This thesis has been submitted in fulfilment of the requirements for a postgraduate degree (e.g. PhD, MPhil, DClinPsychol) at the University of Edinburgh. Please note the following terms and conditions of use:

- This work is protected by copyright and other intellectual property rights, which are retained by the thesis author, unless otherwise stated.
- A copy can be downloaded for personal non-commercial research or study, without prior permission or charge.
- This thesis cannot be reproduced or quoted extensively from without first obtaining permission in writing from the author.
- The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the author.
- When referring to this work, full bibliographic details including the author, title, awarding institution and date of the thesis must be given.
Computational Methods for the Interpretation of Ultrafast Photochemical Reactions

Kyle Acheson

Doctor of Philosophy
The University of Edinburgh
2023
"The more I read, the more I acquire, 
the more certain I am that I know nothing"

Voltaire
Abstract

This thesis presents a detailed account of the development and application of computational methods and frameworks, which aim to aid the understanding of the dynamics involved in ultrafast photochemical reactions. Overall, these methods facilitate the interpretation, analysis, and extraction of information from both theoretical simulations and experimental observables. Particular emphasis is placed on molecular structure determination in time-resolved scattering experiments, including ultrafast x-ray scattering and ultrafast electron diffraction.

The thesis begins by exploring how methods from unsupervised machine learning can be used to build reduced dimensionality summary models that identify the key reaction pathways in trajectory simulations of complex photochemical reactions. In a first approach, we consider a time-series treatment, in which we explore the use of distance metrics to provide a single-valued similarity measure for trajectories across all times. We find that based on a reduced number of internal coordinates alone, this approach is capable of providing a rapid first-order identification of several key components of the reaction flux. However, the clustering results are limited since they do not always reflect the relationship between nuclear geometries and electronic state populations, and can struggle to categorise trajectories that form a broad distribution. We therefore develop a new pattern mining approach, comprised of two algorithms that work in tandem. First, a clustering algorithm is applied at each time step independently, yielding a sequence of clusters at each point in time, then a second pattern mining algorithm mines the clustered data for temporal trends, resulting in a set of spatio-temporal clusters. The clustering algorithm is specifically designed to handle the type of trajectory data that emerges from quantum molecular dynamics simulation, and the pattern mining algorithm is sufficiently flexible to identify both groups of trajectories that are strongly correlated in time, and those that evolve together with a smaller degree of temporal correlation.
The two algorithms are applied to trajectory data that models the electrocyclic ring-opening reaction of the molecule 1,3-cyclohexadiene. In this molecule, we demonstrate that the patterns identified reflect the state populations, and match the lifetimes and product channels determined by other independent studies. The algorithm is shown to be capable of identifying the times at which the wavepacket bifurcates at conical intersections, with each of the resulting spatio-temporal clusters corresponding to unique reactive pathways.

Next, we consider the problem of extracting a molecular model from experimentally measured observables through inversion. This is treated as a forward optimisation problem, in which the experimentally measured observable is reconstructed through the iterative refinement of model observables. The molecular model is expressed in a trajectory basis, ensuring temporal continuity. The inversion therefore amounts to identifying the optimal trajectory weights that reproduce the experimental observable through minimisation of a target function. The procedure can account for a wide range of experimental distortions, providing a robust method that allows one to view the dynamics in light of the given experimental constraints. The approach is evaluated for time-dependent, excited state molecular structure determination using experimental data from ultrafast x-ray and electron scattering experiments. In either case, the optimisation filters out a handful of key characteristic trajectories from an ensemble of trajectory basis functions, corresponding to a minimal representation of the dynamics observed in the experiment. The validity of the results is discussed in view of a range of independent validation metrics. Importantly, the best models (solutions) are consistent across numerous metrics and agree with independent results from the literature.

The photodissociation dynamics of \( \text{CS}_2 + h\nu \rightarrow \text{CS}(X^1\Sigma) + \text{S}(^1D_2)/S(^3P_J) \) as measured by ultrafast electron diffraction, is then discussed in depth. Importantly, the resolution in the specific experiment under consideration was limited and noise levels were high. This precluded the analysis of the experimental data using the atom-atom pair distribution function obtained by a sine-transform of the scattering signal, which is otherwise common in electron diffraction. A critical assessment of the limitations of pair distribution function analysis is provided, and instead the observed dynamics are extracted using the forward optimisation approach to inversion. The inversion allows key markers of photodissociation in the signal to be identified, along with the singlet \( S(^1D) \) to triplet \( S(^3P_J) \) branching ratio and the bound state populations at 1 ps, consistent with literature.
The work presented in this thesis explores how computational methods can be used to better understand chemical reactions that are triggered by light. These so-called photochemical reactions are prevalent in natural processes such as photosynthesis, the formation of atmospheric pollutants, and the formation of prebiotic molecules that are thought to be responsible for the origins of life itself. Furthermore, they are crucial to modern technology, with applications to solar cells, light emitting diodes (LEDs), fibre optics, and the manufacturing of semiconductors used in modern computers. Building an understanding of how these processes occur through a combination of experimental and computational methods is therefore imperative to the development of new technologies and materials. Moreover, by studying these reactions at the atomic level, we aim to understand the very nature of light-induced molecular motion itself.

Molecules are comprised of nuclei and electrons. The electrons are distributed around the nuclei, occupying distinct energy levels. If the impending light matches the energy difference between two levels, this causes a transition to a different energy level with a commensurate change in how the electrons are distributed around the nuclei. This results in the nuclei experiencing different electrostatic forces from the electrons, which in turn induces nuclear motion. A photochemical reaction is thus characterised by the combined motion of nuclei and electrons. Once a photochemical reaction is induced, the system can decay through a multitude of different pathways, eventually yielding a distribution of different products or product states.

Photochemical reactions occur on the fundamental timescale of electronic and nuclear motion and are governed by quantum mechanics. The timescales are extremely short, on the order of attoseconds and femtoseconds, which are one quintillionth \( (10^{-18}\text{ s}) \) and one quadrillionth \( (10^{-15}\text{ s}) \) of a second, for electronic and nuclear motion respectively. In the last two decades, new light sources such as free-electron lasers (FELs), have afforded us with ultrabright, fast, and high energy sources of light that can be used to directly probe the dynamics of photochemical reactions. Typically, one initiates a photochemical reaction by applying a *pump* light pulse to the system. After a time delay, a second *probe* pulse is applied to image the evolving reaction over a continuous range of times. In a scattering experiment, the probe pulse (which
could also be a pulse of particles), interacts with the molecule and is scattered at a different angle. The intensity of each of these scattering events is recorded on a detector, and the series of snapshot detector images over time defines the time-evolution of the system. Each of these images corresponds to a scattering pattern, which encodes information on the specific configuration of electrons and nuclei at that given time. Extracting the molecular structure from each of these scattering patterns is challenging and involves use of advanced mathematical and computational methods to compare the experiment with models, which are themselves often constructed from advanced computer simulations. However, once the molecular structures are determined from every recorded detector image, they can be pieced together to create a "molecular movie" representing the reaction dynamics observed in the experiment.

Computational methods provide a way to simulate the photochemical dynamics according to quantum mechanics. However, even with these simulations in hand, interpreting the experimental observations via side-by-side comparisons with theoretical simulations can be difficult. This thesis explores the use of computational methods that take the experimentally recorded scattering patterns and directly extract the structural information encoded within then, without the need for performing side-by-side comparisons with theoretical models. The development of an optimisation based algorithm is explored, and applied to two example molecules, $\text{CS}_2$ and 1,3-cyclohexadiene (CHD). The method is found to be robust with respect to inherent noise in the experiment and the limitations of experimental resolution, identifying the key structural dynamics that the detector images contain. In both cases, the results are commensurate with other independent theoretical and experimental studies.

Moreover, the simulations that are used to model the dynamics are themselves often highly complex. To aid the interpretation of simulations, this thesis develops unsupervised machine learning algorithms that categorise the time-dependent trends in the simulation data. This allows one to identify the dominant reaction pathways, and provides a concise summary of complex dynamics that can be used to help analyse experiment and to better understand the underlying physics and chemistry. Initially, we explore the application of existing algorithms to simulations that define the dynamics of CHD. However, we then proceed to develop two new clustering algorithms that go beyond the capabilities of existing algorithms and perform better in the specific case of quantum mechanical simulations.
First, I would like to express my deepest gratitude to my supervisor, Prof. Adam Kirrander, who has proved to be a continual source of support and inspiration. Throughout my PhD, he has encouraged exploration into new unfamiliar realms, nurturing my curiosity and ultimately provided me with the tools to become a better scientist. Moreover, in challenging times Adam exhibited a great deal of patience and empathy, helping me through the trials and tribulations of doctoral studies. I will sincerely miss working alongside him, and hope that our relationship will continue to grow as I embark on the next journey in my scientific career.

I would also like to acknowledge the other members of the Kirrander research group - past and present: Dr Mats Simmermacher, Dr Andrés Moreno Carrascosa, Dr Nikola Zotev, Dr Thomas Northey, Joseph Cooper, Lauren Bertram, Jakub Kara, and Eirik Moen Liane. All of whom have provided support, facilitated engaging scientific discussion, and created a stimulating and friendly working environment throughout my time with them. A special thanks goes to Mats and Andrés, who on numerous occasions allowed me to stay with them upon my visits to Oxford and provided significant support, both professionally and personally. I am also thankful for having the opportunity to collaborate with Jakub Kara, who was instrumental in the development of new computational methods. Working with him was a pleasure, and I am grateful for the stimulating discussions he provided.

Furthermore, I am thankful for my friendship with Dr Karl-Michael Ziems, who first came to the group as an external visitor. I am especially grateful for his ability to precisely brew a cup of coffee while engaging in scientific discussion. I would also like to thank Dr Thomas Lambert, who I first met calming down a stressed out student writing their thesis on one of my first days in the department. Tom soon became a candid voice of clarity during trying times of my own, and I am grateful to be able to call him a close friend.

A deep and heartfelt appreciation is reserved for my partner, Anna. Throughout my doctoral studies, her unwavering support, understanding, and belief, have been invaluable. While hunched over the computer screen late at night, she was often a continual source of light. An extended thank you is owed to Anna and her family,
Kate, Alan, and Leah, who welcomed me into their home during the final stretch of my PhD when the rest of the group moved to Oxford. I am forever grateful to all of them for providing me with a not only a place to stay, but a home - even complete with my own office to work from.

Finally, I wish to express my gratitude to my family, both my parents Donna and Robin, and my brother Jordan. From a young age, my parents emphasised the importance of education to me, often sacrificing their own self interests in order to support me in my pursuit of science. Without them, none of this would have been possible. In addition, I find myself incredibly proud and inspired by both my parents and brothers resilience and empathy. Even though I have known them my whole life, there is still a great deal for me to learn from them. I am eternally grateful to all three of them.
Declaration

I, Kyle Acheson, declare that thesis, titled "Computational Methods for the Interpretation of Ultrafast Photochemical Reactions", was composed by myself while studying at The University of Edinburgh, and has not been submitted for any other degree or qualification elsewhere. I confirm that the work herein is my own, except where work which has formed part of jointly-authored publications has been included. All publications are listed appropriately in the 'List of Publications' section. However, my own contributions, and those of other authors, are indicated below. I confirm that appropriate credit has been given within this thesis where reference has been made to the work of others. I also confirm that I have acknowledged all sources of help where appropriate.

The work presented in Chapter 3 was previously published in the Journal of Chemical Theory and Computation as "Automatic Clustering of Excited-State Trajectories: Application to Photoexcited Dynamics" by Kyle Acheson and Adam Kirrander. The idea for the work within the Chapter was conceived by myself, and I am completely responsible for the work, under the supervision of Adam Kirrander.

The work presented in Chapter 4 is currently in preparation for publication, I am solely responsible for the manuscript. I independently conceived the initial idea for the project and am responsible for the work therein. The development of the spatial clustering algorithm was a collaborative effort between Jakub Kara and myself, where I assumed a supervisory role. The development of the pattern mining algorithm, and its application to a molecular system, was solely my contribution, however I am deeply thankful to Jakub Kara for being a continual source of insight and providing engaging discussions. The project was conducted under the supervision of Adam Kirrander, who provided his own unique insight.

The work presented in Chapter 5 was previously published in the Journal of Chemical Theory and Computation as "Robust inversion of time-resolved data via forward-optimisation in a trajectory basis" by Kyle Acheson and Adam Kirrander. I am the first author of this paper and the work reported therein is solely my own. I was also solely responsible for the preparation of the manuscript, under the supervision of Adam Kirrander.
The work presented in Chapter 6 was previously published in Physical Chemistry Chemical Physics as "Multichannel Photodissociation Dynamics in CS2 Studied by Ultrafast Electron Diffraction" by Weronika O. Razmus and Kyle Acheson et. al. I share first authorship of the paper, working closely with Weronika O. Razmus. My contributions include, analysis of the experiment, interpreting the data with the aid of my own algorithm, and identifying experimental limitations. I also provided remote support to the wider collaboration during the collection of experimental data. I was instrumental in the preparation of the manuscript, being solely responsible for writing the theoretical sections and supplementary information. In addition, I played a central role in contributing to the introduction, results, and concluding sections. The work in this chapter involved a large collaboration between experiment and theory, with Rauridh Forbes, Russell Minns, and Adam Kirrander, seeding the idea for the work, providing key supervision, and being the main driving force behind the project.

Kyle Acheson
List of Publications

The work contained within this thesis has been published as:


Other work, that is not directly discussed in this thesis, has been published as:

Contents

Abstract iii
Lay Summary v
Acknowledgements vii
Declaration ix
List of Publications xi
List of Figures xvi
List of Tables xxx
List of Abbreviations xxxi

1 Introduction 1
   1.1 Photochemistry ............................................. 3
   1.2 Ultrafast Imaging .......................................... 6
       1.2.1 Spectroscopy ........................................... 6
       1.2.2 Scattering .............................................. 8
   1.3 Computational Foundations ................................. 14
       1.3.1 Simulating Quantum Molecular Dynamics ................. 14
       1.3.2 Interpretation of Simulations and Experiment .......... 17
       1.3.3 Inversion Schemes ..................................... 18
   1.4 Overview of Thesis .......................................... 20

2 Theory 22
   2.1 The Time-Dependent Schrödinger Equation .................. 23
       2.1.1 The Molecular Hamiltonian .............................. 25
       2.1.2 Nonadiabatic Coupling .................................. 27
       2.1.3 Spin-Orbit Coupling .................................... 30
   2.2 Molecular Quantum Dynamics Methods ....................... 32
       2.2.1 Surface Hopping ......................................... 33
## CONTENTS

2.2.2 Multiconfigurational Ehrenfest ........................................ 37
2.3 Scattering Observables .................................................. 41
  2.3.1 X-ray Scattering .................................................... 42
  2.3.2 Electron Scattering ................................................ 44
  2.3.3 Independent Atom Model .......................................... 47
2.4 Data Mining Methods .................................................... 51
  2.4.1 Spatial Clustering .................................................. 51
  2.4.2 Temporal Clustering ................................................ 56

3 Time Series Clustering of Trajectories .................................. 63
  3.1 Introduction .................................................................. 63
  3.2 Model Trajectory Data .................................................. 65
  3.3 Methodology .................................................................. 68
    3.3.1 Feature Selection .................................................. 68
    3.3.2 Similarity Measures ............................................... 71
    3.3.3 Cluster Validity ..................................................... 76
  3.4 Results and Discussion .................................................. 77
    3.4.1 The Multidimensional $L_2$ norm .............................. 77
    3.4.2 Multidimensional Dynamic Time Warping .................. 84
    3.4.3 Comments on Applicability ...................................... 88
  3.5 Periodicity of Angular Coordinates .................................. 89
    3.5.1 Unit Circle Projection ............................................ 91
    3.5.2 A Norm Based on Arc lengths ................................. 92
  3.6 Conclusions .................................................................. 97

4 Pattern Mining Trajectories ............................................... 99
  4.1 Introduction .................................................................. 99
  4.2 A Density Based Spatial Clustering Algorithm ...................... 100
    4.2.1 Problem Definition ............................................... 100
    4.2.2 Algorithm Overview ............................................... 103
    4.2.3 Grid Traversal ...................................................... 105
    4.2.4 Merging Criterion .................................................. 107
    4.2.5 Algorithm Performance .......................................... 110
  4.3 Trajectory Pattern Mining ............................................... 113
    4.3.1 Problem Definition ............................................... 113
    4.3.2 Overcoming the Consecutive Time Step Constraint ......... 116
    4.3.3 Pattern Mining with Relaxed Moving Clusters ............. 117
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.4</td>
<td>Application to 1,3-Cyclohexadiene</td>
<td>125</td>
</tr>
<tr>
<td>4.5</td>
<td>Periodicity of Angular Coordinates</td>
<td>138</td>
</tr>
<tr>
<td>4.6</td>
<td>Conclusions</td>
<td>145</td>
</tr>
<tr>
<td>5</td>
<td>Inversion of Time-Resolved Data via Forward Optimisation</td>
<td>148</td>
</tr>
<tr>
<td>5.1</td>
<td>Introduction</td>
<td>148</td>
</tr>
<tr>
<td>5.2</td>
<td>General Optimisation Framework</td>
<td>151</td>
</tr>
<tr>
<td>5.3</td>
<td>Scattering Observables</td>
<td>154</td>
</tr>
<tr>
<td>5.3.1</td>
<td>Forward Mapping</td>
<td>155</td>
</tr>
<tr>
<td>5.3.2</td>
<td>Apparatus Mapping</td>
<td>156</td>
</tr>
<tr>
<td>5.3.3</td>
<td>Target Function Form</td>
<td>158</td>
</tr>
<tr>
<td>5.3.4</td>
<td>Assessing Fit Quality</td>
<td>158</td>
</tr>
<tr>
<td>5.4</td>
<td>The Optimisation Procedure</td>
<td>160</td>
</tr>
<tr>
<td>5.4.1</td>
<td>Global Optimisation</td>
<td>160</td>
</tr>
<tr>
<td>5.4.2</td>
<td>Two-Step Optimisation</td>
<td>160</td>
</tr>
<tr>
<td>5.5</td>
<td>Data Treatment</td>
<td>162</td>
</tr>
<tr>
<td>5.5.1</td>
<td>Ring-Opening of 1,3-Cyclohexadiene</td>
<td>162</td>
</tr>
<tr>
<td>5.5.2</td>
<td>Photodissociation of CS$_2$</td>
<td>164</td>
</tr>
<tr>
<td>5.6</td>
<td>Results and Discussion</td>
<td>166</td>
</tr>
<tr>
<td>5.6.1</td>
<td>Ring-Opening of 1,3-Cyclohexadiene</td>
<td>166</td>
</tr>
<tr>
<td>5.6.2</td>
<td>Photodissociation of CS$_2$</td>
<td>172</td>
</tr>
<tr>
<td>5.6.3</td>
<td>Convergence Criteria and Weight Sampling</td>
<td>180</td>
</tr>
<tr>
<td>5.6.4</td>
<td>Iterative Weight Generation</td>
<td>181</td>
</tr>
<tr>
<td>5.7</td>
<td>Conclusions</td>
<td>183</td>
</tr>
<tr>
<td>6</td>
<td>Observing Photodissociation by Electron Scattering</td>
<td>185</td>
</tr>
<tr>
<td>6.1</td>
<td>Introduction</td>
<td>185</td>
</tr>
<tr>
<td>6.2</td>
<td>Methodology</td>
<td>187</td>
</tr>
<tr>
<td>6.2.1</td>
<td>Experimental Methods</td>
<td>187</td>
</tr>
<tr>
<td>6.2.2</td>
<td>Data Treatment</td>
<td>188</td>
</tr>
<tr>
<td>6.2.3</td>
<td>Forward Optimisation</td>
<td>190</td>
</tr>
<tr>
<td>6.3</td>
<td>Results and Discussion</td>
<td>191</td>
</tr>
<tr>
<td>6.3.1</td>
<td>Experimental Signal</td>
<td>191</td>
</tr>
<tr>
<td>6.3.2</td>
<td>Difference Pair Distribution Functions</td>
<td>193</td>
</tr>
<tr>
<td>6.3.3</td>
<td>Extraction of Structural Dynamics</td>
<td>199</td>
</tr>
<tr>
<td>6.3.4</td>
<td>Temporal Resolution</td>
<td>207</td>
</tr>
<tr>
<td>6.4</td>
<td>Conclusions</td>
<td>209</td>
</tr>
</tbody>
</table>
List of Figures

1.1 The molecular orbital diagram describing the main bonding interactions that result from the linear combination of atomic orbitals in molecular hydrogen (top panel), and a schematic showcasing the electronic potential energy curves (PEC) of the ground and excited states in a general diatomic molecule (bottom panel). The PECs describe how the potential energy of each state varies with the internuclear distance $R$. .......................... 5

1.2 Jablonski diagram depicting the available radiative and nonradiative decay mechanisms available to a system excited to the second singlet state $S_2$ upon absorption of a photon. Electronic states are represented by bold lines, and vibrational states by thinner lines. .......................... 6

1.3 A typical pump-probe setup used in time-resolved spectroscopy. Here the system is excited by the pump, the resulting dynamics are then probed as a function of the delay time $\Delta t$. The inset demonstrates the main concepts involved in time-resolved photoelectron spectroscopy, where the system is excited to a higher lying electronic state, which exhibits a distribution of vibrational state populations, indicated by the coloured blocks. The probe pulse ionizes the system, and the kinetic energy of the emitted electron is measured. .......................... 7

1.4 Comparison of photon energies and brightness (peak brilliance) achievable by different sources of radiation. Relative to traditional laser systems and synchrotron radiation, XFELs provided light of much higher peak brilliance and energy. Taken from Ref. [55] .......................... 10

1.5 Typical pump-probe setup utilised in XFEL based scattering experiments. The intensity of the scattered x-ray is recorded as a function of the momentum transfer vector $q$ and the angle $\phi$. The momentum transfer vector is given by the difference between the incident and scattered wave vectors, $k_i$ and $k_f$. Here, $q$ is related to the angle $\theta$ at which the x-ray is scattered. .......................... 12

1.6 Schematic representation of the difference between adiabatic and diabatic electronic states. The dashed lines represent the avoided crossing formed in the basis of adiabatic states. .......................... 16
1.7 In the standard approach to analysing experiments (red arrows), the theoretical observable $X_{th}$ is calculated from the wavefunction $|\Psi\rangle$. This is compared with the experimental observable $X_{obs}$ and the molecular structures are inferred. The inverse problem (blue arrow), directly identifies the structures that are encoded in the experimental signal, typically using optimisation or machine learning based methods.

2.1 Schematic showing different shapes of clusters in a two-dimensional feature space, with red, yellow, magenta, and green identifying four different clusters and black indicating outlier points. The left panel shows examples of convex clusters that are linearly separable (as indicated by the dashed lines), with both compact (red, yellow, and green) and diffuse (green) clusters. The right panel shows nonconvex clusters, which may not be linearly separable (as is the case for the red and yellow clusters).

2.2 Depiction of the main concepts in DBSCAN. a) two data points $p$ and $q$ with $\epsilon$ defining the area within which other points are directly density reachable. b) two points $q$ and $m$ are density connected by $p$. c) The building up of core (red), border (yellow) and noise (blue) points. d) Clusters identified by DBSCAN in an example data set. Grey points correspond to noise.

2.3 Result of BANG clustering on an example 2D data set. Colours of the data points represent cluster assignment and the corresponding partitioning of space into a series of blocks is shown. Here the number of levels (i.e. the number of times to perform binary block splitting) is set to 9.

2.4 Example of how similarity between data is measured in the time series approach. There are two clear groups of time series, indicated in blue and orange. In the lower portion of the figure a global distance measure compares each point in time together, giving a single valued measure of similarity over the whole time axis.

2.5 Two sets of 4 2D trajectories $O = [o_1, o_2, o_3, o_4]$ over a set of times $T = [t_1, t_2, t_3]$. Each time-slice represents a different spatial clustering. The grey circle with radius $\epsilon$ represents the size of each cluster. The top panel shows a moving cluster when $\theta_{\text{min}} = 0.75$, each successive spatial cluster shares at least 3 out of 4 member trajectories. The bottom panel shows a convoy of two trajectories $o_1$ and $o_2$ (red) over 3 time steps.
3.1 The photo-induced electrocyclic ring-opening of CHD to HT. Here the states shown are diabatic, with red points indicating conical intersections between them. Following formation of cZc-HT on the ground state, several cis/ trans HT isomers are accessible. 65

3.2 Populations for the set of 100 AIMCE trajectories over the 150 fs simulation period. At early times, the $S_1$ state corresponds to the diabatic 1B state. At 30 fs a maxima in population of the $S_2$ state is observed as the wavepacket passes through the 2A/1B intersection, at subsequent times, the $S_1$ state is of 2A character. Note that in AIMCE a single Gaussian basis function is shared over the three states, thus trajectories with dominant $S_0$ population at later times still share some population with the excited states. 67

3.4 Definitions of high variance internal coordinates and their corresponding maximum values over all time steps of the simulation. The variance of each angular coordinate corresponds to the circular variance, calculated as $\text{var} = 1 - \bar{R}$, where $\bar{R}$ is the mean resultant vector of the angles represented by their coordinates $(\cos, \sin)$. Thus, the variance of each angular feature lies on the range $0 \leq \text{var} \leq 1$. The selected internal coordinates that form the basis of the feature space are, the dihedral angles $\phi_a, \phi_b,$ and $\phi_c$, as well as the bond length $R_{C1-C6}$. 69

3.5 Distribution of the three dihedral angles with colour representing the $R_{C1-C6}$ bond length. Shown over all time steps of the simulation. 70

3.6 Distribution of the data over several time intervals of the simulation. 71

3.7 The main concepts of the DTW algorithm. In the top left we see two univariate time series, both are similar but exhibit a small local variation in frequency and are slightly offset in time. On the right hand side we can see the cost matrix $D$, with elements according to Eqn. 3.2. The elements included in the warping path are coloured according to the type of step taken, and the corresponding alignment can be seen in the bottom left of the figure. Note the magenta vertical steps, which provide a 2:1 mapping of time points. Steps that are coloured green allow the time axis to warp slightly, accounting for small offsets in time. 73
3.8 Results of clustering using the $L_2$ norm as a measure of trajectory similarity. From the top left, the heatmaps correspond to, the average silhouette coefficient $S_{avg}$, the number of resulting clusters $N_C$, the number of trajectories marked as noise $N_{\text{noise}}$. Note, while $S_{avg}$ lies on the range $-1 \leq S_{avg} \leq 1$, the heatmap is scaled to highlight the differences in otherwise small variations. The table corresponds to the better clustering results for $N_C = 3, 4$, which are highlighted by the coloured ellipses in each heatmap.

3.9 The silhouette coefficient ($S_i$) for each trajectory in its respective cluster $C_i$. The size of the shaded area reflects the size of the cluster. Shown are the best results that correspond to $N_C = 3$ and 4, in each of which, 7 trajectories are excluded as noise. The red line indicates the value of $S_{avg}$.

3.10 The set of four product channels identified by clustering. Shown are the final geometries at each of the cluster centres. Here $C_1$ and $C_3$ define the return of the wavepacket to hot CHD, and the ground state $tZc/tZtHT$ product, respectively. At the end of time, $C_2$ likely corresponds to part of the wavepacket that spreads out on the excited state and exhibits larger torsional angles. Finally, when identified, $C_4$ corresponds to a cEc-HT product.

3.11 Plots of the cluster centres in each of the 4 selected features for $\varepsilon = 1.2$ and $\text{MinPts} = 4$ (using the $L_2$ norm). The top left and right panels show the $C_1-C_6$ bond and the central dihedral angle $\phi_b$, whereas the bottom left and right panels show the dihedral angles $\phi_a$ and $\phi_c$. The shaded areas represent the size of the cluster.

3.12 Plots of the cluster centres in each of the 4 selected features for $\varepsilon = 1.15$ and $\text{MinPts} = 3$ (using the $L_2$ norm). The top left and right panels show the $C_1-C_6$ bond and the central dihedral angle $\phi_b$, whereas the bottom left and right panels show the dihedral angles $\phi_a$ and $\phi_c$. The shaded areas represent the size of the cluster.

3.13 Distribution of $S_0$ and $S_1$ state populations at 150 fs. The left panel shows the $N_C = 3$ resulting from $(\varepsilon, \text{MinPts}) = (1.2, 4)$, and the right panel shows the $N_C = 4$ case with $(\varepsilon, \text{MinPts}) = (1.15, 3)$.

3.14 Results of clustering using DTW as a measure of trajectory similarity. Again, the $S_{avg}$ heatmap is scaled to highlight the differences in otherwise small variations. The table corresponds to the better clustering results for $N_C = 3, 4$, which are highlighted by the coloured ellipses in each heatmap.
3.15 The silhouette coefficient ($S_i$) for each trajectory in its respective cluster $C_i$ (using MD-DTW). The left hand panel shows the best $N_C = 4$ case, and the right hand panel, the $N_C = 3$ case. In each case, red lines indicate the value of $S_{avg}$. ................................. 85

3.16 Plots of the cluster centres in each of the 4 selected features for $\varepsilon = 6.75$ and $MinPts = 3$ (using MD-DTW). The top left and right panels show the $C_1-C_6$ bond and the central dihedral angle $\phi_b$, whereas the bottom left and right panels show the dihedral angles $\phi_a$ and $\phi_c$. The shaded areas represent the size of the cluster. ................................. 86

3.17 Distribution of $S_0$ and $S_1$ state populations at time $t = 150$ fs, using MD-DTW. Points colour coded according to their cluster assignment. The left panel shows the $N_c = 3$ case with $(\varepsilon, MinPts) = (7.65, 6)$, and the right panel the $N_c = 4$ resulting from $(\varepsilon, MinPts) = (6.75, 3)$, ................................. 87

3.18 Distribution of the four clusters of trajectories identified using the $L_2$ norm as in section 3.4.1. Note, the plot is with respect to the three dihedral angles and the $C_1-C_6$ bond is omitted from this plot. ................................. 89

3.19 Plots of each of the four selected features as a function of time, with each trajectory coloured according to its cluster. The clusters have been obtained using the original definition of the $L_2$ norm in section 3.4.1. The top left and right panels show the $C_1-C_6$ bond and the central dihedral angle $\phi_b$, respectively, whereas the bottom left and right panels show the dihedral angles $\phi_a$ and $\phi_c$. ................................. 90

3.20 Distribution across the dihedral angles for the four clusters of trajectories identified using the $L_2$ norm in the seven-dimensional space defined by three pairs of angular coordinates $(\cos\phi, \sin\phi)$ and the $C_1-C_6$ distance. Note, that the $C_1-C_6$ bond is omitted from this plot. ................................. 91

3.21 Plots of all clustered trajectories as a function of time, showing each of the four selected features. In this figure, the clusters have been obtained using the $L_2$ norm in a seven-dimensional space defined by three pairs of angular coordinates $(\cos\phi, \sin\phi)$, accounting for periodicity, and the $C_1-C_6$ distance. The top left and right panels show the $C_1-C_6$ bond and the central dihedral angle $\phi_b$, whereas the bottom left and right panels show the dihedral angles $\phi_a$ and $\phi_c$. Note that only minor differences with Fig. 3.19 are apparent. ................................. 92
3.22 Plots of the cluster centres in each of the 4 selected features for \( \varepsilon = 4.3 \) and \( \text{MinPts} = 4 \), using the angles projected onto their \((\cos(\phi), \sin(\phi))\) coordinates on the unit circle. The top left and right panels show the \( C_1-C_6 \) bond length and the central dihedral \( \phi_b \), whereas the bottom left and right panels show the dihedral angles \( \phi_a \) and \( \phi_c \). The shaded areas represent the size of the cluster. Importantly, the centres and distribution of the clusters are near identical to Fig. 3.12.

3.23 Distribution of the four clusters of trajectories identified using the composite norm \( L_{\text{comp}} \) defined according to Eqn. 3.8. Note that the plot shows the three dihedral angles, with the \( C_1-C_6 \) bond omitted.

3.24 Plots showing the clustered trajectories for each of the four selected features, with the clustering done using the composite norm \( L_{\text{comp}} \) defined by Eqn. 3.8. The top left and right panels show the \( C_1-C_6 \) bond and the central dihedral angle \( \phi_b \), whereas the bottom left and right panels show the dihedral angles \( \phi_a \) and \( \phi_c \). Note that the differences compared to the original results in Fig. 3.19 and those shown in Fig. 3.21 are very minor.

3.25 Plots of the cluster centres in each of the 4 selected features for \( \varepsilon = 5.1 \) and \( \text{MinPts} = 4 \), using a composite norm that includes the minimum arclength between a pair of angles. The top left and right panels show the \( C_1-C_6 \) bond length and the central dihedral \( \phi_b \), whereas the bottom left and right panels show the dihedral angles \( \phi_a \) and \( \phi_c \). The shaded areas represent the size of the cluster. Again, note the similarity with Fig. 3.12.

4.1 A two-dimensional example of typical trajectory evolution across three time points \( t = [t_0, t_1, t_2] \). Circled points correspond to areas with considerable density.

4.2 Demonstration of the problems encountered when clustering QMD trajectories. The first two columns correspond to the clustering obtained with parameterised values for the DBSCAN algorithm (column one: \( \varepsilon = 0.4 \) and \( \text{MinPts} = 3 \); column two: \( \varepsilon = 0.3 \) and \( \text{MinPts} = 3 \)). The third column shows the results from the new clustering algorithm introduced in this chapter. Different rows correspond to different times \( t = [t_0, t_1, t_2] \). The colour of each point indicates which cluster the trajectory belongs to, with black representing trajectories marked as noise (and thus not assigned a cluster).
4.3 Demonstration of the total density along a one-dimensional feature space as a sum of contributions from Gaussian kernels centred on individual data points, depicted in blue, orange, green, and magenta. There are two maxima $C_1$ and $C_2$, representing the two clusters to which the data points are assigned. The (kernel) densities for individual data points are shown as dashed lines in the same colour as corresponding data point. 105

4.4 Demonstration of the gradient ascent process used to assign data points to a local maxima, the background heatmap represents the density field. In a) a data point (magenta circle) is projected to its nearest grid point (red cross), b) density is evaluated at each $2 \times N_f$ nearest neighbours (orange crosses), c) a step is taken in the direction that the gradient increases most, d) the process is repeated until a local maxima is reached and the initial data point in a) is assigned to it, as indicated by the magenta cross. 108

4.5 Demonstration of the merging criterion in our method. Panel a) defines the parameters used in the calculation of the merging metric $\Delta_M$. Panel b) shows the results of the initial assignment of maxima based on gradient ascent, where red and green circles indicate semi-isolated points and points assigned to a maxima surrounded by two larger maxima, respectively. Panel c) shows the reassignment of the circled points to the maxima which contribute the most to their density. Panel d) shows the core points used to determine $\rho_0$ for two clusters. Panel e) shows the final clusters after merging. 111

4.6 Results of clustering using the BANG (column one) and DBSCAN (column two) algorithms vs our method (column three). The ground truth, shown in column four, provides the correct labelling of data points according to which of the three distributions they were generated from initially. Clusters are represented by different colours: orange, green, blue, and magenta, and black data points represent noise. Each row corresponds a different data set, demonstrating algorithm performance across different cluster sizes and densities. 112

4.7 Example of a moving cluster for an example comprised of four trajectories at five different times $T = [t_1,t_2,t_3,t_4,t_5]$. With the set parameters, the moving cluster breaks down briefly at $t_3$ before reforming. 116
4.8 Example of a relaxed moving cluster for a series of four trajectories (red and pink circles) over five times ($t_1$–$t_5$). The addition of $\delta_{\text{max}}$ relieves the consecutive time step constraint. Red trajectories correspond to those that are always clustered together. Pink trajectories represent those that travel more loosely with the evolving trajectory ensemble. The relaxed moving cluster criteria allow all given trajectories to be identified as a spatio-cluster.

4.9 Example of six trajectories ($o_1$–$o_6$) moving in time ($t_1$–$t_6$). Each spatial cluster is coloured according to the relaxed moving cluster it should be assigned to, with different relaxed moving clusters represented by the colours green, orange, and purple. The spatial clusters at each time step $t$ are labelled $C_i^t$, where $i$ is the cardinality of the cluster.

4.10 A representation of the input and output of the pattern mining algorithm for the example in Fig. 4.9. The tree of clusters on the left hand side defines the evolution of clusters in time. The nodes are clusters, with levels of the tree corresponding to each time step. The edges include the number of shared trajectories and grey nodes represents clusters with less than $\text{MinPts} = 2$ members. Clusters contained within each output relaxed moving cluster are circled accordingly. The output tree on the right hand side represents the evolution of the relaxed moving clusters, labelled with their initialisation and termination times.

4.11 A more complex example consisting of 20 trajectories over 15 time steps. The input cluster tree on the left hand side contains five branches, which are coloured accordingly. Grey nodes represent small clusters corresponding to short term deviations. At time $t_{13}$ a group of five trajectories briefly splits into two clusters (highlighted in the grey box). The identified relaxed moving clusters in tree form according to the specified parameters is shown on the right.

4.12 The tree of output moving clusters identified by the high separability set of parameters identified in Table 4.2. The times at which moving clusters split into new branches are indicated in red, and the dominant population of the moving cluster is indicated by the blue and yellow edges. The set of structures shown correspond to the geometries at the cluster centres.
4.13 The tree of output moving clusters identified in the medium separability set of parameters provided in Table 4.2. Note the absence of a moving cluster that contains the small amount cEc-HT product. The other moving clusters are almost unchanged compared to the higher separability case, and are all detectable up to $\theta_{\text{min}} = 0.70$.  

4.14 (Top panel) Number of trajectories in each moving cluster $M_1$-$M_7$ as a function of time (fs). The dashed line segments in the red line representing cluster $M_6$ indicates the times when the moving cluster briefly breaks down. (Bottom panel) The initial (solid circle) and final (empty circle) $(x,y) = (S_0, S_1)$ population for each cluster, with the colours of the clusters the same as in the top panel.  

4.15 Changes in $S_0$ (left panel) and $S_1$ (right panel) populations for the trajectories re-assigned from cluster $M_3$ to $M_2$. The $x$-axis in the left (right) panel shows the $S_0$ ($S_1$) population at the time the cluster $M_2$ is initialised ($t_1=34$ fs), and the $y$-axis the corresponding $S_0$ ($S_1$) population at the time when cluster $M_2$ is terminated ($t_2=66$ fs). That most points are above the $y = x$ line in the left panel, and below in the right panel, indicates a net population transfer $S_1 \rightarrow S_0$. For each trajectory, the time at which the transition between the moving clusters occurs is indicated by the color of the data point, according to the colour mapping shown below the two panels.  

4.16 Distribution of the three dihedral angles $\phi_a$, $\phi_b$, and $\phi_c$ over the times each relaxed moving cluster exists. Each row represents one of the relaxed moving clusters corresponding to a leaf node in Fig. 4.12, and each column the dihedral angle. In each case, the set of trajectories plotted correspond to those that are contained within the moving cluster for at least 90% of its duration. Here, the regions of highest density within the clusters are clear, as indicated by a higher concentrate of points, thus observing a darker shade.  

4.17 A series of reduced clusters constructed from the core trajectories within the moving clusters corresponding to leaf nodes. Core trajectories refer to those that are contained within their respective moving cluster for at least 90% of their total duration. The top left and right panels show the $C_1$-$C_6$ bond and the central dihedral angle $\phi_b$, whereas the bottom left and right panels show the dihedral angles $\phi_a$ and $\phi_c$. 
4.18 The tree of output moving clusters identified by projection of the angles onto their $(\cos, \sin)$ coordinates on the unit circle. Note the absence of a moving cluster that contains the small amount cEc-HT product, but overall this result is similar to that previously reported in Fig. 4.13. ........................................ 140

4.19 (Top panel) Number of trajectories in each moving cluster $M_1$-$M_5$ as a function of time (fs). (Bottom panel) The initial (solid circle) and final (empty circle) $(x, y) = (S_0, S_1)$ population for each cluster, with the colours of the clusters the same as in the top panel. ........................................ 141

4.20 Changes in $S_0$ (left panel) and $S_1$ (right panel) populations for the trajectories re-assigned from cluster $M_3$ to $M_2$. The $x$-axis in the left (right) panel shows the $S_0$ ($S_1$) population at the time the cluster $M_2$ is initialised ($t_1=36$ fs), and the $y$-axis the corresponding $S_0$ ($S_1$) population at the time when cluster $M_2$ is terminated ($t_2=76$ fs). That most points are above the $y = x$ line in the left panel, and below in the right panel, indicates a net population transfer $S_1 \rightarrow S_0$. For each trajectory, the time at which the transition between the moving clusters occurs is indicated by the color of the data point, according to the colour mapping shown below the two panels. ........................................ 143

4.21 Distribution of the three dihedral angles $\phi_a$, $\phi_b$ and $\phi_c$ over the times each relaxed moving cluster exists. Each row represents one of the relaxed moving clusters corresponding to a leaf node in Fig. 4.18, and each column the dihedral angle. In each case, the set of trajectories plotted correspond to those that are contained within the moving cluster for at least 90% of its duration. Here, the regions of highest density within the clusters are clear, as indicated by a higher concentrate of points, thus observing a darker shade. ........................................ 144

5.1 Overview of the optimisation procedure in which the target function is minimised with respect to the set of trajectory weights $\mathbf{w}$ revealing the key contributions to the experimental observable. The case shown here represents the problem of the molecular structure determination in scattering experiments. The bank of trajectories from which the model wavefunction is refined is shown in blue, inputs to the optimisation i.e. the experimental and model observables are shown in red, and the output key molecular structures are shown in green. ........................................ 153
5.2 Convergence of a) the ring-closed fraction, b) ring-opened fraction, c) the target function, and d) the relative absolute error (RAE) as a function of the scaling factor $x$ and time-zero shift $t_0$. Each color represents a different value of $x$ ranging from 0.7 to 1.3.  

5.3 The unoptimised initial excitation fraction $\gamma(q)$ and the optimised excitation fraction $\gamma_x(q)$ that results from scaling with $x = 1.3$.  

5.4 The C$_1$–C$_6$ bond length for the eight dominant trajectories with a weight greater than 1% filtered out by the optimisation.  

5.5 The squared distances $D^2_b$ (left panel) from the best set of weights and the variance $v^2_w$ (right panel) of the weights for a series of CHD optimisations with different scaling factors $x$ and $t_0$ shifts, plotted against the relative absolute error (RAE). Each colour represents a different $x$ and shade represents a different $t_0$ value ranging from -38 to -14 fs (the darker the shade, the earlier the $t_0$ shift).  

5.6 Comparison of the experiment $\%\Delta I_{\text{exp}}(q,t)$, unoptimised $\%\Delta I_{\text{th}}(q,t)$, and optimised $\%\Delta I_{\text{mod}}(q,t)$ signals for a series of time delays.  

5.7 Convergence of the $t_0$ fitting procedure with respect to different $t_0$ and $\tau_c$ values. In the top panel is the convergence of the target function $F(w, c)$, the middle panel the RAE, and the bottom the RMSE.  

5.8 The best fit for step one of the optimisation where $t_0 = -83$ fs and $\tau_c = 230$ fs. The fit is performed using integrated signal $\%\Delta I_{\text{int}}$, and results in an optimised excitation value of $\gamma = 3\%$. The shaded area corresponds to the standard deviation of the trajectory ensemble from the model calculated assuming equal weights.  

5.9 Convergence of the target function $F(w, c)$ for a series of different $t_0$ shifts and $p_{\text{min}}^{\text{conf}}$ thresholds.  

5.10 Convergence of the second step on the CS$_2$ optimisation as a function of the confidence threshold $p_{\text{min}}^{\text{conf}}$ and $t_0$ values. Panel a) branching ratio, b) bound fraction at 1 ps, c) excitation fraction $\gamma$, and d) the relative absolute error (RAE).  

5.11 The squared distances $D^2_b$ (left panel) from the best set of weights and the variance $v^2_w$ (right panel) of the weights for a series of CS$_2$ optimisations with different confidence matrix thresholds $p_{\text{min}}^{\text{conf}}$ and $t_0$ shifts, plotted against the relative absolute error (RAE). Each colour represents a different $p_{\text{min}}^{\text{conf}}$ and shades represents a different $t_0$ value ranging from +17 to -83 fs (the darker the shade, the earlier the $t_0$ shift).
5.12 Comparison of the experiment $\%\Delta I_{\text{exp}}(s,t)$, unoptimised $\%\Delta I_{\text{th}}(s,t)$, and optimised $\%\Delta I_{\text{mod}}(s,t)$ signals for a series of time delays.  
5.13 Convergence of the target function with respect to the number of initial conditions $N_{\text{init}}$.  
5.14 Convergence of the target function with respect to the number of iterations (top panel), and the size of step taken at each iteration (bottom panel). The case shown corresponds to the best optimisation with weights $w_{\text{best}}$.  
5.15 Heatmap of trajectory weights $w$ for the best optimisations that result from each of upper bounding limits $U_{\text{bound}}$ in the iterative approach to weight sampling.  

6.1 Heatmaps of the a) experimentally observed $\Delta s M_{\text{exp}}(s,t)$ and b) the theoretical $\Delta s M_{\text{th}}(s,t)$ signal, calculated from the ensemble average of 197 trajectories. Negative time delays correspond to those prior to arrival of the pump pulse, and positive time delays define the change in the signal observed in reference to the system in its ground state equilibrium position.  
6.2 Heatmaps of the a) experimentally observed $\%\Delta I_{\text{exp}}(s,t)$ and b) the theoretically calculated $\%\Delta I_{\text{th}}(s,t)$, calculated from the ensemble average of 197 trajectories. Negative time delays correspond to those prior to arrival of the pump pulse, and positive time delays define the change in the signal observed in reference to the system in its ground state equilibrium position.  
6.3 Heatmaps of the a) the experimentally observed PDF and b) the theoretically calculated PDF, where the theoretical $\Delta s M$ signal is truncated to match the experimental range $0.95 \leq s \leq 12.25 \text{ Å}^{-1}$.  
6.4 The left hand side shows the $\Delta \text{PDF}(R,t)$ calculated from the theoretical $\Delta s M_{\text{th}}(s,t)$ signal. Regions A-D that correspond to the main features are highlighted. On the right hand side one can see the averaged $\Delta \text{PDF}(R,t)$ over each of these four regions. Regions A and C are the two depletions, whereas B and D are the two enhancements. These correspond to the ground state (GS) Wigner distribution and the evolution of the nuclear wavepacket upon the excited state (ES) respectively.  
6.5 Comparison of experimental $\Delta \text{PDF}_{\text{exp}}$ and theoretical difference pair distribution functions for a series of individual time delays. Here $\Delta \text{PDF}_{\text{th}}$ and $\Delta \text{PDF}_{\text{th}}$ (Truncated) refer to the theoretical distribution functions calculated over the ranges $0 \leq s \leq 12.25 \text{ Å}^{-1}$ and $0.925 \leq s \leq 12.25 \text{ Å}^{-1}$ respectively.
6.6 Comparison of the experimentally observed $\Delta s M_{\text{exp}}$ and theoretical $\Delta s M_{\text{th}}$ signals, shown for a series of individual time delays. Also plotted is $\Delta s M_{\text{exp}} e^{-\alpha s^2}$, which corresponds to the $\Delta s M_{\text{exp}}$ signal weighted by the exponential dampening factor included in the sine-transform of Eqn. 6.5. . . . . . . . . . . . . . 198

6.7 Atomic C−S distances (Å) and S−C−S bending angles (degrees) in CS$_2$ for the four most significant dissociating trajectories. The dashed black line indicates the ground state equilibrium C−S bond length. The two top sets of panels show the two dominant singlet trajectories, while the two bottom sets of panels show the two dominant triplet trajectories. All trajectories are shown to the point of dissociation. . . . . . . . . . . . . . 201

6.8 Populations of the bound, singlet ($^1D$), and triplet ($^3P$) channels determined from the optimised set of model trajectories. . . . . . . . . . . . . . 202

6.9 Decomposition of the optimised model $\Delta I_{\text{mod}}(s, t)$ signal (shown in panel a) into b) bound, c) singlet ($^1D$), and d) triplet ($^3P$) contributions. Black lines mark dissociation times calculated from the corresponding optimised trajectories in table 6.1. . . . . . . . . . . . . . . . . . . . . 205

6.10 Rise in the intensity at the centre of the enhancement just below 2 Å$^{-1}$. Shown are the intensity rises determined from experiment $\% \Delta I_{\text{exp}}$, the total signal of the optimised model $\% \Delta I_{\text{mod}}$, and the contributions from both the individual singlet ($^1D$) and triplet ($^3P$) channels. . . . . . . . . . 206

6.11 Heatmaps of the theoretical difference signal $\% \Delta I_{\text{th}}(s, t)$ at a range of values of the Gaussian width $\tau_c$, used in the convolution that approximates the IRF. Note, all difference signals $\% \Delta I_{\text{th}}(s, t)$ are calculated from the ensemble average of 197 trajectories. Panels a) through d) correspond to widths of 0, 80, 150, and 230 fs, with 230 fs the resolution observed in this work. At 80 fs, the vibrational fine structure becomes apparent. . . . . 208

A.1 Effect of the distribution of angular data on identifying the correct clustering result. In the top and bottom row we have two different data sets, both spanning the range $[0, 2\pi]$. In column one we colour the data points according to the Gaussian distribution it was generated from. In columns two and three we see the result of clustering using an $L_2$ norm on the raw angular data $\phi$ and the projected values $(\cos, \sin)$ respectively. . . . . . . . 217
A.2 Effect of the distribution of angular data on identifying the correct clustering result in a 2D data set with two angular coordinates. The top and bottom rows correspond to the angular coordinates $\phi_1$ and $\phi_2$. In column one we colour the data points according to the Gaussian distribution it was generated from. In columns two and three we see the result of clustering using an $L_2$ norm on the raw angular data $\phi$ and the projected values $(\cos, \sin)$ respectively.
List of Tables

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1</td>
<td>The maximum variance over all time for a set of internal coordinates.</td>
<td>69</td>
</tr>
<tr>
<td>4.1</td>
<td>Hashmap of trajectories and their corresponding time sets over which they are contained within the relaxed moving cluster in Fig. 4.8.</td>
<td>117</td>
</tr>
<tr>
<td>4.2</td>
<td>The number of identified relaxed moving clusters, $N_{MC}$, for various sets of parameters (or parameter ranges) for the pattern mining algorithm. The first row lists the initial ranges of guess parameters from the input tree of spatial clusters over time.</td>
<td>125</td>
</tr>
<tr>
<td>5.1</td>
<td>The optimal input parameters for the CHD data set.</td>
<td>168</td>
</tr>
<tr>
<td>5.2</td>
<td>Weights $w_{best}$ of the eight dominant CHD trajectories resulting from the best optimisation with $(t_0, x) = (-38 \text{ fs}, 1.3)$.</td>
<td>169</td>
</tr>
<tr>
<td>5.3</td>
<td>The optimal input parameters for the CS$_2$ data set.</td>
<td>175</td>
</tr>
<tr>
<td>5.4</td>
<td>Weights of the eight dominant CS$<em>2$ trajectories resulting from the best optimisation with $[t_0, \tau_c, \gamma, p</em>{conf}^{\text{min}}] = [-83 \text{ fs}, 230 \text{ fs}, 3.4%, 0]$.</td>
<td>176</td>
</tr>
<tr>
<td>6.1</td>
<td>Summary of the 8 dominant trajectories that result from the best optimisation. These include two bound, three singlet and three triplet trajectories with a total weight of 28 %, 16.45 %, and 55.55 %. Dissociation times are given where applicable.</td>
<td>199</td>
</tr>
</tbody>
</table>
List of Abbreviations

ADC    Algebraic Diagrammatic Construction Theory
AIMCE  Ab-initio Multiconfigurational Ehrenfest
AIMS   Ab-initio Multiple Spawning
ATP    Adenosine triphosphate
BAT    Bra-ket averaged Taylor expansion
BOA    Born-Oppenheimer approximation
CASPT2 Complete Active Space 2\textsuperscript{nd} Order Perturbation Theory
CASSCF Complete Active Space Self-Consistent Field
CCD    Charge coupled device
CEI    Coulomb explosion imaging
CHD    1,3-Cyclohexadiene
CI     Configuration Interaction
CMS    Coherent-mixed scattering
CPA    Chirped pulse amplification
dd-vMCG direct dynamics variational Multiconfigurational Gaussian
DTW    Dynamic time-warping
FWHM   Full width at half maximum
GA     Genetic algorithm
GeV    Giga-electron volt
HT     1,3,5-Hexatriene
IAM    Independent atom model
IC     Internal conversion
IRF    Instrument response function
ISC    Intersystem crossing
LCLS   Linac Coherent Light Source
LCSS   Longest common subsequence
MCE    Multiconfigurational Ehrenfest
MCSCF  MultiConfigurational Self-Consistent Field
MCTDH  Multiconfigurational Time-Dependent Hartree
MD-DTW Multidimensional dynamic time-warping
MeV    Mega-electron volt
<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>MGBF</td>
<td>Multi Gaussian basis function</td>
</tr>
<tr>
<td>MRCI</td>
<td>Multireference Configuration Interaction</td>
</tr>
<tr>
<td>NACME</td>
<td>Nonadiabatic coupling matrix element</td>
</tr>
<tr>
<td>NN</td>
<td>Neural network</td>
</tr>
<tr>
<td>PAD</td>
<td>Photoelectron angular distribution</td>
</tr>
<tr>
<td>PCA</td>
<td>Principal component analysis</td>
</tr>
<tr>
<td>PDF</td>
<td>Pair distribution function</td>
</tr>
<tr>
<td>PEC</td>
<td>Potential energy curve</td>
</tr>
<tr>
<td>PES</td>
<td>Potential energy surface</td>
</tr>
<tr>
<td>QMD</td>
<td>Quantum molecular dynamics</td>
</tr>
<tr>
<td>RAE</td>
<td>Relative absolute error</td>
</tr>
<tr>
<td>RMSE</td>
<td>Root mean squared error</td>
</tr>
<tr>
<td>SASE</td>
<td>Self-amplified spontaneous emission</td>
</tr>
<tr>
<td>SA</td>
<td>Simulated annealing</td>
</tr>
<tr>
<td>SGBF</td>
<td>Single Gaussian basis function</td>
</tr>
<tr>
<td>SH</td>
<td>Surface Hopping</td>
</tr>
<tr>
<td>SOC</td>
<td>Spin-orbit coupling</td>
</tr>
<tr>
<td>TBF</td>
<td>Trajectory basis function</td>
</tr>
<tr>
<td>TDDFT</td>
<td>Time-Dependent Density Functional Theory</td>
</tr>
<tr>
<td>TDSE</td>
<td>Time-dependent Schrödinger equation</td>
</tr>
<tr>
<td>tICA</td>
<td>Time-independent component analysis</td>
</tr>
<tr>
<td>TISE</td>
<td>Time-independent Schrödinger equation</td>
</tr>
<tr>
<td>TRPES</td>
<td>Time-resolved photoelectron spectroscopy</td>
</tr>
<tr>
<td>TRRLS</td>
<td>Trust-region reflective least-squares</td>
</tr>
<tr>
<td>UED</td>
<td>Ultrafast electron diffraction</td>
</tr>
<tr>
<td>UXS</td>
<td>Ultrafast x-ray scattering</td>
</tr>
<tr>
<td>XFEL</td>
<td>X-ray free electron laser</td>
</tr>
<tr>
<td>XMS-CASPT2</td>
<td>Extended Multistate Complete Active Space 2nd Order Perturbation Theory</td>
</tr>
<tr>
<td>XRS</td>
<td>X-ray scattering</td>
</tr>
<tr>
<td>XUV</td>
<td>Extreme ultraviolet</td>
</tr>
</tbody>
</table>
Chapter 1

Introduction

The question of how light interacts with matter holds profound significance for our understanding of the universe and the development of modern technology. Its exploration dates back to ancient times, with Greek philosopher Empedocles proposing one of the earliest known theories of vision around the 5th century BCE [1]. However, a comprehensive understanding of the behaviour of light was not achieved until Maxwell introduced his classical theory of electromagnetism in 1873 [2]. In 1901, Max Planck introduced the concept of quantization, in which electromagnetic energy is absorbed or emitted in discrete packets (quanta) [3], laying the early foundations of quantum theory. Building on Planck’s work, Niels Bohr introduced his planetary model of the atom, in which the electrons orbit the nucleus in quantized energy levels [4]. In Bohr’s model, the electrons may transition between energy levels by absorbing or emitting a photon with an energy equal to the separation between energy levels, allowing the explanation of the spectral emission and absorption lines from small atoms.

A deeper understanding into both the structure of atoms and light-induced electronic transitions, emerged with Erwin Schrödinger ‘s mathematical formulation of quantum mechanics [5]. As a consequence of their small de-Broglie wavelength, quantum particles (such as electrons) exhibit both the behaviour of waves and particles, and can be described by Schrödinger ‘s wave equations [6, 7]. In the quantum model of the atom, electrons occupy probabilistic clouds of electron density, known as atomic orbitals. Each of these orbitals has an associated energy and is characterized by a series of quantum numbers that define its size, shape, and spin. Atomic orbitals are solutions to Schrödinger ‘s famous wave equation, and describe the probability of finding an electron with a given spin at different positions around the nuclei. Just like
in Bohr’s model, the absorption or emission of a photon by an atom can induce an electronic transition. However, Schrödinger’s model provided an explanation for the splitting of energy levels that result from electron-electron and spin interactions, all of which are not captured by Bohr’s model.

The interaction between light and a molecular system is inherently more complex. Upon photoexcitation to a higher lying electronic state, a change in the electronic density of the molecule occurs, inducing a photochemical reaction in which the electrons and nuclei move in space and time. The outcome of the reaction depends on the available pathways through the manifold of excited states, which may lead to the formation or breaking of chemical bonds and result in a range of different products [8]. Consequently, studying the ensuing reaction dynamics plays a crucial role in understanding the formation of different products, and the nature of molecular motion itself. In the past, it was typical to infer reaction pathways from the distribution of products, regardless of whether the reactions were induced by light or not. Often this would involve isotopic labelling, in which certain atoms within the molecule are changed for heavier isotopes, allowing one to confirm the formation or breaking of certain bonds [9, 10]. In addition, studies of reaction kinetics provided information on the rates over which different products formed [11]. In more recent years, advances in time-resolved techniques, such as ultrafast spectroscopy and scattering, have afforded us with the ability to directly track both ground and excited state reaction dynamics [12–16]. These techniques have enabled dynamic observations of protein folding [17], enzymatic reactions such as the hydrolysis of adenosine triphosphate (ATP) by myosin [18], and the photoexcited dynamics of molecules such as rhodopsin [19], which is involved in vision. Now, with next generation light sources such as free-electron lasers, reaction dynamics can be studied at an even deeper level, observing fundamental phenomena such as electron transfer and nuclear vibrations [20–23].

This thesis explores the synergistic combination of ultrafast experimental methods and computational simulations to unravel the intricate dynamics of photochemical reactions. The primary focus lies on the development of new computational methods that facilitate the interpretation of experimental data and simulations, providing unprecedented insights into complex photoexcited processes. We shall explore how methods from unsupervised machine learning can be utilised to identify the key reaction pathways observed in simulations of photoexcited dynamics, the extraction of information from time-resolved experimental data, and the use of ultrafast scattering techniques for imaging photochemical reactions in the gas phase.
1.1 Photochemistry

The ground state of a molecule describes a stable configuration of the electrons around the nuclei, in which the electrons are contained within molecular orbitals. The molecular orbitals can be expressed as a linear combination of atomic orbitals (LCAOs). Such a representation is only formally exact when the molecular orbitals are expanded in a complete basis of atomic orbitals. Practically, error is introduced when the basis is truncated. More broadly, it is worth noting that the molecular orbitals are mathematical rather than physical constructs, as exemplified by different (but equally valid) representations such as natural or canonical orbitals. In the upper panel of Fig. 1.1, we see the molecular orbital diagram that describes the main bonding interactions in molecular hydrogen. Here the constructive interference of the two atomic hydrogen 1s orbitals results in a doubly occupied $\sigma_g$ molecular orbital; whereas the destructive interference of the atomic orbitals leads to an antibonding $\sigma_u^*$ molecular orbital. Upon absorption of a photon with the correct energy, we observe a $\sigma_g \rightarrow \sigma_u^*$ transition, indicated by the red excited electron.

The electronic transition results in excitation to a higher energy electronic state, which is plotted for the general example of a diatomic molecule in the bottom panel of Fig. 1.1. Here, the potential energy of each electronic state as a function of the internuclear distance defines a potential energy curve (PEC). The gradient of the PEC corresponds to the electrostatic force the nuclei experience from the electrons, which is the driving force behind nuclear motion. Turning to the ground state (GS) PEC in the figure, we observe a potential well corresponding to the equilibrium geometry of the system, where it remains unless a source of additional energy is applied. In comparison, the excited state (ES) does not exhibit such a defined well, at least not at the same internuclear distance. The excited states are generally unstable and tend to decay back to a lower energy state. Upon excitation the nuclei move along the PEC in line with the gradient of the populated excited PEC (i.e. according to the changing force of the electrons), as indicated by the red arrows. In this example, this leads to an increase in the internuclear distance, represented by the ball and stick structures shown at various points along the PEC. Once the system reaches the part of the PEC in which the energy separation of the two involved states is small, the states may exhibit a degree of coupling, facilitating population transfer between the excited and ground states. On decay of the excited system to the ground state, two

1. Note, that the potential energy curves in the bottom panel are not intended to reflect the photoexcited dynamics of molecular hydrogen, and instead serve as a conceptual schematic.
1.1 Photochemistry

pathways emerge, corresponding to a return to the ground state equilibrium structure, and further elongation of the bond which ends in dissociation. It is worth noting that in systems with more than one internal degree of freedom, the potential energies as a function of the nuclear coordinates form a multidimensional potential energy surface (PES), and while they are often visualised in terms of internal coordinates, they can be represented in the basis of any alternative coordinate system.

The decay pathways a photoexcited system can take through a manifold of electronic states is complex, with many competing processes occurring over a range of timescales [8]. Generally, there are two classes of decay mechanisms, radiative and nonradiative decay. These mechanisms are commonly summarised in the form of Jablonski diagrams [24], as shown in Fig. 1.2. This shows a schematic representation of the states in the system and can be used to trace the different decay mechanisms available to the system following absorption of a photon from a singlet ground state \( S_0 \) to an excited state (in this case the second singlet excited state \( S_2 \)). The figure also includes the first excited singlet state \( S_1 \) and the lowest-energy triplet state \( T_1 \).

Electronic states are identified by their energy and their total spin angular momentum \( S \), with singlet and triplet states having values of \( S = 0 \) and \( 1 \), respectively. Note that while electronic states may have a value of \( S = 1/2 \), corresponding to a doublet state with an unpaired electron, these states do not have a role in this thesis.

Radiative decay mechanics include fluorescence and phosphorescence, which both involve the emission of photon. Fluorescence refers to radiative decay between singlet states, whereas phosphorescence involves relaxation from an excited triplet state to a singlet state. The latter is symmetry forbidden and therefore occurs on slower time scales than the former. The other class of decay processes, nonradiative decay, does not involve the emission of a photon, and is typically faster than the radiative decay mechanisms. Nonradiative decay includes internal conversion (IC) and intersystem crossing (ISC), referring to relaxation between states of the same and differing spin respectively. However, population transfer between different spin manifolds (for instance singlet and triplet states) requires spin-orbit coupling, which is generally very small unless heavier atoms are present in the molecule. The work in this thesis is concerned primarily with the nonradiative decay pathways available to a molecular ensemble following excitation. Understanding how such photochemical reactions proceed, and
Figure 1.1: The molecular orbital diagram describing the main bonding interactions that result from the linear combination of atomic orbitals in molecular hydrogen (top panel), and a schematic showcasing the electronic potential energy curves (PEC) of the ground and excited states in a general diatomic molecule (bottom panel). The PECs describe how the potential energy of each state varies with the internuclear distance $R$.

the possible range of products that may form from a number of competing decay pathways, is crucial. However, the time and length scales, and the fundamentally quantum mechanical nature of these process, makes their observation and simulation a very challenging task.
1.2 Ultrafast Imaging

1.2.1 Spectroscopy

In the 1950s-70s, imaging of photochemical reactions was achieved through pump-probe based spectroscopy schemes [25–27], in which a pump laser excites the system under study, initiating a reaction, after a suitable delay time, a second laser probes the evolving dynamics, as shown in Fig. 1.3. The choice of pump laser varies, depending on the specific experiment. Notable early examples include transient absorption spectroscopy, time-resolved fluorescence spectroscopy, and time-resolved infrared spectroscopy [28–30]. These allowed for the determination of excited state lifetimes, fluorescence time scales, and quantum yields. Such studies were vital in the early understanding of crystal and in-solution dynamics. However, they were limited by the available temporal resolution of the time, providing insights into processes that typically occur on the nanosecond ($10^{-9}$ s) timescale at best. Donna Strickland and Gerard Mourou’s development of chirped pulse amplification (CPA) in 1985, provided a method for generating high intensity and ultrashort pulses of light on the femtosecond ($10^{-15}$) time scale [31]. Both Strickland and Mourou went on to receive the 2018 Nobel Prize in Physics for their work. The advent of the femtosecond laser

---

**Figure 1.2:** Jablonski diagram depicting the available radiative and nonradiative decay mechanisms available to a system excited to the second singlet state $S_2$ upon absorption of a photon. Electronic states are represented by bold lines, and vibrational states by thinner lines.
made afforded the possibility of observing dynamics occurring on the time scale of single molecule vibrations. This is exactly what Ahmed Zewail then set out to achieve, and he was awarded the 1999 Nobel Prize in Chemistry for directly observing and characterising the transition state structures involved in several chemical reactions through femtosecond spectroscopy [32, 33]. Zewail’s work marked the beginning of the field of femtochemistry, and along with others, he went on to study photoexcited reaction dynamics using the same laser technology [34–36].

Figure 1.3: A typical pump-probe setup used in time-resolved spectroscopy. Here the system is excited by the pump, the resulting dynamics are then probed as a function of the delay time $\Delta t$. The inset demonstrates the main concepts involved in time-resolved photoelectron spectroscopy, where the system is excited to a higher lying electronic state, which exhibits a distribution of vibrational state populations, indicated by the coloured blocks. The probe pulse ionizes the system, and the kinetic energy of the emitted electron is measured.

An important application of femtosecond lasers to pump-probe spectroscopy involved the development of femtosecond time-resolved photoelectron spectroscopy (TRPES), which probes the changes in electronic state populations over time [37–41]. Following excitation with the pump pulse, the system is ionised by the probe pulse and the distribution of the emitted electrons kinetic energies recorded at the detector, indicated by the inset of Fig. 1.3. The kinetic energy $E_k$ of an ejected electron is related to the binding energy $E_B$ of the occupied quantum state as $E_B = h\nu - E_k$, where $h\nu$ is the energy of the absorbed photon. Therefore, the measured kinetic energy is characteristic of the state from which the electron is ejected, providing direct
information on the evolution of state populations. In comparison to other methods of ultrafast imaging, TRPES has the advantage that any of the occupied states may be ionized providing the right probe energy, and therefore there are no dark states. In addition to providing information on the electronic state populations, TRPES can also provide information on the important coupled vibrational modes, which correspond to vibrational states of the given electronic state. Upon excitation to a higher lying electronic state with the pump pulse, the series of populated electronic states can have excess vibrational energy. This results in branching of the observed kinetic energy distribution for a given electronic state, providing information on the coupled vibrational modes that play a role in driving the nuclear dynamics along the electronic states. This concept is demonstrated in the inset of Fig. 1.3, where each distribution of vibrational states (indicated by coloured blocks) has a slightly different binding energy $E_B$.

### 1.2.2 Scattering

Techniques based on the phenomena of scattering provide an alternative method of imaging, and form a central part of this thesis. In the most general sense, scattering involves measuring the material response of a system to an incident wave, which may correspond to electrons, neutrons, protons, or electromagnetic radiation. The interaction between the incident wave and the material system results in the wave being scattered, changing direction, and may involve a transfer of energy between the wave and the system \cite{42}. When energy is transferred, this corresponds to inelastic scattering. In contrast, the case that involves no energy transfer is known as elastic scattering.

Of particular importance to this thesis is the interaction between x-rays and matter. X-ray based techniques have a rich history in molecular imaging, both in a static and time-resolved context. The scattering of x-rays results from the interaction of the electric field of an incident x-ray photon with the electrons in the sample. Most of the x-rays scatter from an interaction with the core electrons, which are more tightly bound to the nuclei in comparison to valence electrons. Therefore, this indirectly yields information on the molecular structure. Early x-ray crystallography experiments, pioneered by William Bragg, involved determining the arrangement of atoms within ordered crystal structures \cite{43}. Through analysing the diffraction pattern observed by
1.2 Ultrafast Imaging

A sodium chloride crystal, Bragg was able to determine the arrangement of atoms within the crystal, paving the way for modern x-ray based imaging techniques. To this day, crystallography plays an important role in solid state physics, and the imaging of biological macromolecules and proteins [44–46].

Generally, early crystallography experiments that employed standard x-ray tubes were only capable of operating at a fraction of the peak brilliance that is possible today. Signals were therefore only collectable for crystalline samples in which the total scattered intensity is amplified by constructive interference of the scattered waves from individual atoms within the crystal. It was not until the development of brighter x-ray sources that imaging outside of the solid state became a reality. The introduction of synchrotrons greatly advanced the capability of x-ray based methods. By accelerating electrons to near light speeds along a circular path, synchrotrons are capable of emitting x-ray radiation that is both brighter and higher in energy in comparison to conventional x-ray tubes [47]. With these brighter, more tunable sources of light, it became possible to image less ordered solids and solution phase samples [48]. In the 1970s-80s, further advancements in temporal resolution and detector technology provided the opportunity to conduct time-resolved x-ray diffraction studies using synchrotron radiation [49–52]. These early time-resolved studies involved the observation of changes in crystalline structures to temperature and light, observing phenomena such as lattice vibrations and phase transitions [53, 54]. The development of pump-probe technology was also instrumental to time-resolved studies, this allowed the investigation of excited state processes occurring on the nano ($10^{-9}$ s) to picosecond ($10^{-12}$ s) timescales. However, direct observation of excited state dynamics occurring on more fundamental timescales was not yet possible with synchrotron radiation.

Following the development of the femtosecond laser, and further advances in accelerator and detector technology, the world saw the introduction of the first x-ray free electron laser (XFEL), the Linac Coherent Light Source (LCLS) [56], in 2009. Since then, a number of other XFEL facilities have been constructed around the globe². XFELs provide highly tunable, high brilliance, ultrashort coherent pulses of light, allowing for imaging on more fundamental timescales [55, 57]. Now, FELs are capable of producing pulses of light in the attosecond regime, providing a way to directly observe not only single molecule vibrations, but also electronic dynamics [58].

---

2. Existing XFEL's include: EuXFEL (Germany), FERMI (Italy), SwissFEL (Switzerland), SACLA (Japan), PAL-XFEL (South Korea), SHINE (China), LCLS-I (USA), LCLS-II (USA). In 2022, £3.2m of funding was also approved to explore the development of an XFEL in the United Kingdom.
With much higher peak brilliance’s, XFELs are capable of not only imaging in the solid phase, but also in the gas phase where there is a significantly lower concentration of atoms [59]. Gas phase experiments allow the observation of the quantum mechanics that governs excited state dynamics, without having to consider bulk properties of the material or solvation effects. Moreover, the higher energy range of the produced x-rays offer several advantages, such as increased penetration depth in crystalline samples, collection of data over a higher experimental range, and improved spatial resolution [57]. A comparison of the peak brilliance and the available range of x-ray energies between different radiation sources is given in Fig. 1.4.

XFELs achieve these benefits through operating based on the principle of self-amplified spontaneous emission (SASE) [57]. Following generation of an electron beam from a radiofrequency electron gun, electrons are accelerated close to the speed of light in a linear accelerator\(^3\). The electrons then follow a slalom course through an un-
1.2 Ultrafast Imaging

dulator, comprised of a series of magnets arranged in an alternating fashion. The alternating magnetic fields results in the oscillation of the electron beam, and the first through bunches that pass through emit x-ray radiation via the process of stimulated emission. The initially produced x-rays are weak and observe no coherence, however through the process of self amplification they eventually result in an highly coherent short pulses of intense light. This occurs as the initial x-rays travel faster than the electrons along their slalom path, overtaking them and interacting with the electron bunches along the way. Consequently, some of the electrons accelerate and others slow down, resulting in the electrons organising themselves into a series of thin microbunches, within which they are tightly bound within the desired wavelength of the produced x-rays. The microbunched electrons emit x-rays that are in phase with each other, resulting in exponential amplification of the x-ray beam. The coherent, high intensity, and ultrashort x-ray microbunches are then focused on the sample. In the context of imaging photochemical reactions, XFELs can also be employed within a typical pump-probe setup, as seen in Fig. 1.5. Here, the pump initialises a photochemical reaction and the x-ray pulses produced by the XFEL probes the changing electronic density over time. At the detector, the intensity of the scattered waves are recorded over a series of time delays as a function of the scattering angle $\theta$ and the azimuthal angle $\phi$. Often the signal is represented in terms of the momentum transfer vector $\mathbf{q}$, which is defined as the difference between the incident and scattered wave vectors $\mathbf{k}_i$ and $\mathbf{k}_f$, and is related to $\theta$ as $\mathbf{q} = 2|\mathbf{k}_i|\sin(\theta/2)$. The series of scattering patterns collected over time encode the electronic and structural information of the system. It is worth noting that while highly valuable, XFELs require extremely long particle accelerators involving sensitive specialised technology, and are very expensive to build and operate. One example is the European XFEL [60], which began operations in 2017. The XFEL spans a total length of 3.4 km and cost approximately €1.22 billion to construct alone. All of this means that access to XFEL facilities is scarce. However, plasma wakefield accelerators have recently shown promise in generating high energy and ultrabright x-ray radiation with pulse duration's on the order of attoseconds [61]. Scientists are now exploring the use of such technology in ultracompact XFEL's that can be incorporated into standard laboratory environments.
1.2 Ultrafast Imaging

Figure 1.5: Typical pump-probe setup utilised in XFEL based scattering experiments. The intensity of the scattered x-ray is recorded as a function of the momentum transfer vector $q$ and the angle $\phi$. The momentum transfer vector is given by the difference between the incident and scattered wave vectors, $k_i$ and $k_f$. Here, $q$ is related to the angle $\theta$ at which the x-ray is scattered.

In addition to scattering by x-rays, another similar imaging technique, ultrafast electron diffraction (UED), also allows for the probing of excited state dynamics on ultrafast timescales [62–65]. In comparison to the generation of light in XFELs, the electron beam that probes the target sample is generated by inducing photoemission from a photocathode with a laser [66]. The resulting emitted electrons are accelerated in a radiofrequency field, forming short pulses of microbunched electrons. One of the advantages of UED is that it is sensitive to both the electronic density and the nuclei. Recently, this has allowed for simultaneous identification of electron and nuclear dynamics in small gas phase systems [23]. Given that the charge of a nucleus is generally much greater than that of an electron, in principle, the intensity of the scattered signal is enhanced in comparison to x-rays. This additional sensitivity makes it possible to directly observe dynamics involving hydrogen atoms with electron scat-

4. Within the ultrafast electron diffraction community, diffraction and scattering are often used interchangeably when referring to gas phase experiments, even though diffraction implies the use of solid state samples.
Ultrafast Imaging

tering, which is typically not possible within x-ray scattering due to the hydrogen atoms small x-ray cross-section [67]. Moreover, electron beam lines are capable of achieving much higher energies, providing better spatial resolution and the ability to collect data over a larger range of momentum transfer [66]. However, the high energies also make the formation of plasma possible at the sample centre, this can interact with the scattered electrons and result in streaking of the signal [68]. While femtosecond temporal resolution is achievable, there is a fundamental limit placed on the pulse width as a result of Coulomb repulsion between electrons in the beam. Ultimately, UED is capable of providing additional information not seen in x-ray scattering, although the benefits have to considered together with the limitations. Therefore, electron and x-ray scattering should be seen as two complementary methods, which together, can provide a more complete view of the dynamics of a given system. One major benefit of UED is that they can more easily accommodated in a standard laboratory environment, making them significantly cheaper and more accessible in comparison to XFELs.

Generally speaking, excited state dynamics is an inherently complex and high dimensional problem. A complete model of a system would comprise the observation of the electronic density, state populations, nuclear coordinates, and their associated momenta. Typically, a given experiment is only sensitive to one or two aspects of the multidimensional dynamics. For example, TRPES is sensitive to the state populations, x-ray scattering to the electronic density, and electron scattering to both the electronic and nuclear density. As a result, building a complete understanding of the dynamics involved is a time consuming and iterative process, in which understanding is built over time by growing a continuing body of evidence comprised of many different data sets. To give a truly complete picture of the dynamics, data from different types of experiments must be collated and analysed together in a community wide effort. Even observing one component of the multidimensional dynamics is difficult. Taking for example, the task of determining the evolution of the nuclear coordinates using scattering techniques, excited state systems generally exhibit a large number of coupled degrees of freedom, making it difficult to identify the different contributions to the signal. Furthermore, the experimental observable is subject to the constraints of the available temporal and spatial resolution within a given setup, and therefore can produce a simplified or distorted view of the underlying dynamics. This further signifies the importance of performing different types of experiments and developing high quality computational methods that aid in interpreting the experimentally observed dynamics.
1.3 Computational Foundations

Theoretical models and computational methods play an important role in understanding the properties and behaviour of molecular systems, a statement that far extends excited state dynamics and resonates throughout all physical sciences. These provide a means of not only aiding in the interpretation of experimental observables, but for also directly advancing understanding of molecular systems through predicting their properties, behaviour, interactions, and dynamics. This allows a level of understanding that cannot be attained from experiment alone, and can also facilitate the design of new experiments that aim to observe new phenomena. Ultimately, the synergy of theoretical and experimental studies helps to develop a depth of insight that would otherwise not be possible.

1.3.1 Simulating Quantum Molecular Dynamics

The observed properties and evolution of photoexcited systems are governed by quantum mechanics, which challenges the deterministic notions of classical mechanics, where the certainty in the position and momentum of a particle is assumed with complete precision. In contrast, quantum particles exhibit wave-particle duality, and there is an inherent uncertainty in knowing both their position and momentum, known as the Heisenberg uncertainty principle. Due to the wave-like behaviour of quantum particles, they can undergo constructive and destructive interference, just like classical waves. This leads to many interesting phenomena, such as the interference pattern observed upon firing electrons through a double slit. The behaviour of quantum particles can be described by Schrödinger’s wave equation, in which the all information about the system is contained by the wavefunction. A wavefunction describes the quantum state of particles (such as electrons), and can be thought of as a probability amplitude that defines the probability of all possible outcomes that may be observed as a result of measurement on the system. This allows for the phenomenon of quantum superposition, in which the wavefunction describing a particle is spread over a linear combination of quantum states, leading to the famous Schrödinger’s cat thought experiment. This has profound implications for excited state dynamics, where the system may exist in a superposition of electronic states at once. Moreover, individual excited electronic states are not solely defined by their own character, and are often highly coupled. In quantum dynamics, the nuclei themselves are not localised at a singular point in space; therefore both the nuclei and the electrons are represented by a wavefunction. As a result, the observed properties and dynamics
of a photoexcited system may actually be the result of interference between different quantum states. All of this makes the simulation of excited dynamics an inherently complex and computationally expensive task. At the core of the computational expense is the exponential scaling of quantum mechanics with the number of degrees of freedom.

Simulating quantum dynamics involves solving Schrödinger’s set of coupled differential equations. The first of which, the time-independent Schrödinger equation (TISE), describes the energy of a system in the basis of time-independent states. Often one solves this for the energies of a series of electronic states by separating the treatment of the nuclei and electrons, allowing an easier computation in which the electronic energies are calculated for a given fixed set of nuclear coordinates. However, this is an approximation, known as the Born-Oppenheimer approximation (BOA) [69], and the nuclear and electronic degrees of freedom can in fact exhibit strong coupling in regions where the energy separation of states is small. This coupling leads to population transfer between the electronic states, and therefore has to be calculated in order to simulate excited state decay mechanisms accurately. Once the electronic states are calculated, the whole system can be propagated forward in time according to the time-dependent Schrödinger equation (TDSE). Within this, the evolution of the nuclei is described by a nuclear wavepacket, which defines the idea that the nuclear probability distribution evolves over a superposition of time-independent electronic states.

The manner in which the dynamics are simulated depends on the specific method of choice. A fully quantum treatment of both electrons and nuclei is generally only possible for small systems. These methods usually require one to precalculate the electron wavefunctions, and subsequently propagate the nuclear wavepacket along a grid of potential energies. For systems larger than a few atoms, one usually employs semi-classical methods, involving a quantum treatment of the electrons and a classical treatment of the nuclei. These approaches tend to approximate the behaviour of the true nuclear wavepacket by employing an ensemble of trajectories, each of which follows its own set of position and momentum coordinates. While some of these methods involve the trajectories evolving in a coupled fashion, some take the more approximate approach of assuming they evolve independently. In these semi-classical approaches one relies on the collective behaviour of the trajectory ensemble to model the wavepacket. While semi-classical methods can be employed using a grid based representation of the potential energies, they tend to utilise on-the-fly approaches
to calculating the electronic wavefunctions and associated energies. This involves solving the TISE for a fixed set of nuclear coordinates at one time step, and then propagating the system forward to the next time step, where the TISE is solved again for the updated nuclear coordinates. This has the advantage that one does not have to sample the available space that the trajectories may span before the simulation. As the electronic states are usually calculated assuming the BOA, one must perform additional calculations of the couplings between states in order to allow population transfer. In cases of systems of considerable size, a further simplification is often made, and the direct calculation of the couplings is omitted. Instead, the probability of population transfer occurring is approximated by considering the overlap of two electron states.

![Figure 1.6](Image)

**Figure 1.6:** Schematic representation of the difference between adiabatic and diabatic electronic states. The dashed lines represent the avoided crossing formed in the basis of adiabatic states.

Generally, fully quantum and semi-classical methods often differ by the representation of electronic states that they employ. Most quantum chemistry codes that provide a means of calculating the electronic wavefunctions operate in the basis of adiabatic states. Such states result from separation of the nuclear and electronic degrees of freedom when solving the TISE, and are therefore defined with respect to the given specified nuclear coordinates. In the adiabatic representation, states of the same symmetry and spin cannot cross, leading to avoided crossings and conical intersections. Avoided crossings refer to the point where the electronic states come into close contact within one degree of freedom. However, conical intersections correspond to the point where two states intersect in more than one dimension, forming a conical intersection.
funnel between them. Both avoided crossings and conical intersections correspond to regions of high coupling, where large amounts of population transfer can occur. It is also possible to define a different representation of states from a unitary transform of the adiabatic states, known as the diabatic representation. Here states may cross, and the transformation is usually defined such that the states are independent of the nuclear geometry. These states more strongly track the character of the dominant molecular orbital that they correspond to, and are often referred to as spectroscopic states as they define the electronic states that are experimentally observed in spectroscopy. In fully quantum methods of simulating the dynamics, diabatic states are often preferred, however diabatization is a non-trivial task. The difference between these two representations is shown in Fig. 1.6, where the coloured lines correspond to the diabatic states that intersect, and the dashed lines indicate where the avoided crossing would lie in the adiabatic representation.

1.3.2 Interpretation of Simulations and Experiment

In this thesis, semi-classical trajectory based simulations are employed to aid both theoretical and experimental understanding. From these simulations one can calculate observables such as scattering patterns and photoelectron spectra, which can be compared to experimental observations. Such comparisons are important as experiments are typically only sensitive to limited aspects of the dynamics occurring, contain limited information content, and have inherent resolution constraints. This is especially important in the case of ultrafast imaging experiments that utilise particle accelerators such as XFELs. Generally, access to these facilities is competitive and scarce, with scientists often having only a few days to complete their experimental investigations. In these circumstances, theoretically calculated observables serve as an important first order check on the experimental observations. Moreover, this allows one to make informed decisions on whether the setup of the experiment should be modified or adjusted to meet the aims of the experiment. In recent years, the optimisation of the experimental setup based on theoretical models has become an active area of research [70–73]. The methods developed in this thesis aim to aid such optimisation problems by affording additional insight into complex photochemical processes.
Often, we wish to identify the key reaction pathways that form the dominant contributions to the dynamics and experimentally measured observables. Due to the large number of degrees of freedom, multiple coupled electronic states, and strong competition between different decay mechanics, trajectory data is complex and often difficult to categorise. Usually, trajectory data is sorted by eye in order to determine the dominant pathways involved in the dynamics. However, this is normally only based on the one or two aspects of the multidimensional dynamics. In recent years, the field of computational chemistry has seen an increase in the interest of machine learning based algorithms. Important applications include the calculation of potential energy surfaces, hybrid machine learning based quantum dynamics simulations, and predictions of molecular properties [74]. Moreover, unsupervised clustering algorithms that aim to identify key patterns and predict future observations, have seen wide success in numerous fields. Applications to time-dependent data include, the analysis of GPS data [75], tracking trends in animal migration [76], categorising medical time series data [77], and financial forecasting [78]. One of the central themes of this thesis involves the use and development of unsupervised clustering algorithms capable of automatically categorising trajectory data used to model photochemical reaction dynamics, providing a more complete and holistic means of identifying the key trends observed in data.

1.3.3 Inversion Schemes

The majority of experiments are analysed by comparing theoretically calculated observables with the experimental observation in a side-by-side fashion. This is demonstrated by steps with red arrows in Fig. 1.7. Upon computation of the wavefunction $|\Psi\rangle$ that defines the system, the theoretical observable $X_{th}$ can be calculated by solving a set of physically motivated equations that correspond to the problem at hand. Considering the specific case of molecular structure determination in scattering experiments, if there exists sufficient agreement between the theoretically calculated ($X_{th}$) and experimentally observed ($X_{obs}$) scattering signals, one can indirectly infer the structures that the signal likely corresponds to from the simulated data. Note, this is generally an iterative process, in which one tries to assess the different contributions to the signal and determine if parts of there are any experimental observations that do not match the prediction. If parts of the theoretical and experimental observations do not match, this has to be further investigated. However, if agreement is not achieved, or one can not explain the lack of agreement, there is generally little that one can do.
In the standard approach to analysing experiments (red arrows), the theoretical observable $X_{th}$ is calculated from the wavefunction $|\Psi\rangle$. This is compared with the experimental observable $X_{obs}$ and the molecular structures are inferred. The inverse problem (blue arrow), directly identifies the structures that are encoded in the experimental signal, typically using optimisation or machine learning based methods.

**Figure 1.7:** In the standard approach to analysing experiments (red arrows), the theoretical observable $X_{th}$ is calculated from the wavefunction $|\Psi\rangle$. This is compared with the experimental observable $X_{obs}$ and the molecular structures are inferred. The inverse problem (blue arrow), directly identifies the structures that are encoded in the experimental signal, typically using optimisation or machine learning based methods.

In contrast to this, so called inversion methods aim to directly invert the observed signal to uncover the information encoded in the signal, demonstrated by the blue arrows in Fig. 1.7. Generally, the inverse approach aims to identify a set of model parameters that reconstructions the experimental observations. Inverse problems find applications in a wide range of scientific fields and are particularly important in image processing [79], for example in the deconvolution of blurred images or the reconstruction of a 3D object from a photograph [80, 81]. The inverse problem is of significant importance in the case of molecular structure determination, where one wishes to identify the molecular structures corresponding to the observed scattering signals. This forms a central theme of this thesis, in which we explore the inversion of time-resolved x-ray and electron scattering data in uncovering observed photochemical reaction dynamics. Typically, the inversion of scattering signals in structural determination involves the use of both machine learning and numerical optimisation based methods. One example of machine learning approaches includes the use of neural networks (NN), which aim to learn the relationship between the signal and the corresponding structures from a set of training data, typically a pool of sample structures or trajectories [82]. Once the network is trained, it can be applied to invert the experimental observable, uncovering the corresponding molecular structures. In optimisation approaches, one attempts to minimise the error between the set of models calculated from a pool of structures and the experiment, aiming to identify the models that are commensurate with the experiment [83]. The benefit of inversion methods lies in the
fact that they are not based on a side-by-side comparison with simulated data, and can provide a more quantitative analysis of the experiment. Moreover, they may also allow one to identify the structures encoded in the signal in light of the resolution constraints of the experiment, as we shall later explore. However, the inverse problem is difficult, largely due to the fact that it is under determined and there exists a lack of unique of solutions.

1.4 Overview of Thesis

In Chapter 2, we introduce the main theoretical concepts that form the basis of this thesis. This includes a detailed summary of how both the time-independent and time-dependent Schrödinger equations are solved in order to simulate quantum mechanical reactions dynamics. Moreover, the quantum framework of both x-ray and electron scattering is covered, ending with a discussion of a series of approximations that are often employed in analysing scattering experiments. Finally, we introduce the concepts of data clustering, exploring how such algorithms can be applied to categorise data both in space and time. In Chapter 3, we introduce a model trajectory data set describing the photochemistry of 1,3-cyclohexadiene, which reoccurs throughout the thesis. In this chapter, we explore one possible framework in which temporal data can be categorised, employing existing algorithms to the model trajectories. In Chapter 4, we extend on the work of Chapter 3. Here a different approach to categorising temporal data is taken, and we discuss the development of two algorithms that work together to categorise data both spatially and temporally. Moreover, this chapter addresses how such algorithms can be used to provide insight into complicated data, aid in the analysis of experiment, and facilitate more expensive and accurate methods of observable calculations. The later two chapters of the thesis then shift the focus to the direct analysis of time-resolved experiments. In Chapter 5, we discuss a general framework for the inversion of time-resolved observables. This involves an optimisation based approach, in which the experimental observable is reconstructed by refinement of a model in light of the constraints of experimental resolution. The framework is then applied in the specific context of scattering experiments. We explore its application to the dynamics of 1,3-cyclohexadiene as measured by x-ray scattering, and the photodissociation dynamics of CS$_2$, as studied by ultrafast electron diffraction. In Chapter 6, we discuss the CS$_2$ UED experiment conducted as part of a larger collaboration in more detail. Limitations of the nuclear pair distribution
functions that are commonly employed in UED experiments are discussed in the context of this specific experiment. The methods introduced in Chapter 5 are then used to analyse the experiment, and suggestions are made for future UED studies of ultrafast photodissociation dynamics in light of observed experimental limitations.
Chapter 2

Theory

This chapter provides a concise summary of the theoretical foundations of time-dependent quantum mechanics, light-matter interaction, and the unsupervised clustering algorithms on which the subsequent chapters are based. The motion induced by photoexcitation is generally complex, with the nuclei evolving along a series of electronic potential energy surfaces (PESs) that may possess numerous near-degeneracies and strong couplings between them. The photoexcited dynamics that ensue are non-equilibrium processes, wherein the wavepacket explores a large portion of the available phase space (defined by position and momentum coordinates). This, combined with the exponential scaling inherent to quantum mechanics, makes simulating the full quantum nature of the dynamics difficult. Most often a series of approximations are made, allowing one to numerically solve the time-dependent Schrödinger equation (TDSE) that describes the evolution of the system. One approximation that is central to this thesis is the Born-Oppenheimer approximation (BOA) [69]. In this chapter we discuss the origins of the BOA, its breakdown, and the coupling terms that can be included in simulations to account for the nonadiabatic effects that are not captured by the BOA, such as population transfer. In addition, we introduce the theoretical foundations of two methods that are used to simulate the excited state dynamics of two polyatomic molecules that appear later in the thesis.

A large part of this thesis revolves around how computational methods can aid the interpretation of experimentally measured observables. Hence, this chapter also outlines the underlying quantum theory involved in the experimental techniques that are used to image reactions in space and time. In particular, we focus on ultrafast x-ray scattering (UXS) and ultrafast electron diffraction (UED). While there now exist ways to calculate scattering observables in the a full quantum framework [84–87], these calculations are expensive and often impractical. We therefore introduce a series of approximations that are commonly used in the analysis of scattering experiments.
A summary of existing clustering and data mining methods used within unsupervised machine learning is also provided at the end of this chapter. In subsequent chapters the development of these methods for application to complex simulated data will be explored. These provide a tool for classifying and quantifying complex highly dimensional data, allowing for the construction of simplified models of simulated photochemical dynamics as well as the identification of the most important features in the dynamics.

2.1 The Time-Dependent Schrödinger Equation

The molecular wavefunction $|\Psi(t)\rangle$ describes the time-dependent evolution of a system in terms of its nuclear and electronic coordinates, momentum and spin. Thus, it completely defines what is knowable about the system at any point in time, and its evolution is governed by the time-dependent Schrödinger equation (TDSE),

$$i\hbar \frac{\partial |\Psi(t)\rangle}{\partial t} = \hat{H} |\Psi(t)\rangle,$$

(2.1)

where $i$ is the imaginary unit, $\hbar$ is the reduced Planck’s constant, and $\hat{H}$ is the Hamiltonian operator. In the case of a time-independent Hamiltonian, the TDSE can be solved by separation of variables. Insertion of the ansatz $|\Psi(t)\rangle = |\Xi(t)\rangle |\phi\rangle$ into the TDSE yields,

$$i\hbar \langle \Xi(t)|\ddot{\Xi}(t)\rangle = \langle \phi|\hat{H}|\phi\rangle = E,$$

(2.2)

where we use dot notation to indicate the time derivative (i.e. $\frac{\partial |\Xi\rangle}{\partial t} = |\ddot{\Xi}\rangle$). Now the left hand side of Eqn. 2.2 is a function of time, while the right hand side is a function of position. Since these are independent variables, the left hand side and right hand side must be equal to a constant, the energy of the system $E$. Leading to a set of simultaneous equations that can be solved independently,

$$\hat{H}|\phi\rangle = E|\phi\rangle,$$

(2.3)

$$i\hbar |\dot{\Xi}(t)\rangle = E|\Xi(t)\rangle.$$

(2.4)

---

1. Here we refer to the Hamiltonian in the most general sense such that it may define any quantum system or interaction.
2.1 The Time-Dependent Schrödinger Equation

Eqn. 2.3 is the time-independent Schrödinger equation (TISE), an eigenvalue equation where the total energy $E$ of the system is an eigenvalue of the eigenvector $|\phi\rangle$.

Therefore, the TISE defines the allowed energies of a system at a given point in time. Solutions to Eqn. 2.4 can be solved by integration to,

$$|\Xi(t)\rangle = \Xi_0 e^{-iEt/\hbar},$$

therefore the total solutions of the TDSE are of the form,

$$|\Psi(t)\rangle = |\phi\rangle e^{-iEt/\hbar},$$

where the constant $\Xi_0$ has been absorbed by $|\phi\rangle$ under the normalisation condition.

Here the solutions correspond to stationary states that oscillate in the complex plane according to a time-dependent phase factor. This oscillating wave function has a stationary probability distribution, i.e. $p(x') = |\langle x' | \phi \rangle|^2$, which has no explicate time-dependence. However, if the system is prepared in a superposition of two or more stationary states, interference between them introduces time-dependence. Mathematically, due to the linearity of both Eqn. 2.3 and Eqn. 2.4, the solutions $|\Psi(t)\rangle$ can be expanded in the basis of a complete set of orthonormal eigenstates of $\hat{H}$ such that,

$$|\Psi(t)\rangle = \sum_i c_i(t) e^{-i\omega_i t} |\phi_i\rangle,$$

where $c_i(t)$ are time-dependent expansion coefficients, $\omega_i = E_i/\hbar$ is the angular frequency on which the phase factors depend, and $|\phi_i\rangle$ are eigenstates of the Hamiltonian, i.e. they are solutions of the TISE in Eqn. 2.3. The solution to the TDSE in Eqn. 2.7 is known as a wavepacket - a non-stationary superposition of time-independent eigenstates, $|\phi_i\rangle$. Now the probability distribution is given by [7],

$$p(x', t) = \sum_i \left| c_i(t) \right|^2 \left| \langle x' | \phi_i \rangle \right|^2 + 2 \sum_{i>j} \text{Re} \left[ c_i^*(t) c_j(t) \langle \phi_i | x' \rangle \langle x' | \phi_j \rangle e^{i\omega_{ij} t} \right],$$

with $\omega_{ij} = \omega_i - \omega_j$. Here, again, we see that the probability distribution is no longer stationary, and that time-dependence of the wavepacket arises from the interference of time-independent eigenstates and their complex phase factors. The fact that the dynamics of a quantum system is inherently an interference effect is central to quantum mechanics and leads to a number of phenomena, for example, the oscillation of population between nearly degenerate states i.e. quantum beats. We conclude this section
by noting that while the equations introduced here can be solved analytically for small quantum systems with one or two degrees of freedom, additional approximations must still be introduced in order to numerically solve these equations for larger molecular systems, as we shall see in the following sections of this chapter.

2.1.1 The Molecular Hamiltonian

Throughout this thesis, we are concerned with solving the TDSE in the basis of eigenstates that correspond to quantised electronic and vibrational states of a polyatomic molecules. Using a Born-Huang expansion [88], the time-dependent state of the entire molecular system can be expanded in the basis of a complete set of electronic eigenstates $|\psi_\alpha\rangle$ as,

$$|\Psi(t)\rangle = \sum_{\alpha=1}^{\infty} |\chi_\alpha(t)\rangle |\psi_\alpha\rangle,$$

(2.9)

where $|\chi_\alpha(t)\rangle$ are time-dependent nuclear wavepackets of each corresponding electronic state $|\psi_\alpha\rangle$. The wavepackets $|\chi_\alpha(t)\rangle$ are akin to the time-dependent coefficients and phase factors in Eqn. 2.7. When the sum over electronic states $\alpha$ is infinite in Eqn. 2.9, the expansion is formally exact. In practice, the sum is truncated to the minimum number of electronic states that provides an accurate description of the wavepacket and is computationally feasible. The electronic states $|\psi_\alpha\rangle$ are given as solutions to the TISE, and are eigenstates of the time-independant molecular Hamiltonian, which in atomic units, is given by,

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

(2.10)

where $M_A$ is the mass of nucleus $A$, $Z_A$ and $Z_B$ are the atomic numbers of nucleus $A$ and $B$ respectively, $r_{AB}$ refers to the distance between nucleus $A$ and $B$, $r_{iA}$ the distance between electron $i$ and nucleus $A$, $r_{ij}$ the distance between the two electrons $i$ and $j$, and the Laplacian operators $\nabla_i^2$ and $\nabla_A^2$ involve differentiation with respect to the coordinates of the $i^{th}$ electron and $A^{th}$ nucleus. The first term in Eqn. 2.10 again corresponds to the kinetic energy of the nuclei ($\hat{T}_n$), the second to the kinetic energy of the electrons ($\hat{T}_e$), and the following three terms to the Coloumbic potentials of the nuclear-electron ($\hat{V}_{ne}$), electron-electron ($\hat{V}_{ee}$), and nuclear-nuclear ($\hat{V}_{nn}$) interactions respectively. In a theoretically exact treatment, one would treat the nuclear and electronic terms in Eqn. 2.10 together and solve the TISE for solutions of the full Hamiltonian. However, such a simultaneous treatment of the nuclear and
electronic coordinates is difficult and computationally intractable. Instead we separate
the nuclear kinetic energy term from the Hamiltonian and define the clamped-nuclei
Hamiltonian $\hat{H}_e$ as,
\[
\hat{H} = \hat{T}_N + \hat{H}_e = -\sum_{A=1}^{M} \frac{1}{2M_A} \nabla_A^2 + \hat{H}_e,
\]
where $\hat{H}_e$ contains the remaining terms in Eqn. 2.10. This separation of electronic
and nuclear degrees of freedom is known as the Born-Oppenhiemer approximation
(BOA), and under certain conditions is valid given that the nuclei are much heavier
than electrons and thus move slower. Within the BOA we can solve the electronic
TISE for the eigenstates of $\hat{H}_e$ that only depend on $R$ parametrically,
\[
\hat{H}_e |\psi_{el}(r; R)\rangle = E_{el} |\psi_{el}(r; R)\rangle,
\]
where the semicolon represents the parametric dependence of the total electronic
wavefunction $|\psi_{el}(r; R)\rangle$ on the nuclear coordinates. For brevity, this notation will
be dropped in all subsequent equations. The electronic wavefunction $|\psi_{el}\rangle$ can be
expanded in the basis of a complete set of electronic states $|\psi_{\alpha}\rangle$, each with their
respective potential energy $E_{\alpha}$. Together, the states $|\psi_{\alpha}\rangle$ and the time-dependent
nuclear wavepackets $|\chi_{\alpha}(t)\rangle$ define the total molecular wavefunction $|\Psi(t)\rangle$ as given
through the Born-Huang expansion in Eqn. 2.9. The energies $E_{\alpha}$ define the idea of
a potential energy surface (PES), the gradient of which corresponds to the force that
the nuclei feel from the electronic potential, and thus governs how the nuclei evolve
on them. Typically, the methods used for propagating $|\Psi(t)\rangle$ in this thesis, involve
the calculation of the electronic states $|\psi_{\alpha}\rangle$ and energies $E_{\alpha}$ within the BOA, the
nuclear wavepackets $|\chi_{\alpha}(t)\rangle$ are then propagated nonadiabatically according to the
relevant equations of motion, and the electronic wavefunction is then recalculated at
the next set of nuclear coordinates. In order to propagate the nuclei nonadiabatically,
additional couplings must be calculated as the BOA breaks down in the vicinity of
conical intersections and avoided crossings. These must be included in order to
allow for nonadiabatic population transfer between states, as we shall discuss in the
following section.

For polyatomic molecules, the states and energies are numerically calculated through
ab-initio electronic structure codes, given a fixed set of nuclear coordinates within the
BOA. Such programs contain a wide range of methods, the choice of which depends
on the problem at hand. An in-depth review of these different methods is outside the
scope of this thesis, however we shall proceed with a brief discussion on methods
commonly employed in excited state dynamics\(^2\). Generally, for problems pertaining to excited state dynamics, one requires a multiconfigurational description of the electronic wavefunction, such as those based on multiconfigurational self-consistent field theory (MCSCF) [90–93]. In these methods, the electronic wavefunction is described by a linear combination of Slater determinants - each of which represents a specific electron configuration. This accounts for some of the static electron correlation associated with the linear combination of atomic orbitals from which the molecular orbitals are constructed. Including all possible electronic configurations corresponds to full configuration interaction (CI) [94], which is computationally intractable. The complete active space self-consistent field (CASSCF) method gets around this by only considering the possible electronic configurations that are the result of excitation from a subset of user chosen orbitals, known as an active space. Crafting an active space is a difficult process, requiring some chemical intuition for the system at hand, and generally involves some amount of trial and error. The choice in active space is critical to the description of the electronic wavefunction, and therefore can have a profound effect on the dynamics. Multireference methods go beyond typical MCSCF methods by constructing the total electronic wavefunction from multiple multiconfigurational references. These methods are much more computational expensive, although they account for some of the additional dynamic electron correlation. Such methods include multireference configuration interaction (MRCI) [95], complete active space perturbation theory to the second order (CASPT2), and its extended multistate variant (XMS-CASPT2) [96]. For large molecular systems, these methods are often not viable, and thus one often has to resort to more approximate methods such as time-dependent density functional theory (TDDFT) [97] or algebraic diagrammatic construction theory (ADC) [98], however these cannot describe the excited electronic states and the dynamics that evolve on them to the same level of accuracy.

### 2.1.2 Nonadiabatic Coupling

So far we have introduced the general framework within which the molecular wavefunction \(|\Psi(t)\rangle\) is propagated in time, assuming an expansion of the wavefunction into a complete basis of electronic states \(|\psi_\alpha\rangle\), and separation of the nuclear and electronic degrees of freedom. Such Born-Oppenheimer states are much simpler to calculate, however they are unable to capture the coupling between the electronic and nuclear degrees of freedom when the energy separation of states is small.

---

2. An extensive introduction into foundational ab-initio quantum chemistry can be found in Ref. 89
2.1 The Time-Dependent Schrödinger Equation

These interstate couplings give rise to the nonadiabatic transfer of population between electronic states, and thus must be accounted for when considering excited state dynamics. We shall now introduce the mathematical origin of these couplings and how their exclusion defines the BOA. Insertion of expansion in Eqn. 2.9 and the Hamiltonian in Eqn. 2.11 \( \hat{H}_e + \hat{T}_n \) into the TDSE yields,

\[
\int h \frac{\partial}{\partial t} |\chi_\alpha(t)\rangle |\psi_\alpha\rangle = [\hat{H}_e + \hat{T}_n] |\chi_\alpha(t)\rangle |\psi_\alpha\rangle. \tag{2.13}
\]

Now multiplying the left hand side through by \( \langle \psi_\beta | \) we arrive at,

\[
\int h \sum_{\alpha=1}^\infty \frac{\partial}{\partial t} |\chi_\alpha(t)\rangle \langle \psi_\beta | \psi_\alpha \rangle = \int h \frac{\partial}{\partial t} |\chi_\beta(t)\rangle, \tag{2.14}
\]

where we have used the orthonormality of the electronic eigenstates, \( \langle \psi_\beta | \psi_\alpha \rangle = \delta_{\alpha\beta} \).

Multiplying the right hand side in the same fashion yields,

\[
\sum_{\alpha=1}^\infty \langle \psi_\beta | \hat{T}_n |\chi_\alpha(t)\rangle |\psi_\alpha\rangle + \sum_{\alpha=1}^\infty \langle \psi_\beta | \hat{H}_e |\chi_\alpha(t)\rangle |\psi_\alpha\rangle. \tag{2.15}
\]

The Hamiltonian \( \hat{H}_e \) involves derivatives with respect to only the electronic coordinates, and thus the second term in Eqn. 2.15 reduces to \( E_\beta |\chi_\beta(t)\rangle \) given that \( \langle \psi_\beta | \hat{H}_e |\psi_\alpha \rangle = \delta_{\alpha\beta} E_\beta \). In evaluating the first term, we note that the nuclear kinetic energy operator \( \hat{T}_n \), defined in Eqn. 2.11, contains the Laplacian \( \nabla^2 A \) which acts upon both the nuclear wavepackets and electronic states. Applying the product rule for second derivatives and using the definition of \( \hat{T}_n \), the net expression for Eqn. 2.15 becomes,

\[
[\hat{T}_n + E_\beta] |\chi_\beta(t)\rangle = \sum_{\alpha=1}^\infty \sum_{A=1}^{N_A} M_{A^{-1}}^{-1} \left[ \langle \psi_\beta | \nabla_A |\psi_\alpha \rangle \cdot \nabla_A |\chi_\alpha(t)\rangle \right] + \sum_{\alpha=1}^\infty \left[ \langle \psi_\beta | \hat{T}_n |\psi_\alpha \rangle \right] |\chi_\alpha(t)\rangle. \tag{2.16}
\]
Equating the left-hand and right-hand sides given by Eqn’s. 2.14 and 2.16, respectively, we arrive at the full expression for the TDSE in the Born-Huang basis as a set of coupled differential equations for each of the nuclear wave packets $|\chi_{\beta}(t)\rangle$,

$$i\hbar \frac{\partial}{\partial t} |\chi_{\beta}(t)\rangle = \left[ \hat{T}_n + E_{\beta} \right] |\chi_{\beta}(t)\rangle - \sum_{\alpha=1}^{\infty} \sum_{A=1}^{N_N} \frac{1}{2M_A} \left[ 2A_{\beta\alpha A} \cdot \nabla_A + B_{\beta\alpha A} \right] |\chi_{\alpha}(t)\rangle,$$

where we have defined the matrix elements as follows,

$$A_{\beta\alpha A} = \langle \psi_{\beta} \nabla_A | \psi_{\alpha} \rangle,$$

$$B_{\beta\alpha A} = \langle \psi_{\beta} \nabla_A^2 | \psi_{\alpha} \rangle.$$

The first term on the right-hand side of Eqn. 2.17 contains the eigenvalue $E_{\beta}$, which is the energy of the respective electronic state $|\psi_{\beta}\rangle$. In this equation, the physical relationship between the energies of the states $E_{\alpha}$ and the nuclear wavepackets $|\chi_{\alpha}(t)\rangle$ is apparent, with the energies defining a series of PES’s on which the nuclei evolve. The second term contains the matrix elements $A_{\beta\alpha A}$ and $B_{\beta\alpha A}$, named the derivative and kinetic non-adiabatic coupling matrix elements (NACME) respectively. These couple the nuclear wavepackets of one electronic state $|\chi_{\beta}(t)\rangle$ to the wavepackets $|\chi_{\alpha}(t)\rangle$ of all other states. Generally the derivative coupling is significantly greater than the kinetic coupling, and therefore the later is often ignored. The diagonal elements of the derivative coupling are zero, and the off-diagonal terms are important in describing the interstate coupling. It is also worth mentioning that the NACME’s are antisymmetric, i.e. $A_{\alpha\beta} = -A_{\beta\alpha}$. Through application of the Hellmann-Feynman theorem [99] the derivative NACME can be expressed as,

$$A_{\beta\alpha A} = \frac{\langle \psi_{\beta} | \nabla_A \hat{H}_e | \psi_{\alpha} \rangle}{E_{\alpha} - E_{\beta}}.$$

This important result emphasises relationship between the coupling and the separation of the electronic states. The smaller the difference in energy between states, the larger magnitude of the coupling. With large interstate separation, $A_{\beta\alpha A}$ becomes negligible, and thus is often omitted under these circumstances. Through exclusion of all off-diagonal elements within the derivative $A_{\beta\alpha A}$ and kinetic $B_{\beta\alpha A}$ couplings, we invoke the adiabatic approximation. Note the adiabatic approximation is distinct from
the adiabatic representation, for which it is often confused. The latter corresponds to a convenient basis in which states of the same spin multiplicity and symmetry do not cross. One can go a step further and also exclude the diagonal terms of the kinetic NACME $B_{\beta \alpha A}$, the TDSE in Eqn. 2.17 then simplifies to,

$$i\hbar \frac{\partial}{\partial t} \left| \chi_{\beta}(t) \right\rangle \approx \left( \hat{T}_n + E_{\beta} \right) \left| \chi_{\beta}(t) \right\rangle.$$  \hfill (2.21)

This is the TDSE in the framework of the BOA, in which the electronic states and their energies are completely decoupled from the nuclear motion. [69] This approximation is valid as long as the interstate separation is large and the couplings remain small. However, when concerned with the evolution of an excited state wavepacket as in photochemical reaction dynamics, the BOA soon breaks down due to often dense manifolds of excited states that are highly coupled. Population transfer between states is inherently a non-adiabatic effect that occurs where interstate separation is small and the coupling large. Without the NACME's, such effects cannot be modelled. Modern electronic structure programs are capable of providing numerical and sometimes analytical calculations of the NACME's. However, as previously touched on, the electronic eigenstates normally calculated are the adiabatic states, for which there is a singularity in the NACME at degeneracies as $E_{\alpha} - E_{\beta} \to 0$, as can be seen in Eqn. 2.20. How this is resolved differs between different quantum dynamics methods which propagate the nuclear wavepacket in time, as we shall see in subsequent sections of this chapter.

### 2.1.3 Spin-Orbit Coupling

The nonadiabatic couplings introduced in the previous section define the coupling between electronic states of the same spin multiplicity, and thus describe the non-radiative decay of an excited wavepacket through internal conversion (IC). However, the NACME’s do not capture the process of intersystem crossing (ISC), in which population is transferred between states of different spin multiplicity. For example, between singlet and triplet states. Typically, such transitions are forbidden by spin selection rules, although under certain conditions, spin forbidden transitions occur through spin-orbit coupling (SOC) [100]. This phenomenon is rooted in special relativity, and arises from the coupling of an electron’s spin angular momentum and the angular momentum of the orbital that it occupies [101, 102]. The electron spin is inherent to its intrinsic magnetic moment and the orbital angular momentum is related
2.1 The Time-Dependent Schrödinger Equation

to the motion of the electron in its orbital. The coupling is a result of the presence of an external electromagnetic field, such as the effective nuclear charge. High nuclear charges subject the electron to relativistic effects, for example length contraction, which cause the coupling of spin. This can be understood by the fact that in high nuclear charge cases, the electrons move at a sizeable fraction of the speed of light, where the mass of the electron is not constant. The relativistic mass $m_{\text{rel}}$ of a particle moving at approximately 10% the speed of light is defined as,

$$m_{\text{rel}} = \frac{m_{\text{rest}}}{(1 - (v/c))^2},$$

(2.22)

with $m_{\text{rest}}$ the rest mass of the electron, $v$ the particle’s speed relative to the observer, and $c$ the speed of light. As the Bohr radius $a_0$ depends on the electrons mass, any increase in mass will decrease the Bohr radius,

$$a_0 = 4\pi \varepsilon_0 \frac{\hbar^2}{m},$$

(2.23)

where $\varepsilon_0$ is the permittivity of free space. The ratio of the relativistic Bohr radius $a_{\text{rel}}$ to the stationary Bohr radius $a_0$ can be defined as,

$$\frac{a_{\text{rel}}}{a_0} = \sqrt{1 - \left(\frac{Z}{nc}\right)^2},$$

(2.24)

with $Z$ the nuclear charge, and $n$ the principle quantum number. Here Eqn. 2.24 emphasises the fact the contraction of the Bohr radius increases with nuclear charge $Z$. This in turn means that degree of SOC depends on the strength of the electromagnetic field i.e. the nuclear charge. Typically, SOC only becomes significant in systems containing heavier atoms in which the electrons move at a considerable fraction of the speed of the light. Therefore, in many systems it can be ignored. To fully account for relativistic effects, one must turn to the fully relativistic Dirac or Breit equations [103, 104]. However, the Dirac equation is not generalised for many body systems and the Breit equation is usually not computationally feasible. Instead, the inclusion of relativistic effects in the TISE is typically achieved by augmentation of the Hamiltonian with additional terms. For example to account for SOC, we can define the Dirac-Coulomb Breit Hamiltonian [105],

$$\hat{H}_{\text{DCB}}^{\text{SOC}} = \frac{1}{2c^2} \left[ \sum_i \sum_A Z_A \left( \hat{r}_{iA} \times \hat{p}_i \right) \hat{S}_i - \sum_{j \neq i} \frac{1}{r_{ij}^3} \left( \hat{r}_{ij} \times \hat{p}_i \right) \left( \hat{S}_i + 2\hat{S}_j \right) \right],$$

(2.25)
where the first term is a sum over all nuclei $A$ and electrons $i$, and contains the position operator $\hat{r}_{iA}$ of electron $i$ with respect to nucleus $A$, the electron’s momentum operator $\hat{p}_i$, and the electron’s spin operator $\hat{s}_i$. This term accounts for the interaction between the electron’s orbital angular momentum and the nuclear charge. The second term in Eqn. 2.25 includes a sum over all pairs of electrons, with $\hat{r}_{ij}$ defining the position operator for electron $i$ with respect to electron $j$, and $\hat{S}_i$ and $\hat{S}_j$ the spin operators for each individual electron. This term accounts for the interaction between one electron’s orbital angular momentum and the magnetic momentum of another electron. Both of these terms can be included in the definition of the Hamiltonian given in Eqn. 2.11 as,

$$\hat{H} = \hat{H}_e + \hat{T}_n + \hat{H}_{SOC}^{DCB}. \quad (2.26)$$

The addition of these corrections allow the Hamiltonian to describe SOC in polyatomic molecules, and can often be calculated in the basis of the BOA using the aforementioned ab-initio electronic structure theory programs. This allows us to solve the TDSE while accounting for both internal conversion and intersystem crossing. However, the calculation of the SOC’s using Born-Oppenheimer states introduces some additional problems with the representation of the states in some of the methods used to numerically solve the TDSE, as we shall see in the next section.

### 2.2 Molecular Quantum Dynamics Methods

Having introduced the general framework in which the set of Schrödinger equations are solved and the couplings required to account for both IC and ISC, we now turn to a discussion of the different methods that provide numerical solutions to the TDSE in the case of polyatomic molecules. Generally, a fully quantum treatment of both the nuclei and electrons is limited to smaller systems and is typically not applicable to most systems of interest. For example, fully quantum numerical grid propagators [106, 107] and the multiconfigurational time-dependent Hartree (MCTDH) [108, 109] method require potential energy surfaces to be pre-calculated. Thus, they are only feasible for systems with a small number of degrees of freedom, or those for which a reduced dimensionality model can be built. This is due to the exponential scaling with the number of degree of freedom that results from the non-local nature of quantum mechanics. For molecules comprised of more than a few atoms, approximate direct dynamics methods deal with the non-locality of quantum mechanics through a series of trajectories for which the PES is calculated on-the-fly, and thus sample only the
2.2 Molecular Quantum Dynamics Methods

most crucial parts of phase space. These methods tend to be semi-classical in nature, employing a quantum treatment of the electrons and a classical treatment of the nuclei. Such methods can be classified into two groups. Those which model the nuclear wavefunction as a swarm of trajectories that are propagated independently, and those which expand the nuclear wavefunction as a set of fixed-width Gaussian's which form a moving grid of coupled Gaussian's evolving in time. The former includes surface hopping (SH) [110], and the latter ab-initio multiple spawning (AIMS) [111–113], multiconfigurational Ehrenfest (MCE) [114–116], ab-initio multiconfigurational Ehrenfest (AIMCE) [117–119], and direct dynamics variational multi-configurational Gaussian (dd-vMCG) [120]. Gaussian based methods can further be split into single Gaussian basis function (SGBF) methods and multi Gaussian basis function (MGBF) methods. MGBF methods such as AIMS require that a single Gaussian evolves on one PES at a time, and thus uses a set of multiple basis functions. Whereas, SGBF methods such as MCE and its variants allow for a single Gaussian to evolve over a superposition of multiple electronic states at any time. The following subsections will detail the specifics of the methods most pertinent to this thesis. Namely, SH and AIMCE.

2.2.1 Surface Hopping

Surface Hopping (SH) aims to approximate the evolution of the nuclear wavepacket in a semi-classical fashion, within which the nuclei are propagated according to Newtonian equations of motion and the electronic wavefunction is described in a quantum fashion by the ab-initio method of choice [110]. Here the nuclei are modelled by a series of independent trajectories and at any given time step are located at a fixed set of phase space coordinates i.e. the nuclear wavefunction is a $\delta$-function. Moreover, the result is largely dependent on the choice of ab-initio method used to calculate the electronic wavefunction, as this defines the PES on which the nuclei move. For an accurate treatment of the wavefunction at far from equilibrium geometries and population transfer between electronic states, one really requires a multiconfigurational method. However, in larger systems with many electrons, one may have to settle for more approximate methods. SH assumes that for the most part, the nuclei evolve adiabatically on the PES's, and that nonadiabatic population transfer can be approximated stochastically through instantaneous switches or hops between the PES's where the NACME's are large. Thus, it is non-deterministic and relies on the
collective behaviour of swarms of independent trajectories to approximate nonadiabatic dynamics. The classical treatment of the nuclei means that quantum effects such as nuclear interference and tunnelling are not included. However, its simplicity makes it computationally viable for large polyatomic molecules.

In SH, the nuclei are propagated along the PES’s according to Newton’s second law,

\[ M_A \frac{\partial^2}{\partial t^2} R_A (t) = - \nabla_A E_\beta (R_A (t)), \]  

(2.27)

where \( M_A \) refers to the mass of nucleus \( A \), \( R_A (t) \) the coordinates of nucleus \( A \) at time \( t \), and the right hand side is the gradient of electronic state \( \beta \) i.e. the force acting on the nucleus. At each time step, Eqn. 2.27 is integrated using an appropriate algorithm, with the velocity-Verlet algorithm a common choice [121, 122]. The velocity-Verlet algorithm gives the updated nuclear coordinates at time \( t + \Delta t \) as,

\[ R_A (t + \Delta t) = R_A (t) + v_A (t) \Delta t + \frac{1}{2} a_A (t) \Delta t^2, \]  

(2.28)

where \( v_A (t) \) is the nuclear velocity, and \( a_A (t) \) the acceleration at time \( t \), defined as,

\[ a_A (t) = - \frac{1}{M_A} \nabla_A E_\beta (R_A (t)). \]  

(2.29)

Eqn. 2.28 only depends on quantities known from the current time-step, and so the nuclear positions are updated first. Next, the velocities are updated as,

\[ v_A (t + \Delta t) = v_A (t) + \frac{1}{2} \left[ a_A (t) + a_A (t + \Delta t) \right] \Delta t, \]  

(2.30)

where the acceleration at time \( t + \Delta t \) is,

\[ a_A (t + \Delta t) = - \frac{1}{M_A} \nabla_A E_\beta (R_A (t + \Delta t)). \]  

(2.31)

Since the updated velocity in Eqn. 2.30 depends on the gradient of electronic state \( \beta \) at time \( t + \Delta t \), the electronic structure calculation must be performed following the update of the nuclear positions, but prior to the update of the velocity.

The electronic equation of motion is obtained by insertion of the expansion,

\[ |\psi (t)\rangle = \sum_{\alpha=1}^{N_e} c_\alpha (t) |\psi_\alpha (t)\rangle, \]  

(2.32)
into the electronic TDSE, yielding the matrix equation [123],

\[
\frac{\partial}{\partial t} \mathbf{c}^{\text{rep}} = -[\mathbf{H}^{\text{rep}} + \mathbf{K}^{\text{rep}}] \mathbf{c}^{\text{rep}},
\]

(2.33)

where \( \mathbf{c}^{\text{rep}} \) is the matrix of time-dependent wavefunction coefficients, \( \mathbf{H}^{\text{rep}} \) the electronic Hamiltonian matrix, and \( \mathbf{K}^{\text{rep}} \) the dot product of the velocity vector \( \mathbf{v} \) with the NACME's, i.e. \( K_{\alpha \beta} = \mathbf{v} \cdot \langle \psi_\alpha | \nabla | \psi_\beta \rangle \). As there are several representations of the electronic states in which this equation can be integrated, the superscript \( \text{rep} \) in Eqn. 2.33 is used to denote two possible different representations in which SH can be performed. Typically the electronic wavefunction is calculated within the BOA with a standard electronic Hamiltonian \( \hat{H}_e \) that contains no external fields or relativistic corrections, in this representation the Hamiltonian matrix \( \mathbf{H}^{\text{BOA}} \) is diagonal and the elements correspond to the a series of adiabatic Born-Oppenheimer states. Moreover, the elements of the nonadiabatic coupling matrix \( \mathbf{K}^{\text{BOA}} \) are nonzero, often sharply peaked, and describe the nonadiabatic transfer of population between states of the same spin multiplicity, i.e. IC. However, in systems for which the transfer of population between states of different spin is important, i.e. ISC, additional corrections must be included in the Hamiltonian to account for SOC. The total electronic Hamiltonian then becomes \( \hat{H}_{\text{tot}} = \hat{H}_e + \hat{H}_{\text{SOC}} \), meaning that the Hamiltonian matrix in Eqn. 2.33 is no longer diagonal as eigenstates of \( \hat{H}_e \) and \( \hat{H}_{\text{tot}} \) are not identical. The SOC’s contained within the off-diagonal terms are delocalised over the PES, increasing the trajectory sampling requirement. In addition, the inclusion of the spin-orbit term lifts the degeneracy of the different spin multiplets within states of the same spin multiplicity. For example, a triplet state with spin \( S = 1 \) splits into three states corresponding to spin multiplets \( M_s = -1, 0, +1 \). In the SHARC [124] implementation of SH with which we are concerned with in this thesis, this is addressed by performing a unitary transform to a new basis in which Hamiltonian is diagonal i.e. \( \mathbf{H}^{\text{diag}} \). In the diagonal representation all couplings between diagonal states are described by the nonadiabatic coupling matrix \( \mathbf{K}^{\text{diag}} \). This also deals with the degeneracy lifting issues as in the diagonal basis states are split into their individual spin multiplets. However, performing surface hopping directly in this basis is beyond current capabilities since electronic structure codes do not operate in it. Thus calculations from electronic structure codes in the BOA basis are transformed into the diagonal basis. The propagation equation for the coefficients in the diagonal basis is [124],

\[
\mathbf{c}^{\text{diag}} (t + \Delta t) = \mathbf{U}^\dagger (t + \Delta t) \mathbf{P}^{\text{BOA}} (t + \Delta t, t) \mathbf{U} \mathbf{c}^{\text{diag}} (t) = \mathbf{P}^{\text{diag}} (t + \Delta t, t) \mathbf{c}^{\text{diag}} (t), \quad (2.34)
\]
where $U$ is the unitary transformation matrix, and $P^{\text{BOA}}$ the exponential propagator matrix in the BOA representation, defined as,

$$P^{\text{BOA}} = \exp\left[ - (iH^{\text{BOA}} + K^{\text{BOA}}) \Delta t \right]. \quad (2.35)$$

Practically, Eqn. 2.34 achieves the propagation of the coefficients in the diagonal basis in three steps. First, the coefficients $c^{\text{diag}}(t)$ are transformed to $c^{\text{BOA}}(t)$. Second, The propagator matrix $P^{\text{BOA}}$ acts on $c^{\text{BOA}}$ at time $t$ to propagate them to $t + \Delta t$. Finally, the resulting coefficients $c^{\text{BOA}}(t + \Delta t)$, are transformed back into the diagonal basis.

SH implementations require each trajectory to occupy one electronic state at a time. Thus, the gradient of the active state at a specific time governs the evolution of the nuclei. Whether population transfer occurs between two states is stochastic. Therefore, the total population of the active state $\beta$ at time $t$ is given by the fraction of trajectories in the ensemble that are in active state $\beta$ at time $t$. Deciding when population transfer occurs differs between flavours of SH. However, most are based on Tully’s fewest switches criterion, [125] which requires that the total population is maintained using the fewest number of hops between states. This is fulfilled by never hopping from state $\beta$ if $|c_\beta|^2$ is increasing. If $|c_\beta|^2$ is decreasing, the probability of hopping out of state $\beta$ is proportional to the increase of $|c_\beta(t + \Delta t)|^2$,

$$p_{\beta\to\alpha} = 1 - \frac{|c_\beta(t + \Delta t)|^2}{|c_\beta(t)|^2}. \quad (2.36)$$

This however does not specify how the probability is split over a series of target states $\alpha$. The SHARC approach is to partition the probability based on the propagator $P^{\text{diag}}$ as,

$$p_{\beta\to\alpha} = 1 - \frac{|c^{\text{diag}}_\beta(t + \Delta t)|^2}{|c^{\text{diag}}_\beta(t)|^2} \times \frac{\Re \left[ c^{\text{diag}}_\alpha(t + \Delta t) \left(P^{\text{diag}} \right)^* c^{\text{diag}}_\beta(t) \right]}{|c^{\text{diag}}_\beta(t)|^2 - \Re \left[ c^{\text{diag}}_\beta(t + \Delta t) \left(P^{\text{diag}} \right)^* c^{\text{diag}}_\beta(t) \right]}. \quad (2.37)$$

This is fully compatible with the diagonal basis in which SH is conducted for systems that exhibit ISC. At each time step the hopping probabilities are calculated using Eqn. 2.37, a random number $n$ is then generated on the interval $0 \leq n \leq 1$, and a hop from state $\beta$ to $\alpha$ occurs if,

$$\sum_{i}^{\alpha-1} h_{\beta\to\alpha} < n \leq \sum_{i}^{\alpha-1} h_{\beta\to\alpha} + h_{\beta\to\alpha}. \quad (2.38)$$
Unless the states are completely degenerate, the new active state’s potential energy will differ from the old states. To preserve total energy, after a hop occurs the nuclear kinetic energy is adjusted via the nuclear velocity vector \( \mathbf{v} \) in order to conserve the total energy of the system.

### 2.2.2 Multiconfigurational Ehrenfest

Ehrenfest based methods are all based on a Gaussian expansion of the nuclear wavefunction, and are distinguished by the fact that a single Gaussian is distributed across all electronic states [118]. While the nuclei are still propagated classically, the Gaussian expansion and the inclusion of coupling between Gaussian’s in multiconfigurational approaches, provides a more accurate representation of the wavepacket. Furthermore, the transfer of population between states is modelled as a change in amplitude of the Gaussian over all electronic states, providing a description of nonadiabatic transitions that does not rely on stochastic hops like in SH. In Ehrenfest methods, the total molecular wavefunction is expanded in a basis of \( \mathcal{N}_{\text{TB}} \) Ehrenfest wavepackets [114, 115],

\[
|\Psi(t)\rangle = \sum_{n=1}^{\mathcal{N}_{\text{TB}}} c_n(t) |\psi_n(t)\rangle ,
\]

with each Ehrenfest wavepacket,

\[
|\psi_n(t)\rangle = \left[ \sum_{\alpha=1}^{N_e} a_{\alpha}^n(t) |\psi_\alpha\rangle \right] |\chi_n(t)\rangle ,
\]

and the nuclear wavepacket \( |\chi_n(t)\rangle \) defined as a Gaussian in a harmonic potential,

\[
|\chi_n(t)\rangle = \left( \frac{\sigma}{\pi} \right)^{N_d/4} \exp \left( -\frac{\sigma}{2} (\mathbf{R} - \mathbf{R}_n(t))^2 + \frac{i}{\hbar} \mathbf{p}_n(t) (\mathbf{R} - \mathbf{R}_n(t)) + \frac{i}{\hbar} \gamma_n(t) \right) ,
\]

where \( \mathbf{R}_n \) and \( \mathbf{p}_n \) are the position and momentum coordinates at the centre of the Gaussian, and \( \mathbf{R} \) the set of nuclear coordinates over the width of the Gaussian. The Gaussian has dimensionality \( N_d \), corresponding to the number of degrees of freedom of the system. Each of the widths \( \sigma \) are assumed fixed, and are specified for each individual degree of freedom according to the type of atom. The phase of the
Gaussian $\gamma_n(t)$ in Eqn. 2.41, is given by,

$$\gamma_n(t) = \frac{1}{2} \mathbf{P}_n(t) \mathbf{R}_n(t).$$  \hfill (2.42)

The coefficients $a^n_{\alpha}(t)$ in Eqn. 2.40 correspond to the contribution of each quantum state $\alpha$ to its respective Ehrenfest wavepacket. These are normalised to unity for each wavepacket and $|a_{\alpha}(t)|^2$ gives the electronic population of state $\alpha$ at time $t$. The transfer of population between electronic states is therefore modelled as a change in the amplitude of the coefficients $a^n_{\alpha}(t)$. Note, the coefficients $a^n_{\alpha}(t)$ are not coupled between each of the $n$ trajectories. Each individual Ehrenfest wavepacket is propagated independently according to classical equations of motion. The Ehrenfest wavepacket’s are subsequently coupled through the complex coefficients $c_n(t)$ in Eqn. 2.39, which account for the contribution of each wavepacket to the molecular wavefunction [116]. While there exist formulations of MCE in which the evolution of the Ehrenfest wavepackets are coupled at all time steps, this thesis is concerned with the after the fact approach to coupling of AIMCE. [117, 118] This is more conducive to ab initio direct dynamics. First, we shall consider the propagation of the individual wavepackets. Within which, the time evolution of the coefficients $a^n_{\alpha}$ is defined as,

$$a^n_{\alpha} = -\frac{i}{\hbar} \sum_{\beta=1}^{N} \langle \chi_n(t) | \langle \psi_{\alpha} | [\hat{H}_e + \hat{T}_n] | \psi_{\beta} \rangle | \chi_n(t) \rangle a^n_{\beta}(t) - \langle \chi_n \left| \frac{\partial \chi_n}{\partial t} \right| \rangle a^n_{\alpha}(t),$$  \hfill (2.43)

where we have split the total Hamiltonian into electronic $\hat{H}_e$ and nuclear kinetic energy $\hat{T}_n$ terms for ease of determining their contributions to first term of Eqn. 2.43, each of which we shall discuss in turn. First let us consider the contributions from the electronic Hamiltonian $\hat{H}_e$, which is diagonal in the adiabatic basis,

$$\langle \psi_{\alpha} | \hat{H}_e | \psi_{\beta} \rangle = E_{\alpha}(\mathbf{R}_n) \delta_{\alpha\beta},$$  \hfill (2.44)

as $\mathbf{R}$ refers to the set of nuclear coordinates over the whole width of the Gaussian, we require the energies $E_{\alpha}$ to be calculated over these complete set of coordinates simultaneously. Instead, the diagonal contribution from $\hat{H}_e$ to Eqn. 2.43 is approximated by computing the energies at the nuclear coordinates $\mathbf{R}$, which correspond to the centre of each Gaussian only,

$$\langle \chi_n(t) | E_{\alpha}(\mathbf{R}) | \chi_n(t) \rangle \approx E_{\alpha}(\mathbf{R}_n),$$  \hfill (2.45)
thus reducing the computational cost and providing a way to compute the energies $E_\alpha$ at central coordinates $R_n$ on-the-fly. The evaluation of the off-diagonal contribution of $\hat{H}_e$ to Eqn. 2.43 relies on the use of a first-order bra-ket averaged Taylor (BAT) expansion,

$$
\langle \chi_m(t) | E_\alpha(R) \chi_n(t) \rangle \approx \frac{1}{2} \langle \chi_m(t) | \chi_n(t) \rangle (E_\alpha(R_m) - E_\alpha(R_n)) + \frac{1}{2} \langle \chi_m(t) | R - R_m | \chi_n(t) \rangle \frac{dE_\alpha(R_m)}{dR_m} + \frac{1}{2} \langle \chi_m(t) | R - R_n | \chi_n(t) \rangle \frac{dE_\alpha(R_n)}{dR_n},
$$
(2.46)

where we observe that the first order term relies on only energies and gradients at the central coordinates of the Gaussian $R_n$, meaning the computation of coupling between wavepackets is particularly efficient. Now turning to the contributions to the first term in Eqn. 2.43 from nuclear kinetic energy operator $\hat{T}_n$, the approximate diagonal elements are,

$$
\langle \chi_n(t) | \langle \psi_\alpha | \hat{T}_n | \psi_\beta \rangle \chi_n(t) \rangle \approx \frac{P_n(t)^2}{2M} \delta_{\alpha\beta} - i\hbar \frac{P_n(t)^2}{2M} d_{\alpha\beta}(R_n),
$$
(2.47)

with $M$ representing the diagonal matrix of the masses for each degree of freedom, and $d_{\alpha\beta}$ the NACME's. The corresponding off diagonal elements are approximately given by,

$$
\langle \chi_m(t) | \langle \psi_\alpha | \hat{T}_n | \psi_\beta \rangle \chi_n(t) \rangle \approx \langle \chi_m(t) | \hat{T}_n | \chi_n(t) \rangle \delta_{\alpha\beta} - i\hbar \frac{P_n(t)}{2M} (d_{\alpha\beta}(R_m) + d_{\alpha\beta}(R_n)),
$$
(2.48)

which relies on the coupling between states calculated at the Gaussian centres $R_m$ and $R_n$.

The second term of Eqn. 2.43 corresponds to the derivative overlap of two Gaussian basis functions, which when broken down to its position and momentum contributions is,

$$
\langle \chi_n \frac{\partial \chi_n}{\partial t} \rangle = \frac{dR_n}{dt} \langle \chi_n \frac{\partial \chi_n}{\partial R_n} \rangle + \frac{dP_n}{dt} \langle \chi_n \frac{\partial \chi_n}{\partial P_n} \rangle + \frac{i}{\hbar} \frac{d\gamma_n(t)}{dt}
$$

$$
= \frac{dR_n}{dt} \left[ -\frac{i}{\hbar} P_n(t) \right] + 0 + \frac{P_n}{2} \frac{dR_n}{dt}
$$
(2.49)
which depends only on the derivative of the central nuclear coordinates $R_n$, and the corresponding momentum space coordinates $P_n$.

Having discussed all major contributions to the equation of motion governing the evolution of the coefficients $a_{\alpha}^n$ for each individual Ehrenfest wavepacket, we now address the propagation of the wavepacket’s phase space coordinates. The time evolution of these coordinates is governed by the Ehrenfest force, defined as,

$$F_n = -\sum_{\alpha,\beta=1}^{N_s} a_{\alpha}^n(t) a_{\beta}^n(t) \frac{\partial}{\partial R_n} \langle \chi_n(t) | \langle \psi_\alpha | \hat{H}_e | \psi_\beta \rangle | \chi_n(t) \rangle$$

$$= -\sum_{\alpha=1}^{N_s} |a_{\alpha}^n(t)|^2 \frac{dE_\alpha}{dR_n} - \sum_{\alpha \neq \beta}^{N_s} a_{\alpha}^n(t) a_{\beta}^n(t) d_{\alpha\beta} (E_\beta - E_\alpha) ,$$

where the end result on the second line is obtained through application of the approximations in Eqn.’s 2.45 and 2.46. The first term in Eqn. 2.50 represents the gradient of the electronic states weighted by their respective populations, meaning that the nuclei evolve on PES’s in a mean-field fashion. The second term corresponds to the gradient in the direction of the NACME’s, and thus accounts for changes in the force due to nonadiabatic population transfer. In the mean-field treatment of the force, the larger contributions are therefore from the states with the most significant population at a given point in time. Note that in some cases this can lead difficulties with describing the evolution of the nuclei after passing through a region of high nonadiabatic coupling, for example a conical intersection.

Having defined the evolution of each independent Ehrenfest wavepacket, we turn to the time evolution of the coupling coefficients $c_n(t)$ in Eqn. 2.39, which couple the individual wavepackets together. The equation of motion for $c_n(t)$ is defined as,

$$\sum_n \Omega_{mn} c_n(t) = -i \left[ \langle \psi_m(t) | \hat{H} | \psi_n(t) \rangle - i\hbar \left\langle \psi_m(t) \left| \frac{\partial \psi_n(t)}{\partial t} \right. \right] c_n(t) ,$$

where the first term on the left hand side is the overlap of two Ehrenfest wavepackets,

$$\Omega_{mn} = \langle \psi_n(t) | \psi_m(t) \rangle = \left( \sum_{\alpha=1}^{N_s} a_{\alpha}^n(t) a_{\alpha}^m(t) \right) \langle \chi_n(t) | \chi_m(t) \rangle ,$$

(2.52)
which we can see has a dependence on the population coefficients \( a^m_{\alpha} \) of each wavepacket. Looking at the right hand side of Eqn. 2.51, the Hamiltonian matrix elements in the first term can again be evaluated according to the approximations in Eqn.'s. 2.45-2.48. Moreover, the second term on the right hand side of Eqn. 2.51 is the derivative overlap of two Ehrenfest wavepackets,

\[
\left\langle \psi_m(t) \left| \frac{\partial \psi_n(t)}{\partial t} \right| \right. = \left\langle \chi_m(t) \left| \frac{\partial \chi_n(t)}{\partial t} \right| \right. \sum_{\alpha=1}^{N_s} a^m_{\alpha}(t) a^n_{\alpha}(t) \\
+ \left\langle \chi_m(t) \right| \chi_n(t) \right\rangle \sum_{\alpha=1}^{N_s} a^m_{\alpha}(t) \frac{d a^n_{\alpha}(t)}{dt}, \tag{2.53}
\]

where the derivative overlap of two Gaussian’s is given as in Eqn. 2.49. As a final aside, note that Ehrenfest wavepackets are not orthonormal, adding some additional complexity into the computation. At each point in time, normalisation must be ensured by calculation of the norm of the total wavefunction, \( \left\langle \psi_n(t) \right| \hat{1} \left| \psi_n(t) \right\rangle \), where \( \hat{1} \) is the identity operator,

\[
\hat{1} = \sum_{n,m} \left| \psi_n(t) \right\rangle \Omega^{-1}_{nm}(t) \langle \psi_m(t) \right|, \tag{2.54}
\]

with \( \Omega^{-1}_{nm}(t) \) the inverse of the overlap matrix first introduced in Eqn. 2.52.

### 2.3 Scattering Observables

Time-resolved scattering techniques provide a valuable tool for imaging the quantum dynamics of molecular systems. In this thesis, experiments that use both x-ray and electron scattering appear. The following section addresses the underlying theory required to predict scattering cross sections, arriving at a series of approximations used to interpret experiment in subsequent chapters. This involves using semiclassical trajectories obtained by the methods laid out in the previous section to calculate model scattering observables. In experiments, the observable recorded on the detector corresponds to the differential scattering cross-section \( (d\theta / d\Omega) \), which is the number of photons or electrons scattered through the solid angle \( d\Omega \).
2.3 Scattering Observables

2.3.1 X-ray Scattering

In the classical picture of electromagnetism, x-ray scattering (XRS) occurs as a result of the electric field of a photon interacting with the electric field of an electron. This interaction induces a dipole moment in the molecule, causing the electron density distribution to oscillate along the axis of the electric field. [42]. For a free atom, the scattered wave is spherically isotropic with a $180^\circ$ phase shift due to the Thomson scattering. The magnitude of the radiated field at an observation point depends on the polarisation vector and the direction of propagation of the incident photon. Semi-classically, this can be understood in terms of the acceleration $a(t')$ of the charge. The observed acceleration depends on the observation angle $\theta$ as $\sin(\theta) = -\varepsilon \cdot \varepsilon'$, with $\varepsilon$ and $\varepsilon'$ the polarisation vectors of the incident and scattered waves, respectively. Note that technically, the photon-electron interaction is more accurately described in the framework of quantum electrodynamics, where the photons are treated as quantized electromagnetic fields. However, the scattering equations provided later in this section can be derived from the quantum electrodynamics description of light-matter interaction, or also via a semiclassical description [126]. Given that XRS is sensitive to the electronic density this provides a tool for the direct observation of electron dynamics. Moreover, changes in the electronic density affect the motion of the nuclei, and thus XRS indirectly provides information on the nuclear dynamics. The double differential XRS cross-section is given by [127, 128],

$$\frac{d\theta}{d\Omega} = \left( \frac{d\theta}{d\Omega} \right)_{\text{Th}} \sum_{i,j} \sum_{k} \int_{-\infty}^{\infty} I(t-\tau) N_{ki j}(t) \left| L_{ki}(q,R) L_{kj}(q,R) \right| \chi_i(t) \rangle \langle \chi_j(t) \right| dt, \quad (2.55)$$

where $I(t-\tau)$ is the pulse intensity profile, $L_{ki}(q,R)$ the scattering amplitudes, with $i$ and $k$ the initial and final states, and $\left( \frac{d\theta}{d\Omega} \right)_{\text{Th}}$ is the Thomson cross-section, representing the scattering of photon from a free electron. The momentum transfer vector $q$ corresponds to the difference between incident and scattered wave vectors $k_0$ and $k_s$, such that $q = k_0 - k_s$. The window function $N_{ki j}$ acts to weight the contribution from different quantum states. Generally, we use energy integrating detectors, meaning that we observe the total scattering signal, corresponding to sum of contributions from all possible quantum states. These contributions, in order of their magnitude are, the elastic, inelastic and coherent-mixed scattering (CMS) contributions. To better understand these contributions to the total scattering signal, we shall first consider the theoretical case in which they are resolvable. The scattering amplitudes $L_{ki}(q,R)$ in Eqn. 2.55 are expectation values of the one electron scattering operator $\hat{L}$ and the
2.3 Scattering Observables

Electronic states $\psi_i$,

$$L_{ki}(q, R) = \langle \psi_k | \hat{L}(q) | \psi_i \rangle,$$

which correspond to probability amplitudes of single electron scattering from the system and can be written as a Fourier transform of the electron density,

$$L_{ki}(q, R) = \int_{-\infty}^{\infty} e^{iq \cdot r} \rho_{ki}(r, R) dr,$$

where $\rho_{ki}(r, R)$ is the single electron density at electronic coordinate $r$. Turning again to differential cross section in Eqn. 2.55, we can now look to understand each of the terms in the equation. Elastic scattering corresponds to the case where $i = j = k$ in Eqn. 2.55, thus no energy is transferred from the incident photon to the molecule and the matrix elements in the final term become $\langle \chi_i (t) | | L_{ii}(q, R) |^2 | \chi_i (t) \rangle$. Inelastic scattering occurs when there is an electronic transition from state $i \rightarrow k$, meaning energy is transferred between the incident photon and molecule. This corresponds to the case where $i = j \neq k$, with matrix elements $\langle \chi_i (t) | | L_{ki}(q, R) |^2 | \chi_i (t) \rangle$. CMS occurs when two scattering amplitudes from a wavepacket in a superposition of electronic states interfere. This corresponds to $i \neq j$ in Eqn. 2.55, with matrix elements $\text{Re} \left[ \langle \chi_j (t) | | L_{kj}^*(q, R)L_{ki}(q, R) | \chi_i (t) \rangle \right]$. While CMS provides potential of an exciting tool for probing the dynamics that occurs directly at regions of strong nonadiabatic coupling, it is yet to be observed experimentally due to its small cross-section. From here on we discard the contribution of CMS to the scattering signal as it does not play a central role in this thesis.

As previously touched on, the detectors employed do not not posses any energy resolution, and therefore can not resolve different electronic transitions. This results in all contributions to the signal being detected together in the form of the total scattering. Therefore, the single electron matrix elements $L_{ki}(q, R)$ in Eqn. 2.55 are replaced by matrix elements corresponding to the scattering probability from a molecule [129],

$$\Lambda_{ki}(q, R) = \langle \psi_k | \hat{L}^i(q) \hat{L}(q) | \psi_i \rangle$$

$$= \sum_{a,b} \langle \psi_k | e^{iq \cdot (r_a - r_b)} | \psi_i \rangle$$

$$= N_{el} \delta_{ik} + \Lambda_{ki}^{2e}(q, R),$$
this involves the product of two one electron scattering operators and thus can be split into a single and two electron component. The first term on line three in Eqn. 2.58 is the one electron part where \( a = b \). This only contributes if the initial and final states are the same (i.e. \( i = k \)), and corresponds to the number of electrons. The second two electron term is a Fourier transform of the two electron transition density \( \rho_{ki}(r_1, r_2, R) \),

\[
\Lambda_{ki}^{2e}(q, R) = 2 \int \int_{-\infty}^{\infty} e^{i(q \cdot (r_1 - r_2))} \rho_{ki}(r_1, r_2, R) dr_1 dr_2, \tag{2.59}
\]

where \( r_1 \) and \( r_2 \) are the coordinates of electron one and two respectively. Again, the matrix element in Eqn. 2.58 has three contributions. These are detected together and add incoherently as,

\[
\frac{d \theta}{d \Omega} = \left( \frac{d \theta}{d \Omega} \right)_{1e} + \left( \frac{d \theta}{d \Omega} \right)_{td}^{2e} + \left( \frac{d \theta}{d \Omega} \right)_{CMS}^{2e}, \tag{2.60}
\]

with the first term corresponding to a time-independent background that is proportional to the number of detectable electrons and photons, this background term is associated with the one electron part of Eqn. 2.58. The second term corresponds to the diagonal part (i.e. \( i = k \)) of the two electron component \( \Lambda_{ki}^{2e}(q, R) \), yielding a structure dependant signal that changes in time. The final term, reflects the contribution from CMS, here the off diagonal of the two electron component. Together the first two terms of Eqn. 2.60 correspond to the elastic and inelastic contributions, detected together due to the lack of energy resolution. The fact both elastic and inelastic are integrated together on the detector makes their separation impossible. One can however try to identify inelastic contributions to the integrated total signal by a comparison with theory, although the calculation of this can be extremely computationally demanding.

### 2.3.2 Electron Scattering

In contrast to XRS, electron scattering is sensitive to both the electronic density and the nuclei, thus providing a mechanism to directly probe both the electronic and nuclear dynamics simultaneously. Furthermore, the sensitivity to nuclei results in a larger scattering cross-section, meaning it is more sensitive to lighter atoms such as Hydrogen that are typically not detected in XRS. Thus, in principle, one can collect more data. The high energies available in electron beams generally means that data can also be collected over a wider range of momentum transfer than in XRS. However, as we shall explore throughout this thesis, electron scattering is not
2.3 Scattering Observables

without its limitations, for example it typically suffers from poorer temporal resolution. The electron scattering differential cross-section is given by,

\[
\frac{d\theta}{d\Omega} = \left( \frac{d\theta}{d\Omega} \right)_{Ru} \sum_{i,j} \sum_{k} \int_{-\infty}^{\infty} I(t - \tau) N_{kij} \langle \chi_j(t) | Z_{kj}^{*}(s, R) Z_{ki}(s, R) | \chi_i(t) \rangle dt , \tag{2.61}
\]

where in comparison to the XRS case in Eqn. 2.55, the Thomson cross-section has been replaced with the Rutherford cross-section \( \left( \frac{d\theta}{d\Omega} \right)_{Ru} \) describing the scattering of an electron from a charged particle. As usual, the momentum transfer vector is now defined as \( s \), and \( Z_{ki}(s, R) \) is the one electron scattering amplitude,

\[
Z_{ki}(s, R) = L_{ki}(s, R) - \delta_{ki} \sum_{A} Z_{A} e^{iR_{A}}, \tag{2.62}
\]

where \( L_{ki}(s, R) \) is the XRS amplitude introduced previously. The addition of the term which is a sum over all \( N_{A} \) nuclei describes the scattering by a nuclear potential. \( Z_{A} \) is the atomic number of nuclei \( A \). Multiplication by the Kronecker delta \( \delta_{ki} \) reflects the fact that the nuclei only contribute to elastic scattering\(^3\). The addition of the Rutherford cross-section in Eqn. 2.61 results in the intensity being scaled by a factor of \( \frac{1}{s^4} \), causing it to decay faster at large values of \( s \). In contrast, the signal is enhanced at smaller values of \( s \), where it is predominately elastic. This means the small to intermediate \( s \) range captures the nuclear motion particularly well.

The cross-section in Eqn. 2.61 can again be understood in terms of its three components. The dominant elastic contribution corresponds to matrix elements,

\[
\langle \chi_i(t) | |Z_{ii}(q, R)|^2 | \chi_i(t) \rangle ,
\]

and the elastic scattering probability now contains additional terms in comparison to its x-ray equivalent,

\[
|Z_{ii}(q, R)|^2 = |L_{ii}(q, R)|^2 + \sum_{A=1}^{N_{A}} Z_{A}^2 + 2 \sum_{A=1}^{N_{A}} \sum_{B>A} Z_{A} Z_{B} \cos (s \cdot R_{AB}) - \sum_{A=1}^{N_{A}} Z_{A} \text{Re} \left( L_{ii}(s, R)e^{-isR_{A}} \right) , \tag{2.63}
\]

\(^3\) The nuclei only contribute elastically when considering only electronic transitions. In the case of a rovibrational wavepacket, the nuclei may contribute inelastically.
where $R_{AB}$ is the Euclidean distance between nuclei $A$ and $B$. Here the term on line one, $(|L_{ii}(q, R)|^2)$ is the elastic scattering probability from the electrons only. Lines two and three correspond to the elastic scattering from the nuclei. The final term on line four represents the interference between scattering amplitudes of the electrons and nuclei. As inelastic scattering is inherently electronic, apart from the scaling by the Rutherford cross-section, its contribution in electron scattering is identical to its x-ray equivalent. While we are not concerned with the CMS component, it is interesting to note that it is not purely electronic and also includes an electron-nuclei cross term.

Considering the case of the energy integrated detector’s typically used, the matrix elements in Eqn. 2.61 can be replaced with the total scattering probability from a molecule. In terms of its one and two particle components this is,

$$\Lambda^{el,1}_{kl}(s, R) = \delta_{ik} \left(N_e + \sum_{A=1}^{N_A} Z_A^2 \right) + \Lambda^{el,2p}_{ki}(s, R), \quad (2.64)$$

with the two particle part,

$$\Lambda^{el,2p}_{ki}(s, R) = 2 \delta_{kl} \sum_{A=1}^{N_A-1} \sum_{B=1}^{N_A} Z_A Z_B \cos \left(s \cdot R_{AB} \right) + \Lambda^{2e}_{ki}(s, R) - 2 \sum_{A=1}^{N_A} Z_A \text{Re} \left[ L_{ki}(s, R) e^{-i s \cdot R_A} \right]. \quad (2.65)$$

The one particle part of Eqn. 2.64 accounts for the incoherent elastic scattering from the electrons and nuclei. This is static, and leads to a time-independent background as in XRS. Turning to the two particle component in Eqn. 2.65, the first term also accounts for the incoherent elastic scattering from both electrons and nuclei. However, this term is time-dependent and provides structural information. The second term in the two particle component contains both the one and two electron scattering matrix elements, as given in Eqn. 2.56 and Eqn. 2.59 respectively. The two electron matrix element only describes scattering from electrons. However, the one electron matrix element describes the interference of the electron’s scattering amplitude with the nuclei through a cross term. Integrated together with the background from the one particle component, the two particle component (when $i = k$) contains the elastic scattering from electrons and nuclei, as well as the inelastic scattering from electrons. Again, the off-diagonal elements of $\Lambda^{el,2p}_{ki}(s, R)$ correspond to the yet to be detected contribution from CMS.
2.3 Scattering Observables

2.3.3 Independent Atom Model

The majority of experimental studies are typically analysed by comparison with signals calculated assuming the independent atom model (IAM), originally formulated by Debye [130]. The IAM is a convenient, computationally efficient approximation, shown to be sufficiently accurate in many circumstances. This approximation assumes that the electronic density does not change significantly, and therefore can be approximated by the ground state equilibrium electron density of individual atoms, averaged over all possible orientations. This works well as a large part of the scattering is dominated by tightly bound core electrons, while valence electrons contribute little to the total scattering. Within the IAM, the scattered intensity from a molecule can be simply calculated from the interference between all individual atomic scattering amplitudes. However, this only accounts for the elastic contribution to the scattering signal, which is often sufficient given that it is the dominant contribution. While equations allowing the calculation of the total scattering signal from multiconfigurational electronic wavefunctions have been derived, properly accounting for the electron density and transitions from all quantum states in this fashion is a phenomenal computational task [86, 87]. These calculations generally scale with the number of basis functions used to describe an electronic wavefunction, and thus are usually only feasible for small molecules in which a reduced set of sample structures can be identified. Although a recent seminal study was successful in using this approach to simultaneously observe electronic and nuclear dynamics by electron scattering, showing that a side-by-side comparison of theoretically calculated inelastic contributions from different electronic states to the experimental signal can provide insight into electronic dynamics [23].

Within the assumptions of electron density described above, the elastic XRS amplitude in Eqn. 2.56 reduces to,

\[
L_{ii}(q, R) = \langle \psi_i | \hat{L} (q) | \psi_i \rangle \approx \sum_{A=1}^{N_A} f_A(q) e^{i q \cdot R_A},
\]

(2.66)

where \( f_A(q) \) is known as the atomic form factor. The form factor is the rotationally averaged scattering amplitude of a single atom \( A \) in its ground state, and corresponds to the Fourier transform of the ground state electron density. The atomic form factors are known values that have been parameterised experimentally [131], thus making the calculation of the elastic scattering intensity straightforward. The elastic scattering
probability now becomes,
\[
|L_{ii}(q, R)|^2 = \sum_{A,B=1}^{N_A} f_A(q) f_B(q) e^{i q R_{AB}},
\] (2.67)

which amounts to a coherent summation of the scattering amplitudes from individual atoms. Here all terms with \(A = B\) contribute a static background to the intensity - known as the atomic scattering term, whereas all terms \(A \neq B\) correspond to the interference of amplitudes from different atoms, often referred to as the molecular scattering term. The molecular term is sensitive to changes in geometry and therefore indirectly probes the evolution of nuclei. While this often provides satisfactory results, it does not allow for the probing in changing electronic density. In some cases, these assumptions are warranted since the inelastic contribution to the total intensity is smaller and many experiments do not possess the required resolution to observe these contributions. In addition, the IAM does not capture the delocalization of valence electrons as the density is assumed isotropic and localised on the individual nuclei.

All of this means that the IAM cannot capture either the delocalization of electrons in bond formation, or the redistribution of electrons between excited states. Both of which are inherent to photo-chemical processes. However, as experimental technology continuous to develop, one can look forward to performing experiments that are capable of directly observing these phenomena.

In the case of electron scattering, the scattering amplitude in Eqn. 2.62 within the framework of the IAM is [132],
\[
Z(s, R) = \sum_{A=1}^{N_A} f_A^e (s) e^{i s R_A},
\] (2.68)

where the electron scattering form factor is defined as,
\[
f_A^e (s) = f_A (s) - Z_A,
\] (2.69)

with \(Z_A\) the atomic number of nuclei \(A\), and \(f_A (s)\) the corresponding x-ray scattering form factor describing scattering from the electrons. With these definitions in hand, the elastic electron scattering probability is then,
\[
|Z_{ii}(s, R)|^2 = \sum_{A,B=1}^{N_A} f_A^e (s) f_B^e (s) e^{i s R_{AB}}.
\] (2.70)
Note that in practice, the electron scattering form factors are often approximated from the tabulated x-ray form factors using the Mott-Bethe formula [132, 133],

$$f_{MB}^A(s) = \frac{Z_A - f_A(s)}{s^2},$$  \hspace{1cm} (2.71)

which is comparable to Eqn. 2.69 apart from a factor of -1 and $s^{-2}$, otherwise accounted for by the scaling of the Rutherford cross-section.

Gas phase scattering experiments are typically performed assuming an initially isotropic molecular ensemble, within which excitation is equal for all transition dipole moments. The initial distribution of population on the excited states is therefore dependent only on the magnitude of their respective transition dipole moments. However, excitation along a transition dipole moment does introduce a small degree of alignment in the case of single photon excitation. This can introduce some anisotropy in the signal. Ideally, the signal should be decomposed into a series of Legendre polynomials, allowing the completely isotropic component to be extracted. In addition, the higher order Legendre terms may provide additional information on the character of the specific excited states involved. Often it is the case that experiments suffer from noise, meaning that the Legendre decomposition cannot be performed and instead one simply averages the total signal recorded on the detector, yielding a roughly rotationally averaged pattern. In this case, rotational averaging of the x-ray scattering probability in Eqn. 2.67 yields,

$$|L_{ii}(q, R)|^2_{\text{Debye}} = N_A \sum_{A,B} f_A(q) f_B(q) \frac{\sin(qR_{AB})}{qR_{AB}},$$  \hspace{1cm} (2.72)

which is the Debye scattering formula [130], well known to the x-ray diffraction community. In the case of electron scattering, the scattering probability in Eqn. 2.73 reduces to,

$$|Z_{ii}(s, R)|^2_{\text{Debye}} = \sum_{A,B=1}^{N_A} f_A^e(s) f_B^e(s) \frac{\sin(sR_{AB})}{sR_{AB}}.$$  \hspace{1cm} (2.73)

These equations provide a practical solution for rotationally averaging gas phase experiments that include no explicit alignment. However, it is possible to explicitly introduce alignment of the gas ensemble by application of an additional electromagnetic field. This provides a tool for selectively preparing the initial distribution of states that the wavepacket populates. The resulting signal is anisotropic and reflects the
angular distribution of the excited states involved, allowing one to probe specific excited states. While such aligned experiments have provided seminal contributions to the understanding of excited state processes [134], they are typically non-trivial to perform, and thus are usually not attempted.

The assumptions within the IAM framework mean it only directly describes the elastic contribution, additional inelastic corrections can be included. This is achieved by approximating the inelastic terms as a sum of tabulated incoherent scattering functions for each individual atom. The total XRS probability is then approximated as,

$$\Lambda_{ki}(q, R) \approx |L_{ii}(q, R)|^2_{\text{Debye}} + \sum_{A=1}^{N_A} S_{A}^{\text{inel}}(q),$$  \hspace{1cm} (2.74)  

and the electron scattering probability,

$$\Lambda_{el}^{ki}(s, R) \approx |Z_{ii}(s, R)|^2_{\text{Debye}} + \sum_{A=1}^{N_A} S_{A}^{\text{inel}}(s),$$  \hspace{1cm} (2.75)  

where the incoherent scattering functions $S_{A}^{\text{inel}}$ are independent of the nuclear geometry and only contribute to the total signal as a time-independent background. Therefore, they do not capture the changes in electron density associated with inelastic scattering.

In the case of electron scattering, it is worth noting that additional relativistic effects may need to be accounted for in some circumstances. For example, in the case of giga-electron volt (GeV) and mega-electron volt (MeV) experiments, the latter of which we shall explore in later chapters. At these energies the electron’s speed reach a significant portion of the speed of light - $\approx 90\%$ in the case of MeV beams. As we previously discussed in the context of SOC, at these speeds the electrons mass on which the relativistic energy $E = mc^2$ depends, is not constant. This in turn, effects the intensity of the scattered electrons. To account for this, the atomic electron scattering form factors can be replaced with relativistic form factors calculated for a given beam energy.
2.4 Data Mining Methods

Complex and high-dimensional data sets, such as those involved in ultrafast photochemical imaging and QMD simulations, are inherently difficult to work with. Data mining methods, which aim to build summary models of large complex data sets, have seen wide success in applications to many scientific fields such as in the biological sciences, social sciences, and engineering [135]. In light of this, this thesis addresses both the application and development of such methods in the context of ultrafast photochemistry. As we shall later explore, methods from data science can be used to analyse and classify trajectories from QMD trajectory simulations, and to identify distinct reaction or decay pathways available to a photoexcited molecule. This improves the interpretation and understanding of complex simulations and the photochemistry that they describe. We envision that such methods will aid the construction of reduced dimensionality models, comparison to experiment, and aid inversion and interpretation. Moreover, as experiments improve, increasingly sophisticated (and expensive) calculations of accurate observables are necessary. An example of this is going beyond the IAM and calculating the total scattering from ab-initio wavefunctions, as discussed in the previous section. By identifying reduced dimensionality models, data mining methods provide a tool for making accurate observable calculations possible for larger systems. In the remainder of this chapter, we shall take a close look at algorithms that aim to classify simulation data both spatially and temporally.

2.4.1 Spatial Clustering

Given a data set of \( N_{\text{obs}} \) observations (i.e. data points) and \( N_f \) features (i.e. the number of variables), a clustering algorithm aims to group the observations into \( k \) distinct clusters based on some measure of similarity across the \( N_f \) dimensional feature space [135, 136]. Generally, one aims to achieve a clustering where each of the \( k \) clusters are sufficiently different and all points within each cluster are similar across the feature space - i.e. the intracluster variance is minimised and the intercluster variance is maximised. An example of different types of clusters in an example two dimensional feature space is given in Fig. 2.1. In each case, cluster membership is represented by the shared colours of data points, with black points corresponding to data points that cannot be classified and are thus marked as noise. Note, some algorithms may not allow for the identification of outliers, and the definition of noise depends on the algorithm of choice. The left hand panel represents the simple case of a series of convex clusters that can all be separated by a straight line - i.e. they are
clusters are considered to be convex if, when all pairs of points in the cluster are connected by straight lines, these lines never exit the cluster. Here the group of clusters have different sizes and densities, often it can be difficult to separate clusters it data sets that contain highly variable density. In the right hand panel, we see the more complex case of nonconvex clusters, which may or may not be linearly separable. In the case of the two overlapping “U” shaped clusters in the upper portion of the panel, these cannot be separated by a linear boundary, and are therefore nonlinearly separable. It is worth noting that not all clustering algorithms are capable of identifying all types of clusters simultaneously. Generally each algorithm is limited in one way or another, for example, one algorithm may be good at identifying clusters that are linearly separable, but not nonlinearly separable. Therefore, the choice in clustering algorithm often depends on the specific type of data set involved.

![Convex vs Nonconvex Clusters](image)

**Figure 2.1**: Schematic showing different shapes of clusters in a two-dimensional feature space, with red, yellow, magenta, and green identifying four different clusters and black indicating outlier points. The left panel shows examples of convex clusters that are linearly separable (as indicated by the dashed lines), with both compact (red, yellow, and green) and diffuse (green) clusters. The right panel shows nonconvex clusters, which may not be linearly separable (as is the case for the red and yellow clusters).

When dealing with high dimensional and correlated data, one typically has to identify a subset of less correlated features to cluster on. This is largely a result of the "curse of dimensionality", as the number of features increases, the volume of the space the data spans increases exponentially. In addition to increasing the computational complexity, this leads to a host of other issues. For example, the decreased ability of distance metrics to distinguish different clusters, biasing of the clustering result, and overfitting. Biasing typically occurs when a number of highly correlated features
are included, increasing their relative weighting in the partitioning of data. Moreover, overfitting can occur when features that are irrelevant to the important trends are included, preventing the algorithm from identifying the underlying structure of the data and grouping points based on what is essentially noise. Fortunately, there exist many feature selection and dimensionality reduction techniques to combat this [135]. These aim to identify the smallest number of features that capture most of the variance in the entire data set. A particularly important technique is principal component analysis (PCA), which constructs a series of new basis vectors (principal components) that are linear combinations of the original features [137]. Clustering can then be performed directly in the basis of a small number of principle components which capture a high degree of the total variance. In addition, time-independent component analysis (tICA) is another method of dimensionality reduction that has been widely used within the classical molecular dynamics community in identifying slow order parameters from a large number of features [138].

Clustering algorithms can loosely be categorised into three main types, 1) those based on the use of a distance metric as a measure of similarity, 2) those based on generative probabilistic models, and 3) density based algorithms [135]. The first category, distance based algorithms, include algorithms which iteratively refine the partitioning of the data into a user specified number of clusters and methods that form clusters hierarchically at various levels of granularity. This includes the \( k \)-means [139] and agglomerative clustering algorithms [140, 141]. The \( k \)-means algorithm proceeds by initialising \( k \) centroids randomly, data points are then assigned to the closest centroid and these are updated based on the mean of all points within the respective centroid. This procedure is repeated until the change in location of the centroids does not vary between iterations. While easy to use, \( k \)-means is highly sensitive to the initialisation of the centroids, requires the user to specify the number of clusters, and cannot identify nonlinearly separable or nonconvex clusters.

Agglomerative clustering on the other hand, starts by considering all data points as individual clusters, these are then merged based on a distance metric of choice [140, 141]. This allows clusters to be built up at multiple levels of granularity, however it also struggles with both nonconvex and to a lesser extent, nonlinearly separable clusters. Moreover, it can often lead to imbalanced cluster sizes as it prioritises the merging of smaller clusters first. The distance based agglomerative algorithms are among the most commonly employed due to their simplicity and general applicability, given a suitable distance metric is defined.
2.4 Data Mining Methods

The second category, probabilistic models [142], assume that the data can be modelled by a probability distribution, such as a mixture of Gaussian distributions [143]. The model parameters are refined using the observed data to achieve a maximum likelihood fit, and from this the fit the probabilities of each data point in the model are determined. Such approaches can suffer due to the assumed form of the probability distribution, the need for the refinement of a large number of model parameters, and may struggle with convergence in relation to the algorithms used to estimate these parameters.

Finally, the category of density based clustering algorithms assign clusters based on the local density of data points in the feature space [144]. The definition of density varies, from a number of points in a given local volume to kernel estimates of the density. A major advantage of density based methods is that they can easily detect clusters that are nonconvex and nonlinearly separable\(^4\), as well as those that are irregular in size, they also often include forms of outlier detection therefore making them fairly robust with respect to noise. For these reasons, we focus on density based algorithms in our application to QMD trajectory data. However, it is worth noting at this point that these algorithms can sometimes struggle to identify multiple clusters of highly varying density, which can potentially cause issues in applications to QMD simulation data, as we shall explore in subsequent chapters.

We shall now take a deeper look at density clustering algorithms, reviewing the basic concepts of the most widely used algorithm, DBSCAN [145]. This algorithm involves specifying a neighbourhood around each individual data point, with points of "core density" including a minimum number of other data points within the radius of this neighbourhood. Clusters are then built up from the regions of core density that overlap and any additional points that border them. Key to DBSCAN is the idea of density reachability. A data point is said to be directly density reachable from a point \( p \) if it is within a radius \( \varepsilon \) from \( p \), where \( \varepsilon \) defines the neighbourhood and its size. A point \( q \) outside the radius \( \varepsilon \), is said to be density reachable from \( p \) if there exist a series of successive directly density reachable points between them. Both of these definitions can be seen in Fig. 2.2a, where all points within \( \varepsilon \) of \( p \) are directly density reachable with respect to \( p \), and \( q \) is density reachable from \( p \) as they are connected by a series of directly density reachable points. Furthermore, a point \( m \) is said to be density connected to a point \( q \) outside of its radius \( \varepsilon \) if there exists a point \( p \) from which both \( q \)

\(^4\) This is also true of spectral clustering, which is based on finding the optimal cuts of the data represented as an undirected weighted graph.
Figure 2.2: Depiction of the main concepts in DBSCAN. a) two data points \( p \) and \( q \) with \( \varepsilon \) defining the area within which other points are directly density reachable. b) two points \( q \) and \( m \) are density connected by \( p \). c) The building up of core (red), border (yellow) and noise (blue) points. d) Clusters identified by DBSCAN in an example data set. Grey points correspond to noise.

and \( m \) are density reachable, as in Fig. 2.2b. Using these rules, the algorithm identifies so-called core and border points in the data set. Core points are surrounded by at least \( MinPts \) directly density reachable data points within their radius \( \varepsilon \). Border points do not require \( MinPts \) data points, but must be density connected to a core point. All other points are classified as noise points and marked as outliers. Figure 2.2c depicts a series of core (red), border (orange) and noise points (blue) for a simple example. Once all points are built up, clustering is complete as in Fig. 2.2d. The two parameters \( \varepsilon \) and \( MinPts \) are therefore crucial to the DBSCAN algorithm. By increasing the size of the neighbourhood defined by \( \varepsilon \), one can detect more sparsely arranged data at a
higher level of granularity. However, this might mean that localised and highly dense clusters may not be detected. Since the size of the density neighbourhood depends on $\varepsilon$, which is defined for all data points, the algorithm is not always capable of detecting clusters of varying densities.

Naturally, density based approaches are well suited to grid based representations of the feature space. Algorithms such as GRIDCLUS [146] and BANG [147] partition the feature space into a grid of $N$-dimensional blocks by successive binary divisions along each dimension. The clustering result depends on the number of times that the feature space can be successively split, i.e. the number of levels. In addition, blocks with a density $< \rho_{\text{min}}$ can be treated as noise. An example of the BANG algorithm applied to a simple 2D data set can be seen in Fig. 2.3. The partitioning of space into blocks is shown, one can see the non-uniform grid of blocks preserves the topological features of the data. Once the feature space is partitioned, blocks are sorted in order of increasing density and those with the highest density become cluster centres. A recursive neighbour search merges adjacent blocks until all points are searched. The main advantage of these methods is that due to hierarchical fashion in which blocks are merged, they can capture the data are various levels of granularity, meaning that it can be better suited to identifying multiple clusters of varying density.

We shall soon explore the application of both BANG and DBSCAN to QMD trajectory data, however we see that the fixed definition of density in these methods means that they are generally not well suited to identifying clusters as a function of time, at least without parameterisation of $\varepsilon$ and $\text{MinPts}$ in the case of DBSCAN, and the number of successive binary divisions (levels) of the feature space in BANG clustering. Ultimately this leads to the development of a new density clustering algorithm, better suited to the large changes in variance observed over time for QMD trajectory data.

### 2.4.2 Temporal Clustering

When considering time-dependent data, such as QMD trajectories, temporal clustering aims to group data points based on their observed temporal patterns and characteristics. For example, identifying high traffic routes [148, 149], studying human behaviour patterns [150], tracking disease progression [151], monitoring animal migration patterns [76], and identifying climate trends [152]. Generally, temporal data can be clustered within two main frameworks, corresponding to the time series and spatio-temporal approaches, each of which are similar but differ slightly in their treatment of the data.
First, the time series approach involves treating time-dependent data as a sequence of data points measured over a time interval, with each time point corresponding to an observation. This treatment is typically applied to weather, financial market, and sensor data [153, 154]. One can however, consider trajectory data as a special case of time series data in which the trajectories are considered as a sequence of spatial coordinates over time, i.e. a time series of locations. Although not always the case, time series approaches tend to use global measures of similarity over all time. This allows one to compare the whole sequence of data points over time at once, and can be thought of as being largely based on the shape of the time-dependent data. This is demonstrated in Fig. 2.4, which shows four different one dimensional time series colored according to their most similar counterpart over all time. One could for example, use the Euclidean distance over the time axis to provide a single valued measure of similarity. This can be applied to all pairs of trajectories to build a similarity matrix that can be input into a clustering algorithm that takes precomputed distances,
such as DBSCAN. There are a wide selection of similarity measures to be chosen from for application to time series data, the choice of which depends on the data. Many of these are also easily extendable to multidimensional time series data, as we shall see in the following chapter.

Figure 2.4: Example of how similarity between data is measured in the time series approach. There are two clear groups of time series, indicated in blue and orange. In the lower portion of the figure a global distance measure compares each point in time together, giving a single valued measure of similarity over the whole time axis.

In contrast to the largely shape based time series approach, the spatio-temporal approach considers trajectories as moving objects, directly accounting for the co-ordinates of the objects in space and their patterns over time [155, 156]. Generally, this involves identifying what objects are spatially close together over certain time steps. This can be considered a form of pattern mining in which a suitable spatial clustering algorithm is applied to each time step independently, a second algorithm is then applied to the set of identified spatial clusters in order to determine what trajectories are clustered together in time. This approach is generally more computationally demanding than the time series approach due to the application of a spatial clustering algorithm at each individual step. However, it can provide a more complete picture of the trends in the data, due to the fact it is not only concerned with the shape of the trajectory evolution over all time. For example, the spatio-temporal approach can also provide information on how frequently different locations are visited, how long the objects spend at these locations, and at what times do objects travelling together diverge. A typical example of where this is useful is in studying patterns in vehicle traffic [75]. For example, a group of transport and city planners may wish to identify what local amenities are often visited by people, what amenities people from different neighbourhoods visit, how long do people stay there, what routes are taken to different
amendities, and do these share any common pathways [149]. This information can then be used to tackle problems such as optimising the flow of traffic, identifying where to build new amendities, and reducing the number of excessive long distance journeys by ensuring each neighbourhood has basic local amendities available to them. One can see the parallels with studying excited state dynamics here, where we may wish to identify the key regions of the phase space explored, if there are any excited state minima that trajectories get stuck in, what different products are available, and through which series of excited states does the system decay.

The area of trajectory pattern mining [155] has seen the development of several definitions of moving object clusters such as flocks [157, 158], moving clusters [159], convoys [160], and swarms [161]. Generally, these all define a spatio-temporal cluster as a group trajectories moving closely together for at least $t_{\text{min}}$ time steps, however they also include additional parameters that vary between the definitions. Flocks for example, are defined by a minimum number of trajectories being spatially grouped within a certain distance threshold for a number of consecutive time steps. This definition requires spatial clusters between time steps to have a 100% overlap, and are therefore most commonly employed to problems in which the moving objects are very tightly bound. Moving clusters differ from the definition of a flock by requiring spatial clusters between consecutive time steps to have an overlap greater than a user specified threshold, and can therefore identify groups of more loosely travelling objects. However, if the overlap of clusters between time steps or the minimum number of trajectories falls below the specified thresholds, the definition breaks down.

Convoys involve a slightly different treatment where one is not concerned with the idea of an overlap between spatial clusters in time. Instead one aims to identify a minimum set of trajectories that are continuously clustered together in time. Considering two temporally consecutive spatial clusters, this amounts to taking the intersection of the members between the two time steps. Therefore, once a trajectory diverges from the convoy, it cannot rejoin the same convoy at a later time and must instead be considered separately.

These definitions can be better understood by examining Fig. 2.5, in which we consider an example set of four trajectories $O = [o_1, o_2, o_3, o_4]$ over three time steps $T = [t_1, t_2, t_3]$. Each slice represents an individual time step and the ellipses represent the neighbourhood that defines a spatial cluster of data points in a 2D feature space. The top panel shows the moving cluster approach, which requires spatial clusters containing at least $\text{MinPts}$ data points to have an overlap $\geq \theta_{\text{min}}$ for a minimum
of $t_{\text{min}}$ successive time steps. Here, setting $\theta_{\text{min}} = 0.75$, $\text{MinPts} = 3$, and $t_{\text{min}} = 3$, results in the formation of a moving cluster over the whole series of time steps. Moving clusters can thus be thought of as spatio-temporal clusters that exhibit a shared spatial character in time. The bottom panel defines the convoy approach, which identifies a group of moving objects as a set of at least $\text{MinPts}$ trajectories moving together for $t_{\text{min}}$ consecutive time steps. Setting $\text{MinPts} = 2$ and $t_{\text{min}} = 3$ results in a convoy of two trajectories (shown in red in the bottom panel). Here it is apparent that convoys are built up by taking the intersection of the current convoys members with the spatial cluster at the next step in time. For example, at time $t_1$ a convoy is initialised containing three trajectories $[o_1, o_2, o_3]$, at time $t_2$ trajectory $o_3$ leaves the convoy, leaving only trajectories $o_1$ and $o_2$ remaining over the two subsequent time steps. At time step $t_2$, trajectory $o_4$ joins the spatial cluster, but since it was not included at the initialisation of the convoy it cannot be assigned to it. While not shown in Fig. 2.5, flocks correspond to spatial clusters having 100% overlap for $t_{\text{min}}$ consecutive time steps, and thus a flock could only be identified if all four of the trajectories were members of the same spatial cluster over the three time steps.

While widely applicable to a range of different types of data, all the aforementioned methods require that trajectories are clustered together for consecutive time steps. Therefore, they cannot describe small term deviations from the ensemble, which poses a problem for QMD trajectory data.

In contrast, swarms take a vastly different approach to the mining of temporal patterns. A swarm is defined as a minimum number of trajectories clustered together for a minimum number of time steps, however the time steps do not have to be sequential. First, a spatial clustering algorithm is applied to each time step, and the membership of every trajectory over the whole set of times is stored. Swarms are then mined from the search space by expanding a tree of all possible trajectory combinations in depth-first search\(^5\) order. The tree is initialised from a root node, from which a series of nodes corresponding to each of the individual $n$ total trajectories are spawned. Each of these nodes are expanded into a series of other child nodes, in which each node corresponds to a set of possible trajectory combinations. For example, starting from the first node of the first branch $O_{1,1}$ containing trajectory $o_1$, this is expanded

---

\(^5\) Depth-first search is a tree traversal algorithm that starts at the root node and visits all nodes within the tree by exploring branches sequentially. This involves visiting the first child of the current node, and continuing down that branch until there are no more unvisited children or a node with no children is reached. At this point, one traces back to the last node with an additional unvisited child node, corresponding to another branch of the tree to be searched.
Figure 2.5: Two sets of 4 2D trajectories $O = [o_1, o_2, o_3, o_4]$ over a set of times $T = [t_1, t_2, t_3]$. Each time-slice represents a different spatial clustering. The grey circle with radius $\varepsilon$ represents the size of each cluster. The top panel shows a moving cluster when $\theta_{\text{min}} = 0.75$, each successive spatial cluster shares at least 3 out of 4 member trajectories. The bottom panel shows a convoy of two trajectories $o_1$ and $o_2$ (red) over 3 time steps.

by sequentially adding trajectories until the node $O_{1,n} = [o_1, o_2, \ldots, o_n]$ is reached. For each node the algorithm identifies and stores the set of times $T$, over which the contained combination of trajectories are clustered together. If at any point a node has a set of times $T$ smaller than the specified minimum number of time steps, its expansion is truncated and the depth-first search continues with the next branch of the tree. A series of additional pruning rules eliminate branches of the tree that correspond to repeated combinations and that are unlikely to contain swarms that are closed, limiting the size of the search space. A swarm is considered closed if its set of times $T$ is greater than the minimum number of time steps, and providing $T$ can not be expanded by adding another additional trajectory to its set i.e. there exists no supersets of the trajectory set that are clustered together for a longer set of times. In the end, the algorithm outputs a set of closed swarms corresponding to the maximal combinations of trajectories that are clustered together for a minimum num-
ber of (potentially non-sequential) time steps. While this overcomes the requirement of time steps being consecutive, the temporal evolution of spatial coordinates is not considered directly and there is no guarantee the swarms are continuous in time. For example, one may identify a swarm in which the group of trajectories are clustered together at a few beginning and end time steps only. In addition, the algorithm scales as $O(2^{n \times t})$, where $n$ and $t$ are the number of trajectories and time steps respectively. Such scaling quickly becomes prohibitive.

In subsequent chapters we shall explore the development of a new trajectory pattern mining algorithm, capable of identifying spatio-temporal clusters in QMD trajectories. The new algorithm does not suffer the consecutive time step issue observed in the case of flocks, moving clusters, and convoys, making it possible to account for short term deviations in the trajectories. Moreover, it is loosely based on the idea of a moving cluster and considers not only the times at which trajectories are clustered together, but also the explicit evolution of the spatial coordinates in time. Thus, it is better suited to trajectory simulations than the swarms algorithm.
Chapter 3

Time Series Clustering of Trajectories

3.1 Introduction

QMD simulations that approximate the evolution of the nuclear wavepacket as a series of trajectories can be difficult to interpret. While the collective ensemble of trajectories defines the complete model and allows for a side-by-side comparison to experimental observables, we often wish to identify the key set of reaction pathways that contribute to the observable. In cases where there are many redundant degrees of freedom or when previous experimental studies provide insight into key aspects of the dynamics, differences in 1 or 2D aspects of the nuclear geometry can usually be identified by eye. However, for separation of trajectory data in higher dimensions, or for systems that are less well understood, this can be challenging. Clustering algorithms are capable of automatically identifying the important trends in complex data, and compared to simple separation of nuclear geometries by inspection, can provide a more complete overview of the multidimensional dynamics. Such algorithms are useful tool for identifying the main decay or reaction pathways available to an excited system, with clusters of similar trajectories representing concentrations of reaction flux. This aids mechanistic understanding of simulations, helps the interpretation of experimental observations, and allows for cheaper calculation of observables. The complexity and computational cost for calculating accurate observables from simulations explains the persistence of simple approximations such as the IAM [130] or Dyson norm calculations [162] for scattering and photoelectron spectroscopy, respectively. Highly accurate calculations of total scattering signals from ab-initio wavefunctions [87] and photoelectron spectra using sophisticated R-matrix codes [163], are computationally expensive, often being limited to small molecules, or a small subset trajectories. By automatically identifying the most significant reaction pathways, and representative trajectories for each pathway, clustering provides a reduced sample of trajectories on which to perform these highly accurate calculations. However, the clustering of
3.1 Introduction

trajectories that approximate the quantum dynamics of a system is not without complications. This includes the high dimensionality of the problem, the potential of there existing overlapping concentrations of trajectories with highly variable density, and subtle differences in the temporal structure of similar trajectories i.e. small variations in vibrational frequencies and the speed at which they follow a given pathway.

In this chapter we explore the time series treatment of trajectory clustering, which was published in the *Journal of Chemical Theory and Computation* as "Automatic Clustering of Excited-State Trajectories: Application to Photoexcited Dynamics" [164]. We employ two different multidimensional distance measures that provide a measure of how similar a pair of trajectories are over all times. We investigate the application of the existing DBSCAN algorithm to a set of trajectories describing the electrocyclic ring-opening reaction of 1,3-cyclohexadiene (CHD) to form 1,3,5-hexatriene (HT).

First we discuss how to select a reduced number of internal coordinates on which we perform the clustering. The identification of a reduced number of features that capture the key underlying trends in the data is important, and essential for good performance. We then explore the use of two different multidimensional distance measures, which are used to construct a similarity matrix that provides a single-valued measure of how alike every pair of trajectories are over all times. The resulting matrix of pairwise distances serves as the input for the DBSCAN algorithm [145], which identifies a number of clusters based on specified parameters. Following publication, it was pointed out that the similarity measures originally employed do not explicitly account for the periodicity of angular coordinates. A discussion of this issue is provided towards the end of the chapter, where we propose two alternative treatments of the angular coordinates. Ultimately, we find that in the application to CHD considered here, the inability of the original similarity measures to account for periodicity has no significant effect on the clustering results. However, we note that this will not always be the case, and best practice would be to ensure any similarity measure accounts for periodicity.

The results help us identify limitations to the time series approach when applied to quantum molecular dynamics trajectories, setting the scene for next chapter which presents a new, alternative, approach. The work in the current chapter can thus be considered a first attempt at clustering trajectories, based on internal coordinates alone and using an established and widely applicable algorithm. This provides a
3.1 Introduction

foundation for the exploration of a pattern mining approach to trajectory clustering in the subsequent chapter, and leads to the development of a new spatial density clustering algorithm and an additional pattern mining algorithm that identifies spatio-temporal clusters in QMD trajectories.

3.2 Model Trajectory Data

The ring-opening of 1,3-cyclohexadiene (CHD) to 1,3,5-hexatriene (HT), depicted in Fig. 3.1, is a prototypical Woodward-Hoffmann electrocyclic reaction [165], that has been subject of an extensive list of theoretical and experimental studies. [22, 166–187] To test the application of clustering algorithms to trajectory data, we utilise an ensemble of 100 semi-classical AIMCE trajectories obtained from Ref. 174. The trajectories are propagated at the 3SA-CASSCF(6,4)/cc-pVDZ level of theory for up to 150 fs. In the adiabatic basis in which the simulations are performed, only the first two singlet states are formally involved. However, the $S_1$ state with initial 1B character gains 2A character from the adiabatic $S_2$ state as the reaction progresses, thus the inclusion of the first three singlet states is crucial to describing the dynamics.

![Figure 3.1](image)

**Figure 3.1:** The photo-induced electrocyclic ring-opening of CHD to HT. Here the states shown are diabatic, with red points indicating conical intersections between them. Following formation of cZc-HT on the ground state, several cis/ trans HT isomers are accessible.
3.2 Model Trajectory Data

The wavepacket evolution proceeds on the 1B electronic state, following a $\pi \rightarrow \pi^*$ transition initiated by absorption of a 267 nm (4.64 eV) photon. The acceleration of the wavepacket out of the Franck-Condon region is fast, with the conrotatory twist of the terminal carbon atoms already taking place before the 2A/1B conical intersection is reached [169, 181], and a strong overall ballistic character to the dynamics. The lifetime of the 1B state is somewhat contested. Many experimental studies report that the wavepacket is on the 2A state by 50-70 fs [166, 169, 172, 181], but recent studies suggest the decay of population occurs within 30 fs [173].

Turning to Fig. 3.2, we observe the populations for the trajectory ensemble used here. Following excitation to the $S_1$ (1B) state, Fig. 3.2 shows a rapid decrease and increase in the $S_1$ and $S_2$ populations starting at around 20 fs, corresponding to the wavepacket reaching the 2A/1B conical intersection. There is a clear maxima in the $S_2$ population at around 30 fs, This agrees with a recent surface-hopping study at the XMS-CAS(6,6)-PT2/cc-pVDZ level of theory [186] and indicates that the lifetime of the 1B state is approximately 30 fs. At the 2A/1B conical intersection the wavepacket bypasses a cusp between the surfaces, forcing it to break symmetry via a set of symmetry equivalent 2A/1A conical intersections [181–184].

Furthermore, calculations support the existence on an extended seam of intersection points within the 2A/1A intersection space that is accessible to the wavepacket [181]. The rapid acceleration from the slope of the cusp region and the extended intersection seam allows the wavepacket to decay to the ground state rapidly. In Fig. 3.2, we observe the onset in $S_0$ population rise at 25 fs. The $S_0$ population continues to rise as the $S_1$ and $S_2$ states stabilise at around 70 fs, by which time a large portion of the wavepacket has reached the $S_0$ ground state. The lifetime of the 2A state has been reported in the range of 60-80 fs [169, 170, 181]. The majority of experimental and theoretical studies estimate the total excited state lifetime to be on the order of 130-140 fs [170, 173, 178, 181]. It is thought that on first pass of the conical intersections, the rapid acceleration of the wavepacket does not allow it to disperse on the 2A state, resulting in a localised portion passing directly through the conical intersection [182]. Following nonadiabatic population transfer to the ground state, the wavepacket can then return to vibrationally hot CHD, or form cZc-HT. At times greater than 70 fs, Fig. 3.2 shows the populations stabilise with significant $S_1$ population. At these times the $S_1$ state is of 2A character, and the remainder of the wavepacket undergoes several more oscillations along the near flat 2A surface, causing it to disperse. This dispersion results in a broader distribution of geometries which have a higher C$_1$–C$_6$ torsion angle. It has been theorized that subsequent passes of the conical intersection
3.2 Model Trajectory Data

...seam by this disperse wavepacket, may result in a small amount of more open E-HT isomers being formed on the ground state [185]. Since the formation of cZc-HT occurs with an energy in excess of the barrier to isomerisation on the ground state, a range of cZc, cZt and tZt HT isomers are possible. This isomerisation takes several ps to reach thermal equilibrium, however studies have observed the presence of tZt-HT after several hundred femtoseconds [178]. Generally the quantum yield of HT has been reported to lie in the range of 40-60% [167, 170, 172, 186].

Figure 3.2: Populations for the set of 100 AIMCE trajectories over the 150 fs simulation period. At early times, the $S_1$ state corresponds to the diabatic 1B state. At 30 fs a maxima in population of the $S_2$ state is observed as the wavepacket passes through the 2A/1B intersection, at subsequent times, the $S_1$ state is of 2A character. Note that in AIMCE a single Gaussian basis function is shared over the three states, thus trajectories with dominant $S_0$ population at later times still share some population with the excited states.
3.3 Methodology

3.3.1 Feature Selection

For a general data set defined by a number of observations (data points) as a function of numerous variables, the performance of clustering algorithms highly depends on the construction of a reduced-dimensionality feature space that captures the underlying variance in the data set. This is due to the poor performance of Euclidean distance metrics in higher-dimensional spaces, the increased sampling requirement due to the exponential scaling of the space the data occupies with the number of dimensions, biasing of the result from including highly correlated features, and overfitting from including irrelevant features [135, 188]. In addition to these issues, there is an associated increase in computational cost with the inclusion of more features. In the specific case of trajectories that model the quantum dynamics of a given system, the evolution of the wavepacket is defined by 3N-5 or 3N-6 nuclear degrees of freedom, as well as electronic populations, velocities, and momenta. Thus, the selection of a reduced number of features that define the dynamics can be challenging. In some cases, it may only be possible to resolve one or two aspects of the multimodal dynamics e.g. nuclear geometries and/or populations. The difficulty in identifying nuclear coordinates that represent the motion of the nuclei is of course, system dependent. When the nuclear dynamics are defined by a small number of normal modes, the selection may be trivial, and for example correspond to, a small set of bond lengths, angles, or simply interatomic distances. In other cases, one may be forced to define a different distance measure that captures the displacements of several nuclei at once.

While there exists a range of algorithms aimed at dimensionality reduction and feature selection, e.g. principal component analysis (PCA) [137] and tICA [138], we opt for a more straightforward approach in which we select the features that correspond to the highest variance over time. In the case of CHD, we are mostly concerned with the large amplitude ‘chemical’ motions of the heavier carbon atoms, thus we do not consider the faster and weakly correlated motion of the hydrogen atoms in feature selection. In Fig 3.4, we observe the maximum variance over time for a set of internal coordinates. Apart from the $R_{C1-C6}$ bond that is broken, all other C–C bond lengths have negligible variance. The same can be said for the angles, with the exception of $\theta_a$ and $\theta_b$ included in table 3.1. However, these angles are highly correlated with the dihedral angles that describe the out-of-plane motion of the molecule. The variance of the four dihedral angles suggests that the molecule exhibits a large degree of torsion.
3.3 Methodology

out of the plane, and thus some combination of them should be selected. For this reason we select $\phi_b$, as well as the two dihedrals $\phi_a$ and $\phi_c$ which capture the ground state isomerisation to trans-HT products. Given that these three dihedrals do not allow distinction between cZc-HT and hot CHD, we also include the $R_{C1-C6}$ bond. This set of four internal coordinates forms the basis of the feature spaces in which clustering is performed in the subsequent sections of this chapter.

![Diagram](image)

(a) High variance internal coordinates.

Table 3.1: The maximum variance over all time for a set of internal coordinates.

<table>
<thead>
<tr>
<th>Internal coordinate</th>
<th>Maximum Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>$R_{C1-C6}$</td>
<td>1.973 Å²</td>
</tr>
<tr>
<td>$\theta_a$</td>
<td>0.105</td>
</tr>
<tr>
<td>$\theta_b$</td>
<td>0.115</td>
</tr>
<tr>
<td>$\phi_a$</td>
<td>0.239</td>
</tr>
<tr>
<td>$\phi_b$</td>
<td>0.574</td>
</tr>
<tr>
<td>$\phi_c$</td>
<td>0.248</td>
</tr>
<tr>
<td>$\phi_d$</td>
<td>0.455</td>
</tr>
</tbody>
</table>

Figure 3.4: Definitions of high variance internal coordinates and their corresponding maximum values over all time steps of the simulation. The variance of each angular coordinate corresponds to the circular variance, calculated as $\var = 1 - \bar{R}$, where $\bar{R}$ is the mean resultant vector of the angles represented by their coordinates $\cos, \sin$. Thus, the variance of each angular feature lies on the range $0 \leq \var \leq 1$. The selected internal coordinates that form the basis of the feature space are, the dihedral angles $\phi_a$, $\phi_b$, and $\phi_c$, as well as the bond length $R_{C1-C6}$.

The distribution of the trajectories over all time steps can be seen in Fig. 3.5, with each axis representing one angle and the colour bar the $C_1-C_6$ bond. Furthermore, in Fig. 3.6, we see the distribution of trajectories over four different time intervals. In Fig. 3.6a, we can see that the trajectories occupy a small volume of space as they are all localised within the Franck-Condon region following excitation. Turning to panel b, the trajectories notably start to disperse along all four internal coordinates. At this point, the trajectories begin to separate, with some reaching $C_1-C_6$ bond lengths on
the order of 4 Å. In the bottom two panels (c and d), we observe the emergence of four key groups of trajectories. The most easily identifiable of these being those that correspond to vibrationally hot CHD, with small angular coordinates and bond lengths below 2 Å. In addition, one can identify two groups of trajectories that observe large dihedral angles and \( C_1 - C_6 \) bond lengths on the order of 5-6 Å at opposing ends of the hypercube. Finally, there remains a broad distribution of trajectories approximately centred around \( \phi_c = 100^\circ \) with more intermediate \( C_1 - C_6 \) bond lengths between 3-5 Å. We note that one would expect these to converge towards one of the other three aforementioned groups in the limit of a longer simulation in which the excited trajectories fully decay to the ground state.

Figure 3.5: Distribution of the three dihedral angles with colour representing the \( R_{C1-C6} \) bond length. Shown over all time steps of the simulation.

It is worth noting that in the current example, the end products are known from previous experimental and theoretical studies, allowing us to employ chemical intuition for what features to include. However, in cases where the end products are less certain, one could envision employing a spatial clustering algorithm solely at the end time step of the simulation to identify the distribution of different product geometries. With the time-dependence removed, the conformations of different product geometries can be identified using common spatial clustering algorithms in conjunction with PCA. Such approaches are commonly employed in the clustering of protein conformations in mo-
3.3 Methodology

Figure 3.6: Distribution of the data over several time intervals of the simulation.

Molecular dynamics simulations, where no explicit time-dependence is usually included [189–193]. Once the distribution of different products is known, this information can be used to provide some constraints on the selection of features that are used in temporal clustering.

3.3.2 Similarity Measures

In the time series model of trajectories we employ here, we construct a similarity matrix of dimensions \( N_{\text{traj}} \times N_{\text{traj}} \), with \( N_{\text{traj}} \) the total number of trajectories. Each element of this matrix corresponds to a scalar value of the similarity between a pair of trajectories over all \( N_t \) total time steps. Of course, to avoid redundant calculations and excessive memory usage, one stores the strictly upper triangular form of the similarity matrix. Given the wide span of data types on which time series clustering is performed, there are a number of different measures of similarity. For example, longest
common subsequence (LCSS), dynamic time warping (DTW), and the $L_p$ norm [194, 195]. All of these have multidimensional generalisations. The LCSS distance involves identifying the longest subsequence that is common to two time series, the distance is then given by the difference between the length the two time series considered and the length of the common subsequence. However, since LCSS does not consider the order of points in time, we do not consider it in our application. The $L_p$ norm considers the distance between each point in the time series, and involves a 1:1 mapping of time points. The multidimensional generalisation of the $L_p$ norm for two trajectories $o_i$ and $o_j$ is defined as,

$$L_p - \text{norm} (o_i, o_j) = \left( \sum_{t} \sum_{f} (o_{it} - o_{jt})^p \right)^{\frac{1}{p}}, \quad (3.1)$$

with $N_t$ the total number of time steps, and $N_f$ the total numbers of features included. Note, the case where $p = 2$ corresponds to the $L_2$ norm, i.e. the generalisation of the Euclidean distance, which we shall consider in our future clustering application. The multidimensional $L_p$ norm requires that both trajectories are the same length in time, and therefore might not be applicable to systems which include dissociation. Moreover, it has been proven to generally underperform in higher dimensions [196, 197]. Due to the 1:1 mapping of time points, $L_p$ norm cannot account for small variations within the local time axis. For example, two trajectories may be highly similar but offset in time slightly, or exhibit a small local variation in frequency. In some cases this may result in the misclassification of trajectories as the full extent of their similarity may not be captured.

Some of these shortcomings are overcome by dynamic time-warping (DTW), [198, 199] which allows for some stretching or contraction of the time axis. DTW warps the trajectories non-linearly in time and determines the minimum warping path required for alignment subject to a set of constraints. This allows for the potential of a 1:many mapping of temporal indices, meaning one can compare two time series that are similar but observe local variations in frequency and time offset. In addition, this makes it possible to compare time series data of differing lengths, unlike the $L_p$ norm.

In DTW, the warping path describes how a pair of trajectories are warped in time so that they are optimally aligned subject to certain conditions. Therefore, the warping path describes a mapping of temporal indices that minimises the cost of alignment. To demonstrate how the warping path is calculated, let us consider two trajectories $o_1$
and $o_2$, with time sets $T_1 = [t_1, t_2, \ldots, t_n]$ and $T_2 = [t_1, t_2, \ldots, t_m]$ respectively. Furthermore, one does not necessarily require $n = m$. Next, we construct a cost matrix $D$ of dimensions $n+1 \times m+1$, in which the calculation of the matrix elements are treated as a dynamic programming problem. The cost matrix contains all possible paths that can be taken to warp the two trajectories in time, with each element comparing two temporal indices. The first row $D(0, 1 : m)$ and column $D(1 : n, 0)$ are initialised as infinity, and $D(0, 0)$ is set to zero. This ensures that the first time step of both time series are always matched together, and is shown for a simple working example on the right hand side of Fig. 3.7. The elements of the matrix are then calculated as,

\[
D(i, j) = |o_{1i} - o_{2j}| + \min[D(i - 1, j - 1), D(i - 1, j), D(i, j - 1)],
\]

(3.2)
with $i$ and $j$ the time indices of each trajectory, running to $i \rightarrow n$ and $j \rightarrow m$ respectively. Here, the first term is the norm between trajectories $o_1$ and $o_2$ at the respective temporal indices. The second term is referred to as a stepping pattern, corresponding to the minimum value of previously calculated elements. Once the matrix elements have been calculated, the warping path is determined by starting at element $D(n,m)$ and traversing back through the matrix to $D(0,0)$, indicated by the red arrow at element $D(n,m)$ in Fig. 3.7. The traversal proceeds by selecting the minimum value of elements $D(n-1,m-1)$, $D(n-1,m)$, $D(n,m-1)$ as the next element to be selected in the traversal. Taking the set of three elements in order, these correspond to diagonal, horizontal, and vertical steps through the cost matrix. Taking a diagonal step allows the warping path to bend in time, accounting for temporal distortions such as one trajectory being shifted slightly in time, indicated by both the green and yellow matrix elements in Fig. 3.7. Here green elements result in a warping of the temporal axis as $i \neq j$, whereas yellow corresponds to the case where the indices matched are at the same point in time i.e. $i = j$. In contrast, horizontal and vertical steps allow one trajectory to skip or repeat a temporal index, respectively. These steps either speed a trajectory up or slow it down, allowing the potential for a 1:many mapping of time points. In Fig. 3.7 vertical steps are indicated in magenta. This provides a way to account for changes in local frequency between a given pair of trajectories. After a step is taken, the direction of the movement is recorded and the procedure is repeated until element $D(0,0)$ is reached. The path taken through the cost matrix defines the warping path, with the sum of all its elements the total cost of alignment. This total cost corresponds to the measure of similarity between the two trajectories, with lower values indicating they are more similar. For the worked example in Fig. 3.7, the alignment of the two time series defined by the warping path is shown in the lower left hand portion of the figure, with the points coloured according to the type of step taken through the cost matrix.

While the example in Fig. 3.7 shows the univariate case, DTW can be extended to handle multidimensional time series [199, 200]. Generally there are two approaches, the simplest involves independent application of the univariate DTW algorithm to each dimension, the total cost is then the sum of the costs of each independent dimension. Alternatively, one can employ full multidimensional DTW (MD-DTW), which ensures all dimensions warp together by construction of a higher order cost tensor. The warping path is calculated in the same fashion, except now one has to trace through a higher order tensor instead of a matrix. As a result, the computational complexity increases due to the associated increase in the number of possible warping paths.
with the number of dimensions. It has been shown that the case where independent
treatment of the dimensions is assumed can lead to wrong classification, and really
one should use full MD-DTW where possible [200]. For this reason, we only consider
the use of full MD-DTW in this thesis.

It is worth noting that DTW was originally developed for applications in speech re-
cognition [198]. Since then, DTW has been employed in a wide range of fields, being
applied to many different types of time series data [78, 152, 153, 201–203]. In some
applications one may not desire frequency invariance or the ability to warp the tem-
poral axis to account for offsets in time. It is possible to tune the degree of this by
modifying the stepping pattern given in Eqn. 3.2. For example, one may not consider
horizontal steps at all, or limit the number of successive horizontal and vertical steps
taken. Moreover, one can bias against certain steps by adding an additional penalty
cost to them. In the context of quantum molecular dynamics, we may wish to limit
the flexibility of the warping with respect to frequency, since molecular vibrations of
distinct frequency generally correspond to distinct motions. Moreover, one might want
to account for the slight offset of two trajectories in time as they follow the same path
through phase space, just separated slightly in time. However, a large degree of this
could render different time scales undetectable and indistinguishable.

We conclude this section by noting that the the trajectory data must undergo normal-
alisation prior to construction of a similarity matrix and subsequently, clustering. This
is due to the fact the the numerical magnitude of the features may vary greatly. For
example, in our application, the dihedral angles span a significantly larger range of
values than the bond lengths. Through normalisation, one limits the risk that features
which span a larger range dominate the clustering results. Here we utilise a min-max
normalisation, in which we map each feature to the range \([0, 1]\) as,

\[
M : o_{if} \rightarrow \frac{o_{if} - \min(O_f)}{\max(O_f) - \min(O_f)},
\]

where \(o_{if}\) refers to the trajectory \(o_i\) along the feature dimension \(f\), and \(o_{if} \in O_f\)
forms the set of all trajectories across feature \(f\). Once each of the identified features
are normalised appropriately, a similarity matrix is constructed from one of the afore-
mentioned methods, this is then provided as input to the clustering algorithm.
3.3 Methodology

3.3.3 Cluster Validity

Following the discussion of the distribution of the trajectories in the feature space in section 3.3.1, we expect the identification of three of four clusters. We assess the validity of the resulting $N_C$ clusters by the percentage of trajectories classified as noise ($N_{\text{noise}}$)\(^1\), and the silhouette coefficient ($S_i$), which defines a statistical measure of how well a trajectory is assigned to its respective cluster $C_I$. The silhouette coefficient for a trajectory $i$ is defined as \[^{[204]}\]

$$S_i = \frac{b_i - a_i}{\max\{a_i, b_i\}}$$

(3.4)

where $a_i$ is the mean distance between trajectory $i$ and all other trajectories within its cluster $C_I$,

$$a_i = \frac{1}{|C_I| - 1} \sum_{j \in C_I, i \neq j} D(i, j),$$

(3.5)

where $|C_I|$ is the number of members in cluster $I$, and $b_i$ the minimum distance of trajectory $i$ to all other trajectories in its nearest-neighbour cluster $C_J$,

$$b_i = \min \frac{1}{|C_J|} \sum_{j \in C_J} D(i, j).$$

(3.6)

The values of $a_i$ and $b_i$ define how well trajectory $i$ is assigned to cluster $C_I$, and how dissimilar $i$ is to its nearest neighbouring cluster. If the value of $a_i$ is small, the trajectory is well assigned to cluster $C_I$, and if $b_i$ is small, this indicates that the trajectory is similar to trajectories in its nearest neighbouring cluster. Together $a_i$ and $b_i$ define the silhouette coefficient $S_i$, whose values are in the range $S_i \in [-1, 1]$. The closer $S_i$ is 1, the more appropriately the trajectory is clustered. However, if $S_i < 0$, then trajectory $i$ would be more appropriately assigned to its nearest neighbouring cluster. The average value of the silhouette coefficient over all clustered data points ($S_{\text{avg}}$), provides a global measure that can be used to assess the validity of the clustering of all trajectories.

\(^1\) We aim to classify no more than roughly 10\% (i.e. 10/100 trajectories) as noise.
3.4 Results and Discussion

We shall now turn to a discussion of the results obtained by the clustering of the 100 CHD trajectories discussed previously, using the four identified features, i.e. the three dihedral angles ($\phi_a, \phi_b, \phi_c$), and the $R_{C1-C6}$ bond length) shown in Fig. 5.5. The similarity of every trajectory pair is calculated using both the $L_2$ norm and MD-DTW. The resulting similarity matrix is then provided as a precalculated input into the DBSCAN algorithm\(^2\). In each case, we scan a suitable range of possible $\epsilon$ and $MinPts$ values and analyse the clustering result.

3.4.1 The Multidimensional $L_2$ norm

With the $L_2$ norm as a similarity measure, the parameters $\epsilon$ and $MinPts$ are scanned in the ranges $\epsilon \in [1, 1.5]$ and $MinPts \in [3, 7]$ respectively. The result of clustering for each pair of parameters can be seen in Fig. 3.8, which shows the number of clusters $N_C$, the average value of the silhouette coefficient $S_{avg}$ and the percentage of trajectories marked as noise $N_{noise}$. Generally, the highest values of $S_{avg}$ result in either $N_C = 3$ or $N_C = 4$ clusters being identified. Note, for many combinations of $\epsilon$ and $MinPts$, there is little variation in the value of $S_{avg}$, making the identification of the ideal number of clusters difficult. The best combinations of parameter values are indicated by the circled points and table in Fig. 3.8. Generally, the region corresponding to $\epsilon > 1.3$ and $MinPts < 6$ yields the smallest values of $S_{avg}$. Here, the definition of density neighbourhood is too large and diffuse to provide any meaningful clustering, and all trajectories are grouped into $N_C \leq 2$ broad clusters. However, increasing the value of $MinPts$ to $\geq 6$ means core points must have a higher local density and thus can result in the identification of $N_C = 3$ clusters. In this region, the best result corresponds to $(\epsilon, MinPts) = (1.35, 7)$, indicated by the orange circled points. As highlighted in the table of Fig. 3.8, this yields $N_{noise} = 5$ and $S_{avg} = 0.386$. Moreover, the intermediate range of $\epsilon$ where $1.175 \leq \epsilon \leq 1.3$ yields $N_C = 3$ with the highest value of $S_{avg}$ across all values of $MinPts$, indicating optimal cluster separability. Note that here the number of trajectories marked as noise increases above 10 for high values of $MinPts$. The best $N_C = 3$ case defined by $(\epsilon, MinPts) = (1.2, 4)$, is indicated by the blue points in each panel of Fig. 3.8. In this case, 7 trajectories are marked as noise, with a slightly higher value of $S_{avg} = 0.389$ in comparison to $(\epsilon, MinPts) = (1.35, 7)$. Finally, when $\epsilon \leq 1.15$, the density neighbourhood is small enough to identify $N_C = 4$ clusters.

---

2. Recall DBSCAN depends on two parameters, $\epsilon$ and $MinPts$
Figure 3.8: Results of clustering using the $L_2$ norm as a measure of trajectory similarity. From the top left, the heatmaps correspond to, the average silhouette coefficient $S_{avg}$, the number of resulting clusters $N_C$, the number of trajectories marked as noise $N_{noise}$. Note, while $S_{avg}$ lies on the range $-1 \leq S_{avg} \leq 1$, the heatmap is scaled to highlight the differences in otherwise small variations. The table corresponds to the better clustering results for $N_C = 3, 4$, which are highlighted by the coloured ellipses in each heatmap.

In this region, as many as 15-20 trajectories are marked as noise. However, the lower value of $MinPts = 3$ generally observes reasonable levels of noise. Here, when $(\varepsilon, MinPts) = (1.15, 3)$, $N_C = 4$ clusters are identified with as few as 7 trajectories marked as noise. This yields $S_{avg} = 0.37$, which is slightly lower than the $N_C = 3$ region. This point is indicated in green in Fig. 3.8.
To better understand the clustering results, we turn to Fig. 3.9, which shows the values of the individual silhouette coefficients $S_i$ for each trajectory in its respective cluster $C_i$. Looking first at the $(\varepsilon, MinPts) = (1.15, 3)$ case in the left hand panel, we notice the $S_i$ values of $C_1$, $C_3$, and $C_4$ are all above the average of $S_{avg} = 0.37$, indicating that the trajectories therein are well clustered. However, the broad shaded area corresponding to $C_2$ indicates the existence of a large cluster that contains some trajectories for which $S_i < S_{avg}$. In addition, $C_2$ contains a number of trajectories for which $S_i < 0$, meaning that they are unfavourably clustered and would be better assigned to the nearest neighbouring cluster to $C_2$. Turning to the $(\varepsilon, MinPts) = (1.2, 4)$ case in the right panel of Fig. 3.9, we see that $C_4$ disappears and $C_2$ becomes larger as it absorbs these trajectories. Note the reduction in the amount of trajectories with $S_i < 0$. This, and the slight increase in $S_{avg}$ from 0.37 to 0.389, indicates that $(\varepsilon, MinPts) = (1.2, 4)$ results in a marginally statistically improved clustering. The set of potential reaction products identified by each of the cluster centres at the end of time, can be seen in Fig. 3.10. In the case where $N_c = 3$ clusters are identified, these correspond to vibrationally hot CHD ($C_1$), a set of structures that exhibit larger torsional angles ($C_2$), and tZt-HT ($C_3$), which forms from the isomerisation of cZc-HT. The $N_c = 4$ case results in the identification of an additional product ($C_4$), corresponding to cEc-HT.
3.4 Results and Discussion

Figure 3.10: The set of four product channels identified by clustering. Shown are the final geometries at each of the cluster centres. Here $C_1$ and $C_3$ define the return of the wavepacket to hot CHD, and the ground state $tZc/tZt$ HT product, respectively. At the end of time, $C_2$ likely corresponds to part of the wavepacket that spreads out on the excited state and exhibits larger torsional angles. Finally, when identified, $C_4$ corresponds to a cEc-HT product.

Having identified the $N_c = 3$ case with $(\varepsilon, \text{MinPts}) = (1.2, 4)$ as the statistically better result, although marginally so, we now turn to discuss the distribution and evolution of the clusters in time. The resulting clusters $C_1$ through to $C_3$ can be seen in Fig. 3.11, where each panel refers to one of the four features on which clustering was performed. Before approximately 40 fs, all clusters exhibit a similar evolution, with the CHD ring structure being strained upon excitation. The strained geometry is indicated by the increase in the dihedral angles $\phi_a$ and $\phi_b$, as well as the $C_1-C_6$ bond length. The straining of the ring system continues, with oscillations in $\phi_a$ and $\phi_b$, and a simultaneous sharp increase in the dihedral $\phi_b$. Following this, the clusters centres begin to separate more clearly as the trajectories within them disperse, evolving along a series of different reaction pathways. Here it is clear to see that cluster $C_1$ returns back to vibrationally hot CHD after several oscillations of the $C_1-C_6$ distance, which dampen in time. Turning to cluster $C_3$, we observe the $C_1-C_6$ bond length increase at a faster rate than the other clusters. This likely occurs as cZc-HT forms on the ground state. Towards times exceeding 100 fs, $C_3$ begins to observe an increase in
both the $\phi_a$ and $\phi_b$ dihedral angles, commensurate with the isomerisation of cZc-HT to tZt-HT. Identifying the pathway that cluster $C_2$ corresponds to is less clear, with the cluster exhibiting significant broadening in all four features. This broadening is most significant for the $C_1-C_6$ distance and $\phi_b$ dihedral, which span almost the whole range of available distances and angles. This is likely the result of a very disperse wavepacket limiting the separability of the data, at least based on nuclear geometries alone. However, we do note that the cluster is centred on values of $\phi_b$ around 100 degrees, and $\phi_a/\phi_c$ around zero degrees. Thus, it is likely the densest part of $C_3$ corresponds to part of the excited wavepacket that experiences dispersion along the $\phi_b$ coordinate, resulting in a greater degree of torsion as the trajectories remain trapped on the $S_1$ state that exhibits small gradients.

![Figure 3.11](image)

**Figure 3.11**: Plots of the cluster centres in each of the 4 selected features for $\epsilon = 1.2$ and $\text{MinPts} = 4$ (using the $L_2$ norm). The top left and right panels show the $C_1-C_6$ bond and the central dihedral angle $\phi_b$, whereas the bottom left and right panels show the dihedral angles $\phi_a$ and $\phi_c$. The shaded areas represent the size of the cluster.

Turning to the $N_c = 4$ case with $(\epsilon, \text{MinPts}) = (1.15, 3)$, plotted in Fig. 3.12, we observe the same general trends as Fig. 3.11. However, we see that some trajectories are redistributed from the broad $C_2$ cluster to a new cluster, $C_4$. While the dihedral angles $\phi_b$ still span a large range, this is somewhat reduced. Recall that here cluster
$C_2$ contains a greater number of trajectories with negative values of the Silhouette coefficient $S_i$, previously addressed in Fig. 3.9. It is therefore likely that the very broad nature of the wavepacket here results in clusters that have a degree of overlap. Thus, some trajectories within $C_2$ may be misclassified. The additional cluster $C_4$ observes very small values of the dihedral angles $\phi_a/\phi_c$. In addition, the sharper increase in the $C_1-C_6$ bond length, and the fact the dihedral angle $\phi_b$ exceeds 200 degrees, suggests this cluster is commensurate with the cEc-HT product channel. Here we note that both $\phi_a$ and $\phi_c$ exhibit strong oscillations before the central dihedral $\phi_b$ rapidly increases, suggesting that the initial ring strain transfers energy to the dihedral $\phi_b$.

![Figure 3.12: Plots of the cluster centres in each of the 4 selected features for $\varepsilon = 1.15$ and $\text{MinPts} = 3$ (using the $L_2$ norm). The top left and right panels show the $C_1-C_6$ bond and the central dihedral angle $\phi_b$, whereas the bottom left and right panels show the dihedral angles $\phi_a$ and $\phi_c$. The shaded areas represent the size of the cluster.](image-url)

We now shift our focus to the distribution of populations observed within each identified cluster. Although populations are not included in the clustering process here, they provide additional insight into the nature of each identified cluster. In Fig. 3.13 we see the populations of the $S_1$ and $S_0$ state plotted for every trajectory in each cluster at 150 fs. The dashed line $x = y$ is drawn to indicate that trajectories below
and above this line are mostly $S_0$ and $S_1$ in character, respectively. Moreover, the left and right hand panels show the $N_c = 3$ and $N_c = 4$ clustering results respectively. In both panels, we see that clusters $C_1$ and $C_3$ are mostly $S_0$ in character, confirming they correspond to the return of the wavepacket to ground state CHD, and the ground state HT product channel that isomerise from cZc-HT to cZt- and tZt-HT. However, the character of $C_2$ is again less clear. While most trajectories are dominated by $S_1$ population, there exist some with higher $S_0$ population. This supports the idea that the densest part of $C_2$ defines part of the wavepacket that remains trapped on the excited state, dispersing into a broader range of $\phi_b$ angles due to the low gradients. Although, it suggests that some of the trajectories are clustered unfavourably, a fact also supported by the presence of negative $S_1$ values. In the right hand panel, we see that in the $N_c = 4$ case, some of the $C_2$ trajectories that are on the ground state are relabelled as $C_4$, corresponding to cEc-HT.

![Figure 3.13](image)

**Figure 3.13**: Distribution of $S_0$ and $S_1$ state populations at 150 fs. The left panel shows the $N_c = 3$ resulting from $(\epsilon, \text{MinPts}) = (1.2, 4)$, and the right panel shows the $N_c = 4$ case with $(\epsilon, \text{MinPts}) = (1.15, 3)$. 
3.4 Results and Discussion

3.4.2 Multidimensional Dynamic Time Warping

In an attempt to improve on the misclassification of some $C_2$ trajectories by the $L_2$ norm, we now turn to the use of MD-DTW. Given that MD-DTW can account for small variations in offsets between trajectories and small local changes in frequency, it can provide a more accurate measure of similarity between temporal series. The idea here being that it may improve the separability of the data in comparison to the $L_2$ norm.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure3.14.png}
\caption{Results of clustering using DTW as a measure of trajectory similarity. Again, the $S_{\text{avg}}$ heatmap is scaled to highlight the differences in otherwise small variations. The table corresponds to the better clustering results for $N_C = 3, 4$, which are highlighted by the coloured ellipses in each heatmap.}
\end{figure}

<table>
<thead>
<tr>
<th>$N_C$</th>
<th>$N_{\text{noise}}$</th>
<th>$S_{\text{avg}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>4</td>
<td>11</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td>0</td>
</tr>
</tbody>
</table>
Here, the parameters $\varepsilon$ and $\text{MinPts}$ are scanned in the ranges $\varepsilon \in [6.5, 8]$ and $\text{MinPts} \in [3, 7]$ respectively. The result of clustering in terms of $N_C$, $N_{\text{noise}}$, and $S_{\text{avg}}$ can be seen in Fig. 3.14. Generally, performing DBSCAN clustering using MD-DTW yields similar results to the $L_2$ norm, where either 3 or 4 clusters are identified. In this case, the highest values of $S_{\text{avg}}$ are observed as $\text{MinPts}$ increases and $\varepsilon$ lies in the range $7.6 \leq \varepsilon \leq 8$. This results in $N_C = 3$ clusters. The highest value of $S_{\text{avg}} = 0.396$ corresponds to $(\varepsilon, \text{MinPts}) = (7.65, 6)$, indicated by the purple circled points in each panel. The intermediate range where $6.9 \leq \varepsilon \leq 7.6$ consistently results in $N_C = 3$ clusters. Higher values of $\text{MinPts}$ here result in an excessive number of trajectories being marked as noise. However, values of $\varepsilon \leq 7.7$ and $\text{MinPts} \leq 6$ consistently results in $N_C = 4$ clusters being identified. The number of trajectories excluded as noise increases with $\text{MinPts}$. In addition the value of $S_{\text{avg}}$ decreases slightly in comparison to the $N_C = 3$ results. Although, the set of values $(\varepsilon, \text{MinPts}) = (6.75, 3)$ results in $N_C = 4$ clusters being identified with lower noise levels and a reasonable value of $S_{\text{avg}} = 0.332$, as indicated by the green circled points.

Turning to Fig. 3.15, which shows for the best $N_C = 3$ and 4 cases the silhouette coefficient $S_i$ for each trajectory in its respective cluster $C_i$ (using MD-DTW). The left hand panel shows the best $N_C = 4$ case, and the right hand panel, the $N_C = 3$ case. In each case, red lines indicate the value of $S_{\text{avg}}$.

Figure 3.15: The silhouette coefficient ($S_i$) for each trajectory in its respective cluster $C_i$ (using MD-DTW). The left hand panel shows the best $N_C = 4$ case, and the right hand panel, the $N_C = 3$ case. In each case, red lines indicate the value of $S_{\text{avg}}$. 

Turning to Fig. 3.15, which shows for the best $N_C = 3$ and 4 cases the silhouette coefficient $S_i$ for each trajectory in its respective cluster, we see that the $C_4$ cluster is significantly larger in comparison to the $N_C = 4$ case in Fig. 3.9. Here, additional trajectories are acquired from the $C_2$ cluster. Furthermore, $C_2$ appears to have a smaller number of trajectories with negative $S_i$ values. For the set of parameters $(\varepsilon, \text{MinPts}) = (7.65, 6)$, which has the highest $S_{\text{avg}}$, the $C_4$ cluster is again absorbed by the $C_2$ cluster. Note that apart from a small number of trajectories, the $N_C = 3$ case
is practically indistinguishable from the corresponding $N_C = 3$ results identified by the $L_2$ norm. However, the $N_C = 4$ case as identified by MD-DTW, is unique. We shall therefore focus the rest of our discussion on this result, noting that the $C_4$ cluster is always absorbed by the $C_2$ cluster in the set of $N_C = 3$ results.

![Figure 3.16: Plots of the cluster centres in each of the 4 selected features for $\varepsilon = 6.75$ and $\text{MinPts} = 3$ (using MD-DTW). The top left and right panels show the $C_1$-$C_6$ bond and the central dihedral angle $\phi_b$, whereas the bottom left and right panels show the dihedral angles $\phi_a$ and $\phi_c$. The shaded areas represent the size of the cluster.](image)

The plot of the cluster centres for the $N_C = 4$ case is shown in Fig. 3.16. Here we observe that the $C_4$ cluster now includes a broader range of dihedral angles $\phi_b$ in comparison to the $L_2$ norm case. At the end of time, this cluster is centred at a slightly lower value of $\phi_b$ (170 degrees). However, it does still contain a range of trajectories that exceed angles $\phi_b > 200$ degrees. Also note how that included trajectories with lower values of $\phi_b$ come from the $C_2$ cluster, which is now slightly smaller than the $L_2$ norm case. The broad range of available dihedral angles $\phi_b$, makes the separation of these clusters difficult. Based on nuclear geometries alone, it is not immediately clear if the additional trajectories within $C_4$ correspond to ground state cEc-HT. Without additional information, the separation of two groups of trajectories,
3.4 Results and Discussion

amounting to ground state cEc-HT and the remainder of the excited wavepacket with larger torsional angles, may not be possible. As with the $L_2$ norm, the $N_C = 3$ case corresponds to clusters $C_2$ and $C_4$ in Fig. 3.16 merging together, forming a very broad and large cluster $C_2$, as in Fig. 3.11.

Turning to Fig. 3.17, which shows the final-time ($t = 150$ fs) distribution of populations observed within each cluster, we see that the $N_C = 3$ case (left panel) is near identical to the $L_2$ norm results in Fig. 3.13. In the right panel, we see the $N_C = 4$ result that is unique to MD-DTW. Here, we see a number of additional $C_4$ trajectories with a population above the $x = y$ line. Thus, the additional trajectories contained within the $C_4$ cluster are dominated by the $S_1$ state amplitude. These trajectories belong to the remainder of the excited wavepacket that exhibits a larger degree of torsion, and do not correspond to cEc-HT formation on the ground state. However, we note that these trajectories have a population that is near equal for the $S_0$ and $S_1$ state. Where the simulations run longer, these may completely decay to the ground state and yield the cEc-HT product. Moreover, we see that while the $C_2$ cluster is comprised mostly of trajectories with dominant $S_1$ character, it still contains a number of trajectories that are mostly $S_0$.

![Figure 3.17](image.png)

**Figure 3.17**: Distribution of $S_0$ and $S_1$ state populations at time $t = 150$ fs, using MD-DTW. Points colour coded according to their cluster assignment. The left panel shows the $N_c = 3$ case with $(\varepsilon, MinPts) = (7.65, 6)$, and the right panel the $N_c = 4$ resulting from $(\varepsilon, MinPts) = (6.75, 3)$.
3.4 Results and Discussion

3.4.3 Comments on Applicability

In both the application of the multidimensional $L_2$ norm and MD-DTW as a similarity measure, we have seen that based on nuclear geometries alone, the time series approach struggles to separate out trajectories from a broad distribution of dihedral angles. Using MD-DTW we were able to achieve a higher level of separability, although this still resulted in the misclassification of some trajectories. However, both metrics were able to identify four different reaction channels successfully, providing some basic insight into the decay pathways available to the excited system. The applicability of this approach is of course system dependent. In applications to other systems that observe a less disperse wavepacket, it may provide more fruitful results. The current approach is easy to implement and fast, and therefore may prove a useful tool for quickly identifying the key product channels. These could then be used for highly accurate observable calculations. However, there are some limitations. Generally, the clusters identified have a degree of overlap, and vary in both size and density. These issues are likely a result from both the limited ability of DBSCAN to identify clusters of varying density, a known issue, and that we apply the algorithm to a system in which the the wavepacket is quite disperse at later times. However, given the nature of the data and that this approach requires an algorithm capable of utilising a precomputed similarity matrix, DBSCAN is the best choice. Other algorithms that take a precomputed similarity matrix such as $k$-means or hierarchal clustering are generally not suitable, and in testing we found them to yield noninterpretable and nonsensical results.

The time series approach suffers from several additional issues in itself. First, the similarity measures are global over time, this means that smaller amplitude motions at earlier times are effectively washed out in systems that exhibit large amplitude motion, such as CHD. The clustering result is therefore highly biased towards later time steps, where the system settles into a distribution of different products. This is likely the cause of the similar results obtained with both the multidimensional $L_2$ norm and MD-DTW. Secondly, lets consider the scenario where an initial large swarm of trajectories decays through a conical intersection, leaving a number of trajectories behind that undergo several additional oscillations on the excited state before decaying. The straggling trajectories slowly leak through the conical intersection over a range of times, before converging to a given product. Given that the time series approach depends largely on the time at which a significant number of trajectories assume a different geometry, it is only able to capture large swarms of trajectories that decay at
the same time. The use of MD-DTW may compensate for this to a greater extent than the $L_2$ norm, due to its ability to account for offsets in time. However, as we observed here, in practice the choice of method had little effect on the clustering result. Finally, for very long simulations, the use of MD-DTW is not practical due to its scaling with the length of the temporal series. The use of other metrics such as the multidimensional $L_2$ norm is also limited, given its poor performance in high dimensional spaces.

### 3.5 Periodicity of Angular Coordinates

Following publication, it was pointed out to us that both the distance measures based on the $L_2$ norm do not directly account for the periodicity of angles. Ignoring the periodicity may lead to erroneous clustering results if the distribution of angles spans a range greater than $\pi$ radians (or equivalently, 180°). For example, the pair of angles $(0^\circ, 270^\circ)$ are only $90^\circ$ apart when accounting for the periodicity. However, the use of an $L_2$ norm as defined by Eqn. 3.1 results in the larger distance of $270^\circ$.

![Figure 3.18: Distribution of the four clusters of trajectories identified using the $L_2$ norm as in section 3.4.1. Note, the plot is with respect to the three dihedral angles and the $C_1-C_6$ bond is omitted from this plot.](image)

---

3. Specifically, we are grateful to the Examiners of the PhD thesis, Dr Ryan Coffee and Dr Antonia Mey, for raising this important issue.
In Fig. 3.18 we show the distribution of the dihedral angles for the four clusters identified using the $L_2$ norm in section 3.4.1. In this net distribution, which accounts for all times, the angles indeed span more than $180^\circ$. However, inspection of Fig. 3.19, which shows the clustered features as a function of time for each cluster, reveals that the range of dihedral angles only exceeds $180^\circ$ at later times, towards the end of the simulation. Also note that while each angle spans a range greater than $180^\circ$, the data is multidimensional and one might anticipate relatively minor effects on the clustering provided that each cluster is separable in more than one dimension, as we shall see.

![Figure 3.19: Plots of each of the four selected features as a function of time, with each trajectory coloured according to its cluster. The clusters have been obtained using the original definition of the $L_2$ norm in section 3.4.1. The top left and right panels show the $C_1-C_6$ bond and the central dihedral angle $\phi_b$, respectively, whereas the bottom left and right panels show the dihedral angles $\phi_a$ and $\phi_c$.](image)
3.5 Periodicity of Angular Coordinates

3.5.1 Unit Circle Projection

In order to account for the periodicity of angles, we evaluate two alternative approaches. First, we consider the projection of each dihedral angle \( \phi \) onto the unit circle with coordinates \((\cos \phi, \sin \phi)\). In this treatment we can simply continue to use the \(L_2\) norm as given by Eqn. 3.1. However, the dimensionality of the feature space increases to seven as each of the three dihedral angles are defined by a pair of coordinates on the unit circle. As each of the angular coordinates \((\cos \phi, \sin \phi)\) spans the range \([-1, 1]\), we modify the min-max normalisation previously used in section 3.4.1 accordingly and normalise the \(C_1-C_6\) distances to span the range \([-1, 1]\).

In the original non-periodic treatment of angles in Section 3.4.1, we identified \(N_C = 4\) clusters with \((N_{\text{noise}}, \bar{S}_{\text{avg}}) = (7\%, 0.37)\), as summarised in Fig. 3.8. Using this new definition of a distance metric, we also identify \(N_C = 4\) clusters with very similar values \((N_{\text{noise}}, \bar{S}_{\text{avg}}) = (11\%, 0.32)\) for input parameters \((\varepsilon, \text{MinPts}) = (4.3, 4)\). While the fraction of trajectories marked as noise increases slightly and the value of \(\bar{S}_{\text{avg}}\) decreases somewhat, we note that only six trajectories end up being reclassified, with five out of these reclassified as outliers. Upon comparison of the distribution of clustered trajectories in the original treatment, shown in Figs. 3.18-3.19, with the newly identified clusters in Figs. 3.20-3.21, it is apparent that the clustering result
is largely unchanged. This is also reflected by the similarity of Figs. 3.22 and 3.12, which show the centre and distribution of each cluster. Note that in Fig. 3.20 some of the tZt-HT trajectories within $C_3$ become visually clearer than in Fig. 3.18, this is due to the reclassification of three $C_2$ trajectories as noise and does not indicate any significant deviation.

**Figure 3.21:** Plots of all clustered trajectories as a function of time, showing each of the four selected features. In this figure, the clusters have been obtained using the $L_2$ norm in a seven-dimensional space defined by three pairs of angular coordinates $(\cos \phi, \sin \phi)$, accounting for periodicity, and the $C_1-C_6$ distance. The top left and right panels show the $C_1-C_6$ bond and the central dihedral angle $\phi_b$, whereas the bottom left and right panels show the dihedral angles $\phi_a$ and $\phi_c$. Note that only minor differences with Fig. 3.19 are apparent.

### 3.5.2 A Norm Based on Arc lengths

We now consider a second alternative treatment of the periodicity of the angles, which does not increase the dimensionality of the feature space. Consider the angular coordinates $x_i$ and $y_i$ of a pair of trajectories $x$ and $y$. While these angles independently may span the range $0$ to $2\pi$, the maximum pairwise distance between the two angles corresponds to the minimum arc length around the unit circle in both the clockwise
Figure 3.22: Plots of the cluster centres in each of the 4 selected features for $\varepsilon = 4.3$ and $\text{MinPts} = 4$, using the angles projected onto their $(\cos(\phi), \sin(\phi))$ coordinates on the unit circle. The top left and right panels show the $C_1-C_6$ bond length and the central dihedral $\phi_b$, whereas the bottom left and right panels show the dihedral angles $\phi_a$ and $\phi_c$. The shaded areas represent the size of the cluster. Importantly, the centres and distribution of the clusters are near identical to Fig. 3.12.

and anti-clockwise directions, which is given by,

$$l_i^m = \min\{l_i, 2\pi - l_i\}, \quad (3.7)$$

with $l_i = |x_i - y_i|$ being the absolute difference between the two angles in radians. Using this definition, we define a composite norm as,

$$L_{\text{comp}}(x, y) = \left( \sum_i N_i \sum_t (l_i^m)^2 + \sum_j N_j \sum_i (x_{ij} - y_{ij})^2 \right)^{\frac{1}{2}}, \quad (3.8)$$

where the first sum with index $i$ accounts for the angular distances and the second sum with index $j$ represent the inter-atomic distances. Given that the values of $l_i^m$ span the range $[0, \pi]$, we again modify the normalisation of the $C_1-C_6$ inter-atomic distances to match this so that all features are treated uniformly.
Using this new distance metric and input parameters \((\varepsilon, \text{MinPts}) = (5.1, 4)\), we again identify \(N_C = 4\) clusters with \((N_{\text{noise}}, S_{\text{avg}}) = (7\%, 0.38)\), providing a marginal improvement on both previous results shown in sections 3.4.1 and 3.5.1. Moreover, turning to Figs. 3.23-3.25, we see that the clustering result is nearly identical with only a very small number of trajectories reclassified.

**Figure 3.23:** Distribution of the four clusters of trajectories identified using the composite norm \(L_{\text{comp}}\) defined according to Eqn. 3.8. Note that the plot shows the three dihedral angles, with the \(C_1-C_6\) bond omitted.
3.5 Periodicity of Angular Coordinates

Figure 3.24: Plots showing the clustered trajectories for each of the four selected features, with the clustering done using the composite norm $L_{\text{comp}}$ defined by Eqn. 3.8. The top left and right panels show the $C_1-C_6$ bond and the central dihedral angle $\phi_b$, whereas the bottom left and right panels show the dihedral angles $\phi_a$ and $\phi_c$. Note that the differences compared to the original results in Fig. 3.19 and those shown in Fig. 3.21 are very minor.

At this point, we conclude our discussion on the treatment of the periodicity of angular features. In terms of the two approaches presented and evaluated, we note that although the approach based on $(\cos, \sin)$-coordinates in section 3.5.1 and the approach based on arc lengths in section 3.5.2 yield very similar results, more generally the former approach may struggle to identify clusters due to the increase in dimensionality as the same number of trajectories span a much larger volume in the feature space. Therefore, while we favour the composite norm based on arc lengths, both approaches work well and are essentially equivalent in this case. Finally, we explain the similarity of the original results in section 3.4.1, which did not account for periodicity, with those in this section by examining the features of the trajectories. For example, while the dihedral angle $\phi_b$ may independently reach values exceeding $200^\circ$ or $-50^\circ$ (equivalent to $310^\circ$), all four clusters identified are separable with respect to more than one independent feature. This is apparent in Fig. 3.12, where while clusters...
3.5 Periodicity of Angular Coordinates

Figure 3.25: Plots of the cluster centres in each of the 4 selected features for $\varepsilon = 5.1$ and $\text{MinPts} = 4$, using a composite norm that includes the minimum arclength between a pair of angles. The top left and right panels show the $C_1-C_6$ bond length and the central dihedral $\phi_b$, whereas the bottom left and right panels show the dihedral angles $\phi_a$ and $\phi_c$. The shaded areas represent the size of the cluster. Again, note the similarity with Fig. 3.12.

$C_1$, $C_2$, and $C_4$ largely overlap in the $\phi_a$ and $\phi_c$ coordinates, they have distinct density maxima centred on different values of $\phi_b$ and $C_1-C_6$. Furthermore, one can see that $C_3$ can easily be identified from the other clusters due to its larger values of $\phi_a$ and $\phi_c$. An analysis of the currently evaluated trajectory data-set thus explains why the results are robust whether the periodicity is accounted for or not. However, this may not be the case more generally, and thus unless the angles only span the range $[0, \pi]$, then one should explicitly account for periodicity using one of the approaches described in this section.

---

4. A discussion on the limits at which not accounting for periodicity in the $L_2$ norm fails is given in Appendix A.
3.6 Conclusions

In this chapter, we have explored the application of the DBSCAN algorithm to semi-classical trajectories that model the electrocyclic ring-opening of CHD. We have discussed the general framework and approach to applying clustering algorithms to time-dependent trajectory data that models quantum dynamics. First, one must identify a reduced number of features on which to carry out the clustering, with the most straightforward approach being to look at the variance of different coordinates over time. One must ensure that the selected features are capable of capturing the different available products. In the event that the products are not known apriori, these can be most easily identified by removing the time-dependence and applying the clustering algorithm to the final time steps of the simulation. After normalisation of the data, a similarity matrix containing all pairwise distances between trajectories can be constructed. In this chapter we first discussed the application of both the multidimensional $L_2$ norm, and MD-DTW. In an application to CHD, these yielded similar results, primarily due to the fact the clustering result is dominated by large amplitude motion. However, we saw that MD-DTW yielded a unique result, providing some greater separability of a broad range of data. Both similarity measures struggled with correctly classifying some trajectories in terms of their populations, based on nuclear geometries alone, but given the complicated and indirect relationship between population and nuclear geometry, this is not surprising. Following publication it was pointed out that both the multidimensional $L_2$ norm and MD-DTW do not explicitly account for the periodicity of angles. Two alternative treatments of the angular coordinates were then explored. First, the dihedral angles were linearised by projection onto their corresponding coordinates on the unit circle. Second, we proposed a composite norm in which inter-atomic distances are treated in the usual fashion and the distance between angular coordinates correspond to the minimum arc length between them. Crucially, it was shown that the lack of periodicity in the original treatment of angles had little effect on the clustering of the data considered here. However, as this is not generally guaranteed, one should ensure the periodicity of angles is correctly accounted for.

Overall, modelling the trajectories as a multidimensional time series has the potential to provide a fast first order identification of the pathways that lead to product formation. In comparison to separating trajectories by eye, this provides a way to automatically identify the key product channels, and accounts for the multidimensional nature of the dynamics to a greater extent. At the very least, it is a powerful complement to
3.6 Conclusions separation and clustering by eye and can aid the discovery process when interpreting a large simulation data set. Moreover, in the clustering approach, the regions of highest density correspond to the cluster centres. This provides for easier sampling of trajectories from the ensemble. One can simply sample trajectories from around each cluster centre, with the number of trajectories sampled proportional to the size of each cluster. However, this approach is not without issues, including not being sensitive to subtler changes at early times or not capturing the steady decay of excited state population. Thus, its ability to provide more in-depth insights into the simulated dynamics is limited. As mentioned in the introduction to this chapter, the contents of this chapter constitute a first attempt at performing time-dependent clustering using existing algorithms. It therefore forms the basis of the next chapter, in which we move beyond the time series approach. This involves considering the spatial coordinates of each trajectory at every time step, rather than just considering the shape of the time series corresponding to internal coordinates. By applying a spatial clustering algorithm at each time step, one can subsequently mine the spatial correlations between trajectories in time. This provides a way to build a more complete summary model of the dynamics. As we shall see, the application to QMD trajectory data constitutes a special case, with several additional complications.
Chapter 4

Pattern Mining Trajectories

4.1 Introduction

The mining of spatio-temporal clusters is distinct from the time series approach discussed in the preceding chapter. Here we consider trajectory data as a group of moving objects, each with their own spatial and temporal coordinates. By directly considering the spatial coordinates at each time step, we allow for the identification of objects that have both a similar degree of spatial overlap, and temporal coherence. Generally, the mining of spatio-temporal clusters involves the independent application of a clustering algorithm to the data at each time step, the temporal correlations are then mined by identifying the groups of trajectories that are continually clustered together in time. This approach to clustering is generally better suited to time-dependent data in which the spatial evolution is important, for example in the analysis of vehicular GPS data, monitoring migration patterns, and epidemiology [75, 76, 148–151, 156, 205, 206]. When applied to semi-classical trajectories that model quantum dynamics, we wish to identify the groups of trajectories that share a similar evolution on the excited potential energy surfaces. The identified spatio-temporal clusters can be thought of as representing distinct concentrations of the wavepacket flux. The benefit of this approach is that it allows us to not only identify similar trajectories, but also the time-segments during which they share a similar spatial evolution and the time-points at which they diverge.

As observed in Chapter 3, the trajectory data considered here is challenging due to the presence of clusters that vary significantly in density and size. Moreover, with quantum dynamics being inherently high dimensional, balancing the size of the selected feature set to include sufficient information for effective and physically relevant clustering while avoiding linear dependencies and keeping the feature set as small as possible, can be difficult. In this chapter we shall develop two new algorithms that are well suited to the type of data that emerges from trajectory-based QMD simulations.
While these constitute two separate algorithms, combined they provide an effective tool for spatio-temporal clustering. We will first introduce the subtleties of pattern mining QMD trajectories, and then explore the new spatial clustering algorithm. This algorithm aims to model the distribution of trajectories at any time step by reproducing their associated density field over the feature space. Second, we introduce a new pattern mining algorithm, which mines the temporal correlations of trajectories across the spatial clusters identified at every time step. To achieve this, we propose a new definition of a spatio-temporal cluster, which allows for short term deviations between trajectories. Using the same set of trajectories from the previous chapter, which model the ring-opening dynamics of 1,3-cyclohexadiene, we demonstrate the application of both algorithms. Together, the two developed algorithms provide a much more detailed and comprehensive analysis of the dynamics than the time series approach, and can identify shared segments between trajectories, bifurcation, and changes in state populations over time. This results in a valuable tool capable of identifying decay or reaction pathways, providing mechanistic insight into complex simulations, as well as aiding in the interpretation of experimental observations and the identification of a reduced set of trajectories that can be used to calculate highly accurate observables.

Following submission of this thesis, it was pointed out that the first version of the newly developed spatial clustering algorithm does not explicitly account for the periodicity of angular features. In the revised Chapter 3, we demonstrate that periodicity has little effect when clustering the CHD trajectories due to the underlying distribution of the data. Similarly, at the end of this chapter we analyse the effect of explicitly accounting for periodicity when mining spatio-temporal patterns. We show that the same key trends emerge as in the original clustering without periodicity, but note that further investigation is necessary to determine the optimal manner in which to account for periodic features.

4.2 A Density Based Spatial Clustering Algorithm

4.2.1 Problem Definition

The application of clustering to trajectories that model quantum dynamics presents several problems for existing algorithms. Quite generally in photoinduced dynamics, the entire ensemble of trajectories is localised in the Franck-Condon region immediately following excitation. At these early times, the molecule may only undergo
small changes in geometry. However, as the wavepacket evolves, possibly splitting into different pathways, this results in an inhomogenous set of clusters of differing density, in which trajectories do not span equal-sized volumes. This can be seen in Fig. 3.6 of Chapter 3, where a number of pathways with varying distributions emerge as time evolves. To see how this situation arises, consider two reaction pathways available upon bifurcation of the wavepacket. The first pathway involves population transfer to a lower electronic state which has a steep gradient, resulting in a localised group of trajectories that evolve in a ballistic manner with little dispersion. The second pathway, instead, follows a relatively flat PES, meaning that trajectories on the second pathway evolve similarly, but end up being more dispersed. In order to capture the key concentrations of reaction flux, the clustering algorithm must be flexible and capable of fulfilling the following four criteria, preferably without requiring intervention or re-optimisation of parameters. First, it must be capable of grouping trajectories together at time zero, where there is little variance. Second, it much be able to capture the initial subtle changes as the wavepacket begins to move out of the Franck-Condon region. Third, as the wavepacket begins to bifurcate and evolve along different regions of the PES, it must be able to detect the locations that exhibit the highest density within overlapping clusters of points. Fourth, as the system converges to a series of different product channels, it must be able to identify well-separated clusters of varying density and size.

Figure 4.1: A two-dimensional example of typical trajectory evolution across three time points $t = [t_0, t_1, t_2]$. Circled points correspond to areas with considerable density.
These requirements are best demonstrated by the two-dimensional example in Fig. 4.1. Initially, at time $t_0$, all trajectories are concentrated within a localised region corresponding to the initial conditions prepared on an excited state. At the next time point, $t_1$, we see that the trajectories have dispersed and begun to follow different pathways. While there is a degree of overlap, the highest density regions of space are circled. Continuing to time point $t_2$, we see that the trajectories have continued to evolve along their respective pathways, each with a different size and internal density distribution. The trajectories that continually move together are indicated by circles of the same colour.

**Figure 4.2:** Demonstration of the problems encountered when clustering QMD trajectories. The first two columns correspond to the clustering obtained with parameterised values for the DBSCAN algorithm (column one: $\varepsilon = 0.4$ and $MinPts = 3$; column two: $\varepsilon = 0.3$ and $MinPts = 3$). The third column shows the results from the new clustering algorithm introduced in this chapter. Different rows correspond to different times $t = [t_0, t_1, t_2]$. The colour of each point indicates which cluster the trajectory belongs to, with black representing trajectories marked as noise (and thus not assigned a cluster).
4.2 A Density Based Spatial Clustering Algorithm

In order to identify the time-segments over which trajectories share a similar evolution through phase space, we require a flexible and general algorithm that meets the four aforementioned requirements and that is effective across the entire dynamics explored by the trajectories. This presents a challenge for many existing density based clustering algorithms, which will typically require different parameterisations at different times. This is demonstrated by the example in Fig. 4.2, where in the first two columns we see how the DBSCAN [145] algorithm fails to identify key regions of density. In the first column, we see the clusters identified by DBSCAN with the parameters set to $\varepsilon = 0.4$ and $\text{MinPts} = 3$ over all time steps. While most trajectories are contained within one cluster at $t_0$, at subsequent time steps the algorithm either misses regions of density or merges them. Reducing the value of $\varepsilon$ to 0.3 in the second column, the algorithm manages to separate out the two regions of density at the final time step. However, it still fails to detect other regions of density. Without optimising the parameters at each time individually, it is difficult to determine where there is a concentration of reaction flux. In addition, even if one performed a dynamic re-parameterisation across all times, it is not particularly well suited to identifying clusters of highly variable density. In comparison, in the third column we show the results from the new algorithm presented in this chapter. This algorithm does not have to be re-parameterised at different times, and fulfils the main requirements of being well behaved in both the Franck-Condon region and at later times, being capable of identifying the key regions of density within overlapping data points, and clusters of variable density and size. This is achieved by constructing a density field that describes the distribution of points on a discretized feature space, amounting to a mapping of the wavepackets probability density to a lower dimensionality space. Therefore, providing a suitable set of features are identified, the grouping of trajectories reflects the probability density of the wavepacket.

4.2.2 Algorithm Overview

The key component of a density-based clustering algorithm is how it defines density. In our approach, the distribution of trajectories at each time point is represented by a density scalar field. Following discretization of the feature space, a series of density kernels are placed on each data point. From this, the density at any point $x(1: N_f)$ in the $N_f$ dimensional feature space that the data spans, can be calculated as,

$$\rho (x) = \sum_{i=1}^{N_i} K (||x - x_i||), \tag{4.1}$$
where \( x_i \) is the position of data point \( i \) and \( K \) the density kernel. Generation of the density field therefore corresponds to a summation of contributions from a total of \( N_i \) density kernels, where \( N_i \) is the total number of data points - i.e. the total number of trajectories at a given time step. In principle, one is free to choose any kernel. We opt for a Gaussian kernel due to the fact the wavepacket observes a Gaussian probability distribution, and for its mathematical convenience, notably the absence of discontinuities or singularities. The \( N_f \) dimensional Gaussian kernel can simply be defined as the product of a series of one-dimensional Gaussians,

\[
K(x) = \prod_{f=1}^{N_f} \exp \left( -\frac{(x_f - x_{if})^2}{2\sigma_f^2} \right),
\]

with \( x_{if} \) the \( f^{th} \) component (feature) of data point \( x_i \), and \( \sigma_f^2 \) the variance of the data along the feature \( f \). The Gaussian width can be scaled separately for each feature (dimension) via its standard deviation \( \sigma_f \), allowing one to capture the distribution of the data in the multidimensional feature space. Note, the total set of \( N_f \) widths remains constant for each data point on which the Gaussian kernel is centred. While the clustering results ultimately depend on the width of the Gaussian and the discretization process, setting the Gaussian width to reflect the standard deviation at each time step allows for an adaptive definition of the density neighbourhood. This achieves a suitable resolution throughout and ensures that the algorithm does not require that the parameters are tuned at each time step. In Fig. 4.3, we see a simple one dimensional demonstration of these concepts. Here, the total density field, indicated by the solid grey line, is generated by summation of the contributions from four density kernels. Maxima in the field correspond to regions of high density, providing the location and distribution of the most concentrated points in the feature space.

Clusters are constructed by taking each data point and assigning it to the local maximum via a gradient ascent traversal of the density field. This provides an initial grouping of data points, from which the final series of clusters are identified by merging maxima based on their separation and local density landscape. The benefit of this approach is that it allows for data points to be assigned to their corresponding maxima in data sets that exhibit clusters of highly varying density, and in which there may be a degree of overlap. This enables clustering which reflects the true probability density
4.2 A Density Based Spatial Clustering Algorithm

Figure 4.3: Demonstration of the total density along a one-dimensional feature space as a sum of contributions from Gaussian kernels centred on individual data points, depicted in blue, orange, green, and magenta. There are two maxima $C_1$ and $C_2$, representing the two clusters to which the data points are assigned. The (kernel) densities for individual data points are shown as dashed lines in the same colour as corresponding data point.

distribution of the wavepacket, without redefining clustering parameters at each time step. In addition, the algorithm is capable of identifying nonconvex clusters and does not rely on distance metrics that typically under-perform in higher dimensional spaces, at least for the initial grouping of points based on maxima assignment.

4.2.3 Grid Traversal

First we address the discretization process and the subsequent assignment of data points to local maxima. The algorithm proceeds by discretizing the $N_f$ dimensional feature space into $N_G$ grid points. In a naive approach, one could compute and store the value of density at every point in the grid prior to the assignment of maxima. However, such an approach is not feasible due to the amount of memory required, and the associated scaling of order $O(N_fN_G)$. Instead, each data point is treated sequentially, with the path through the density field to a local maxima generated on-the-fly for each data point. While this approach is not as scalable as widely used algorithms such as DBSCAN, within the context of QMD trajectories, we are concerned with a comparably small number of data points, usually on the order of several hundred and rarely more than a thousand. As such, the sacrifice of somewhat poorer scaling in comparison to widely used algorithms is justified by the improvement in the quality of the results. Crucially, within the relative size of QMD data sets, the dynamic approach almost always allows a suitable grid resolution capable of capturing the trends in the data.
Prior to discretization, the data must be normalised, as discussed in Chapter 3.3.2, for which we utilise a min-max normalisation that scales each feature to the range $[0, 1]$. In principle, the feature space can be discretized into $N_G$ total grid points with a variable number of grid points along each feature (dimension) $N^f_G$. However, for a simpler treatment, we opt to partition the feature space equally along each of its $N_f$ dimensions. There are a number of ways in which the feature space can be discretized, for example, the number of grid points may be specified by the user themselves, providing the value of $N_G$ selected is suitable, this can be set globally over all times. Alternatively, the value of $N_G$ may be automatically set such that the grid spacing corresponds to a value smaller than the minimum distance between data points. In this work, we however define the number of grid points along each dimension based on the standard deviation of the data as,

$$N^f_G = \frac{N}{\min\{\sigma_f\}}, \quad (4.3)$$

where the denominator corresponds to the minimum value of the set of standard deviations along each min-max normalised feature $\sigma_f = [\sigma_1, \sigma_2, \ldots \sigma_{N_f}]$. The value of $N$ can be tuned, however a value of $N = 20$ provides a sufficient grid resolution in all of the utilized test and real data sets. Once the number of grid points is known, we scale the component of the data points along each feature $x_f$ such that $x_f \in [0, N^f_G]$, allowing a simpler integer representation of the grid. As a consequence of scaling the data points, the standard deviation $\sigma_f$ of each feature must be scaled in the same fashion to ensure that the Gaussian widths are consistent with the scaled data.

Due to the dynamic generation of the density field, we require the density to be independent of the number of data points, and to fulfil the normalisation condition,

$$\int \rho(x) d^N x = 1, \quad (4.4)$$

where the left hand side evaluates to,

$$\int \rho(x) d^N x = \sum_{i=1}^{N_f} \int K(||x - x_i||) d^N x = N_i \times \prod_{f=1}^{N_f} \sqrt{2\pi\sigma_f}, \quad (4.5)$$
4.2 A Density Based Spatial Clustering Algorithm

making use of the fact that the value of $\sigma_f$ is the same for each of the Gaussian kernels. The normalisation condition can then be ensured through a mapping of each Gaussian width $\sigma_f$,

$$M_{\text{var}} : \sigma_f \rightarrow \sigma_f \times (N_i)^{-1/N_f} \left(2\pi\right)^{-1/2} \left(\prod_{f=1}^{N_f} \sigma_f \right)^{-1/N_f}.$$  \hspace{1cm} (4.6)

Following discretization and normalisation of the density field, we proceed with the assignment of data points to local maxima. A visual representation of the process in a simple two-dimensional feature space is shown in Fig. 4.4. Here the panels (a-d) correspond to different steps in the assignment procedure, which is performed for every data point sequentially. In each case, the discretized grid is indicated by the grey points and the total density field by the background heatmap. Starting with panel (a), the selected data point, shown as a magenta circle, is projected onto its nearest coordinate in the grid, indicated by the red cross. Next, in panel (b), the density at the initial grid point and its $2 \times N_f$ nearest neighbours is evaluated. Here, the nearest neighbours are indicated by orange crosses. The nearest neighbour with the highest value of density is then selected as the next active grid point, as indicated in panel (c). This gradient ascent style traversal of the density field continues until all nearest neighbours of the active grid point are lower in density, and thus a local maxima has been detected, as in panel (d) where the final grid-point assignment of the initial magenta data point in panel (a) is shown by a magenta cross. This procedure is repeated until each data point has been assigned to a local maxima, providing an initial grouping of data points. Crucially, the independent and sequential assignment of data points has no order dependence.

4.2.4 Merging Criterion

The gradient ascent allows for the identification of local maxima, providing an initial grouping of points based on the path taken. However, the clustering process is not yet complete and we must address two potential issues with the grouping of data points. First, in the event that there exists a number of semi-isolated data points, this can lead to maxima that are very small in magnitude, ultimately resulting in clusters that contain a single data point. Second, since it is not guaranteed that regions of high density correspond to a single finely peaked maximum, the topology of the density
4.2 A Density Based Spatial Clustering Algorithm

Figure 4.4: Demonstration of the gradient ascent process used to assign data points to a local maxima, the background heatmap represents the density field. In a) a data point (magenta circle) is projected to its nearest grid point (red cross), b) density is evaluated at each $2 \times N_f$ nearest neighbours (orange crosses), c) a step is taken in the direction that the gradient increases most, d) the process is repeated until a local maxima is reached and the initial data point in a) is assigned to it, as indicated by the magenta cross.

The field may include overlapping maxima with saddle points between them. As a result, points that in fact correspond to a single region of high density, may be split between adjacent maxima. To address both of these issues, we now introduce a two-step merging procedure, which merges maxima to build the final set of clusters.

The first step deals with issue of semi-isolated data points by considering the density contribution of all other maxima to the data point in question. This involves the notion that if the assignment of data point $x_i$ to a specific maximum is correct, then the contribution of all other maxima to the total density at $x_i$ (excluding the data points own density) should be lower in comparison to the contribution from the maximum it was originally assigned to. In the event it is not, the data point is reassigned to
reflect this. The effect of this contribution-based adjustment can be seen in Fig. 4.5(b-c), where the isolated data points circled in red are reassigned to the maxima that contributes the most to their own density. The green circled region corresponds to an area where there are three maxima that are in very close proximity. Upon adjustment, the points contained within the smaller middle maximum are reassigned to the larger adjacent maximum. However, note that following the contribution-based adjustment in Fig. 4.5c, we still observe that some data points within high density areas are split between two overlapping maxima, due to the absence of a narrowly peaked maximum. The second step of the procedure therefore involves the introduction of a merging metric, which decides if the points within two maxima should be merged together. This merging metric is defined as,

\[ \Delta_M = \Delta_\rho \Delta_d, \] (4.7)

which corresponds to the product of two components, \( \Delta_\rho \) and \( \Delta_d \), which describe the density profile and separation between two maxima respectively. The density component is given by,

\[ \Delta_\rho = \frac{\min(\rho_1, \rho_2) - \rho_0}{\rho_0}, \] (4.8)

where \( \rho_1 \) and \( \rho_2 \) correspond to the density at the nearest neighbouring points between the maxima, and \( \rho_0 \) is the minimum value of density along the straight line that connects them. Each of the values used in the calculation of the density component are shown in Fig. 4.5(a) for a simple one-dimensional example containing two maxima, labelled \( C_1 \) and \( C_2 \). If the density between these two maxima drops significantly relative to the height of the smallest maximum, then the value of \( \Delta_\rho \) increases as the maxima are well defined. In the case of a multidimensional Gaussian, the determination of \( \rho_0 \) is less clear. Here, we find the value of \( \rho_0 \) by considering the closest pairs of core points between maxima. Core points are defined as those that have a density greater than the average of all points assigned to the maxima. This is demonstrated in Fig. 4.5(d), where only the core points assigned to each maxima are coloured. For each pair of neighbouring core points we identify the minimum density along the straight line that connects them, and the value of \( \rho_0 \) is then taken to be the minimum value from all pairs of points considered. By considering a set of nearest neighbouring points in this fashion, we ensure adequate sampling of the space between the two multidimensional Gaussian’s. The second distance component of the merge metric is
4.2 A Density Based Spatial Clustering Algorithm

then calculated as,

\[ \Delta_d = \frac{d_0}{\min(d_1, d_2)}, \]  

(4.9)

where \( d_0 \) is the minimum distance between the pairs of points assigned to each maxima, and \( d_1 \) and \( d_2 \) correspond to the minimum nearest neighbour separation within each maxima, again shown in Fig. 4.5(a). If the merge metric \( \Delta_M < 1 \), the points within each maxima are merged together. This process is repeated for every pair of groups until merging no longer occurs and the final set of clusters are identified, as in Fig. 4.5(e).

4.2.5 Algorithm Performance

As discussed before, the main requirements for our algorithm are that it is capable of identifying the key regions of density within potentially overlapping data, is able to detect multiple clusters of varying density and size, and is capable of automatically capturing the large changes in variance seen in QMD trajectories over time without the need for tuning parameters at each time step. While the success of our algorithm was briefly shown in Fig. 4.2, we further demonstrate that the algorithm meets these criteria by turning to a series of synthetic data sets, generated using \textit{scikit-learn} [207]. Here data is generated from a series of Gaussian distributions with different standard deviations.
Figure 4.5: Demonstration of the merging criterion in our method. Panel a) defines the parameters used in the calculation of the merging metric $\Delta_M$. Panel b) shows the results of the initial assignment of maxima based on gradient ascent, where red and green circles indicate semi-isolated points and points assigned to a maxima surrounded by two larger maxima, respectively. Panel c) shows the reassignment of the circled points to the maxima which contribute the most to their density. Panel d) shows the core points used to determine $\rho_0$ for two clusters. Panel e) shows the final clusters after merging.

The application of our algorithm and two other common density-based clustering algorithms (DBSCAN and BANG) to this data, is shown in Fig. 4.6. Each row corresponds to a different data set and each of the columns one through three show the result achieved using a different algorithm. The fourth column shows the ground
Figure 4.6: Results of clustering using the BANG (column one) and DBSCAN (column two) algorithms vs our method (column three). The ground truth, shown in column four, provides the correct labelling of data points according to which of the three distributions they were generated from initially. Clusters are represented by different colours: orange, green, blue, and magenta, and black data points represent noise. Each row corresponds a different data set, demonstrating algorithm performance across different cluster sizes and densities.

truth, where the colours correspond to the distribution from which the data points were sampled. In row one, we see that all three algorithms perform well when the clusters are highly localised and of equal size. In row two, we consider a more varied data set, which contains two localised clusters and one central more diffuse cluster. Again, all algorithms perform well, however both BANG and DBSCAN require that the parameters are tuned to achieve optimal results, as indicated by the values shown in the lower right corner of each panel. Without this adjustment, these two algorithms can only capture the most central points of the diffuse cluster. Both the BANG algorithm and our method show the presence of an isolated two-member cluster. In our case, the contribution-based adjustment performed before merging is not capable of
reassigning these points to the diffuse central cluster. In the third row, we consider an even more varied data set in which points between clusters overlap. Here both BANG and DBSCAN fail to identify the clusters present, despite parameter tuning. BANG results in semi-isolated points localised within the broader distribution being identified as a series of small clusters. In the DBSCAN case, the central region is too delocalised to be identified as its own cluster, and instead merges with a smaller cluster. In contrast, the new method can handle the high variance in this data set and correctly identifies the central cluster. However, we note the presence of the small two-membered cluster and a small number of points that do not reflect the ground truth. Crucially, in all cases examined, the new algorithm achieves sensible results without the need for any tuning of parameters.

4.3 Trajectory Pattern Mining

Having developed a spatial clustering algorithm capable of identifying the key concentrations of density in overlapping data, and clusters of highly variable density over time, we now turn to the problem of mining the temporal correlations within these spatial clusters over time. The following sections explore the issue with existing definitions of spatio-temporal clusters, an alternative definition of a spatio-temporal cluster, and the application of an algorithm based on this new definition to the set of CHD trajectories introduced in Chapter 3.2.

4.3.1 Problem Definition

Given a series of trajectories that are spatially clustered together in time, there may exist trajectories that exhibit short term deviations from the spatial cluster they are most strongly associated with. Existing pattern mining algorithms that rely on definitions of a spatio-temporal cluster such as a moving cluster [159] or convoy [160], first introduced in Chapter 2.4.2, are all subject to the constraint that they exist for at least $t_{\text{min}}$ successive time steps. In the case of convoys, one wishes to identify sets of trajectories that are always spatially clustered together in time. This means

---

1. We refer to the clusters identified at each independent time step by application of our density clustering algorithm as spatial clusters. In addition, spatio-temporal clusters refer to the groups of trajectories that are contained within the same spatial clusters over a series of time steps; this is used a general term, as we shall introduce different definitions of spatio-temporal clusters used in pattern mining.
that if a trajectory leaves a convoy, it cannot rejoin at a later time. Furthermore, if at any time the number of trajectories within the convoy drops below a minimum threshold, $MinPts$, then the convoy is terminated. This occurs even if the same group of trajectories are clustered together at a series of later time steps. On the other hand, a moving cluster defines a spatio-temporal cluster through an overlap requirement $\theta_{\text{min}}$. In this framework trajectories may leave the moving cluster and rejoin at a later time step as long as the overlap requirement is continuously satisfied. However, if at any time the overlap between consecutive time steps falls below $\theta_{\text{min}}$, or the subsequent spatial cluster contains less than $MinPts$ trajectories, then the moving cluster is terminated. Again, this breakdown occurs even if there exists a spatial cluster at later times that fulfils these requirements. While moving clusters provide a greater degree of flexibility in identifying temporal patterns compared to convoys, they are still subject to the consecutive time step constraint.

In Algorithm 1, we outline the basic algorithm that allows for the identification of moving clusters. At line 3 we begin the main loop over time, at line 6 we proceed to perform spatial clustering on the trajectory set at time $t$. While originally designed for DBSCAN, in principle one is flexible to choose any spatial algorithm. From lines 7-14 we compare every combination of spatial clusters at time $t$ to the active moving clusters. If a cluster is assigned to one of the moving clusters (i.e. it has an overlap $\geq \theta_{\text{min}}$ and contains at least $MinPts$ points), we update it and append it to $M_{\text{next}}$. On lines 15 and 16, if a spatial cluster is not assigned then it must be initialised as a new moving cluster for the next time step. On lines 17-19, if a moving cluster is not extended to the subsequent time step but its duration is $\geq t_{\text{min}}$, we output it.

The consecutive time step constraint in the context of moving clusters is best demonstrated in Fig. 4.7. Here we observe a set of four trajectories $O = [o_1, o_2, o_3, o_4]$ evolving over five times $T = [t_1, t_2, t_3, t_4, t_5]$. At each time, the trajectories that are spatially clustered are contained within the grey ellipse. Setting the parameters $(\theta_{\text{min}}, t_{\text{min}}, MinPts) = (0.75, 2, 3)$, we observe a valid moving cluster over the first two time steps, $t_1$ and $t_2$. However, at time $t_3$, two trajectories leave the cluster, and the requirements defined by $\theta_{\text{min}}$ and $MinPts$ are not met. As a result, the moving cluster is terminated, and the algorithm cannot detect that the same group of trajectories at time $t_4$ are clustered together again. Applying Algorithm 1 to this example therefore results in the identification of two different moving clusters, despite both describing a similar trajectory evolution. The first moving cluster contains the spatial clusters at times $t_1$ and $t_2$, and the second contains the spatial clusters at $t_4$ and $t_5$. Ultimately, the
Algorithm 1 A simple moving cluster algorithm

1: \textbf{procedure} MOVINGCLUSTERS\((O, T, \text{MinPts, } \theta_{\text{min}}, t_{\text{min}}, \ldots)\)
2: \hspace{1em} \(M := \emptyset\) // Set of current active moving clusters
3: \hspace{1em} \textbf{for all} time steps \(t \in T\) \textbf{do}
4: \hspace{2em} \(M_{\text{next}} := \emptyset\) // Set of moving clusters to be carried forward
5: \hspace{2em} \(O_t := O[t]\) // Trajectory set at time \(t\)
6: \hspace{2em} \(S_t := \text{SpatialClustering}(O_t, \ldots)\) // Spatial clustering algorithm of choice
7: \hspace{2em} \textbf{for all} current spatial clusters \(c \in S_t\) \textbf{do}
8: \hspace{3em} \(\text{assigned := False}\)
9: \hspace{3em} \textbf{for all} active moving clusters \(m \in M\) \textbf{do}
10: \hspace{4em} \textbf{if} \(\frac{|m \cap c|}{|m|} \geq \theta_{\text{min}}\) \textbf{then}
11: \hspace{5em} \(m.\text{extend} := \text{True}\)
12: \hspace{5em} \(M_{\text{next}} := M_{\text{next}} \cup (m \circ c)\) // Update the moving cluster
13: \hspace{5em} \(m.\text{duration} ++\)
14: \hspace{5em} \(\text{assigned := True}\)
15: \hspace{3em} \textbf{if} \text{not assigned} \textbf{then}
16: \hspace{4em} \(M_{\text{next}} := M_{\text{next}} \cup c\)
17: \hspace{3em} \textbf{for all} active moving clusters \(m \in M\) \textbf{do}
18: \hspace{4em} \textbf{if} \text{not } m.\text{extend } \& \text{m.duration} \geq t_{\text{min}} \textbf{then}
19: \hspace{5em} \textbf{output } m
20: \hspace{1em} \(M := M_{\text{next}}\) // Carry clusters to next iteration

algorithm cannot capture the presence of trajectories that are more loosely travelling within the spatio-temporal cluster. Algorithms based on the idea of a swarm [161], first introduced in Chapter 2.4.2, overcome the consecutive time step constraint by determining the set of times at which every combination of trajectories are grouped together. However, on the flip side, there is no guarantee that the swarms capture the full temporal evolution of the trajectories, as they may only be clustered together at a series of highly separated time steps.

In the following sections we shall explore a new definition of a spatio-temporal cluster, which is not subject to the consecutive time step constraint. This allows one to capture groups of trajectories that are strongly correlated in time, and those which share a looser a temporal correlation. This new definition is based on the idea of a moving cluster, therefore all spatio-temporal clusters identified have a consistent degree of correlation over the times they exist.
4.3 Trajectory Pattern Mining

Figure 4.7: Example of a moving cluster for an example comprised of four trajectories at five different times $T = [t_1, t_2, t_3, t_4, t_5]$. With the set parameters, the moving cluster breaks down briefly at $t_3$ before reforming.

### 4.3.2 Overcoming the Consecutive Time Step Constraint

To overcome the limitations of the consecutive time step constraint, we introduce a new definition of a spatio-temporal cluster, which we shall refer to as a relaxed moving cluster. A relaxed moving cluster includes a fourth parameter $\delta_{\text{max}}$, that defines the maximum number of successive time steps for which a moving cluster can fail to exceed $\theta_{\text{min}}$ and $\text{MinPts}$. Turning to the example comprised of four trajectories introduced in Fig. 4.7, keeping the same parameters as before and setting $\delta_{\text{max}} = 1$ we identify a relaxed moving cluster that exists over all five time steps, as seen in Fig. 4.8. While the overlap between spatial clusters at times $t_2$ and $t_3$ briefly falls below the overlap threshold of $\theta_{\text{min}} = 0.75$, we see that the overlap requirement is again fulfilled at the next time step, $t_4$. Here, trajectories $o_2$ and $o_4$ (red points) are tightly clustered at all times, whereas $o_1$ and $o_3$ (pink points) briefly leave the clusters at times $t_1$ and $t_3$, and are therefore more loosely associated with the spatio-temporal cluster although they clearly travel with the same group. Thus, relaxed moving clusters are sufficiently flexible to capture spatio-temporal patterns in which some trajectories are highly temporally correlated, while others are more loosely associated.

While a standard moving cluster is simply a list of correlated cluster objects, a relaxed moving cluster is defined by a hashmap $M_i$, where the hash function $h$ maps each trajectory object $o_i \in M_i$ as $h: o_i \rightarrow T_i$, with the elements $T_i$, the set of time steps that trajectory $o_i$ exists in $M_i$. For the example in Fig. 4.8, the identified relaxed moving cluster has the hashmap shown in Table 4.1. Therefore, every trajectory within the relaxed moving cluster has its own corresponding time set showing the times at which
4.3 Trajectory Pattern Mining

Figure 4.8: Example of a relaxed moving cluster for a series of four trajectories (red and pink circles) over five times ($t_1$-$t_5$). The addition of $\delta_{\text{max}}$ relieves the consecutive time step constraint. Red trajectories correspond to those that are always clustered together. Pink trajectories represent those that travel more loosely with the evolving trajectory ensemble. The relaxed moving cluster criteria allow all given trajectories to be identified as a spatio-cluster.

![Diagram of trajectories and time steps]

Table 4.1: Hashmap of trajectories and their corresponding time sets over which they are contained within the relaxed moving cluster in Fig. 4.8

| $o_1$ | $[t_1, t_2, t_4, t_5]$ |
| $o_2$ | $[t_1, t_2, t_3, t_4, t_5]$ |
| $o_3$ | $[t_2, t_4, t_5]$ |
| $o_4$ | $[t_1, t_2, t_3, t_4, t_5]$ |

it is a member of the spatio-temporal cluster. This allows for easier identification of the trajectories that are most strongly correlated in time, the degree to which more loosely travelling trajectories are correlated, and for identification of the times at which trajectories leave and join the cluster. In addition, this representation provides the user with a simple way of filtering trajectories based on their degree of temporal correlation.

4.3.3 Pattern Mining with Relaxed Moving Clusters

While our method is based on Algorithm 1, it differs in two main ways. The first is a result of the introduction of the $\delta_{\text{max}}$ parameter in the definition of a relaxed moving cluster. This inherently complicates the identification of spatio-temporal clusters. Due to the fact that at any time the relaxed moving cluster breaks down, it must continue to be propagated in time for at least $\delta_{\text{max}}$ time steps. We call this the $\delta$-propagation stage. During this stage, the trajectories from the highest overlapping spatial cluster...
are temporarily added as candidates to the hashmap associated with the relaxed moving cluster. If the moving cluster reforms during the $\delta$-propagation, these candidate trajectories are promoted to full membership on the condition that they are members before and after the $\delta$-propagation. Otherwise, the candidates are deleted and the relaxed moving cluster terminated.

The second difference lies in the interpretation of what the spatio-temporal cluster represents. In the standard moving cluster algorithm in Algorithm 1, the current set of active moving clusters are continually carried forward in time until the overlap requirement or the threshold defining the minimum number of points are no longer met. If at any time an additional moving cluster forms as a result of a subset of trajectories breaking away, the original moving cluster is still carried forward. The set of identified moving clusters thus defines the maximal set of times over which groups of correlated trajectories move together. In the mining of relaxed moving clusters, we take the approach that if a group of trajectories diverges from a larger group, and are continually clustered together in a different set of spatial clusters, the set of trajectories should be split into two separate relaxed moving clusters. The mining problem can therefore be understood as trying to identify the common subsequences that groups of trajectories may share. Physically, this provides a mechanism for identifying the approximate times at which the wavepacket bifurcates upon reaching a conical intersection or a cusp in the PES.

Practically, this can be achieved by representing the evolution of the clusters in time as a tree, where each relaxed moving cluster is represented by a node. Consider the case where one relaxed moving cluster contains the majority of trajectories following photoexcitation; this is initialised as the root node in the tree. Upon nonadiabatic population transfer to a lower-lying electronic state which causes different nuclear dynamics, the trajectories that undergo a transition should emerge as a new cluster. The relaxed moving cluster initialised as the root node may then spawn two child nodes, one corresponding to the group of trajectories that undergo population transfer to a different state, and the other to the trajectories that remain on the same excited state. The ability of the algorithm to detect this situation depends on the separability of the data, and nuclear geometries alone may not be sufficient. This is due to the fact that there is time delay between the change in electronic character and the resulting
change in nuclear motion. It is therefore recommended that one includes electronic state populations in the features considered for the clustering at each time step. The final output of the algorithm is a tree that comprises the relaxed moving clusters that meet the conditions set by the specified parameters.

The main concepts of our pattern mining algorithm can be understood by considering the example in Fig. 4.9, in which we observe a series of spatial clusters for a set of six trajectories over six time steps. Clearly, over all time steps there exist three main shared common sequences between the trajectories, these are coloured accordingly in the figure and correspond to the desired identification of relaxed moving clusters. The first of these include the set of six trajectories clustered together up until $t_3$, the second and third correspond to the trajectories that split into two separate clusters at times $t_4$ to $t_6$. Note that in the standard definition of a moving cluster, Algorithm 1 is not guaranteed to identify three different moving clusters. Looking at the clusters $C_{31}$ and $C_{41}$, containing six and four trajectories respectively, we see that they have an overlap of $2/3$. Therefore, if the overlap requirement is set to any value $\theta_{\text{min}} < 2/3$, then only two moving clusters are identified. The first corresponding to the set of all combined spatial clusters that are coloured green or orange, and the second to the set of three two membered clusters that are coloured purple. In this case, one can see that standard moving clusters define the maximal set of times over which a group of trajectories are clustered together, subject to the specified parameters. In contrast, our implementation of relaxed moving clusters, which focuses on identifying the common subsequences groups of trajectories share, is guaranteed to identify the three groups of trajectories even in the event that $\theta_{\text{min}} < 2/3$.

Setting the parameters $(\theta_{\text{min}}, \text{MinPts}, t_{\text{min}}, \delta_{\text{max}}) = (0.65, 2, 3, 0)$, the correct group of patterns can be identified. The algorithm starts at time $t_1$ by initialising the cluster labelled $C_{11}$ as a potential relaxed moving cluster, which we label $M_1$. At times $t_2$ and $t_3$ the overlap exceeds 0.65 and the number of trajectories exceeds two. Thus, $M_1$ is carried forward to time step $t_3$. At this point, $M_1$ has existed for at least three time steps, and therefore should result in output from the algorithm. Hence, $M_1$ is initialised as the root node in the tree of relaxed moving clusters. Once time $t_4$ is reached, the cluster $C_{21}$ separates out, containing two trajectories. This new cluster contains more than two trajectories, and therefore is initialised as a new relaxed moving cluster, $M_2$. Since the trajectories in $M_2$ derive from $M_1$, $M_1$ is assigned as the parent moving cluster.
4.3 Trajectory Pattern Mining

Figure 4.9: Example of six trajectories \((o_1-o_1)\) moving in time \((t_1-t_6)\). Each spatial cluster is coloured according to the relaxed moving cluster it should be assigned to, with different relaxed moving clusters represented by the colours green, orange, and purple. The spatial clusters at each time step \(i\) are labelled \(C_i^t\), where \(i\) is the cardinality of the cluster.

Note that the parent relaxed moving cluster \(M_1\) is not terminated at time \(t_4\), since cluster \(C_1^4\) shares 4 trajectories with it and exhibits an overlap greater than \(\theta_{\text{min}} = 0.65\). Therefore, \(M_1\) is carried forward to the next time step. Looking at times \(t_5\) and \(t_6\), it is clear that \(M_1\) and \(M_2\) continue to co-exist for the remainder of time. Once the time \(t_6\) is reached, \(M_2\) has existed for at least three time steps, qualifying it as output from the algorithm. However, as \(M_1\) is the parent of \(M_2\) and is still active, a check is performed, and \(M_1\) is then split at time \(t_3\) where \(M_2\) was spawned. The set of clusters \(C_1^4, C_1^5,\) and \(C_1^6\), which were assigned to \(M_1\) after \(t_3\), are then reassigned to a new relaxed moving cluster \(M_3\), which is assigned the now shorter \(M_1\) cluster as a parent. Both \(M_2\) and \(M_3\) are then added to the tree of final relaxed moving clusters given as output. Note that we have hereby introduced the condition that if any active relaxed moving exists for at least \(t_{\text{min}}\) time steps and has a parent assigned to it that is still active, the parent must undergo splitting in order to reflect that the trajectories diverge.
We perform this splitting only once the child relaxed moving cluster has existed for the required $t_{\text{min}}$ time-steps to avoid splitting for nonviable relaxed moving clusters. However, this means that if a relaxed moving cluster with several children assigned to it splits, then the assignment of parents must be updated to reflect this.

The tree that results from applying the algorithm to the example introduced in Fig. 4.9 is shown on the right hand side of Fig. 4.10. Here, the times shown at the nodes indicate when the relaxed moving clusters are initialised and terminated. On the left hand side of the figure, we see a tree representation of the clusters shown in Fig. 4.9. Here
4.3 Trajectory Pattern Mining

each node corresponds to a cluster at a given time step, with clusters containing more than two trajectories indicated in red. Each level of the tree therefore contains the spatial clusters identified at each time step from application of our density clustering algorithm. The edges represent the number of trajectories shared between clusters at consecutive time steps. Instead of calling our clustering algorithm at each time-step within the pattern mining algorithm, we identify all spatial clusters before performing pattern mining. The resulting list of cluster objects over time are then represented as a tree, which is constructed by an external procedure. This tree representation serves as input to the pattern mining algorithm, where we iterate over time by taking the clusters within each level of the tree in turn. We take this approach as the input tree provides a way to gauge what parameter values are suitable, as will be illustrated in the next example.

For the simple example in Fig. 4.9, the identification of suitable parameter values is trivial. To demonstrate how the input tree of clusters can be used to identify parameters, we now consider a more complex example, comprised of 20 trajectories moving over 15 time steps. The left hand side of Fig. 4.11 shows the input tree constructed by an external procedure which takes the list of cluster objects identified by our density clustering algorithm. We note the presence of a small number of trajectories that briefly separate from a larger cluster before re-merging, these are indicated as grey nodes in the tree. All of these small clusters contain no more than three trajectories at any one time. Due to the nature of the trajectories, such deviations are inevitable. By inspection it is immediately apparent that five significant branches exist within the tree, with almost all the clusters therein containing at least five trajectories. Moreover, these branches all exist for at least four time steps. Therefore, in order to capture the main trends in the data, we identify \( t_{\text{min}} = 4 \) and \( \text{MinPts} = 5 \) as a suitable set of parameter values. The smallest overlap between two consecutive clusters within the main branch is \( \theta = 0.6 \), however most contain an overlap of at least \( \theta = 0.75 \). This provides a set of suitable bounds on the value of the parameter \( \theta_{\text{min}} \).

Turning to time step \( t_{13} \), highlighted by the grey box, we observe that one branch briefly breaks down as the cluster containing five trajectories at the previous time step splits into two clusters with two and three trajectories, respectively. However, at the following time step, we see that these trajectories merge into a single cluster. Since they still exhibit a degree of correlation in time, we can set \( \delta_{\text{max}} = 1 \) to capture this phenomenon. The right hand side of Fig. 4.11 shows the output tree for this example, identifying five key relaxed moving clusters. Generally, the identification of parameters
may not be as straightforward for real trajectory data, and their specification should be viewed as an iterative process in which the initial tree of clusters provides a starting point. With $\delta_{\text{max}}$ providing the degree of relaxation allowed, to provide meaningful clustering, this value should be small in comparison to $t_{\text{min}}$. In the event it is not, the features selected may not be suitable for clustering. Of course, in some extreme cases, the trajectory data may simply be unsuitable for clustering.
Figure 4.11: A more complex example consisting of 20 trajectories over 15 time steps. The input cluster tree on the left hand side contains five branches, which are coloured accordingly. Grey nodes represent small clusters corresponding to short term deviations. At time $t_{13}$ a group of five trajectories briefly splits into two clusters (highlighted in the grey box). The identified relaxed moving clusters in tree form according to the specified parameters is shown on the right.

$MinPts = 5$

$\theta_{min} = 0.70$

$t_{min} = 4$

$\delta_{max} = 1$
4.4 Application to 1,3-Cyclohexadiene

We shall now apply the new algorithms presented to the set of 100 AIMCE trajectories first introduced in Chapter 3.2, which describe the electrocyclic ring-opening of the molecule 1,3-cyclohexadiene (CHD) to 1,3,5-hexatriene (HT). We perform the clustering on a seven-dimensional feature space. This contains the four previously identified internal coordinates, comprised of three dihedral angles $\phi_a$, $\phi_b$, and $\phi_c$, and the $C_1-C_6$ bond length, as discussed in Chapter 3.3.1. In addition to internal coordinates, we also include the populations of the three adiabatic electronic states. As discussed in Chapter 3.3.2, the data is normalised using a min-max normalisation. Although recall the data does undergo a series of scalings according to the discretization process used in our density clustering algorithm, introduced earlier in this chapter within section 4.2.3.

The resulting tree of clusters over time is inherently more complex than the two examples discussed in the previous section. However, it provides us with a set of reasonable parameter ranges as a starting point, summarised in the first row of Table 4.2. While there exist several key branches of the tree that consistently contain more than ten trajectories for at least 30 fs, there also exists a branch with significantly fewer trajectories. For this reason, we identify four trajectories as the lower bound of $MinPts$. By tuning the parameters within the ranges specified in the first row of Table 4.2, we identify three significant cases, corresponding to high, medium and low separation. These parameters yield seven, five, and four relaxed moving clusters, respectively. The allowed parameter values in each of these cases are summarised in rows two to seven of Table 4.2.

<table>
<thead>
<tr>
<th>$N_{MC}$</th>
<th>$t_{\text{min}}$</th>
<th>$MinPts$</th>
<th>$\theta_{\text{max}}$</th>
<th>$\delta_{\text{max}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>-</td>
<td>30-40 fs</td>
<td>4-10</td>
<td>0.6-0.75</td>
<td>4-8 fs</td>
</tr>
<tr>
<td>7</td>
<td>30-34 fs</td>
<td>4</td>
<td>0.50-0.60</td>
<td>6 fs</td>
</tr>
<tr>
<td>7</td>
<td>30 fs</td>
<td>4</td>
<td>0.625</td>
<td>8 fs</td>
</tr>
<tr>
<td>5</td>
<td>30-34 fs</td>
<td>5-12</td>
<td>0.60</td>
<td>4-6 fs</td>
</tr>
<tr>
<td>5</td>
<td>30-34 fs</td>
<td>4-12</td>
<td>0.65</td>
<td>4-6 fs</td>
</tr>
<tr>
<td>5</td>
<td>30 fs</td>
<td>4-10</td>
<td>0.70</td>
<td>4-6 fs</td>
</tr>
<tr>
<td>4</td>
<td>30-34 fs</td>
<td>4-12</td>
<td>$&gt;0.70$</td>
<td>4-6 fs</td>
</tr>
</tbody>
</table>

Table 4.2: The number of identified relaxed moving clusters, $N_{MC}$, for various sets of parameters (or parameter ranges) for the pattern mining algorithm. The first row lists the initial ranges of guess parameters from the input tree of spatial clusters over time.
In all cases, we generally identify 30-34 fs as a suitable range for the number of
time steps over which the moving clusters must exist, as evident in Table 4.2. With
MinPts = 4 and a relatively small overlap of $\theta_{\text{min}} \leq 0.625$, the highest separabil-
ity case of seven relaxed moving clusters is achieved. The corresponding tree of
relaxed moving clusters output from our algorithm can be seen in Fig. 4.12, with
the branches of the tree coloured according to their dominant state population and
the corresponding changes in geometry overlayed. Here we observe the majority of
trajectories evolving together within $M_1$ up until 34 fs. This corresponds largely to the
evolution of the wavepacket out of the Franck-Condon region on the $S_1(1B)$ state.
This then splits into two branches as the wavepacket bifurcates. The first branch
is $M_2$, comprised of trajectories that have decayed to the ground state. However,
at this point, separation of the ground state trajectories into the different product
channels is not achieved, and the trajectories within $M_2$ correspond to a range of
different ground state products. Upon decay to the ground state, the initial difference
in geometry between cZc-HT and hot CHD is small, essentially amounting to a subtle
variation in the $C_1-C_6$ bond length. Such differences are too subtle to be identified
by the clustering algorithm at this point in time, and the different ground state reaction
channels are indistinguishable based on the provided feature set. The second branch
corresponds to $M_3$, consisting of trajectories that remain on the excited $S_1$ state after
first passing the conical intersection. At later times, the adiabatic $S_1$ state gains 2A
electronic character, and the trajectories that remain on this state undergo a degree
of torsion around the central C–C bond, bending out of the plane. At 66 fs the
differences between ground state trajectories becomes more apparent, with $M_4$ and
$M_5$ separating out as vibrationally hot CHD and cZc-HT respectively. Moving cluster
$M_5$ continues to evolve in time, as the $C_1-C_6$ bond length shown in its representative
structure further increases. This again branches into $M_6$ and $M_7$ at 102 fs, these
represent cEc-HT and a cluster of trajectories that isomerise from cZc-HT to tZt-HT,
respectively.

The lower values of $\theta_{\text{min}}$ and MinPts in the highest separability case ($N_{NC} = 7$) are
required as the relaxed moving cluster $M_6$, containing cEc-HT trajectories, is small
and difficult for the algorithm to identify. Increasing the value of MinPts to above four
renders these trajectories undetectable and results in the medium separability case
($N_{NC} = 5$) in Table 4.2. It is worth noting that increasing the value of the overlap
above 0.65 to the range $0.625 < \theta_{\text{min}} \leq 0.70$, achieves the same result. This is due
to the fact that all other relaxed moving clusters are well defined, and the spatial
clusters they encompass generally have a higher degree of overlap than those that
form $M_6$. The corresponding tree output in the medium separability case ($N_{NC} = 5$) can be seen in Fig. 4.13, noting that $M_5$ does not split into two separate branches like in the high separability case ($N_{NC} = 7$). The trajectories contained within $M_5$ therefore correspond to the union of relaxed moving clusters $M_5$ and $M_7$, as in the high separability case. In the medium separability case ($N_{NC} = 5$), the cEc-HT trajectories that were contained in $M_6$ in the high separability case ($N_{NC} = 7$), are excluded from all relaxed moving clusters after 102 fs.

Increasing the value of $\theta_{\text{min}}$ to above 0.70 results in the lowest separability case where only four moving clusters are identified ($N_{NC} = 4$). At this point, the algorithm is incapable of detecting early separation of ground state trajectories defined by moving cluster $M_2$. For this reason, we exclude the low separability case from further analysis. In each case, a large number of trajectories still remain in $M_3$ at 150 fs. For a longer duration simulation, these trajectories would eventually decay onto the ground state via the 2A/1A conical intersection and result in different ground-state products.
Figure 4.12: The tree of output moving clusters identified by the high separability set of parameters identified in Table 4.2. The times at which moving clusters split into new branches are indicated in red, and the dominant population of the moving cluster is indicated by the blue and yellow edges. The set of structures shown correspond to the geometries at the cluster centres.
Figure 4.13: The tree of output moving clusters identified in the medium separability set of parameters provided in Table 4.2. Note the absence of a moving cluster that contains the small amount cEc-HT product. The other moving clusters are almost unchanged compared to the higher separability case, and are all detectable up to $\theta_{\text{min}} = 0.70$. 
Focusing on the highest separability case ($N_{NC} = 7$), we now turn further analyse the changes in the size and populations of the identified relaxed moving clusters. The size of each relaxed moving cluster over time can be seen in the top panel of Fig. 4.14. In addition, the change in population between the $S_0$ and $S_1$ states is given in the bottom panel of Fig. 4.14. The solid circles represent the population of the states at the time of initialisation, and the empty circles represent the populations at the time each relaxed moving cluster is terminated. As expected, virtually all trajectories are contained within $M_1$ before 25 fs, which observes an $S_1$ (1B) population equal to one. At 25 fs, the trajectory ensemble observes a maximum in the $S_2$ population$^2$ as trajectories pass through the 1B/2A conical intersection, combined with a marked increase in $S_0$ population at 25-35 fs. After 25 fs, the size of $M_1$ decreases, reaching just under 80 trajectories at 30-34 fs. While the trajectories within $M_1$ are still predominately $S_1$ in character, some of them are on the diabatic 2A state, which still corresponds to the $S_1$ state in the adiabatic representation of states employed in the simulation.

The decrease in the size of $M_1$ and $S_1$ state after 25 fs results from a small number of trajectories continuing on the 1B state (adiabatic $S_2$ after 25 fs) state or decaying rapidly to the $S_0$ (1A) state. At this point in time, the small number of trajectories that begin to populate the $S_0$ state are undetectable. Due to the rapid acceleration of the wavepacket towards the 2A/1A conical intersection by the steep gradients at a cusp in the PES, the trajectories are left largely unaffected by the 2A state and hence observe little change in geometry. This contributes to the difficulty the algorithm faces in identifying the trajectories that rapidly decay to the $S_0$ state at early times. Moreover, the algorithm cannot distinguish the trajectories on the $S_2$ state due to both its short lifetime and the fact that the overlap between consecutive clusters is consistently small. The latter results from the fact that there is a steady flux of trajectories between the $S_2$ and $S_1$ states. Arguably, these shortcomings are a result of the representation of states employed in on-the-fly trajectory based simulations, rather than the pattern mining algorithm itself.

At 34 fs, the group of trajectories rapidly decaying through the 2A/1A conical intersection then becomes detectable as $M_2$, which is initialised with near equal $S_1$ and $S_0$ population, as indicated by the orange point in the bottom panel of the figure. From 34-56 fs the size of $M_2$ increases from approximately 10 to 35 trajectories, with a marked increase in its $S_0$ population, indicated by the empty orange circle in the bottom panel of the figure. Meanwhile, $M_3$ decreases from around 70 to 45 trajectories, as

---

2. Seen in Fig. 3.2 within Chapter 3.2.
4.4 Application to 1,3-Cyclohexadiene

more trajectories pass through the 2A/1A conical intersection and move from $M_3$ to $M_2$. These relative changes are commensurate with the decay of $S_1$ population and levelling off of $S_0$ population at around 50-60 fs in Fig. 3.2 (Chapter 3.2). Drawing attention to moving cluster $M_3$ in the region of 60-75 fs, we see that its size then begins to increase again until levelling off of around 55-60 trajectories. This is due to the aforementioned undetectable trajectories that continue to evolve adiabatically on the $S_2$ state before decaying through the 1B/2A conical intersection, resulting in their assignment to $M_3$. Following this point, $M_3$ is completely defined by the 2A state, explaining the minor change in $S_1$ population seen in the bottom panel of the figure.

Moving to the remaining set of ground state moving clusters, $M_4$ which contains trajectories that return to hot CHD, consistently contains 12-14 trajectories and is generally the most stable moving cluster. Moving cluster $M_5$, corresponding to trajectories that originally emerge as cZc-HT (prior to isomerisation), contains 15-20 trajectories except from at 80 fs where it briefly drops to 10 trajectories due to some small deviations in the clusters. This trajectory set then splits between $M_6$ and $M_7$, the former of which corresponds to a cEc-HT pathway and initially consists of only 4 trajectories. Generally, as seen in the bottom panel of the figure, moving clusters $M_4 - M_7$ observe little change in their relative populations over the time they exist as they define trajectories that have settled into a range of ground state isomers. On average, the cEc-HT trajectories within $M_6$ reach a point where their $S_0$ population exceeds their $S_1$ population at 36 fs. However, despite the fact the ground state is dominant, the magnitude of the $S_1$ population is notably higher for $M_6$ than the other ground state clusters\(^3\).

---

3. Recall that the nuclear wavefunction evolves on a superposition of all electronic states in AIMCE. This is performed in a mean-field fashion, as discussed in Chapter 2.2.2
Figure 4.14: (Top panel) Number of trajectories in each moving cluster $M_1$-$M_7$ as a function of time (fs). The dashed line segments in the red line representing cluster $M_6$ indicates the times when the moving cluster briefly breaks down. (Bottom panel) The initial (solid circle) and final (empty circle) $(x, y) = (S_0, S_1)$ population for each cluster, with the colours of the clusters the same as in the top panel.
4.4 Application to 1,3-Cyclohexadiene

Figure 4.15: Changes in $S_0$ (left panel) and $S_1$ (right panel) populations for the trajectories re-assigned from cluster $M_3$ to $M_2$. The $x$-axis in the left (right) panel shows the $S_0$ ($S_1$) population at the time the cluster $M_2$ is initialised ($t_1=34$ fs), and the $y$-axis the corresponding $S_0$ ($S_1$) population at the time when cluster $M_2$ is terminated ($t_2=66$ fs). That most points are above the $y=x$ line in the left panel, and below in the right panel, indicates a net population transfer $S_1 \rightarrow S_0$. For each trajectory, the time at which the transition between the moving clusters occurs is indicated by the color of the data point, according to the colour mapping shown below the two panels.

To better understand the re-assignment of trajectories from moving cluster $M_3$ to $M_2$, we turn to Fig. 4.15, which shows the changes in populations for each trajectory that undergoes a transition between the two relaxed moving clusters. In both panels, the $x$-axis corresponds to the population of the state at the time which $M_2$ is initiated (34 fs), and the $y$-axis the population at the time which $M_2$ is terminated (66 fs). Points above the $y=x$ line therefore reflect an increase in the respective state population, and those below reflect a decrease in population. In the left panel, with the exception of one trajectory, every trajectory has a larger magnitude $S_0$ population following its transition to $M_2$, while the right panel shows a significant decrease in $S_1$ population, corresponding to a net $S_1 \rightarrow S_0$ population transfer. Thus, the algorithm is capable of detecting how trajectories transition between electronic states and can redistribute trajectories between relaxed moving clusters defined by electronic state and nuclear dynamics.
Figure 4.16: Distribution of the three dihedral angles $\phi_a$, $\phi_b$, and $\phi_c$ over the times each relaxed moving cluster exists. Each row represents one of the relaxed moving clusters corresponding to a leaf node in Fig. 4.12, and each column the dihedral angle. In each case, the set of trajectories plotted correspond to those that are contained within the moving cluster for at least 90% of its duration. Here, the regions of highest density within the clusters are clear, as indicated by a higher concentrate of points, thus observing a darker shade.
To further understand the distribution of nuclear geometries within each relaxed moving cluster we turn to Fig. 4.16. Here, the columns correspond to the three key dihedral angles, and the rows to each relaxed moving cluster that corresponds to a leaf node in Fig. 4.12. A leaf node refers to a node with no child nodes, i.e. $M_3$, $M_4$, $M_6$, and $M_7$. Each of these therefore represent one of the unique channels that the system settles into towards the end of the simulation. While $M_3$ corresponds to trajectories that remain on the excited state and are yet to decay, the other three moving clusters represent a unique product channel. Each panel of the figure shows the angles that the core trajectories of a moving cluster observes, where we define a core trajectory as those that are contained within the cluster for at least 90% of its lifetime. Turning to the largest moving cluster $M_3$, we see that the values of the dihedral angles $\phi_a$ and $\phi_b$ are narrowly distributed around 50 degrees at 34 fs. However, as time evolves, these quickly begin to decrease towards 0 degrees, undergoing several small oscillations and dispersing, as indicated by the broadening band of highly concentrated blue points. In addition, there appears to be a small number of trajectories that lie just outside this high density band, deviating slightly from the general oscillatory motion. In contrast, the dihedral angle $\phi_b$ is narrowly distributed just below zero degrees at 34 fs, with the trajectories showing an initially coherent increase in $\phi_b$ before dispersing. The dispersion of $\phi_b$ after approximately 90 fs is commensurate with the trajectories spreading out a near flat portion of the excited 2A ($S_1$) PES, where the majority of trajectories observe values in the approximate range $75 \leq \phi_b \leq 150$ at 150 fs.

Turning to the set of ground state product channels, in moving cluster $M_4$, all three dihedral angles are more narrowly distributed, with most trajectories undergoing oscillations that are relatively low in frequency. Although, there exists a small number of trajectories that fall outside the highest density regions of the plot, observing slightly different oscillatory behaviour, these still the general trend. As the degree of ring strain reduces and the trajectories converge back towards hot CHD, the dihedral angle $\phi_b$ reduces from approximately 75 degrees at 66 fs, to around 0 degrees at 150 fs. In moving cluster $M_6$, corresponding to trajectories that form cEc-HT, the large dihedral angles $\phi_b$ are immediately apparent, reaching values exceeding 200 degrees while the other dihedral angles remain small. As previously discussed, this moving cluster is largely comprised of only four trajectories, making its identification difficult. While the small number of trajectories therein may separate earlier on, it is not feasible for the algorithm to identify this. Finally, moving cluster $M_7$, observes a relatively broad distribution of angles. Here, $\phi_b$, the narrowly distributed of all angles, is consistently small. In contrast, both $\phi_a$ and $\phi_c$ show a relative increase in time, as the trajectories
therein isomerise from cZc-HT towards a range of trans-HT isomers. While the highest density part of the cluster corresponds to a near tZt-HT geometry, as previously shown by the overlaid structures in Fig. 4.12, both angles observe significant broadening, encompassing a range of tZt-HT and cZt-HT structures.

Figure 4.17: A series of reduced clusters constructed from the core trajectories within the moving clusters corresponding to leaf nodes. Core trajectories refer to those that are contained within their respective moving cluster for at least 90% of their total duration. The top left and right panels show the C\(_1\)–C\(_6\) bond and the central dihedral angle \(\phi_b\), whereas the bottom left and right panels show the dihedral angles \(\phi_a\) and \(\phi_c\).

We now turn to address how one can identify a reduced sample of trajectories based on the identified relaxed moving clusters. To generate a suitable set of trajectories to sample from, we first consider the core trajectories assigned to the moving clusters that correspond to a leaf node. Taking the set of core trajectories over the whole length of the simulation and not just over the times their respective relaxed moving cluster exists, we construct a set of four spatio-temporal clusters that are continuous in time. The centre of each of these reduced spatio-temporal clusters, along with
their distribution, is plotted in Fig. 4.17 as a function of the four key internal coordinates. These continuous spatio-temporal clusters are labelled according to the relaxed moving cluster they are generated from. Here we see that the trajectories are remarkably similar until around 50 fs. At this point, subtle differences in the C$_1$–C$_6$ bond length emerge and the trajectories within $M_6$ observe a larger magnitude oscillation of both $\phi_a$ and $\phi_b$ dihedral angles. At later times, the difference in C$_1$–C$_6$ bond length between ring-opened and ring-closed trajectories becomes apparent. We can also see the broadening of dihedral $\phi_b$ for excited state trajectories within $M_3$, as previously discussed in the context of Fig. 4.16. Here, the highest density of trajectories is observed just below 100 degrees at 150 fs. A considerable degree of dispersion is also seen for the trajectories contained within $M_7$, where the dihedral angles $\phi_a$ and $\phi_b$ increase as the cZc-HT product isomerises to cZt/tZt-HT. In addition, the small set of core trajectories within $M_6$ form a narrow distribution of cEc-HT product. It is worth noting that while the trajectories in $M_6$ undergo a larger magnitude oscillation around 50 fs, our pattern mining algorithm does not separate out the cEc-HT pathway until 102 fs, as previously discussed. This is due to the fact the trajectories largely overlap around 50 fs, and based on the internal coordinates and populations alone, the separability of the data is poor here. Without additional information, or exploring alternative ways to represent the nuclear geometries, the subtle differences in these trajectories cannot be detected earlier than 102 fs.

With the continuous set of spatio-temporal clusters in Fig. 4.17, one can then identify a reduced set of trajectories that capture the evolution of the ensemble in terms of both populations and nuclear geometry. This subset can then be used to perform highly accurate observable calculations that are too computationally demanding for the whole ensemble of trajectories. In the simplest approach, this could correspond to taking the trajectories that are located at each of the cluster centres, however this will not reflect the quantum yield of each product channel. In some circumstances, this may still be a viable option, for example in providing a quick first-order check of the potential different contributions to an experimental observable, or for use in the forward optimisation based approach to inversion that we shall discuss in the subsequent chapter. In a more accurate approach, one could sample from each of the distributions plotted in Fig. 4.17, taking the minimum number of trajectories required to maintain the correct quantum yield. Providing a suitable clustering result is achieved, this trajectory subset can then be used in place of the ensemble in highly accurate observable calculations.
4.5 Periodicity of Angular Coordinates

Following submission of this thesis, it was pointed out that the density kernel in Eqn. 4.1 does not directly account for periodicity. In Chapter 3, we saw that clustering using metrics that explicitly account for periodicity changes the results very little. When analysing the reason for this, we found that the characteristics of the CHD data set used are such that periodicity has little impact on the clustering. While we use the same CHD data in the current chapter, and therefore anticipate little effect on the clustering from periodicity, we nevertheless undertake to evaluate the effect on the results. This is achieved by projecting each dihedral angle onto its pair of \((\cos, \sin)\) coordinates on the unit circle. In the previous chapter, it was found that this approach produced almost identical results to the alternative approach of using the minimum arc length between pairs of angles. In the discrete representation of the density used in this chapter, the former approach is more convenient as it does not require the density kernel to be redefined to utilise arc lengths.

Earlier in the current chapter, in section 4.4, we found that performing spatial clustering followed by pattern mining resulted in the identification of up to seven spatio-temporal clusters. Three distinct cases labelled as high, medium, and low separability were reported depending largely on the value of the minimum overlap threshold \(\theta_{\text{min}}\). The highest separability case with seven relaxed moving clusters was identifiable with \(\theta_{\text{min}} \leq 0.625\), whereas the medium separability case with five relaxed moving clusters was identifiable up to \(\theta_{\text{min}} \leq 0.70\). The low separability case with only four moving clusters was excluded from further analysis. In the high separability case of Fig. 4.12 we were able to distinguish the very low probability cEc-HT pathway \(\langle M_7 \rangle\) that emerged on the ground state at later times. However, in the medium separability case of Fig. 4.13, this cEc-HT pathway was rendered undetectable. In both these cases, the initial bifurcation of the approximate model wavepacket following non-adiabatic population transfer from the \(S_1\) state to \(S_0\) was identified at 34 fs. Moreover, the trajectories that return to hot CHD and those that proceed to form a range of ground state HT isomers became separable at 66 fs.

Upon rerunning the clustering and pattern mining algorithms by projecting the angles onto their coordinates \((\cos, \sin)\) we find we that the spatio-temporal patterns become more challenging to identify, with the already difficult to identify low probability cEc-HT pathway becoming undetectable. However, crucially, the same key trends are identifiable with \(\theta_{\text{min}} = 0.50\). The resulting tree of relaxed moving clusters is shown.
4.5 Periodicity of Angular Coordinates

in Fig. 4.18, where the overlayed structures correspond to coordinates at the cluster centres for representative time-steps at their initialisation and termination. Furthermore, the dominant population of each cluster is represented by the colour of the edges, with blue representing the $S_0 (^1A)$ ground state and yellow the adiabatic $S_1$ state that corresponds to the diabatic $^1B$ and $^2A$ states at early and late times respectively. The times at which the relaxed moving clusters are initialised and terminated are labelled in red at each node. Importantly, this result is near-equivalent to the previously reported medium separability case in Fig. 4.13. Some minor differences show up, for example the splitting of $M_1$ into $M_2$ and $M_3$, resulting from some trajectories decaying to the ground state, now occurs at 36 fs instead of 34 fs. Furthermore, the splitting of $M_2$ into $M_4$ and $M_5$ as the ground state product channels become separable now occurs at 76 fs instead of 66 fs. Finally, the relaxed moving cluster $M_5$, corresponding to the HT ground state product channel that isomerises towards tZt-HT, is detectable up to 142 fs.
Figure 4.18: The tree of output moving clusters identified by projection of the angles onto their (cos, sin) coordinates on the unit circle. Note the absence of a moving cluster that contains the small amount cEc-HT product, but overall this result is similar to that previously reported in Fig. 4.13.
Figure 4.19: (Top panel) Number of trajectories in each moving cluster $M_1-M_5$ as a function of time (fs). (Bottom panel) The initial (solid circle) and final (empty circle) $(x,y) = (S_0,S_1)$ population for each cluster, with the colours of the clusters the same as in the top panel.
Turning to Fig. 4.19, we now analyse the changes in the size and populations of the five identified relaxed moving clusters. Here, the top panel shows the number of member trajectories in each cluster over the times it exists, whereas the bottom panel shows the change in population between the $S_0$ and $S_1$ states with the solid and empty circles representing the population at the time of initialisation and termination respectively. In terms of moving clusters $M_1$-$M_4$, the size and relative populations of each moving cluster are largely the same as reported previously in Fig. 4.14. In both cases, the majority of the trajectories are contained within $M_1$ initially as the model wavepacket moves out of the Franck-Condon region, the moving cluster $M_2$ then emerges and continues to grow in size as an increasing number of trajectories decay to the ground state. Again, $M_2$ observes a net decrease in $S_1$ population and an increase in $S_0$ population as before. The moving cluster $M_3$ containing the trajectories that remain trapped on the excited $S_1$ state continues to have a dominant $S_1$ population. As previously seen, this moving cluster remains fairly stable and consistently contains on the order of 55-60 trajectories. In section 4.4, we explained that the decrease in the size of $M_3$ at earlier times followed by an increase to around 60 trajectories was due to an undetectable branch of trajectories that continue to evolve adiabatically on the $S_2$ state before decaying through the 1B/2A conical intersection. This feature is also observed in Fig. 4.19. Looking at moving cluster $M_4$, which corresponds to the pathway that returns to hot CHD, we see it consistently contains on the order of 10-15 trajectories as previously seen in Fig. 4.14. Moreover, $M_4$ again corresponds to trajectories with dominant $S_0$ character, with the composition changing little over time. However, we note that there are a small number of time steps at which the cluster breaks down before reforming again. In comparison to Fig. 4.14, moving clusters $M_6$ and $M_7$ are not observed as they result from the splitting of $M_5$ at 102 fs. As the cEc-HT is undetectable here, in Fig. 4.19, $M_5$ exists over a longer time period. Previously we saw that moving clusters $M_5$-$M_7$ were more consistent in terms of the number of trajectories they contained at any one time step. Here, moving cluster $M_5$ fluctuates to a greater degree, observing a decrease in size followed by a subsequent increase again. However, this moving cluster still corresponds to a pathway that isomerises towards tZt-HT from cZc-HT when looking at the most strongly correlated trajectories.

In Fig. 4.20 we plot the changes in $S_0$ (left panel) and $S_1$ (right panel) populations for trajectories that transition from moving cluster $M_3$ to $M_2$ as they decay to the ground state. In both panels, the $x$-axis corresponds to the population of the state at the time which $M_2$ is initialised (36 fs), and the $y$-axis the population at the time $M_2$ is
4.5 Periodicity of Angular Coordinates

Figure 4.20: Changes in $S_0$ (left panel) and $S_1$ (right panel) populations for the trajectories re-assigned from cluster $M_3$ to $M_2$. The $x$-axis in the left (right) panel shows the $S_0$ ($S_1$) population at the time the cluster $M_2$ is initialised ($t_1$=36 fs), and the $y$-axis the corresponding $S_0$ ($S_1$) population at the time when cluster $M_2$ is terminated ($t_2$=76 fs). That most points are above the $y=x$ line in the left panel, and below in the right panel, indicates a net population transfer $S_1 \rightarrow S_0$. For each trajectory, the time at which the transition between the moving clusters occurs is indicated by the color of the data point, according to the colour mapping shown below the two panels.

terminated (76 fs). As previously discussed in Fig. 4.15, that the majority of points are above the $y=x$ line in the left panel reflects the fact the trajectories undergo a net increase in $S_0$ population as they transition from $M_3$ to $M_2$. Furthermore, in the right panel, we see the majority of points are again below the $y=x$ line, reflecting a net decrease in $S_1$ state population.

Finally, in Fig. 4.21 we plot the distribution of each relaxed moving cluster corresponding to a leaf node in the tree of Fig. 4.18, thus representing one of the unique product channels the system settles into at later times. Importantly, we see that the geometries within $M_3$, $M_4$, and $M_5$ here are in good agreement with those previously seen in Fig. 4.16, with $M_3$ containing trajectories that remain dispersed on the excited $S_1$ state and rotate around $\phi_b$, while $M_4$ corresponds to those that relax back to hot CHD. Again, recall that $M_5$ here corresponds to $M_7$ in the previously reported result.
However, here $M_5$ is initialised earlier at 76 fs as the splitting of the cEc-HT pathway is not detected. At around 76 fs the two dihedral angles $\phi_a$ and $\phi_c$ are distributed around 0-50 degrees, these then increase as time evolves and isomerisation from cZc-HT towards tZt-HT occurs.

![Graph of dihedral angles](image)

**Figure 4.21:** Distribution of the three dihedral angles $\phi_a$, $\phi_b$, and $\phi_c$ over the times each relaxed moving cluster exists. Each row represents one of the relaxed moving clusters corresponding to a leaf node in Fig. 4.18, and each column the dihedral angle. In each case, the set of trajectories plotted correspond to those that are contained within the moving cluster for at least 90% of its duration. Here, the regions of highest density within the clusters are clear, as indicated by a higher concentrate of points, thus observing a darker shade.

To conclude our discussion on the treatment of periodicity in this Chapter, we note that while projecting the angles onto their ($\cos, \sin$)-coordinates reproduces the same qualitative trends, the ground state channel is now distinguishable at 36 fs compared to 34 fs previously, and that the hot CHD and HT products are distinguishable at 76 fs compared to 66 fs previously. Generally speaking, the spatio-temporal clusters become more difficult to identify, although the results are highly similar to the medium
4.5 Periodicity of Angular Coordinates

separability case previously seen in Fig. 4.13. Overall, the differences are rather small and confirm that the results are reasonably robust upon inclusion of periodicity for this particular data set. That the changes, however minor, are greater for the method evaluated in this chapter compared to the time-series method in Chapter 3 is unsurprising given that the clustering of the entire time-series provides greater stability to local deviations in the distance metric. Furthermore, in this chapter we aim to tackle a higher dimensional problem on inclusion of state populations.

As discussed in section 3.5 of the previous chapter, although the data set includes angles that span a range larger than $[0, \pi]$, the unique reactive pathways that are identified are all separable with respect to more than one independent feature. While projecting the angles onto $(\cos, \sin)$ coordinates directly accounts for periodicity, increasing the dimensionality while the number of data points remains the same may make it more difficult to identify a sensible clustering result as the volume of space the data spans increases exponentially with the number of dimensions. Moreover, introducing two correlated features per angle may risk biasing the result towards periodic features when used in combination with non-periodic features. As previously stated, implementing a density kernel to include arc lengths for the angles is less straightforward in the discrete representation used here. However, we are now in the process of implementing a continuous representation of the density field, which we favour. In the future, we aim to incorporate into this a kernel that uses arc lengths and to evaluate the effect this has compared to the projection of the angles onto $(\cos, \sin)$-coordinates. In applications to high-dimensional data sets that include a range of both non-periodic and periodic features, we anticipate that this will perform better.

4.6 Conclusions

In summary, this chapter has presented two new algorithms, which together provide a means of mining the spatio-temporal patterns present within ensembles of trajectories. The algorithms have been applied to trajectories describing the electrocyclic ring-opening dynamics of CHD. Compared to the time series approach presented in Chapter 3, the new pattern mining approach is shown to be capable of providing a more in-depth understanding of the dynamics. It provides not just information on which trajectories contain similar molecular geometries over all times, but also shared patterns, the times at which groups of trajectories diverge, and captures the changes in electronic population over time. The new density clustering algorithm, applied at
4.6 Conclusions

Each time step, provides an adaptive clustering that reflects the probability density of the wavepacket. The benefit of this approach is that it can identify the core regions of density in overlapping data, can detect nonconvex and nonlinearly separable clusters, and can identify cluster of highly variable density and size without the need for re-parameterisation at each time step. However, in comparison to existing algorithms that are designed to be applicable to a wider range of data types, it does sacrifice some scalability in favour of performance in the specific case of QMD trajectory data.

The generated clusters are then used as input for a novel pattern mining algorithm. While this pattern mining algorithm is based on the idea of a moving cluster, it is not subject to the consecutive time constraint. Redefining a spatio-temporal cluster as a relaxed moving cluster, which includes an additional parameter $\delta_{\text{max}}$, we are able to account for brief deviations between trajectories, capturing the presence of trajectories that are both strongly correlated in time and less so. Moreover, the unique tree representation which we employ allows one to easily identify the times at which groups of trajectories diverge, describing the phenomenon of wavepacket bifurcation.

In our original application of these algorithms to CHD, we were able to detect the presence of seven unique relaxed moving clusters, each describing a unique evolution shared by groups of trajectories. This disclosed the presence of a series of product channels, corresponding to vibrationally hot CHD, cEc-HT, and cZc-HT, which ultimately isomerises to a range of cis/trans isomers. In addition, a broad distribution of trajectories that remain on the excited $S_1$ (2A) state and exhibit a high degree of out of plane torsion were observed. For a longer duration simulation, these longer-lived excited state trajectories would eventually decay to form ground state products.

As well as capturing the changes in nuclear geometry, the pattern mining algorithm identified that the excited wavepacket initially bifurcates at around 34 fs, on the first pass at the 2A/1A conical intersection. This is in agreement both with the observed populations of the ensemble, other simulations, and experiments. While identification of the small amount of ground state cEc-HT products was difficult and required a lower value of the overlap threshold, all other moving clusters exhibit a larger degree of stability and were detectable with higher threshold values.

While following submission of this thesis it was pointed out that the definition of the density kernel does not explicitly account for periodicity, we found that when accounting for this by projecting each angle onto its $(\cos, \sin)$ coordinates on the unit circle, all reactive pathways excluding the low probability cEc-HT pathway were detectable. As also addressed in Chapter 3, the distribution of the CHD data set used
4.6 Conclusions

in this application is such that the results have a high degree of similarity whether periodicity is explicitly accounted for or not. However, we are grateful to the examiners of this thesis for pointing out potential issues with periodicity as we hope to further develop the methods introduced here into a generally applicable tool for the QMD community to employ.

Looking forward, further improvements to the clustering algorithm can be envisioned. At the time of writing, we are in the process of implementing a continuous definition of the density. In this new approach, we shall look to identify the best approach to dealing with periodicity in QMD data sets comprised of a mix of periodic and non-periodic features. Thus helping us towards our goal of developing a general pattern mining tool for the community. Furthermore, the identification of spatio-temporal patterns could be improved by the inclusion of more information in the clustering algorithm, such as momenta or velocities. However, in the current approach where we include geometric information in the form of internal coordinates, extending the feature set to include momenta is non-trivial. This would require one to map the Cartesian representation of momenta used in the simulation to internal coordinates. Alternatively, one could choose a representation of nuclear geometries that is better suited for the mapping of momenta. While we have seen that this approach can be applied to a rather complex set of trajectories with several coupled and broadly distributed degrees of freedom, the applicability to other types of reactions such as photodissociation and charge transfer should be explored. In other systems, depending on how active each of the degrees of freedom are in the dynamics, more geometric information may have to be included in the feature space. At some point, it becomes infeasible to include a large number of internal coordinates in the set of features, future work should therefore explore alternative representations of the geometries. Perhaps, leveraging methods that are capable of constructing a reduced dimensionality representation of the feature space, encoding the the important characteristics within it. Although this encoding space would have to be valid over the whole duration of the trajectory simulation.
Chapter 5

Inversion of Time-Resolved Data via Forward Optimisation

5.1 Introduction

The standard approach to the interpretation of time-resolved experiments involves a side-by-side comparison with theory, in which the time-evolution of an observable $X$ is predicted from the molecular wavefunction as solutions to a set a physically motivated equations through a forward mapping,

$$|\Psi(t)\rangle \xrightarrow{M_j} X^j(t),$$  \hfill (5.1)

where $M_j$ is the mapping function and the index $j$ denotes the type of experiment. The form of the mapping will vary depending on the type of experiment, theoretical approximations, and computational frameworks. For example, in the case of TRPES, the photoelectron spectra may be predicted from approximate Dyson orbital calculations [162, 176], or from more accurate R-matrix methods [163]. For scattering based techniques, the scattered intensity may be predicted simply from the IAM [130], or using more sophisticated ab-initio total scattering calculations [87]. Typically, observables are predicted using a trajectory model of the wavefunction, using either the ensemble average, or a small sample of key trajectories selected on intuition. Usually, the solutions are well defined and allow for a qualitative analysis of the experiment in which different contributions to the signal can be inferred from theory. However, such an approach is highly dependent on the accuracy of the simulation data, and thus in order to provide a useful interpretation of experiment, one must focus on ensuring that the simulation is correct. Furthermore, despite these efforts, agreement between
5.1 Introduction

theory and experiment is not guaranteed due to approximations inherent in simulation, the calculation of the observables, and the constraints placed on the observable by a given experimental setup. This leaves little recourse if agreement between theory and experiment is not reached.

The inverse problem is fundamentally different from forward mapping and aims to retrieve from the experimental signal the key molecular characteristics that give rise to the observed signals. This is called inversion. Quite generally, the inverse problem deals with determining the set of model parameters $p$ that reproduce a given observation $X_j$ such that,

$$X_j = M_j(p).$$

Inverse problems are ubiquitous in science and far extend the applications considered in this chapter. Important applications can be found in engineering [208], material characterisation [79], and medical imaging [80, 81]. In most cases, including those addressed in this chapter, the forward map $M_j$ is nonlinear. In comparison to the forward case, the inverse problem is mathematically ill-defined, suffering from instability and a lack of unique solutions. Typical approaches to solving the inverse problem include regularization, optimisation methods, stochastic and probabilistic methods, and machine learning.

In the case of molecular structure determination, the inverse problem is essential. Here the molecular structure is typically under determined by the experimental data, and so the structure is determined by the iterative refinement of a model subject to auxiliary constraints. Examples of this include the fitting of spectroscopic data to a Morse oscillator, or Dunham coefficients in the case of diatomic [209]. Techniques such as x-ray crystallography, NMR, and electron diffraction all involve the refinement of the observed structure using molecular geometry constraints from supplementary experimental and/or computational data [42, 210]. While the constraints on near equilibrium ground state structures are well known, for excited state dynamics such constraints are much harder to define. In this chapter, we shall explore a solution to this problem. In principle, the introduced method can be applied to a range of experiments, such as TRPES, x-ray and electron scattering, and Coulomb explosion imaging (CEI), although we shall focus predominantly on the case of time-resolved scattering. In this case, a minimal inversion model would consist of a sequence of continuous structures, while a complete model would capture the full evolution of the molecular wavefunction, including all nuclear and electronic distributions and state populations.
5.1 Introduction

Inversion algorithms offer the potential to decouple experiment and theory, providing insight into what range of theoretical models are commensurate with the experiment, allowing a more quantitative analysis and the ability to identify discrepancies between the two.

The inversion of experiments that track excited state dynamics is challenging enough that it is usually not attempted. However, there has been considerable growth in their development in recent years. One notable example involves the use of a recurrent neural network (NN) to invert time-resolved scattering data, revealing a series of interatomic distances while ensuring all continuity and conservation constraints are fulfilled [82]. Another approach utilises a convolutional NN, capable of reconstructing the 3D structure of a molecule from electron diffraction data [211]. However, this has only seen success in static applications, although the authors comment on extending the approach to time-resolved inversion, the network does not inherently ensure that the continuity and conservation constraints are met. Other approaches include optimisation based methods, such a Monte-Carlo approach to fitting sampled structures to detector images [83, 212]. However, this requires a large number of sample structures and continuity is not guaranteed. Within the UED community, genetic algorithms have been employed to invert the signal via consecutive in silico mutations [213]. In addition, there exist approaches that incorporate phase retrieval algorithms, aiming to reconstruct the 3D coordinates of static molecular structures [214–217], although this require at least partial alignment of the molecular ensemble.

The development of a diverse range of inversion algorithms is crucial in providing the ultrafast imaging community with the means to invert different types of experiments. The following sections of this chapter detail a new optimisation based approach to inversion, which exploits that the forward mapping is well defined. We define the optimisation in the basis of trajectories, providing a set of reasonable constraints on far from equilibrium structures while accounting for continuity constraints. In essence, the procedure involves the optimisation of a model wavefunction in light of given experimental constraints, with the aim of determining the set of parameters that reproduce the experimental observable accurately. The work herein loosely follows the manuscript "Robust Inversion of Time-Resolved Data via Forward-Optimisation in a Trajectory Basis", previously published in the Journal of Chemical Theory and
Computation [218]. The purpose of this chapter is to provide a general framework within which this approach can be applied to different observables, establish the best practice for performing the optimisation procedure, and to critically evaluate its performance specifically in relation to scattering observables.

5.2 General Optimisation Framework

We proceed to tackle the inverse as a forward optimisation problem, where we aim to find a model wavefunction \( \Psi_{\text{mod}}(t) \) that yields a predicted observable \( \bar{X}_j(t) \), which reproduces the measured experimental observation \( X_{\text{exp}}(t) \). The prediction of the model observable is comprised of two main steps,

\[
\Psi_{\text{mod}}(t) \xrightarrow{M_j} X_{\text{mod}}^j(t) \xrightarrow{S_j} \bar{X}_j(t),
\]

where the first step corresponds to the forward mapping as previously introduced in Eqn. 5.1, and the second step is an additional apparatus mapping step \( S_j \) that reproduces the effect of the experimental apparatus on the observable. For example, accounting for the effects of observed temporal or spatial resolution and other experimental distortions. Such issues with the experimental data are inevitable despite the post-processing that the data typically undergoes. We define the parameterised model wavefunction as,

\[
\Psi_{\text{mod}}(t) = \sum_{n=1}^{N_{\text{TBF}}} w_n(t) c_n(t) |\psi_n(t)\rangle,
\]

where \( c_n(t) \) are the expansion coefficients, and \( w_n(t) \) the weights that rescale the trajectory basis functions (TBFs) \( |\psi_n(t)\rangle \) to bias the theoretical model towards the experiment, subject to the normalisation condition \( \langle \Psi_{\text{mod}}(t) | \Psi_{\text{mod}}(t) \rangle = 1 \). The wavefunction for each of the \( N_{\text{TBF}} \) total TBFs is given by,

\[
|\psi_n(t)\rangle = \left[ \sum_{k=1}^{N_s} a_k^n(t) |\psi_k\rangle \right] |\chi_n(t)\rangle.
\]

Here the square brackets contain each of the \( N_s \) electronic states \( |\psi_k\rangle \) and their coefficients \( a_k^n(t) \), which correspond to the their excited state populations as \( |a_k^n(t)|^2 \). The nuclear basis functions \( |\chi_n(t)\rangle \) describe the evolution of the nuclei in time, and follow phase-space coordinates \( (R_n(t), P_n(t)) \) where \( R_n(t) \) and \( P_n(t) \) are the nuc-
lear positions and momenta, respectively. Ultimately the form of the nuclear basis functions, their equations of motion, populations, and additional coefficients such as wavepacket widths, will vary between propagation methods. In methods such as AIMCE, AIMS, and MCG, \( \chi_n(t) \) takes the form of a Gaussian [113, 118, 120], whereas in SH it simply reduces to a \( \delta \)-function [110]. For SH and AIMS, only one electronic state is populated at any given time, whereas in MCE based methods a given nuclear basis function can occupy several states at once [115, 116, 118, 119]. In AIMS, the number of TBFs increases in time due to spawning/cloning [113, 219]. While the TBFs may generally be computed from any of these methods, in this thesis we focus on the use of TBFs propagated using AIMCE and SH, which were introduced in detail in Chapter 2.2.

In this framework, the model observable predicted from the mapping of the model wavefunction is iteratively refined until the error between the model and experiment is minimised, according to the target function,

\[
F(w) = \sum_j \alpha_j \int |X_{\text{exp}}^j(t) - \tilde{X}_{\text{mod}}^j(t, w)|^2 dt,
\]  

(5.6)

where \( w = [w_1, w_1, \ldots w_n] \) are the TBF weights to be optimised, the sum over \( j \) runs over the number of different experimental data sets, and \( \alpha_j \) is a regularisation factor that scales the data so that each experiment can be compared in a balanced manner. This factor is determined based on the magnitude of each type of experiment included. In the current work, we shall focus on data sets that only include a single type of experiment. However, this provides a framework within which one can combine complementary data sets from different experiments. In addition, we note that this approach could also be applied to cases involving a density matrix representation of the model by augmentation with additional wavefunctions weighted by their respective populations. Our approach therefore amounts to an iterative refinement of the model wavefunction in light of the constraints placed on the observable by the experimental apparatus, accounted for by the apparatus mapping step \( S_j \) in Eqn. 5.3. An overview of the optimisation procedure, applied to the case of molecular structure determination using scattering, can be seen in Fig. 5.1.

The benefit of minimising the target function in Eqn. 5.6 in a trajectory basis is that it account for all continuity constraints and provides a set of physically reasonable constraints on a far from equilibrium system. The identification of the contributions to the experimental observable depends largely on the accuracy of the forward mapping
5.2 General Optimisation Framework

Figure 5.1: Overview of the optimisation procedure in which the target function is minimised with respect to the set of trajectory weights $w$ revealing the key contributions to the experimental observable. The case shown here represents the problem of the molecular structure determination in scattering experiments. The bank of trajectories from which the model wavefunction is refined is shown in blue, inputs to the optimisation i.e. the experimental and model observables are shown in red, and the output key molecular structures are shown in green.

and the trajectories themselves. Thus, when highly accurate observable calculations are not possible, the use of more approximate methods must be justified on a case by case basis and any potential features observed in the experiment that are not present in theory isolated prior to the procedure. Since the accuracy of the trajectories depends on the chosen electronic structure and nuclear propagation method, one should ensure an accurate description of the electronic wavefunction, nonadiabatic couplings, and nuclear equations of motion where possible. While the accuracy of the trajectories is important, and ultimately for the optimisation to pick out a particular contribution it must be present in the TBF bank, the re-parametrisation of the model wavefunction allows for some flexibility in discrepancies between theory and experiment. For example, allowing the correction of quantum yields and branching ratios. For this, generous sampling of the initial conditions is required so that a broad distribution of trajectories are included.

The minimisation of the target function in Eqn. 5.6 can be performed using various optimisation methods, the choice of which depends on the dataset, target function topology, and the dimensionality of the problem. Typically, the optimisation of the model wavefunction is nonconvex and there may exist many local minima. For this reason, gradient descent by itself, is not recommended. Stochastic methods such genetic
5.2 General Optimisation Framework

algorithms (GA) or simulated annealing (SA) are more suitable, allowing the exploration of large portions of the solution space and being less liable to get stuck in local minima [220, 221]. However, they may struggle to identify stable solutions, resulting from their sensitivity to initial conditions, sampling requirement, and large number of parameters. We find that a multi-start approach to the nonlinear trust-region reflective least-squares (TRRLS) algorithm provides both a reasonable computational overhead and suitable convergence [222]. One unique benefit of TRRLS is that is specifically designed to minimise sum of square error target functions, such as in Eqn. 5.6. Briefly, the nonlinear TRRLS algorithm constructs a local quadratic approximation to the target function at each step, requiring computation of the gradients and Hessian around the current parameter values (TBF weights). A trust region is constructed around this area, limiting the allowed step size. The optimal step size is determined by minimising the quadratic approximation within the trust region and the corresponding change in the target function is evaluated. If the step results in a decrease the trust region is expanded to further exploration of the solution phase. Conversely, if the step taken does not result in a sufficient decrease, the trust region is decreased to focus on exploitation and the current parameters are refined until a minima is located. Along with the multi-start implementation employed, this provides a balance between exploration and exploitation, addressing the challenges in a nonconvex optimisation.

5.3 Scattering Observables

We now turn to the specific case of inversion of x-ray and electron scattering observables for molecular structure determination. Although for gas phase ensembles it is possible to calculate the scattered intensity from first principles [87], and recent seminal studies have shown that with sufficient resolution and accuracy electronic effects can be resolved [21–23], here we focus on the use of the simpler IAM [130] owing to the available resolution of the data sets we shall later explore. A detailed discussion of scattering and the approximations within the IAM was given in Chapter 2.3, here we introduced the general expressions required for optimisation of the model wavefunction.
In the case of time-resolved scattering, the forward mapping yields the total scattering cross-section as the raw observable (i.e. $\chi_{\text{mod}}^{j} (t)$), which has not yet been subject to any apparatus mapping that mimics the act of experimental observation. Within the IAM, the instantaneous energy-integrated total scattering cross-section at a given point in time is $\langle d\theta / d\Omega \rangle_x$, where $q$ is the momentum transfer vector. In the x-ray scattering case, the prefactor $(d\theta / d\Omega)_x$ refers to the differential Thomson scattering cross-section $(d\theta / d\Omega)_{\text{Th}}$, which includes the polarisation factor $|e_0 \cdot e_1^*| [42]$, while for electron scattering it is the Rutherford scattering cross-section $(d\theta / d\Omega)_{\text{Ru}}$, which includes the scaling of the cross-section by $1/s^4 [132, 133]$. Note, here Eqn. 5.7 does not account for the duration of the incident pulses.

While more general expressions that account for the nonlocality of the trajectories from which scattering observables are calculated have been derived, we assume the diagonal zeroth-order bracket-averaged Taylor expansion (BAT) approximation [223]. Within which, the TBF weights can be taken to be time-independent, and the normalisation condition is simply given by $\| w \| = 1$. As such, the model scattering intensity is simply given by a weighted sum over each TBFs contribution,

$$ I_{\text{mod}} (q, t, w) = \sum_{n=1}^{N_{\text{TBF}}} w_n I_n (q, R_n (t)) , $$

with the total scattering intensity for a single TBF,

$$ I_n (q, R_n (t)) = |f (q, R_n (t))|^2 + S_{\text{inel}} (q) , $$

where the first and second terms are the elastic and inelastic scattering probabilities. In the IAM, the inelastic scattering is approximated as being independent of nuclear geometry, and is given by an incoherent sum over individual atomic contributions [130],

$$ S_{\text{inel}} = \sum_{A=1}^{N_{A}} S_A (q) , $$

1. Defined as $s$ in the case of electron scattering.
5.3 Scattering Observables

with $N_{at}$ the total number of atoms. The dominant elastic contribution is given by a coherent sum over the individual atomic form factors $f_A (q)$ as,

$$f (q, R_n (t)) = \sum_{A=1}^{N_{at}} f_A (q) e^{iqR_{nA}(t)}, \quad (5.11)$$

where $R_{nA} (t)$ is the position vector of atom $A$ in TBF $n$. As both the atomic form factors $f_A (q)$ and the incoherent scattering functions $S_A (q)$ are tabulated [131], the computation of the signal is straightforward. In the electron scattering case, the corresponding form factors are given by the Mott-Bethe formula as

$$f_e A (q) = \frac{Z_A}{s^2} \left[ f_x A \right], \quad [132, 133],$$

where $Z_A$ is the atomic number and $f_x A$ the tabulated x-ray scattering form factors. Note that in the case of high energy electron scattering, one may require fully relativistic form factors [225, 226]. In the event the system under study is an anisotropic distribution of molecules in the gas phase (as is the case in this chapter), rotational averaging of Eqn. 5.9 yields [130],

$$I_n (q, R_n (t)) = \sum_{A \neq B}^{N_{at}} f_A (q) f_B (q) \frac{\sin (q, R_{nAB}(t))}{qR_{nAB}(t)} + \sum_{A=B}^{N_{at}} |f_A (q)|^2 + S_{inel} (q), \quad (5.12)$$

with $R_{nAB} (t) = ||R_{nA} (t) - R_{nB} (t)||$ the Euclidean distance between atoms $A$ and $B$ in TBF $n$. Here the first term, known as the molecular term, contains all the structural dependence of the signal. Whereas, the second atomic term and third inelastic term only contribute a static background to the total signal. Note, that the amount of structural information in the signal is limited by the available $q$-range of a given experiment [174, 224].

5.3.2 Apparatus Mapping

To match the forward mapped observable (i.e. $X_{mod}^j (t)$) to the experimentally observed signal, we now consider the second apparatus mapping step, $S_j$ in Eqn. 5.3. The identification of time-zero (i.e. when excitation occurs from arrival of the pump pulse) in scattering experiments can be challenging, usually being roughly determined through observation of the signal. However, without independent validation, one must check the alignment of the temporal axis in the experiment and model. The corresponding relationship between the experimental and model time axes can be
5.3 Scattering Observables

defined as $t' = t + t_0$, with $t'$ the experiment, $t$ the theoretical model, and $t_0$ the time shift required for alignment. To ensure an accurate alignment and determination of time-zero, we include $t_0$ as an additional parameter to be indirectly optimised in our procedure.

As the determination of time-zero depends on incident pulse duration, we reproduce the effect of this on the model signal through a Gaussian convolution,

$$I_{\text{conv}}^{\text{mod}} = \int_{-\infty}^{\infty} I_{\text{mod}}(q, \tau) G(t - \tau) d\tau,$$

(5.13)

where the Gaussian kernel $G(t) = b_c \exp(-a_c t^2)$ approximates the instrument response function (IRF), within which the normalisation constant is given by $b_c = \sqrt{4 \log 2 / (\tau_c \pi)}$, where $\tau_c$ defines the full-width half maximum (FWHM). This effectively amounts to the cross-correlation of the pump and probe pulses, and to an extent compensates for the assumed $\delta$-function like excitation model assumed in trajectory simulations. Note that in some cases an additional convolution and binning in $q$ may be required to mimic the effect of the resolution of the detector.

Typically, the scattering signal is represented in terms of the percentage difference relative to the ensemble prior excitation. This percentage difference signal highlights the otherwise small modulations in the signal that result from structural changes over time, and is given by,

$$\% \Delta I_{\text{conv}}^{\text{mod}}(q, t, w, \gamma) = 100 \frac{I_{\text{conv}}^{\text{mod}}(q, t, w) - I_{\text{off}}^{\text{th}}(q)}{I_{\text{off}}^{\text{th}}(q)},$$

(5.14)

where the excitation fraction $\gamma$ scales the intensity of the signal according to the relative fraction of the ensemble that is excited, and $I_{\text{off}}^{\text{th}}(q)$ corresponds to the theoretical laser off signal. The laser off defines the signal prior to excitation by the pump pulse and is calculated from a Wigner distribution of the near equilibrium ground state ensemble. In some cases, the functional form of Eqn. 5.14 may vary slightly, however this provides a general description for all systems within the IAM. The corresponding experimental difference signal is normally defined as,

$$\% \Delta I_{\text{exp}}(q, t') = 100 \frac{I_{\text{on}}(q, t') - I_{\text{off}}(q, t' \ll t_0)}{I_{\text{off}}(q, t' \ll t_0)},$$

(5.15)
where $I_{\text{on}}(q,t')$ and $I_{\text{off}}(q,t' < t_0)$ represent the optically pumped signal and the static laser off signal respectively. Usually, the laser off signal is taken to correspond to an average of the signal at time delays well before any excitation occurs. While other representations of the signal are possible, the added benefit of using this explicit form is the additional degree of systematic error cancellation due to the division by the laser off signal.

### 5.3.3 Target Function Form

With the experimental and model observables defined as $\%\Delta I_{\text{exp}}(q,t')$ and $\%\Delta I_{\text{mod}}(q,t)$, the general target function in Eqn. 5.6 becomes,

$$F(w, c) = \sum_{i,j} \left| \%\Delta I_{\text{exp}}(q_i, t_{i}') - \%\Delta I_{\text{mod}}(q_i, t_{i}', w) \right|^2 p_{\text{conf}}(q_i, t_{i}') ,$$  

(5.16)

where $c$ consists of additional parameters to be optimised, such as the excitation fraction $\gamma$, time-zero shift $t_0$, and the Gaussian width $\tau_c$ that defines the temporal convolution. Note that the integral in Eqn. 5.6 has been replaced with a double summation over all available experimental data points, with $q_i$ and $t_{i}'$ identifying the momentum transfer and temporal coordinates respectively. To minimise the risk of overfitting to inherent noise in the experiment, we include a confidence matrix $p_{\text{conf}}(q_i, t_{i}')$ in the target function. This matrix weights each point in momentum transfer and temporal coordinates according to the confidence in the accuracy of the experimental data point. The form of this may vary between data sets, but generally points with poor statistics or those that are subjected to known systematic errors, are given a lower weight than others. In challenging cases, one can define a confidence threshold $p_{\text{conf}}^{\text{min}}$ such that all points where $p_{\text{conf}}(q_i, t_{i}') \leq p_{\text{conf}}^{\text{min}}$ are set to zero and excluded from the optimisation procedure. Of course, any exclusion of low quality data points must be balanced with the effects of a reduced sample size on which the fit is performed.

### 5.3.4 Assessing Fit Quality

Due to the lack of unique solutions, risk of overfitting, and the fact the molecular structure is underdetermined by the signal, it is advisable to assess the quality of solutions based on several different metrics in addition to the target function. The improvement in the overall agreement by the optimised an unoptimised model can be
5.3 Scattering Observables

assessed by the relative absolute error (RAE),

\[
    \text{RAE} = \frac{1}{N_t} \sum_t \frac{1}{N_q} \sum_q \left| \frac{\% \Delta I_{\text{mod}}(q,t)}{\% \Delta I_{\text{exp}}(q,t)} - \frac{\% \Delta I_{\text{th}}(q,t)}{\% \Delta I_{\text{exp}}(q,t)} \right|,
\]

(5.17)

where \( N_t \) is the number of time steps, \( N_q \) the number of points in \( q \), and \( \% \Delta I_{\text{th}}(q,t) \) the theoretical signal calculated from the unoptimised theoretical wavefunction (i.e. \( w_n = 1/N_{\text{TBF}} \)). Thus, quantifying how much better (or worse) the optimised model fits the experimental observable relative to the ensemble average, with values below one indicating an improvement. Since the RAE is independent of the \( N_q \times N_t \) grid, this allows comparisons between optimisations for which the values of \( t_0, \tau_c, \) and \( P_{\text{conf}}^{\text{min}} \) may vary. Furthermore, we also define the root mean squared error (RMSE) between the optimised model and the experiment as,

\[
    \text{RMSE} = \frac{1}{N_t} \sum_t \sqrt{\frac{1}{N_q} \sum_q \left( \frac{\% \Delta I_{\text{mod}}(q,t)}{\% \Delta I_{\text{exp}}(q,t)} \right)^2}.
\]

(5.18)

The resulting solutions can be characterized by the variance of all \( N_{\text{TBF}} \) weights, allowing one to determine the spread of the optimised weights that result from a particular initial condition, given by,

\[
    v_w^2 = \sum_{n=1}^{N_{\text{TBF}}} |w_n - \langle w \rangle|^2,
\]

(5.19)

where \( \langle w \rangle \) is the mean of the optimised weights. In addition, the distance \( D_b \) from the overall best set of optimised weights \( w_{\text{best}} \), can be used to quantify the distance of a particular solution \( b \) from the identified best result of a set of optimisations, and is defined as,

\[
    D_b^2 = |w_b - w_{\text{best}}|^2.
\]

(5.20)

Crucially, in identifying the best set of weights, one looks for consistency between a number of different metrics, steady convergence of the target function, and convergence of the actual values of the weights between different optimisations that are deemed "good" - i.e. the best sets of optimisations should converge to similar weights.
5.4 The Optimisation Procedure

We now turn to the specific procedural details. Generally, the optimisation can be performed in two ways; globally, where all parameters are optimised simultaneously, or using a two-step procedure, in which the optimisation of additional global parameters (such as $t_0$, $\tau_c$, and $\gamma$) and the TBF weights are separated.

5.4.1 Global Optimisation

In the ideal situation, the experimental resolution is sufficient to allow a global one-shot optimisation in which all model parameters can be optimised simultaneously. For this, the experiment must meet the requirements of having a combined pump-probe pulse duration short enough to observe fundamental molecular vibrations, a large enough $q$-range over which data is collected, and an excellent signal-to-noise ratio. In general, the number of global parameters defined by $c$ in Eqn. 5.16 is small. Therefore, a straightforward approach involves direct optimisation of the TBF weights $w$ for a set of fixed global parameters $c$, which are scanned over. For each set of fixed parameters, a pool of $N_{\text{init}}$ TBF weights $w$ are sampled in a Monte-Carlo fashion, with the values of each weight $w_n \in [0, 1]$. The target function in Eqn. 5.16 is then minimised with respect to the weights $w$ using the nonlinear TRRLS algorithm. Of the initial conditions that converge to a minima, those with the lowest values of the target function are selected as candidate solutions and compared using the different metrics defined in section 5.3.4. Convergence is typically defined by a minimum number of iterations for which the change in the target function is below a set tolerance threshold.

5.4.2 Two-Step Optimisation

Unfortunately the kind of ultrafast and high energy experiments that probe excited state dynamics are not always so well behaved, largely due to the sensitive nature of the particle accelerators at which they are performed. In cases where the experimental data is of lower quality, more stable solutions can be found through a two-step procedure, in which the global parameters $c$ are first approximately determined independent of the TBF weights.
5.4 The Optimisation Procedure

Step One: Global Parameter Determination

In this case, the global parameters are the time-zero shift $t_0$, the excitation fraction $\gamma$, and the convolution width defined by $\tau_c$. To determine the optimal values, we minimise the error between the experimental and theoretical net integrated percentage difference signal, defined as,

$$\% \Delta I_{\text{int}}(t) = \int_{p_{\text{conf}}^\text{min}}^{p_{\text{conf}}^\text{max}} \% \Delta I(q,t) \, dq,$$  \hspace{1cm} (5.21)

where the limits of integration define the bounds on the region of highest confidence, typically identified from confidence matrix $p_{\text{conf}}(q_i,t')$. This ensures the effect of low quality data points is minimised. With the aim of aligning the experimental and model time axes, we fit the net model integrated signal $\% \Delta I_{\text{int}}^{\text{mod}}(\gamma,t)$ to its experimental equivalent $\% \Delta I_{\text{exp}}^{\text{int}}(t')$, through minimisation of the modified target function,

$$F_{\text{int}}(\gamma,c) = \sum_j \left| \% \Delta I_{\text{exp}}^{\text{int}}(t'_j) - \% \Delta I_{\text{mod}}^{\text{int}}(\gamma,t_j) \right|^2,$$  \hspace{1cm} (5.22)

scanning over combinations of fixed global parameters $c$. Typically, the fixed parameters include $t_0$ and $\tau_c$, which is inherently linked to the determination of time-zero. The scaling of the intensity by the excitation fraction $\gamma$ also has an effect, thus we include this as a parameter to directly optimised. Note the independence of Eqn. 5.22 from the TBF weights $w$, which is ensured by assuming the $t \approx t_0$ wavefunction can be approximated by an equally weighted sum of TBFs (i.e. $w_n = 1/N_{\text{TBF}}$). Given the limited degree of dispersion and dephasing of the wavepacket immediately following excitation, such an assumption is reasonable. The set of values $(t_0, \tau_c, \gamma)$ that correspond to the lowest value of the target function, and the metrics described in section 5.3.4, are selected as inputs for the second step of the optimisation, the determination of the trajectory weights.

Step Two: Trajectory Weight Determination

With the best set of $(t_0, \tau_c, \gamma)$ values from step one, we then optimise the target function in Eqn. 5.16 with respect to the weights $w$. In this step, the value of $t_0$ and $\tau_c$ are fixed. However, the value of $\gamma$ is re-optimised along with the TBF weights $w$, using the value identified in step one as an initial guess. If the final value of $\gamma$ is significantly different from its initial value, this may indicate that the global parameters determined in step one are sub-optimal. The optimisation proceeds by generating a pool of $N_{\text{init}}$
5.4 The Optimisation Procedure

initial weights, which are Monte-Carlo sampled. Solutions with a sufficiently low target function value are then analysed with respect to the different metrics of fit quality. To ensure self-consistency between the two steps is achieved, the optimisation is repeated for other values of \((t_0, \tau_c, \gamma)\) that correspond to good fits from step one. Upon agreement between the values from both step one and two, the procedure is complete. The best solutions should be consistent over several different quality measures and result in a model that is commensurate with other independent experimental observations, where available. In the event these criteria are not met, possible issues must be evaluated on case by case basis.

5.5 Data Treatment

The optimisation procedure was tested on both x-ray and electron scattering data sets, in which the treatment of the data varied. Therefore, the general expressions in section 5.3 take a more specific form in each case, as we shall now explore.

5.5.1 Ring-Opening of 1,3-Cyclohexadiene

First, we turn to the electrocyclic ring-opening of 1,3-cyclohexadiene to 1,3,5-hexatriene, measured by time-resolved x-ray scattering. The experimental signal was obtained from Ref. 174, in which excitation to the optically bright 1B state was achieved with a 267 nm pump pulse. Data was collected over the temporal range, \(-1 \leq t' \leq 4\) ps, with the corresponding range in momentum transfer \(0.925 \leq q \leq 4.175\) Å\(^{-1}\). The duration of the UV pump and x-ray probe pulse were experimentally measured as 60 fs and 30 fs respectively.

To perform the forward optimisation procedure, we utilise a set 100 AIMCE trajectories, first introduced in Chapter 3.2. As the pulse duration is known, we fix the Gaussian convolution width \(\tau_c\), used in the apparatus mapping step, to \(\tau_c = 90\) fs. With the length of the simulated trajectories, and accounting for both the value of \(\tau_c = 90\) fs and the \(\Delta t = 25\) fs binning of time frames, the effective time interval over which the fit is performed is 275 fs.
In this case, we scale the percentage difference signal according to the observed fluctuations in the x-ray beam intensity. Thus, the percentage difference signal in Eqn. 5.15 becomes,

\[
\% \Delta I_{\text{exp}} (q, t') = \frac{100}{1 + \beta (t', \gamma)} \left( \gamma(q) \left[ \frac{I_{\text{on}} (q, t')}{{I_{\text{off}}} (q)} - 1 \right] - \beta (t', \gamma) \right),
\]

with \( \beta (t', \gamma) \) given by,

\[
\beta (t', \gamma) = \frac{Q'(t')}{Q_{\text{off}}},
\]

where \( Q_{\text{off}} \) and \( Q'(t') \) refer to the integrated signal with the laser off and on respectively. Note the additional dependence of the excitation fraction \( \gamma(q) \) on \( q \) in Eqn. 5.23, which we include to account for the nonuniformity of the scattered intensity across \( q \) - a result of the long interaction region in the experiment. The initial estimate of \( \gamma(q) \) was calculated from Beer-Lambert law, and is then included as an additional parameter to be optimised. The integrated laser on intensity in Eqn. 5.23 is given by,

\[
Q'(t') 2\pi \int_{q_0}^{q_1} F_{\text{instr}} (q) \gamma(q) \left[ I_{\text{on}} (q, t', \gamma(q)) - {I_{\text{off}}} (q) \right] q dq,
\]

where \( F_{\text{instr}} \) is the instrument response function that accounts for the variation in detector sensitivity across \( q \), and the limits of integration correspond to the radial limits of the detector. The corresponding integrated laser off intensity is defined as,

\[
Q_{\text{off}} = 2\pi \int_{q_0}^{q_1} F_{\text{instr}} (q) {I_{\text{off}}} (q) q dq.
\]

Rather than include \( \gamma(q) \) as a parameter to be directly optimised over the whole range of \( q \), we optimise the scaling factor \( x \) that uniformly scales \( \gamma(q) \) across \( q \) such that it yields a scaled excitation fraction \( \gamma_x (q) \), which is related to the initial estimate as \( \gamma_x (q) = \gamma(q) x \). Thus, the model signal now takes the specific form,

\[
\% \Delta I_{\text{conv}}^{\text{mod}} (q, t, w, x) = 100 \gamma_x (q) \frac{I_{\text{conv}}^{\text{mod}} (q, t, w) - {I_{\text{th}}} (q)}{{I_{\text{th}}} (q)}. \]

Here, all form factors used to calculate the theoretical model signal are obtained from Ref. 131. Moreover, to account for the fact the x-rays are linearly polarised, the model signal is also scaled by a polarisation factor of \( \frac{1}{2} \left( 1 + \cos^2 (\theta) \right) \).
5.5 Data Treatment

In this case, with available statistics on the number of photon hits per frame count, the confidence matrix included in the target function takes the form,

\[ p_{\text{conf}}(q_i, t'_j) = \frac{N(t'_j)}{N_{\text{max}}} p_q(q_i), \quad (5.28) \]

where \( N(t'_j) \) and \( N_{\text{max}} \) refer to the number of detector hits per frame and the maximum number of hits over all frames respectively. The accuracy of each point in \( q \) is given by \( p_q(q_i) \). Here, the quality of the experimental data is sufficient to allow the optimisation to be performed using the global procedure, outlined in section 5.4.1. For this, we scan over fixed values of \( x \) and \( t_0 \) in the ranges \( x \in [0.7, 1.3] \) and \( t_0 \in [-38, -14] \) fs, generating \( N_{\text{init}} = 100 \) initial TBF weights for each combination of parameters.

5.5.2 Photodissociation of CS\(_2\)

For our second data set, we turn to the photodissociation dynamics of CS\(_2\), observed using MeV-UED. Upon excitation of the system to the bright \(^1B_2\) (\(^1\Sigma^+_u\)) state complex dynamics ensue on a dense manifold of coupled electronic states, with the excited wavepacket favouring nonlinear geometries as it experiences a combination of bending and stretching modes [227, 228]. The system exhibits strong competition between internal conversion (IC) and intersystem crossing (ISC), ultimately resulting in two dissociation channels that yield either triplet (\(^3P_J\)) or singlet (\(^1D_2\)) sulphur as the dissociation product. The singlet dissociation occurs on faster timescales, however the triplet (\(^3P_J\)) product dominates [229, 230]. Note, the MeV-UED dataset obtained from Ref. 231 also features in the subsequent chapter, where we address the specifics of the experiment and underlying chemistry in more detail. Here, we focus on the convergence and performance of the optimisation procedure.

In this experiment, the system was excited to the optically bright \(^1\Sigma^+_u\) state with a 202 nm UV-pump pulse, and the resulting dynamics were probed using a 3.7 MeV electron beam. The experimental data was collected over the momentum transfer range \( 0.95 \leq s \leq 12.25 \) Å\(^{-1}\), with the corresponding temporal range \(-1.26 \leq t' \leq 1.73\) ps [231]. In this case, we leave the value of \( \tau_c \) to be determined by the optimisation procedure. Both the experimental and model difference signals are defined according to the more general expressions given in Eqn. 5.15 and Eqn. 5.14. As the experiment is conducted in the MeV domain, we use fully relativistic form factors calculated from the ELSEPA program [226]. However, since the relativistic and Mott-
Data Treatment

Bethe\(^2\) form factors are near identical over the experimental range of momentum transfer, we find that the choice in form factors has no effect on the optimisation result. Note that this is not guaranteed, and it is therefore advisable to check the difference between the fully relativistic and more approximate Mott-Bethe form factors.

The model signal is constructed from a set of 197 SH trajectories calculated using the SHARC surface-hopping package [123, 124]. The trajectories are prepared assuming a 200 nm excitation energy and with initial conditions sampled according to a ground state Wigner distribution. The electronic wavefunctions, forces, and non-adiabatic couplings were calculated at the SA8-CASSCF(10,8)/SVP level, specific details can be found in Ref. 227. While the trajectories are propagated for up to 1 ps, they often crash shortly after dissociation occurs due to instabilities in the electronic structure at large internuclear separations. Note, the dissociation time is defined as the time at which a trajectory reaches a C–S bond length greater than 3.4 Å, which is the point of no return in the simulations. To account for the contribution of the remaining vibrating C–S fragment to the model signal, we extend all crashed dissociative trajectories to 1 ps using a harmonic model of the remaining fragment, and linearly extend the dissociated sulphur according to its last known velocity. Accounting for the sampling of time delays in accordance with the experiment and the Gaussian convolution width \(\tau_c\), the effective time interval over which the fit is performed ranges from 1.05 to 1.15 ps.

In this case, we define the confidence matrix based on estimated experimental standard deviations from a bootstrapping procedure [231]. This is converted to a probability by normalisation as follows,

\[
p_{\text{conf}}(q_i, t'_j) = \left( \frac{\sigma(q_i, t'_j)}{\min \{\sigma(q_i, t'_j)\}} \right)^{-1},
\]

where \(\sigma(q_i, t'_j)\) is the estimated experimental standard deviation. Here, the experimental data does not support the one-step global optimisation procedure, we therefore utilise the two-step procedure outlined in section 5.4.2. In the first step, the net integrated signal (Eqn. 5.21) is calculated using the integration limits \([p_{\text{min}}^{\text{conf}}, p_{\text{max}}^{\text{conf}}] = [2.8, 4.2] \text{ Å}^{-1}\), identified from the confidence matrix \(p_{\text{conf}}(q_i, t'_j)\). Here, fixed values of \(t_0\) and \(\tau_c\) are scanned in the ranges \(t_0 \in [-16, 83] \text{ fs}\) and \(\tau_c \in [150, 250] \text{ fs}\), respectively. For the best set of parameters \((t_0, \tau_c, \gamma)\), \(N_{\text{init}} = 100\) initial conditions are

\(^2\) Calculated from tabulated x-ray form factors [131].
5.5 Data Treatment

generated for use in the multi-start optimisation of TBF weights. Given the poorer signal-to-noise ratio and resolution of the experiment, we also explore the use of a confidence threshold $p_{\text{conf}}^{\min}$ that filters the lower quality data points from the optimisation, values are scanned in the range of $p_{\text{conf}}^{\min} \in [0, 0.45, 0.50, 0.55, 0.60, 0.65]$. The higher the value of $p_{\text{conf}}^{\min}$, the more data points are excluded from the optimisation. Note, the case where $p_{\text{conf}}^{\min} = 0$ corresponds to including all data points, but weighing them according to their respective value in the matrix $p_{\text{conf}}(q_i; t'_j)$.

5.6 Results and Discussion

5.6.1 Ring-Opening of 1,3-Cyclohexadiene

First, we turn to analyse the results of the optimisation in terms of the target function, RAE, and the fractions of ring-opened and ring-closed product, shown in Fig. 5.2. Larger values of $x$ and more negative $t_0$ shifts result in lower values of both the target function $F(w, c)$ and the RAE, as in Fig.5.2c-d. Generally, the optimisation is more sensitive to changes in the scaling factor $x$ than $t_0$. In terms of the target function, values of $x \in [0.7, 0.85, 1]$ indicate that a $t_0$ of around -26 fs results in marginally lower values. In contrast, values of $x \in [1.15, 1.3]$ converge towards values of $t_0 \in [-34, -38]$ fs, with $(t_0, x) = (-38 \text{ fs}, 1.3)$ resulting in the lowest overall value of $F(w, c)$.

Turning to the RAE, the differences in fit quality are more apparent. Here, across all values of $x$, $t_0 = -38$ fs results in the biggest improvement, with $x = 1.3$ yielding the lowest value again. Figures 5.2a-b show the change in the fraction of ring-opened trajectories is small as $t_0$ changes, which is unsurprising since the majority of the ring-opening occurs within 140 fs [170]. Existing literature generally reports values in the range of 40-60% [167, 170, 172, 186]. With increasing $x$, we observe a redistribution of the ring-opened trajectory weights to ring-closed. Physically, one does not expect the fraction of ring-opened to closed trajectories to change with the excitation fraction. This artifact results from the fact that the percentage difference signal is stronger for ring-opened molecules than ring-closed, due to the reference structure corresponding to near equilibrium CHD in its ground electronic state. In principle, this could be addressed by a different choice in reference structure. The optimal pair of parameters selected correspond to $(t_0, x) = (-38 \text{ fs}, 1.3)$, due to their lower values of $F(w, c)$ and RAE. In addition, the observed fractions of ring-opened and closed products
are 48% and 52% respectively, commensurate with literature values. The value of $x = 1.3$ results in the global scaling of the $q$ dependent excitation fraction $\gamma(q)$, as seen in Fig. 5.3. The key input parameters and the values that correspond to the best optimisation are summarized in table 5.1.

The best result of the optimisation yields eight dominant trajectories that account for almost all the weight, summarised in table 5.2. While the procedure converges on a small set of trajectories, these capture each of the different possible product pathways i.e. hot CHD, cZc-Ht, cZt-HT/tZt-HT, and cEc-HT. Together, the selected trajectories correspond to a minimal basis representation of the dynamics observed in light of the experimental constraints. Thus, the optimisation procedure can be thought of as a useful tool that distills out the most significant contributions to the observed signal. The set of eight key trajectories can be visualised in terms of their characteristic $C_1-C_6$ bond length as shown in Fig. 5.4, and can be separated into three
5.6 Results and Discussion

Figure 5.3: The unoptimised initial excitation fraction $\gamma(q)$ and the optimised excitation fraction $\gamma_x(q)$ that results from scaling with $x = 1.3$.

Main groups, direct ring-opening, indirect ring-opening, and ring-closed. The indirect ring-opening trajectories correspond to those that observe several oscillations of the $C_1 - C_6$ bond before breaking. Moreover, the ring-closed trajectories observe several initial strong oscillations in $C_1 - C_6$ following excitation, these then dampen as energy disperses across other available motions and the trajectories return to the ground state of CHD. More specifically, the set of ring-opened trajectories are comprised of a range of different cis/trans isomers that form on the ground state, out of the 48% HT product we observe cZc-HT (34.5%), cEc-HT (8.7%), and a configuration between cZt/tZt (3.4%). Note that due to the available time scale, only a small amount of the tZt-HT product is detected, with most of 3.4% observing a configuration between cZt/tZt-HT. The dominance of the cZc-HT isomer is as expected as it forms immediately after relaxation to the ground state [167, 170], given a longer time range one would expect this to redistribute to a thermal equilibrium between cZc, cZt, and tZt-HT isomers.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$t_0$</td>
<td>-38 fs</td>
</tr>
<tr>
<td>$x$</td>
<td>1.3</td>
</tr>
<tr>
<td>$\tau_c$</td>
<td>90 fs</td>
</tr>
</tbody>
</table>

Table 5.1: The optimal input parameters for the CHD data set.
Table 5.2: Weights $w_{\text{best}}$ of the eight dominant CHD trajectories resulting from the best optimisation with $(t_0, x) = (-38 \text{ fs}, 1.3)$.

<table>
<thead>
<tr>
<th>Weight (%)</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>28.4</td>
<td>open (indirect)</td>
</tr>
<tr>
<td>27.3</td>
<td>closed</td>
</tr>
<tr>
<td>15.1</td>
<td>closed</td>
</tr>
<tr>
<td>9.70</td>
<td>closed</td>
</tr>
<tr>
<td>8.70</td>
<td>open</td>
</tr>
<tr>
<td>6.05</td>
<td>open (indirect)</td>
</tr>
<tr>
<td>3.38</td>
<td>open</td>
</tr>
<tr>
<td>1.35</td>
<td>open</td>
</tr>
</tbody>
</table>

Figure 5.4: The $C_1-C_6$ bond length for the eight dominant trajectories with a weight greater than 1% filtered out by the optimisation.

The whole set of solutions to the optimisation problem can be further characterised by the distance of each solution from the best set of weights $w_{\text{best}}$ (Eqn. 5.20), and the variance in each particular solutions set of weights (Eqn. 5.19), as shown plotted against the RAE in Fig. 5.5. Here, the best solution where $(t_0, x) = (-38 \text{ fs}, 1.3)$, is defined by $D^2_b = 0$ and an RAE = 0.775. Generally, in left panel of Fig. 5.5, there is little variation from the best optimisation if one examines solutions corresponding to more negative $t_0$ shifts and larger values of $x$. For these very good optimisations there is a clear cluster of solutions found within a small radius of $D^2_b = 0$, and with RAE’s in the range of 0.77-0.82. Within this cluster of solutions, the dominant trajectories are the same and variations between them only correspond to a slight re-weighting of the
5.6 Results and Discussion

trajectories, indicating the solutions converge towards the best set of weights \( w_{\text{best}} \). As the value of \( x \) decreases and/or the \( t_0 \) shift increases to more positive values, the solutions become further removed from the best result. The large distances in the case of optimisations where \( x \in [0.7,0.85] \) can be explained by the redistribution of weight to a number of ring-closed trajectories, and the selection of different set of ring-opened trajectories. In the right panel of Fig. 5.5, we observe that variance of the weights tends to increase for the better optimisations, as indicated by their RAE value. Thus, as the solution improves, the optimised set of weights assume a larger spread of values. While there is no observable trend between the weight variance and the scaling factor \( x \), we note that more negative values of \( t_0 \) tend to result in a higher variance.

![Figure 5.5](image)

**Figure 5.5:** The squared distances \( D^2_b \) (left panel) from the best set of weights and the variance \( \nu^2_w \) (right panel) of the weights for a series of CHD optimisations with different scaling factors \( x \) and \( t_0 \) shifts, plotted against the relative absolute error (RAE). Each colour represents a different \( x \) and shade represents a different \( t_0 \) value ranging from -38 to -14 fs (the darker the shade, the earlier the \( t_0 \) shift).

In Fig. 5.6, we can see the result of the optimisation in terms of the percentage difference signal for a series of time delays. Here we plot the optimised model signal \( \% \Delta I_{\text{mod}}(q,t) \), with weights \( w_{\text{best}} \) (as in table 5.2), and global parameters \( (t_0,x) = (-38 \text{ fs}, 1.3) \), against the experimental signal \( \% \Delta I_{\text{exp}}(q,t) \), and the theoretically calculated signal that is not subjected to any bias \( \% \Delta I_{\text{th}}(q,t) \) - i.e. assuming all 197 trajectories have an equal weighting. From this it is apparent that the optimisation results in better agreement with the experiment both qualitatively in terms of the signal, and quantitatively in terms of the previously discussed metrics. An improvement is seen
in terms of both the intensity of the signal, and the centre of the main peaks, with the peak around $2 \, \text{Å}^{-1}$ shifting in $q$. For the unoptimised case, this peak is observed at slightly lower values of $q$. Furthermore, there is clear improvement in agreement for the optimised model in the region of 2.75-3.25 $\text{Å}^{-1}$ between 100 and 150 fs.

**Figure 5.6:** Comparison of the experiment $\% \Delta I_{\text{exp}}(q,t)$, unoptimised $\% \Delta I_{\text{th}}(q,t)$, and optimised $\% \Delta I_{\text{mod}}(q,t)$ signals for a series of time delays.
5.6 Results and Discussion

5.6.2 Photodissociation of CS\textsubscript{2}

We now turn to the more challenging MeV-UED CS\textsubscript{2} dataset, where the resolution forces use to employ the two-step procedure. The result of the first step of the optimisation can be seen in Fig. 5.7, where for each combination of fixed parameters \( t_0 \) and \( \tau_c \), we plot the value of the target function, RAE, and RMSE. Across all values of \( \tau_c \), the more negative \( t_0 \) shifts perform better, with \( t_0 = -83 \) fs providing the best results. Note, for \( t_0 = -83 \) fs, we also observe smoother convergence across all values of \( \tau_c \). The optimal solution is identified as \( t_0 = -83 \) fs and \( \tau_c = 230 \) fs, which yields an optimised value of \( \gamma = 3\% \), commensurate with the experimental estimate of \( \gamma = 3 - 4\% \). Crucially, this solution results in not only the lowest value of the target function, but also the other two complementary metrics.

![Figure 5.7: Convergence of the \( t_0 \) fitting procedure with respect to different \( t_0 \) and \( \tau_c \) values. In the top panel is the convergence of the target function \( F(w, c) \), the middle panel the RAE, and the bottom the RMSE.](image)

For the identified optimal parameters \( (t_0, \tau_c, \gamma) = (-83 \text{ fs}, 230 \text{ fs}, 3\%) \), the result of the optimisation on the integrated percentage difference signal, can be seen in Fig. 5.8. Here, the signal is centred on \( t = 0 \) fs such that it corresponds to \( t' = -83 \) fs along the experimental time axis. In addition to showing good agreement in terms of the quantitative metrics previously discussed, the model also provides qualitative agreement with the experimental signal around time-zero. At later times the experiment is
within one standard deviation of the model, which refers to the standard deviation of the trajectory ensemble relative to the equally weighted model. The small standard deviation around time-zero shows the validity of the equal weight approximation used here. At times \( t < 0 \) fs, the noise floor in the experiment is apparent. While the noise floor is evenly distributed around zero, this highlights the importance of a method that is robust with respect to noise, allowing for the interpretation of challenging datasets.

![Graph showing %ΔI_{int} and %ΔI_{model} over time.](image)

**Figure 5.8:** The best fit for step one of the optimisation where \( t_0 = -83 \) fs and \( \tau_c = 230 \) fs. The fit is performed using integrated signal %ΔI_{int}^\text{mod}, and results in an optimised excitation value of \( \gamma = 3\% \). The shaded area corresponds to the standard deviation of the trajectory ensemble from the model calculated assuming equal weights.

Taking the best global parameters \((t_0, \tau_c, \gamma) = (-83 \text{ fs}, 230 \text{ fs}, 3.0\%)\) from step one as a starting point, we then proceed with the determination of the trajectory weights \(w\). However, to check that consistency between step one and step two is achieved, we repeat the optimisation for \( t_0 \in [+17, -33] \) fs. The convergence of the target function \( F(w, c) \) as a function of different confidence matrix thresholds \( p^{\min}_{\text{conf}} \) and \( t_0 \) shifts, can be seen in Fig. 5.9. Note that generally, the values of the target function between different \( p^{\min}_{\text{conf}} \) thresholds are not comparable, since the confidence matrix itself affects the value of the target function. However, we observe that across all values of \( p^{\min}_{\text{conf}} \), \( t_0 = -83 \) fs results in lower values of the target function. The fact the second step of the procedure consistently converges on \( t_0 = -83 \) fs, is commensurate with the first step of the procedure. It is also worth noting that the difference between the values of \( F(w, c) \) are more pronounced for lower values of \( p^{\min}_{\text{conf}} \), which correspond to the inclusion of more data points.
5.6 Results and Discussion

Figure 5.9: Convergence of the target function $F(w, c)$ for a series of different $t_0$ shifts and $p_{\text{min}}^{\text{conf}}$ thresholds.

The convergence of the second step of the optimisation with respect to the triplet to singlet branching ratio, the remaining bound fraction at 1 ps, the excitation fraction $\gamma$, and the RAE, can be seen in Fig. 5.10. Generally, across all values of $t_0$, higher $p_{\text{min}}^{\text{conf}}$ thresholds are detrimental to the RAE, as in Fig. 5.10d. In addition, the best (lowest) value of the RAE is observed when $p_{\text{min}}^{\text{conf}} = 0$ and $t_0 = -83$ fs. The re-optimised values of the excitation fraction $\gamma$ are given in Fig. 5.10c, where their relative stability is apparent. This indicates the initial estimate of $\tau_c$ from step one is sufficient.

Figures 5.10a-b show the convergence with respect to the observable branching ratio and the bound fraction remaining at 1 ps. While these observable quantities are not explicitly included in the optimisation, they are indirectly a function of the trajectory weights. Hence, a good model should reproduce values that are in agreement with other independent experiments. The branching ratio varies significantly across the values of $t_0$ and $p_{\text{min}}^{\text{conf}}$, with convergence towards literature values of $\approx 3$ seen only for $t_0 = -83$ fs and lower values of $p_{\text{min}}^{\text{conf}}$ [230, 232–239]. Generally, increasing $t_0$ causes earlier onset of the features correlated with dissociation in the signal, thus the contribution of the singlet dissociation channel is underestimated. In contrast, the convergence of the bound fraction is more uniform across the range of parameter values. As the value of $p_{\text{min}}^{\text{conf}}$ decreases, values in the range of 20-30% are observed, congruent with independent spectroscopic results. Increasing the value of $p_{\text{min}}^{\text{conf}}$ to greater than 0.50 results in the bound fraction being increasingly overestimated, which is again caused the omission of data correlated with dissociation. The poorer results
observed for higher values of $p_{\text{conf}}^{\text{min}}$ highlight the importance of finding a balance between minimising the risk of fitting to noise, and ensuring a large enough sample of data points on which the fit is performed. The key input parameters and their corresponding values which result in the best fit, are summarized in table 5.3

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$t_0$</td>
<td>-83 fs</td>
</tr>
<tr>
<td>$\tau_c$</td>
<td>230 fs</td>
</tr>
<tr>
<td>$\gamma$ (initial)</td>
<td>3%</td>
</tr>
<tr>
<td>$p_{\text{conf}}^{\text{min}}$</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 5.3: The optimal input parameters for the CS$_2$ data set.

![Figure 5.10: Convergence of the second step on the CS$_2$ optimisation as a function of the confidence threshold $p_{\text{conf}}^{\text{min}}$ and $t_0$ values. Panel a) branching ratio, b) bound fraction at 1 ps, c) excitation fraction $\gamma$, and d) the relative absolute error (RAE).](image-url)
Based on the above, the identified best optimisation yields the final values \([t_0, \tau_c, \gamma, p_{\text{conf}}^{\text{min}}] = [-83 \text{ fs}, 230 \text{ fs}, 3.4\%, 0]\). This results in the re-optimisation of initial guess \(\gamma = 3\%\) from step one, to \(\gamma = 3.4\%\). The comparatively small change in \(\gamma\) indicates a stable solution. This is further supported by the fact that across step one and two, \(t_0 = -83\) fs consistently provides the best results in terms of both the target function, and additional metrics. The best optimisation yields eight dominant trajectories with a weight greater than 1\%, as summarised in table 5.4. Again, these trajectories make up almost all of the total weight. Two of these trajectories are bound and together make up a total of 28\%. Out of the dissociative trajectories, three are singlet and two triplet, resulting in a branching ratio of 1:3.38 in favour of the triplet.

<table>
<thead>
<tr>
<th>Weight (%)</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>44.7</td>
<td>Triplet</td>
</tr>
<tr>
<td>17.8</td>
<td>Bound</td>
</tr>
<tr>
<td>10.4</td>
<td>Singlet</td>
</tr>
<tr>
<td>10.2</td>
<td>Bound</td>
</tr>
<tr>
<td>7.42</td>
<td>Triplet</td>
</tr>
<tr>
<td>4.42</td>
<td>Singlet</td>
</tr>
<tr>
<td>3.39</td>
<td>Triplet</td>
</tr>
<tr>
<td>1.67</td>
<td>Singlet</td>
</tr>
</tbody>
</table>

Table 5.4: Weights of the eight dominant CS\(_2\) trajectories resulting from the best optimisation with \([t_0, \tau_c, \gamma, p_{\text{conf}}^{\text{min}}] = [-83 \text{ fs}, 230 \text{ fs}, 3.4\%, 0]\)

In Fig. 5.11 we see the solutions characterised by their square distances \(D_b^2\) and variances \(v_w^2\). In the left panel, the set of good solutions, as judged by their lower RAE values, cluster around the best solution \(w_{\text{best}}\) as measured by the distance \(D_b^2\). This trend is more pronounced for values of \(t_0 \in [-83, -33] \text{ fs}\) and \(p_{\text{conf}}^{\text{min}} \in [0, 0.45]\). Furthermore, we see that for values of \(p_{\text{conf}}^{\text{min}} \in [0, 0.45, 0.5]\), the earlier time shifts deviate less from the best set of weights \(w_{\text{best}}\). The behaviour of larger values of \(p_{\text{conf}}^{\text{min}}\) is less well defined. However, we note that the increase in distance from \(w_{\text{best}}\) can be explained by the fact that as the value of \(p_{\text{conf}}^{\text{min}}\) increases, the bound population at 1 ps is overestimated. In the right hand panel we see that generally the better optimisations result in a higher spread of weights, as indicated by their variance \(v_w^2\).

The optimisation tends to converge on the same subset of 16 trajectories, with the case where \(p_{\text{conf}}^{\text{min}} = 0\) converging on the same eight, as in table 5.4. Generally, different input values result in most of the weights varying by a small amount, however some trajectories experience a significant re-weighting. For example, the dominant triplet trajectory with a 45\% weight appears only once the value of \(p_{\text{conf}}^{\text{min}}\) is lowered
5.6 Results and Discussion

Figure 5.11: The squared distances $D^2_b$ (left panel) from the best set of weights and the variance $v^2_w$ (right panel) of the weights for a series of CS$_2$ optimisations with different confidence matrix thresholds $p^\text{min}_\text{conf}$ and $t_0$ shifts, plotted against the relative absolute error (RAE). Each colour represents a different $p^\text{min}_\text{conf}$ and shades represents a different $t_0$ value ranging from +17 to -83 fs (the darker the shade, the earlier the $t_0$ shift.

and a wider range of data points included. This effect is also seen for singlet trajectory with 4.5% weight in the best case, which increases from near 0% as the confidence threshold is decreased. The opposite is also observed for two different triplet and bound trajectories, which see their weight decrease from 20% $\rightarrow$ 7.5% and 32% $\rightarrow$ 18% respectively. Thus, when the confidence matrix threshold is too high and the optimisation performed on a smaller subset of the data, the contribution of some trajectories can be over or under estimated. Moreover, the distribution of singlet and triplet trajectories is sensitive to the value of $t_0$. Given that singlet and triplet dissociation is separated temporally, with the singlet typically occurring at earlier times, this is to be expected. By adjusting the value of $t_0$ shift, the alignment of the experimental and model temporal axes shifts, and the optimisation attempts to compensate for this by changing the relative distribution of singlet and triplet dissociation.

The result of the optimisation on the signal can be seen in Fig. 5.12, while the improvement is less clear than in comparison to the CHD case, one can see that the optimised signal shows better agreement in relation to the main enhancement in the signal, which is initially centred in the experiment at 3.79 Å$^{-1}$ and shifts to 3.45 Å$^{-1}$ at 1 ps. This region of momentum transfer $s$ is where the highest confidence in the experimental signal is observed. For the optimised model, the peak is initially
centred at 3.65 Å\(^{-1}\) and shifts to 3.43 Å\(^{-1}\), in contrast it is centred at 3.55 Å\(^{-1}\) and shifts to 3.39 Å\(^{-1}\) in the unoptimised case. At just below 2 Å\(^{-1}\) we observe a small rise in intensity that appears to grow in time. Both the shift in main enhancement and the rise in the low \(s\) signal appear to be correlated with dissociation occurring\(^3\).

3. The correlation of these features with dissociation and the structural dynamics will be discussed in depth in the subsequent chapter.
Figure 5.12: Comparison of the experiment $\%\Delta I_{\text{exp}}(s,t)$, unoptimised $\%\Delta I_{\text{th}}(s,t)$, and optimised $\%\Delta I_{\text{mod}}(s,t)$ signals for a series of time delays.
5.6.3 Convergence Criteria and Weight Sampling

Generally, the optimisation procedure is robust if the data is treated appropriately and the required geometries are contained within the bank of trajectories. However, given the multi-start nature of the optimisation, the sampling of initial conditions is important in ensuring the correct minima is located. In the top panel of Fig. 5.13 we see the value of the target function for a series of different initial condition samplings, where the number of initial conditions \( N_{\text{init}} \) is increased from 10 to 200. The lowest value of the target function is observed when \( N_{\text{init}} = 200 \). However, convergence is generally quick and increasing the value of \( N_{\text{init}} \) generally results in a greater number of initial conditions reaching the identified lowest minima. In the case where \( N_{\text{init}} = 100 \), which is the value used in the optimisations, 46% of the initial conditions within the sample pool converge within \( 1e^{-3} \) of the lowest value.

![Figure 5.13](image)

*Figure 5.13:* Convergence of the target function with respect to the number of initial conditions \( N_{\text{init}} \).

The convergence of the best optimisation with respect to the iteration number can be seen in the top panel of Fig. 5.14. Here we observe that the target function converges within 30 iterations. The bottom panel of Fig. 5.14 shows the corresponding change in the step size taken. The stopping condition for the optimisation is defined by the change in the target function value and step size between iterations. Once the target function changes by less than \( 1e^{-7} \) and the step size is smaller than \( 1e^{-5} \), the optimisation stops. If the target function is non-decreasing in all directions, a minima has been located.
5.6 Results and Discussion

Figure 5.14: Convergence of the target function with respect to the number of iterations (top panel), and the size of step taken at each iteration (bottom panel). The case shown corresponds to the best optimisation with weights \( w_{\text{best}} \).

5.6.4 Iterative Weight Generation

In addition to the standard Monte-Carlo sampling procedure, it may sometimes be beneficial to instead sample weights through an iterative procedure, in which the sampling is initially constrained and at each iteration constraint is lifted. For this we define an increasing set of 35 upper bounding limits \( U_{\text{bound}} \), where \( U_{\text{bound}} \in [0.075, 1] \).

For each limit within the set, the TBF weights are sampled on the internal \( w_n \in [0, U_{\text{bound}}] \). As the value of \( U_{\text{bound}} \) increases, so does the sampling requirement of the respective weight space. Hence, we increase the number of initial conditions from \( N_{\text{init}} = 10 \) to \( N_{\text{init}} = 100 \) over the series of bounds. Note that the \( U_{\text{bound}} = 1 \) case corresponds to the unconstrained sampling case, from which the best set of weights \( w_{\text{best}} \) were previously determined. Through iteratively increasing the size of the space from which weights are sampled, the approach allows for a combination of local and global search. At each iteration, the best solutions from the previous iteration are added to the pool of new initial conditions to be generated, avoiding loss of good solutions.

Turning to Fig. 5.15, we see the heatmap of the best trajectory weights for each upper bound. For highly constrained optimisations, the weights are more evenly distributed between a larger number of trajectories. As the upper bound increases, the optimisation picks out a smaller range of trajectories that have a larger variance in
weight between them. Once \( U_{\text{bound}} = 0.30 \) is reached, the trajectory weights vary little and begin to converge as the minimal basis representation of the observed dynamics is revealed. At \( U_{\text{bound}} = 1 \), corresponding to the unconstrained case, eight dominant trajectories with a weight greater than 1% are filtered out. Note that the iterative sampling approach used here converges to the same minima as in the global approach, and the eight trajectories correspond to those in table 5.4. Moreover, it is worth noting that the excitation fraction \( \gamma \) tends to vary only by a small amount between different values of \( U_{\text{bound}} \). Typically values of \( \gamma \) are always in the range 3.4-3.7%, again demonstrating that the value of \( \gamma \) is robust. In addition, the bound fraction at 1 ps tends to lie in the range of 20-30%, before converging to 28% as the constraint is lifted and \( U_{\text{bound}} \to 1 \). In contrast, the same trend is not observed for the branching ratio, which converges to the value of 1:3.38 once \( U_{\text{bound}} \geq 0.40 \). At this point, some weight is redistributed to give a triplet trajectory in Fig. 5.15 a value of 44.7%.

**Figure 5.15:** Heatmap of trajectory weights \( w \) for the best optimisations that result from each of upper bounding limits \( U_{\text{bound}} \) in the iterative approach to weight sampling.
5.7 Conclusions

In summary, we have introduced a forward optimisation based approach to the inversion of time-resolved data, and specifically evaluated its application to scattering observables. By exploiting that the forward mapping of observables is well defined, this approach succumbs less to the lack of definition in typical direct inversion methods. The parameterised model wavefunction reconstructs the experimental observable in light of the given constraints of the apparatus by optimising a set of semi-classical trajectory weights. The trajectories provide a set of reasonable constraints on far from equilibrium structures, and ensure that all continuity relations are fulfilled. Moreover, the trajectory basis allows one to partially overcome the highly underdetermined nature of inversion, and yields physically sensible solutions. While the accuracy of the trajectory ensemble is important, and both the experiment and simulation must probe the same space, our approach allows for the optimisation procedure to overcome some potential shortcomings of the simulation. For example, quantum yields and branching ratios. On the inclusion of suitable confidence matrix, the risk of over fitting is reduced, and thus the procedure is robust with respect to noise. Thresholding of poor quality data points must be balanced against ensuring a large enough sample, and generally better results are observed for lower values of the confidence matrix threshold. Furthermore, the two-step optimisation procedure allows one to accurately determine time-zero in noisy datasets. The iterative weight generation procedure provides a way to perform a series of local and global searches of the target function, which may be useful in datasets with poor signal-to-noise due to the likely increased number of possible minima.

In an application the ring-opening of CHD, and the photodissociation of CS$_2$, we demonstrate the use of our procedure on both high and lower resolution datasets. Generally the method is robust, and provides a tool for the extraction of information from otherwise challenging datasets. In both studied reactions, we have seen that the method is capable of reproducing observed values of branching ratios from other independent experiments. In addition, it allows for identification of key dynamics motifs, for example the three distinct pathways that contribute to the signal in the case of CHD, and the presence of a key dissociation marker in the case of CS$_2$. The model tends to converge on a handful of trajectories that define the minimal basis representation of the dynamics observed, which reflects the limited momentum space resolution of the experiments.
5.7 Conclusions

Looking forward, the framework laid out in this chapter should be explored in applications to other observables, such as TRPES and CEI, as well as multimodal datasets where complementary information from different datasets are combined. For example, by combining population information from TRPES with structural information from scattering experiments. Such an approach would reduce the risk of overfitting and allow for the resolution of a broader range of information contained within the multidimensional evolution of the wavefunction. For scattering observables in particular, higher quality data could potentially allow for the inclusion of state-specific scattering, inelastic effects, coherent-mixed scattering, and molecular alignment in the inversion.

In terms of the optimisation itself, further exploration of different algorithms other than TRRLS could provide potential ways to increase the efficiency and quality of the solutions identified. While in the examples studied in this reaction, we find TRRLS performs well, this may vary depending on the observable and dataset.

The current approach fits nicely into the typical workflow of an ultrafast experiment, with the forward optimisation being performed at little extra cost, allowing for a robust interpretation of the experimental data, and for easier identification of discrepancies between theory and experiment. Generally, the analysis of such experiments are challenging, with the identification of subtle effects requiring high quality low noise data. Thus, the ultrafast community benefits from a broad arsenal of methods that provide a good interpretation of the data. In recent years, interest in inversion schemes has peaked, and they are thought to play an increasingly important role in experimental interpretation [82, 83, 211–217]. Such methods are clearly capable of yielding physically credible provisional answers and accelerate the process of experimental analysis.
Chapter 6

Observing Photodissociation by Electron Scattering

6.1 Introduction

In this chapter we proceed to take a closer look at the CS$_2$ data set briefly introduced in Chapter 5, focusing on the specifics of the ultrafast electron diffraction (UED) experiment in which the photodissociation dynamics were observed. The work herein constitutes the collaborative effort of both experiment and theory and was previously published in Physical Chemistry Chemical Physics as "Multichannel Photodissociation Dynamics in CS$_2$ Studied by Ultrafast Electron Diffraction" [231].

Despite the relative simplicity of its ground state equilibrium structure, the excited state dynamics that ensue following excitation to the $^1\Sigma_u^+$ ($^1B_2$) state are incredibly complex in nature, making the system a key target for ultrafast scientists [229, 230, 240, 241]. However, the majority of existing studies focus on the electronic state dynamics and a complete picture of the evolving nuclear wavepacket eludes us. With advances in mega-electronvolt (MeV) electron sources providing both sub-Å spatial resolution and femtosecond temporal resolution, MeV-UED has proven to be successful in probing the structural changes involved in complex photoexcited dynamics. Moreover, MeV-UED studies have a high potential for providing complementary information to spectroscopic studies [173, 242], and has recently been employed to image both electronic and nuclear dynamics simultaneously [23]. In this chapter we explore the use of MeV-UED in directly observing the nuclear dynamics at play following excitation with a 200 nm pump laser. In principle, the observed scattering signal can be transformed to reveal the changes in the molecular structure. However, this is not guaranteed yield meaningful insights due to the limited range of momentum...
6.1 Introduction

transfer over which data is collected, dispersion of the wavepacket, and the existence of many different product channels, making the identification of representative molecular structures difficult [243]. The extraction of accurate structural information from experiment therefore often relies on close comparison to accurate theoretical models.

Previous TRPES studies of CS$_2$ exploited molecular alignment to target the early internal conversion (IC) dynamics [228, 240, 241], highlighting the initial bending and stretching that occurs in the initially populated $^1\Sigma_u^+$ state. As the wavepacket experiences rapid bending the $^1\Sigma_u^+$ state gains $^1\Pi_g$ character, with the photoelectron angular distributions (PADs) showing a significant contribution from a $\sigma^*$ orbital. More recently, two extreme ultraviolet (XUV) TRPES studies tracked the complete reaction pathway, observing the competition between IC and ISC [229, 230]. Both studies report similar time constants and branching ratios, observing the triplet $^3P_J$ dissociation channel dominates the fast singlet ($^1D_2$) channel.

Furthermore, quantum beats associated with the bending and asymmetric stretching modes were observed in both time-resolved ion yield [244] and UV photoelectron spectroscopy experiments [230, 245]. The quantum beats have been shown to be the result of large changes in the ionisation potential associated with the bending and stretching modes [230]. Here, the signal observed significant broadening due to the rapid damping of the oscillations, suggesting a highly disperse wavepacket and an associated broad distributions of structures.

In the following sections, we explore an alternative perspective on the photodissociation dynamics of CS$_2$, providing complementary information to existing spectroscopic studies by focusing on the structural dynamics. In comparison to x-ray scattering, the higher available range of momentum transfer over which data is collected and the larger cross-section in principle facilitates the transformation of the signal into real space, allowing analysis of the structural dynamics. We explore the feasibility of such real space transformation of the signal in analysing the structural dynamics, and provide an alternative method for extracting the structural dynamics in the event the transform cannot be performed. For this, we employ the forward optimisation based approach to inversion introduced in Chapter 5, which focused on the performance and convergence of the algorithm. In addition, we provide an outlook on the additional observations that are likely resolvable in future UED experiments with improved temporal resolution.
6.2 Methodology

6.2.1 Experimental Methods

The experiment was performed using a relativistic MeV electron beam line at SLAC national accelerator laboratory [68]. Both the UV pump and electron probe pulses were generated using a Ti:Sapphire laser system, with the electron bunch being generated from a photocathode using the third harmonic of the laser. The emitted electron bunches are accelerated to 3.7 MeV with a radiofrequency gun operating at a repetition rate of 360 Hz. The bunches are focused to a spot size of 200 µm FWHM using a holey mirror, with a pulse duration of ~125 fs. The linearly polarised UV pump pulse is generated using a nonlinear crystal based scheme, with a pulse duration of ~100 fs and an on-target energy of 2 µJ. Separation of the 200 nm light is achieved using dielectric mirrors. Both the electron beam and the UV pulse are then recombed at small angles and focused to a 250 µm spot in the gas-sample cell. In the experiment, the excitation fraction $\gamma$ was approximately determined to lie in the range of 3-4%. Moreover, comparison of the static and time-dependant signals as a function of pump laser intensity confirmed that the experiment was performed in the linear absorption range.

The liquid CS$_2$ sample (Sigma Aldrich, >99%) was placed in a bubbler and delivered to the gas-cell, without the use of a seed gas. A computer controlled flow control was used to control the sample density and thus the strength of the signal. The path length of the gas-cell was ~4 mm, with an inner diameter of ~2 mm, and a wall thickness of ~1 mm. To prevent sample clogging and clipping of the pulses a hole diameter of 0.4 mm was used. The sample cell was heated to 100°C to further prevent clogging and condensation. The in situ sample pressure was ~1 Torr, corresponding to an average collision time of approximately 100 ns, which is significantly longer than the timescales of the dynamics under investigation.

The scattered electrons were detected using a phosphor screen and a charge coupled device (CCD) camera. Images were recorded over 20 seconds with 7200 electron bunches and stored independently for post-processing. To limit the effect of unscattered electrons contributing a significant background to the signal, the phosphor screen contains a hole in the centre, allowing the unscattered electrons to be transmitted through it. The final dataset used in the analysis is the result of multiple time scans added together, with each time delay corresponding to approximately 185 combined individual frames.
6.2 Methodology

6.2.2 Data Treatment

The recorded scattering patterns were first checked for any obvious unphysical changes in intensity over all pump-probe delays. The overall scattering intensity was calculated as a sum of all hits over the full image. Any images with an intensity outside of 2 standard deviations of the mean were removed. Following these quality checks, the centre of each recorded image was identified independently, given the potential shifting in the image centre due to the electron beam drifting [68]. In addition, further variations may occur at high pump energies due to the formation of plasma around the centre. The electric field of the plasma can interact with electron beam, causing streaking and shifting of the image centres. To avoid these plasma effects, we kept the pump energy low. Typically, deviations of the image centres were small and on the order of only a few pixels. The centre of each image was found using a least squares fitting algorithm [246], which given an initial guess of the centre finds an average value in a ring at an input radius.

After performing quality control and centering each image, the rotationally averaged scattering signal as a function of pump-probe delay was obtained by integrating over all azimuthal angles and subsequently combining the signal at each delay time. To account for changes in the absolute electron beam charge, the scattering signal was normalised assuming the total scattered signal in the range of \( 2 \leq s \leq 8 \ \text{Å}^{-1} \) remains constant.

The momentum transfer \( s \) axis was calibrated based on the static CS\(_2\) signal, and was performed by conversion of the signal to \( sM(s) \),

\[
sM(s) = s \frac{I_{\text{mol}}(s)}{I_{\text{at}}(s)}, \quad (6.1)
\]

where \( I_{\text{mol}}(s) \) is the molecular scattering term, arising from the interference between the scattering amplitudes of different atoms, and \( I_{\text{at}}(s) \) is the atomic background. Experimentally, \( I_{\text{mol}}(s) \) was obtained by subtraction of an initial uniform background and the atomic term \( I_{\text{at}}(s) \) from the measured scattering intensity. The atomic term is defined as,

\[
I_{\text{at}}(s) = \sum_{A=B} |f_A(s)|^2, \quad (6.2)
\]
where $f_A(s)$ is the form factor of nucleus $A$, calculated using ELSEPA program [225, 226] assuming a 3.7 MeV electron beam energy. The obtained experimental $sM(s)$ was then compared to the corresponding static CS$_2$ signal calculated from a Wigner distribution of the ground state system, with $I_{\text{mol}}(s)$ defined within the IAM as,

$$I_{\text{mol}}(s) = \sum_{A \neq B} N_{\text{at}} f_A(s) f_B(s) \frac{\sin(sR_{AB})}{sR_{AB}}$$  \hspace{1cm} (6.3)$$

with $N_{\text{at}}$ the number of atoms, and $R_{AB}$ the distance between atoms pairs $A$ and $B$. The comparison allowed the calibration from pixels to momentum transfer in the range $0.95 \leq s \leq 12.25$ Å$^{-1}$. Note, as collecting data down to $s = 0$ Å$^{-1}$ is not possible due to the presence of the detector hole, the signal is extrapolated to $s = 0$ Å$^{-1}$ linearly.

In gas-phase UED it is common to highlight the structural changes observed as a function of pump-probe delay as,

$$\Delta sM(s,t) = \frac{I_{\text{on}}(s,t) - I_{\text{off}}(s,t \ll 0)}{I_{\text{at}}(s)}$$  \hspace{1cm} (6.4)$$

where $I_{\text{off}}(s,t \ll 0)$ and $I_{\text{on}}(s,t)$ correspond to the unpumped and pumped signals respectively. The benefit of this representation of the signal is that the division by $I_{\text{at}}(s)$ highlights the structural differences that result from the interference atomic contributions. In principle, $\Delta sM(s,t)$ can be transformed from reciprocal to real space, to obtain the distance pair distribution function $\Delta \text{PDF}(R,t)$,

$$\Delta \text{PDF}(R,t) = \int_{s_{\text{min}}}^{s_{\text{max}}} \Delta sM(s,t) \sin(sR) e^{-\alpha s^2} ds,$$  \hspace{1cm} (6.5)$$

where $\alpha$ is a dampening factor that acts to smoothly reduce the intensity of the observed signal towards zero at $s_{\text{max}}$ and minimise high-frequency artefacts from the truncation of the signal at $s_{\text{max}}$. Moreover, the dampening factor acts to minimise the effects of noise at higher values of $s$, since the signal scales with $s^{-4}$. Ideally, one would set the limits $s_{\text{min}} = 0$ and $s_{\text{max}} = \infty$, however this is not possible in practice.

Alternatively, one can represent the time-dependent signal in the form of $\%\Delta I_{\text{exp}}(s,t)$, defined as,

$$\%\Delta I_{\text{exp}}(s,t) = \frac{I_{\text{on}}(s,t) - I_{\text{off}}(s,t \ll 0)}{I_{\text{off}}(s,t \ll 0)},$$  \hspace{1cm} (6.6)$$

where $I_{\text{off}}(s,t \ll 0)$ and $I_{\text{on}}(s,t)$ correspond to the unpumped and pumped signals respectively.
where we take $I_{\text{off}}(s, t \ll 0)$ to be the averaged scattering signal measured at 20 discrete time delays $-1200 \leq s \leq -120$ fs. This has the advantage of being completely independent of the $I_{\text{at}}(s)$ term calculated from theory, potentially resulting in some additional error cancellation by division with $I_{\text{off}}(s, t \ll 0)$, and also reduces the effect of scaling the signal with $s$. For reasons to be addressed later in the chapter, we tend to perform most of our analysis in the basis of $\% \Delta I_{\text{exp}}(s, t)$.

Based on experiment alone, the absolute value of time zero $t_0$ could not be determined exactly. However, it was first determined to lie in the range of 100 fs based on traces of the observed signal. The exact value of $t_0$ was then more accurately determined from the forward optimisation method discussed at length in Chapter 5. Moreover, the statistical uncertainty in each experimental data point $(s, t)$ was determined using a bootstrapping (jackknifing) procedure. For this 200 datasets with 185 images per time delay were constructed by random sampling with replacement. The mean and standard deviation of each point in momentum transfer $s$ and time delay $t$ were then obtained.

### 6.2.3 Forward Optimisation

The characteristic structural dynamics are extracted from the experimental signal using the optimisation procedure from Chapter 5, where the convergence and performance of the method was assessed in depth. To briefly recap, the optimisation was performed in the basis of 197 SH trajectories with energies, forces, and nonadiabatic couplings calculated at the SA8-CASSCF(10,8)/SVP level of theory [247]. Due to the signal-to-noise ratio, the optimisation procedure was carried out using the two-step procedure outlined in Chapter 5.4.2. Self-consistency between the two steps of the optimisation was observed, and the best optimisation yielded the parameters $(t_0, \tau_c, \gamma) = (-83 \text{ fs}, 230 \text{ fs}, 3.4\%)$, with both the values of $t_0$ and $\gamma$ being consistent with the approximate values deduced from experiment. The temporal resolution of 230 fs precludes direct observation of any vibrational fine structure in the signal. However, we are able to observe the broad dynamics associated with the singlet and triplet dissociation channels, analysing the structural dynamics in light of eight characteristic trajectories filtered out by the optimisation. Moreover, the best result reproduces observations consistent with the spectroscopic literature, including a singlet ($^1D_2$) to triplet ($^3P_J$) branching ratio of 1:3.38 and a bound population of 28% at 1 ps.
6.3 Results and Discussion

6.3.1 Experimental Signal

Figure 6.1: Heatmaps of the a) experimentally observed $\Delta s M_{\text{exp}}(s,t)$ and b) the theoretical $\Delta s M_{\text{th}}(s,t)$ signal, calculated from the ensemble average of 197 trajectories. Negative time delays correspond to those prior to arrival of the pump pulse, and positive time delays define the change in the signal observed in reference to the system in its ground state equilibrium position.

A comparison of the experimental and theoretically calculated signals represented in the form of $\Delta s M(s,t)$ (Eqn. 6.4) is given in Fig. 6.1. Here the left heatmap corresponds to the experimental signal $\Delta s M_{\text{exp}}(s,t)$, and the right heatmap the theoretical signal $\Delta s M_{\text{th}}(s,t)$ calculated from the ensemble average of 197 SH trajectories. Qualitatively, the experimental and theoretical signals show broad agreement in terms of the main enhancement and depletion observed in the 3.5-6 Å$^{-1}$ region. However, outside of this region the poor signal-to-noise ratio of the experiment is apparent. Turning to the negative time delays, a significant noise floor in the experimental signal is observed. Moreover, the signal increases to exceedingly large values as $s \to 12$ Å$^{-1}$. Generally, poorer signal-to-noise with increasing momentum transfer is to be expected due to the $s^{-4}$ scaling of the scattered intensity. At lower values of momentum transfer ($s \leq 2$ Å$^{-1}$), the theoretical signal posses a small enhancement that is delayed in onset, emerging at around 400 fs. While the experiment observes a clear enhancement just below 2 Å$^{-1}$, it is less well defined in comparison to theory and appears to be
significantly broader. Furthermore, the theoretical signal predicts an additional enhancement with a delayed onset around 6 Å⁻¹, which is not observed in experiment, likely due to the poorer statistics at higher s. We suggest these delayed enhancements correspond to dissociation, however in the representation of ∆sM(s,t) further analysis is challenging due to the quality of the experimental signal.

Figure 6.2: Heatmaps of the a) experimentally observed %ΔI_{exp}(s,t) and b) the theoretically calculated %ΔI_{th}(s,t), calculated from the ensemble average of 197 trajectories. Negative time delays correspond to those prior to arrival of the pump pulse, and positive time delays define the change in the signal observed in reference to the system in its ground state equilibrium position.

In contrast, the experimentally observed signal represented in the form of %ΔI_{exp}(s,t) shows an improved signal-to-noise ratio, as seen in the left panel of Fig. 6.2. Moreover, the increasing intensity as s → 12 Å⁻¹ is not observed here. While the ∆sM_{exp}(s,t) signal depends on the calculated atomic term I_{at}(s) and is scaled by s, the %ΔI_{exp}(s,t) is dived by the laser-off signal I_{off}(s,t ≪ 0) (Eqn. 6.6), taken over multiple time delays. This division likely leads some additional cancellation of errors, removing any of the additional static background that is independent of excitation. Despite the existence of significant noise floor at negative time delays, the signal here is evenly distributed around zero.¹ Again, in terms of the main enhancement and depletion in the range of 3.5-6 Å⁻¹, the experimental signal shows broad agreement with its theoretical

¹. As previously shown by Fig. 5.8 in Chapter 5.6.2, which shows the distribution of the noise levels prior to time-zero, determined to lie at -83 fs.
equivalent in the right hand panel of the figure. At time zero, this enhancement is initially centred on 3.79 Å\(^{-1}\), shifting to 3.45 Å\(^{-1}\) at 1 ps. We also observe broadening of this enhancement from an initial width of 1.4 Å\(^{-1}\) to 1.7 Å\(^{-1}\) at 1 ps, where the given widths are based on the value at which the peaks reach 50% of the maximum. The timescales of the observed shifting and broadening of the main enhancement are indicative of dissociation. Turning to the region just below 2 Å\(^{-1}\), we see better agreement between the delayed enhancement present in both experiment and theory. However, determining the time at which the enhancement first appears is still not clear due to a lack of structure at early times and the observed signal-to-noise. In comparison to Fig. 6.1, the enhancement around 8 Å\(^{-1}\) is better defined, although the experimental signal again shows no clear delayed enhancement around 6 Å\(^{-1}\), unlike its theoretical counterpart.

### 6.3.2 Difference Pair Distribution Functions

![Figure 6.3](image-url)

**Figure 6.3:** Heatmaps of the a) the experimentally observed PDF and b) the theoretically calculated PDF, where the theoretical $\Delta sM$ signal is truncated to match the experimental range $0.95 \leq s \leq 12.25$ Å\(^{-1}\).

First, we turn to address the challenges associated with the extraction of structural information from the experimental signal through the sine-transform in Eqn. 6.5. The resulting experimental difference pair distribution function $\Delta PDF_{exp}(R,t)$, is shown in Fig. 6.3a. It reveals three main enhancements and two depletions that reflect
changes in the distribution of internuclear distances over time. However, the experimental $\Delta PDF_{exp}$ observes several issues, such as high background noise levels, the presence of an enhancement around 1 Å that suggests unphysical C–S distances, and a lack of intensity streaking out to larger values of $R$ as dissociation occurs. For these reasons, further analysis of the structural dynamics through the $\Delta PDF_{exp}$ is not feasible. Interestingly, if one truncates the theoretical signal to the range $0.95 \leq s \leq 12.25 \text{ Å}^{-1}$ to match the experimental range of momentum transfer, the resulting sine-transform observes similar features, as shown in Fig. 6.3b. Here, we also observe the presence of an unphysical enhancement at $R \leq 1$ Å. In addition, while the enhancement around 4 Å exhibits a shift to larger distances and broadens in time, there is distinct lack of any streaking of the intensity out to large values of $R$ as one would expect.

Figure 6.4: The left hand side shows the $\Delta PDF(R, t)$ calculated from the theoretical $\Delta sM_{th}(s, t)$ signal. Regions A-D that correspond to the main features are highlighted. On the right hand side one can see the averaged $\Delta PDF(R, t)$ over each of these four regions. Regions A and C are the two depletions, whereas B and D are the two enhancements. These correspond to the ground state (GS) Wigner distribution and the evolution of the nuclear wavepacket upon the excited state (ES) respectively.

Upon inclusion of data down to $s = 0$ Å$^{-1}$ the theoretical difference pair distribution function $\Delta PDF_{th}$ observes no unphysical enhancement below 1 Å, as seen in Fig. 6.4. The issues previously encountered with the experimental $\Delta PDF_{exp}(R, t)$ are therefore likely the combined result of the poor signal-to-noise and the absence of reliable data below $s = 0.95$ Å$^{-1}$, with the latter a consequence of the hole in the centre of
the detector. Although $\Delta \text{PDF}_{\text{th}}$ is a theoretically calculated entity, it provides a broad overview of the dynamics implied by the model trajectories in light of the experiments 230 fs temporal resolution. The two depletions that show a net decrease at early times correspond to internuclear separations of $1.2 < R < 1.85$ Å (band A) and $2.7 < R < 3.4$ Å (band C), matching the ground state equilibrium $\text{C}--\text{S}$ and $\text{S}--\text{S}$ distances of 1.55 and 3.1 Å, respectively. The corresponding averaged intensity of bands A and C over their respective intervals can be seen on the right hand side of Fig. 6.4, showing the depletion of these regions in time. Adjacent to the two depletions, are two regions in which a transient increase in intensity is observed, identified as band B (1.9-2.6 Å) and band D (3.45-4.2 Å) in the left hand side of Fig. 6.4. These enhancements correspond to a net increase of the internuclear separations. The average values of these bands can also be seen in the right hand side of the figure. Here we observe a rapid increase in the intensity of both band B and D up until around 200 fs, commensurate with the elongation of the bonds from excited state stretching and bending. At its peak, band B has an intensity roughly double that of band D, due to the fact there are double the amount of $\text{S}--\text{S}$ distances vs $\text{C}--\text{S}$ distances. Moreover, the fact that the intensity of band D has a peak intensity just over half that of band B results from some molecules dissociating. As dissociation occurs, the intensity of bands B and D drop, and we observe clear streaking of the intensity of band D to values of $R > 4.2$ Å. As previously discussed, the expected streaking of intensity is absent from the experimental $\Delta \text{PDF}_{\exp}$ in Fig. 6.3.

The issues with the experimental distribution function can be further understood by examining Fig. 6.5, in which we plot each of the $\Delta \text{PDF}$’s discussed so far at a series of individual time delays. In all three cases, the largest magnitude troughs and peaks are observed at similar positions. However, both $\Delta \text{PDF}_{\exp}$ and $\Delta \text{PDF}_{\text{th}}$ (Truncated) are clearly shifted in comparison to $\Delta \text{PDF}_{\text{th}}$, which is calculated over the momentum transfer range $0 \leq s \leq 12.25$ Å$^{-1}$. At earlier times (150-350 fs), the magnitude of the two depleted regions of intensity centred on 1.55 Å and 3.1 Å appear to be near equal for both $\Delta \text{PDF}_{\exp}$ and $\Delta \text{PDF}_{\text{th}}$ (Truncated). In contrast, the magnitude of these troughs appear in a ratio closer to 2:1 in the $\Delta \text{PDF}_{\text{th}}$ case. Since these regions correspond to a depletion of near equilibrium ground state $\text{C}--\text{S}$ and $\text{S}--\text{S}$ internuclear distances, one expects these troughs to occur in a ratio near 2:1. Again, the fact that truncation of the theoretical transform window to match the experimentally available $s$ range recreates these artifacts casts further doubt on the validity of $\Delta \text{PDF}_{\exp}$. From a mathematical
viewpoint, the lack of low $s$ signal where all scattered wavelets scatter in phase as $s \to 0 \, \text{Å}^{-1}$ results in the omission of low frequencies in the sine-transform. This explains why the experimentally observed $\Delta \text{PDF}_{\text{exp}}$ appears to be shifted by a near constant term.

![Graph showing experimental and theoretical PDFs](image)

**Figure 6.5:** Comparison of experimental $\Delta \text{PDF}_{\text{exp}}$ and theoretical difference pair distribution functions for a series of individual time delays. Here $\Delta \text{PDF}_{\text{th}}$ and $\Delta \text{PDF}_{\text{th}}$ (Truncated) refer to the theoretical distribution functions calculated over the ranges $0 \leq s \leq 12.25 \, \text{Å}^{-1}$ and $0.925 \leq s \leq 12.25 \, \text{Å}^{-1}$ respectively.
6.3 Results and Discussion

Figure 6.6 shows the corresponding signal from which each of the ∆PDFs are calculated, again plotted at a series of individual time delays. The experimental signal ∆sM_{exp} can be seen over the range of momentum transfer 0.925 \leq s \leq 12.25 \text{ Å}^{-1}. In comparison, the theoretically calculated signal ∆sM_{th} includes data down to the \( s = 0 \text{ Å}^{-1} \) limit. We also plot ∆sM_{exp}e^{-αs^2}, showing the effect of the exponential dampening function \( e^{-αs^2} \) on minimising the noise at high \( s \). Here the increasing intensity at higher values of \( s \) in the experimental signal is even more apparent, thus showing the need for a heavy dampening function. In UED studies the inclusion of such dampening terms are common, and while they minimise the effect of noise they also inherently reduce the information content of the experiment. Moreover, the theoretically calculated signal shows the presence of a trough just below 1 \text{ Å}^{-1}, which appears to grow in magnitude with time. The timescales over which this feature emerges suggest it is indicative of dissociation. The low frequencies included in this region likely capture the broad dissociative motion of two atoms being pulled apart. Of course, collecting data down to \( s = 0 \text{ Å}^{-1} \) is not experimentally feasible due to the detector hole, and the omission of low \( s \) data is a common problem encountered when transforming the signal into real space. Generally, other studies that manage to perform this transform deal with this issue through an unbiased extrapolation of the signal, or by filling in the low \( s \) region with theoretically calculated values. In this case, due to the fact the low \( s \) region exhibits a trough that grows in magnitude with time, we find the linear extrapolation of the experimental signal to be insufficient as it cannot capture this feature. Moreover, given that artificially filling the experimental signal with theory can potentially bias the experimental observations towards the theoretical model used for comparison, we choose not to perform this step.
Figure 6.6: Comparison of the experimentally observed $\Delta s_{M_{\text{exp}}}$ and theoretical $\Delta s_{M_{\text{th}}}$ signals, shown for a series of individual time delays. Also plotted is $\Delta s_{M_{\text{exp}}} e^{-\alpha s^2}$, which corresponds to the $\Delta s_{M_{\text{exp}}}$ signal weighted by the exponential dampening factor included in the sine-transform of Eqn. 6.5.
6.3 Results and Discussion

6.3.3 Extraction of Structural Dynamics

Given the unphysical artifacts encountered in the experimental $\Delta$PDF, we turn to analyse the structural dynamics using the forward optimisation procedure discussed in Chapter 5. Another advantage of this is that the optimised model can be more easily disentangled in comparison to the experimental $\Delta$PDF, which is a convolution of all geometries observed over the duration of the experiment. While structural insight can be gleaned from the pair distribution functions by looking at the transient increases and decreases in intensity, the experimental $\Delta$PDF would have to be deconvolved in order to gain more in-depth insight into the specific structures present.

The optimisation yields eight dominant trajectories, their character at the end of the simulation, weights, and time taken to reach the dissociation limit, are summarised in table 6.1. Each of these trajectories are representative of a characteristic dynamic evolution of the system, and together they define the minimal representation of the dynamics observed in light of the experiment. The set of weighted trajectories observe a singlet to triplet branching ratio of 1:3.38, with a bound population of 28% remaining at 1 ps. These values are in reasonable agreement with recent femtosecond photo-electron spectroscopy measurements [230, 232], and nanosecond experiments [233–239], providing branching ratios around 1:2.5 and 1:3 respectively. The 1:3 ratio from nanosecond experiments is based on an average of a number of observed values ranging between 1:4 and 1:6. Note that for both previous nano and femtosecond experiments, these values are based on the asymptotic values rather than the ratio at 1 ps. Given that triplet dissociation occurs on a longer timescale, the small variation of our observations from some of these reported values is expected.

<table>
<thead>
<tr>
<th>Weight (%)</th>
<th>Type</th>
<th>Dissociation Time (fs)</th>
</tr>
</thead>
<tbody>
<tr>
<td>10.20</td>
<td>Bound</td>
<td>-</td>
</tr>
<tr>
<td>44.73</td>
<td>Triplet</td>
<td>433</td>
</tr>
<tr>
<td>10.35</td>
<td>Singlet</td>
<td>132</td>
</tr>
<tr>
<td>17.81</td>
<td>Bound</td>
<td>-</td>
</tr>
<tr>
<td>4.43</td>
<td>Singlet</td>
<td>523</td>
</tr>
<tr>
<td>1.66</td>
<td>Singlet</td>
<td>112</td>
</tr>
<tr>
<td>3.40</td>
<td>Triplet</td>
<td>776</td>
</tr>
<tr>
<td>7.43</td>
<td>Triplet</td>
<td>949</td>
</tr>
</tbody>
</table>

Table 6.1: Summary of the 8 dominant trajectories that result from the best optimisation. These include two bound, three singlet and three triplet trajectories with a total weight of 28 %, 16.45 %, and 55.55 %. Dissociation times are given where applicable.
Turning to Fig. 6.7, which shows the C–S distances and associated S–C–S angle for the four most dominant trajectories, we proceed to analyse the structural dynamics implied by the fitted trajectories. Here, the two upper panels correspond to singlet S\(^{(1}\text{D})\) dissociation and the two bottom panels to triplet S\(^{(3}\text{P})\) dissociation. Across all key trajectories, we immediately observe rapid bending of the S–C–S angle following excitation, this is coupled with C–S stretching that is initially largely symmetric in nature. The top panel of Fig. 6.7 shows the dominant singlet trajectory (10.35%) which exhibits rapid dissociation. In this trajectory, both C–S bonds undergo one cycle of elongation and shortening before dissociation. However, one of the C–S bonds (shown in dark green) exhibits a significantly larger elongation than the other. After this initial stretching, the molecule returns to a near-linear geometry and momentum is transferred from the more rapidly moving sulphur to the other sulphur, which is then ejected during a simultaneous sharp bending motion resulting in an angle of 111 degrees at the time of dissociation. The remaining C–S fragment continues to undergo small amplitude vibrations while rotating. The next most significant singlet trajectory (4.43%) in the second uppermost panel, dissociates on a much slower timescale. Following the initial largely symmetric vibration, this trajectory then proceeds to vibrate in a mostly asymmetric fashion, with one of the C–S bonds (shown in light green) undergoing low-frequency large amplitude motion. This C–S bond initially extends to lengths of 2.3 \(\text{Å}\) at around 150 fs, nearly reaching the dissociation limit. However, the sulphur atom remains bound, and the trajectory then explores a much wider range of available S–C–S angles until dissociation.

We now shift our focus to the two most significant triplet trajectories in the bottom two panels of Fig. 6.7. The second lowest panel, showing the dominant triplet trajectory (44.73%), initially observes a largely asymmetric vibration which then becomes more symmetric as momentum is transferred between the C–S bonds after a return to a near-linear geometry just before 200 fs. The bottom panel shows a triplet trajectory (7.43%) that is significantly longer lived. A series of smaller amplitude vibrations are observed up until dissociation occurs at 949 fs. For both triplet trajectories, we note that the initial bending motion is stronger in comparison to the singlets, reaching angles of 121 and 112 degrees respectively within the first vibration period. This observation is commensurate with a previous theoretical study where the majority of ISC occurs at strongly bent conformations [247].
Figure 6.7: Atomic C–S distances (Å) and S–C–S bending angles (degrees) in CS₂ for the four most significant dissociating trajectories. The dashed black line indicates the ground state equilibrium C–S bond length. The two top sets of panels show the two dominant singlet trajectories, while the two bottom sets of panels show the two dominant triplet trajectories. All trajectories are shown to the point of dissociation.
The optimisation procedure is capable of capturing three different classes of dynamics based on their electronic character (bound, singlet, and triplet), these can be differentiated largely on their dissociation timescales. The majority ($\approx 73\%$) of the singlet class is defined by two trajectories that represent ballistic dissociation, occurring at 132 fs and 112 fs. Meaning the singlet dissociation is dominated by rapid breaking of the C–S bond following IC. The remaining 27% of the singlet contribution comes from a trajectory with a much longer dissociation time of 523 fs. As such, we are able to capture both the rapid singlet dissociation, and the slower leakage of a smaller amount of singlet product at later times. In addition, 80% of the total triplet class is defined by a trajectory with a dissociation time of 433 fs, with the remaining 20% coming from two trajectories with longer dissociation times of 776 and 949 fs, significantly longer than any of the trajectories comprising the singlet contribution. From the set of optimised trajectories we determine the weighted average for the dissociation times of the singlet and triplet product channels to be 235 and 523 fs, respectively.

**Figure 6.8:** Populations of the bound, singlet ($^1D$), and triplet ($^3P$) channels determined from the optimised set of model trajectories.
While the dominant triplet dissociation occurs on longer timescales, there is a degree of overlap with longer lived singlet dissociation. This can be observed in Fig. 6.8, which shows the changes in the population of the two dissociation channels. While the majority of the singlet dissociation occurs within the first 200 fs, small increases in the singlet dissociation channel can be seen at later times where the triplet dissociation dominates. It is worth noting that the optimised model does not predict any triplet dissociation before approximately 250-300 fs, however evidence of this can be seen in previous spectroscopic studies [229]. The model's failure to capture the small amount of early-time triplet dissociation can be attributed to two main factors. Firstly, the delayed enhancement around 2\,Å$^{-1}$, which is correlated with dissociation, only appears faintly around 200 fs before becoming more apparent at later times (as illustrated in Fig. 6.2a). Secondly, at early time delays, the molecular geometries of the singlet and triplet dissociation channels are similar, making it challenging to differentiate between them. While some scattering experiments have allowed for the identification of the electronic states directly involved, we do not see any evidence of such signatures in our experiment. Hence, we are unable to discern further differences in the singlet and triplet character of the states involved.

Using the set of optimised trajectories we decompose the model scattering signal into its triplet, singlet, and bound contributions, as shown in Fig. 6.9. Here, panel (a) shows the total model signal, and panels (b-d) show each of the three different contribution. Generally, within the constraints of the experimental resolution, the bound contribution to the signal shows little time dependence following excitation. Here we observe small undulations in the enhancements centred just below 4 and 8 \,Å$^{-1}$, which we attribute to the combination of bending and stretching vibrational modes in the excited states. In contrast, the dissociative contributions exhibit distinct changes over time, characterized by the delayed enhancements just below 2, 6, and 10 \,Å$^{-1}$, which are absent in the bound contribution to the signal. The onset of these enhancements correlates with the calculated dissociation times for the archetypal trajectories (represented by black lines), providing further evidence that they represent signatures of C$-$S bond breaking. The combined effect of these contributions, as observed in the experiment, is shown in panel (a), where the delayed enhancement only becomes noticeable at 400 fs. As the intensity of this enhancement increases, the main enhancement of the model signal shifts from 3.65 to 3.43 \,Å$^{-1}$ at 1 ps, suggesting this is also a consequence of dissociation. The observed shift in the model signal is consistent with the previously discussed shift observed in the experiment, where it shifts from 3.79 to 3.45 \,Å$^{-1}$. The enhancement centred at 3.65 \,Å$^{-1}$ can also be thought about in terms
of its singlet and triplet contributions, which observe a total shift of 0.15 and 0.25 Å\(^{-1}\) respectively, as seen in panels (c-d). The shifting of the enhancement results from a combination of bending and stretching modes, but ultimately mostly depends on the asymmetric stretch. Therefore, we attribute the greater shift of the triplet signal due to the larger amount of dissociation that occurs through this channel.

Furthermore, the presence of the enhancement just below 2 Å\(^{-1}\) in the model signals of Fig. 6.9 is commensurate with the experimentally observed signal in Fig. 6.2a. Although both dissociation channels are detected simultaneously in experiment, the initial increase in intensity likely corresponds to S\(^{1}D\) formation, while the subsequent rise results from a combination of S\(^{3}P\) and a minor amount of slower S\(^{1}D\) formation. This is supported by Fig.6.10, where we plot the rise of the enhancement observed in experiment and the optimised model, along with the model's individual singlet and triplet contributions. Qualitatively, the rise of the model signal (\%)\(\Delta I_{\text{mod}}\) closely mirrors the experimental data (\%)\(\Delta I_{\text{exp}}\). In addition, the singlet \(^{1}D\) contribution demonstrates a rapid increase before stabilizing around 500fs, while the triplet \(^{3}P\) takes longer to emerge, only becoming positive after 400 fs. Together, the known time scales of the dissociation channels and the fact that the rise in intensity is correlated with dissociation, provide evidence to support the conclusion that the early rise in intensity is dominated by the singlet channel, whereas the later rise is mostly dominated by the triplet. This feature is therefore a key target for future scattering experiments that may be capable of resolving the different electronic states involved.
Figure 6.9: Decomposition of the optimised model $\Delta I_{\text{mod}}(s,t,w)$ signal (shown in panel a) into b) bound, c) singlet ($^1D$), and d) triplet ($^3P$) contributions. Black lines mark dissociation times calculated from the corresponding optimised trajectories in table 6.1.
6.3 Results and Discussion

Figure 6.10: Rise in the intensity at the centre of the enhancement just below 2 Å\(^{-1}\). Shown are the intensity rises determined from experiment \(\%\Delta I_{\text{exp}}\), the total signal of the optimised model \(\%\Delta I_{\text{mod}}\), and the contributions from both the individual singlet (\(^1D\)) and triplet (\(^3P\)) channels.

Ultimately the optimisation procedure allows us to not only analyse the structural dynamics implied by the characteristic trajectories in light of the experimental constraints, but also provides a means for disentangling the different contributions to the experimental signal that are not immediately obvious from the experimental signal or \(\Delta\text{PDF}\) alone. Here, the procedure was able to distinguish the differences in the observed shifting of the signal and the rise of the delayed enhancement just below 2 Å\(^{-1}\), as well as the final state populations and dissociation life times. This provides a powerful tool that greatly aids in the analysis of challenging experimental data sets, particularly in cases where data at sufficiently low momentum transfer cannot be collected. As a result, it overcomes the previously discussed limitations in analysing the structural information through the transformation of the signal into real space.
6.3 Results and Discussion

6.3.4 Temporal Resolution

As a final aside, we note that the observation of the photodissociation dynamics of \( \text{CS}_2 \) would have benefited from improved temporal resolution. In Fig. 6.11 we see the calculated theoretical signal at a range of different \( \tau_c \) values that mimic the IRF, ranging from 0 fs to 230 fs, as observed in the current work. It is evident that more detailed features that capture the vibrational fine structure only become apparent with temporal resolutions less than 100 fs. Given the slowest vibrational mode, the S–C–S bending mode, has a period of 84 fs, this is unsurprising. Comparing Fig. 6.11a and b, corresponding to absolute and 80 fs temporal resolution, it is clear that improvements in time-resolution are required to further resolve aspects of dissociation, although the highly disperse nature of the wavepacket and nearly statistical dissociation rate also fundamentally limit the observation of dissociation markers in the signal. With increased temporal resolution, one could look to exploit the periodicity of different vibrational modes to determine their contributions to the observed signal. However, the asymmetric stretch that is fundamentally responsible for the shifting of the main enhancement occurs on a timescale of approximately 22 fs. This fine structure can only be seen in the main enhancement of Fig. 6.11a, however lower frequency vibrations that result from the bending mode can be faintly seen in Fig. 6.11b. Therefore, with a temporal resolution of around 80 fs, additional information on the bending mode should be available.
Figure 6.11: Heatmaps of the theoretical difference signal $\% \Delta I_{\text{th}}(s,t)$ at a range of values of the Gaussian width $\tau_c$, used in the convolution that approximates the IRF. Note, all difference signals $\% \Delta I_{\text{th}}(s,t)$ are calculated from the ensemble average of 197 trajectories. Panels a) through d) correspond to widths of 0, 80, 150, and 230 fs, with 230 fs the resolution observed in this work. At 80 fs, the vibrational fine structure becomes apparent.
6.4 Conclusions

In this chapter, we have investigated the photodissociation dynamics of \( \text{CS}_2 \) using UED following excitation with a 200 nm pump pulse. Although we collected data across a wide range of momentum transfer, we encountered challenges in analyzing the structural dynamics through the transformed pair distribution function. These challenges were primarily due to the lack of low \( s \) data and the degradation of signal-to-noise ratio at higher \( s \) values. The absence of low \( s \) data prevented us from capturing the low-frequency components associated with the broad dissociation motion of the two atoms being pulled apart. Consequently, the resulting \( \Delta \text{PDF} \) exhibited not only a constant shift but also unphysical internuclear separations of less than 1 Å. These issues could not by resolved by simple linear extrapolation to zero. In addition, we do not fill in the low momentum transfer region with theory to avoid biasing the experimental observations towards theory.

Direct extraction of the structural dynamics from the experimental signal was challenging, due to competing reaction pathways being observed together and the available temporal resolution. However, by coupling the experimental analysis with the optimisation procedure from Chapter 5, we are able to deduce the characteristic motions from the set of trajectories filtered out by the optimisation, taking into account the experimental constraints. The optimised model yields branching ratios, dissociation timescales, and populations that are congruent with other experimental studies. Furthermore, by decomposing the model signal into its different contributions, we were able to disentangle differences in features correlated with dissociation.

The rapid dispersion of the wave packet in this system inherently imposes limitations on the resolution of the structural dynamics, a phenomenon that may also occur in other molecular systems [248]. Another factor to consider is that the current conclusions are somewhat limited by the available temporal resolution. With improved temporal resolution, the vibrational fine structure in the signal should be apparent, providing the opportunity for a more detailed analysis of the dissociation dynamics in terms of the available bending and stretching modes. Improvements in temporal resolution and overall signal levels would justify a theoretical treatment that goes beyond the IAM, potentially allowing one to unravel electronic effects in the dynamics. [23, 84–87, 249]
In this thesis, we presented new computational methods to interpret trajectory-based quantum molecular dynamics simulations and ultrafast imaging experiments on photoexcited processes in molecules. The methods improved the workflow of experimental and theoretical analysis, and provided valuable insights. We considered three main aspects; identifying concentrations of reaction flux in complex trajectory simulations, extracting information from time-resolved observables, and interpreting the photodissociation dynamics of CS₂ as observed by MeV-UED.

In Chapter 3, we explored a time-series approach to clustering semi-classical trajectory simulations based on a single-valued measure of similarity over all time. In an application to CHD, the time-series approach was shown to provide a reasonably easy way of identifying the dominant pathways that lead to product formation. Furthermore, due to the underlying distribution of the data, we saw that this was the case whether we explicitly accounted for periodicity or not. However, the algorithm struggled to separate trajectories that remain on the excited S₁ (2A) state from ground state cEc-HT trajectories, and the results did not always reflect actual electronic state populations. Ultimately, the ability of the algorithm to provide deep insights into the photoexcited dynamics was limited. This however, laid the foundations for Chapter 4.

Next, in Chapter 4, we considered a pattern mining approach to identifying spatio-temporal clusters in trajectory data. This involved the development of two new algorithms that work in tandem. The first algorithm identified the spatial clusters at each time step by clustering trajectories based on their assignment to maxima in a discretized density field, represented by a series of Gaussian kernels. This amounted to mapping the probability density to a lower dimensional feature space, and allowed identification of key concentrations of reaction flux at each time step without the need for re-parameterisation, unlike existing algorithms. The second algorithm focused on mining the temporal patterns between spatially clustered trajectories. For this we
introduced a new definition of a spatio-temporal cluster, known as a relaxed moving cluster. This overcame the consecutive time constraint and the tree representation we employed facilitated the detection of the moment when groups of trajectories diverge, capturing the times at which the wavepacket bifurcates upon reaching an avoided crossing, conical intersection, or cusp in the PES. Both algorithms were originally applied to CHD without explicitly accounting for periodicity. However, the results obtained when accounting for periodicity by projecting each angle onto its coordinates \((\cos, \sin)\) on the unit circle were largely similar, as we anticipated from the findings of Chapter 3 and the underlying distribution of the data. Importantly, in both cases this provided deeper insight in comparison to the time-series approach, capturing changes in state populations and the initial bifurcation of the wavepacket around 34-36 fs. While we originally saw improved identification of trajectories representing the minor ground state cEc-HT product, this low probability pathway was not detectable in the periodic treatment of angles. However, this is not surprising given the introduction of two correlated features for each angle, thus increasing the dimensionality. Finally, we demonstrated how the relaxed moving clusters can be used to identify a reduced yet representative sample of trajectories, appropriate for accurate observable calculations too computationally expensive to perform on the entire ensemble.

In Chapter 5 we introduced an optimisation-based approach to inverting time-resolved data, focusing on its application to scattering observables. This was applied to both x-ray and electron scattering data sets, describing the ring-opening of CHD and the photodissociation of \(\text{CS}_2\), respectively. The procedure reconstructed the experimental observable by minimising a target function with respect to a set of trajectory weights, utilising a multi-start nonlinear TRRLS algorithm. In the case of high-resolution x-ray data, we discussed a global approach in which we scanned a range of global parameters and optimised the trajectory weights in each case. For the lower-quality electron scattering data, a two-step procedure was introduced in which the global parameters were first determined independent of the trajectory weights. In both examples, the optimisation converged on a handful of trajectories, corresponding to a minimal basis representation of the dynamics in light of experimental distortions. This minimal subset captured key dynamic motifs, such as the pathways to ring-opening and the distribution of HT isomers in CHD or dissociation channels in \(\text{CS}_2\). Moreover, the optimisation reproduced observed values from other independent experiments. Importantly, the method was shown to be robust with respect to experimental noise, providing a tool for extracting useful information from challenging data and potentially identifying and bridging shortcomings in both experiments and theory.
Finally, in Chapter 6, we took a closer look at the photodissociation dynamics of CS$_2$ as studied by MeV-UED. A lack of low $s$ data and the poor signal-to-noise at higher $s$, precluded analysis of structural dynamics using the nuclear pair distribution function. Instead, we employed the optimisation procedure from Chapter 5, identifying key markers of dissociation in the signal, as well as other observable quantities such as the singlet ($^1D_2$) to triplet ($^3P_J$) branching ratio, dissociation timescales, and the bound population remaining at 1 ps, all of which agreed with other independent spectroscopic measurements. Following excitation, rapid S–C–S bending was observed, which continued as the molecule entered a combination of stretching modes. The singlet dissociation was found to occur on significantly faster time scales, with rapid C–S bond cleavage. While the experiment constitutes an important first application of scattering based experimental techniques to CS$_2$, further insights will ultimately require better time resolution.

Looking ahead, there are many exciting avenues to explore based on the methods introduced in this thesis. Currently, the spatial clustering algorithm introduced in Chapter 4 is in the process of being rewritten in a continuous form that can treat periodicity in a variety of ways. With this in hand, we look to assess different treatments of periodicity and aim to identify the best approach in the context of QMD trajectory data. Another potentially interesting avenue of exploration could involve the use of Wasserstein distances [250]. Intuitively, the Wasserstein distance defines the cost of transporting one probability distribution into another, and could potentially be used as a metric in merging different Gaussian distributions defined on the feature space. In this thesis, we addressed the question of how to physically identify spatiotemporal patterns in trajectories. So far, this has involved the use of simplistic internal coordinates and populations. However, in applications to larger systems, other representations of the nuclear geometries may allow one to describe several vibrational modes at once. Alternative coordinate systems may provide easier mapping of the momenta, which is not straightforward in the case of internal coordinates. Constructing a feature space using dimensionality reduction techniques is another route that should be explored. However, the feature space used must account for the time-dependence of the data, meaning that we favour the exploration of tICA over standard PCA. Alternatively, autoencoders may be suited to the problem at hand, with their usage in data compression and feature selection well documented [251, 252]. These NN’s use an encoder to map the data to a lower dimensional latent space representation, capturing the key features and excluding irrelevant information. The original input is then reconstructed from the compressed latent space using a
7. Conclusions and Outlook

decoder, minimising the error between the reconstructed and input data. This latent space representation can then be used as a basis for clustering [253–255]. While it is possible to generate latent space structures that are representative of the distances in the original data space [253], care has to be taken due to the non-linearity of the latent space. Therefore, Euclidean based distances cannot be applied to the latent space. Such an approach may provide a way to encode a wider range of information such as geometries, momenta, and state populations. However, introducing too much complexity into the construction of the feature space could result in a computationally expensive and time consuming method. There is an argument that if the aim is to improve the interpretation and workflow of trajectory analysis, the approach should have a low barrier to entry.

Regarding the forward optimisation procedure, exploring alternative algorithms beyond a multi-start approach to nonlinear TRRLS is worth considering [220, 221]. While we tried to utilise a genetic algorithm (GA), it struggled to find stable and meaningful solutions compared to TRRLS. GA optimisations were resource-intensive, required large population sizes and extensive parameter tuning, without converging to as low minima as TRRLS. This was likely due to the inherently random search and non-deterministic nature of GAs. However, combining TRRLS with GAs could offer a powerful tool by leveraging local and global search capabilities. In a combined procedure, TRRLS could quickly identify a range of local minima, and the resulting trajectory weights could serve as input to a GA, confirming global minima. Moreover, reduced samples of trajectories identified by our pattern mining algorithm provide potential to extend analysis of scattering experiments beyond IAM, unravelling more intricate electronic dynamics. Given that our method is capable of refining the wavefunction, correcting branching ratios and quantum yields, one does not have to worry about the distribution of trajectories in the identified sample set. Furthermore, as each particular experimental technique is typically sensitive only to a subset of features of the evolution of the excited system, combining data from several different types of experiments in the optimisation procedure should allow for a more complete description of the dynamics. More generally, ongoing research into inversion methods that do not require trajectories is promising. However, due to the lack of unique solutions and the under-determined nature of inversion, it is not clear if these approaches are viable for experiments involving molecules with limited alignment. Currently, even advanced NN...
approaches rely on trajectories for the training phase [82]. One exciting approach, currently in development, attempts to unwrap the dynamics based on sampling the normal modes of the initial time-zero structure using simulated annealing. However, applications to real experimental data have yet to be demonstrated.

With advancing UED technology, the future of these experiments lies in going beyond the IAM to observe both nuclear and electronic dynamics. However, analysing the resulting pair distribution functions is challenging, given that the collection of data at low momentum transfer remains problematic and may result in unphysical distribution functions. Extrapolation of the data is often not sufficient, and while several high-impact publications take the approach of filling in the experimental data with theory [23, 256], the impact of this on observations needs further careful assessment. We advocate for developing inversion methods covering the experimental data range or algorithms that improve data extrapolation accuracy, rather than risk biasing experimental observations towards theoretical values. Furthermore, scattering from randomly orientated ensembles of gas-phase molecules greatly reduces the available information content. In exploiting molecular alignment UED shows promise for full 3D structural retrieval, without the need for comparison to theoretical models. Such approaches have been successful in applications to simple molecules, however full 3D structural retrieval of complex molecules requires the further development of experimental technology and retrieval algorithms [65].

While spectroscopic methods have provided valuable insights into CS$_2$, scattering allows for the direct observation of the structural evolution. The timescales associated with the asymmetric stretch pose challenges for UED, making x-ray scattering an attractive alternative for capturing the fundamental vibrational motion. A recent x-ray scattering study, which was not discussed in this thesis, successfully revealed the distribution of dissociation velocities in both the singlet and triplet channels, providing further evidence that the triplet channel dominates at later times and exhibits higher dissociation velocities [257]. However, direct observation of the vibrational motion still eludes us. On the theory side, there is still much to uncover about the dynamics of CS$_2$, with existing semi-classical simulations either only considering the early time singlet dynamics or not including the full range of excited states accessible at the excitation energy in experiments [228, 247]. Addressing these limitations, ongoing efforts by the author involve the calculation of full-dimensional PESs. However, determining a suitable active space in CASSCF calculations remains a challenge, with an accurate representation of both the excitation energies and dissociation barriers
requiring a large active space [247]. Additionally, the mixing of character between electronic states necessitates including as many as 18 electronic states to ensure smooth potentials. Consequently, this poses a formidable computational task. With the availability of these PESs, fully quantum grid-based simulation methods become accessible, offering deeper theoretical insights into the quantum dynamics at play.

The computational methods presented in this thesis serve to aid the analysis and interpretation of both simulations and experimental data, further advancing our understanding of photoinduced quantum dynamics in molecular systems. Photoexcited reaction dynamics is inherently complex and future research will undoubtedly involve a combination of complementary experimental techniques, theoretical models, and computational tools. Together, these integrated approaches hold the potential to unravel the intricacies of photoexcited processes.
In Chapter 3 and 4, we concluded the reported results were relatively robust whether the periodicity of angular coordinates were accounted for or not. However, this may not always be the case, and in this section we proceed to discuss the limits at which periodicity may have a larger effect on the results.

Unless one can guarantee that all angles lie on the interval $[0, \pi]$, we note that best practice would be to ensure periodicity is accounted for directly. However, the risk of issues with periodicity arising also depends on the distribution of the angular data. To briefly demonstrate this point, we turn to Fig. A.1, where we generate two different periodic data sets on the unit circle, in both cases three Gaussian's are sampled on the interval $[0, 2\pi]$. Here, the top row corresponds to a data set that spans a range larger than $[0, \pi]$ on the unit circle, but is largely separable. In the first column we see the result of clustering using DBSCAN with an $L_2$ norm on the raw angle $\phi$ without account for periodicity. In the second column, we see the result of using the $L_2$ norm on the projected $(\cos, \sin)$ coordinates of the angle. Despite the fact that the angles span a range larger than $[0, \pi]$, both approaches produce identical results, identifying three clusters as one would expect. However, in the bottom row, we see a data set where two distributions broadly overlap over the top portion of the unit circle. Here, the correct result corresponds to the identification of two clusters as the two distributions spread out over the top of the unit circle are inseparable when correctly accounting for the periodicity of angles. As a result, using the $L_2$ norm on the raw angles $\phi$ fails and three clusters are identified. In the second column, we see that only two clusters are identified when projecting the angle onto its $(\cos, \sin)$ coordinates, as periodicity is accounted for here.
A. Failures of Non-periodic Distance Metrics

Figure A.1: Effect of the distribution of angular data on identifying the correct clustering result. In the top and bottom row we have two different data sets, both spanning the range $[0, 2\pi]$. In column one we colour the data points according to the Gaussian distribution it was generated from. In columns two and three we see the result of clustering using an $L_2$ norm on the raw angular data $\phi$ and the projected values $(\cos, \sin)$ respectively.

It is worth noting, however, that while the simple example in Fig. A.1 demonstrates the situations where non-periodic similarity measures fail in 1D, it also likely overstates the problem when considering multidimensional data. In the multidimensional case, while some angular coordinates may overlap, the distribution may still be separable with respect to the other dimensions. This is demonstrated in Fig. A.2, which shows two 2D distributions on the unit circle. Here, the top row corresponds to the first coordinate ($\phi_1$), and the bottom row to the second coordinate ($\phi_2$). In column one the data points are again coloured according to the distribution they are generated from. Note that while the angles span a range larger than $[0, \pi]$ and overlap in terms of $\phi_1$ (top row), they are separable on the hyper-sphere in terms of $\phi_2$ (bottom row). As a result of this, we see that the clustering results shown in the second and third columns, corresponding to non-periodic and periodic treatments, give identical results and correctly predict two clusters.

We conclude this discussion by noting that while using distance metrics based on the $L_2$ norm on the raw angular values does not account for periodicity and does not correctly reflect the distances between angles if they span a range greater than $[0, \pi]$, the clustering result identified may be equivalent or highly similar to using a
Figure A.2: Effect of the distribution of angular data on identifying the correct clustering result in a 2D data set with two angular coordinates. The top and bottom rows correspond to the angular coordinates \( \phi_1 \) and \( \phi_2 \). In column one we colour the data points according to the Gaussian distribution it was generated from. In columns two and three we see the result of clustering using an \( L_2 \) norm on the raw angular data \( \phi \) and the projected values \((\cos, \sin)\) respectively.

periodic treatment. This depends on the distribution of the data and its separability in higher dimensions. In cases where the data spans the unit circle and the data overlaps significantly over this range, then one is likely to encounter issues with periodicity. However, as a matter of best practice and ensuring a general treatment that can applied to any angular data set, one should directly account for the circular statistics of angles as this removes the risk of identifying erroneous clustering results.


