D. Comparat. *Phys. Rev. Lett.*, **114**(21):213001, (2015).
- <sup>72</sup> M. Nötzold, R. Wild, C. Lochmann, and R. Wester. *Phys. Rev. A*, **106**(2):023111, (2022).
- <sup>73</sup> D. López-Durán, E. Bodo, and F. A. Gianturco. *Comput. Phys. Commun.*, **179**(11):821–838, (2008).
- <sup>74</sup> K. K. Baldridge, M. S. Gordon, R. Steckler, and D. G. Truhlar. *J. Phys. Chem.*, **93**(13):5107–5119, (1989).
- <sup>75</sup> R. L. Bell and T. N. Truong. *J. Chem. Phys.*, **101**(12):10442–10451, (1994).
- <sup>76</sup> W. Chen, W. L. Hase, and H. B. Schlegel. *Chem. Phys. Lett.*, **228**(4-5):436–442, (1994).
- <sup>77</sup> W. T. Duncan, R. L. Bell, and T. N. Truong. *J. Comput. Chem.*, **19**(9):1039–1052, (1998).
- <sup>78</sup> J. M. Millam, V. Bakken, W. Chen, W. L. Hase, and H. B. Schlegel. *J. Chem. Phys.*, **111**(9):3800–3805, (1999).
- <sup>79</sup> G. Li and W. L. Hase. *J. Am. Chem. Soc.*, **121**(30):7124–7129, (1999).
- <sup>80</sup> D. J. Mann and W. L. Hase. *J. Am. Chem. Soc.*, **124**(13):3208–3209, (2002).
- <sup>81</sup> L. Sun and W. L. Hase. *J. Chem. Phys.*, **121**(18):8831–8845, (2004).
- <sup>82</sup> H. Tachikawa. *J. Phys. Chem. A*, **108**(39):7853–7862, (2004).
- <sup>83</sup> H. G. Yu, J. T. Muckerman, and J. S. Francisco. *J. Phys. Chem. A*, **109**(23):5230–5236, (2005).
- <sup>84</sup> K. Park, K. Song, and W. L. Hase. *Int. J. Mass Spectrom.*, **265**(2-3):326–336, (2007).

- <sup>85</sup> K. C. Thompson, M. J. T. Jordan, and M. A. Collins. *J. Chem. Phys.*, **108**(2):564–578, (1998).
- <sup>86</sup> A. Brown, A. B. McCoy, B. J. Braams, Z. Jin, and J. M. Bowman. *J. Chem. Phys.*, **121**(9):4105–4116, (2004).
- <sup>87</sup> Z. Li, J. R. Kermode, and A. D. Vita. *Phys. Rev. Lett.*, **114**(9):096405, (2015).
- <sup>88</sup> Y. Guan, S. Yang, and D. H. Zhang. *Mol. Phys.*, **116**(7-8):823–834, (2018).
- <sup>89</sup> L. M. Raff, M. Malshe, M. Hagan, D. I. Doughan, M. G. Rockley, and R. Komanduri. *J. Chem. Phys.*, **122**(8):084104, (2005).
- <sup>90</sup> R. Dawes et al. *J. Chem. Phys.*, **130**(14):144107, (2009).
- <sup>91</sup> B. Kolb, P. Marshall, B. Zhao, B. Jiang, and H. Guo. *J. Phys. Chem. A*, **121**(13):2552–2557, (2017).
- <sup>92</sup> H. Sugisawa, T. Ida, and R. V. Krems. *J. Chem. Phys.*, **153**(11):114101, (2020).
- <sup>93</sup> G. Schmitz, E. L. Klinting, and O. Christiansen. *J. Chem. Phys.*, **153**(6):064105, (2020).
- <sup>94</sup> S. Seo, M. Wallat, T. Graepel, and K. Obermayer. In *Mustererkennung 2000: 22. DAGM-Symposium.*, pages 27–34, (2000).
- <sup>95</sup> N. Artrith and J. Behler. *Phys. Rev. B*, **85**(4):045439, (2012).
- <sup>96</sup> N. Ramakrishnan, C. B. Kellogg, S. Tadepalli, and V. N. Pandey. In *Proceedings of the 2005 SIAM International Conference on Data Mining*, pages 427–438, (2005).
- <sup>97</sup> A. Krause, A. Singh, and C. Guestrin. *J. Mach. Learn. Res.*, **9**(2), (2008).
- <sup>98</sup> E. Bosch, M. Moreno, J. M. Lluch, and J. Bertrán. *J. Chem. Phys.*, **93**(8):5685–5692, (1990).
- <sup>99</sup> S. Barethiya, U. Sankar, S. Rana, and U. Lourderaj. *Master of Science Thesis*. NISER Bhubaneswar, (2022).
- <sup>100</sup> P. J. Haley and D. Soloway. *Proc. Int. Jt. Conf. Neural Netw.*, **4**:25–30, (1992).
- <sup>101</sup> X. Zhang, S. Zou, L. B. Harding, and J. M. Bowman. *J. Phys. Chem. A*, **108**(41):8980–8986, (2004).
- <sup>102</sup> R. W. Zwanzig. *J. Chem. Phys.*, **32**(4):1173–1177, (1960).
- <sup>103</sup> H. Saltsburg and J. N. Smith Jr. *J. Chem. Phys.*, **45**(6):2175–2183, (1966).

- <sup>104</sup> J. E. Hurst Jr., L. Wharton, K. C. Janda, and D.J. Auerbach. *J. Chem. Phys.*, **83**(3):1376–1381, (1985).
- <sup>105</sup> C. B. Mullins and W. H. Weinberg. *J. Vac. Sci. Technol.*, **8**(3):2458–2462, (1990).
- <sup>106</sup> C. T. Rettner, D. S. Bethune, and D. J. Auerbach. *J. Chem. Phys.*, **91**(3):1942–1943, (1989).
- <sup>107</sup> L. M. Raff, J. Lorenzen, and B. C. McCoy. *J. Chem. Phys.*, **46**(11):4265–4274, (1967).
- <sup>108</sup> J. Lorenzen and L. M. Raff. *J. Chem. Phys.*, **52**(3):1133–1142, (1970).
- <sup>109</sup> J. Lorenzen and L. M. Raff. *J. Chem. Phys.*, **49**(3):1165–1177, (1968).
- <sup>110</sup> J. Lorenzen and L. M. Raff. *J. Chem. Phys.*, **52**(12):6134–6140, (1970).
- <sup>111</sup> A. W. Kleyn and T. C. M. Horn. *Phys. Rep.*, **199**(4):191–230, (1991).
- <sup>112</sup> H. Schlichting, D. Menzel, T. Brunner, W. Brenig, and J. C. Tully. *Phys. Rev. Lett.*, **60**(24):2515, (1988).
- <sup>113</sup> P. G. Barratt, B. C. Krüger, S. Meyer, A. Kandratsenka, A. M. Wodtke, and T. Schäfer. *Phys. Chem. Chem. Phys.*, **19**(30):19904–19915, (2017).
- <sup>114</sup> M. E. Saecker, S. T. Govoni, D. V. Kowalski, M. E. King, and G. M. Nathanson. *Science*, **252**:1421–1424, (1990).
- <sup>115</sup> D. Kulginov, M. Persson, C. Åkerlund, I. Zorić, and B. Kasemo. *J. Vac. Sci. Technol.*, **13**(3):1511–1516, (1995).
- <sup>116</sup> H. Ambaye and J. R. Manson. *J. Chem. Phys.*, **125**(8):084717, (2006).
- <sup>117</sup> B. C. Krüger, G. B. Park, S. Meyer, R. J. V. Wagner, A. M. Wodtke, and T. Schäfer. *Phys. Chem. Chem. Phys.*, **19**:19896–19903, (2017).
- <sup>118</sup> S. Nave, A. K. Tiwari, and B. Jackson. *J. Phys. Chem. A*, **118**(41):9615–9631, (2014).
- <sup>119</sup> S. B. M. Bosio and W. L. Hase. *J. Chem. Phys.*, **107**(22):9677–9686, (1997).
- <sup>120</sup> D. Beck, U. Ross, and W. Schepper. *Phys. Rev. A*, **19**(6):2173, (1979).
- <sup>121</sup> C. T. Rettner, E. K. Schweizer, and H. Stein. *J. Chem. Phys.*, **93**(2):1442–1454, (1990).
- <sup>122</sup> J. M. Bowman. *Chem. Phys. Lett.*, **62**(2):309–311, (1979).
- <sup>123</sup> R. Schinke. *Chem. Phys.*, **47**(2):287–294, (1980).
- <sup>124</sup> M. B. Faist and R. D. Levine. *J. Chem. Phys.*, **64**(7):2953–2970, (1976).

- <sup>125</sup> J. D. McClure. *J. Chem. Phys.*, **51**(5):1687–1700, (1969).
- <sup>126</sup> A. W. Kleyn. *Surf. Rev. Lett.*, **1**(01):157–173, (1994).
- <sup>127</sup> D. Beck. *Chem. Phys.*, **126**(1):19–26, (1988).
- <sup>128</sup> A. C. Wight and R. E. Miller. *J. Chem. Phys.*, **109**(5):1976–1982, (1998).
- <sup>129</sup> B. D. Kay, T. D. Raymond, and M. E. Coltrin. *Phys. Rev. B*, **36**(12):6695, (1987).
- <sup>130</sup> M. E. Coltrin and B. D. Kay. *J. Chem. Phys.*, **89**(1):551–561, (1988).
- <sup>131</sup> A. P. Sutton and J. Chen. *Philos. Mag. Lett.*, **61**(3):139–146, (1990).
- <sup>132</sup> S. Arcidiacono, J. H. Walther, D. Poulidakos, D. Passerone, and P. Koumoutsakos. *Phys. Rev. Lett.*, **94**(10):105502, (2005).
- <sup>133</sup> A. Berg, C. Peter, and K. Johnston. *J. Chem. Theory Comput.*, **13**(11):5610–5623, (2017).
- <sup>134</sup> W. L. Hase *et al.* *Quan. Chem. Prog. Exch. Bull.*, **16**:671, (1996).
- <sup>135</sup> C. T. Rettner, J. Kimman, and D. J. Auerbach. *J. Chem. Phys.*, **94**(1):734–750, (1991).
- <sup>136</sup> O. Bünermann, A. Kandratsenka, and A. M. Wodtke. *J. Phys. Chem. A*, **125**(15):3059–3076, (2021).
- <sup>137</sup> R. Nieman, J. A. A. Aquino, and H. Lischka. *J. Phys. Chem. A*, **125**(5):1152–1165, (2021).
- <sup>138</sup> T. Greenwood and S. P. K. Koehler. *J. Phys. Chem. C*, **125**(32):17853–17860, (2021).
- <sup>139</sup> T. Greenwood, H. AlSalem, and S. P. K. Koehler. *J. Phys. Chem. A*, **127**(5):1124–1129, (2023).
- <sup>140</sup> T. Greenwood and S. P. K. Koehler. *Chem. Phys. Chem.*, **23**(22):e202200216, (2022).
- <sup>141</sup> D. S. Leela and U. Lourderaj. *Phys. Chem. Chem. Phys.*, **15**(40):17479–17486, 2013.
- <sup>142</sup> D. A. Case *et al.* Amber 10. Technical report, University of California, 2008.

# **Appendices**

# Appendix A

Table A.1: Statistics of the trajectories considered in the study for different initial conditions for gold-formaldehyde system.

Surface temperature, $T_s$ (K)	Rotational temperature (K)	Orientation	Collision energy (eV)	No. of trajectories	Scattered (ZPE corrected)	Trapped
140	15	random	1.21	2560	2504 ( <b>1346</b> )	56
			0.88	2560	2387 ( <b>1141</b> )	173
			0.47	2592	1737 ( <b>605</b> )	855
			0.39	2876	1401 ( <b>437</b> )	1475
			0.33	2560	931 ( <b>266</b> )	1629
200	15	random	1.21	2560	2500 ( <b>1371</b> )	60
			0.88	2560	2401 ( <b>1171</b> )	159
			0.47	2560	1662 ( <b>578</b> )	898
			0.39	2286	1109 ( <b>346</b> )	1177
			0.33	2560	1019 ( <b>303</b> )	1541
	0	OR <sub>1</sub>	1.21	1024	996 ( <b>424</b> )	28
			0.88	1024	955 ( <b>420</b> )	69
			0.47	1024	790 ( <b>250</b> )	234
			0.39	1024	700 ( <b>229</b> )	324
			0.33	1024	622 ( <b>162</b> )	402
		OR <sub>2</sub>	1.21	1024	1024 ( <b>551</b> )	0
			0.88	1024	1022 ( <b>525</b> )	2
			0.47	1024	798 ( <b>380</b> )	226
			0.39	1024	618 ( <b>311</b> )	406
			0.33	1024	498 ( <b>221</b> )	526
		OR <sub>3</sub>	1.21	1024	1024 ( <b>554</b> )	0
			0.88	1024	1007 ( <b>515</b> )	17
			0.47	1024	472 ( <b>138</b> )	552
			0.39	1024	279 ( <b>67</b> )	745
			0.33	1024	196 ( <b>45</b> )	828
			1.21	1024	1024 ( <b>575</b> )	0
			0.88	1024	1017 ( <b>499</b> )	7

300		OR <sub>4</sub>	0.47	1024	754 ( <b>316</b> )	270
			0.39	1024	599 ( <b>236</b> )	425
			0.33	1024	467 ( <b>169</b> )	557
		OR <sub>5</sub>	1.21	1024	1024 ( <b>940</b> )	0
			0.88	1024	1022 ( <b>889</b> )	2
			0.47	1024	784 ( <b>528</b> )	240
			0.39	1024	670 ( <b>414</b> )	354
			0.33	1024	563 ( <b>313</b> )	461
	15	random	1.21	11880	11580 ( <b>6196</b> )	300
			0.88	9600	8935 ( <b>4262</b> )	665
			0.47	9600	6198 ( <b>2260</b> )	3402
			0.39	8807	4409 ( <b>1457</b> )	4398
			0.33	14982	5869 ( <b>1876</b> )	9113
		OR <sub>1</sub>	1.21	1024	968 ( <b>420</b> )	56
			0.88	1024	938 ( <b>369</b> )	86
			0.47	1024	767 ( <b>238</b> )	257
			0.39	1024	661 ( <b>196</b> )	363
			0.33	1024	589 ( <b>154</b> )	435
300	0	OR <sub>2</sub>	1.21	1024	1021 ( <b>550</b> )	3
			0.88	1024	1005 ( <b>521</b> )	19
			0.47	1024	742 ( <b>351</b> )	282
			0.39	1024	594 ( <b>251</b> )	430
			0.33	1024	480 ( <b>218</b> )	544
		OR <sub>3</sub>	1.21	1024	1024 ( <b>567</b> )	0
			0.88	1024	984 ( <b>471</b> )	40
			0.47	1024	519 ( <b>169</b> )	505
			0.39	1024	347 ( <b>109</b> )	677
			0.33	1024	260 ( <b>85</b> )	764
		OR <sub>4</sub>	1.21	1024	1018 ( <b>568</b> )	6
			0.88	1024	1000 ( <b>499</b> )	24
			0.47	1024	724 ( <b>305</b> )	300
			0.39	1024	636 ( <b>244</b> )	388
			0.33	1024	460 ( <b>191</b> )	564
		OR <sub>5</sub>	1.21	1024	1024 ( <b>903</b> )	0
			0.88	1024	1017 ( <b>865</b> )	7
			0.47	1024	792 ( <b>555</b> )	232
			0.39	1024	668 ( <b>412</b> )	356
			0.33	1024	560 ( <b>348</b> )	464
	15	random	1.21	5276	5012 ( <b>2645</b> )	264
			0.88	5120	4696 ( <b>2168</b> )	424
			0.47	5256	3272 ( <b>1169</b> )	1984
			0.39	5096	2651 ( <b>902</b> )	2445

400	0		0.33	5120	2182 ( <b>685</b> )	2938
			1.21	1024	958 ( <b>414</b> )	66
			0.88	1024	928 ( <b>379</b> )	96
			0.47	1024	722 ( <b>223</b> )	302
			0.39	1024	639 ( <b>232</b> )	385
			0.33	1024	576 ( <b>170</b> )	448
		OR <sub>1</sub>	1.21	1024	1020 ( <b>482</b> )	4
			0.88	1024	998 ( <b>496</b> )	26
			0.47	1024	725 ( <b>312</b> )	299
			0.39	1024	587 ( <b>266</b> )	437
			0.33	1024	483 ( <b>194</b> )	541
		OR <sub>2</sub>	1.21	1024	1020 ( <b>550</b> )	4
			0.88	1024	974 ( <b>456</b> )	50
			0.47	1024	551 ( <b>210</b> )	473
			0.39	1024	381 ( <b>134</b> )	643
			0.33	1024	333 ( <b>107</b> )	691
		OR <sub>3</sub>	1.21	1024	1021 ( <b>551</b> )	3
			0.88	1024	993 ( <b>515</b> )	31
			0.47	1024	720 ( <b>311</b> )	304
			0.39	1024	589 ( <b>233</b> )	435
			0.33	1024	503 ( <b>173</b> )	521
		OR <sub>4</sub>	1.21	1024	1023 ( <b>887</b> )	1
			0.88	1024	1006 ( <b>839</b> )	18
			0.47	1024	764 ( <b>530</b> )	260
			0.39	1024	670 ( <b>465</b> )	354
			0.33	1024	555 ( <b>386</b> )	469
		OR <sub>5</sub>	1.21	1024	2492 ( <b>1339</b> )	68
			0.88	2560	2340 ( <b>1078</b> )	220
			0.47	2560	1703 ( <b>592</b> )	857
			0.39	2958	1589 ( <b>538</b> )	1369
			0.33	2728	1125 ( <b>344</b> )	1603
480	15	random	1.21	2560	2492 ( <b>1339</b> )	68
			0.88	2560	2340 ( <b>1078</b> )	220
			0.47	2560	1703 ( <b>592</b> )	857
			0.39	2958	1589 ( <b>538</b> )	1369
			0.33	2728	1125 ( <b>344</b> )	1603

# Appendix B

Table B.1: Statistics of the trajectories considered in the study for different initial conditions for graphene-formaldehyde system.

Surface temperature, $T_s$ (K)	Rotational temperature (K)	Orientation	Collision energy (eV)	No. of trajectories	Scattered (ZPE corrected)	Trapped
300	15	random	1.21	2500	988 ( <b>447</b> )	1512
			0.88	3200	810 ( <b>388</b> )	2365
			0.47	3840	259 ( <b>112</b> )	3581
			0.39	3840	173 ( <b>73</b> )	3667
			0.33	7680	271 ( <b>113</b> )	7409
	0	OR <sub>1</sub>	1.21	6400	2607 ( <b>1096</b> )	3793
			0.88	6400	1301 ( <b>515</b> )	5099
			0.47	6400	288 ( <b>74</b> )	6112
			0.39	6400	175 ( <b>46</b> )	6225
			0.33	20591	128 ( <b>0</b> )	20463
		OR <sub>2</sub>	1.21	6400	2281 ( <b>1134</b> )	4119
			0.88	6400	2192 ( <b>1061</b> )	4208
			0.47	6400	911 ( <b>435</b> )	5489
			0.39	6400	482 ( <b>205</b> )	5918
			0.33	20923	649 ( <b>160</b> )	19644
		OR <sub>5</sub>	1.21	6400	3178 ( <b>2625</b> )	3222
			0.88	6400	1566 ( <b>1285</b> )	4834
			0.47	6400	337 ( <b>177</b> )	6063
			0.39	6400	176 ( <b>123</b> )	6224
			0.33	20255	563 ( <b>530</b> )	19692