COULOMB EXPLOSION AS A HIGH-DIMENSIONAL PROBE OF SINGLE MOLECULES

Dissertation zur Erlangung des Doktorgrades
an der Fakultät für Mathematik, Informatik und Naturwissenschaften
Fachbereich Physik der Universität Hamburg

vorleget von

BENOÎT RICHARD

Hamburg
2024

Gutachter/innen der Dissertation:	Dr. Ludger Inhester Prof. Dr. Robin Santra
Zusammensetzung der Prüfungskommission:	Dr. Ludger Inhester Prof. Dr. Robin Santra Prof. Dr. Daniela Pfannkuche Prof. Dr. Markus Ilchen Dr. Kartik Sankaran Ayyer
Vorsitzende/r der Prüfungskommission:	Prof. Dr. Daniela Pfannkuche
Datum der Disputation:	15.10.2024
Vorsitzender des Fach-Promotionsausschusses PHYSIK:	Prof. Dr. Markus Drescher
Leiter des Fachbereichs PHYSIK:	Prof. Dr. Wolfgang J. Parak
Dekan der Fakultät MIN:	Prof. Dr.-Ing. Norbert Ritter

Declaration on oath

I hereby declare and affirm that this doctoral dissertation is my own work and that I have not used any aids and sources other than those indicated.

If electronic resources based on generative artificial intelligence (gAI) were used in the course of writing this dissertation, I confirm that my own work was the main and value-adding contribution and that complete documentation of all resources used is available in accordance with good scientific practice. I am responsible for any erroneous or distorted content, incorrect references, violations of data protection and copyright law or plagiarism that may have been generated by the gAI.

July 27, 2024 B. Richard

Date Signature of doctoral candidate

ACKNOWLEDGEMENTS

The research presented in this thesis has been conducted with constant and insightful feedbacks from my supervisor Ludger Inhester, for which I am deeply grateful.

I thank my thesis director Robin Santra for his insights and his help in shaping this project.

In addition, I wholeheartedly thank

Zoltan Jurek, for his unyielding support with the simulations and tireless willingness to help,

Rebecca Boll and Till Jahnke, for sharing their experimental data and their repeated enthusiasm for my work and my vision of the problem,

Anthony, Eva, Ilke, John, Lea, Maude, Mukthar, Nidin, Niels, Sebastiaõ, Sourav, and Xavier, for their valuable, and always supporting, comments and suggestions during the writing of this thesis.

This work is funded by the Cluster of Excellence 'CUI: Advanced Imaging of Matter' of the Deutsche Forschungsgemeinschaft (DFG) - EXC 2056 - project ID 390715994.

ZUSAMMENFASSUNG

In den vergangenen Jahrzehnten wurden zahlreiche Methoden zur Untersuchung einzelner Moleküle auf der Femtosekunden-Zeitskala entwickelt. Eine dieser Methoden ist die Röntgenstrahlen-induzierte Coulomb-Explosion, die durch der Einsatz von Röntgen-Freie-Elektronen-Lasern ermöglicht wurde. Hierbei werden intensive, ultrakurze Röntgenpulse auf einen Gasstrahl aus Molekülen geschossen. Die Wechselwirkung eines Röntgenpulses mit einem einzelnen Molekülen verursacht dessen schnelle und starke Ionisation und führt dadurch zu einer ultraschnellen Dissoziation in seine atomaren Fragmente. Danach werden die Impulse der Atomfragmente gleichzeitig gemessen. Diese Methode wurde bereits erfolgreich auf Moleküle, die aus nur wenigen Atomen bestehen, angewandt, die Ausweitung auf größere Systeme bleibt jedoch eine Herausforderung. In dieser Arbeit nutzen wir Simulationen und fortgeschrittene Analysemethoden, um das Potenzial der Coulomb-Explosion und die Anwendbarkeit auf größere Moleküle zu untersuchen. Dabei konzentrieren wir uns auf das 2-Iodpyridin-Molekül (C_5H_4NI), das experimentell untersucht wurde. Wir zeigen, dass die gleichzeitige Messung von nur wenigen Fragmenten genügt, um mittels Coulomb-Explosions-Bildgebung aussagekräftige Bilder des untersuchten Moleküls zu erhalten. Durch eine detaillierte Analyse unserer Simulationen stellen wir fest, dass Kollisionen zwischen Atomen während der Explosion messbare Korrelationen zwischen den Impulsen der gemessenen Fragmente hervorrufen. Dies verdeutlicht den großen, in der hohen Dimensionalität der Daten verborgenen Informationsanteil. Die Berücksichtigung von Grundzustandsfluktuationen des Moleküls in den Simulationen zeigt, dass diese einen Fingerabdruck in den Ionen-Ionen-Korrelationen hinterlässt. Dies weist darauf hin, dass die Coulomb-Explosion genutzt werden kann, um die die komplexen kollektiven Merkmale einer Verteilung von Molekülstrukturen abzubilden. Um diese Analysen auf Experimente anzuwenden, entwickeln wir einen Algorithmus zum systematischen Auffinden der Korrelationen in experimentellen Daten. Die entwickelte Methode ist trotz der zufälligen Orientierung des Moleküls im Laborsystem und der begrenzten Detektoreffizienz erfolgreich und eröffnet somit neue Wege, Coulomb-Explosionsdaten zu interpretieren und einen Teil ihres bisher ungenutzten Potenzials auszuschöpfen.

ABSTRACT

The past decades saw the development of X-ray induced Coulomb explosion imaging as one of several methods to perform single-particle molecular imaging on a femtosecond timescale. It has been enabled by the advent of x-ray free electron lasers. It works by firing intense ultrashort x-ray pulses on a gas jet of molecules. A single pulse rapidly and strongly ionizes a single molecule, resulting in its ultrafast dissociation, after which the momenta of the atomic fragments are measured in coincidence. This method has been successfully applied to molecules composed of few atoms, but its extension to larger systems remains a challenge. In this thesis, we use simulation and advanced analysis techniques to explore the potential of Coulomb explosion imaging and its applicability to larger molecules, focusing on the 2-iodopyridine molecule (C_5H_4NI), which has been measured experimentally. We show that, despite detecting only few fragments in coincidence, Coulomb explosion imaging provides meaningful images of the molecule. Analyzing our simulations in detail, we establish that collisions between atoms during the explosion induce measurable correlations between the momenta of the measured fragments, indicating that a large amount of information is hidden in the high dimensionality of the data. We then include the vibrational ground-state fluctuations of the molecule in our simulations, and show that they also leave a fingerprint in the ion-ion correlations, demonstrating that Coulomb explosion can be used to image the complex collective features of a distribution of molecular structures. To apply those analyses to the experiment, we develop an algorithm to systematically find the correlations in the experimental data. Our method succeeds despite the random orientation of the molecule in the lab frame and the limited efficiency of the detectors, thereby opening new ways to interpret Coulomb explosion data, and allowing to exploit some of its untapped potential.

When going back from work, late a night, I went through a park, unlit and dark like most public parks in Hamburg. When I exited on the other side, I saw a man, under one of the streetlights, searching the ground. He burst out suddenly, holding up a banknote:

"What a luck, to find 50 euros while looking for my lost keys."

"But what about your keys?" I asked "Did you lose them around here?"

"No, not at all. I lost them in the middle of that park you came from."

"But then" I asked, confused, "why are you not searching the park?"

He answered as if it was evident:

"The park is too dark to find anything, but, here, there is light."

PREFACE

There is no scientific method.

This is not a new idea, but a rather old realization [23]. Even so it may sound like a criticism, this statement is in fact a celebration of science. To highlight the absence of a single method is to praise the vitality of the scientific world, and the diversity of thoughts it nurtures. Science is a living animal evolving to face the new challenges that it itself uncovers. It also means there is no sure path to progress, that our methods should be constantly questioned and evaluated, changed if failing, abandoned if unsalvageable. Yet, according to my experience during my PhD, the question of the scientific method is rarely discussed, if ever acknowledged, even when its absence is blatant.

This thesis, for example, is rooted in a situation that was barren of any such consideration. The seed from which it is now blooming was a seemingly simple request: "Find something interesting in Coulomb explosion data". Generated from simulations, the data was available and plentiful (1 TB!), so the project seemed very reasonable to me. Back then I did not notice that it did not contain any hint of the methods I had been taught were scientific: there was no theory to be tested, the agreement with the experiment was already considered satisfactory, and, therefore, there was no plan to challenge any previous hypothesis or conclusion. Only later, when the project proved harder than I had envisioned, I realized that the absence of method was one of the difficulties that I was encountering.

A key question had gone unnoticed, however central it was: what does "interesting" mean? When I brought it up, the question was left unanswered, and even discarded. It needed no answer, I was told, because when we will find something interesting, we will recognize it, and everything will become clear. That was an unexpected turn of events, far remote from the Popperian ideal that I naively believed was defining science. In this enterprise I had believed constructed solely on reason, intuition and instinct were taking a pivotal and explicit role. Unconvinced, I continued my research.

But the unanswered questions did not die, it bred two new ones, one epistemic, the other practical. On the front of epistemology: how can science accept this big hole in the definition of its work? All scientists are looking for interesting phenomena and results, yet interesting-ness appears to be unquantified, and its study carefully, but probably unconsciously, avoided. I do not have an answer, only a bit of optimism: I believe that proposing a solution to this question is possible, but will require a significant amount of dedicated and deliberate effort.

The second question, in a sense, is a practical example of the importance of the first one: how long should we keep digging before giving up? When do we accept that there is simply nothing interesting to be unearthed from the available data? Had we defined what we are searching for, we could try to quantify it, estimate maybe the probability to eventually stumble upon something worth our time. We could get an idea on the value of continuing the search, on whether it is still meaningful. Despite its practical and organizational importance, especially as data analysis is becoming more prominent in physics research, I do not have an answer to this question either. And, in fact, I did not need such answer. For a simple reason: as it was foretold, I found something interesting enough in the data, and that unlocked the rest of my thesis work. Then, I was happy to accept that the method had been right, and to forget about the path that led me there.

This is a common pattern, I realized, when presenting one's work to others, especially in the form of manuscript. First you struggle, you doubt, you run blindly into walls, you fall on your knees, bash your head on a corner of the living room table, and once you hit the ground you notice something shiny under the sofa, a discovery worth sharing with your peers. Then you pretend that everything was planned from the beginning and that your actions followed an unwavering irrepressible logic. You did not stumble, you pretend, you decided with great foresight to go look under this specific piece of furniture. Instead of presenting the method that led you to your result, you tell the story of how an imaginary scientist found what they were destined to discover.

The struggles are part of a scientist's life, but absent from the mythology of science that we write for ourselves.

This thesis is no exception.

However, I would not like to start on a negative note. Once I found something interesting, my work fell into to the scientific mold that I was initially expecting. The analysis led to unexpected results, so we put forward hypotheses, and then tested them. This thesis is also the story of this process, of science on a smaller scale, of the countless small steps necessary to make the whole project sound. A story of the humbling experience of systematically pursuing objectivity.

Hopefully, with success.

CONTENTS

ACRONYMS

CEI Coulomb explosion imaging

COLTRIMS cold target recoil ion momentum spectroscopy

EuXFEL European XFEL

KL divergence Kullback-Leibler divergence

L-BFGS limited-memory Broyden–Fletcher–Goldfarb–Shanno

MCP microchannel plate

PC principal component

PCA principal component analysis

PEMSP per-element moments of the scalar products

ReaxFF reactive force field

REMI reaction microscope

SQS small quantum systems

VGSF vibrational ground-state fluctuation

XFEL x-ray free-electron laser

1 INTRODUCTION

We want a femtosecond movie to understand the core of a megaton industry.

Imagine that you stand on a tall hill, overseeing somehow the whole of earth industry. You see countless plants, gargantuan, devouring train after train of raw materials, drinking avidly water, oil and other liquids from gigantic pipes, and digesting it all into new chemicals. Then they pour their production back into the industrial network that connects it with hundreds of other plants, to get refined, mixed and combined with others, creating compounds of increasing complexity. Their journey continues until they become everything that makes the modern world: plastic, fertilizer, medication or other directly usable chemicals that get distributed to where they are needed. You need time to discern the logic, to trace in your head the synthesis routes going through this labyrinth, to understand the path of sulfur, petrol, water and air, to name a few.

As you observe the processes from your vantage point, you are mesmerized by its scale and complexity. Suddenly, your contemplation is interrupted by your own cough, so violent that you nearly fall on your knees. Once you manage to pull yourself together, you look up again and notice the black smoke rising from mountains of coals being burnt and the bleak color of the air, stained by a variety of poisonous fumes ejected as byproducts. You follow the propagation of the toxic clouds backward to their sources and, after thinking for a moment, realize that by rerouting some of the pathways you identified, the necessary products may be produced without requiring to go through the most polluting plants. Maybe this would require finding new synthesis routes altogether, but maybe it would suffice to reduce the cost of some known reactions. Then the invisible hand of the economy would seize the opportunity to redraw the network, excluding the obsolete plants, building modern ones, connecting them with new routes, and changing the life of millions in the process.

It would not be the fist time that discovering a new synthesis method has a significant impact on the industry. The production of acetic acid is one of many examples. It is a chemical produced at an estimated rate of 14 megatons per year [13]. From the onset of the production in 1957 [86], the process of fabrication has been improved twice, in 1973 with the Monsanto process [43] and in 1995 with the Cativa process [4], which is more economical and environment friendly. Thus, today, most of the acetic acid is produced by the Cativa process. Yet, despite these already great improvements, the production of this

chemical is still an active subject of research [61] due to its importance and consequent environmental footprint.

What could guide you to reproduce such a feat? From the top of the hill, you search for inspiration. You identify methanol and carbon monoxide going into a huge plant and acetic acid coming out. But you still need more details to really understand what is going on. You take out your binoculars and look with more scrutiny. First, you single out one plant, and start zooming to see the inside (your binoculars are a bit magical), you focus on one of the big vats where the reactions are happening. You continue to zoom in, seeing inside the tank, and soon beyond the uniform bulk of the liquid, suddenly discerning the chaos of uncountably many individual molecules. They are interacting, reacting incredibly fast with each other once they move close enough. Things are happening so fast that your eyes are orders of magnitudes too slow to make sense of them, as it takes only few hundred femtoseconds for atoms to rearrange themselves into their new configurations. So in addition to your magical binoculars, you dream of a slow motion camera to get a chance to analyze and really understand what is going on.

From your dream you realize that the potential of such movie is huge. By following the details of a reaction, you could find subtleties in the process and refine greatly your chemical intuition. Then, you would exploit it to guide the overall research effort, leading it in the direction of efficiency. Due to the shear volume of products involved, even small optimization of the bulk processes could change the industrial landscape significantly and affect all the producers and consumers of the chemical industry. That is, it could ultimately affect everyone.

Now it is time to leave this metaphorical hill, and ask pragmatically: How could such movie of a chemical reaction be filmed? To answer, let us look at the logic that led some of our predecessors to success. In this case, we can take inspiration from the Lumière brothers, scientists and pioneers of the film industry, who developed one ancestor of modern cameras.

However, they did not start by developing a method to make a movie. They were aware that capturing an image takes time, and that the picture produced is averaged over the duration of the capture, known as the *exposure time*[1]. Thus, if an element of the image moves far enough over the duration of the exposure, the picture is blurry. To get sharp images they invented a method to make *instantanés*, literally instantaneous pictures, in 1881. The name is of course an overstatement, but their photographic plates nonetheless worked with an extremely short exposure time, shorter than the timescale of

1 The name *exposure time* refers to the physical exposure of the photographic plate to the light, which causes the photochemical reaction that then produces the image on the plate.

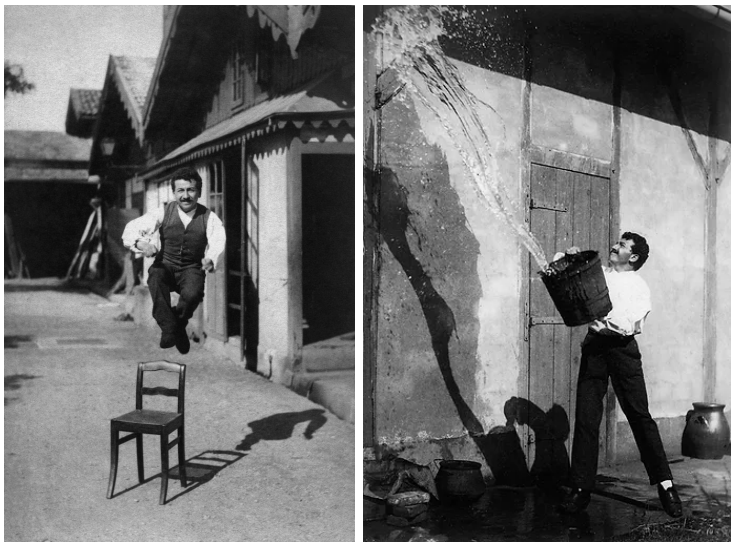

Figure 1: **Auguste Lumière photographed by his brother Louis, circa 1888 (Public Domain).** These two pictures showcase the capability of the photographic plates invented and sold by the Lumière brothers. In particular, they produce pictures using a very short exposure time, "freezing" the imaged motion.

the dynamic that we can observe in everyday life. To showcase their work, they took pictures like the ones shown in Fig. 1. For example, one of the brothers jumping from a chair or throwing a bucket of water into the air. The pictures are crisp, proving the quality of their product.

Then, fourteen years later, in 1895, they combined their technique for *instantanés* with other technological advances and created the ancestor of the modern camera, the *caméra cinématographique*.

In this thesis, I study one technique to apply the same strategy to chemical reactions.

My work focuses on the first part, producing crisp still images, with an exposure time of only few femtoseconds. This is achieved by submitting the system to a light pulse (*the probe*) that is itself few femtoseconds long.

Before going into the details of how to generate these snapshots, I briefly describe how such individual images can be assembled into a movie. The accepted strategy is to trigger the dynamic of interest with another short light pulse (*the pump*). By controlling the time delay between the pump and the probe, different stages of a reaction can be recorded and then recombined into a coherent movie. This process is shown schematically in Fig. 2A and Fig. 2B, while the blurry result of using a long exposure time is shown in Fig. 2C.

Importantly, note that chemical reactions are repeatable, as two molecules of the same species are indistinguishable. Therefore, the concept of pump-probe imaging can be used on independent but identical systems for each snapshot, allowing to use a destructive

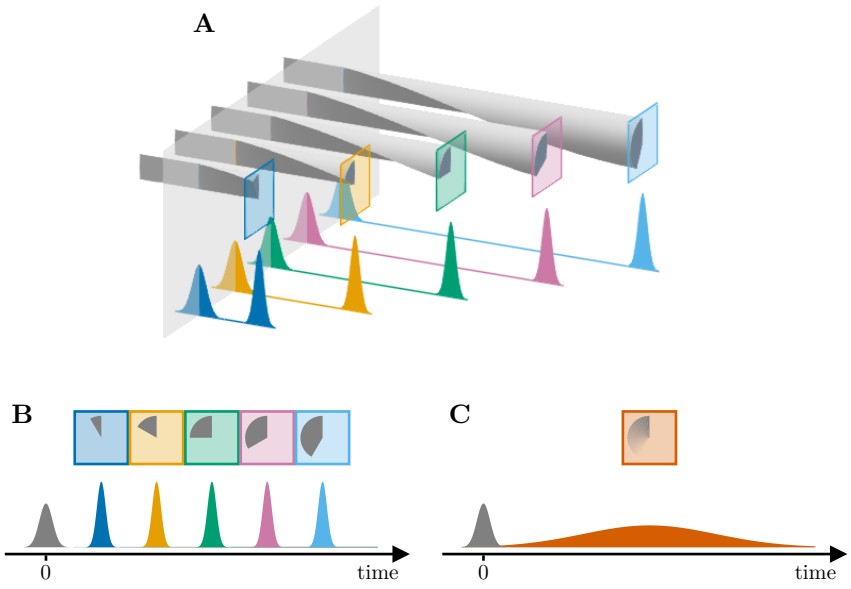

Figure 2: **Pump-probe scheme.** A. The same process is repeated independently multiple times. The dynamic is triggered (gray plane) by the pump pulse, a snapshot of the state of the system (colored squares) is recorded later by the probe pulse. B. Combining the snapshot together based on the delay between the pump and prob pulses, a movie of the dynamics is built ; here a disk filling counterclockwise. C. Using a long probe pulse results in a blurry image, where the state of the system at a specific time can not be resolved.

imaging technique. The analogy with the macroscopic cinema technology stops here: if for each picture Louis Lumière had to severely damage his brother Auguste, their movies would have been much more sinister.

Now, we know how the molecular snapshots are combined into a movie, but how exactly would we generate the snapshot itself? First, we want to film individual systems, for this thesis even focusing on individual molecules. Imaging from an ensemble can be done, but the information is averaged over all the particles in the sample. Thus, in general, it is difficult to untangle the dynamics of the constituting molecules. To mitigate this problem, bulk methods often rely on ordering the constituent of the sample, allowing the signal to be treated as arising from a coherent material. For example, the molecules can be aligned with a laser field [28] or arranged in a crystal. In the latter case, a diffraction pattern is produced when a wave (typically electrons or light) passes through the crystal. Studying those patterns led to a well established method known as crystallography. Fabricating the crystal is however known to be challenging[31, 57].

Currently, there are several candidates for performing femtosecond snapshots of single molecules: Single-particle x-ray scattering [49], laser-induced electron diffraction [5, 90] and Coulomb explosion imaging (CEI), the main topic of this thesis. The idea of CEI is to fragment the molecule and measure the ejected particles.

The goal of this thesis is to explore the current state, limitations and prospects of x-ray induced CEI, with a special focus on data analysis. In particular, I show that CEI data is intrinsically high dimensional, which is both an opportunity and challenge. On the one hand, the high dimensionality contains a lot of information about the dynamics of the explosion and the state of the molecule before the explosion, information that is not always accessible to other methods. On the other hand, it implies that the data has a high complexity and must be analyzed using advanced methods taken or inspired from the field of machine learning.

2 COULOMB EXPLOSION IMAGING

Sections 2.3 and 2.4 of this chapter are based on:

Rebecca Boll, Julia Schäfer, **Benoît Richard** *et al.*
"X-ray multiphoton-induced Coulomb explosion images complex single molecules."
Nat. Phys. **18**, 423–428 (2022).

Contribution I performed a detailed analysis of the dynamics of the Coulomb explosion in the simulation, in particular describing the atomic collision between C2 and I (Fig. 8) and the trapping of the electrons on hydrogen atoms (Fig. 7).

2.1 BASIC CONCEPT

The idea of Coulomb explosion imaging (CEI) is straightforward [84]: take a molecule and remove many of its valence electrons. This weakens the chemical bonds that pull the atoms together and, simultaneously, increases the positive charge on the atoms. When enough electrons are removed, the repulsive Coulomb interaction dominates the interactions between the atoms, and the system explodes into atomic fragments. Once they are far away from each other, they reach asymptotic states that contain information about the molecule just before the explosion was triggered. Therefore, by measuring the dynamical properties of the fragments, one images the original molecule.

In this chapter, I first discuss how the electrons can be removed from a molecule on the ultrafast femtosecond timescale required to generate movies of chemical reactions. Then, I describe the measurement of the fragments, and, especially, the adaptations necessary to image large molecules. Finally, I detail the simulations performed for x-ray induced Coulomb explosion of molecules, and what they can tell us about the process. The description of the simulation is rather simple and more advanced analyses will be presented in the next chapters.

Nowadays, two methods are typically used to induce an ultrafast Coulomb explosion. One way is *strong field induced Coulomb explosion*, in which an intense laser pulse ionizes valence electrons through tunnel ionization. The exposure time corresponds to the duration of the laser pulse, which can be as short as 6 fs [32]. However, strong field ionization does not naturally scale to large molecules. When an electron is removed, the remaining electrons feel a stronger Coulomb potential from the nuclei and removing more electrons gets increasingly less likely, as the ionization potential increases as a function of

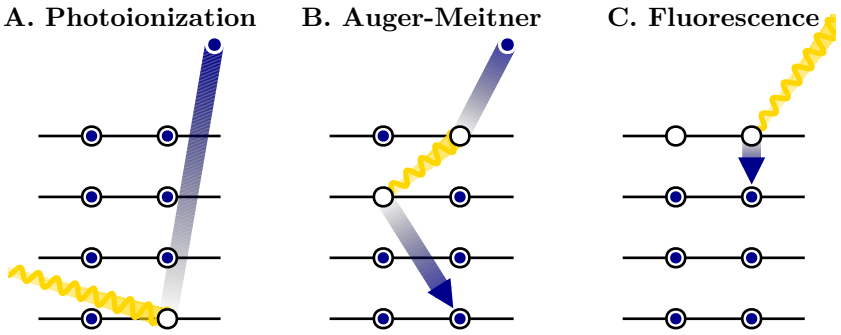

Figure 3: **Cartoon representation of the main atomic processes.** Photoionization: An external photon is absorbed by the atom and a bound electron is ejected. Auger-Meitner decay: An electron decays to a lower energy hole, and the excess energy is transferred to another electron that is ejected. A photon is used to depict the transfer of energy from the decaying electron to the ejected one. Fluorescence: An electron decays to a lower energy hole by releasing the excess energy as a photon.

the charge state. Nonetheless, the method has been applied successfully for the determination of the geometry of small molecules [5, 58, 76], and to follow various time dependent processes [20, 30, 32, 34, 51, 59].

In the other method, *x-ray induced Coulomb explosion*, the explosion is induced by an intense ultrashort x-ray pulse. An x-ray photon has enough energy to remove a core electron through direct photoionization, creating a core hole, as depicted schematically in Fig. 3A. This brings the ionized atom to a highly excited state that relaxes mainly through two fundamental processes: Auger-Meitner and fluorescence decays. In an Auger-Meitner decay, an electron decays into a hole present in a lower energy shell. Its excess energy is transferred to another electron, which gets ejected, as shown in Fig. 3B. Therefore, Auger-Meitner decays contribute to the charging up of the molecule.

Then, during fluorescence an electron decays to a hole while releasing a photon, as shown in Fig. 3C. Fluorescence does not contribute to the charging up dynamics of the molecule. For the photon energy and elements that we consider, fluorescent decays are much rarer than Auger-Meitner decays. For example[1], for an iodine atom with a 3d hole, the characteristic time for fluorescence is 2719.1 fs, while it is only 1.1 fs for an Auger-Meitner decay. Therefore, in the following, we focus on direct photoionizations and Auger-Meitner decays.

For a light atom, the full charge-up dynamic induced by an x-ray photon has two stages: first a core electron is ejected by a direct pho-

1 All quantities in this section are computed using XATOM [74]

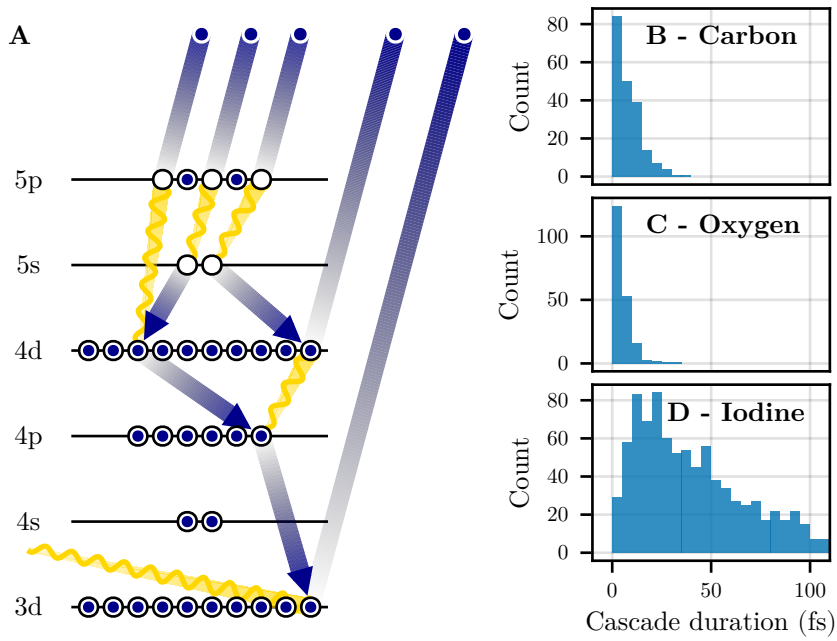

Figure 4: **Auger-Meitner cascade.** A. Example cascade for iodine. A single photoionization of the 3d shell leads to four successive Auger-Meitner decays, resulting in five electrons ejected in total. B - D. Distribution of cascade duration for various atoms and a single initial hole, calculated based on Monte-Carlo sampling of the decay pathways using XATOM [74]. B. Carbon with a 1s hole. C. Oxygen with a 1s hole. D. Iodine with a 3d hole.

toionization, then another one is removed through an Auger-Meitner decay. For heavy atoms, a single core hole can relax through a sequence of multiple Auger-Meitner decays, a process known as an *Auger-Meitner cascade*. We show an example of such cascade for an iodine atom in Fig. 4A. Hence, a single photoabsorption can lead to the ejection of several electrons, and the absorption of few photons is sufficient to efficiently charge up a molecule.

To quantify the efficiency of the charging up, we look at the duration of the cascade. It intrinsically depends on the elements of the atoms being ionized. For example, if we consider isolated atoms with a single 1s hole, the median decay time (i. e., the half-life of the excited state) is 6.5 fs for carbon and 3.6 fs for oxygen. For those light atoms, the pathway is simple as it stops after a single Auger-Meitner decay. Hence, its duration directly follows an exponential distribution.

As an example of a heavy atom, consider iodine after the ionization of an electron from the 3d shell (the most likely process for 2 keV x-ray photons). The median time for the whole cascade is 41.4 fs, with a median of 5 Auger-Meitner decays, a significantly longer cascade than for carbon and hydrogen (Fig. 4D). In both cases, the charging up of the atom happens on a femtosecond timescale.

In this thesis, we focus on Coulomb explosion induced by using x-ray pulses generated at an x-ray free-electron laser (XFEL). The technology is now mature enough to provide ultrashort pulses that have high enough intensity to reliably trigger Coulomb explosion of single molecules via multiple photon absorption. Several facilities offer pulses with a duration of few tens of femtoseconds, with the prospect to get even shorter ones, in the attosecond range, thanks to the continued improvement of the XFEL technology.

2.2 EXPERIMENTAL SETUP

We examine an experiment that took place at the small quantum systems (SQS) scientific instrument of the European XFEL (EuXFEL). This endstation offers 10 fs pulses at 1 keV photon energy with 1 mJ energy per pulse, focused on a 1 μm^2 area yielding an average fluence of 3.12×10^{12} photons$/\mu m^2$. At this photon energy, the photoionization cross-section of a core 1s electron of an isolated carbon atom is $4.2 \times 10^{-12}\,\mu m^2$. With these conditions, we expect on average all carbon atoms in a molecule to be photoionized once and to acquire a second charge due to a subsequent Auger-Meitner decay. The induced charge state is sufficient to fully dissociate most carbon-based molecules. In addition, note that the photon energy (1 keV) is much higher than the 1s orbital energy of the 1s shell (291 eV), so the x-ray photon will still have enough energy to ionize core electrons, even when their energy is shifted by the presence of other holes. Therefore, a Coulomb explosion can be reliably induced by the x-ray pulse, even in large molecules.

The molecules are brought in the focus of the x-ray beam in a supersonic gas jet. The density of the molecular gas is tuned to ensure that, at most a single molecule is present in the focus at a time. Thus, the data collected by each shot[2] refers to a single molecule.

Once the explosion has been triggered, the asymptotic 3D momenta of the fragments are measured in coincidence, using the cold target recoil ion momentum spectroscopy (COLTRIMS) reaction microscope (REMI) [19, 36, 83], schematically depicted in Fig. 5. A bias field directs the fragments toward a microchannel plate (MCP) [89]. When an ion hits one of the channels, it acts as an electron multiplier, producing an electron cloud. This cloud is picked up by pairs of parallel wires, on which voltages of opposite directions are applied. Thus, the signal from the electron cloud travels in both wires simultaneously, in opposite directions. Measuring the time difference between the arrival of each electrical pulse, the position along the pair of wires is determined. To increase accuracy, the electrical signal is delayed to form a *delay line detector* [35]. In this arrangement, each wire section

2 A shot is the result of a single x-ray pulse hitting the molecular gas jet, and the corresponding measured data.

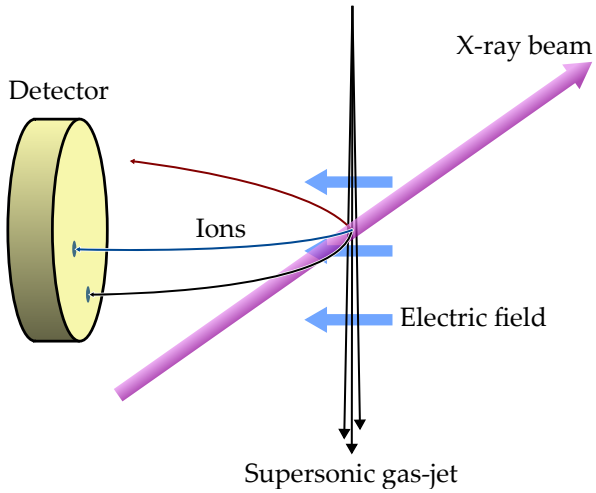

Figure 5: **Schematic of a COLTRIMS.** The x-ray beam crosses a molecular gas-jet, triggers the Coulomb explosion of the molecules. An electric field then directs the ions toward a detector plate, where the fragments are detected in coincidence.

corresponds to a slice of the MCP area. While the time resolution is not high enough to determine the position inside a single wire section, the slice of the MCP which is hit by the ion is determined accurately. By putting a second identical delay line detector perpendicular to the first one, a 2D image is obtained.

The detection on the plate is synchronized with the x-ray pulse, allowing to measure the time that the ion takes to fly from the interaction region to the detector (known as the *time of flight of the ion*). By measuring the time of flight, the mass-over-charge ratio of the detected ions can be determined. Analyzing the distribution of mass-over-charge ratio, the charge and the element of the ion are inferred. Finally, combining the time of flight and the position of the plate, the asymptotic 3D momentum of each of the detected ions is reconstructed.

From the explosion of a single molecule, multiple fragments can be detected. In this case, we say that the fragments are detected *in coincidence*. Coincident measurement is a specific advantage of CEI, giving access to high dimensional data about the imaged molecule. This high dimensionality comes with both challenges and opportunities. Addressing them for CEI is the main question discussed in this thesis.

2.3 LARGE MOLECULES

In this section, we address a major experimental challenge of CEI: the detector has a finite detection efficiency and on average only few fragments are measured in coincidence from a single shot. The probability of detecting all atoms in coincidence from the explosion of a

molecule composed of n atoms is p^n, where p is the detection probability of each ion. For example, with $p = 60\%$ (reported e.g. in [9]), the probability for a complete detection of all atoms is, for molecules composed of 5, 10 and 15 atoms respectively, $p^5 = 7.7\%$, $p^{10} = 0.6\%$ and $p^{15} = 0.047\%$. Hence, gathering sufficient statistics for large molecules is impractical, as the time to collect full-coincidence measurements grows exponentially with the molecule size. Accordingly, we use the term "large" to describe molecules for which all fragments can not be detected in coincidence due to a prohibitively long data acquisition time. Currently, this corresponds to molecules composed of 5 or more atoms, a boundary that may be incrementally pushed by developing more efficient detectors or experiments with a higher repetition rate. This is, however, unlikely to be a fast development. Another idea is needed to overcome the barrier and reliably address larger systems.

Therefore, we propose a different strategy: Analyze the data based on detecting a limited number of atoms in coincidence. In this work, we demonstrate this idea for the 2-iodopyridine molecule (C_5H_4IN), depicted in Fig. 6A. Later, we refer to this molecule as simply iodopyridine. The Coulomb explosion of this molecule was realized at the SQS instrument of EuXFEL [8].

In iodopyridine, two of the composing atoms, iodine and nitrogen, can be uniquely identified by their element. Thus, we use them to define a *molecular frame* for the measured momenta. In the case of iodopyridine, we define the x-axis as the direction of the momentum of the iodine ion and the xy-plane as the plane in which the iodine and nitrogen momenta lie (see Fig. 6B), defining a reference frame known as the *recoil frame*. Additional ions measured in coincidence with iodine and nitrogen are placed relative to these axes, as they define the orientation of the system. Combining the data for every shot where the iodine atom, the nitrogen atom and at least one other ion were detected in coincidence[3], we obtain clear plots of the momenta, shown in Fig. 6C. Looking at the in-plane projection of the data, the resemblance with the geometry of the molecule is striking, particularly for the hydrogen ions, which are clearly distinguishable. The contribution from each carbon ion can also be determined, but requires more care, either applying further filtering or fitting the 2D data with a mixture of Gaussian distributions. The clarity of the in-plane Newton diagram is particularly remarkable considering that the momenta spread significantly more widely out-of-plane than in-plane, forming, in 3D, a "tube" for each ion. The molecular recoil frame allows us to look at them from the right angle, so that their separation is apparent.

3 Additional filtering is applied to the raw detector data to avoid false coincidences, see [9] for more details and [8] for the data itself.

We understand this clear separation of the ions, as a consequence of the fast charging up of the molecule, resulting in a clean direct explosion, without the formation of intermediate molecular fragments.

Applied to a more complex or non-planar molecule, we would expect a similar distinguishability of the ions. However, without a clear molecular recoil frame in general, exploiting it is likely to require a more general analysis technique. In Chapter 5, we present such a method, building on the same idea to retrieve the full distribution of 3D momenta from partial coincidence data.

2.4 SIMULATIONS

We now describe our strategy for the simulation, as implemented in XMDYN [39]. The molecule is approximated as a group of atoms with independent electronic structures. A Monte-Carlo scheme follows the temporal changes of the electronic orbitals using a discrete timestep propagation. The rate of the main atomic processes induced by x-ray (direct photoionization, Auger-Meitner decays and fluorescence) are computed from first principle for each atom using XATOM [74]. The probability that a given event happens during a timestep is computed from those rates. When electrons are emitted, the emitting ion recoils to conserve the total momentum. The emitted electrons and the ions are both modelled as point charges in real space interacting through the classical Coulomb field.

The bound electrons are allowed to jump from atom to atom based on a semi-classical over-the-barrier model [72]. In this model, an electron moves to a neighboring atom if the classical Coulomb barrier between the two atoms is lower than the electron energy. When the atoms get further away, the barrier increases, until it becomes high enough that the electrons are no longer allowed to move around, as represented schematically in Fig. 7. This simple model is in good agreement with experimental results [21].

This procedure yields trajectories that are stochastic due to the randomness of the atomic processes. We repeat the simulation many times and consider all trajectories together as an ensemble.

We simulate the Coulomb explosion of iodopyridine, using the parameters reported in Table 1, starting from the fixed ground-state equilibrium geometry with zero velocity and with the x-ray pulse always polarized perpendicular to the molecular plane. Over the course of this thesis, we generated several other ensembles of trajectories, summarized in Table 2. The current chapter exclusively uses Simulation Set I. The properties of the other sets will be detailed in the chapters that introduce them.

The result of the simulations is shown in Fig. 6D, where we see a reasonable agreement with the experiment. The simulated momenta have, on average, the correct direction, with however a norm that is

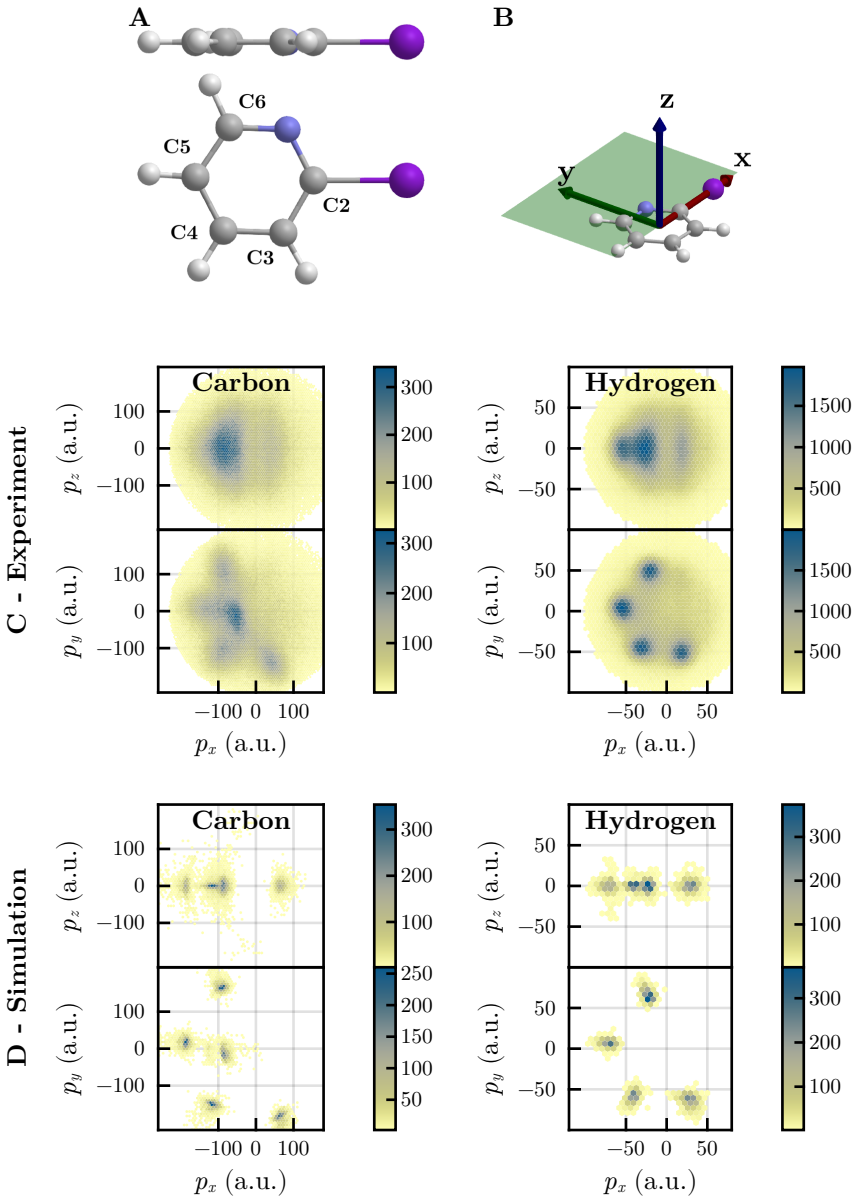

Figure 6: **Coulomb explosion of iodopyridine.** A. Side and top view of the iodopyridine molecule in its ground-state equilibrium geometry using a ball-and-stick model. B. Definition of the molecular recoil frame. The x-axis is defined by the recoil direction of the iodine atom and the positive xy half-plane, in green, is spanned by the recoil direction of the nitrogen atom. C. Newton diagrams of the measured ions in the molecular recoil frame, from the experiment [9]. Only carbon and hydrogen atoms with charge +1 are shown. D. Newton diagrams of the ions in the molecular recoil frame, according to our simulation. Only carbon and hydrogen atoms with charge +1 at the end of the simulation are shown.

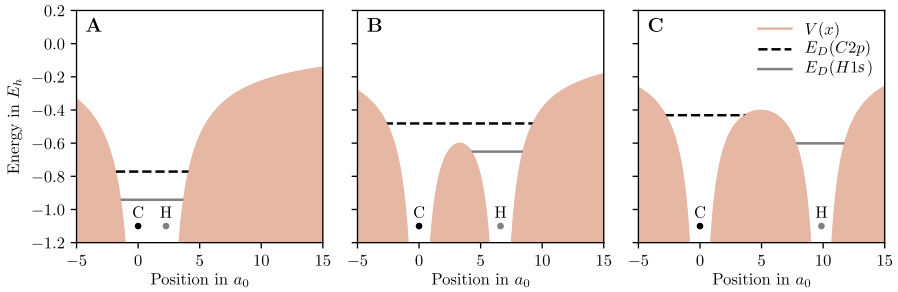

Figure 7: **Over-the-barrier model between a carbon and hydrogen atom.** A. The Coulomb barrier between the two atoms is lower than their valence energies. An electron can hop in both directions, when a hole is present. B. The Coulomb barrier is higher than the valence energy of the hydrogen atom, but lower than the valence energy of the carbon atom. An electron can hop from the carbon atom to the hydrogen, but not in the opposite direction. C. The Coulomb barrier is higher than both valence energies. No charge transfer is possible. *Reproduced from [9], Creative Commons CC BY.*

Simulation software	XMDYN
Molecule	2-iodopyridine
Time step	1 as
Photon energy	2 keV
Pulse shape	Gaussian
Pulse duration	10 fs (FWHM)

Table 1: **Core parameters of the simulations.** Parameters common to all simulations performed for this thesis.

	I	II	III	IV	V
Polarization	Perp.	Rand.	Rand.	Rand.	Rand.
VGSF	No	No	No	Yes	Yes
Force field	No	No	Yes	Yes	Yes
Fluence (10^{10}photons/μm^2)	7.5	7.5	15	15	7.5
Trajectories	15000	15000	10000	61652	16000
Performed by	JS	BR	BR, SB	BR, SB	BR, SB

Table 2: **Different simulation sets used in this thesis.** Polarization: polarization of the x-ray pulse, perpendicular to the molecular plane (Perp.), or randomly oriented (Rand.). Vibrational ground-state fluctuation (VGSF): not considered with the simulation starting from the equilibrium geometry of the molecule with all atoms at rest (No), or included through the sampling of the ground-state Wigner distribution (Yes). Force field: whether a force field is used to model the chemical bonds. Pauli repulsion: whether the Pauli repulsion between atoms is considered. Trajectories: the number of independent trajectories generated for the given parameters. Performed by: Julia Schäfer (JS), Sourav Banerjee (SB), or myself (BR).

slightly too high. At the time of publication of these results [9], this was deemed sufficient.

However, as we developed a better understanding of the data, we understood that variance of the momenta also contains crucial information about the explosion process and the state of the molecule before explosion. We discuss these aspects in Chapters 3 and 4. For now, we discuss further what we can learn from the current set of simulations, keeping in mind that they do not fully model all physical processes involved.

The main advantage of the simulation is that we can analyze the time evolution of the explosion in full detail, which is impossible from the experiment as only the asymptotic final momenta are measured. In Fig. 8, we show the distribution of momenta (Fig. 8a) and positions (Fig. 8b) at given times, using the center of the pulse as time zero. Interestingly, we can see that the C2 atom first moves towards the iodine atom (between $t = +10\,\text{fs}$ and $t = +30\,\text{fs}$), before being pushed away (from $t = 30\,\text{fs}$ onwards). This explains why the C2 atom has an asymptotic momentum close to zero (Fig. 6): it transfers part of its energy to the iodine atom when it collides with it and then is trapped inside the aromatic ring.

To get a quantitative view of the explosion, in Fig. 8c, we consider the average distances between pairs of atoms and plot their evolution in time during the explosion. Three different timescales are present. First, the hydrogen atoms are ejected, as indicated by the rapid increase of the carbon-hydrogen distances (4 Å at $t = +10\,\text{fs}$). Hydrogen atoms move the earliest due to their light mass compared to the other atoms. Second, the carbon-nitrogen ring explodes, as indicated by the increase in the carbon-carbon and carbon-nitrogen distances (4 Å at $t = +30\,\text{fs}$). Third, the C2 atom separates from iodine (4 Å at $t = +50\,\text{fs}$). Before the dissociation, the two atoms collide, so that approximately 30 fs after the center of the pulse, the distance between the two atoms reaches its minimum (around 1 Å, half of the equilibrium bond length). We see that the Coulomb explosion is not a trivial process, it happens in multiple stages and involves strong transient interactions between the atoms, like the C2-iodine collision.

The simulations also allow to analyze the charge-up dynamics. Figure 8d shows the evolution of the average charge for each atom. More than 90% of the photoabsorption events happen on the iodine site, in accordance with its cross-section being much larger than the one of the other atoms for 2 keV photons. In consequence, the timescale of the charging-up of the molecule matches the timescale of the Auger-Meitner cascade on an independent iodine atom (Fig. 4), typically between 30 fs and 40 fs.

Once a hole is created, the charges quickly rearrange, resulting in an on-average uniform charge distribution in the molecule. In the simulation, the speed of the redistribution of charges is likely

overestimated since an electron hops at every attosecond timestep, as long as it has enough energy to jump over the classical electrostatic barrier. Therefore, the charges can hop over a 1 Å gap between two neighboring atoms in 1 as, which corresponds to a charge migration speed of 1000 Å/fs, and charge migrates from one end of the iodopyridine molecule to the other in about 0.01 fs. This speed is significantly higher than predicted by density functional theory calculations, which estimate the charge migration speed to be around 4 Å/fs in linear molecules [25], roughly requiring 2 fs for a charge to redistribute over the whole molecule. In both cases, however, the charge migration is significantly faster than the atomic motion and the charging up of the molecule, which both happen over tens of femtoseconds, as shown in Fig. 8c and Fig. 8d. Thus, we expect the difference in charge migration speed to only cause a small discrepancy.

We see in Fig. 8d, that the charges on each atom are equal, as expected, until 10 fs, when the hydrogen atoms dissociate. At this point, the average charge on the hydrogen atom drops. This effect is caused by the asymmetry of the over-the-barrier model (see Fig. 7): there is a range of distance for which the valence electrons on a carbon atom have enough energy to jump over the barrier towards a hydrogen ion, landing on a valence state of lower energy, and then are unable to jump back. Thus, electrons tend to become trapped on the hydrogen atoms.

Finally, we see on Fig. 8d that the charge on iodine does not seem to converge to exactly $+1$, which is unexpected as we only consider trajectories that yield an I+ ion. However, at the end of the simulation (after 1 ps), some emitted electron may not have enough energy to be free. If an electron kinetic energy is lower than the electrostatic potential it feels, we recombine it with the latter, leading to a slight decrease of the charges of the ions.

In conclusion, numeric simulations are a valuable tool to understand the dynamics of a Coulomb explosion. In particular, they show that the explosion happens on several timescales, overlapping with the charging-up of the molecule by the x-ray pulse. Yet, there are clear structural similarities between the initial geometry of the molecule before the explosion and the distribution of the asymptotic momenta. Therefore, despite its complexity, the explosion maps features of the initial state of the molecule to the measured momenta distribution of the fragments. Understanding this mapping is essential to use Coulomb explosion for imaging. This is an ongoing research question that we further discuss in Chapter 4.

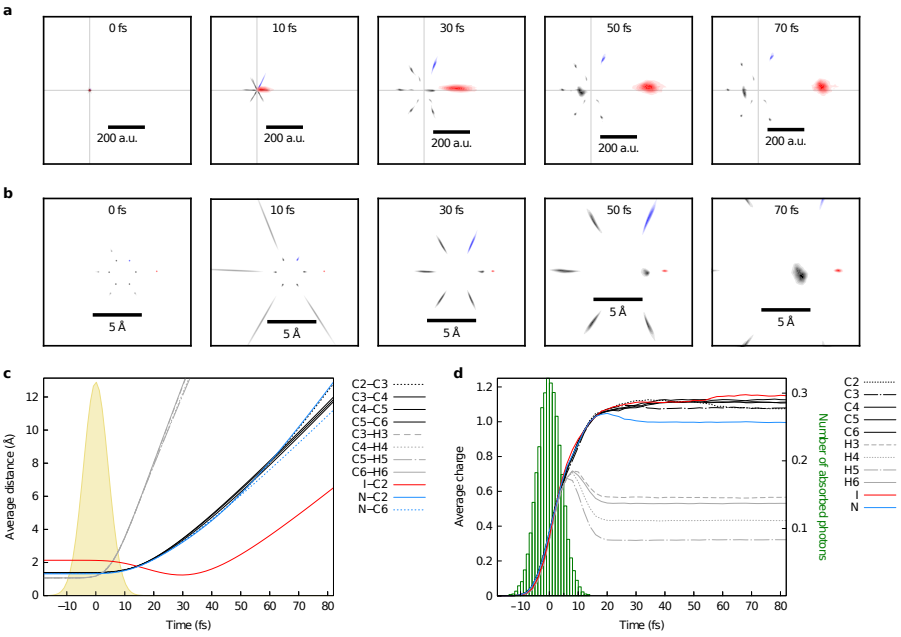

Figure 8: **Dynamics of the Coulomb explosion of iodopyridine.** Only trajectories that yield I+ and N+ are considered. a. Snapshots of the distribution of the momenta at different times. b. Snapshots of the distribution of positions at different times. c. Average internal distances between atoms as a function of time. The filled yellow curve represents the intensity of the pulse. d. Average interatomic distances as a function of time. The green bars show the average number of photons absorbed. *Reproduced from [9], Creative Commons CC BY.*

3

ATOM-ATOM CORRELATIONS

This chapter reproduces with few adaptations:

Benoît Richard *et al.*
"Statistical analysis of correlations in the x-ray induced Coulomb explosion of iodopyridine."
J. Phys. B: At. Mol. Opt. Phys. **54** 194001 (2021)

Contribution to the manuscript I established the correlations in the simulation data, proposed to use PCA on the simulation to study them further, and performed the related analyses. I drafted the manuscript with input from other coauthors.

The interaction between ions during a Coulomb explosion can lead to non-trivial collision effects [9, 91] and rotation dynamics that complicate the interpretation of the final momentum data [56]. Due to the possibly large number of ions emitted from a single molecule, the generated data are high-dimensional, and it is often hard to understand the information they contain.

In order to facilitate the interpretation of Coulomb explosion data, we conduct a statistical analysis of simulation data for the x-ray induced Coulomb explosion of the iodopyridine molecule. We show how relevant patterns of the dynamics occurring during the explosion can be uncovered through principal component analysis (PCA) of the final ion momentum data and the forces. We demonstrate that individual ion pairs exhibit particularly strongly correlated momenta due to their collision during the molecular explosion. Furthermore, we decompose the explosion dynamics into a few collective coordinates that we obtain from PCA components of the forces.

3.1 PRINCIPAL COMPONENT ANALYSIS

PCA is a standard analysis technique for high-dimensional data [77]. It decomposes its input space into multiple uncorrelated components by diagonalizing the covariance matrix. Given the input data $\mathbf{x}$, the elements of the covariance matrix Σ_{ij} are given by

$$\Sigma_{ij} = \left\langle \left(x_i - \langle x_i \rangle\right)\left(x_j - \langle x_j \rangle\right)\right\rangle, \tag{1}$$

where $\langle . \rangle$ represents the ensemble average. PCA decomposes the covariance matrix $\mathbf{\Sigma}$ as

$$\mathbf{\Sigma} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T \tag{2}$$

where the columns of matrix $\mathbf{U}$ are the principal components $\mathbf{u}_i$ and the diagonal matrix $\mathbf{\Lambda}$ contains the eigenvalues λ_i, known as the explained variances [38]. The principal components $\mathbf{u}_i$ are ordered according to λ_i in decreasing order. Projecting variations of a data set onto its first principal components thus describes the data with maximal variance for a given reduced dimensionality. The study of the first few components is therefore often sufficient to capture most of the variations. Moreover, thanks to $\mathbf{\Lambda}$ being diagonal, projecting the original data on the PCA components results in uncorrelated quantities. The algorithm can thus be viewed as linearly decorrelating the input data. PCA is a useful tool to analyze MD simulations, for both steady state [3, 33, 77] and non-equilibrium [54, 55, 68] dynamics.

In this work, we apply PCA to two different inputs: the final momenta and the dynamical forces. Given a set of N atoms, we first perform PCA on the $3N$-dimensional set of asymptotic momenta of the ions. This data can in principle be measured in coincident CEI experiments [9]. Therefore, final momentum PCA may be performed on experimental data in the near future. We note, however, that high-order coincident detection of ions (in particular of ions with very different masses) is experimentally very challenging, and that PCA methods working with missing data [78] are likely to be needed to compensate for the finite detection efficiency. Understanding how it can be interpreted and used to elucidate the dynamics of the Coulomb explosion is thus of prime interest.

Moreover, we apply PCA to the dynamical data, taking into account the full time-dependent data from the explosion simulation. This way, we aim to describe not only the variation of the ensemble in the final state of the explosion (i. e., the final momenta), but also cover temporal variations during the explosion dynamics. A common choice for analysis of MD trajectories is to use atom positions as input. However, in the current context, the topology of the positional data is problematic due to the far-from-equilibrium nature of Coulomb explosions. The atoms are evolving towards constant velocities, and thus, depending on the length of the simulation, the resulting components may be dominated by the asymptotic behavior of the system. This suppresses information about the dynamics during the ionization and explosion and defeats the purpose of using dynamical data. As a more natural alternative, we employ forces as input. Forces are bounded, and they are zero at the start and end of the simulation. They are also directly proportional to the charge of each atom, allowing us to emphasize the effect of the charge distribution on the explosion. From the simulation data, we compute the forces on the

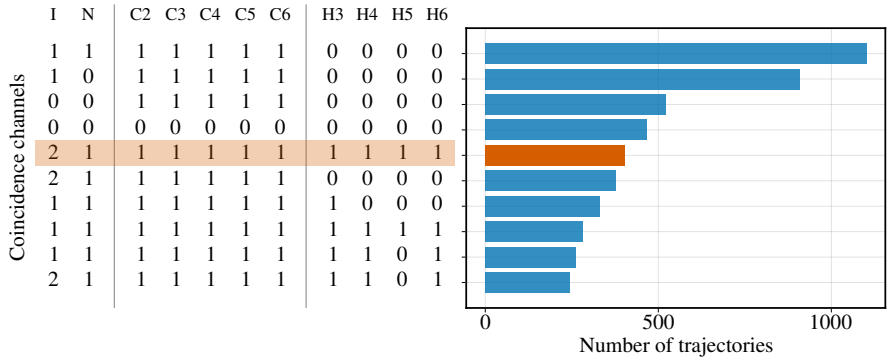

Figure 9: Number of occurrences of the most abundant final charge distributions obtained in the simulation. The numbers are the final charges of each atom. The coincidence condition used for the analysis is highlighted in orange.

individual atoms by second-order finite differences from the position data employing a femtosecond time step.

3.2 RESULTS

3.2.1 *Description of the simulation data*

The simulation used in this chapter is described in Section 2.4, and corresponds to dataset I in Table 2. The data consists of 15,000 trajectories describing the time-dependent position and charge of all atoms. In the final simulation frame, the trajectories show a variety of different charge states for a given atom. The simulations result in 1739 distinct combinations of final charges. The trajectory counts for the most abundant ones are depicted in Fig. 9. If not stated otherwise, we restrict our analysis to a single coincidence condition, i. e., we select trajectories where the atoms have the final charge state highlighted in Fig. 9. As can be seen, this set of coincident ion charges is the most abundant in which all atoms have a charge of at least +1, which is a prerequisite for their experimental detection. This coincidence channel contains 404 trajectories. For simplicity, our analysis focuses on this subset of trajectories because the data for other final charge coincidence conditions exhibit only *quantitative* differences. For example, the 8th channel, which also fulfills the experimental prerequisite, displays on average approximately 10% slower ions due to the lower total charge in the molecule but *qualitatively* very similar dynamics. Analysis of the other sets of coincident ion charges yields analogous results.

We start by describing the general properties of the data. Iodopyridine is a planar molecule and in the following the molecular plane is defined by the xy plane. The equilibrium geometry of the neutral molecule is represented in Fig. 10(b) together with the atom la-

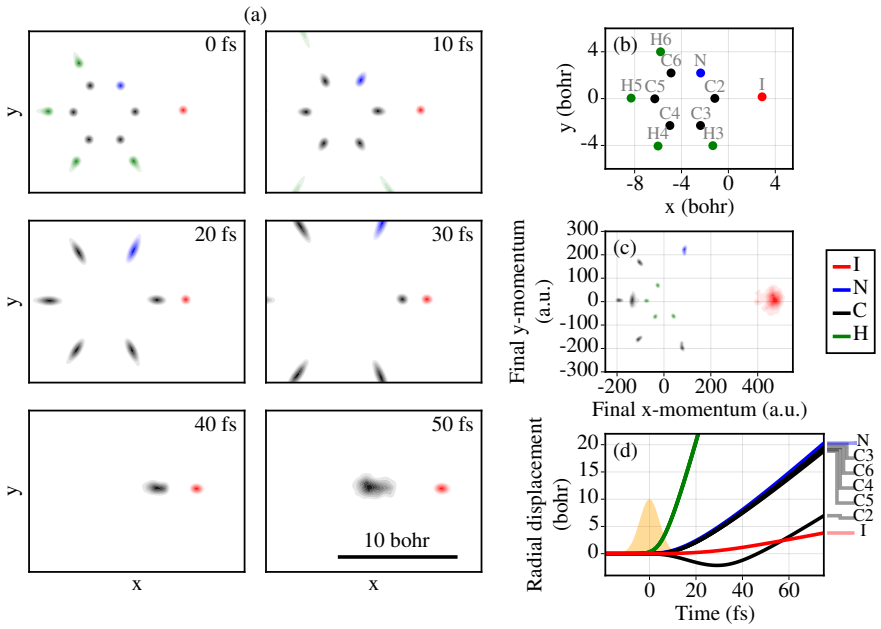

Figure 10: (a) Distribution of the ion positions at selected times. The center of the pulse is at $t = 0$ fs. (b) Iodopyridine equilibrium geometry used as the initial condition for the simulations. (c) Final momentum distribution. (d) Radial displacements. The yellow area indicates the temporal shape of the x-ray pulse.

bels. The ionization cross-section for ionization on iodine is significantly larger compared to ionization on the other atoms such that the charge is primarily created on the iodine through photoionization and subsequent Auger decays. The created charge is then distributed among the atoms by charge rearrangement. After the molecule has been exposed to the x-ray pulse it rapidly dissociates into individual atomic ions. Snapshots of the position distribution at selected times are shown in Fig. 10(a). Here and throughout this work, the center of the x-ray pulse is at $t = 0$ fs. After ~ 100 fs the momenta of all atomic ions converge to their final asymptotic values. The distribution of these final momenta is shown in Fig. 10(c). As can be seen, the final ion momentum distribution of each atom is relatively sharp and most of the momenta can be directly linked to the geometrical position of the atom in the initial geometry. Compared to the other heavier atoms, the C2 momentum stands out by pointing to the left, whereas in the initial geometry, the C2 atom is next to the iodine atom on the right-hand side of the molecule. The C2 momentum also displays a somewhat higher variance compared to other carbon atoms.

As indicated by the snapshots in Fig. 10(a), the explosion is mainly radial. In Fig. 10(d) we show the radial displacement of all atoms as a function of time, with the radial direction defined by the final average momentum of each atom. As can be seen, already during the

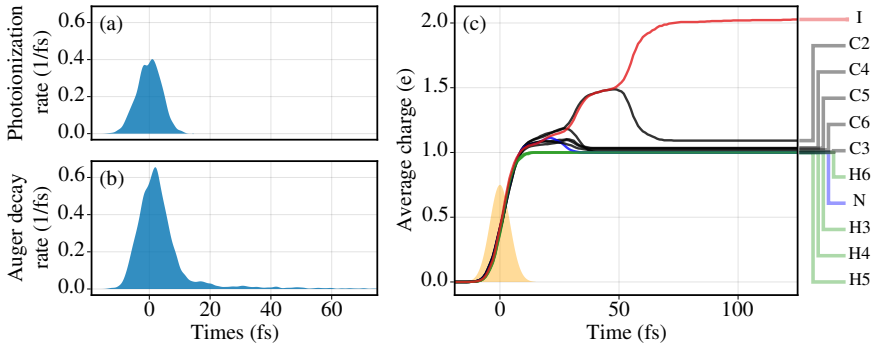

Figure 11: (a) Average rate of photoionization events as a function of time. (b) Average rate of Auger decay events as a function of time. (c) Average charge of each atom as a function of time. The yellow area indicates the temporal shape of the x-ray pulse.

pulse exposure, the hydrogens are expelled. This is followed by the radial dynamics of the atoms N, C3, C4, C5, C6, and, eventually, the I atom. An exception to this purely radial explosion is the C2 atom that undergoes a reversion of its motion during the explosion. As depicted in Fig. 10(a) and (d) this reversion occurs at around 30 fs.

We conclude this section by discussing the charge dynamics of the atoms in the molecule. In Fig. 11 (a) and (b) we show the average rate of photoionization and Auger decay in time over our ensemble of trajectories. The distribution of the electron production through photoionization is centered around the pulse center (0 fs), while the distribution of Auger decay is slightly delayed. Most of the electrons are emitted through Auger decays (8.0 per trajectory on average) rather than direct photoionization (4.2 per trajectory on average). Slightly more electrons are ejected than the total final charge in the system because, at the end of the simulation, we count electrons trapped in the electrostatic potential of one of the atomic ions as bound to it. Fig. 11(c) shows the average charge as a function of time for the individual atoms. The charging up of the molecule starts immediately during the pulse. Initially, all atoms charge up uniformly. Differences between atoms appear at later times when the interatomic distances become larger and the growing potential barriers between the atoms prohibit charge migration. Accordingly, the average charges of atoms C2 and I become distinct from the charges of the other carbon atoms and N at $t > 30$ fs when the interatomic distance between these groups of atoms becomes too large for the continuation of charge migration. Since the C2 and I separate from each other later, charge migration between these two atoms remains possible, and, consequently, the charges on C2 and I become distinct for times $t > 50$ fs.

3.2.2 *Out-of-plane dynamics*

We note that while the molecule is initially planar, a substantial momentum spread appears along the out-of-plane direction for the I and C2 atoms. Because in the simulation the initial conditions do not contain any out-of-plane distortion, this motion must be attributed to the recoil momentum caused by the emission of electrons from the iodine atom. Because ionization occurs predominantly on the iodine atom, the momentum perpendicular to the molecular plane is originally only on the I atom. It is then transferred to other atoms and further enhanced by the repulsion between the atomic ions, especially for I and C2 which come close to each other. These two processes can be associated with different time periods during the simulation. This is illustrated in Fig. 12(a) which shows the out-of-plane forces for I and C2. Two distinct patterns can be recognized in the figure by coloring the forces according to time. For early time $< 10\,\mathrm{fs}$ (blue lines), i. e., during or shortly after the x-ray pulse, the out-of-plane forces dominantly appear for I and are considerably lower for the C2 atom. For later times (orange lines), i. e., after the x-ray pulse, the out-of-plane forces for the two atoms show a strong anti-correlation as a consequence of their mutual repulsion. This out-of-plane momentum is particularly strong for I and C2 and the involved collision dynamics eventually result in a strong anti-correlation in the final momenta that is shown in Fig. 12(b). This strong anti-correlation in the final momenta can thus be seen as a strong fingerprint of the collision between the two atoms, in which the two atoms are initially pushed against each other by the Coulomb forces due to the surrounding ions and later repel each other. The observation of out-of-plane momenta that are amplified due to the collision between the two colliding atoms represents a remarkable manifestation of the photoelectron recoil [17], whose effects have been studied before via photoelectron and Auger spectra [12, 26, 42, 44–48, 73, 79–82]. We note, however, that in the current simulation data, zero-point vibrations are neglected. They can also lead to out-of-plane dynamics that can be further amplified by the collision. In light of these effects, experimental identification of photoelectron recoil effects may require additional information that could be provided via coincident ion and photoelectron detection.

3.2.3 *PCA of final momenta*

To further elucidate how the explosion is mapped into correlations in the final momentum data, we now inspect the principal components of the final momenta. The first four components are illustrated in panels (a)-(d) of Fig. 13 showing projections in the xy and xz planes. The components are depicted by arrows on each atom of the iodopyridine molecule at its ground-state equilibrium geometry. The relative

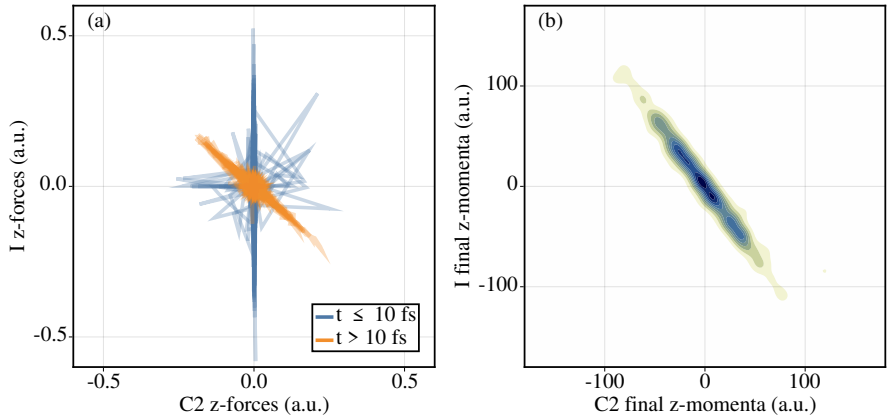

Figure 12: (a) Out-of-plane forces for all trajectories and all times for the C2 and I atoms. Each point on a trajectory is colored according to time. (b) Distribution of the final momentum of the I atom along the z axis against the final momentum of the C2 atom along the z axis.

length of the arrows and their direction indicate the respective amplitude of each atom for the component. We choose to orient the components such that the projection of the final velocities on the component is on average positive.

In Fig. 13(e) we report the ratio of explained momentum variance for the first PCA components. In the following, we discuss the components individually.

The first component describes an anti-correlated out-of-plane movement of the C2 and I atom that has already been discussed in Fig. 12. As can be seen, the variance explained by this component is quite significant compared to the total variance in the final momentum data. The predominant occurrence of this component points to the strong interactions between the two atoms arising from their collision described above. The strong interaction between the C2 and I atoms is further emphasized by the two next components. Component 2 reflects anti-correlated variance of the C2 and I atom along the x-axis, i. e., along the direction of the I-C2 collision. Component 3 corresponds to variations of the I and C2 momenta in the xy plane transverse to the collision direction. Component 4 corresponds to the expulsion of the carbon atoms. A notable feature absent from the first four components is the momenta of the hydrogen atoms.

The predominant occurrence of correlated I-C2 momenta in the dominant components indicates that the collision dynamics of the two atoms strongly impact the final momentum variations. This result suggests that studying covariances of final ion momenta (e.g., via PCA decomposition) gives a clear indication of collisions during the explosion.

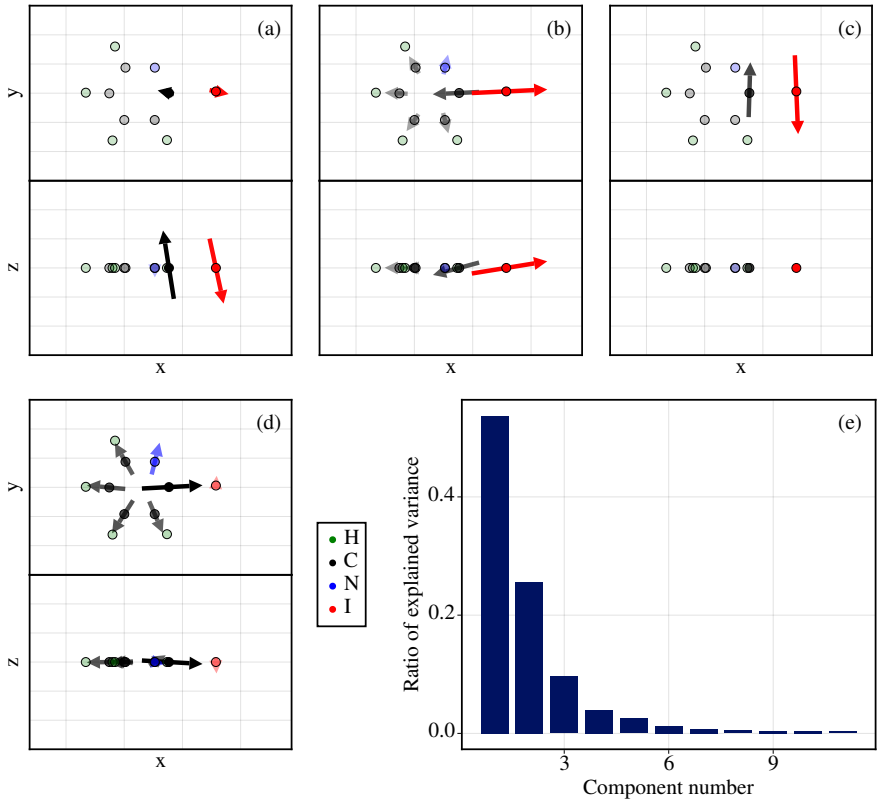

Figure 13: Results of final momentum PCA. (a)-(d): First four PCA components, superimposed on the equilibrium atom positions and ordered by component number. The transparency of the color is set relative to the magnitude of the component at the given atom. Arrows smaller than the atom marker are omitted. (e) Ratio of explained variance of the PCA components.

3.2.4 *PCA of forces*

We now turn to the study of the PCA used on forces. We applied the same procedure as for the final momentum PCA, but now also take into account the dynamical information at each femtosecond time step. Specifically, we consider forces for the time frames between 20 fs before and 150 fs after the pulse center and for all trajectories fulfilling the coincidence condition highlighted in Fig. 9. Each snapshot is considered equally when constructing the covariance matrix. Thus the variance captures not only the ensemble variations but also the changes in the forces in time during each trajectory. The results are reported in Fig. 14 similarly as in Fig. 13. As can be seen, the first component corresponds to the radial expulsion of the iodine, nitrogen, and all the carbon atoms, but C2. The next component is very similar to the final momentum component 2, as it mostly represents the C2-I collision. Component 3 describes the ejection of the hydrogen atoms, coupled with a slight push of the carbon atoms toward the center. This anti-correlated force between the hydrogens and the carbons can be associated with the recoil momentum that the hydrogens imprint on the carbon atoms when they are radially expelled out of the molecule. Accordingly, the corresponding inward contribution for the nitrogen atom that has no hydrogen attached to it is absent. Finally, component 4 depicts the out-of-plane dynamics for C2 and I that have been discussed already in Fig. 12 and are rooted in the electron emission recoil and the C2-I collision.

The first four components obtained by the force PCA exhibit a quite complete picture of the dynamics during the Coulomb explosion. They separate the observed dynamics into remarkably insightful contributions.

3.2.5 *Decomposition of the dynamics*

To underline the advantage of the principal components obtained by force PCA in decomposing the explosion dynamics, we demonstrate to what degree the overall dynamics can be reconstructed using only the four components presented in Fig. 14. In Fig. 15 we describe the overall dynamics employing projections of time-dependent quantities onto the first four components. The left panel shows the average projected displacements, velocities, and forces.

Fig. 15(a) shows the displacements projected onto the components. We see that displacement along the first 3 components increases steadily. Displacement along component 3 increases more rapidly compared to components 1 and 2. Component 4 only shows a marginal increase. The corresponding velocities are shown in Fig. 15(b). As can be seen, they increase up to a certain threshold. For component 2 we see an intermediate maximum followed by a slight decrease until 30 fs. The

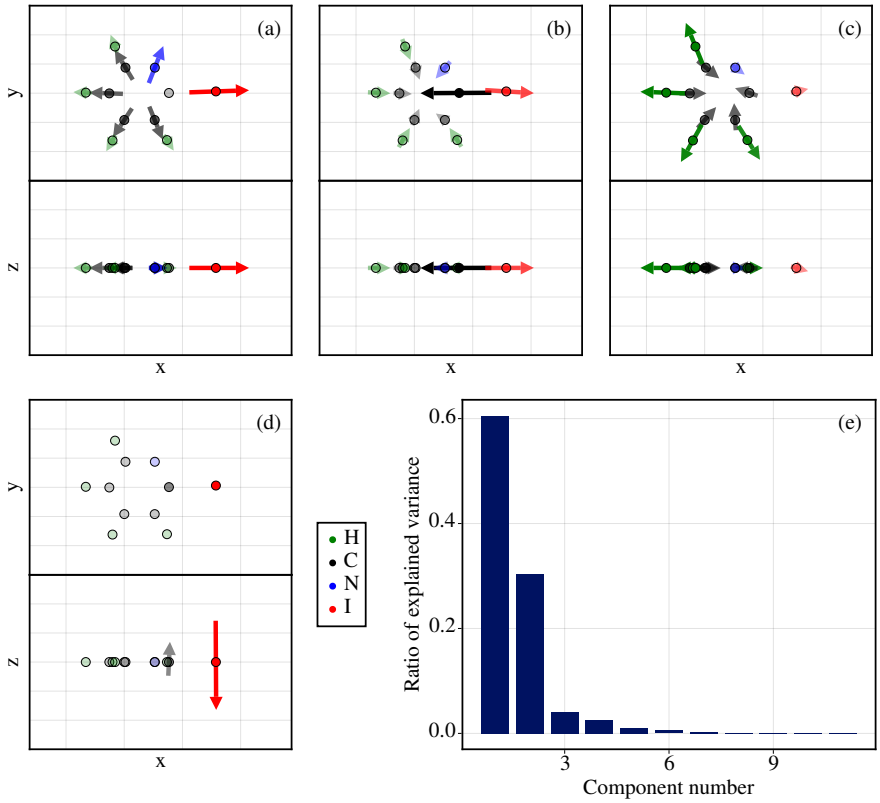

Figure 14: Results of force PCA. (a)-(d) First four PCA components, superimposed on the equilibrium geometry and ordered by component number. The transparency of the color is set relative to the size of the component at the given atom. Arrows smaller than the atom marker are omitted. (e) Ratio of explained variance of the PCA components.

velocities along component 4 remain very low. The forces projected along the components are shown in Fig. 15(c). Whereas the force projected along component 1 rapidly increases at the onset of the explosion and decreases again afterward, components 2 and 3 change their sign during the explosion.

In accordance to what we have discussed in Fig. 10, the decomposition of the dynamics shown in Fig. 15(a), Fig. 15(b), and Fig. 15(c) shows that the Coulomb explosion of the molecule can be mostly described by the concerted acceleration of the atoms into the direction of component 1 (i. e., the radial explosion) having the largest absolute force at all times. On top of that, the average dynamics are modulated by components 2 and 3. For component 2, the change in sign of force reflects the reversion of movement of C2. The modulation by component 3 adds the temporal sequence in which hydrogens are emitted first, followed later by the heavier carbon-nitrogen ring, and finally by the C2-I pair. The additional component 4 describes the out-of-plane dynamics of the I and C2 atoms. This component should be zero by symmetry; the relatively small number (below 0.03 a.u.) must be attributed to finite sampling.

The right panel in Fig. 15 shows the variances in displacement [Fig. 15(d)], velocity [Fig. 15(e)], and force [Fig. 15(f)] at each femtosecond time step. The fraction of the variance that can be explained with the first four components obtained by force PCA is indicated in dark purple, whereas the variance that is not covered by the first four components is indicated in light purple. For the position [Fig. 15(d)], we can see that due to the explosion the variances steadily increase. At the same time, the variance in velocity [Fig. 15(e)] peaks shortly after the pulse and then decreases to a constant plateau value. The variance in force [Fig. 15(f)] shows two peaks, one at $\sim$ 0 fs, where the molecule starts to build up charge, and the second is at $\sim$ 30 fs, where the collision occurs between C2 and I.

As indicated by the color code in the right panels of Fig. 15, a large contribution of variance in the trajectories occurs within the first four components obtained by force PCA and the variance outside the space spanned by these components (the *unexplained* variance indicated in light purple) is relatively small. This result demonstrates that the first four components obtained by force PCA cover most of the dynamics during the explosion. We are thus able to describe the relevant mechanisms during the explosion with a reduction in dimensionality from 33 to four.

3.2.6 *Relation of ion kinetic energy to charge buildup time*

Finally, we look at the relation between the charging up of the molecule and the final momenta. As a measure of the time an atom needs to charge up, we use the time difference between the first and last charge

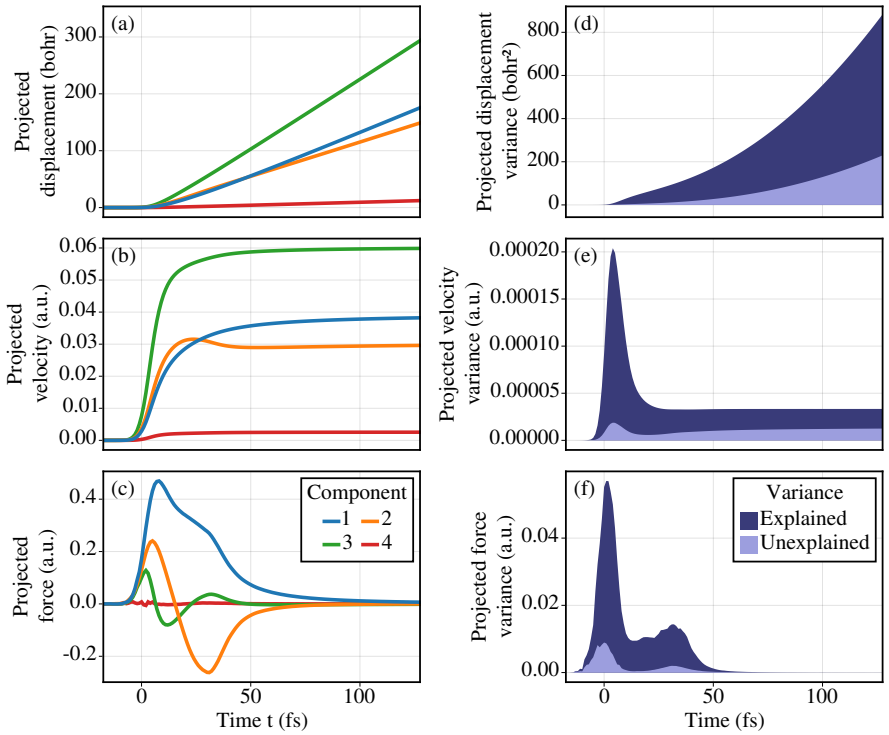

Figure 15: Left panel: Average of the dynamics projected on the first 4 force PCA components in time. (a) Projected displacements. (b) Projected velocities. (c) Projected forces. Right panel: Variance explained by the projections on the first 4 force PCA components in time. (d) Variance of projected displacements. (e) Variance of projected velocities. (f) Variance of projected forces.

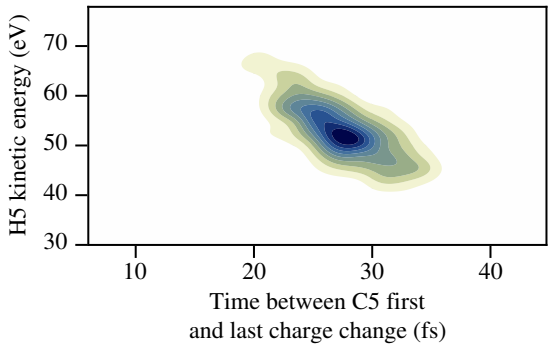

Figure 16: Kinetic energy release of the H5 atom against the time difference between the first and last charge transfer events to the C5 atom for final charge +2 on C5 and +1 on H5.

transfer events from or to a given atom as modeled in our simulation via over-the-barrier charge rearrangement processes. We show in Fig. 16 the relation between the charge buildup time of atom C5 and the kinetic energy of the connected hydrogen atom H5. To ensure a strong effect, we look at higher charges and in contrast to the data inspected before, the figure contains data from all trajectories filtered for a final charge of +2 on C5 and +1 on H5 (irrespective of the charge states of all other atoms). As can be seen, the hydrogen momentum shows a strong anti-correlation to the charge buildup time of the associated heavier atom. For a larger charge buildup time of C5, we can see that the H5 momentum is considerably smaller, whereas with fast charging-up of C5 the H5 momentum is larger. Given the fact that the charge is predominantly created on the iodine atom on the opposite side of the molecule, the relation of the kinetic energy of H5 and charge timing of the carbon atom C5 suggests an opportunity to probe the timescale of charge-buildup on specific atoms and charge redistribution in the molecule.

3.3 CONCLUSION

We have performed a detailed analysis of the simulated dynamics of the x-ray induced Coulomb explosion of iodopyridine. The x-ray induced Coulomb explosion of the molecule involves an interplay between the expulsion of hydrogens, the expulsion of heavier atoms, and the collision dynamics of the C2 atom with the iodine atom. Among other effects, the collision amplifies out-of-plane motion, producing strongly correlated out-of-plane momenta for the I and C2 ions.

PCA applied to final momenta results in components that contain fingerprints of this atomic collision. This result suggests that colliding atoms may be identified via correlation in their final momenta. It is

important to note that ground-state vibrations in the molecule have been neglected in the simulation data considered here. In Chapter 4, we go one step further by including the ground-state vibrations and showing how they affect the resulting principal components.

Studying the data employing force PCA reveals components that facilitate the description of the dynamics during the explosion. They allow us to interpret the overall dynamics using a limited number of components. Employing only 4 components, we are able to follow the main average dynamics and describe the dominant part of the variances during the Coulomb explosion.

Our analysis demonstrates that representing data from the simulations of Coulomb explosion employing components obtained from force PCA is a useful tool allowing to highlight the most relevant dynamical motifs and achieve a decomposition of the dynamics in reduced dimensionality. The method discussed here can be applied to other molecules, also non-planar ones. We expect that the number of components needed to comprehensively reconstruct the dynamics grow with the complexity of the explosion dynamics, which increases considerably when more ion collisions occur.

4

MAPPING OF THE INITIAL STATE THROUGH CORRELATIONS

This chapter adapts a manuscript that was later accepted for publication, after modifications:

Benoît Richard, Rebecca Boll, Sourav Banerjee, Julia M. Schäfer, Zoltan Jurek, Gregor Kastirke, Kilian Fehre, Markus S. Schöffler, Nils Anders, Thomas M. Baumann, Sebastian Eckart, Benjamin Erk, Alberto De Fanis, Reinhard Dörner, Sven Grundmann, Patrik Grychtol, Max Hofmann, Markus Ilchen, Max Kircher, Katharina Kubicek, Maksim Kunitski, Xiang Li, Tommaso Mazza, Severin Meister, Niklas Melzer, Jacobo Montano, Valerija Music, Yevheniy Ovcharenko, Christopher Passow, Andreas Pier, Nils Rennhack, Jonas Rist, Daniel E. Rivas, Daniel Rolles, Ilme Schlichting, Lothar Ph. H. Schmidt, Philipp Schmidt, Daniel Trabert, Florian Trinter, Rene Wagner, Peter Walter, Pawel Ziolkowski, Artem Rudenko, Michael Meyer, Robin Santra, Ludger Inhester, and Till Jahnke

"Imaging correlated quantum fluctuations of the structure of a complex molecule"
Science **389** 650-654 (2025)

The content of Section 4.2 has been previously published as the documentation of the package:

Benoît Richard
NormalModes.jl
`https://github.com/Kolaru/NormalModes.jl`

Contribution to the manuscript I ran the simulations using an updated version of XMDYN, proposed the analysis of the mapping from vibrational modes to principal components, and performed it. I proposed, developed, implemented, optimized and tested the reconstruction algorithm. I cleaned the experimental data, and applied the algorithm to it and to the simulation data, and compared the results. I drafted the manuscript with input from other coauthors.

4.1 GROUND-STATE FLUCTUATIONS

In the previous chapters, we have pictured molecules as mostly classical objects. In doing so, we have neglected one of their most important quantum property: their vibrational ground-state fluctuation (VGSF), also known as zero-point vibrations. Even when a molecule is in its lowest energy state, its wavefunction spreads in both position and momentum space, leading to a distribution of structures that can be interpreted as fluctuations of the equilibrium geometry. The VGSF are responsible for phenomena such as quantum phase transitions [85], have dramatic consequences for molecular reaction dynamics [14], and can play a decisive role for hydrogen bonds [53].

The first experimental demonstration of this quantum phenomenon was presented one hundred years ago by Robert S. Mulliken, when he compared energy levels of different isotopologues [66]. By now, the measurement of discrete molecular energy levels and the identification of zero-point energies are routine [62]. However, typically one only measures the indirect consequences of the VGSF on molecular energy levels, but cannot observe the molecular structure variations themselves, which would add important insight to understanding collective quantum phenomena such as the ones mentioned above.

Imaging the VGSF of a molecule is out of reach for traditional imaging techniques such as crystallography, because the VGSF is intrinsically a high-dimensional effect. To get access to ground-state fluctuations, it is necessary to target individual molecules and access the entire distribution of structures. From this distribution, the structural dependencies among the atoms constituting the molecule can be determined. Thanks to its ability to measure atomic fragments in coincidence, CEI is uniquely suited to perform such single-molecule measurements.

In this Chapter, we briefly review the harmonic approximation for ground-state fluctuations. In particular, introduce the decomposition of the ground state into normal modes (also known as vibrational modes), and how it is used to sample of the ground-state distribution through the Wigner quasiprobability distribution. Finally, we perform simulations that include ground-state fluctuations, and show that they are mapped to the principal components of the momenta measured after the Coulomb explosion of the molecule.

4.2 NORMAL MODES' DECOMPOSITION

We study a molecule composed of N atoms. The atom with index a has position $\mathbf{x}_a$ and momentum $\mathbf{p}_a$. It is however not convenient to

work with individual atoms, therefore we concatenate the positions and momenta into single vectors of dimension $3N$, namely

$$\mathbf{x} = \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_N \end{pmatrix} \qquad \text{and} \qquad \mathbf{p} = \begin{pmatrix} \mathbf{p}_1 \\ \mathbf{p}_2 \\ \vdots \\ \mathbf{p}_N \end{pmatrix}. \tag{3}$$

In this thesis we use indices $a, b, c, \ldots$ to represent indices of *atoms* (ranging from 1 to N) and indices $i, j, k, \ldots$ to represent indices in the concatenated vectors (ranging from 1 to $3N$).

Using this notation, the molecular Hamiltonian is

$$\mathcal{H} = \mathcal{T}(\mathbf{p}) + \mathcal{V}(\mathbf{x}), \tag{4}$$

where $\mathcal{T}(\mathbf{p})$ is the kinetic energy operator, depending only on the momenta $\mathbf{p}$ of the atoms, and $\mathcal{V}(\mathbf{x})$ is the potential energy operator, depending only on the positions $\mathbf{x}$ of the atoms.

The potential depends on the position of the atoms, as well as the electronic state of the system. To simplify the picture, we expand the potential $\mathcal{V}(\mathbf{x})$ around the equilibrium geometry where the gradient of the potential, $\nabla\mathcal{V}(\mathbf{x})$, is zero, yielding

$$\mathcal{V}(\mathbf{x}) = V_0 + \mathbf{x}^T \nabla\mathcal{V}(\mathbf{x}) + \frac{1}{2}\mathbf{x}^T \mathbf{H} \mathbf{x} + \mathcal{O}(\mathbf{x}^3). \tag{5}$$

Here V_0 is a constant shift that we set to zero without loss of generality, $\nabla\mathcal{V}(\mathbf{x})$ is the gradient of the potential, which is also zero as mentioned above, and $\mathbf{H}$ is the Hessian matrix of the potential. The expansion to the leading nonzero term gives the quadratic form

$$\mathcal{V}(\mathbf{x}) \approx \frac{1}{2}\mathbf{x}^T \mathbf{H} \mathbf{x}. \tag{6}$$

In the position representation, the Hamiltonian then reads

$$\mathcal{H} = \sum_i \frac{-\hbar^2}{2m_i} \frac{\partial^2}{\partial x_i^2} + \frac{1}{2}\mathbf{x}^T \mathbf{H} \mathbf{x}, \tag{7}$$

where m_i is the mass of the atom corresponding to dimension i. This Hamiltonian describes a multidimensional harmonic oscillator. The idea of normal mode decomposition is to change the basis to express turn the Hamiltonian into a set of *uncoupled* harmonic oscillators.

To this end, we first introduce the diagonal mass-weighting matrix $\mathbf{M}$, with elements $\mathbf{M}_{ii} = m_i^{-\frac{1}{2}}$. We decompose the weighted Hermitian as

$$\mathbf{MHM} = \mathbf{V}\boldsymbol{\Omega}^2\mathbf{V}^T, \tag{8}$$

where $\mathbf{V}$ is the orthogonal matrix whose columns are the eigenvectors and $\boldsymbol{\Omega}^2$ the diagonal matrix of eigenvalues. Since the Hessian matrix

$\mathbf{H}$ is semi-positive definite, all the eigenvalues are positive or zero[1], allowing us to write the eigenvalues as $\mathbf{\Omega}_{jj}^2 = \omega_j^2$.

We define the *normal coordinates* $\mathbf{z}$ as

$$\mathbf{z} = \mathbf{V}\mathbf{M}^{-1}\mathbf{x} \qquad \text{or, equivalently,} \qquad \mathbf{x} = \mathbf{M}\mathbf{V}\mathbf{z}. \tag{9}$$

Inserting this expression for $\mathbf{x}$ in the Hamiltonian, we get the decoupled system of harmonic oscillators (see Appendix A for the derivation):

$$\mathcal{H} = \sum_j \frac{-\hbar^2}{2}\frac{\partial^2}{\partial z_j^2} + \frac{1}{2}\omega_j^2 z_j^2. \tag{10}$$

The columns of $\mathbf{V}$ are the *normal modes* of the ground state. According to Eq. (9), the corresponding *vibrational modes* in position space are the columns of $\mathbf{MV}$. Vibrational modes are often used to depict the ground-state fluctuations, as they can be interpreted as a distortion of the ground-state equilibrium geometry of the molecule.

4.3 WIGNER SAMPLING OF THE GROUND STATE

To better understand the ground-state fluctuations of the molecule, we look at its wavefunction. The solution to the time-independent Schrödinger equation is

$$\Psi(\mathbf{z}) = \prod_{j=1}^{3N} \psi_j(z_j), \tag{11}$$

where $\psi_j(\mathbf{x})$ is an eigenfunction of the j-th harmonic oscillator, fulfilling the corresponding Schrödinger equation

$$\left[\frac{-\hbar^2}{2}\frac{\partial^2}{\partial z_j^2} + \frac{1}{2}\omega_j^2 z_j^2\right]\psi_j(z_j) = E_j\psi_j(z_j), \tag{12}$$

where E_j is the energy. The ground state is the solution with lowest E_j and is a Gaussian,

$$\psi_j(z_j) = \frac{1}{\sqrt{2}}\left(\frac{\omega_j}{\pi\hbar}\right)^{\frac{1}{4}} \exp\left[-\frac{\omega_j z_j^2}{2\hbar}\right]. \tag{13}$$

Its associated Wigner quasiprobability distribution, using p_{zj} for the momenta corresponding to coordinate z_j, is

$$W_j(z_j, p_{zj}) = \frac{1}{\pi\hbar}\exp\left[-\frac{\omega_j z_j^2}{\hbar}\right]\exp\left[-\frac{p_{zj}^2}{\hbar\omega_j}\right]. \tag{14}$$

1 Frequencies very close to zero (corresponding to translations and rotations) can be negative when computed numerically, due to the finite precision of the computation. This is compensated by setting them to zero.

This is simply a product of two independent Gaussian distributions, therefore z_j and p_{zj} are independent and normally distributed, with variances

$$\mathrm{var}\left[z_j\right] = \frac{1}{2}\frac{\hbar}{\omega_j} \qquad \text{and} \qquad \mathrm{var}\left[p_{zj}\right] = \frac{1}{2}\hbar\omega_j. \tag{15}$$

The distribution for the vectors $\mathbf{z}$ and $\mathbf{p}_z$ are multivariate Gaussian, with the diagonal covariance matrices given by

$$\mathrm{var}[\mathbf{z}] = \frac{\hbar}{2}\boldsymbol{\Omega}^{-1} \qquad \text{and} \qquad \mathrm{var}[\mathbf{p}_z] = \frac{\hbar}{2}\boldsymbol{\Omega}. \tag{16}$$

The real space coordinates $\mathbf{x}$ and $\mathbf{p}$ are linear combinations of the normal coordinates, and are thus normally distributed, with covariance matrices

$$\begin{aligned} \mathrm{var}[\mathbf{x}] &= \frac{\hbar}{2}\mathbf{M}\mathbf{V}\boldsymbol{\Omega}^{-1}\mathbf{V}^T\mathbf{M}^T \qquad \text{and} \\ \mathrm{var}[\mathbf{p}] &= \frac{\hbar}{2}\mathbf{M}^{-1}\mathbf{V}\boldsymbol{\Omega}\mathbf{V}^T\mathbf{M}^{-T}. \end{aligned} \tag{17}$$

Therefore, to sample the Wigner distribution of the ground state in the harmonic approximation, we first determine the equilibrium geometry and the hessian matrix of the potential. Then, we generate the random samples according to Eq. (17).

4.4 NORMAL MODES AS PRINCIPAL COMPONENTS

PCA and the decomposition into normal modes are related. This is not immediately evident, because $\mathbf{MV}$ in Eq. (17) is not an orthogonal matrix.

We can fix that by weighting the momenta by the inverse of the square root of the mass of each atom, defining $\mathbf{Mp}$ to be the *weighted momenta*. These also follow a multivariate Gaussian distribution, with covariance matrix

$$\mathrm{var}[\mathbf{Mp}] = \frac{\hbar}{2}\mathbf{V}\boldsymbol{\Omega}\mathbf{V}^T. \tag{18}$$

In comparison, performing PCA on the weighted momenta, we get

$$\mathrm{var}[\mathbf{Mp}] = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^T. \tag{19}$$

Thus, in this case, $\mathbf{U} = \mathbf{V}$ and $\boldsymbol{\Lambda} = \frac{\hbar}{2}\boldsymbol{\Omega}$. In other words, the normal modes (the columns of $\mathbf{V}$) are the principal components of the distribution of weighted momenta, and the frequencies of the vibrational modes are related to the explained variances as $\lambda_j^2 = \frac{\hbar}{2}\omega_j$.

4.5 SIMULATION WITH GROUND-STATE FLUCTUATIONS

To elucidate how the ground-state fluctuations relate to the observed correlations in the measured ion momenta, we perform simulations similar to the one presented in Chapter 2, (Simulation Set IV in Table 2). The main differences are the sampling of the ground-state fluctuations and the inclusion of chemical bonds, necessary to avoid the dissociation of the molecule due to the nonzero momenta present in the ground state.

Chemical bonds are approximated by using the reactive force field (ReaxFF) method [71, 92]. We interface the sPuReMD [2] implementation of ReaxFF with XMDYN, integrating the forces provided by the forcefield to the dynamic computed by XMDYN. The forcefield is tuned for iodopyridine based on existing parameter tables [22, 60, 65], with the additional help of the JAX-ReaxFF framework [40, 41], to fit the forcefield to reference data computed with XMOLECULE [27], employing the Hartree-Fock method and the 3-21G basis set [6, 16], and to available spectroscopic data [75]. During the explosion, the forcefield is linearly damped proportionally to the charges in the system, until totally disabled once 4 electrons have been removed. This simulates the weakening of the chemical bonds as the charging up of the molecule progresses.

For the initial states of the trajectories, we draw 61 652 samples of positions and momenta from the ground-state Wigner distribution, as described in Section 4.3, using the ReaxFF Hessian.

We restrict our analysis to the I^{4+}/N^{2+} coincidence channel. To increase the number of trajectories reaching this final charge state, we use a higher fluence than previously, 1.5×10^{11} photons/μm^2. This coincidence channel is chosen to ensure quick removal of the valence electrons, such that the bonding plays a minimal role during the explosion dynamics, while still being experimentally achievable.

For completeness, the parameters of this set of simulations are detailed in Table 3.

4.6 CORRELATION BETWEEN NORMAL MODES AND PRINCIPAL COMPONENTS

In Chapter 3, we describe how correlations between the asymptotic momenta of the atom carry information about the dynamics of the explosion. Now, we study how they carry information about the ground state of the molecule when the x-ray pulse hits it. Again, we use PCA, to describe the correlations between asymptotic momenta. To weight the momentum coordinates appropriately, we perform PCA on the momenta weighted by the inverse of the square root of the mass of the atoms. The principal components are numbered in decreasing order of their explained variance.

Simulation software	XMDYN
Number of trajectories	61 652
Initial state	Geometry and momenta sampled from the ground-state Wigner distribution
Chemical bonds	ReaxFF
Time step	1 as
Simulation duration	1 ps (10^6 time steps)
Photon energy	2 keV
Pulse shape	Gaussian
Pulse length	10 fs (FWHM)
Polarization	Random
Fluence	1.5×10^{11} photons/μm^2

Table 3: **Parameters of the simulations.**

To determine the relation between the vibrational modes and the principal components of the weighted momenta, we compute the normal coordinates $\mathbf{z}$ and $\mathbf{p}_z$ of the initial state by projecting them on the normal modes according to Eq. (9). For each trajectory, the initial state is the simulation frame recorded just before the first photoionization.

Then, we project the asymptotic weighted momentum $\mathbf{Mp}$ on the principal components to obtain the principal coordinates $\mathbf{c} = \mathbf{U}^T\mathbf{Mp}$. We study the relation between mode i and component j using the Pearson correlation coefficient, defined as

$$\mathrm{cor}[X, Y] = \frac{\mathrm{cov}[X, Y]}{\mathrm{std}[X]\,\mathrm{std}[Y]}, \tag{20}$$

where cov[.,.] is the covariance and std[.] the standard deviation. We consider

$$r_{ij}^2 = \mathrm{cor}\left[z_i, c_j\right]^2 + \mathrm{cor}\left[p_{zi}, c_j\right]^2, \tag{21}$$

which includes the contribution of both the initial positions and initial momenta. When the initial positions z_i and momenta p_{zi} are in a perfect linear relation to the principal coordinates c_j, we have

$$\sum_i r_{ij}^2 = 1. \tag{22}$$

This is a consequence of the properties of the correlation coefficient, which provides an absolute scale to judge the quality of the linear mapping of the vibrational modes into the principal components.

Moreover, assuming a linear mapping, the independence of the z_i and p_{zi} implies (see Appendix B for details)

$$\frac{c_j}{\mathrm{std}\left[c_j\right]} = \sum_i \mathrm{cor}\left[z_i, c_j\right] \frac{z_i}{\mathrm{std}[z_i]} + \mathrm{cor}\left[p_{zi}, c_j\right] \frac{p_{zi}}{\mathrm{std}[p_{zi}]}. \tag{23}$$

Therefore, the r_{ij}^2 quantifies how much vibrational mode i affects principal component j in a linear model, it is the *mode contribution* of a vibrational mode on a principal component.

The values of r_{ij}^2 are shown in Fig. 17A. The contributions are particularly strong between out-of-plane vibrational modes and out-of-plane principal components, as indicated by darker squares. They are much stronger than their in-plane counterparts and, as expected due to the planar symmetry of the molecule, there is no cross-contribution between out-of-plane and in-plane modes and components.

This shows that the vibrational modes are mapped to principal components of the asymptotic momentum distribution. To get a more quantitative look at the linear relation between the two, we plot the sum of all mode contributions in Fig. 17B. We see that it is excellent for the out-of-plane components, as the total contribution is close to 1 (perfect linear relation), while significantly smaller for the in-plane components. We attribute this difference to the simplicity of the out-of-plane dynamic, in which small distortions are amplified while no complex interaction, such as a collision, is possible. Two effects can explain the lower correlations in the in-plane case. On the one hand, a Coulomb explosion is in general complex and highly nonlinear. This can not be captured in the present analysis as the correlation coefficient we are using only captures linear dependencies. On the other hand, some information may be lost due to sensitivity to the stochastic nature of the charging-up process induced by a pulse of finite duration, which is again a result of the nonlinear nature of the explosion that manifests more strongly in-plane than out-of-plane.

In the rest of this thesis, we focus on the out-of-plane features, that are more prominent and thus more likely to be found in the experiment. In particular, we see in Fig. 17A that some vibrational modes have a significant contribution to some dominant out-of-plane principal components, the latter are therefore acting as a fingerprint of the former. To study in more detail, we show three of the dominant out-of-plane components in Fig. 18A, E, and I, where the arrows represent the relative variations in the weighted momenta of each atom. The contributions of the vibrational modes are reported in the corresponding panel, Fig. 18B, F, and J. Finally, the right panels show the two vibrational modes with the highest contribution to each of the depicted principal components. The vibrational modes are represented in position space weighted by the inverse square root of the mass of the atoms, which is the natural space for normal coordinates (Eq. (9)) and captures the relative importance of atoms of different masses on the overall dynamic.

The connection between the vibrational modes and the principal components is visually evident: they tend to have similar features, as, for example, between principal component (PC) 3 (Panel A) and the vibrational mode at $1212\,\mathrm{cm}^{-1}$ (Panel D). In this case, the initial

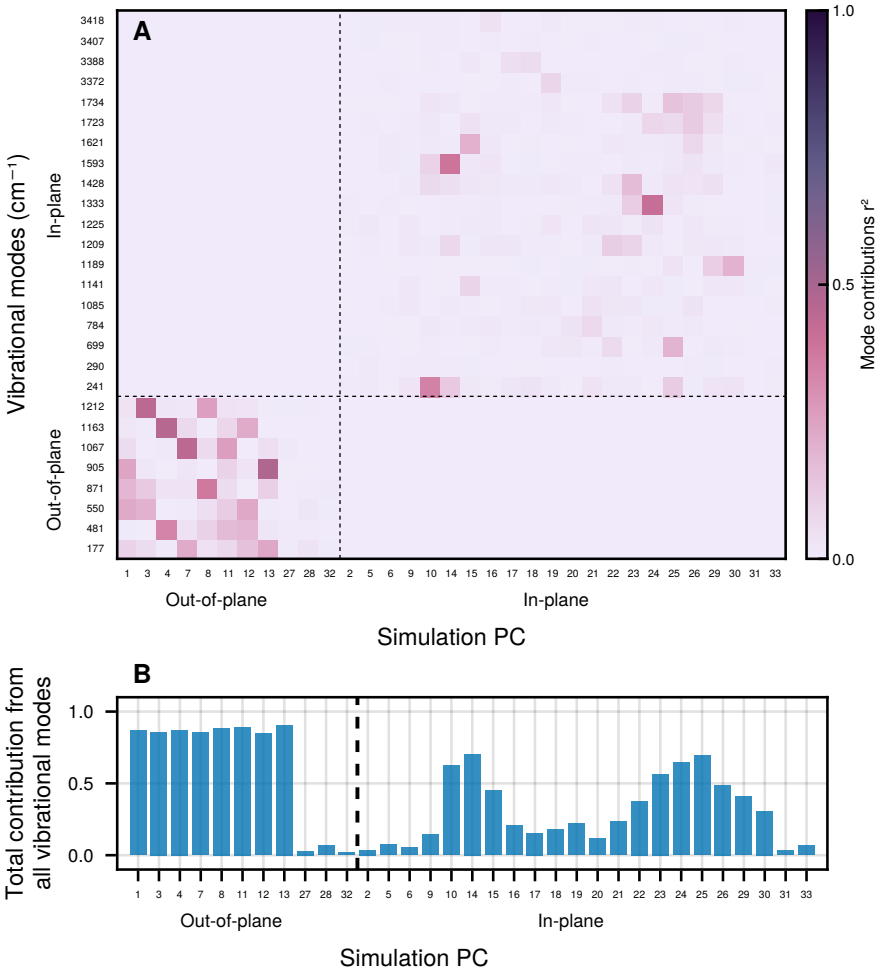

Figure 17: **Contribution of the ground-state vibrational modes to the principal components of the asymptotic momenta.** A. Contribution of each vibrational mode to each principal component (PC). B. Total contribution from all vibrational modes to each principal component (PC).

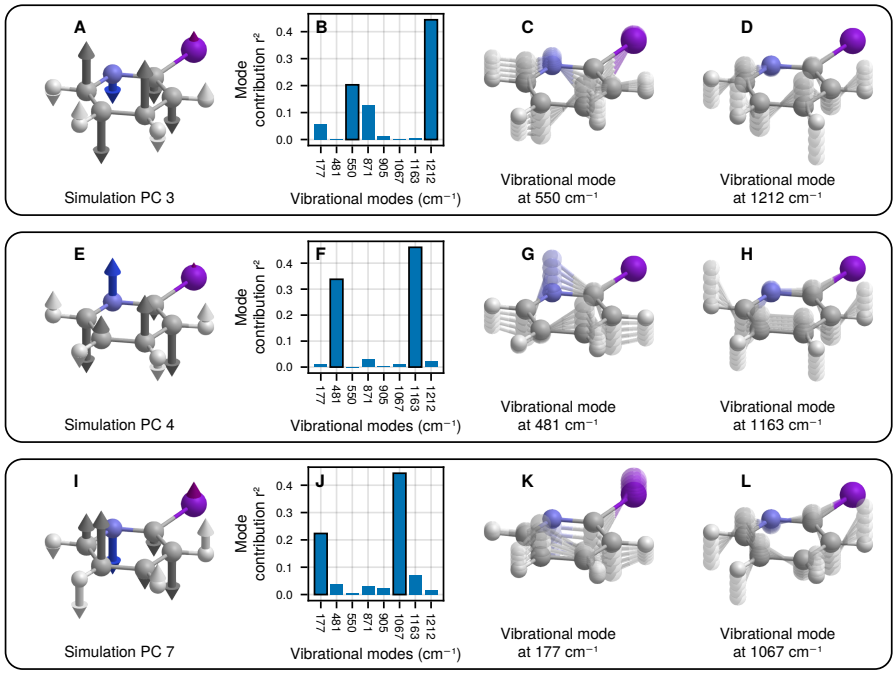

Figure 18: **Mapping of the principal components to the ground-state vibrational modes.** A, E, I. Principal components computed from the simulation. B, F, J. Contribution r^2 of each vibrational mode to the principal components. The two leading contributions are highlighted and the corresponding vibrational modes are shown in the next panels. C-D, G-H, K-L. Vibrational modes with the most dominant contribution. The vibrational modes are shown in position space weighted by the inverse square root of the mass of the atoms.

displacement along each vibrational mode is mapped to the final momenta with a similar pattern. The similarity is lower for the PC 3 (panel A) and the vibrational mode at 550 cm^{-1} (Panel C), because while the correlated motion of the heavy atoms follow a similar pattern, it is different for the hydrogen atoms. In this case, the distortion caused by the vibrational mode cannot be directly transferred to the corresponding principal component.

Moreover, in PC 3, PC 4 and PC 7 (Fig. 18A, E and I) we see patterns where neighboring atoms move in opposite directions. They tend to be favored by the explosion dynamics: since the atoms overwhelmingly interact through the Coulomb repulsion, they push themselves away from each other, resulting in alternating patterns that are more likely to appear in the leading principal components of the asymptotic momenta.

In conclusion, we show that principal components carry information about the VGSF of the molecule, and that we can quantify it in our simulations. For the iodopyridine molecule, the main and clearest contributions are between out-of-plane vibrational modes and out-of-

plane principal components, indicating that our analysis of the experiment for this molecule should focus on this part of the data.

4.7 CONCLUSION

In this Chapter, we discuss the concept of the ground-state fluctuations of a molecule and how it can be described in the harmonic approximation. To solve the approximate Schrödinger equation for the ground state, the problem is decomposed into normal modes. Knowing the solution, we can approximate the distribution of initial geometries using Wigner sampling.

Then, we sample the ground-state distribution of iodopyridine and add quantum fluctuations to the simulations of its Coulomb explosion. To be able to incorporate quantum fluctuation in both positions and momenta, we approximate the chemical bond through ReaxFF. By analyzing the result of the simulations, we demonstrate that the fluctuation along vibrational modes of the molecule are mapped into variation of the asymptotic momenta along specific principal components. This result extends the horizon of possibilities for CEI, as it shows that the method has the power to image the collective behavior of the atom in the distribution of initial geometries of the molecule, a unique property unavailable to other imaging techniques.

The ground-state fluctuation is a single example of such collective behaviors. Understanding them may bring unprecedented insight into the property of the molecular systems and may help to untangle the complex dynamics in ensemble of molecules following diverse pathways.

5

ALGORITHM TO FIND CORRELATION IN EXPERIMENTAL DATA

This chapter adapts a manuscript that was later accepted for publication, after modifications:

Benoît Richard, Rebecca Boll, Sourav Banerjee, Julia M. Schäfer, Zoltan Jurek, Gregor Kastirke, Kilian Fehre, Markus S. Schöffler, Nils Anders, Thomas M. Baumann, Sebastian Eckart, Benjamin Erk, Alberto De Fanis, Reinhard Dörner, Sven Grundmann, Patrik Grychtol, Max Hofmann, Markus Ilchen, Max Kircher, Katharina Kubicek, Maksim Kunitski, Xiang Li, Tommaso Mazza, Severin Meister, Niklas Melzer, Jacobo Montano, Valerija Music, Yevheniy Ovcharenko, Christopher Passow, Andreas Pier, Nils Rennhack, Jonas Rist, Daniel E. Rivas, Daniel Rolles, Ilme Schlichting, Lothar Ph. H. Schmidt, Philipp Schmidt, Daniel Trabert, Florian Trinter, Rene Wagner, Peter Walter, Pawel Ziolkowski, Artem Rudenko, Michael Meyer, Robin Santra, Ludger Inhester, and Till Jahnke

"Imaging correlated quantum fluctuations of the structure of a complex molecule"
Science **389** 650-654 (2025)

Contribution to the manuscript I ran the simulations using an updated version of XMDYN, proposed the analysis of the mapping from vibrational modes to principal components, and performed it. I proposed, developed, implemented, optimized and tested the reconstruction algorithm. I cleaned the experimental data, and applied the algorithm to it and to the simulation data, and compared the results. I drafted the manuscript with input from other coauthors.

Note. In this chapter, we use $\mathbf{p}$ to denote the momenta of the atoms weighted by the inverse of the square root of their mass, as defined in Section 4.4.

5.1 GENERAL STRATEGY

In the previous Chapters, we showed that important phenomenons leave a fingerprint in CEI data, in the form of correlations between the

momenta of the atoms revealed by using PCA on the simulated data. Can the same features be identified in the experimental data?

To answer this question, we need to apply PCA to the experiment, but three major limitations prevent the direct use of the method:

1. The orientation of each molecule in the laboratory frame is random. Therefore, the molecular frame has to be determined based on the recorded ion momenta.

2. Only a subset of all ions created by the x-ray pulse is detected. In the experiment that we study, described in Chapter 2, about half of the ions (5.3 detected ions out of 11 atoms on average) are detected in coincidence on average, due to the finite efficiency of the detector.

3. While the mass-to-charge ratio of each detected ion is measured in the experiment, its initial position within the molecule is unknown for those cases where multiple atoms of the same element exist. We say that the *identity* of the atoms is unknown.

To overcome these three challenges, our general strategy is to fit a $3N$-dimensional Gaussian distribution ($N = 11$ for iodopyridine) to the partial coincidence data available from the experiment and then compute PCA from the fitted covariance matrix according to Eq. (2). This fitting procedure is as follows.

First, to make the fit independent of the initial (laboratory-frame) orientation of the molecule, we use the scalar products between the measured weighted momenta as coordinates. We consider all possible scalar products $(\mathbf{p}_a \cdot \mathbf{p}_b)$, where the ions a and b are detected in coincidence. These coordinates are independent of the rotation (i. e., laboratory-frame orientation) of the molecule, solving the first challenge.

Second, to perform the fit with partial coincidence data, we use the statistical moments of the distribution of scalar products. As an example, one of the scalar-product moments of order two is

$$\langle (\mathbf{p}_a \cdot \mathbf{p}_b)\ (\mathbf{p}_c \cdot \mathbf{p}_d) \rangle. \tag{24}$$

Since the moments only involve a limited number of ions (in the given example, the ions a, b, c, and d), they can be computed as long as a sufficient amount of coincident measurements containing all necessary atoms have been recorded, and regardless of whether any of the other atoms are missing, thus solving the second challenge.

Finally, we compute the moments *per element*, rather than per atom. Therefore, not knowing the identity of the atoms of the same element in the experiment is not a problem, and the third challenge is solved.

We name these moments the per-element moments of the scalar products (PEMSP). In the next sections, we show how they are computed in practice, and we describe how we compare the PEMSP computed from the experimental data and from a Gaussian model to

perform the fit. Finally, we use the simulations to confirm that this method allows to recover the principal components of the distribution of weighted-momenta.

5.2 COMPUTING THE PEMSP FROM EXPERIMENTAL DATA

The PEMSP are calculated in the following steps. We compute the scalar products between the weighted momenta $\mathbf{p}_a$ of each pair of ions in a single measurement, and label them by the elements forming the pair. As an example, consider shots in which 3 ions are detected in coincidence, (I^+, C^+, C^+). From each such measurement, we build 6 scalar products, namely

$$\begin{aligned}&(\mathbf{p}_\mathrm{I}\cdot\mathbf{p}_\mathrm{I}),(\mathbf{p}_\mathrm{I}\cdot\mathbf{p}_\mathrm{C}),(\mathbf{p}_\mathrm{I}\cdot\mathbf{p}_{\mathrm{C}'})\\&(\mathbf{p}_\mathrm{C}\cdot\mathbf{p}_\mathrm{C}),(\mathbf{p}_\mathrm{C}\cdot\mathbf{p}_{\mathrm{C}'}),(\mathbf{p}_{\mathrm{C}'}\cdot\mathbf{p}_{\mathrm{C}'}).\end{aligned} \tag{25}$$

Some scalar products, in the example $(\mathbf{p}_\mathrm{C}\cdot\mathbf{p}_\mathrm{C})$ and $(\mathbf{p}_{\mathrm{C}'}\cdot\mathbf{p}_{\mathrm{C}'})$, are indistinguishable, because we do not know the identity of the atoms. We can, however, distinguish between the scalar product of one carbon ion momentum with itself ($(\mathbf{p}_\mathrm{C}\cdot\mathbf{p}_\mathrm{C})$) and the scalar product between the momenta of two different carbon ions ($(\mathbf{p}_\mathrm{C}\cdot\mathbf{p}_{\mathrm{C}'})$), since they are detected in coincidence.

We write $M_\mathbf{i}$ the moments of the scalar-product distribution. A moment of order n is specified by the label $\mathbf{i}$, which is a set of pairs of measured ions,

$$\mathbf{i}=\{(a_1,b_1),(a_2,b_2),\ldots,(a_n,b_n)\}. \tag{26}$$

A moment is computed as

$$M_\mathbf{i}=\frac{1}{Y_\mathbf{i}}\sum_k\sum_{\substack{\mathbf{j}\,\mathrm{in}\,k\\ \mathbf{j}\sim\mathbf{i}}}\prod_{(a,b)\in\mathbf{j}}(\mathbf{p}_a\cdot\mathbf{p}_b)^{(k)}, \tag{27}$$

where $(\mathbf{p}_a\cdot\mathbf{p}_b)^{(k)}$ is a scalar product recorded in the kth in-coincidence measurement, the product $\prod_{(a,b)\in\mathbf{j}}$ runs over all ion pairs within label $\mathbf{j}$, the sum over $\mathbf{j}$ runs over all labels that are indistinguishable from $\mathbf{i}$, the sum k runs over all measurements, and $Y_\mathbf{i}$ is the total number of contributions to this moment, from all measurements.

For example, considering N_ICC shots where the ions (I^+, C^+, C^+) are detected, one of the PEMSP is

$$\begin{aligned}&M_{\{(I,C'),(C,C')\}}=\\&\frac{1}{2N_\mathrm{ICC}}\sum_k(\mathbf{p}_I\cdot\mathbf{p}_{C'})^{(k)}(\mathbf{p}_\mathrm{C}\cdot\mathbf{p}_{C'})^{(k)}+(\mathbf{p}_I\cdot\mathbf{p}_C)^{(k)}(\mathbf{p}_\mathrm{C}\cdot\mathbf{p}_{C'})^{(k)}.\end{aligned} \tag{28}$$

Due to the high number of shots from the experiment and the large number of possible ion combinations, the number of products to consider in Eq. (27) is very large when the coincidence order and the

PEMSP order grows. As a result, it is not possible to hold them all in memory simultaneously. Instead, we use OnlineStats.jl[15] to compute the PEMSP and the standard error on their value with minimal memory footprint.

This procedure yields a set of PEMSP that characterize the distribution of ion momenta as measured in the experiment.

5.3 COMPUTING THE PEMSP FROM A GAUSSIAN MODEL

We consider a $3N$ multivariate Gaussian model, defined by the vector of mean values $\boldsymbol{\mu}$ and the covariance matrix $\boldsymbol{\Sigma}$. The probability density of a $3N$-dimensional weighted-momentum vector drawn from this distribution is

$$g(\mathbf{p}) = \frac{1}{Z} \exp\left[-\frac{1}{2}(\mathbf{p}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{p}-\boldsymbol{\mu})\right], \tag{29}$$

with the normalization constant $Z = \sqrt{(2\pi)^{3N}\det\boldsymbol{\Sigma}}$. The joint distribution of scalar products generated from this distribution is

$$g_{\mathrm{SP}}(\mathbf{s}) = \int \mathrm{d}^{3N}p \prod_{a\le b\le N} \delta(s_{ab} - \mathbf{p}_a^T \mathbf{p}_b) g(\mathbf{p}), \tag{30}$$

where $\mathbf{s}$ is an $N(N+1)/2$-dimensional vector containing scalar product values s_{ab} for all combinations of ions a and b. The product in Eq. (30) runs over all pairs of ions. The integration domain is the full $3N$-dimensional momentum space.

By taking the Fourier transform of Eq. (30) and carrying out the integration, the characteristic function becomes

$$\begin{aligned}\tilde{g}_{SP}(\mathbf{k}) = &\left[\det\boldsymbol{\Sigma}\,\mathbf{M}(\mathbf{k})\right]^{-\frac{1}{2}} \exp\left[\frac{i}{2}\boldsymbol{\mu}^T\mathbf{A}(\mathbf{k})\,\boldsymbol{\mu}\right] \\ &\times \exp\left[-\frac{1}{2}\boldsymbol{\mu}^T\mathbf{A}(\mathbf{k})\,\mathbf{M}^{-1}(\mathbf{k})\,\mathbf{A}^T(\mathbf{k})\,\boldsymbol{\mu}\right],\end{aligned} \tag{31}$$

where $\mathbf{k}$ is an $N(N+1)/2$-dimensional vector in Fourier space with elements k_{ab}. We also introduced the following two matrices

$$\mathbf{M}(\mathbf{k}) = \boldsymbol{\Sigma}^{-1} - i\mathbf{A}(\mathbf{k}), \tag{32}$$

and

$$\mathbf{A}(\mathbf{k}) = \begin{pmatrix} 2k_{11}\mathbb{1} & k_{12}\mathbb{1} & \cdots & k_{1N}\mathbb{1} \\ k_{12}\mathbb{1} & 2k_{22}\mathbb{1} & \cdots & k_{2N}\mathbb{1} \\ \vdots & \vdots & \ddots & \vdots \\ k_{1N}\mathbb{1} & k_{2N}\mathbb{1} & \cdots & 2k_{NN}\mathbb{1} \end{pmatrix}, \tag{33}$$

where $\mathbb{1}$ is the 3×3 identity matrix. The matrix $\mathbf{A}(\mathbf{k})$ allows expressing sums of scalar products between 3-dimensional weighted momenta in a matrix form, as

$$\frac{1}{2}\mathbf{p}^T\mathbf{A}(\mathbf{k})\mathbf{p} = \sum_{a\le b\le N} k_{ab}\mathbf{p}_a^T\mathbf{p}_b. \tag{34}$$

Using Eq. (31), any statistical moment of the $N(N+1)/2$-dimensional distribution of scalar products is retrieved by differentiating the characteristic function,

$$\begin{aligned} m_{\{(a_1,b_1),(a_2,b_2),\dots\}} &= \left\langle \left(\mathbf{p}_{a_1}\cdot\mathbf{p}_{b_1}\right)\left(\mathbf{p}_{a_2}\cdot\mathbf{p}_{b_2}\right)\cdots \right\rangle \\ &= \frac{(-i)\,\partial}{\partial k_{a_1,b_1}}\frac{(-i)\,\partial}{\partial k_{a_2,b_2}}\cdots \tilde{g}_{\mathrm{SP}}(\mathbf{k})\Big|_{\mathbf{k}=\mathbf{0}}. \end{aligned} \tag{35}$$

To take into account the unknown identity of the ions, we average these moments over all equivalent labels. This yields the per-element moments of the scalar products (PEMSP),

$$\widehat{M}_{\mathbf{i}} = \langle m_{\mathbf{j}} \rangle_{\mathbf{j}\sim\mathbf{i}}, \tag{36}$$

where $\langle\cdot\rangle_{\mathbf{j}\sim\mathbf{i}}$ means averaging over equivalent labels. In the following, we consider the PEMSP as explicit functions of the parameters of the Gaussian weighted-momentum distribution, i. e., $\widehat{M}_{\mathbf{i}}(\boldsymbol{\mu},\boldsymbol{\Sigma})$.

5.3.1 *Equivalent moments*

To compute the PEMSP according to Eq. (36), we need to identify all equivalent labels. Two labels are equivalent either because the associated moments are equal or because they are indistinguishable.

For example, the following products are equal

$$(\mathbf{p}_{\mathrm{I}}\cdot\mathbf{p}_{\mathrm{C2}})\,(\mathbf{p}_{\mathrm{C3}}\cdot\mathbf{p}_{\mathrm{C2}}) = (\mathbf{p}_{\mathrm{C2}}\cdot\mathbf{p}_{\mathrm{C3}})\,(\mathbf{p}_{\mathrm{C2}}\cdot\mathbf{p}_{\mathrm{I}})\,. \tag{37}$$

Thus, the associated moments are also equal, and their label are equivalent, i. e.,

$$\{(\mathrm{I},\mathrm{C2}),(\mathrm{C3},\mathrm{C2})\} \sim \{(\mathrm{C2},\mathrm{C3}),(\mathrm{C2},\mathrm{I})\}. \tag{38}$$

We can identify equivalent labels corresponding to equal products by ordering the scalar product and the atom in lexicographic order. For example, once ordered, the two previous labels both become $\{(\mathrm{C2},\mathrm{C3}),(\mathrm{C2},\mathrm{I})\}$.

Then, we also need to take into account the fact that atoms of the same elements are indistinguishable. Using prime as before to denote different atoms in a single coincident measurement, the label $\{(\mathrm{C2},\mathrm{C3}),(\mathrm{C2},\mathrm{I})\}$ can be written in two different forms with undistinguishable atoms,

$$\{(\mathrm{C},\mathrm{C}'),(\mathrm{C},\mathrm{I})\} \quad \text{or} \quad \{(\mathrm{C},\mathrm{C}'),(\mathrm{C}',\mathrm{I})\}. \tag{39}$$

To make it unique, we consider all possibilities and define the first one, in lexicographic order, to be the *canonical label*. In the example, the canonical label of $\{(\mathrm{C2},\mathrm{C3}),(\mathrm{C2},\mathrm{I})\}$ and $\{(\mathrm{C2},\mathrm{C3}),(\mathrm{C3},\mathrm{I})\}$ is in both case $\{(\mathrm{C},\mathrm{C}'),(\mathrm{C},\mathrm{I})\}$. If two labels have the same canonical form, like in this example, they are equivalent.

5.3.2 *Explicit formula for the moments*

Equations (35) and (36) fully determine how to compute the PEMSP from a Gaussian model. However, to implement them numerically, we derive an explicit formula.

Rather than computing the moments directly from Eq. (35), it is, in practice, easier to first compute the *cumulants* defined as

$$C_{\mathbf{i}} = (-i)^n \, \partial_{i_1} \cdots \partial_{i_n} \ln \left[\tilde{g}(\mathbf{k})\right]_{\mathbf{k}=0}, \tag{40}$$

where $\partial_{i_\ell} = \partial/\partial k_{i_\ell}$, and we are writing the elements of $\mathbf{i}$ as single indices instead of pairs for convenience, i. e.,

$$\mathbf{i} = (i_1, i_2, \ldots, i_n) = ((a_1, b_1), (a_2, b_2), \ldots, (a_n, b_n)), \tag{41}$$

with $i_\ell = (a_\ell, b_\ell)$.

From Eq. (31),

$$\begin{aligned} \ln \tilde{g}_{SP}(\mathbf{k}) = &-\frac{1}{2} \ln \det \left[\boldsymbol{\Sigma}\mathbf{M}(\mathbf{k})\right] \\ &+ \frac{i}{2} \boldsymbol{\mu}^T \mathbf{A}(\mathbf{k}) \boldsymbol{\mu} - \frac{1}{2} \boldsymbol{\mu}^T \mathbf{A}(\mathbf{k}) \, \mathbf{M}^{-1}(\mathbf{k}) \, \mathbf{A}^T(\mathbf{k}) \, \boldsymbol{\mu}. \end{aligned} \tag{42}$$

The derivatives in Eq. (40) are computed through the Taylor expansion of $\ln \tilde{g}_{SP}(\mathbf{k})$. The first term in Eq. (42) expands to

$$\begin{aligned} -\ln \det \left[\boldsymbol{\Sigma}\mathbf{M}(\mathbf{k})\right] &= -\mathrm{tr} \ln \left[\boldsymbol{\Sigma} \left(\boldsymbol{\Sigma}^{-1} - i\mathbf{A}(\mathbf{k})\right)\right] \\ &= -\mathrm{tr} \ln \left[\mathbb{1} - i\boldsymbol{\Sigma}\mathbf{A}(\mathbf{k})\right] \\ &= \sum_{j=1}^{\infty} \frac{i^j}{j} \mathrm{tr} \left[\boldsymbol{\Sigma}\mathbf{A}(\mathbf{k})\right]^j \end{aligned} \tag{43}$$

where we used the identity

$$\ln \det \mathbf{X} = \mathrm{tr} \ln \mathbf{X}, \tag{44}$$

and the Taylor expansion of the logarithm around $\mathbf{k} = \mathbf{0}$.

The formal solution for the inverse of the matrix $\mathbf{M}(\mathbf{k})$,

$$\mathbf{M}^{-1}(\mathbf{k}) = \sum_{j=0}^{\infty} i^j \left[\boldsymbol{\Sigma}\mathbf{A}(\mathbf{k})\right]^j \boldsymbol{\Sigma}, \tag{45}$$

leads to

$$\begin{aligned} & i\boldsymbol{\mu}^T\mathbf{A}(\mathbf{k})\boldsymbol{\mu} - \boldsymbol{\mu}^T\mathbf{A}(\mathbf{k})\mathbf{M}^{-1}(\mathbf{k})\mathbf{A}^T(\mathbf{k})\boldsymbol{\mu} \\ &\quad = i\boldsymbol{\mu}^T\mathbf{A}(\mathbf{k})\boldsymbol{\mu} - \boldsymbol{\mu}^T\mathbf{A}(\mathbf{k})\left[\sum_{j=0}^{\infty} i^j \left[\boldsymbol{\Sigma}\mathbf{A}(\mathbf{k})\right]^j \boldsymbol{\Sigma}\right]\mathbf{A}(\mathbf{k})\boldsymbol{\mu} \\ &\quad = i\boldsymbol{\mu}^T\mathbf{A}(\mathbf{k})\boldsymbol{\mu} + i\sum_{j=0}^{\infty} i^{j+1}\boldsymbol{\mu}^T\mathbf{A}(\mathbf{k})\left[\boldsymbol{\Sigma}\mathbf{A}(\mathbf{k})\right]^{j+1}\boldsymbol{\mu} \\ &\quad = i\sum_{j=0}^{\infty} i^j\boldsymbol{\mu}^T\mathbf{A}(\mathbf{k})\left[\boldsymbol{\Sigma}\mathbf{A}(\mathbf{k})\right]^j\boldsymbol{\mu} \\ &\quad = \sum_{j=0}^{\infty} i^{j+1}\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\left[\boldsymbol{\Sigma}\mathbf{A}(\mathbf{k})\right]^{j+1}\boldsymbol{\mu} \\ &\quad = \sum_{j=1}^{\infty} i^j\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\left[\boldsymbol{\Sigma}\mathbf{A}(\mathbf{k})\right]^j\boldsymbol{\mu}. \end{aligned} \tag{46}$$

Together, the two Taylor expansions yield

$$\ln \tilde{g}_{sm}(\mathbf{k}) = \frac{1}{2}\sum_{j=1}^{\infty} i^j \left(\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\left[\boldsymbol{\Sigma}\mathbf{A}(\mathbf{k})\right]^j\boldsymbol{\mu} + \frac{1}{j}\mathrm{tr}\left[\boldsymbol{\Sigma}\mathbf{A}(\mathbf{k})\right]^j\right). \tag{47}$$

From Eqs. (40) and (47), the cumulants are

$$\begin{aligned} C_{\mathbf{i}} = \frac{1}{2}\Big(&\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\left[\partial_{i_1}\cdots\partial_{i_n}\left[\boldsymbol{\Sigma}\mathbf{A}(\mathbf{k})\right]^n\right]_{\mathbf{k}=0}\boldsymbol{\mu} \\ &+ \frac{1}{n}\mathrm{tr}\left[\partial_{i_1}\cdots\partial_{i_n}\left[\boldsymbol{\Sigma}\mathbf{A}(\mathbf{k})\right]^n\right]_{\mathbf{k}=0}\Big) \end{aligned} \tag{48}$$

The sum over j disappears, as only terms of order n in $\mathbf{k}$ are nonzero after computing the derivatives and setting $\mathbf{k}$ to zero. The remaining terms become

$$\partial_{i_1}\cdots\partial_{i_n}\left[\boldsymbol{\Sigma}\mathbf{A}(\mathbf{k})\right]^n = \sum_{\sigma}\prod_{\ell=1}^{n}\boldsymbol{\Sigma}\left[\partial_{i_{\sigma(\ell)}}\mathbf{A}(\mathbf{k})\right], \tag{49}$$

where the sum runs over all permutation σ of length n. The constant matrices $\mathbf{U}_{i_\ell}$ is defined as the derivatives of the $\mathbf{A}(\mathbf{k})$ matrices, namely

$$\mathbf{U}_{i_\ell} = \partial_{i_\ell}\mathbf{A}(\mathbf{k}) = \begin{pmatrix} 0 & \cdots & 0 & \cdots & 0 & \cdots & 0 \\ \vdots & & \vdots & & \vdots & & \vdots \\ 0 & \cdots & 0 & \cdots & \mathbb{1} & \cdots & 0 \\ \vdots & & \vdots & & \vdots & & \vdots \\ 0 & \cdots & \mathbb{1} & \cdots & 0 & \cdots & 0 \\ \vdots & & \vdots & & \vdots & & \vdots \\ 0 & \cdots & 0 & \cdots & 0 & \cdots & 0 \end{pmatrix}. \tag{50}$$

The two nonzero blocks are the ones using variable $k_{a_\ell b_\ell}$ in Eq. (33), i. e., the (a_ℓ, b_ℓ) and the (b_ℓ, a_ℓ) blocks.

With the notation $\sigma(\mathbf{i}) = (i_{\sigma(1)}, \ldots, i_{\sigma(n)})$ for the permutation of the label $\mathbf{i}$, and the definition of the $\mathbf{Q}(\mathbf{i})$ matrix as

$$\mathbf{Q}(\mathbf{i}) = \mathbf{U}_{i_1} \prod_{\ell=2}^{n} \mathbf{\Sigma} \mathbf{U}_{i_\ell}, \tag{51}$$

the cumulants simplify to

$$C_{\mathbf{i}} = \frac{1}{2} \sum_{\sigma} \left(\boldsymbol{\mu}^T \mathbf{Q}(\sigma(\mathbf{i})) \boldsymbol{\mu} + \frac{1}{n} \mathrm{tr} \left[\mathbf{\Sigma} \mathbf{Q}(\sigma(\mathbf{i})) \right] \right). \tag{52}$$

The moments are related to the cumulants through [18]

$$m_{\mathbf{i}} = \sum_{\nu=1}^{n} \sum_{\zeta \in \{\mathcal{P}_\nu(n)\}} \prod_{\mathbf{k} \in \zeta} C_{\mathbf{i}_\mathbf{k}}, \tag{53}$$

where $\{\mathcal{P}_\nu(n)\}$ is the set of all non equivalent partitions of length ν of the integer n, $\mathbf{i}$ is a multi-index and $\mathbf{i}_\mathbf{k} = (i_{k_1}, \ldots, i_{k_r})$ is a short notation for subindexing $\mathbf{i}$ with a partition $\mathbf{k}$.

Equations (52) and (53) provide an explicit expression to compute the statistical moments $m_{\mathbf{i}}$ from momenta following a Gaussian distribution.

5.4 FITTING

In order to fit the PEMSP $\widehat{M}_{\mathbf{i}}(\boldsymbol{\mu}, \mathbf{\Sigma})$ of the model Gaussian distribution to the experimental PEMSP $M_{\mathbf{i}}$, we define a loss function. First, we define the contribution of an individual PEMSP to the loss as

$$D_{\mathbf{i}}(\boldsymbol{\mu}, \mathbf{\Sigma}) = \left(\frac{\left(\widehat{M}_{\mathbf{i}}(\boldsymbol{\mu}, \mathbf{\Sigma}) - M_{\mathbf{i}} \right)^2 + \delta M_{\mathbf{i}}^2}{\delta M_{\mathbf{i}}^2} \right)^{\frac{1}{n_{\mathbf{i}}}} - 1, \tag{54}$$

where $n_{\mathbf{i}}$ is the order of moment $M_{\mathbf{i}}$ and $\delta M_{\mathbf{i}}$ is the standard error on the experimental PEMSP. Normalizing by the squared standard error reduces the contribution of the moments whose value is not well known due to low statistics. Taking the $n_{\mathbf{i}}$-th root ensures that all contributions have the same unit and similar magnitudes, even with different moment orders. The shift by -1 enforces the convention that perfect agreement ($\widehat{M}_{\mathbf{i}} = M_{\mathbf{i}}$) has a loss of exactly zero.

When combining the loss contribution of the individual moments, we reduce weights on moments that are already very well approximated. We consider individual moments computed from the model to be well approximated when the absolute difference to the experimental moment is smaller than twice its standard error. Taken together, we employ the total loss function

$$L(\boldsymbol{\mu}, \mathbf{\Sigma}) = \left[\frac{1}{N_{\mathbf{i}}} \sum_{\mathbf{i}} \alpha \left(\widehat{M}_{\mathbf{i}}(\boldsymbol{\mu}, \mathbf{\Sigma}), M_{\mathbf{i}} \right) \; D_{\mathbf{i}}(\boldsymbol{\mu}, \mathbf{\Sigma}) \right]^{\frac{1}{2}} \tag{55}$$

where $N_{\mathbf{i}}$ is the number of moments considered, and $\alpha(\widehat{M}_{\mathbf{i}}, M_{\mathbf{i}})$ is the dampening term

$$\alpha(\widehat{M}_{\mathbf{i}}, M_{\mathbf{i}}) = \begin{cases} 0.01 & \text{if } |\widehat{M}_{\mathbf{i}} - M_{\mathbf{i}}| < 2\delta M_{\mathbf{i}}, \\ 1 & \text{otherwise.} \end{cases} \tag{56}$$

which reduces the effect of the moments that are close to the experimental value compared to the statistical error.

We minimize this loss function using the limited-memory Broyden–Fletcher–Goldfarb–Shanno (L-BFGS) algorithm implemented in the Optim.jl package [64]. This requires the gradient of the loss function with respect to all of the parameters, namely the mean $\boldsymbol{\mu}$ and the covariance matrix $\boldsymbol{\Sigma}$ of the Gaussian distribution. Using the $\boldsymbol{\Sigma}$ matrix directly causes problems, when the matrix becomes nonsymmetric or not positive definite. Therefore, we employ the Cholesky decomposition of the covariance matrix, $\boldsymbol{\Sigma} = \mathbf{L}\mathbf{L}^T$ and use the lower triangular matrix $\mathbf{L}$ as parameter, ensuring that $\boldsymbol{\Sigma}$ is always symmetric and positive definite. The gradient is computed using Zygote.jl [24], and implementing custom differentiation rules for most of the internal functions with ChainRules.jl [88] (see Section 5.5 for details).

The number of moments grows extremely rapidly with moment order, making the usage of high moment order unpractical and computationally very expensive. Therefore, for the results shown in the main text, we restrict ourselves to a maximal moment order of 3. Because this involves maximally three mutual scalar products, combined information of up to 6 ions measured in coincidence is considered in the fit.

5.5 EFFICIENT COMPUTATION OF THE GRADIENT

The L-BFGS algorithm requires the gradient of the loss function. When possible, the derivatives are computed using automatic differentiation through the Zygote.jl [24] package. We combined it with optimized implementations of the derivatives when necessary. Notably, automatic differentiation is incompatible with the optimizations necessary for the fast computation of the cumulants. Therefore, we implement custom explicit differentiation rules for them, as described below.

Differentiating Eq. (52) yields

$$\mathrm{d}C_{\mathbf{i}} = \frac{1}{2}\sum_{\sigma}\left(2\boldsymbol{\mu}^T\mathbf{Q}\mathrm{d}\boldsymbol{\mu} + \boldsymbol{\mu}^T\mathrm{d}\mathbf{Q}\boldsymbol{\mu} + \frac{1}{n}\mathrm{tr}\left[\mathrm{d}\boldsymbol{\Sigma}\mathbf{Q}\right] + \frac{1}{n}\mathrm{tr}\left[\boldsymbol{\Sigma}\mathrm{d}\mathbf{Q}\right]\right). \tag{57}$$

The differentials $\mathrm{d}\boldsymbol{\mu}$ and $\mathrm{d}\boldsymbol{\Sigma}$ are straightforward to compute. From Eq. (51), the $\mathrm{d}\mathbf{Q}$ term is

$$\mathrm{d}\mathbf{Q}(\mathbf{i}) = \sum_{d=2}^{n}\mathbf{U}_{i_1}\prod_{\ell=2}^{n}\mathbf{D}_{d\ell}\mathbf{U}_{i_\ell} \tag{58}$$

where

$$\mathbf{D}_{d\ell} = \begin{cases} \mathrm{d}\mathbf{\Sigma} & \text{if} \quad d = \ell \\ \mathbf{\Sigma} & \text{otherwise.} \end{cases} \tag{59}$$

5.5.1 *Recursive computation of the derivatives*

Computing all $\mathrm{dC}_{\mathbf{i}}$ of order n requires many matrix multiplications. However, a lot of them stem from repeated computations and can be avoided. To see it, we define

$$\mathbf{Q}_{.d}(\mathbf{i}) = \mathbf{U}_{i_1} \prod_{\ell=2}^{n} \mathbf{D}_{d\ell} \mathbf{U}_{i_\ell}, \tag{60}$$

With this definition, Eq. (58) reads

$$\mathrm{d}\mathbf{Q}(\mathbf{i}) = \sum_d \mathbf{Q}_{.d}(\mathbf{i}). \tag{61}$$

We use $d = 0$ for undifferentiated terms as

$$\mathbf{Q}(\mathbf{i}) = \mathbf{Q}_{.0}(\mathbf{i}). \tag{62}$$

To perform the computation of $\mathbf{Q}_{.d}(\mathbf{i})$ recursively, we represent it as a set of trees. The root node of each tree is $\mathbf{Q}_{.0}((i_1)) = \mathbf{U}_{i_1}$, where i_1 can be any index. Then, the children of a node $\mathbf{Q}_{.0}(\mathbf{i})$, with $\mathbf{i}$ of length n, is

$$\begin{aligned} &\mathbf{Q}_{.0}((\mathbf{i}, i_{n+1})) = \mathbf{Q}_{.0}(\mathbf{i}) \mathbf{\Sigma} \mathbf{U}_{i_{n+1}} \quad \text{and} \\ &\mathbf{Q}_{.(n+1)}((\mathbf{i}, i_{n+1})) = \mathbf{Q}_{.0}(\mathbf{i}) \mathrm{d}\mathbf{\Sigma} \mathbf{U}_{i_{n+1}}, \end{aligned} \tag{63}$$

for each possible index i_{n+1}. When $d \neq 0$, the children of a node $\mathbf{Q}_{.d}(\mathbf{i})$ are

$$\mathbf{Q}_{.d}((\mathbf{i}, i_{n+1})) = \mathbf{Q}_{.d}(\mathbf{i}) \mathbf{\Sigma} \mathbf{U}_{i_{n+1}}, \tag{64}$$

again for each possible index i_{n+1}.

In Fig. 19, we show part of a tree as an example (children are shown only for the purple nodes to avoid clutter). This tree is relatively small, as we only consider two atoms, with index 1 and 2. The number of children of a node is not constant, 6 for nodes that do not contain the differential $\mathrm{d}\mathbf{\Sigma}$, and 3 otherwise. The matrix $\mathbf{U}_{21}$ does not appear because it is equal to the $\mathbf{U}_{12}$, the corresponding terms are therefore already computed.

In practice, we precompute the structure of the tree and then compute all elements using Eqs. (63) and (64). When we traverse the tree depth-first, it is only necessary to store the nodes that are ancestors of the current node being computed, making it advantageous to limit the amount of memory required, which is crucial when parallelizing the code. The parallelization itself is straightforward: since each branch of a tree independent of the others, branches are distributed to be computed on different cores.

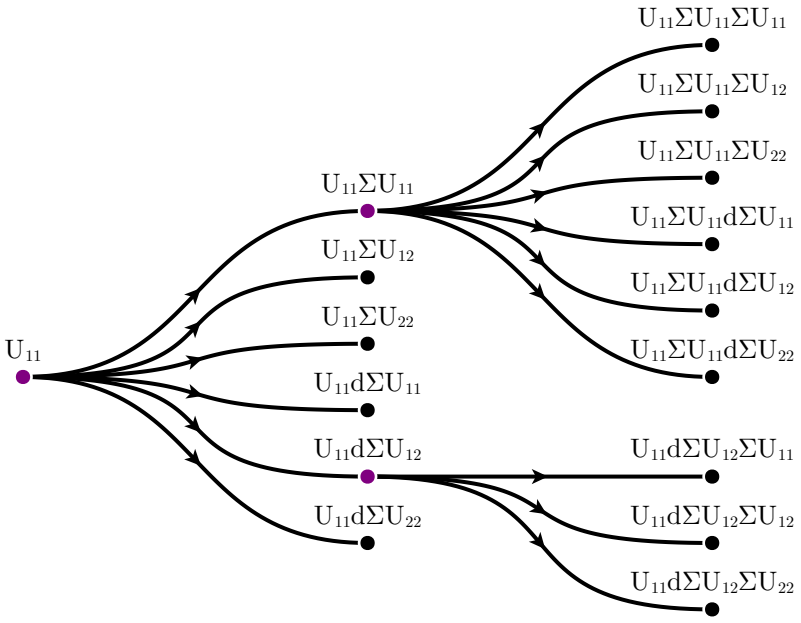

Figure 19: **Example of a tree for the recursive computation of the derivatives of the cumulants.** The figure shows part of a tree for a system composed of two atoms (1 and 2). We show all children of the purple nodes. The children of the other nodes are skipped for brevity, as the number of nodes grows quickly with the depth of the tree.

5.5.2 *Custom matrix operations*

According to their definition in Eq. (50), the **U** matrices are extremely sparse. Similarly, the products in Eqs. (63) and (64) always involve sparse matrices, which have at most 6 nonzero columns. We only store the nonzero blocks or columns of those two kind of matrices, and define a fast sparse matrix multiplication using Tullio.jl [1].

This package is also used to speed up the computation of the traces and bilinear products in Eq. (57).

5.5.3 *Benchmark*

To see the gain from the optimizations presented in Sections 5.5.1 and 5.5.2, we compute the PEMSP with two different implementations. The naive implementation, that does not utilize any optimizations, implements Eqs. (52) and (57) directly, while the optimized implementation uses all the methods described in the sections to speed up the code and limit the memory usage. In Fig. 20, we report the time to compute a set of cumulant with a maximum cumulant order of 3 and various numbers of ions measured in coincidence. In this benchmark, we use a fully random distribution of scalar products for the iodopyridine molecule. The tests are run on a single thread, so the further improvements thanks to the parallelization of the code are not shown.

Already, we see that the optimized version is in all cases about 16 times faster than the naive implementation, with a computation cost growing fast with the coincidence order. This shows that the optimization is necessary to compute the PEMSP in a reasonable time, especially considering that the fit needs to compute them a few thousand times before converging.

5.6 TEST OF THE ALGORITHM ON SIMULATED DATA

5.6.1 *Simulations*

To assess the quality of the fitting procedure, we first test it on the simulated data. We use the simulation described in Section 4.5 (Simulation set IV in Table 2). To describe it briefly, those simulations use XMDYN with the inclusion of molecular bonds through the use of ReaxFF. From the ensemble of simulated trajectories, we only consider those that end up with the charge state I^{+4}/N^{+2}, while the other atoms can have any charge. This coincidence channel contains 1 738 trajectories (out of initially 61 652 trajectories).

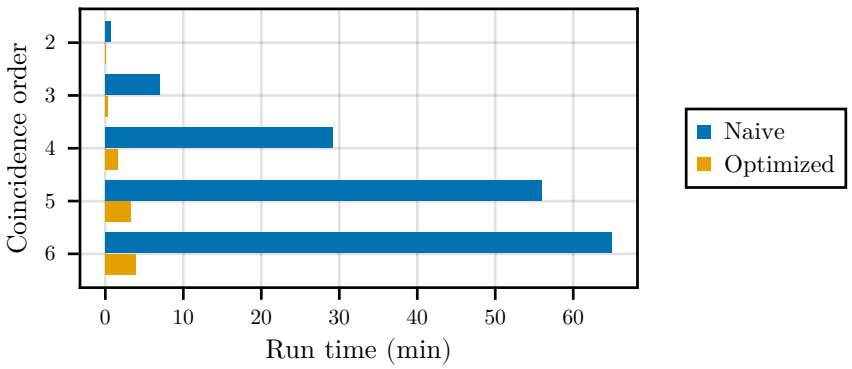

Figure 20: **Computation time for a set of cumulants.** Time required to compute all cumulants for a maximum cumulant order of 3 and the given coincidence order, for the iodopyridine molecule using a single thread, for the optimized version and a naive implementation.

5.6.2 *Initial guess*

To determine an initial guess for the algorithm, we compute the mean values and the covariance matrix of the data. The mean values are used directly as the guess for $\boldsymbol{\mu}$. The off-diagonal entries of the covariance matrix are set to zero and the remaining diagonal matrix is used as the guess for $\boldsymbol{\Sigma}$. By removing the covariance terms, we ensure that the guess does not carry prior information about the correlations between the atoms.

5.6.3 *Comparison with PCA on full data*

First, Figs. 21A and 21B show the dominant in-plane and out-of-plane PCs, respectively, computed from the full simulated data set. This is the reference for the quality of the reconstruction by the algorithm.

Then, we apply the algorithm. We artificially introduce limitations in the simulation data set, similar to the ones known to affect the experimental data: we consider all possible 6-ion coincidence subsets of the data, randomly orient each trajectory and remove the identity of ions of the same elements[1]. We perform the fit by computing PEMSP up to order 3.

The dominant principal components reconstructed from these limited data are shown in Figs. 21C and 21D. The three dominant in-plane and the three dominant out-of-plane components explain more than 80% of the variance in the data (see Fig. 21E). The fit reproduces the original distribution very well. We quantify the similarity between the principal components retrieved from both analyses

1 In practice, computing the PEMSP already remove this information, so no further processing is needed.

by computing their overlap S, defined as the absolute value of their scalar product. Since the principal components are normed, when the agreement is perfect, we have $S = 1.0$ and the matrix of all overlaps is diagonal. Panels F and G of Fig. 21 show S for in-plane and out-of-plane components respectively. The quality of the reconstruction is remarkable considering that there are only 1 738 trajectories in the I^{+4}/N^{+2} coincidence channel.

We see that the reconstruction retrieves the out-of-plane components better than the in-plane ones. We attribute this to the higher complexity of the dynamic in the molecular plane, resulting in a more complex distribution of momenta that suffers more from the limited amount of data.

To confirm the effect of the low number of trajectories, we go to higher statistics by bootstrapping the data. We fit a Gaussian to the full simulation data and draw 10^5 new samples. Then we apply the methodology described above and report the result in Fig. 21H-L. Panels H and I show the reconstruction of the leading principal components. They are nearly perfectly reconstructed. There are some small discrepancies in the reconstructed explained variance for each of them, as seen in Fig. 21J. The high quality of the reconstruction does not extend to components with lower explained variance, as shown in Figs. 21K and 21L. However, there is a clear improvement compared to the reconstruction with lower statistics, especially for the in-plane components.

When the number of samples is sufficient, all the important features, as measured by the explained variance, are correctly reconstructed. This success establishes that it is possible to extract the main characteristics of the full momentum distribution of 11 ions from data including only coincidence information on a much smaller number of ions, in this case up to 6.

5.6.4 *Convergence*

To investigate the convergence of the algorithm, we show in Fig. 22A how the loss function varies during the optimization. The value of the loss function converges after around one thousand iterations.

Since we are applying the algorithm to simulation, we can use the full data as a reference distribution. We estimate the dissimilarity compared to the reference using the Kullback-Leibler divergence (KL divergence) D_{KL}, defined as

$$D_{\mathrm{KL}}(g \| g_{\mathrm{ref}}) = \int \mathrm{d}\mathbf{p}\; g(\mathbf{p}) \ln\left(\frac{g(\mathbf{p})}{g_{\mathrm{ref}}(\mathbf{p})}\right) \tag{65}$$

for a momentum probability distribution $g(\mathbf{p})$ and a reference distribution $g_{\mathrm{ref}}(\mathbf{p})$. The KL divergence is 0 when the probability distribution $g(\mathbf{p})$ matches the reference perfectly; otherwise, it gives a measure of the information difference between the two distributions.

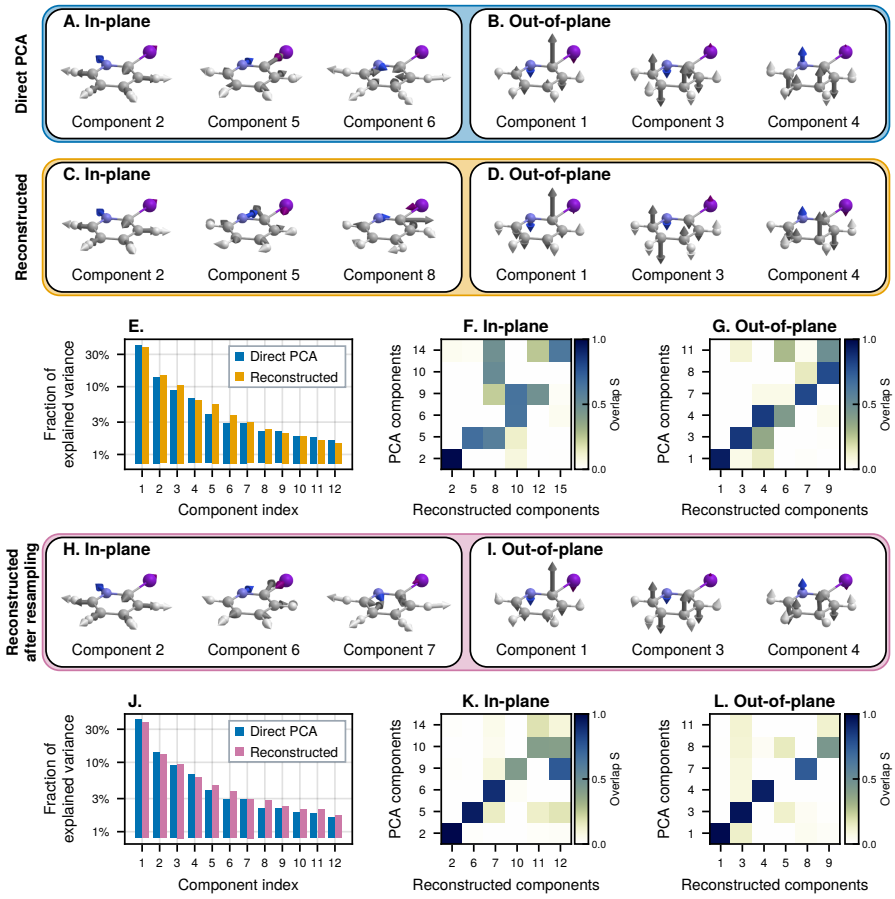

Figure 21: **Comparison of the fits using the full and the limited simulated data sets.** A-B. Principal components of the full momentum distribution from the simulations, computed directly from the complete data, where the laboratory-frame rotation and the ion identity are known. C-D. Principal components of the full momentum distribution, reconstructed using our algorithm, with information similar to that available in the experiment, i.e., only information on 6 ions is covered, the original ion location inside the molecule is unknown, and the molecule has a random and unknown rotation. E. Fraction of variance explained by each principal component for both procedures. F-G. Similarity of the principal components obtained with the two methods, computed as the scalar product between the components. H-I. Principal components of the full momentum distribution, based on 100 000 samples drawn from the best Gaussian fit to the full data. This reconstruction uses our algorithm, with information similar to that available in the experiment, i.e., only information on 6 ions is covered, the original ion location inside the molecule is unknown, and the molecule has a random and unknown rotation. J-L. Comparison between the components obtained from the reconstruction algorithm after resampling and the components computed directly from the complete data (shown in A-B). J. Fraction of variance explained by each principal component. K-L. Similarity of the principal components obtained with the two methods.

In Fig. 22B, we show the KL divergence between the fitted model and the reference. As expected, the dissimilarity steadily decreases as the loss function gets smaller.

We also use the KL divergence to explore the effect of using different values for the maximum coincidence order and maximum moment orders. In Fig. 22C, we show how the quality of the fitted model changes when we limit the coincidence order. The plot shows that a higher coincidence order generally corresponds to a better or equally good fit. We note, however, that we do not observe any significant improvement of the fit beyond coincidence order 3, indicating that this coincidence order already contains enough information to fully characterize the distribution. We speculate that when a more general distribution (i.e., a non-Gaussian distribution) is fitted, the fit procedure might benefit from a higher coincidence order.

In Fig. 22D we show the effect of changing the maximum moment order. The effect is relatively small. Going to even higher moment order may be beneficial, but increasing its value beyond 3 is prohibitively computationally expensive. Therefore, we do not extend our tests further.

In conclusion, we confirm that our algorithm converges to the desired distribution that best characterizes the input data with a Gaussian distribution, if the coincidence and maximum moment order are sufficiently high (coincidence order ≥ 4, moment order $= 3$).

5.7 CONCLUSION

In this chapter, we develop an algorithm to perform PCA on the momenta of fragments detected in a CEI experiment. We show that we can fit a multivariate Gaussian to the data in a way that overcomes the challenges faced by the experimental data. With this procedure, we obtain a covariance matrix that is decomposed into the principal components of the distribution.

By testing the algorithm on simulation data, we demonstrate that it converges satisfactorily. Based on these results, we conclude that the algorithm can be meaningfully applied to experimental data to obtain a view of the full dimensional distribution of the asymptotic momenta from measurement with partial coincidence.

We perform such analysis in the next chapter.

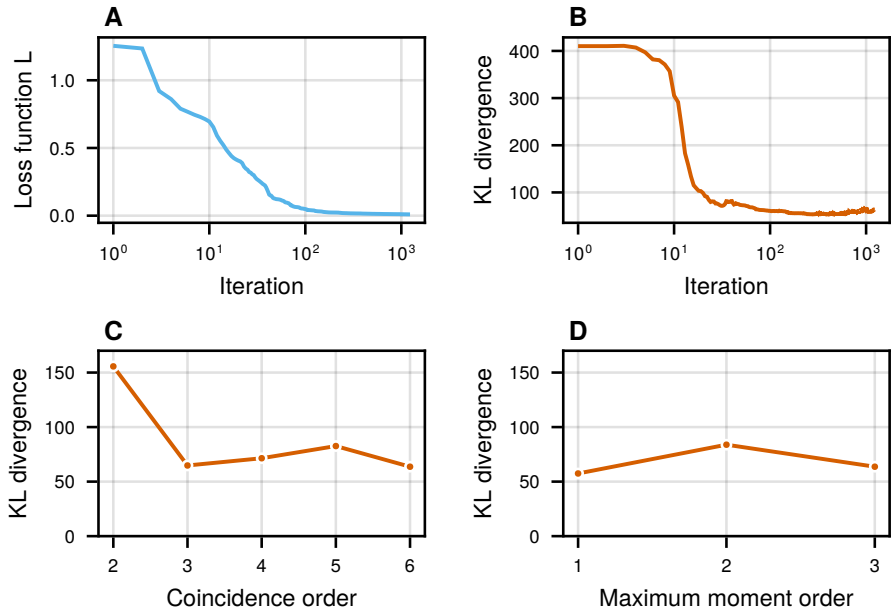

Figure 22: **Convergence of the reconstruction algorithm.** A. Loss function during the fitting. B. KL divergence from the ground truth during the fitting of the simulation data. C. KL divergence of the Gaussian model fitted to the simulation data, for various coincidence orders and moment order 3. D. KL divergence of the Gaussian model fitted to the simulation data, for various moment orders and coincidence order 6.

6 APPLICATION OF THE CORRELATION ANALYSIS

This chapter adapts a manuscript in preparation:

Benoît Richard, Rebecca Boll, Sourav Banerjee, Julia M. Schäfer, Zoltan Jurek, Gregor Kastirke, Kilian Fehre, Markus S. Schöffler, Nils Anders, Thomas M. Baumann, Sebastian Eckart, Benjamin Erk, Alberto De Fanis, Reinhard Dörner, Sven Grundmann, Patrik Grychtol, Max Hofmann, Markus Ilchen, Max Kircher, Katharina Kubicek, Maksim Kunitski, Xiang Li, Tommaso Mazza, Severin Meister, Niklas Melzer, Jacobo Montano, Valerija Music, Yevheniy Ovcharenko, Christopher Passow, Andreas Pier, Nils Rennhack, Jonas Rist, Daniel E. Rivas, Daniel Rolles, Ilme Schlichting, Lothar Ph. H. Schmidt, Philipp Schmidt, Daniel Trabert, Florian Trinter, Rene Wagner, Peter Walter, Pawel Ziolkowski, Artem Rudenko, Michael Meyer, Robin Santra, Ludger Inhester, and Till Jahnke

"Imaging correlated quantum fluctuations of the structure of a complex molecule"
In preparation

Contribution to the manuscript I ran the simulations using an updated version of XMDYN, proposed the analysis of the mapping from vibrational modes to principal components, and performed it. I proposed, developed, implemented, optimized and tested the reconstruction algorithm. I cleaned the experimental data, and applied the algorithm to it and to the simulation data, and compared the results. I drafted the manuscript with input from other coauthors.

6.1 EXPERIMENT

In this chapter, we go back to the experiment and apply the algorithm developed in Chapter 5 to real experimental data.

This chapter uses data coming from the same experiment as described in Chapter 2 and reference [9]. The main difference, is that we are using data with the high charge state I^{+4}, N^{+2}, C^{+2} and H^{+}, corresponding to an extremely fast charging up of the molecule during the 10 fs x-ray pulse.

Since we aim to relate the experiment to quantum ground-state effects, we note that the frequency of the lowest excitation of iodopyridine is $\nu = 177\,\mathrm{cm}^{-1}$. This low frequency corresponds to a temperature of $h\nu/k_B = 256\,\mathrm{K}$, which implies that, under normal conditions, there are some thermal excitations in this vibrational mode. However,

the temperature of the molecules in the supersonic gas jet is considerably lower than before the expansion. The temperature of the gas can be estimated by [63]

$$\frac{T}{T_0} = \left(1 + \frac{\gamma - 1}{2} M^2\right)^{-1}, \tag{66}$$

where T/T_0 is the temperature ratio before and after the nozzle, M is the mach number after the nozzle, and $\gamma = c_p/c_V$ is the ratio of heat capacities. Assuming $M = 1$, which is reached at the nozzle exit, $c_p/c_V = 5/3$, and $T_0 = 298\,\mathrm{K}$, an upper estimate for the gas-jet temperature is $T = 220\,\mathrm{K}$. At this temperature, we estimate the relative occupation of the first excited vibrational state to be $n_1 = e^{-h\nu/k_B T}/(1 + e^{-h\nu/k_B T}) \simeq 0.24$ (assuming a microcanonical ensemble of two-level entities). Accordingly, we expect that even for this lowest-frequency mode, the fluctuations are dominated by quantum fluctuations of the vibrational ground state (also known as zero-point vibrations). All the other out-of-plane modes have considerably higher frequencies, confirming that they are mostly not thermally populated for molecules in the gas jet.

6.2 DIRECT ANALYSIS OF THE EXPERIMENTAL DATA

Before any further analysis, a certain amount of background events is filtered out by restricting the data set to cases where all measured ions have a momentum within a sensible range. A measurement is only included if the individual ion momenta are in the range of $|p_H| < 130\,\mathrm{a.u}$ $|p_C| < 500\,\mathrm{a.u.}$, $|p_N| < 650\,\mathrm{a.u.}$ and $|p_I| < 1000\,\mathrm{a.u.}$ In addition, a rather relaxed restriction on the sum of the momenta of the measured fragments of $|p_{\mathrm{sum}}| < 1500\,\mathrm{a.u.}$ is applied to further suppress background and false coincidences. Also, ions with a very similar mass-over-charge ratio may overlap partially in their time-of-flight distributions. We therefore needed to discard the part of the N^{2+} ion momentum distribution with $|p_{lab,z}| > 100\,\mathrm{a.u.}$, which corresponds to a subset of nitrogen ions being emitted in the laboratory frame towards the ion detector with high kinetic energy.

For reference, we plot the geometry of the iodopyridine molecule together with the associated iodine-nitrogen recoil frame in the top left panel of Fig. 23 (see Section 2.3 for more details on the definition of the recoil frame). The Newton diagrams of the asymptotic momenta of the hydrogen and carbon ions in the recoil frame are shown in Fig. 23A-D.

To further study the collective features of the experimental data, it is beneficial to use spherical coordinates, as sketched in Fig. 23E. The angular emission distribution of the carbon ions in this frame is shown in Fig. 23F.

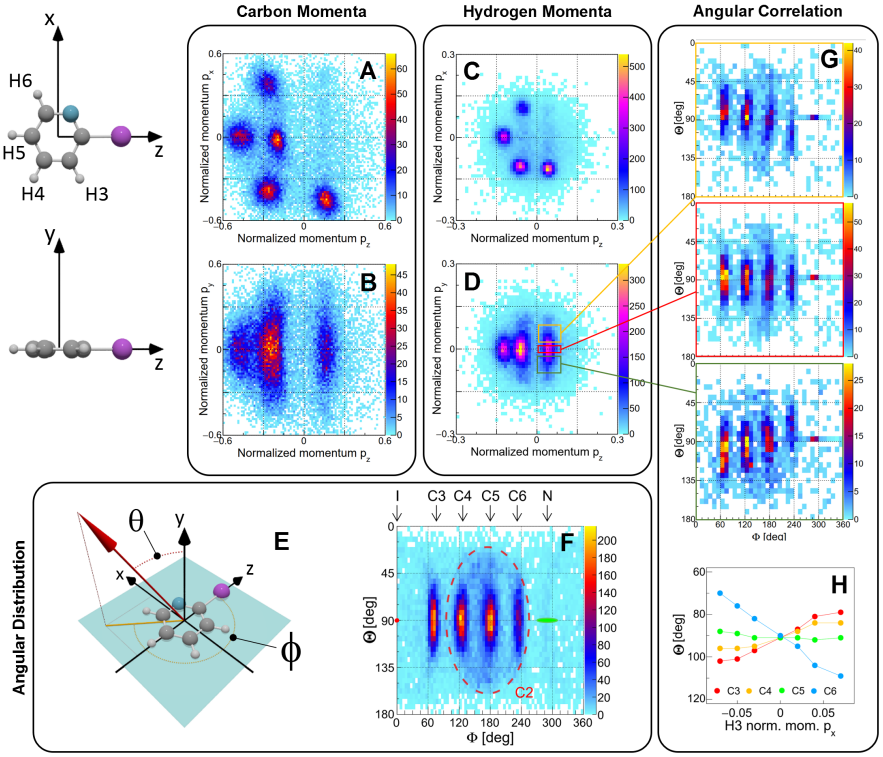

Figure 23: **Experimental data with high charge state.** A. Ground-state average structure of the iodopyridine molecule. B. Definition of the iodine-nitrogen recoil frame. C. Newton diagrams of the asymptotic momenta for C^{+2} (left) and H^{+} (right). *Figure by Till Jahnke, reused with his kind permission.*

The four pronounced peaks correspond to the C3, C4, C5, and C6 carbon ions. The C2 ion has very low momentum (see Fig. 23A) and lies close to the origin of the spherical coordinate system ; thus appearing as a broad feature in Fig. 23F, (marked by the dashed red line). This representation allows visualizing the dependence between the carbon and hydrogen out-of-plane momenta: The panels in Fig. 23G depict the distributions of the angular emission direction of the carbon ions, corresponding to specific emission directions of the H3 hydrogen ion (marked with colored boxes in Fig. 23D). If the H3 ion is emitted within the molecular plane (red box), the carbon ions are on average also emitted mostly in plane. If the H3 ion is either emitted *upwards* (yellow box) or *downwards* (green box) with respect to the molecular plane, the carbon angular distribution becomes tilted, with C3 and C4 being emitted in the same direction as H3, and C6 in the opposite direction. C5 remains centered in plane. The full dependence between the H3 emission direction and the out-of-plane emission angle of C3 to C6 is summarized in Fig. 23H where we depict the mean emission angle Θ of the four carbon ions C3 to C6 in dependence of the out-of-plane momentum of the H3 proton.

6.3 OUT-OF-PLANE VARIANCE

To understand the source of the correlations found in the experimental data, we compare them to theory. We simulate the explosion as described in Section 4.5 (corresponding to Simulation Set III and IV in Table 2). Briefly, we use XMDYN simulations that model the molecular bonds through the use of ReaxFF. In Simulation Set III all trajectories are started from the fixed equilibrium geometry, thus neglecting the vibrational ground-state fluctuation (VGSF) of the molecule. They are included in Simulation Set IV, through the sampling of initial positions and momenta from the ground-state Wigner distribution, according to Section 4.3.

Figure 24 shows the out-of-plane part of the resulting distributions. We look at the I^{+4}/N^{+2} coincidence channel of the experiment and show the asymptotic momenta of the H^{+} and C^{+2} ions in Fig. 24A and Fig. 24B, respectively. Next to those panels, in Fig. 24C and Fig. 24D, we show the corresponding distributions from the simulation that include ground-state fluctuations. The experiment and the simulation agree both qualitatively and quantitatively, except for the diffuse background present in the experimental data. We note that the out-of-plane magnitude of the momenta is large, of the order of 100 a.u. for the hydrogen ions and 400 a.u. for the carbon ions. In both cases, this is about twice the size of each of the in-plane peaks, as seen in Fig. 23C.

We attribute those large variations to the presence of fluctuation in the vibrational ground state of the molecule, that are amplified by the

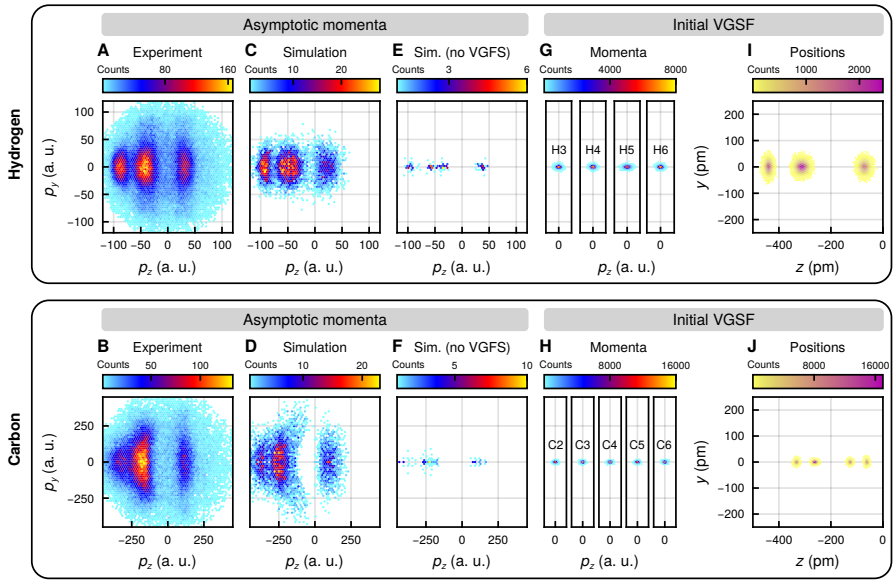

Figure 24: **Out-of-plane momenta of the hydrogen and carbon ions, in experiment and theory.** A-B. Experiment. C-D. Simulation with initial states sampled from the ground-state Wigner distribution. E-F. Simulation from fixed static geometries. G-H. VGSF of the momenta. I-J. VGSF of the positions.

explosion. To highlight this, we consider simulations that do not include VGSF, thus in which the only source of out-of-plane movement is the recoil from the electrons on the atoms when they get ionized. We show the resulting distribution of asymptotic momenta in Fig. 24E and Fig. 24F. Their amplitudes are minuscule in comparison with the one found in the experiment and the simulation with VGSF, confirming that the out-of-plane momenta that we see in the experiment is a consequence of the VGSF of the molecule.

This indicates that the magnitude of the VGSF is large enough to be probed the Coulomb explosion imaging. For reference, we estimate this magnitude using the standard deviation of the Wigner distribution, which is, on average, 13.9 pm in position space and 1.9 a.u. in momentum space for the hydrogen atoms, and 4.6 pm in position space and 4.9 a.u. in momentum space for the carbon atoms. Additionally, in Figs. 24G and 24H, we show the ground-state out-of-plane momenta for hydrogen and carbon ions, and in Figs. 24I and 24J, ground-state positions for hydrogen and carbon ions.

6.4 APPLICATION OF THE ALGORITHM TO EXPERIMENTAL DATA

6.4.1 *Data cleaning*

To further analyze the correlations among the out-of-plane momenta described in Section 6.2, we apply the fitting method described in

Chapter 5 to the experimental data, obtaining the principal components of the distribution of asymptotic momenta.

However, before we can apply the algorithm, we need to clean the experimental data, which has a substantial, widely spread, low-density background. It may originate from false coincidences that were not filtered out initially or belong to cases where the charge-up induced by the x rays was, for example, less violent or occurring along different and rare fragmentation pathways. It turns out that this contribution has a strong effect on the quality of the results of our fitting procedure. We therefore removed this low-density region from the data, using the procedure described below.

We clean the data by estimating the density of measured data points and excluding data points whose estimated density falls under a given threshold. The procedure is described here for the carbon ions, but the analogous procedure was used for protons. We filter out subsets of ions measured in coincidence that involve the detection of iodine (I), nitrogen (N), and any number of carbon (C) ions. For each of those in-coincidence measurements, we built all possible triplets of the form $(\mathbf{p}_\mathrm{I}, \mathbf{p}_\mathrm{N}, \mathbf{p}_\mathrm{C})$, where $\mathbf{p}_\mathrm{C}$ is the momentum of any of the measured C ions. From these momentum triplets, we generate sextets of internal coordinates containing the three momentum magnitudes and three relative angles,

$$(\|\mathbf{p}_\mathrm{I}\|, \|\mathbf{p}_\mathrm{N}\|, \|\mathbf{p}_\mathrm{C}\|, \angle(\mathbf{p}_\mathrm{I}, \mathbf{p}_\mathrm{N}), \angle(\mathbf{p}_\mathrm{I}, \mathbf{p}_\mathrm{C}), \angle(\mathbf{p}_\mathrm{N}, \mathbf{p}_\mathrm{C}))\,, \tag{67}$$

where $\angle(\cdot, \cdot)$ specifies the angle between two momenta. Since the norms and angles have different units, we whiten the data. That is, we center each internal coordinate around its mean value and divide it by its standard deviation. The whitened internal coordinates are unitless and have a standard deviation equal to 1. We estimate the density at each point in this internal-coordinate space by counting the number of neighbors in a 0.5 radius by representing the data as a k-d tree and performing a nearest neighbor search, as implemented in NearestNeighbor.jl [11]. If all possible sextets generated from a single measurement has more than 30 neighbors, we keep it, otherwise, we discard it.

We started with a data set with 34 336 shots, after cleaning the low-density regions, 8 060 remain. The effect of the procedure is illustrated in Fig. 25. Notably, it removes most of the large halo visible in the raw data (panel A) under the carbon ions and protons.

6.4.2 *Initial guess*

We build a reasonable guess to start the reconstruction by fitting a mixture of Gaussian distributions to the cleaned momenta in the iodine-nitrogen recoil frame (Fig. 25B). The fit is performed with the expectation–maximization algorithm, as implemented in the package

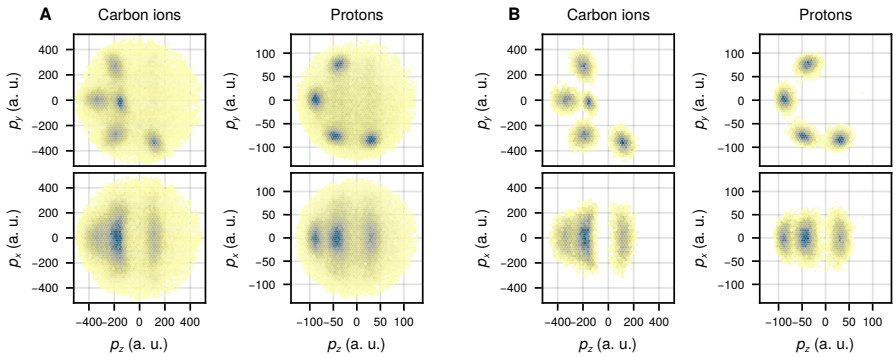

Figure 25: Effect of cleaning the data. A. Experimental data before cleaning. B. Experimental data after cleaning.

GaussianMixtures.jl [50]. The Gaussian mixtures approximate the covariance by a 3×3 block diagonal matrix of the covariance matrix. Each block corresponds to a single atom only, therefore there are no correlations between different ions in the guess.

Some elements of the guess covariance matrix are zero, namely the components corresponding to the variance of the iodine momentum along the x and y axes and the variance of the nitrogen momentum along the x axis, due to the definition of the frame. For these elements, we take 10% of the total variance of the respective ions along the other directions as the initial guess. This prevents the covariance matrix guess $\mathbf{\Sigma}$ from being singular.

6.4.3 *Results*

We apply the reconstruction algorithm to the cleaned experimental data, and get the set of principal components that describe it. The leading out-of-plane principal components are shown in Fig. 26A-D. They are in good qualitative agreement with the principal components from the simulation, shown in Fig. 26E-H. However, there is some discrepancy in the explained variance of each principal component, that leads us to reorder them according to their similarity, for an easier visual comparison.

We quantify the similarity by computing the overlap S defined as the scalar product between a principal component reconstructed from the experiment and a principal component of the simulation. We report the overlap for all pairs of the dominant out-of-plane components in Fig. 26I. The matrix of overlaps is close to a diagonal, which would correspond to a perfect agreement. We conclude that the simulation reproduces the experiment well enough to extend the conclusions drawn in Section 4.6 to the experimental data: the measurement of the asymptotic momenta probes the VGSF in the molecule.

In addition, in Section 6.5, we show that we can reach an even better agreement by taking into account the finite detection efficiency

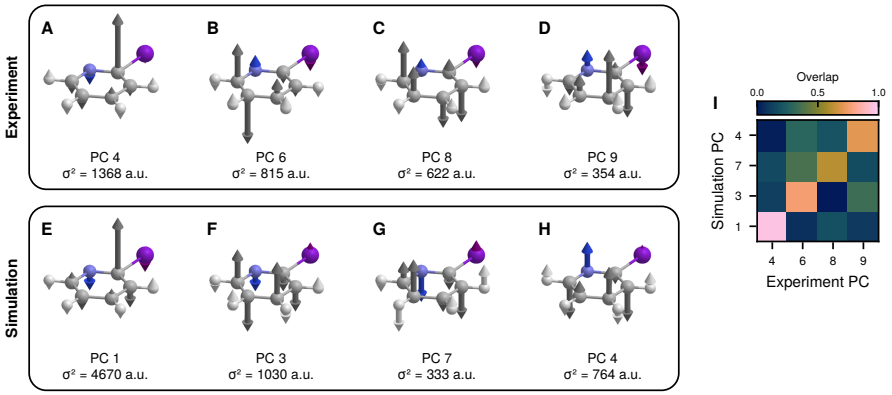

Figure 26: **Principal components of the distribution of asymptotic momenta.** A-D. Reconstructed from the experiment. E-H. Computed from simulation. σ^2 represents the variance explained by the component. I. Overlap between the experimental and simulation principal components. For a perfect agreement, the matrix would be diagonal.

of the detector in the experiment, and the diversity of possible charge states of the molecule, and in Section 6.6, we reproduce the same analysis on data with lower charges, demonstrating that a high charge state is necessary to achieve good agreement between experiment and simulation.

6.4.4 *Reproducing the collective features depicted in Fig. 23*

In Fig. 23, we show that the H3 proton emission direction is correlated with the emission direction of the carbon ions, when the data is inspected in the iodine-nitrogen recoil frame. We can, however, not directly relate these correlations to the principal components shown in Fig. 26, because the choice of the recoil frame has an impact on the variances of the data[1].

To give a simple example, consider an ion emission pattern where only the iodine ion is emitted out-of-plane, while all other atoms stay perfectly in-plane, as illustrated in Fig. 27A. By construction, the variance in this example is described by a single principal component illustrated in Fig. 27B. When the same data is rotated into the iodine-nitrogen molecular recoil frame, it enforces a zero out-of-plane variance for the iodine and nitrogen momenta. As a side effect, most of the other ions now get a significant out-of-plane momentum, as can be seen in Fig. 27C. In this frame, the principal component exhibits a correlated fluctuation in the out-of-plane momenta, shown in Fig. 27D.

This example demonstrates that using the recoil frame gives a qualitatively different picture of the correlations between the momenta.

1 It also affects the mean, but, in the present data, this effect is negligible.

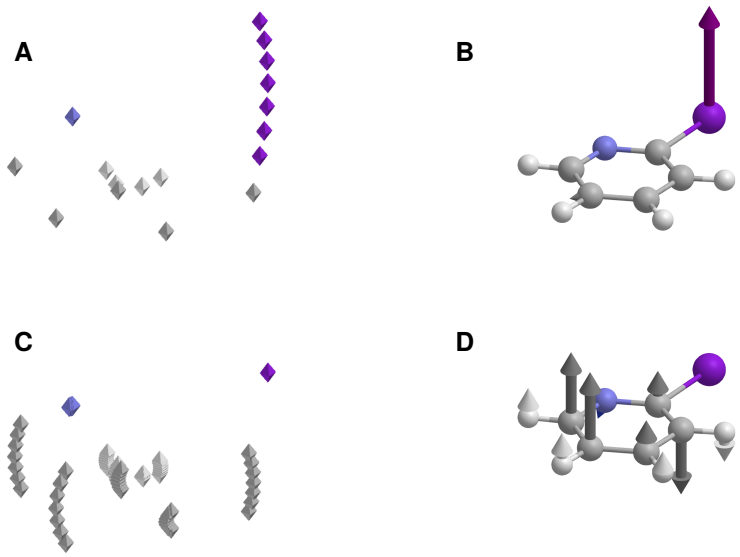

Figure 27: **Effect of the choice of molecular frame on principal components.** A and B. Artificial measurements in momentum space (A), where only the iodine ions are emitted out-of-plane and the corresponding principal component (B). C and D. Same data as in panel A, but each measurement is rotated into the iodine-nitrogen molecular frame (C) and the corresponding principal component (D).

In contrast, our approach for fitting a model distribution is based on mutual momentum scalar products as internal coordinates, and therefore the frame around which the fluctuations are considered is fitted to the data. By minimizing the discrepancy with the data, we obtain an unbiased frame for addressing the momentum fluctuations.

Therefore, to reproduce the pattern shown Fig. 23 we have to rotate the components in the iodine-nitrogen frame. Once done, we find one principal component that has a matching pattern, simulation PC 7 shown in Fig. 28A. The other panels show how the component matches the experimental result. In Fig. 28B, we show the experimental out-of-plane distribution of the asymptotic momenta of hydrogen ions. Each box is used to filter the hydrogen data, the corresponding experimental momentum distribution of the carbon ions is shown In Fig. 28C-E. The red overlay shows the expected distribution of momentum according to the displayed principal components. The agreement is excellent, indicating that this component is indeed the main contributor to the effect reported in Fig. 23.

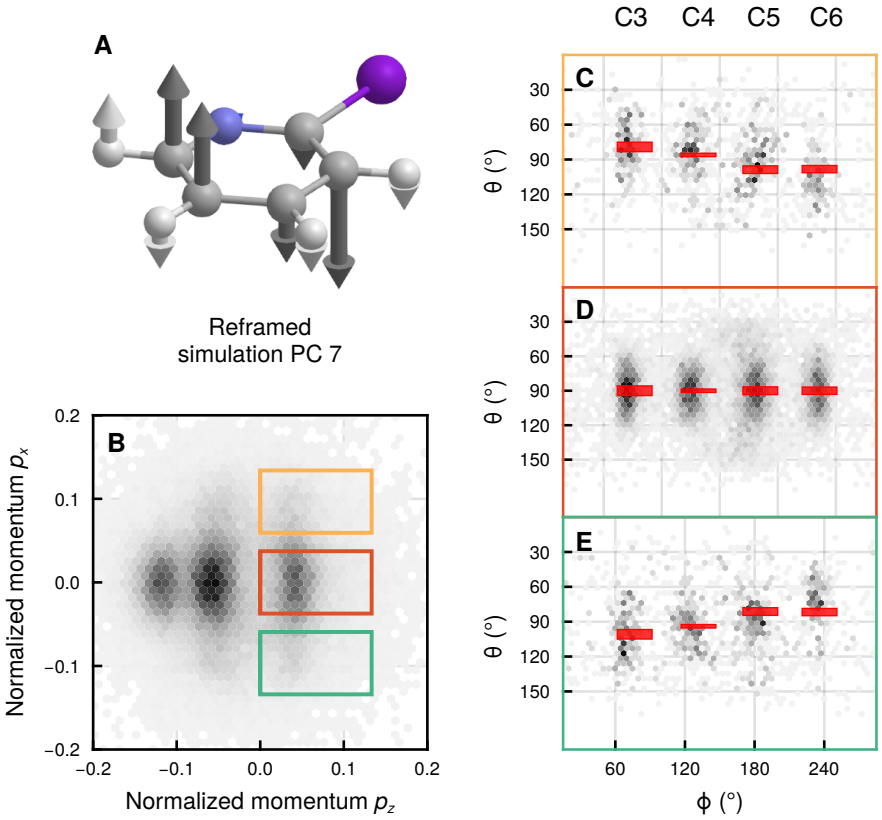

Figure 28: **Correlations between the emission direction of the H3 proton and the carbon ions, as captured by simulation PC 7.** A. Simulation PC 7 in the iodine-nitrogen frame. B. Experimental momentum distribution of the hydrogen ions. The boxes are used to filter the data with respect to H3 to see the correlation with carbon ions, as explained in Section 6.2. C-E. Angular distribution of the carbon momenta in spherical coordinates. The red overlays represent the correlation expected based on simulation PC 7.

6.5 EFFECT OF FINITE DETECTION EFFICIENCY ON THE RECONSTRUCTION

In Fig. 26, we analyze the experiment and the simulation in two subtly different ways. In this section, we describe the differences in more detail and show how they affect the result of the reconstruction algorithm.

When treating the experimental data, we use the $I^{+4}/N^{+2}/C^{+2}/H^{+}$ coincidence channel, i. e., a measurement is included in the analysis if all the following are detected in coincidence:

- the iodine ion with charge $+4$,
- the nitrogen ion with charge $+2$,
- at least one carbon ion with charge $+2$,
- at least one proton.

Additional C^{+2} and H^{+} ions detected in coincidence are also included in the measurement, while additional carbon and hydrogen ions with different charges are discarded. Also, some atoms can remain undetected, either due to the finite detection efficiency of the apparatus or because they are neutral in their final state. Therefore, the exact final charge state of the ions in the shot is unknown.

In contrast, with the simulation data, we always know the final charge of every atom for every trajectory. This means that a choice is necessary: when a measurement contains atoms that have a charge incompatible with the $I^{+4}/N^{+2}/C^{+2}/H^{+}$ coincidence channel, we can either

1. discard the measurements entirely when they contain atoms with a charge that does not match the coincidence channel (for example C^{+1}),
2. include all atoms in the analysis, regardless of their charge,
3. remove the non-matching atoms (e. g., neutral hydrogen) and create partial coincidence measurements.

Option (1) causes two problems. First, in the experiment, the measured molecules have a distribution of final charge states, and it can not be assumed to be $I^{+4}N^{+2}5C^{+2}4H^{+}$. Second, only 90 trajectories, out of the initial 61 652, have this exact charge state, a sample size too small to perform PCA meaningfully. Therefore, we decided never to use this option.

With Option (2), we expect the distribution of charge states to match the one found in the experiment. However, we also include on an equal footing some atoms incompatible with the $I^{+4}/N^{+2}/C^{+2}/H^{+}$ coincidence channel. We used this version in Chapters 4 and 6, as

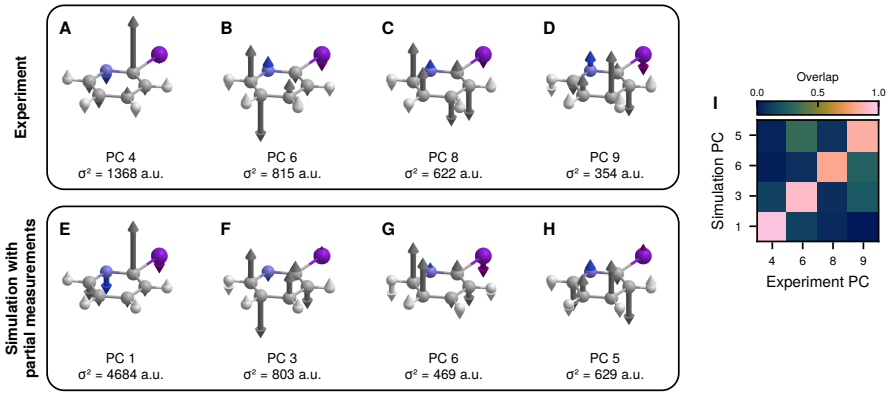

Figure 29: **Principal components of the distribution of asymptotic momenta.** A-D. Reconstructed from the experiment. E-H. Reconstructed from simulation by assuming a 60% detection efficiency, and only taking into account the ions with the correct charge. σ^2 represents the variance explained by the component. I. Overlap between the experimental and simulation principal components. For a perfect agreement, the matrix would be diagonal.

the statistic is sufficient (1 738 trajectories) and it preserves full 33-dimensional measurements. Having full measurements is necessary to determine the principal coordinate and establish the linear mapping between the vibrational modes and the principal components of the distribution of asymptotic momenta, as described in Section 4.6.

Finally, Option (3) is the closest to the treatment of the experiment, and similarly requires the use of our reconstruction algorithm to determine the principal components of the distribution. But, as mentioned above, we can not perform all the analyses we want on this dataset, therefore, to keep the main argument consistent, we preferred to always use Option (2) until this point.

We now show how the experiment and simulation compare, when we are as close as possible to the experiment. In addition to using Option (3), we only include atoms in a measurement with a 60% probability, corresponding to the detection efficiency of the detector in the experiment. We then apply the algorithm described in Chapter 5 to reconstruct the principal component of the distribution of asymptotic momenta.

We show the resulting principal components in Fig. 29E-H, together with the principal components of the experimental data in Fig. 29A-D. In this case, the agreement is nearly perfect, as confirmed by the matrix of overlap shown in Fig. 29I, a notable improvement from the results described in Sections 4.6 and 6.4.3.

This shows that the principal components are also sensitive to more subtle effect stemming from the charge state distribution. This represents a challenge for both the theory and the data analysis, as it is hard to gather sufficient statistics for a specific charge state, but it is

also necessary to acknowledge that the charge of the undetected ion have an impact on the results.

6.6 RECONSTRUCTION WITH LOW-CHARGE STATES

So far in this Chapter, we have studied the $I^{+4}/N^{+2}/C^{+2}/H^{+}$ coincidence channel. Here, we perform the same analysis on the $I^{+}/N^{+}/C^{+}/H^{+}$ coincidence channel, and show that this leads to significant discrepancies between experiment and simulation.

First, we apply the algorithm described in Chapter 5 to the experimental data with charge $I^{+}/N^{+}/C^{+}/H^{+}$, which is detailed in Chapter 2. The resulting leading out-of-plane components are shown in Fig. 30A-D.

Then, we simulate the explosion using the same parameters described in Section 4.5, but with a lower fluence, to increase the probability of getting the charge state of interest. In short, the simulations use XMDYN interfaced with a ReaxFF implementation to model the molecular bonds, and include the vibrational ground-state fluctuation (VGSF) of the molecule. These simulations are summarized in Table 1 as Simulation Set V. We retrieve the principal components of the obtained distribution of asymptotic momenta using our algorithm to get as close as possible to the experiment, as described in Section 6.5. We plot the leading out-of-plane components in Fig. 30E-H. We see that, despite using our best simulation and best analysis procedure, the results do not match with the experiment. This is further confirmed by looking at the matrix of similarity, shown in Fig. 30I, that has little in common with the identity matrix, which would correspond to agreement with the experiment. The only matching components are experimental PC 7 and simulation PC 2, stemming from the collision between C2 and I, as described in Chapter 3. However, none of the principal components that can be linked to the VGSF of the molecule matches, therefore the simulation can not be used to help us interpret the experiment.

We explain this discrepancy by the effect of the molecular bonds. They are not perfectly reproduced in the simulation, as the reactive force field used to model them is optimized for a neutral molecule close to equilibrium, but during a Coulomb explosion, the molecule stays neither neutral, nor close to equilibrium, and accordingly, the force field is switched off. However, to reach a higher final charge state, like in Chapter 6, the charging up of the molecule must be significantly faster. Thus, we expect all electrons relevant to molecular bonding to be removed before they have time to significantly affect the dynamic of the explosion. In this case, accurate modeling of the bonds is therefore less important, and the simulations can match the experiment.

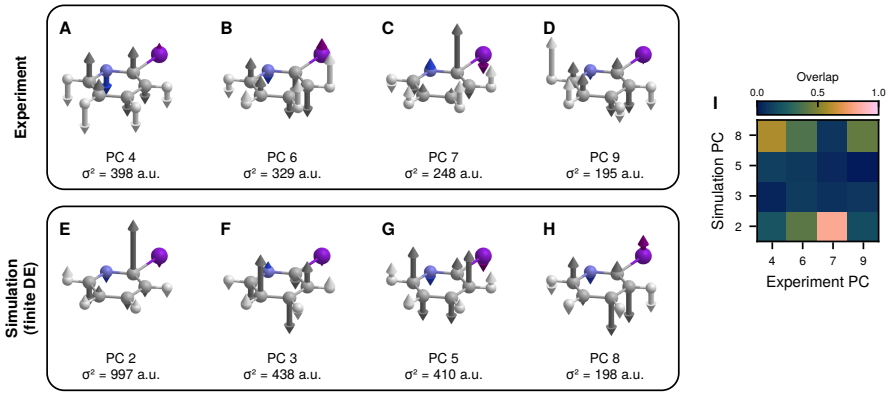

Figure 30: **Principal components of the distribution of asymptotic momenta.** A-D. Reconstructed from the experiment. E-H. Computed from simulation. σ^2 represents the variance explained by the component. I. Overlap between the experimental and simulation principal components. For a perfect agreement, the matrix would be diagonal.

This shows that by considering a higher charge state, we can select measurements where the Coulomb explosion is dominated by the Coulomb forces and the simulations match the experiment.

6.7 CONCLUSION

In this Chapter, we apply our algorithm to experimental data with a high-charge state. The decomposition into principal components directly gives us access to intricate features of the data, that can otherwise only be discovered thanks to a good intuition.

Thanks to the high-charge state considered, which ensure a minimal effect of the molecular bond on the explosion dynamic, our simulations is able to reproduce these fine details of the experiment, for the out-of-plane features of the distribution of asymptotic momenta.

This high agreement between theory and experiment allows us to use the principal components from the simulation to relate the results to the vibrational modes of the molecule, as described in Section 4.6. Accordingly, we can conclude that the current Coulomb explosion experiment is probing the VGSF, and that the main correlations discussed in Fig. 23 result from the vibrational modes with frequency 177 $cm^{-}1$ and 1067 $cm^{-}1$.

7 CONCLUSION AND OUTLOOK

7.1 CONCLUSION

In this thesis, I study Coulomb explosion imaging (CEI) from a data-driven perspective, exploring its potential for imaging by analyzing the simulations, and developing new tools to understand the experiment. I address the explosion of the 2-iodopyridine molecule, but the challenges and opportunities encountered in this specific case generalize to any Coulomb explosion experiment with molecules larger than 6 atoms, for which full coincidence measurements are hard to obtain with sufficient statistic.

My work shows that to exploit CEI to its fullest potential, one needs to understand and exploit the high-dimensional features of the measurements. In particular, in Chapter 3 I demonstrate the usefulness of principal component analysis (PCA), that decomposes the data into components describing the most important correlations between the measured ions. Applied to the dynamics of the explosion, PCA highlights several processes affecting the explosion dynamics. For iodopyridine, the leading contributions are the charging up of the molecule, that causes the total amount of kinetic energy released by the explosion to vary due to its randomness, and an atomic collision between the iodine atom and the closest carbon atom. These contributions also leave measurable fingerprints in the components obtained by applying PCA to the asymptotic momenta of the fragments. Moreover, due to its quantum nature, the molecule undergoes ground-state fluctuations. Including them in the simulations, as I did in Chapter 4, demonstrates that some of their features are mapped to correlations between the measured ions. Therefore, properties of both the explosion dynamics and the initial state of the system can be probed by CEI, provided that we study the measured asymptotic momenta as high-dimensional vectors.

Applying this analysis to experimental data is challenging, due to the amount of information that is missing compared to simulations, namely the orientation of the molecule in the lab frame, the identity of the detected ions, and the fragments that are not observed due to the finite efficiency of the detectors. The new method presented in Chapter 5 addresses these challenges by fitting a Gaussian model to the data by matching the per-element moments of the scalar products (PEMSP). The implementation of the method is heavily optimized to run in a reasonable time. Its convergence behavior is established by testing it on the simulation data for which the full information is

available. In Chapter 6 I apply it to the experimental data of the explosion of iodopyridine, revealing that to obtain a satisfying agreement between our simulations and the experiment, we must use measurement with high charge states. This ensures that the chemical bonds in the molecules are broken during the few first femtoseconds of the explosion, simplifying the explosion dynamics such that we are able to reproduce it. The good agreement between simulation and experiment allows to conclude that the experimental data contain fingerprints of specific ground-state fluctuations of the molecules.

This constitutes the final achievement of my thesis: by combining simulations, data analysis, and experimental data, CEI gives access to an unprecedented amount of information about the initial state of the molecule even though not all atomic fragments can be detected in coincidence.

7.2 OUTLOOK

To see how the findings of this thesis fit in the greater picture, let us first go back to the introduction. There, I presented the goal of this thesis as follows: to advance the imaging of single molecules on a femtosecond timescale using CEI, in the perspective of using it in time-resolved pump-probe experiments. How do my findings fit relative to this stated goal?

I explore the capabilities of CEI as a high-dimensional probe for molecular systems. In particular, I show that Coulomb explosion experiments provide a wealth of information that can, however, only be accessed when the data is treated carefully. For example, the variance present in the data is an important feature of the measured distribution of momenta, and must be matched by the simulations to guarantee a quantitative agreement with the experiment. Likewise, the information about the correlations between ions is precious, but can be hidden by an analysis that relies only on a low-dimensional representation of the data, such as a Newton diagram. This demonstrates that when analyzed with care, CEI is a sensitive *probe*, containing fingerprints of both the processes that happen during the explosion, and of the state of the system at the onset of the explosion.

Reproducing the experimental data with the simulations to this high-level of details is difficult. In general, the explosion of a molecule such as iodopyridine is far from ideal: the features of the initial state can blend with the dynamics of the fragmentation, even when the molecule fully fragments into individual ions. Therefore, the measurement process is not independent of the properties of the system being measured, and CEI only acts as a *probe*, albeit an excellent one. To achieve true *imaging* the explosion must be independent of the chemical dynamics that we want to image with CEI. As I highlight in Chapter 6, this requires a high-charge state that guarantees a fast

removal of enough valence electrons, such that the chemical bonds can be neglected in the description of the explosion.

Fulfilling this condition unlocks the prospect of CEI as an ultrafast high-dimensional imaging method, specifically sensitive to collective behaviors. This thesis moreover establishes that CEI is not limited by the measurement of only a subset of ions, and can therefore be extended to larger molecules as long as the experiment can produce a sufficiently high charge state with a femtosecond pulse. In consequence, I expect the evolution of CEI to continue toward increasingly larger molecules, up to a few tens of atoms, which is enough to image many molecules relevant to chemistry.

Of course, imaging static distributions of molecular structures is only the beginning for CEI. The ideas put forward in this thesis can be exploited along four main avenues: bettering our understanding of the Coulomb explosion process, applying CEI to pump-probe experiments, probing a wide range of static properties of a system, and finally combining CEI with other experimental techniques. We look at each of them in detail.

First, finely understanding how a Coulomb explosion images a molecule is key to interpreting and using the data from CEI experiments. Based on the observation that crucial information is contained in the correlations between atoms, the simulation of the explosion could be pushed further: instead of taking an atomistic view of the molecule, its dynamics could be projected on the dominant principal components and simulated in a space of significantly lower dimensionality.

The reduction of the dimensionality of the data is not limited to PCA. It would be natural to define a reduced space based on the specific features that we want to image, making sure that they are correctly reproduced even when the simulation are performed in much lower dimensionality. This is a challenging task, that can be tackled using modern machine learning methods.

At the same time, learning more about the explosion dynamic will also help to solve the inverse problem: to reconstruct the initial state from the measurements. This could be achieved by running simulations of the explosion on a guess for the initial state, and iteratively improving this guess until the results of the simulations match the measurement. In this context, the results could be summarized by fitting a Gaussian distribution with the algorithm presented in this thesis. This would provide a clear target for the reconstruction of the initial state, more suitable for optimization than the partial and non-homogenous raw experimental data. Note that for such an idea, it is also possible to summarize the data directly with the PEMSP. In some cases, using the PEMSP may even be preferred to avoid an extra intermediate step in the procedure.

Second, as pump-probe CEI experiments [37, 87] become more common, it is natural to apply the methods developed in this thesis to temporal data, either gathered in a time-resolved experiment or generated from simulations. The algorithm can be applied frame by frame, resulting in a high-dimensional movie of the measured momenta, that I expect to be easier to wield and understand than the set of measurements. In particular, the changes in the mean and the principal components will inform on how the probed dynamic affects the distribution of states of the molecule, possibly allowing us to follow the evolution of quantum wavepackets during chemical reactions.

The compact representation of the momentum distribution as a Gaussian, or a set of PEMSP will ease the use of more advanced techniques that infer the dynamic of the system, such as dynamic mode decomposition [10], a standard method to treat high dimensional temporal data. Alternatively, our Gaussian fitting algorithm can be applied to the set of all frames as a whole, an idea also used to deduce reaction coordinates from molecular simulations using PCA [7, 52]. In this case, the principal components highlight the directions with the highest temporal variance, allowing the analysis to focus on what is changing in the data.

Third, the algorithm is likely to be applicable to a wide range of problems, with different interpretations depending on the situation. For example, understanding how atoms correlate may allow for separating overlapping blobs in Newton diagrams, thus resolving individual fragments in more complex molecules, especially if they are flexible. Similarly, when considering a molecule in solution, the ions produced from the surrounding liquid are expected to be less correlated to each other than to the ones originating from the more rigid structure of the solute. By identifying all atom-atom correlations, the solute may be distinguished from the solvent, effectively extending the range of application of CEI beyond isolated gas phase molecules.

Moreover, we are not limited to a purely Gaussian approximation of the distribution of asymptotic momenta. The most straightforward extension to the algorithm is to use a Gaussian mixture to reconstruct more general momentum distributions. This is especially interesting for experiments that contain a mixture of systems in different states, for example, samples containing different isomers of a molecule. Each set of isomers would correspond to a single Gaussian in the mixture, and its properties would thus be retrieved individually.

Together with time-resolved experiments, this would enable following the evolution of complex chemical processes that involves multiple reactions in parallel, and are currently extremely challenging to untangle. Together with modern advances in the triggering of thermal chemistry reactions [69], this could be applied to processes of industrial importance. In the future, I expect this field to be the main

domain of application of x-ray induced CEI, because it fits the unique capability of the technique, ultrafast, sensitive to all atoms, including hydrogen, and able to resolve single molecules. As mention in the introduction, the chemical industry is central to the modern world. Therefore, better understanding the chemical reaction at its center may bring far-reaching progresses by allowing to optimize costly processes or avoid the most polluting ones.

A first potential candidate for this line of study could be the Grignard reactions, a class of chemical reactions where a Grignard reagent is added to an aldehyde or a ketone to produce different sort of alcohol. Despite its importance over the past century, earning Grignard the Nobel Prize in 1912 and being largely used in the industry, its mechanism has only been elucidated using simulation in 2020 [29, 67], and, to my knowledge, no experiment has been able to confirm these theory findings yet.

Fourth, the analysis could be extended to work together with additional information. In a Coulomb explosion experiment, the first candidate is the charge state of the fragments. In this thesis, we have considered a single charge state at a time. Undoubtedly, the charge state is correlated with the measured momenta, and studying them together will provide insights, in particular on the dynamics of the explosion itself.

Then, CEI can be performed in tandem with other measurements, like diffraction, which causes the fragmentation of the molecule anyway. One existing proposition [70] is to infer the orientation of the molecule thanks to CEI, and merge this information with collected diffraction patterns to retrieve the structure of a measured particle. CEI however provides much richer data than simply the orientation of the molecule, determining to which extent it can be merged with data of another kind is likely to open new research avenues.

In fact, considering the experimental challenge that such combined experiments represent, establishing its potential from simulation is likely to be of great interest.

APPENDIX

A

DERIVATION OF THE UNCOUPLED HARMONIC OSCILLATORS

We want to substitute $\mathbf{x} = \mathbf{MVz}$ into

$$\mathcal{H} = \sum_i \frac{-\hbar^2}{2m_i}\frac{\partial^2}{\partial x_i^2} + \mathbf{x}^T\mathbf{Hx}. \tag{68}$$

By the definition of the diagonal matrix $\mathbf{M}$ and the orthogonal matrix $\mathbf{V}$ we have

$$\mathbf{x}^T\mathbf{Hx} = \mathbf{z}^T\mathbf{V}^T\mathbf{MHMVz} = \mathbf{z}^T\mathbf{\Omega z}. \tag{69}$$

For the other term, we use Leibnitz chain rule, starting with the first derivative only

$$\frac{\partial}{\partial x_i} = \sum_j \frac{\partial z_j}{\partial x_i}\frac{\partial}{\partial z_j}. \tag{70}$$

Next we use $\mathbf{z} = \mathbf{V}^T\mathbf{M}^{-1}\mathbf{x}$ to compute

$$\frac{\partial z_j}{\partial x_i} = \frac{\partial}{\partial x_i}\mathbf{e}_j^T\mathbf{V}^T\mathbf{M}^{-1}\mathbf{x} = \mathbf{e}_j^T\mathbf{V}^T\mathbf{M}^{-1}\mathbf{e}_i, \tag{71}$$

where $\mathbf{e}_i$ are the unit vectors of the canonical basis. Since no new dependency in $\mathbf{x}$ or $\mathbf{z}$ appear, we can apply it twice. We transpose one of them however, to make the vector with index i appear together.

$$\begin{aligned}\sum_i \frac{-\hbar^2}{2m_i}\frac{\partial^2}{\partial x_i^2} &= -\frac{\hbar^2}{2}\sum_{i,j,k}\frac{1}{m_i}\frac{\partial z_k}{\partial x_i}\left[\frac{\partial z_j}{\partial x_i}\right]^T \\ &= -\frac{\hbar^2}{2}\sum_{j,k}\mathbf{e}_k^T\mathbf{V}^T\mathbf{M}^{-1}\left[\sum_i \frac{1}{m_i}\mathbf{e}_i\mathbf{e}_i^T\right]\mathbf{M}^{-1}\mathbf{V}\mathbf{e}_j\frac{\partial}{\partial z_j}\frac{\partial}{\partial z_k}.\end{aligned} \tag{72}$$

By definition,

$$\begin{aligned}&\mathbf{M}^2 = \sum_i \frac{1}{m_i}\mathbf{e}_i\mathbf{e}_i^T, \\ &\mathbf{V}^T\mathbf{V} = \mathbb{1}, \qquad \text{and} \\ &\delta_{kj} = \mathbf{e}_k^T\mathbf{e}_j,\end{aligned} \tag{73}$$

where $\mathbb{1}$ is the identity matrix and δ_{kj} is the Kroenecker delta. Using those properties, everything collapses, as

$$\begin{aligned}\sum_i \frac{-\hbar^2}{2m_i}\frac{\partial^2}{\partial x_i^2} &= -\frac{\hbar^2}{2}\sum_{j,k}\mathbf{e}_k^T\mathbf{V}^T\mathbf{V}\mathbf{e}_j\frac{\partial}{\partial z_j}\frac{\partial}{\partial z_k}\\ &= -\frac{\hbar^2}{2}\sum_{j,k}\mathbf{e}_k^T\mathbf{e}_j\frac{\partial}{\partial z_j}\frac{\partial}{\partial z_k}\\ &= -\frac{\hbar^2}{2}\sum_j\frac{\partial^2}{\partial z_j^2},\end{aligned} \tag{74}$$

and we end up with a set of uncoupled harmonic oscillators

$$\mathcal{H} = \sum_j \frac{-\hbar^2}{2}\frac{\partial^2}{\partial z_j^2} + \omega_j^2 z_j^2. \tag{75}$$

B

LINEAR FIT OF INDEPENDENT RANDOM VARIABLES

Consider two random vectors $\mathbf{x}$ and $\mathbf{y}$. Furthermore, assume that

$$\begin{aligned} &\langle x_i \rangle = \langle y_j \rangle = 0 \qquad \text{and} \\ &\text{std}[x_i] = \text{std}\left[y_j\right] = 1, \end{aligned} \tag{76}$$

for all components x_i and y_j of $\mathbf{x}$ and $\mathbf{y}$ respectively. This can always be achieved by *whitening* a variable as

$$x_i \to \frac{x_i - \langle x_i \rangle}{\text{std}[x_i]}. \tag{77}$$

Furthermore, assume that the x_i are independent, $\text{cov}\left[x_i, x_j\right] = 0$, for $i \neq j$.

Then, let $\mathbf{X}$ and $\mathbf{Y}$ be data matrices where each column represents a realization of the random vectors $\mathbf{x}$ and $\mathbf{y}$. Then, if they are in a perfect linear relation to each other we have

$$\mathbf{AX} = \mathbf{Y}, \tag{78}$$

where the matrix $\mathbf{A}$ is

$$\mathbf{A} = \mathbf{A}\text{cov}[\mathbf{x}, \mathbf{x}] = \mathbf{AXX}^T = \mathbf{YX}^T = \text{cov}[\mathbf{x}, \mathbf{y}]\,, \tag{79}$$

where we use the fact that

$$\begin{aligned} &\text{cov}[\mathbf{x}, \mathbf{x}] = \mathbf{XX}^T \qquad \text{and} \\ &\text{cov}[\mathbf{x}, \mathbf{x}] = \mathbb{1}, \end{aligned} \tag{80}$$

which are direct consequences of the properties that we assume for the random vectors $\mathbf{x}$ and $\mathbf{y}$. The unit standard deviation also implies

$$\text{cov}[\mathbf{x}, \mathbf{y}] = \text{cor}[\mathbf{x}, \mathbf{y}]\,, \tag{81}$$

establishing that the correlation of coefficient defines the linear relation between $\mathbf{x}$ and $\mathbf{y}$.

Note that in general, the linear mapping is not perfect and an error term $\mathbf{E}$ must be included as

$$\mathbf{AX} = \mathbf{Y} + \mathbf{E}, \tag{82}$$

leading to

$$\text{cor}[\mathbf{x}, \mathbf{y}] = \mathbf{A} - \mathbf{EX}^T. \tag{83}$$

Therefore, the correlation coefficient is a biased estimator of the linear mapping $\mathbf{A}$. Due to its simplicity, it is nonetheless useful to be considered.

BIBLIOGRAPHY

[1] Michael Abbott, Dilum Aluthge, N3N5, Simeon Schaub, Chris Elrod, Carlo Lucibello, and Johnny Chen. *Mcabbott/Tullio.Jl: V0.3.5*. Zenodo, Sept. 22, 2022. DOI: 10.5281/zenodo.7106192. URL: https://zenodo.org/record/7106192 (visited on 10/26/2022).

[2] Hasan Metin Aktulga, Sagar A. Pandit, Adri C. T. van Duin, and Ananth Y. Grama. „Reactive Molecular Dynamics: Numerical Methods and Algorithmic Techniques." In: *SIAM Journal on Scientific Computing* 34.1 (Jan. 2012), pp. C1–C23. ISSN: 1064-8275. DOI: 10.1137/100808599. (Visited on 07/25/2024).

[3] Andrea Amadei, Antonius B. M. Linssen, and Herman J. C. Berendsen. „Essential Dynamics of Proteins." In: *Proteins: Structure, Function, and Bioinformatics* 17.4 (1993), pp. 412–425. ISSN: 1097-0134. DOI: 10.1002/prot.340170408. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/prot.340170408 (visited on 05/16/2021).

[4] Simon D. Aubigne, Jeremy B. Cooper, Bruce L. Williams, and Derrick J. Watson. „Process for the Production of Acetic Acid." Pat. US5416237A. May 1995. (Visited on 04/02/2024).

[5] Surjendu Bhattacharyya et al. „Strong-Field-Induced Coulomb Explosion Imaging of Tribromomethane." In: *The Journal of Physical Chemistry Letters* (June 21, 2022). DOI: 10.1021/acs.jpclett.2c01007. URL: https://pubs.acs.org/doi/pdf/10.1021/acs.jpclett.2c01007 (visited on 09/22/2022).

[6] J. Stephen Binkley, John A. Pople, and Warren J. Hehre. „Self-consistent molecular orbital methods. 21. Small split-valence basis sets for first-row elements." In: *J. Am. Chem. Soc.* 102 (1980), pp. 939–947. DOI: 10.1021/ja00523a008.

[7] Adam B. Birkholz and H. Bernhard Schlegel. „Coordinate Reduction for Exploring Chemical Reaction Paths." In: *Theoretical Chemistry Accounts* 131.3 (Feb. 2012), p. 1170. ISSN: 1432-2234. DOI: 10.1007/s00214-012-1170-6. (Visited on 08/04/2024).

[8] Rebecca Boll. *Tracing Electron Rearrangement upon Absorption of Very Intense X-ray Pulses*. 2019. DOI: 10.22003/XFEL.EU-DATA-002159-00. (Visited on 04/23/2024).

[9] Rebecca Boll et al. „X-Ray Multiphoton-Induced Coulomb Explosion Images Complex Single Molecules." In: *Nature Physics* (Feb. 21, 2022), pp. 1–6. ISSN: 1745-2481. DOI: 10.1038/s41567-022-01507-0. URL: https://www.nature.com/articles/s41567-022-01507-0 (visited on 03/31/2022).

[10] Steven L. Brunton, Marko Budišić, Eurika Kaiser, and J. Nathan Kutz. *Modern Koopman Theory for Dynamical Systems*. Oct. 2021. DOI: 10.48550/arXiv.2102.12086. arXiv: 2102.12086 [cs, eess, math]. (Visited on 08/02/2024).

[11] Kristoffer Carlsson et al. *KristofferC/NearestNeighbors.Jl: V0.4.16*. Version v0.4.16. Zenodo, Dec. 27, 2023. DOI: 10.5281/zenodo.10435686. URL: https://zenodo.org/records/10435686 (visited on 02/07/2024).

[12] Denis Céolin et al. „Recoil-Induced Ultrafast Molecular Rotation Probed by Dynamical Rotational Doppler Effect." In: *PNAS* 116.11 (Mar. 2019), pp. 4877–4882. DOI: 10.1073/pnas.1807812116.

[13] ChemAnalyst. *Acetic Acid Market Size, Growth | Analysis & Forecast, 2032*. https://www.chemanalyst.com/industry-report/acetic-acid-market-609. May 2023. (Visited on 04/02/2024).

[14] Benjámin Csorba, Péter Szabó, Szabolcs Góger, and György Lendvay. „The Role of Zero-Point Vibration and Reactant Attraction in Exothermic Bimolecular Reactions with Submerged Potential Barriers: Theoretical Studies of the R + HBr → RH + Br (R = CH_3, HO) Systems." In: *J. Phys. Chem. A* 125.38 (2021), pp. 8386–8396. DOI: 10.1021/acs.jpca.1c05839. eprint: https://doi.org/10.1021/acs.jpca.1c05839. URL: https://doi.org/10.1021/acs.jpca.1c05839.

[15] Josh Day. *Joshday/OnlineStats.Jl*. URL: https://github.com/joshday/OnlineStats.jl (visited on 09/13/2023).

[16] Kerwin D. Dobbs and Warren J. Hehre. „Molecular orbital theory of the properties of inorganic and organometallic compounds 4. Extended basis sets for third- and fourth-row, main-group elements." In: *J. Comput. Chem.* 7 (1986), pp. 359–378. DOI: 10.1002/jcc.540070313.

[17] Wolfgang Domcke and Lorenz S. Cederbaum. „Electronic Recoil Effects in High-Energy Photoelectron Spectroscopy." In: *J. Electron Spectrosc. Relat. Phenom.* 13.3 (Jan. 1978), pp. 161–173. DOI: 10.1016/0368-2048(78)85024-5.

[18] Krzysztof Domino, Piotr Gawron, and Łukasz Pawela. „Efficient Computation of Higher Order Cumulant Tensors." In: *SIAM Journal on Scientific Computing* 40.3 (Jan. 2018), A1590–A1610. ISSN: 1064-8275, 1095-7197. DOI: 10.1137/17M1149365. arXiv: 1701.05420 [cs]. URL: http://arxiv.org/abs/1701.05420 (visited on 10/24/2022).

[19] Reinhardt Dörner, Volker Mergel, Ottmar Jagutzki, Lutz Spielberger, Joachim Ullrich, Robert Moshammer, and Horst Schmidt-Böcking. „Cold Target Recoil Ion Momentum Spectroscopy: A 'Momentum Microscope' to View Atomic Collision Dynamics."

In: *Physics Reports* 330.2 (June 1, 2000), pp. 95–192. ISSN: 0370-1573. DOI: 10.1016/S0370-1573(99)00109-X. URL: https://www.sciencedirect.com/science/article/pii/S037015739900109X (visited on 09/22/2022).

[20] Tomoyuki Endo et al. „Capturing Roaming Molecular Fragments in Real Time." In: *Science* 370.6520 (Nov. 2020), pp. 1072–1077. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.abc2960. pmid: 33243885.

[21] Benjamin Erk et al. „Imaging Charge Transfer in Iodomethane upon X-Ray Photoabsorption." In: *Science* 345.6194 (July 2014), pp. 288–291. DOI: 10.1126/science.1253607.

[22] Mark V. Fedkin et al. „Development of the ReaxFF Methodology for Electrolyte–Water Systems." In: *The Journal of Physical Chemistry A* 123.10 (Mar. 2019), pp. 2125–2141. ISSN: 1089-5639. DOI: 10.1021/acs.jpca.8b10453. (Visited on 07/25/2024).

[23] Paul Feyerabend. *Against Method: Outline of an Anarchistic Theory of Knowledge*. Verso Books, May 2020. ISBN: 978-1-78960-097-1.

[24] *FluxML/Zygote.Jl*. FluxML. URL: https://github.com/FluxML/Zygote.jl (visited on 09/13/2023).

[25] Aderonke S. Folorunso, Adam Bruner, François Mauger, Kyle A. Hamer, Samuel Hernandez, Robert R. Jones, Louis F. DiMauro, Mette B. Gaarde, Kenneth J. Schafer, and Kenneth Lopata. „Molecular Modes of Attosecond Charge Migration." In: *Physical Review Letters* 126.13 (Mar. 2021), p. 133002. DOI: 10.1103/PhysRevLett.126.133002. (Visited on 04/24/2024).

[26] Sven Grundmann et al. „Observation of Photoion Backward Emission in Photoionization of He and ${\mathrm{N}}_{2}$." In: *Phys. Rev. Lett.* 124.23 (June 2020), p. 233201. DOI: 10.1103/PhysRevLett.124.233201.

[27] Yajiang Hao, Ludger Inhester, Kota Hanasaki, Sang-Kil Son, and Robin Santra. „Efficient Electronic Structure Calculation for Molecular Ionization Dynamics at High X-ray Intensity." In: *Struct. Dyn.* 2.4 (2015), p. 041707. DOI: http://dx.doi.org/10.1063/1.4919794.

[28] Christopher J. Hensley, Jie Yang, and Martin Centurion. „Imaging of Isolated Molecules with Ultrafast Electron Pulses." In: *Physical Review Letters* 109.13 (Sept. 28, 2012), p. 133202. DOI: 10.1103/PhysRevLett.109.133202. URL: https://link.aps.org/doi/10.1103/PhysRevLett.109.133202 (visited on 03/13/2024).

[29] Nathaniel Herzberg. „La réaction de Grignard, à l'origine des parfums, démasquée par les Arsène Lupin de la chimie." In: *Le Monde* (July 2024). ISSN: 1950-6244. (Visited on 07/30/2024).

[30] Akiyoshi Hishikawa, Akitaka Matsuda, Mizuho Fushitani, and Eiji J. Takahashi. „Visualizing Recurrently Migrating Hydrogen in Acetylene Dication by Intense Ultrashort Laser Pulses." In: *Phys. Rev. Lett.* 99.25 (Dec. 2007), p. 258302. ISSN: 0031-9007, 1079-7114. DOI: 10.1103/PhysRevLett.99.258302. URL: http://link.aps.org/doi/10.1103/PhysRevLett.99.258302 (visited on 01/17/2014).

[31] Joshua Holcomb, Nicholas Spellmon, Yingxue Zhang, Maysaa Doughan, Chunying Li, and Zhe Yang. „Protein Crystallization: Eluding the Bottleneck of X-ray Crystallography." In: *AIMS biophysics* 4.4 (2017), pp. 557–575. ISSN: 2377-9098. DOI: 10.3934/biophy.2017.4.557.

[32] Andrew J. Howard et al. „Filming Enhanced Ionization in an Ultrafast Triatomic Slingshot." In: *Communications Chemistry* 6.1 (1 Apr. 27, 2023), pp. 1–10. ISSN: 2399-3669. DOI: 10.1038/s42004-023-00882-w. URL: https://www.nature.com/articles/s42004-023-00882-w (visited on 12/12/2023).

[33] Jochen S. Hub and Bert L. de Groot. „Detection of Functional Modes in Protein Dynamics." In: *PLOS Computational Biology* 5.8 (Aug. 28, 2009), e1000480. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1000480. URL: https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1000480 (visited on 04/02/2021).

[34] Heide Ibrahim et al. „Tabletop Imaging of Structural Evolutions in Chemical Reactions Demonstrated for the Acetylene Cation." In: *Nature Communications* 5.1 (July 2014), p. 4422. ISSN: 2041-1723. DOI: 10.1038/ncomms5422.

[35] Ottmar Jagutzki, Volker Mergel, Klaus Ullmann-Pfleger, Lutz Spielberger, Ullrich Meyer, Reinhard Doerner, and Horst W. Schmidt-Boecking. „Fast Position and Time-Resolved Read-out of Micro-Channelplates with the Delay-Line Technique for Single-Particle and Photon-Detection." In: *Imaging Spectrometry IV*. Vol. 3438. SPIE, Oct. 1998, pp. 322–333. DOI: 10.1117/12.328113. (Visited on 07/04/2024).

[36] T. Jahnke, Th. Weber, T. Osipov, A.L. Landers, O. Jagutzki, L.Ph.H. Schmidt, C.L. Cocke, M.H. Prior, H. Schmidt-Böcking, and R. Dörner. „Multicoincidence studies of photo and Auger electrons from fixed-in-space molecules using the COLTRIMS technique." In: *J. Electron Spectrosc. Relat. Phenom.* 141.2 (2004), pp. 229–238. ISSN: 0368-2048. DOI: https://doi.org/10.1016/j.elspec.2004.06.010. URL: https://www.sciencedirect.com/science/article/pii/S0368204804003408.

[37] Till Jahnke et al. *X-Ray Coulomb Explosion Imaging Reveals Role of Molecular Structure in Internal Conversion*. May 2024. DOI: 10.48550/arXiv.2405.15367. arXiv: 2405.15367 [physics]. (Visited on 06/19/2024).

[38] „Mathematical and Statistical Properties of Sample Principal Components." In: *Principal Component Analysis*. Ed. by I. T. Jolliffe. Springer Series in Statistics. New York, NY: Springer, 2002, pp. 29–61. ISBN: 978-0-387-22440-4. DOI: 10.1007/0-387-22440-8_3. URL: https://doi.org/10.1007/0-387-22440-8_3 (visited on 08/09/2021).

[39] Z. Jurek, S.-K. Son, B. Ziaja, and R. Santra. „XMDYN and XATOM: Versatile Simulation Tools for Quantitative Modeling of X-ray Free-Electron Laser Induced Dynamics of Matter." In: *Journal of Applied Crystallography* 49.3 (3 June 1, 2016), pp. 1048–1056. ISSN: 1600-5767. DOI: 10.1107/S1600576716006014. URL: http://scripts.iucr.org/cgi-bin/paper?zd5003 (visited on 04/02/2021).

[40] Mehmet Cagri Kaymak, Ali Rahnamoun, Kurt A. O'Hearn, Adri C. T. van Duin, Kenneth M. Jr. Merz, and Hasan Metin Aktulga. „JAX-ReaxFF: A Gradient-Based Framework for Fast Optimization of Reactive Force Fields." In: *Journal of Chemical Theory and Computation* 18.9 (Sept. 2022), pp. 5181–5194. ISSN: 1549-9618. DOI: 10.1021/acs.jctc.2c00363. (Visited on 07/25/2024).

[41] Mehmet Cagri Kaymak, Samuel S. Schoenholz, Ekin D. Cubuk, Kurt A. O'Hearn, Kenneth M. Merz Jr., and Hasan Metin Aktulga. „End-to-End Differentiable Reactive Molecular Dynamics Simulations Using JAX." In: *High Performance Computing*. Ed. by Abhinav Bhatele, Jeff Hammond, Marc Baboulin, and Carola Kruse. Cham: Springer Nature Switzerland, 2023, pp. 202–219. ISBN: 978-3-031-32041-5. DOI: 10.1007/978-3-031-32041-5_11.

[42] M. Kircher et al. „Recoil-Induced Asymmetry of Nondipole Molecular Frame Photoelectron Angular Distributions in the Hard X-Ray Regime." In: *Phys. Rev. Lett.* 123.24 (Dec. 2019), p. 243201. DOI: 10.1103/PhysRevLett.123.243201.

[43] W. Knox, J. Roth, F. Paulik, and A. Hershman. „Production of Carboxylic Acids and Esters." Pat. US3769329A. Oct. 1973. (Visited on 04/02/2024).

[44] K. Kreidi et al. „Photo- and Auger-Electron Recoil Induced Dynamics of Interatomic Coulombic Decay." In: *Phys. Rev. Lett.* 103.3 (July 2009), p. 033001. DOI: 10.1103/PhysRevLett.103.033001.

[45] E. Kukk et al. „Violation of the Franck-Condon Principle Due to Recoil Effects in High Energy Molecular Core-Level Photoionization." In: *Phys. Rev. Lett.* 95.13 (Sept. 2005), p. 133001. DOI: 10.1103/PhysRevLett.95.133001.

[46] E. Kukk et al. „Effects of Molecular Potential and Geometry on Atomic Core-Level Photoemission over an Extended Energy Range: The Case Study of the CO Molecule." In: *Phys. Rev. A* 88.3 (Sept. 2013), p. 033412. DOI: 10.1103/PhysRevA.88.033412.

[47] E. Kukk et al. „Photoelectron Recoil in CO in the X-Ray Region up to 7 keV." In: *Phys. Rev. A* 95.4 (Apr. 2017), p. 042509. DOI: 10.1103/PhysRevA.95.042509.

[48] E. Kukk et al. „Energy Transfer into Molecular Vibrations and Rotations by Recoil in Inner-Shell Photoemission." In: *Phys. Rev. Lett.* 121.7 (Aug. 2018), p. 073002. DOI: 10.1103/PhysRevLett.121.073002.

[49] Jochen Küpper et al. „X-Ray Diffraction from Isolated and Strongly Aligned Gas-Phase Molecules with a Free-Electron Laser." In: *Physical Review Letters* 112.8 (Feb. 28, 2014), p. 083002. DOI: 10.1103/PhysRevLett.112.083002. URL: https://link.aps.org/doi/10.1103/PhysRevLett.112.083002 (visited on 03/13/2024).

[50] David van Leeuwen. *GaussianMixtures*. URL: https://github.com/davidavdav/GaussianMixtures.jl (visited on 09/13/2023).

[51] F. Légaré, Kevin F. Lee, I. V. Litvinyuk, P. W. Dooley, A. D. Bandrauk, D. M. Villeneuve, and P. B. Corkum. „Imaging the time-dependent structure of a molecule as it undergoes dynamics." In: *Phys. Rev. A* 72.5 (Nov. 2005), p. 052717. ISSN: 1050-2947, 1094-1622. DOI: 10.1103/PhysRevA.72.052717. URL: http://link.aps.org/doi/10.1103/PhysRevA.72.052717 (visited on 01/19/2017).

[52] Wenjin Li and Ao Ma. „Recent Developments in Methods for Identifying Reaction Coordinates." In: *Molecular Simulation* 40.10-11 (Aug. 2014), pp. 784–793. ISSN: 0892-7022. DOI: 10.1080/08927022.2014.907898. (Visited on 08/04/2024).

[53] Xin-Zheng Li, Brent Walker, and Angelos Michaelides. „Quantum Nature of the Hydrogen Bond." In: *Proceedings of the National Academy of Sciences* 108.16 (Apr. 19, 2011), pp. 6369–6373. DOI: 10.1073/pnas.1016653108. URL: https://www.pnas.org/doi/10.1073/pnas.1016653108 (visited on 01/10/2024).

[54] Xusong Li, Deping Hu, Yu Xie, and Zhenggang Lan. „Analysis of Trajectory Similarity and Configuration Similarity in On-the-Fly Surface-Hopping Simulation on Multi-Channel Nonadiabatic Photoisomerization Dynamics." In: *The Journal of Chemical Physics* 149.24 (Dec. 26, 2018), p. 244104. ISSN: 0021-9606. DOI:

10.1063/1.5048049. URL: https://aip.scitation.org/doi/full/10.1063/1.5048049 (visited on 04/07/2021).

[55] Xusong Li, Yu Xie, Deping Hu, and Zhenggang Lan. „Analysis of the Geometrical Evolution in On-the-Fly Surface-Hopping Nonadiabatic Dynamics with Machine Learning Dimensionality Reduction Approaches: Classical Multidimensional Scaling and Isometric Feature Mapping." In: *Journal of Chemical Theory and Computation* 13.10 (Oct. 10, 2017), pp. 4611–4623. ISSN: 1549-9618. DOI: 10.1021/acs.jctc.7b00394. URL: https://doi.org/10.1021/acs.jctc.7b00394 (visited on 04/07/2021).

[56] Zheng Li et al. „Ultrafast Isomerization in Acetylene Dication after Carbon K-Shell Ionization." In: *Nature Communications* 8.1 (Sept. 2017), p. 453. ISSN: 2041-1723. DOI: 10.1038/s41467-017-00426-6.

[57] J. A. Littlechild. „Protein Crystallization: Magical or Logical: Can We Establish Some General Rules?" In: *Journal of Physics D: Applied Physics* 24.2 (Feb. 1991), p. 111. ISSN: 0022-3727. DOI: 10.1088/0022-3727/24/2/004. (Visited on 04/04/2024).

[58] F. Légaré et al. „Laser Coulomb-explosion Imaging of Small Molecules." In: *Physical Review A* 71.1 (Jan. 19, 2005), p. 013415. DOI: 10.1103/PhysRevA.71.013415. URL: https://link.aps.org/doi/10.1103/PhysRevA.71.013415 (visited on 04/23/2021).

[59] C. B. Madsen, L. B. Madsen, S. S. Viftrup, M. P. Johansson, T. B. Poulsen, L. Holmegaard, V. Kumarappan, K. A. Jørgensen, and H. Stapelfeldt. „A combined experimental and theoretical study on realizing and using laser controlled torsion of molecules." In: *J. Chem. Phys.* 130.23 (June 2009), p. 234310. ISSN: 0021-9606. DOI: 10.1063/1.3149789. URL: https://pubs.aip.org/aip/jcp/article/130/23/234310/924793/A-combined-experimental-and-theoretical-study-on.

[60] Qian Mao, Yihua Ren, K. H. Luo, and Adri C. T. van Duin. „Dynamics and Kinetics of Reversible Homo-Molecular Dimerization of Polycyclic Aromatic Hydrocarbons." In: *The Journal of Chemical Physics* 147.24 (Dec. 2017), p. 244305. ISSN: 0021-9606. DOI: 10.1063/1.5000534. (Visited on 07/26/2024).

[61] Juan Luis Martín-Espejo, Jesús Gandara-Loe, José Antonio Odriozola, T. R. Reina, and Laura Pastor-Pérez. „Sustainable Routes for Acetic Acid Production: Traditional Processes vs a Low-Carbon, Biogas-Based Strategy." In: *Science of The Total Environment* 840 (Sept. 2022), p. 156663. ISSN: 0048-9697. DOI: 10.1016/j.scitotenv.2022.156663. (Visited on 04/02/2024).

[62] Donald A. McQuarrie and John D. Simon. *Physical Chemistry : A Molecular Approach*. Sausalito, Calif.: University Science Books, 1997. ISBN: 0935702997 9780935702996.

[63] David R Miller and Giacinto Scoles. „Atomic and molecular beam methods.“ In: *Atomic and Molecular Beam Methods* 1 (1988), p. 14.

[64] Patrick Kofod Mogensen et al. *JuliaNLSolvers/Optim.Jl: V1.7.7*. Zenodo, Aug. 17, 2023. DOI: 10.5281/zenodo.8254057. URL: https://zenodo.org/record/8254057 (visited on 09/13/2023).

[65] Susanna Monti, Alessandro Corozzi, Peter Fristrup, Kaushik L. Joshi, Yun Kyung Shin, Peter Oelschlaeger, Adri C. T. van Duin, and Vincenzo Barone. „Exploring the Conformational and Reactive Dynamics of Biomolecules in Solution Using an Extended Version of the Glycine Reactive Force Field.“ In: *Physical Chemistry Chemical Physics* 15.36 (2013), pp. 15062–15077. DOI: 10.1039/C3CP51931G. (Visited on 07/26/2024).

[66] Robert S. Mulliken. „The Band Spectrum of Boron Monoxide.“ In: *Nature* 114.2862 (2862 Sept. 1924), pp. 349–350. ISSN: 1476-4687. DOI: 10.1038/114349a0. URL: https://www.nature.com/articles/114349a0 (visited on 07/27/2023).

[67] Raphael Mathias Peltzer, Jürgen Gauss, Odile Eisenstein, and Michele Cascella. „The Grignard Reaction – Unraveling a Chemical Puzzle.“ In: *Journal of the American Chemical Society* 142.6 (Feb. 2020), pp. 2984–2994. ISSN: 0002-7863. DOI: 10.1021/jacs.9b11829. (Visited on 07/30/2024).

[68] Matthias Post, Steffen Wolf, and Gerhard Stock. „Principal Component Analysis of Nonequilibrium Molecular Dynamics Simulations.“ In: *The Journal of Chemical Physics* 150.20 (May 28, 2019), p. 204110. ISSN: 0021-9606. DOI: 10.1063/1.5089636. URL: https://aip.scitation.org/doi/full/10.1063/1.5089636 (visited on 04/07/2021).

[69] Matthew S. Robinson and Jochen Küpper. „Unraveling the Ultrafast Dynamics of Thermal-Energy Chemical Reactions.“ In: *Physical Chemistry Chemical Physics* 26.3 (Jan. 2024), pp. 1587–1601. ISSN: 1463-9084. DOI: 10.1039/D3CP03954D. (Visited on 08/01/2024).

[70] Matthew Robinson and Nidin Vassery. *Private communication*. 2024.

[71] Michael F. Russo and Adri C. T. van Duin. „Atomistic-Scale Simulations of Chemical Reactions: Bridging from Quantum Chemistry to Engineering.“ In: *Nuclear Instruments and Methods in Physics Research Section B: Beam Interactions with Materials and Atoms*. Computer Simulations of Radiation Effects in

Solids 269.14 (July 2011), pp. 1549–1554. ISSN: 0168-583X. DOI: 10.1016/j.nimb.2010.12.053. (Visited on 07/25/2024).

[72] Hiroshi Ryufuku, Ken Sasaki, and Tsutomu Watanabe. „Oscillatory Behavior of Charge Transfer Cross Sections as a Function of the Charge of Projectiles in Low-Energy Collisions." In: *Physical Review A* 21.3 (Mar. 1, 1980), pp. 745–750. DOI: 10.1103/PhysRevA.21.745. URL: https://link.aps.org/doi/10.1103/PhysRevA.21.745 (visited on 07/12/2021).

[73] Marc Simon et al. „Atomic Auger Doppler Effects upon Emission of Fast Photoelectrons." In: *Nat. Commun.* 5.1 (June 2014), p. 4069. DOI: 10.1038/ncomms5069.

[74] Sang-Kil Son et al. XATOM—*an integrated toolkit for x-ray and atomic physics*. CFEL. DESY. Hamburg. Germany, 2023.

[75] Veenasangeeta Sortur, Jayashree Yenagi, and J. Tonannavar. „Vibrational assignments of 2-iodopyridine." In: *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy* 69.2 (2008), pp. 604–611. ISSN: 1386-1425. DOI: https://doi.org/10.1016/j.saa.2007.05.010. URL: https://www.sciencedirect.com/science/article/pii/S138614250700251X.

[76] Henrik Stapelfeldt, Eric Constant, Hirofumi Sakai, and Paul B. Corkum. „Time-Resolved Coulomb Explosion Imaging: A Method to Measure Structure and Dynamics of Molecular Nuclear Wave Packets." In: *Physical Review A* 58.1 (July 1998), pp. 426–433. DOI: 10.1103/PhysRevA.58.426.

[77] Sarah A. Mueller Stein, Anne E. Loccisano, Steven M. Firestine, and Jeffrey D. Evanseck. „Chapter 13 Principal Components Analysis: A Review of Its Application on Molecular Dynamics Data." In: *Annual Reports in Computational Chemistry*. Ed. by David C. Spellmeyer. Vol. 2. Elsevier, Jan. 1, 2006, pp. 233–261. DOI: 10.1016/S1574-1400(06)02013-5. URL: https://www.sciencedirect.com/science/article/pii/S1574140006020135 (visited on 05/17/2021).

[78] Marc H. Taylor, Martin Losch, Manfred Wenzel, and Jens Schröter. „On the Sensitivity of Field Reconstruction and Prediction Using Empirical Orthogonal Functions Derived from Gappy Data." In: *Journal of Climate* 26.22 (Nov. 15, 2013), pp. 9194–9205. ISSN: 0894-8755, 1520-0442. DOI: 10.1175/JCLI-D-13-00089.1. URL: https://journals.ametsoc.org/view/journals/clim/26/22/jcli-d-13-00089.1.xml (visited on 05/25/2021).

[79] T. D. Thomas et al. „Photoelectron-Recoil-Induced Rotational Excitation of the $B\phantom{\rule{0.2em}{0ex}}{2}\ensuremath{\Sigma}_{u}{+}$ State in $\mathrm{N}_{2}{}{+}$." In: *Phys. Rev. A* 79.2 (Feb. 2009), p. 022506. DOI: 10.1103/PhysRevA.79.022506.

[80] T. D. Thomas et al. „Valence Photoelectron Spectroscopy of N2 and CO: Recoil-Induced Rotational Excitation, Relative Intensities, and Atomic Orbital Composition of Molecular Orbitals.“ In: *J. Chem. Phys.* 133.17 (Nov. 2010), p. 174312. DOI: 10.1063/1.3503658.

[81] T. Darrah Thomas. „Effect of Recoil-Induced Internal Excitation on the Auger Transition Energy.“ In: *Phys. Rev. A* 101.4 (Apr. 2020), p. 043415. DOI: 10.1103/PhysRevA.101.043415.

[82] T. Darrah Thomas et al. „Recoil Excitation of Vibrational Structure in the Carbon 1s Photoelectron Spectrum of CF_4.“ In: *J. Chem. Phys.* 128.14 (Apr. 2008), p. 144311. DOI: 10.1063/1.2897756.

[83] J. Ullrich, R. Moshammer, A. Dorn, R. D. rner, L. Ph H. Schmidt, and H. Schmidt-B. cking. „Recoil-Ion and Electron Momentum Spectroscopy: Reaction-Microscopes.“ In: *Reports on Progress in Physics* 66.9 (Aug. 2003), pp. 1463–1545. ISSN: 0034-4885. DOI: 10.1088/0034-4885/66/9/203. URL: https://doi.org/10.1088/0034-4885/66/9/203 (visited on 09/22/2022).

[84] Z. Vager, E. P. Kanter, G. Both, P. J. Cooney, A. Faibis, W. Koenig, B. J. Zabransky, and D. Zajfman. „Direct Determination of the Stereochemical Structure of C${\mathrm{H}}_{4}^{+}$.“ In: *Physical Review Letters* 57.22 (Dec. 1986), pp. 2793–2795. DOI: 10.1103/PhysRevLett.57.2793.

[85] Matthias Vojta. „Quantum phase transitions.“ In: *Rep. Prog. Phys.* 66.12 (2003), p. 2069. DOI: 10.1088/0034-4885/66/12/R01. URL: https://dx.doi.org/10.1088/0034-4885/66/12/R01.

[86] Reppe Walter and Friederich Herbert. „Acetic Acid Anhydride.“ Pat. US2789137A. Apr. 1957. (Visited on 04/02/2024).

[87] Enliang Wang et al. *Time-Resolved Coulomb Explosion Imaging Unveils Ultrafast Ring Opening of Furan*. Nov. 8, 2023. DOI: 10.48550/arXiv.2311.05099. arXiv: 2311.05099 [physics]. URL: http://arxiv.org/abs/2311.05099 (visited on 03/18/2024). preprint.

[88] Frames White et al. *JuliaDiff/ChainRules.Jl: V1.54.0*. Zenodo, Sept. 4, 2023. DOI: 10.5281/zenodo.8316272. URL: https://zenodo.org/record/8316272 (visited on 09/13/2023).

[89] Joseph Ladislas Wiza et al. „Microchannel plate detectors.“ In: *Nucl. Instrum. Methods* 162.1-3 (1979), pp. 587–601.

[90] B. Wolter et al. „Ultrafast Electron Diffraction Imaging of Bond Breaking in Di-Ionized Acetylene.“ In: *Science* 354.6310 (Oct. 21, 2016), pp. 308–312. DOI: 10.1126/science.aah3429. URL: https://www.science.org/doi/10.1126/science.aah3429 (visited on 03/13/2024).

[91] Tomoyuki Yatsuhashi, Naoya Mitsubayashi, Masako Itsukashi, Masatoshi Kozaki, Keiji Okada, and Nobuaki Nakashima. „Persistence of Iodines and Deformation of Molecular Structure in Highly Charged Diiodoacetylene: Anisotropic Carbon Ion Emission.“ In: *ChemPhysChem* 12.1 (2011), pp. 122–126. ISSN: 1439-7641. DOI: 10.1002/cphc.201000756.

[92] Adri C. T. van Duin, Siddharth Dasgupta, Francois Lorant, and William A. Goddard. „ReaxFF: A Reactive Force Field for Hydrocarbons.“ In: *The Journal of Physical Chemistry A* 105.41 (Oct. 2001), pp. 9396–9409. ISSN: 1089-5639. DOI: 10.1021/jp004368u. (Visited on 07/25/2024).