THE UNIVERSITY *of* EDINBURGH

Ab Initio Molecular Diffraction

Thomas Northey

Doctor of Philosophy

The University of Edinburgh

2016

Declaration

I declare that this thesis was composed by myself and that the work contained therein is my own, except where explicitly stated otherwise in the text.

(*Thomas Northey*)

Acknowledgements

Firstly, I would like to express my sincere gratitude to my supervisor Dr. Adam Kirrander for the continuous support of my Ph.D. study and related research, for his patience, motivation, and immense knowledge. We were both first-timers to Edinburgh in 2012. His guidance helped me during the time of research and writing of this thesis. I could not have imagined having a better advisor and mentor for my Ph.D. study.

I would like to thank the other members of my group (past and present) for their help and companionship. They are Andrés Moreno-Carrascosa, Darren Bellshaw, Dr. Kenichiro Saita, Nikola Zotev and Dr. Maria Tudorovskaya. I also give special thanks to Dr. David Rogers at The University of Edinburgh, for helping a lot when I was a first year Ph.D. student learning the ropes of *ab initio* electronic structure calculations.

Last but not least, I would like to thank my girlfriend Alex, and my parents, brother and sister for supporting me throughout writing this thesis and in my life in general.

Papers

Work from this thesis has been published (or soon will be) in the following papers:

[1]. T. Northey, N. Zotev, and A. Kirrander. *Ab Initio* Calculation of Molecular Diffraction, *Journal of Chemical Theory and Computation*, 10(11):4911-4920, 2014.

[2]. T. Northey, and A. Kirrander. Fragment-based *Ab initio* X-ray diffraction from Biomolecules, (manuscript, 2017)

[3]. T. Northey, A. M. Carrascosa, S. Schäfer, and A. Kirrander. Elastic X-ray scattering from state-selected molecules. *The Journal of Chemical Physics*, 145(15):154304, 2016.

[4]. A. Moreno Carrascosa, T. Northey and A. Kirrander. Imaging rotations and vibrations in polyatomic molecules with X-ray scattering, (submitted, 2017).

Other papers not discussed in detail:

[5]. M. P. Minitti, J. M. Budarz, A. Kirrander, J. S. Robinson, D. Ratner, T. J. Lane, D. Zhu, J. M. Glownia, M. Kozina, H. T. Lemke, M. Sikorski, Y. Feng, S. Nelson, K. Saita, B. Stankus, T. Northey, J. B. Hastings, and P. M. Weber. Imaging Molecular Motion: Femtosecond X-Ray Scattering of an Electrocyclic Chemical Reaction, *Physical Review Letters*, 114, 255501, 2015.

[6]. M. P. Minitti, J. M. Budarz, A. Kirrander, J. Robinson, T. J. Lane, D. Ratner, K. Saita, T. Northey, B. Stankus, V. Cofer-Shabica, J. Hastings and P. M. Weber. Toward Structural Femtosecond Chemical Dynamics: Imaging Chemistry in Space and Time, *Faraday Discuss*, 171, 81-91, 2014.

Abstract

In 1915, Debye derived his well-known equation for the X-ray scattering from a sample of randomly orientated gas-phase molecules. He approximated the molecular scattering by adding the contributions of isolated atomic constituents. This is known as the Independent Atom Model (IAM). However, it omits the redistribution of valence electrons due to bonding, and is limited to the electronic ground state. The main proposition of this thesis is that it is worthwhile going beyond the IAM when interpreting X-ray scattering data. In part, this is motivated by the arrival of new X-ray sources called X-ray Free-Electron Lasers (XFELs).

A new method called *Ab Initio* X-ray Diffraction (AIXRD) is introduced. It calculates the elastic X-ray molecular scattering factor directly from wave functions calculated by *ab initio* electronic structure theory, for instance Hartree-Fock or multiconfigurational self-consistent field. In this way, the valence electrons are correctly taken into account, and calculations based on electronically excited wave functions become possible. The wave functions must be constructed from spatial orbitals made up of *Gaussian-Type Orbitals* (GTOs), giving an analytical solution to the Fourier transform integrals involved, and is key to computationally efficient and accurate results. This is compared to a fast Fourier transform (FFT) method, where the electron density is computed on a 3D grid and an FFT algorithm is used to obtain the elastic X-ray molecular scattering factor.

Inspired by post-crystallography experiments such as serial femtosecond crystallography and single-particle imaging at XFELs, the AIXRD method is expanded to allow accurate X-ray diffraction calculations from large molecules such as proteins. To make the underlying *ab initio* problem tractable, the molecule is split into fragments. In other words, the electron density is constructed by a sum of fragment contributions, as is the corresponding molecular form-factor. In this way, it is analogous to the IAM approach except that instead of isolated atoms, there are isolated fragments. A pairwise summation of fragment contributions is

also used to account for fragment-fragment interactions. Various fragment definitions are compared based on their effect on the X-ray diffraction signal, and are compared to the IAM method.

Finally, X-ray diffraction from molecules in specific quantum states is calculated, revealing a distinct quantum fingerprint in the X-ray diffraction, and a comparison to experiment is made. In particular, the elastic X-ray diffraction is calculated from gas-phase H_2 pumped to various electronic, vibrational, and electronic states. This is expanded upon for polyatomic molecules using the harmonic approximation for the vibrational states.

Lay Summary

X-rays interact with matter via absorption or scattering processes, predominantly via the electrons. This thesis is a theoretical and computational study of gas-phase elastic X-ray diffraction from molecules as small as H_2 to as large as proteins. It is motivated by the arrival of new X-ray sources called X-ray free-electron lasers (XFELs), capable of producing X-ray pulses with unprecedented intensity and very short duration. These characteristics are well suited for improvements in experiments such as gas-phase and time-resolved X-ray diffraction, and crystal-free structure determination of biomolecules.

A recurring theme throughout is the effectiveness of going beyond the conventional method of calculating the X-ray diffraction signal, called the Independent Atom Model (IAM), which treats all atoms in a molecule as isolated from one another, and therefore, does not correctly account for valence and bonding electrons. A new method called *Ab Initio* (Latin: 'from the beginning') X-ray Diffraction (AIXRD) is described, which uses the total molecular wave function to accurately calculate the X-ray diffraction signal. An outline of the underlying methodology is described within, including calculations of X-ray diffraction from molecules in electronically excited states.

This method is expanded upon in two ways. Firstly, to allow larger molecules such as proteins to be treated with the AIXRD method, a fragment-based approach is employed, and is compared to the full AIXRD and IAM methods. It is inspired by experimental developments at XFELs, such as *serial crystallography* and the possibility of single-particle imaging. Lastly, the effect of specific quantum states (electronic, vibrational, and rotational) in molecules is studied, encouraged by progress in experimental state-selection of gas-phase molecular samples, in combination with renewed interest in gas-phase X-ray scattering.

Contents

List of Abbreviations

XFEL	X-ray Free-Electron Laser
IAM	Independent Atom Model
SASE	Self-Amplified Spontaneous Emission
LCLS	Linac Coherent Light Source
XAS	X-ray Absorption Spectroscopy
XANES	X-ray Absorption Near-Edge Structure
EXAFS	Extended X-ray Absorption Fine Structure
XES	X-ray Emission Spectroscopy
TRXRD	Time-Resolved X-ray Diffraction
ESRF	European Synchrotron Radiation Facility
PES	Potential Energy Surface
CHD	1,3-Cyclohexadiene
AIXRD	*Ab Initio* X-ray Diffraction
HF	Hartree-Fock
DFT	Density Functional Theory
MCSCF	Multi-Configurational Self-Consistent Field
MD	Molecular Dynamics
GTO	Gaussian-Type Orbital
SCF	Self-Consistent Field
CI	Configuration Interaction
FFT	Fast Fourier Transform
MO	Molecular Orbital
AO	Atomic Orbital
MOF	Metal Organic Framework
FMO	Fragment Molecular Orbital
IFM	Independent Fragment Model
FT	Fourier Transform
PDB	Protein Data Bank
FF	Diphenylalanine
PEC	Potential Energy Curve

Chapter 1

Introduction

The subject of this thesis is elastic X-ray scattering from molecules, primarily in the gas-phase. It is motivated by the recent construction of a new class of X-ray sources, called X-ray Free-Electron Lasers (XFEL), which open the door for many experimental possibilities including gas-phase X-ray diffraction experiments. This thesis explores the utility of going beyond the conventional independent atom model (IAM) for interpreting elastic X-ray scattering and explicitly taking into account the electronic, vibrational, and rotational states of molecules. An approach is developed here called *ab initio* X-ray diffraction (AIXRD). It is based on the electronic wave function of the molecule, as calculated by modern quantum chemistry methods. This chapter provides a brief overview of the history of X-rays and their sources, their interaction with matter, and finally the experimental possibilities enabled by XFELs and a more detailed reasoning behind the development of AIXRD.

1.1 Brief historical background

X-rays are electromagnetic waves with wavelengths on the order of an Ångström (10^{-10} m), which is the length scale of chemical bonds in molecules. X-rays interact weakly with matter, meaning that they penetrate further through materials than, for instance, electrons. They were discovered in 1895 by Wilhelm Conrad Röntgen in his laboratory at the University of Würzburg, Germany. The first X-ray crystallography experiment determined the structure of rock salt in 1913 [7]. Since then X-rays have become an invaluable tool for probing the structure of

matter, ranging from atoms to proteins, crystals and solids. Their importance in this aspect has been truly staggering; complex molecular structures, such as DNA and proteins, have been solved by the use of X-ray scattering.

There have been many Nobel prizes awarded for work which included the use of X-rays, starting with Max von Laue in 1914 for his discovery of X-ray diffraction from crystals [8], while William Lawrence Bragg, discoverer of Bragg's law of diffraction, and his father William Henry Bragg obtained the 1915 Nobel Prize in Physics "*for their services in the analysis of crystal structure by means of X-ray*" [9]. X-ray diffraction was used for the determination of the structure of DNA by James Watson and Francis Crick in 1953, for which they were awarded the 1962 Nobel Prize in Physiology or Medicine together with Maurice Hugh Frederick Wilkins [10]. The same year the Nobel Prize in Chemistry was awarded jointly to Max Ferdinand Perutz and John Cowdery Kendrew for their studies of the structures of globular proteins [11]. Dorothy Hodgkin pioneered X-ray protein crystallography experiments and determined the structure of penicillin and vitamin B-12, for which she got the 1964 Nobel Prize in Chemistry [11]. Although not directly related to X-ray diffraction, Ahmed Zewail's 1999 Nobel Prize in Chemistry is also notable for his studies of transition states using femtosecond spectroscopy [12].

1.2 X-ray sources

1.2.1 Overview

Since the interaction of X-rays with matter is weak, experiments require a surplus of X-ray photons, making X-ray sources a critical bottleneck in experiments. The first mayor advance in this respect came in 1912 with Coolidge's X-ray tube, followed by the introduction of synchrotrons in the 1970s, and very recently the development of X-ray Free-Electron Lasers (XFELs). Each technological advance in X-ray sources has been followed by corresponding scientific advances.

To describe the comparative quality of various X-ray sources a quantity called brilliance is used, which is a combination of several beam characteristics. It is proportional to the number of photons emitted by the source per second, and inversely proportional to the beam collimation, which describes how much the

beam spreads out as it propagates, measured in units of horizontal and vertical milli-radians (mrad^2); how large the source is in units of mm^2, in other words, how well it can be focused to a small area; and finally, the spectral distribution. This is conventionally defined as the photon energy range as a fixed relative energy bandwidth (0.1% BW). The source brilliance can be expressed by the equation,

$$\text{Brilliance} = \frac{\text{Photons/second}}{(\text{mrad})^2(\text{mm}^2 \text{ source area})(0.1\% \text{ BW})}. \tag{1.1}$$

The brilliance of synchrotron and XFEL sources is shown in Fig. 1.1, and these sources are discussed in Sections 1.2.2 and 1.2.3 below. The Coolidge X-ray tube does not even register on the scale of Fig. 1.1. X-rays can be produced by accelerating electrons into a metal anode, where the sudden deceleration produces electromagnetic radiation This is consequently known as bremsstrahlung (after the German 'bremsen' for brake), and has a maximum energy proportional to the acceleration voltage applied. A sharp line spectrum is superimposed on this broad spectrum. The reason for this is that an incident photon can collide with an inner shell atomic electron and remove it, which causes a vacancy followed by relaxation via an outer shell electron, producing X-ray radiation with characteristic energy equal to the energy difference between the two shells. This is the fluorescent radiation, and is several orders of magnitude more intense than the bremsstrahlung spectrum. The inner, middle, and outer shells are conventionally labelled K, L, and M respectively. The X-ray spectrum produced by relaxation from the L to the K shell is called the K_α line, and the X-ray spectrum produced by relaxation from the M to the K shell is called the K_β line.

1.2.2 Synchrotrons

X-ray sources have come a long way since Coolidge's invention of the X-ray tube in 1912, and the subsequent rotating anode source. Nowadays, there are synchrotrons all over the world, which produce X-rays $\sim 10^{12}$ times more brilliant than the original lab-based sources. These facilities are routinely used to solve the structures of large molecules such as proteins and other biomolecules via X-ray crystallography. As of 28 January 2014, there are 97,362 protein structures in the protein data bank which were predominantly discovered in this way [14].

It is interesting to note that synchrotron radiation occurs naturally and has been observed near stellar nebulae where high magnetic fields cause relativistic charged

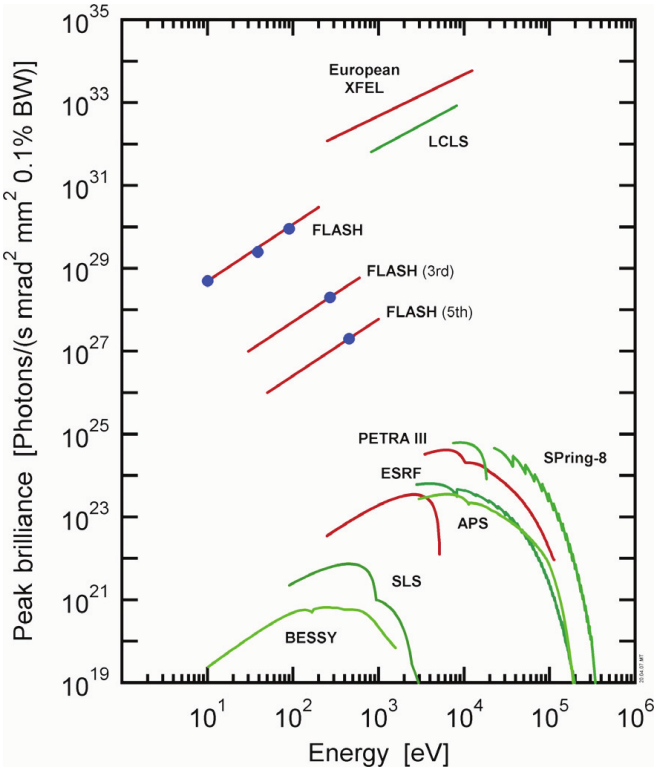


Figure 1.1: Peak brilliance and photon energy comparison between X-ray free electron laser and synchrotron radiation. Figure taken from Ref. [13].

particles to follow curved paths or orbits, producing X-ray radiation. Back on Earth, the same mechanism produces radiation in bending magnets, wigglers, and undulators in synchrotrons. An electron moving through a constant magnetic field experiences a Lorentz force and moves in a curved path. The radial acceleration causes electromagnetic radiation. In a storage ring the synchrotron radiation is produced either in the bending magnets needed to keep the electrons in a closed circuit, or in insertion devices such as wigglers or undulators situated in the straight sections of the storage ring. In the latter devices, magnets with alternating polarity force the electrons to follow oscillating paths rather than moving in a straight line. The undulator is a specifically designed wiggler, constructed such that the radiation from one oscillation is in phase with the radiation from subsequent oscillations. This gives coherent addition of amplitudes and a monochromatic spectrum (with harmonics). However, due to the finite number of periods in the undulator, a quasi-monochromaticity is achieved with typical BW of around 1%.

1.2.3 X-ray free-electron lasers

While synchrotrons play an important and ongoing role as an established X-ray source, recent developments have given rise to a new source called the X-ray Free-Electron Laser (XFEL). Originally proposed by Madey in 1971 [15], this ultra-bright X-ray source is essentially a repurposed linear accelerator, fitted with long ($\sim$100 m) undulators. It distinctly relies upon a process called self-amplified stimulated emission (SASE) [16,17], allowing Lorentz contracted electron bunches accelerated to near light speed to lase in the X-ray regime while passing through the undulators. As electron bunches traverse the long undulators they co-propagate with their own undulator radiation, and form microbunches which radiate coherently. See Fig. 1.2 for a schematic of this process. This builds up exponentially to a saturation point (self-amplification), giving extremely bright and coherent X-ray pulses. For this reason XFELs produce radiation of far higher brilliance, and transverse and longitudinal coherence than synchrotron radiation.

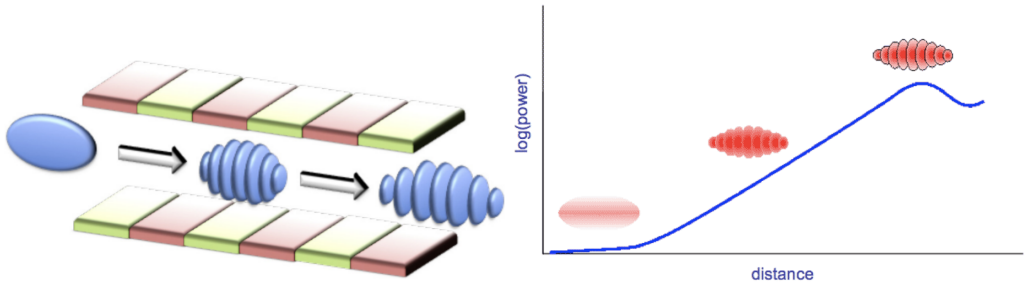


Figure 1.2: Schematic of microbunching: As the electron bunches traverse the long undulators they co-propagate with their own undulator radiation, and form microbunches which radiate coherently. Figure (a) taken from Ref. [18]. There is an ideal undulator length at which the coherence and power is at a maximum, after which space-charge repulsion effects take a role. Figure (b) taken from the Paul Scherrer Institute website, `https://www.psi.ch/swissfel/how-it-works`.

The first soft X-ray free-electron laser was FLASH [19] in Hamburg, which opened in 2005, and delivers radiation in the wavelength range 44-5 nm (28-250 eV). There are now a number of other facilities operating in the soft X-ray, VUV and XUV ranges including FERMI [20] in Italy and Artemis [21] in the UK. The original hard XFEL was the Linac Coherent Light Source (LCLS) in Stanford which first lased in 2009 [22], closely followed in 2011 by SACLA in Japan [23]. Other facilities that are recently completed or near completion include the European XFEL [24] in Hamburg, Korean XFEL [25], and SwissFEL [26] at the Paul Scherrer Institute in Switzerland.

They all have unprecedented peak intensity of up to nine orders of magnitude beyond that of 3rd generation synchrotrons. See again Fig. 1.1 for a comparison between XFEL and synchrotron peak brilliance and photon energy. All XFELs produce ultrashort pulses on the order of femtoseconds, and have highly tunable wavelengths, with their technology covering an enormous bandwidth from the VUV to the hard X-ray regime. The LCLS is capable of producing hard X-ray pulses of < 7 fs duration, with < 25 fs pulses generated routinely [27]. It is interesting to note that the total XFEL photon flux is comparable to that of modern synchrotrons. However, the photons come in highly concentrated pulses with a low repetition rate (for XFELs based on normal conducting cavities this is typically about 120 Hz), leading to much higher intensity. Upcoming XFEL projects which are based on superconductive cavities, such as the European XFEL, are designed to operate at up to 27 kHz, increasing the total flux significantly. The short and intense pulses allow detection of very dilute samples such as cold molecular ions or beams, and aid background suppression. For general interest and a popular description of XFELs, their history and science (with a particular emphasis on the LCLS), see the Scientific American article by Berrah and Bucksbaum in Ref. [28].

1.3 Interaction of X-rays with matter

X-rays can interact with matter in a number of ways. The interaction Hamiltonian for an incident X-ray photon field and an electron in a sample has both an absorption and a scattering term [29],

$$\hat{H}_I = \mathbf{A} \cdot \mathbf{p} + \frac{\mathbf{A}^2}{2}, \tag{1.2}$$

in atomic units, where $\mathbf{A}$ is the vector potential, and $\mathbf{p}$ is the momentum of the electron. Photoelectric absorption, corresponding to the first term in Eq. (1.2), involves the destruction of an incident photon upon interaction with an atom, which for X-rays leads to the ejection of an electron from the atom. If the hole is created in the atom's K shell, this can be filled by an electron from the L or M shell, leading to emission of K_α or K_β radiation respectively. This is called X-ray emission. Additionally, a secondary electron can be emitted, called an Auger electron. This is caused by an outer shell electron receiving energy from the core-hole relaxation. There are various well-known experimental techniques which measure

X-ray emission, or use the X-ray absorption process in some other way. Examples of these methods include X-ray Absorption Spectroscopies (XAS), such as X-ray Absorption Near-Edge Structure (XANES), Extended X-ray Absorption Fine Structure (EXAFS); and resonant and non-resonant X-ray Emission Spectroscopies (XES). X-ray spectroscopies will benefit greatly from XFELs, but are not discussed in this thesis. For an overview of recent experimental and theoretical developments in ultrafast X-ray absorption spectroscopies see Ref. [18].

X-rays can also interact with matter via the scattering term in the interaction Hamiltonian, i.e. the second term in Eq. (1.2). This involves destruction of the incident photon and creation of another photon. If the new photon has exactly the same energy as the incident photon, this is called Thomson or elastic scattering. The photon can also deposit energy onto the sample, which results in a scattered photon with a longer wavelength. This is known as inelastic or Compton scattering. Elastic and inelastic scattering can be treated within the same theoretical framework, as shall be discussed further in Section 2.1. The focus in this thesis is on elastic scattering which provides useful information about molecular structure.

1.4 Opportunities at XFELs using scattering

1.4.1 Towards crystal-free structure determination

In standard synchrotron experiments, radiation damage can be a major obstacle, and protein crystals are routinely kept at cryogenic temperatures ($\sim$ 100 K) to minimise this damage. Prior to their construction, this was a concern for future XFEL sources, as their extremely large radiation dosage per pulse would destroy the sample. However, in an influential paper Neutze et al. [30] theorised that the sample would not degrade on the timescale of the ultrashort XFEL pulses. On the contrary, it would provide a meaningful snapshot of the scattering signal from the undamaged molecule. Of course, the sample would eventually vaporise into plasma, and require replacement of the sample after each snapshot. This situation is known as diffraction before destruction, and has now been confirmed by various experiments. However, inelastic photon-electron collisions can cause electronic damage, predominantly photoionisation, and to a lesser extent Auger emission, at a timescale comparable to the XFEL pulses. Careful tuning of pulse length (1-10 fs) and fluence (10^{13}-10^{15} photons/μm^2) mostly avoids or outruns

even these effects [31, 32], giving a high signal to noise ratio.

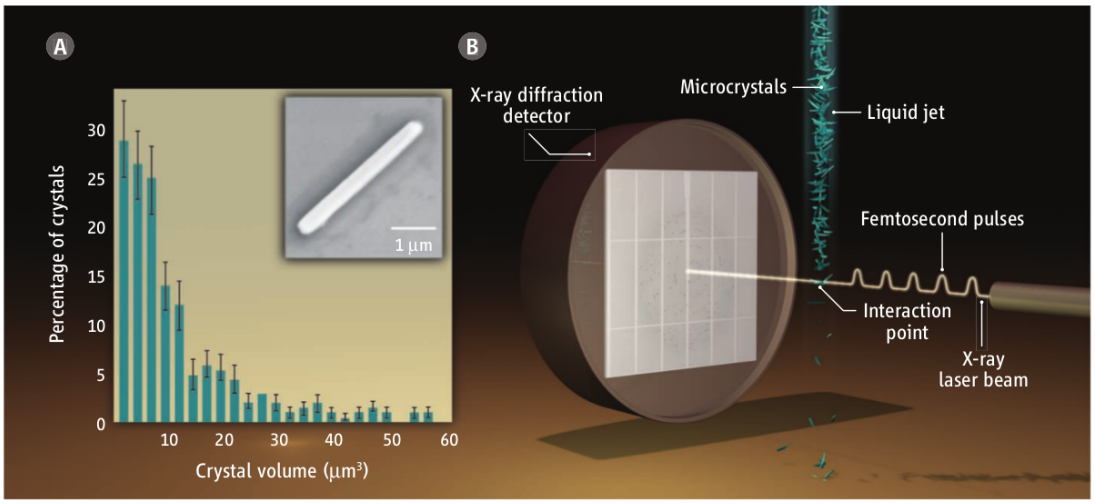


Figure 1.3: An illustration of serial crystallography. (A) Typical in-vivo crystal used in the study. Most crystals used by Redecke et al. [33] had a volume <10 μm^3. (B) Schematic of the femtosecond X-ray crystallography technique, with a liquid jet containing microcrystals, femtosecond X-ray pulses, and detector. Figure taken from Ref. [34].

A structure determination method of increasing importance is called serial femtosecond X-ray crystallography. It utilises the unprecedented peak brilliance of femtosecond pulses at XFEL sources to bypass the usual requirement of growing large crystals and irradiating them with X-rays for long periods of time. Instead, many microcrystals are passed through the XFEL beam in a liquid jet, where each brief irradiation by an X-ray pulse outruns the radiation damage and captures high quality diffraction data which is added to a cumulative dataset, used to solve the molecular structure [14, 33, 34]. The well-known protein lysozyme has been solved by this method with no prior structural information [35], and even single-particle imaging of a mimivirus [36] has been performed, showing no measurable radiation damage. Fig. 1.3 is a schematic for the crystal size data in Redecke et al. [33] and the experimental setup, showing the liquid jet containing microcrystals, the femtosecond X-ray pulses, and the detector. As the technology improves and access to XFEL beam-time becomes easier, it can be expected that more structures will be solved in this way; thus avoiding data deterioration by radiation damage, and crucially, the need for large crystal growth. This is especially important for proteins which are difficult or impossible to crystallise, such as membrane proteins, many of which are important drug targets [37]. New insights into their natural structure, including structures of such proteins encoded by the human genome, could soon be obtained using serial femtosecond crystallography [14].

1.4.2 Gas-phase scattering

Gas-phase scattering is different from scattering from crystalline samples, as there are no Bragg peaks, only a continuous diffraction intensity. In X-ray crystallography, only the Bragg peaks are useful, as the diffraction signal arriving at these points is strongly amplified by coherent summation of the signal. This signal amplification is useful as it mitigates the small scattering cross-section of X-rays, and has been a requirement for many structure determination experiments. In the gas-phase this amplification is absent and X-ray scattering from rotationally-averaged molecules is distributed continuously in concentric rings according to the Debye formula [38].

Historically, between X-ray and electron scattering, the latter has been thought of as the preferred method of probing the molecular structure of gas-phase samples [39]. This is due to two main reasons: The differential cross-section for electrons is much larger than for X-rays; and lab-based electron sources with high enough flux have been developed, whereas X-ray sources such as Coolidge tubes and rotating anodes did not meet flux requirements to be easily used for this purpose. In 1915, Debye theoretically examined the scattering of X-rays from molecules in the gas-phase, and wrote his famous equation $\sin(qr/qr)$ [38], where r is an internuclear distance, and $q = 2k_0 \sin(\theta/2)$, with k_0 the incident wavevector and θ the radial scattering angle. His proposed experiment was realised 13 years later by X-ray scattering from carbon tetrafluoride [40]. There were practical difficulties inherent in the weak scattering of the gaseous sample by X-rays. Later the apparatus was improved [41], allowing X-ray scattering from atoms and molecules such as H_2, He, N_2, O_2, Ar, and CO_2 [42] to be measured, with further investigations by Wollan [43], and Chipman and Jennings [44].

However, already by the 1930s electron scattering was seen as the preferred technique in the gas-phase [45]. It is therefore not surprising that electron scattering dominated the field of ultrafast time-resolved diffraction in the beginning [46,47]. Unfortunately time-resolved electron scattering has inherent problems of its own, as ultrashort electron bunches are difficult to generate due to space-charge repulsion effects, and accurate timing can be difficult to achieve using an optical pump and electron probe pulse due to different propagation velocities. Therefore now, there is renewed interest in gas-phase X-ray diffraction due to the short pulse durations and tremendous increase in the number of photons per pulse at XFEL sources, thus, the promise of ultrafast time-resolved X-ray diffraction [5,6,48,49].

In parallel, there is a surge in gas-phase X-ray scattering experiments at synchrotrons [50–53] which in part inspired the work in Chapter 5.

1.4.3 Time-resolved experiments

Early time resolved X-ray diffraction (TRXRD) experiments were performed at synchrotrons, where a mechanical chopper was fitted along the X-ray beam giving ∼100 ps X-ray pulses. In 2001, Techert et al. [54] probed transient structural changes in an organic solid at the European Synchrotron Radiation Facility (ESRF), in Grenoble, France. They managed to achieve 100 ps X-ray pulse widths using a synchronised chopper; in a pump-probe scheme, photoexciting their organic crystal with a Ti:sapphire laser and subsequently probing with X-ray pulses of variable time-delay. They imaged the structural changes in the organic crystal caused by electronic excitation and relaxation back to the ground state. Later that year, Neutze et al. [30] probed a photoexcited state of molecular iodine in solution with a similar technique and were able to visualise via diffraction using 80 ps X-ray pulses how the solvent cage altered the photodissociation dynamics. Similar TRXRD work at synchrotrons followed in this direction [55–57]. This type of experiment sets itself apart from spectroscopic techniques as it directly probes the time-evolving molecular structure, rather than energy spectra. TRXRD and spectroscopy are complimentary techniques because with information from both types of experiment a full picture can be obtained in terms of time-evolution on potential energy surfaces (PES) and changes in molecular structure [58].

The extreme brevity of the pulses at XFELs are ideal for this type of time-dependent study. The combination of extremely bright and ultrashort pulses is ideal for imaging molecules, as the length and time-scales for molecular motion are Ångströms (C-C single bonds are 1.2-1.5 Å) and femtoseconds (the shortest vibrational periods of molecules are ∼10 fs) respectively, which correspond very well to XFEL pulse characteristics. The unprecedented qualities of XFEL radiation expedite new and challenging areas in the field of TRXRD [59, 60].

In February 2014, the Kirrander group was fortunate enough to participate in an experiment at the LCLS, which imaged the photochemical ring-opening reaction of the 1,3-cyclohexadiene (CHD) molecule in the gas-phase using 30 fs X-ray pulses [5]. See Fig. 1.4 for the experimental schematic taken from Ref. [5]. The reaction proceeds extremely rapidly with a time constant of about 80 fs.

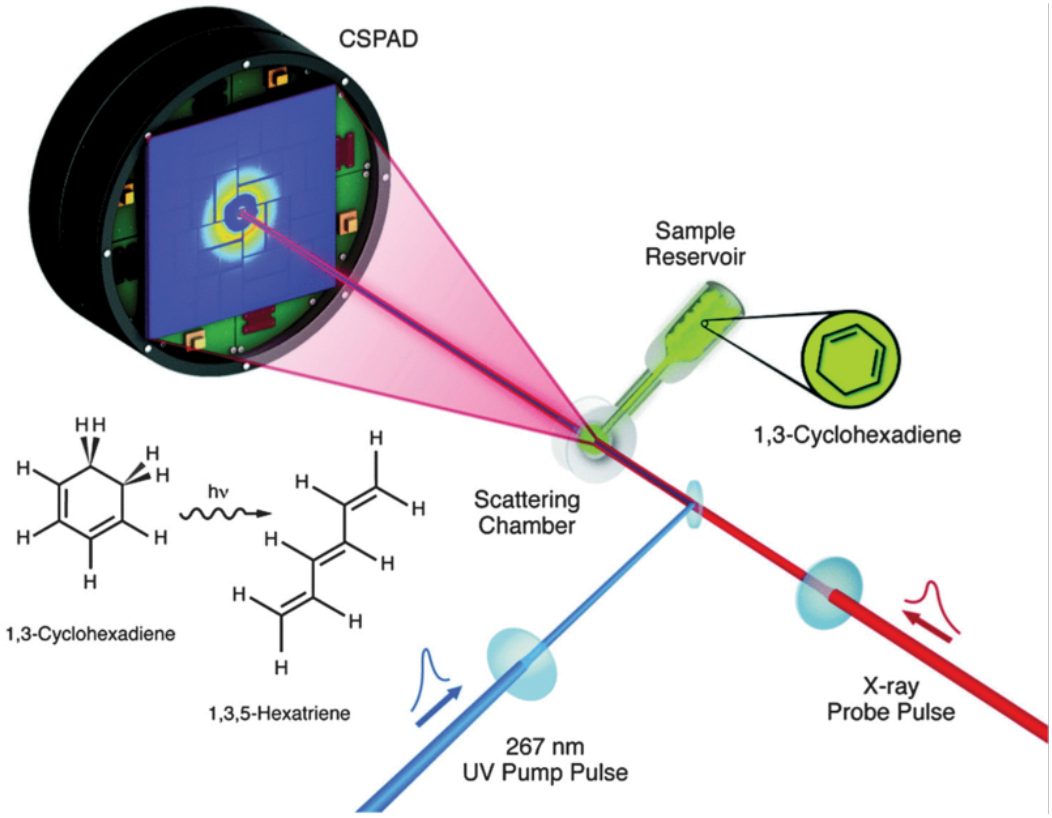


Figure 1.4: Schematic of gas-phase photochemical 1,3-cyclohexadiene reaction to 1,3,5-hexatriene; pumped with UV pulses and probed with X-ray pulses, with diffraction data recorded at the CSPAD detector. As performed by Minitti et al. [5]. Figure taken from this reference also.

After photoexcitation from its ground state (1A) to the Franck-Condon region on its first excited state (1B), the molecule propagates onto another excited state (2A), then to a conical intersection with the ground state, and has approximately 50% probability to undergo symmetry breaking and ring-opening to form 1,3,5-hexatriene. See Fig. 1.5 for a schematic of the reaction. In this experiment X-rays were used to image the average time-evolving molecular geometry during a photochemical reaction of a small molecule in the gas-phase at an ultrashort timescale. It strongly supports the perception that the XFEL is an extremely powerful imaging tool for molecules. As improvements are made in this field and XFELs are upgraded even further (e.g. LCLS II [61]), time-resolved imaging of molecules will become more common and hopefully new and deeper understanding will be gained about the dynamic processes of chemical reactions.

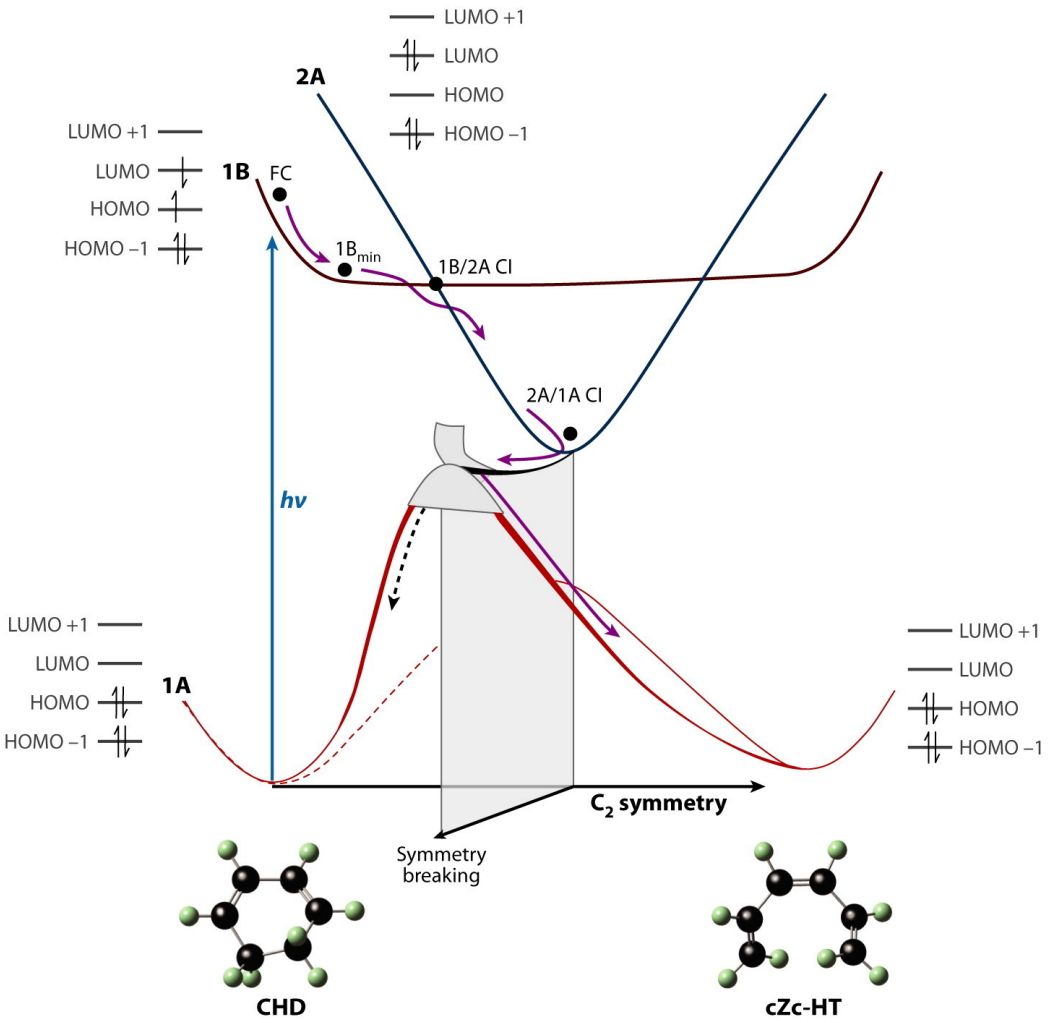


Figure 1.5: Reaction pathway of the photochemical 1,3-cyclohexadiene (CHD) ring-opening reaction to 1,3,5-hexatriene. After photoexcitation from the stationary point on the 1A potential energy surface (PES) to the Franck-Condon region on the 1B PES, the molecule passes through the conical intersection (1B/2A CI), then reaches the conical intersection with the ground state (2A/1A CI) where it has significant probability to symmetry break and undergo ring-opening. Figure taken from Ref. [6].

1.5 Beyond the Independent Atom Model

In X-ray diffraction theory, a simple and commonly used approximation is the independent atom model (IAM). It states that the molecular scattering factor can be approximated by a sum of atomic scattering factors or *form-factors*. Tables of atomic form-factors are listed in the International Tables of Crystallography [62], in which they were calculated using Hartree-Fock level theory [63, 64]. The IAM approximation is equivalent to saying that a sum of isolated-atom electron densities can be summed to approximate the total molecular electron density. Clearly, this omits alterations in the electron density caused by valence electrons

involved in chemical bonding, and is also restricted to the electronic ground state.

One of the main arguments in this thesis is that going beyond the IAM approximation is useful and significant, specifically by deriving the molecular form-factor from the full electron density. This argument is discussed including a description of theoretical calculations and results in Paper 1 (also reference [1]), and in Chapter 3 of this thesis. For example TRXRD experiments involving short-lived electronically excited states may particularly require molecular form-factors computed from electronically excited states. The question of whether X-ray scattering can be used to identify the electronic state of a molecule has been considered theoretically [65, 66]; and X-ray diffraction from exceptionally long-lived excited states in molecules has been observed experimentally [67, 68], although indirectly via changes in geometry. From theoretical results shown in this thesis (Chapters 3 and 5), it appears that electronically excited states could be visible via X-ray diffraction with ultrashort high intensity pulses.

If there is a degree of molecular alignment, the molecular form-factor will be a function of the radial and azimuthal scattering angles, dramatically increasing the information content in the observed scattering signal. Molecular alignment techniques have been developed over the past decades [69, 70], as well as techniques for state-selection [71–74], which have recently been demonstrated in both X-ray [48] and electron [75] gas-phase scattering. In principle, much richer information can be deduced from such non-thermal samples [76].

Using an *ab initio* electronic structure method in combination with X-ray diffraction theory gives rise to a new methodology called *ab initio* X-ray diffraction (AIXRD) to accurately compute molecular form-factors, and is described in detail in Chapter 3. It is not limited to the electronic ground state, as excited state wave functions obtained via *ab initio* electronic structure calculations can be used; and crucially, changes in valence electron distribution caused by chemical bonding are fully taken into account.

1.6 Thesis overview

This thesis is a theoretical study of molecular X-ray diffraction in the context of gas-phase experiments at the new ultrabright femtosecond XFEL sources. It primarily introduces an analytical methodology for calculating elastic X-ray

molecular scattering factors based on *ab initio* electronic structure theory using Gaussian basis sets, entitled *ab initio* X-ray diffraction (AIXRD), and compares this method to the simple IAM approximation. The strength of AIXRD lies in its ability to efficiently calculate scattering factors from molecules in any of their electronically excited states as well as their ground state, and by explicitly taking into account changes in electronic structure due to chemical bonding. This method is discussed in detail in Chapter 3.

With the advent of crystal-free structure determination experiments performed at XFELs, such as serial femtosecond X-ray crystallography [14,34–36], the AIXRD method is expanded to calculate elastic X-ray molecular scattering factors from large molecules, such as proteins and other biomolecules. This method involves dividing the molecule into appropriate fragments which are treated separately with electronic structure theory followed by AIXRD. This allows *ab initio* treatment of molecules with thousands of electrons. A comprehensive comparison between the fragment-based AIXRD method and the IAM method is performed for the well-known protein lysozyme, including temporal and orientational averaging as would occur over many shots in serial crystallography. This work is the topic of Chapter 4.

The AIXRD method is expanded upon by adding vibrational and rotational states, which is discussed in Chapters 5 and 6. A comparison to a recent high-resolution elastic X-ray diffraction experiment on H_2 in the gas-phase [50] is made. The motivation for this work is that the feasibility to detect, identify, and characterise individual quantum states using elastic X-ray scattering is at hand, whether electronic [1, 65, 66], vibrational [77], or rotational [78], as well as aligned molecules [79, 80]. Experimental realisations to date include scattering from partially aligned and somewhat state-selected molecules [75, 81, 82] and indirect detection of metastable electronically excited states via changes in geometry [67, 68]. Due to increasing sophistication of gas-phase X-ray diffraction experiments performed at XFELs [5, 6], combined with the rapid development of experimental molecular orientation and alignment [71, 73, 83], it may become possible to directly image state-selected molecules via X-ray scattering in the not too distant future.

Chapter 2

Theory

The theoretical aspects of this thesis primarily constitute a combination of X-ray scattering theory and *ab initio* electronic structure theory, to construct a *beyond IAM* method (as briefly mentioned in Section 1.5) for elastic X-ray scattering from molecules in the gas-phase. This chapter covers the fundamental expressions and methods involved along with brief discussions. Chapter 3 expands on this with further details on calculating the X-ray molecular form-factor from a wave function calculated by *ab initio* electronic structure theory.

2.1 X-ray scattering

X-rays only interact with the electrons in a molecule. Although the nuclei also carry charge, they are much heavier than the electrons and do not undergo substantial oscillations in response to the high frequency electromagnetic field. Therefore, X-ray scattering probes the spatial positions of the nuclei in an indirect manner, via the electrons associated with the nuclei. This thesis primarily focuses on elastic X-ray scattering from molecules in the gas-phase, with the main theoretical points discussed in the following. For a more general overview of the interactions of X-rays with matter, including scattering, see for instance Refs. [29] and [84].

2.1.1 Elastic scattering from electron density $\rho(\mathbf{r})$

Throughout this thesis the X-ray source is assumed to be monochromatic, which is valid in most modern X-ray scattering experiments at synchrotrons and XFELs; and the incident electric field can be considered as a plane wave,

$$\mathbf{E}(\mathbf{r}, t) = \mathbf{E}_0 e^{i(\mathbf{kr}-\omega t)}, \tag{2.1}$$

where $\mathbf{E}_0 = \hat{\varepsilon} E_0$ for unit vector $\hat{\varepsilon}$ and magnitude E_0, $|\mathbf{k}| = 2\pi/\lambda$ for X-ray wavelength λ, and $\omega = 2\pi/\tau = 2\pi\nu$ for wave period τ and frequency ν. For a non-monochromatic source there is a frequency distribution of incident photons, $E_0 = E_0(\omega)$. This is the case for an XFEL source, however, using the average photon energy $E_0 = \langle E_0 \rangle$ is a valid assumption as the bandwidth is small compared to the absolute photon energy. Elastic scattering from a general atomic or molecular electron density, $\rho(\mathbf{r})$, will now be considered. In a Cartesian coordinate system with the direction of the incident wavevector along the z-axis,

$$\mathbf{k}_0 = |\mathbf{k}| \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}, \tag{2.2}$$

and the scattered wavevector will depend on polar θ and azimuthal ϕ angles,

$$\mathbf{k} = |\mathbf{k}| \begin{pmatrix} \sin\theta\cos\phi \\ \sin\theta\sin\phi \\ \cos\theta \end{pmatrix}. \tag{2.3}$$

The incident and scattered wavevectors are shown in this coordinate system in Fig. 2.1. In a purely classical description, the radiation scattered by an atom or molecule is a superposition of different volume element contributions of the electron density, $\rho(\mathbf{r})$. This superposition depends on the phase of the incident wave as it interacts with the volume element at the origin and at the position $\mathbf{r}$. The phase difference $\Delta\phi$ between these two volume elements is the vector projection of $\mathbf{r}$ on $\mathbf{k}_0$ multiplied by $|\mathbf{k}_0|$, which is simply $\mathbf{r} \cdot \mathbf{k}_0$, minus the vector projection of $\mathbf{r}$ on $\mathbf{k}$ multiplied by $|\mathbf{k}|$, i.e. $\mathbf{r} \cdot \mathbf{k}$. The schematic of this is shown in Fig. 2.2. The momentum transfer or scattering vector can now be defined as,

$$\mathbf{q} \equiv \mathbf{k}_0 - \mathbf{k}, \tag{2.4}$$

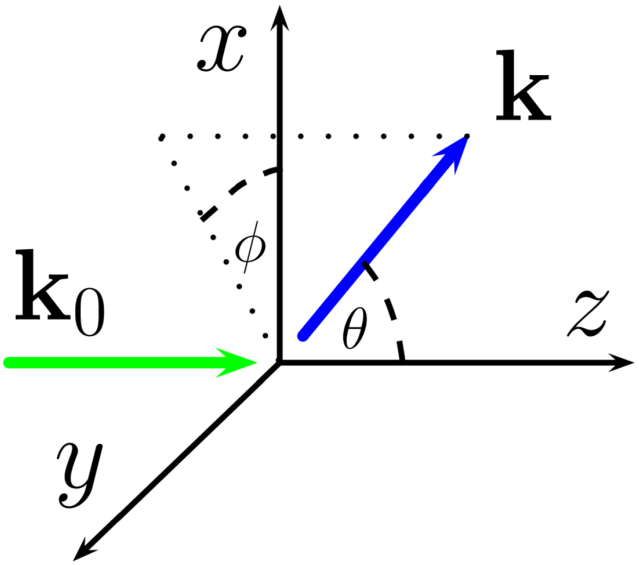


Figure 2.1: 3D kinematic scattering of the wave vector $\mathbf{k}_0$ (green) to scattered wave vector $\mathbf{k}$ (blue), where $|\mathbf{k}_0| = |\mathbf{k}|$. The incoming wave vector $\mathbf{k}_0 = (0, 0, k_z)$ is aligned with the z-axis, and the scattered wave vector is $\mathbf{k} = |\mathbf{k}_0|(\sin\theta\cos\phi, \sin\theta\sin\phi, \cos\theta)$, defined in terms of the radial (azimuthal) and angular (polar) scattering angles (θ, ϕ) both of which are defined relative to the direction of the incoming wave vector $\mathbf{k}_0$.

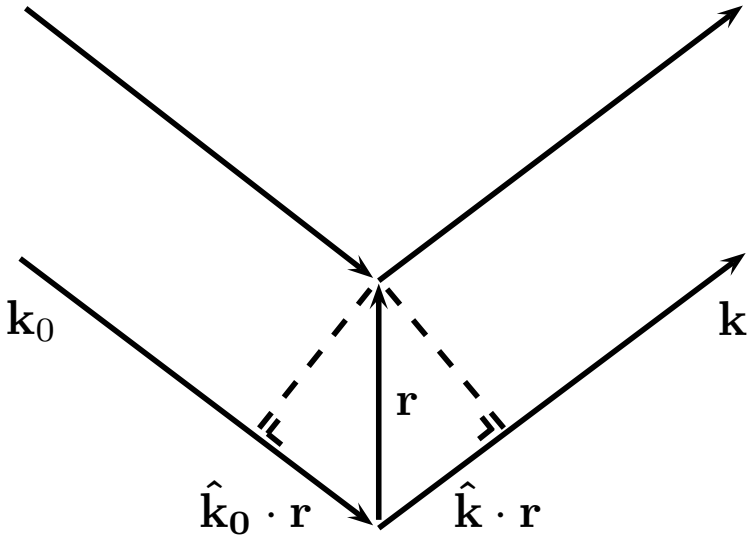


Figure 2.2: Schematic of the phase difference between the wavevector scattering at the origin and the wavevector scattering at position $\mathbf{r}$, which depends on the vector projection of $\mathbf{r}$ on the incident wavevector $\mathbf{k}_0$ and on the scattered wavevector $\mathbf{k}$.

which gives a phase difference of $\Delta\varphi = (\mathbf{k}_0 - \mathbf{k})\cdot\mathbf{r} = \mathbf{q}\cdot\mathbf{r}$. The minus sign appears because there is a phase shift of π upon scattering. The reason for this is that the total amplitude must go to zero at the scattering point, i.e. for an incident wave described as $\sin(kz)$, the scattered wave must be $\sin(kz + \pi) = -\sin(kz)$.

The scattering is elastic with $|\mathbf{k}_0| = |\mathbf{k}|$, and thus $|\mathbf{q}| = 2|\mathbf{k}|\sin(\theta/2)$ for radial scattering angle θ. From Eqs. (2.2) and (2.3), the scattering vector in 3D is,

$$\mathbf{q} = |\mathbf{k}| \begin{pmatrix} -\sin\theta\cos\phi \\ -\sin\theta\sin\phi \\ 1-\cos\theta \end{pmatrix}, \tag{2.5}$$

which can also be written as,

$$\mathbf{q} = 2|\mathbf{k}|\sin(\theta/2) \begin{pmatrix} -\cos(\theta/2)\cos\phi \\ -\cos(\theta/2)\sin\phi \\ \sin(\theta/2) \end{pmatrix}. \tag{2.6}$$

The superposition of volume element contributions to the scattering can now be constructed. A volume element $\mathrm{d}\mathbf{r}$ contributes an amount $-r_0\rho(\mathbf{r})\mathrm{d}\mathbf{r}$ with a phase factor $e^{i\mathbf{q}\cdot\mathbf{r}}$, where $r_0 = e^2/4\pi\epsilon_0 mc^2 \approx 2.82$ fm is the Thomson scattering length or classical radius of the electron. Thus, the total scattering length of the atom or molecule is,

$$-r_0 \int \rho(\mathbf{r}) e^{i\mathbf{q}\cdot\mathbf{r}} \mathrm{d}\mathbf{r} = -r_0 f(\mathbf{q}), \tag{2.7}$$

where $f(\mathbf{q})$ is the atomic or molecular form-factor. Note that the integral in Eq. (2.7) is a Fourier transform, and that $\mathbf{q} = 0$ corresponds to the exactly forward scattering case ($\mathbf{k}_0 = \mathbf{k}$). Then the phase term becomes unity and the integral is $f(\mathbf{q}=0) = \int \rho(\mathbf{r})\mathrm{d}\mathbf{r} = N$. The form-factor is complex,

$$f(\mathbf{q}) = |f(\mathbf{q})| e^{i\varphi(\mathbf{q})}, \tag{2.8}$$

that is, it has an associated phase factor, $e^{i\varphi(\mathbf{q})}$.

2.1.2 The Independent Atom Model

For atoms, the electron density is isotropic and therefore the associated atomic form-factors are also isotropic and can be written as a function of $q = |\mathbf{q}|$ only. There are tabulated real-valued atomic form-factors $f^{\mathrm{at}}(q)$, which generally fit a sum of N_g Gaussian functions of the form,

$$f^{\mathrm{at}}(q) = \sum_{i=1}^{N_g} a_i \exp\left[-b_i \Big(\frac{q}{4\pi}\Big)^2 \right] + c, \tag{2.9}$$

and contained within the references [62,64] are the values of the fitting coefficients $\{a_i, b_i, c\}$ for many atoms and atomic ions. Conventionally, the molecular form-factor is approximated by a sum of N_{at} atomic form-factors with appropriate phase-terms,

$$f(\mathbf{q}) = \sum_{j}^{N_{\mathrm{at}}} f_j^{\mathrm{at}}(\mathbf{q}) e^{i\mathbf{q}\cdot\mathbf{r}_j}, \tag{2.10}$$

where each atom is centred at $\mathbf{r}_j$. This approximation is known as the independent atom model (IAM), and is equivalent to stating that the total molecular electron density can be approximated by a sum of isolated atomic densities,

$$\rho(\mathbf{r}) = \sum_{j}^{N_{\mathrm{at}}} \rho_j^{\mathrm{at}}(\mathbf{r} - \mathbf{r}_j), \tag{2.11}$$

each centred at $\mathbf{r}_j$. Clearly this omits the redistribution of valence electrons due to chemical bonding, therefore works well for molecules containing heavy atoms, where the ratio of valence to bonding electrons is small. The absolute-square of the molecular form-factor gives a probability distribution in $\mathbf{q}$ for each photon, this is proportional to the diffraction intensity,

$$I(\mathbf{q}) \propto |f(\mathbf{q})|^2 = \sum_{i}^{N_{\mathrm{at}}} f_i^{\mathrm{at}}(\mathbf{q})^2 + 2\sum_{j>i}^{N_{\mathrm{at}}} f_i^{\mathrm{at}}(\mathbf{q}) f_j^{\mathrm{at}}(\mathbf{q}) e^{i\mathbf{q}\cdot(\mathbf{r}_i - \mathbf{r}_j)}, \tag{2.12}$$

where the first and second terms are called the atomic and molecular terms respectively.

2.1.3 Rotational-averaging

In a gas-phase experiment with incident X-rays scattering from a randomly orientated thermal ensemble of molecules, Eq. (2.12) must be averaged over all possible molecular orientations. Only the phase term within the molecular term depends on the positions of the nuclei, so using $\mathbf{q}\cdot(\mathbf{r}_i - \mathbf{r}_j) = q|\mathbf{r}_i - \mathbf{r}_j|\cos\theta = qr_{ij}\cos\theta$, i.e. the distance between the centre of atom i and atom j is r_{ij}, and the integral limits are $[0, \pi]$ for θ and $[0, 2\pi]$ for ϕ, which gives,

$$\left\langle e^{i\mathbf{q}(\mathbf{r}_i - \mathbf{r}_j)} \right\rangle_{\theta,\phi} = \frac{\int e^{iqr_{ij}\cos\theta} \sin\theta \mathrm{d}\theta\mathrm{d}\phi}{\int \sin\theta \mathrm{d}\theta\mathrm{d}\phi} = \frac{1}{4\pi}\int e^{iqr_{ij}\cos\theta} \sin\theta \mathrm{d}\theta\mathrm{d}\phi. \tag{2.13}$$

The remaining integral is,

$$\begin{aligned}\int e^{iqr_{ij}\cos\theta}\sin\theta \mathrm{d}\theta\mathrm{d}\phi &= 2\pi\int_0^{\pi} e^{iqr_{ij}\cos\theta}\sin\theta\mathrm{d}\theta \\ &= 2\pi\Big(\frac{-1}{iqr_{ij}}\Big)\int_{iqr_{ij}}^{-iqr_{ij}} e^u du \\ &= 2\pi\Big(\frac{-1}{iqr_{ij}}\Big)\Big(e^{-iqr_{ij}} - e^{iqr_{ij}}\Big) \\ &= 4\pi\frac{\sin qr_{ij}}{qr_{ij}},\end{aligned} \tag{2.14}$$

and hence,

$$\Big\langle e^{i\mathbf{q}(\mathbf{r}_i-\mathbf{r}_j)}\Big\rangle_{\theta,\phi} = \frac{\sin qr_{ij}}{qr_{ij}}. \tag{2.15}$$

Using this result the rotationally-averaged diffraction intensity is,

$$I_{\mathrm{Deb}}(q) = \sum_i^{N_{\mathrm{at}}} f_i^{\mathrm{at}}(q)^2 + 2\sum_{j>i}^{N_{\mathrm{at}}} f_i^{\mathrm{at}}(q) f_j^{\mathrm{at}}(q)\frac{\sin qr_{ij}}{qr_{ij}}, \tag{2.16}$$

which was famously derived by Debye in 1915 [38]. Note this expression is only valid when the molecular form-factor is approximated by a sum of atomic factors, as in the IAM approximation. Throughout this thesis a 'beyond IAM' method is predominantly used, which has been touched upon in Section 1.5. Eq. (2.16) is valid with the IAM method where the tabulated atomic scattering factors can be used. Starting with Eq. (2.7), where the intensity is the absolute square of the molecular scattering factor, $I(\mathbf{q}) = |f(\mathbf{q})|^2$, the rotationally-averaged diffraction intensity is given by the angular integration,

$$I(q) = \int_0^{2\pi}\int_0^{\pi} |f(q,\theta,\phi))|^2 \sin\theta\mathrm{d}\theta\mathrm{d}\phi, \tag{2.17}$$

which is used instead. It is solved via a numerical method such as quadrature, and the term $f(\mathbf{q})$ from Eq. (2.7) is computed on a reciprocal space grid in spherical coordinates (q,θ,ϕ). Alternatively, a Cartesian grid in reciprocal space can be used with the help of Eq. (2.5) or (2.6). In subsequent chapters comparisons are made to the IAM method, that is $I(q)$ from the above equation is compared to $I_{\mathrm{Deb}}(q)$ from Eq. (2.16).

2.1.4 Beyond the Independent Atom Model

The IAM method, as defined by Eq. (2.10), is useful due to its simplicity and computational efficiency. However, it fails to take into account the redistribution of valence electrons due to bonding, and only applies to the electronic ground state. Throughout this thesis, a beyond IAM method is used, beginning with Eq. (2.7). The electron density can be calculated with *ab initio* electronic structure theory, and its description can be such that there is analytical solution to the Fourier integral. This will be discussed in detail in Ch. 3. In this way, the bonding electrons are taken into account, and X-ray scattering from electronically excited states can be calculated. In addition, the molecular wave function within the Born approximation (see Section 2.2.1) has separate nuclear and electronic terms. Averaging over all nuclear rotations has been considered in Eqs. (2.16) and (2.17). However, the molecule can also be pumped to specific rotational states, which induces alignment, and therefore a anisotropic diffraction signal. Vibrational states may also play a role in the X-ray scattering. In short, a molecule in a specific quantum state (electronic, rotational, and vibrational) has a distinct X-ray diffraction fingerprint. This is the topic of Ch. 5.

The discussion so far has only treated the scattering problem classically. There are of course quantum effects which ultimately govern the electrons within atoms and molecules. Such electrons have discrete energy levels, as briefly mentioned in Section 1.2.1, such as the K, L, and M shells. For example, with incident X-ray energy much less than the binding energy of the K shell, the response to the X-ray field will be dampened due to the fact that the electrons are bound. Higher shell electrons (L, M) are less tightly bound, therefore respond more freely to the driving field. This is taken into account in the atomic form-factor by adding an energy-dependent term $f'(\hbar\omega)$. This term becomes zero for X-ray energies much larger than the binding energy. There is also a phase delay between the incident field and the oscillation of the electrons. This is taken into account by another energy-dependent term $if''(\hbar\omega)$. These two correction terms are known as dispersion corrections. These terms have resonance behaviour related to the electron binding energies, and away from so-called absorption edges these terms are close to zero. The X-ray energy is assumed to be such that they are approximately zero, and are therefore ignored throughout this thesis.

2.1.5 Eigenstate scattering and thermal averaging

Although state-specific single-molecule diffraction imaging has not yet been experimentally realised and would pose intricate difficulties, especially for small molecules, the difference between such a case and a thermal or ensemble average in the scattering signal is briefly discussed here. In short, the scattering from a single molecule is perfectly described by the molecular form-factor, $f(\mathbf{q})$. It is up to theory to take into account the effect of the electronic, vibrational, and rotational states in the scattering. For a single molecule, a convolution of $f(\mathbf{q})$ over accurately calculated nuclear distributions is the correct theoretical description. In other words, the molecular superposition coherently scatters, and the phase terms $\{e^{i\mathbf{q}\mathbf{R}_j}\}$ play an important role. This is eigenstate averaging. In the thermal case, with X-rays incident upon an ensemble of molecules, as is the case in gas-phase diffraction experiments, the coherence of each molecule is indistinguishable from the incoherent ensemble-average scattering. At the detector only the average intensity is visible, that is, the convolution of the intensity (not the form-factor) with the nuclear distribution.

To illustrate the difference between eigenstate and thermal averaging, an IAM molecular form-factor, $f(\mathbf{q};\mathbf{R}) = \sum_j^{N_{\text{at}}} f_j(\mathbf{q}) e^{i\mathbf{q}\mathbf{R}_j}$ is used. In the eigenstate averaging case, it is convoluted with a Gaussian ground-state vibrational distribution, defined as,

$$|\Psi_0(\mathbf{R})|^2 = \prod_{i=1}^{N_{at}} \left(\frac{\gamma_i}{\pi}\right)^{\frac{3}{2}} e^{-\gamma_i(\mathbf{R}_i - \mathbf{R}_i^0)^2}, \tag{2.18}$$

where $\{\mathbf{R}_i^0\}$ are the equilibrium or stationary point coordinates for each atom, and γ_i is the Gaussian width. The convolution gives,

$$\langle f(\mathbf{q};\mathbf{R})\rangle_{\text{vib}} \;=\; \sum_{i=1}^{N_{\text{at}}} \left(\frac{\gamma_i}{\pi}\right)^{\frac{3}{2}} f_i(\mathbf{q}) \int e^{-\gamma_i(\mathbf{R}_i - \mathbf{R}_i^0)^2} e^{i\mathbf{q}\mathbf{R}_i} \mathrm{d}\mathbf{R}_i. \tag{2.19}$$

Completing the square, $-\gamma_j(\mathbf{R}_j - \mathbf{R}_j^0)^2 + i\mathbf{q}\mathbf{R}_j = -\gamma_j(\mathbf{R}_j - \sigma_j)^2 + \tau_j$, gives $\sigma_j = \mathbf{R}_j^0 + \frac{i\mathbf{q}}{2\gamma_j}$, $\tau_j = i\mathbf{q}_j\mathbf{R}_j^0 - \frac{\mathbf{q}^2}{4\gamma_j}$, and the new form-factor averaged over the

vibrational distribution is,

$$\begin{aligned}\langle f(\mathbf{q};\mathbf{R})\rangle_{\rm vib} &= \sum_{j=1}^{N_{\rm at}} \left(\frac{\gamma_j}{\pi}\right)^{\frac{3}{2}} f_j(\mathbf{q}) e^{\tau_j} \int e^{-\gamma_j(\mathbf{R}_j-\sigma_j)^2} {\rm d}\mathbf{R}_j \\ &= \sum_{j=1}^{N_{\rm at}} f_j(\mathbf{q}) e^{i\mathbf{q}\mathbf{R}_j^0} e^{-\frac{\mathbf{q}^2}{4\gamma_j}}.\end{aligned} \tag{2.20}$$

The square of which is the X-ray intensity,

$$\begin{aligned}|\langle f(\mathbf{q};\mathbf{R})\rangle_{\rm vib}|^2 &= \sum_{i,j=1}^{N_{at}} f_i(\mathbf{q}) f_j(\mathbf{q}) e^{i\mathbf{q}\mathbf{R}_{ij}^0} e^{-\gamma_{ij}\mathbf{q}^2} \\ &= \Big[\sum_{i=1}^{N_{at}} |f_i(\mathbf{q})|^2 + \sum_{i<j} f_i(\mathbf{q}) f_j(\mathbf{q}) e^{i\mathbf{q}\mathbf{R}_{ij}^0}\Big] e^{-\gamma_{ij}\mathbf{q}^2} \\ &= I(\mathbf{q}) e^{-\gamma_{ij}\mathbf{q}^2}\end{aligned} \tag{2.21}$$

where $I(\mathbf{q})$ is the normal IAM intensity from Eq. (2.12), $\mathbf{R}_{ij}^0 = \mathbf{R}_i^0 - \mathbf{R}_j^0$, and $\gamma_{ij} = (\gamma_i^{-1} + \gamma_j^{-1})/4$. Note that the entire intensity (atomic and molecular terms) is blurred by the reciprocal Gaussian $e^{-\gamma_{ij}\mathbf{q}^2}$.

In the thermal ensemble case the convolution should be with the intensity and the vibrational distribution. The intensity before convolution is the absolute-square of the form-factor,

$$\begin{aligned}|f(\mathbf{q};\mathbf{R})|^2 &= \sum_{i,j}^{N_{\rm at}} f_i(\mathbf{q}) f_j(\mathbf{q}) e^{i\mathbf{q}(\mathbf{R}_i^0 - \mathbf{R}_j^0)} \\ &= \sum_{i}^{N_{\rm at}} |f_i(\mathbf{q})|^2 + \sum_{i<j} f_i(\mathbf{q}) f_j(\mathbf{q}) e^{i\mathbf{q}\mathbf{R}_{ij}^0}.\end{aligned} \tag{2.22}$$

Convolution with the vibrational distribution gives,

$$\begin{aligned}&\langle |f(\mathbf{q};\mathbf{R})|^2\rangle_{\rm vib} = \\ &\sum_{i}^{N_{\rm at}} |f_i(\mathbf{q})|^2 + \sum_{i<j} \left(\frac{\gamma_i\gamma_j}{\pi^2}\right)^{\frac{3}{2}} f_i(\mathbf{q}) f_j(\mathbf{q}) \prod_{k=i,j}^{2} \int e^{-\gamma_k(\mathbf{R}_k - \mathbf{R}_k^0)^2} e^{i\mathbf{q}\mathbf{R}_k^0} {\rm d}\mathbf{R}_k.\end{aligned} \tag{2.23}$$

The integrals are the same as before, thus,

$$\langle |f(\mathbf{q};\mathbf{R})|^2\rangle_{\rm vib} = \sum_{i}^{N_{at}} |f_i(\mathbf{q})|^2 + \sum_{i<j} f_i(\mathbf{q}) f_j(\mathbf{q}) e^{i\mathbf{q}\mathbf{R}_{ij}^0} e^{-\gamma_{ij}\mathbf{q}^2}. \tag{2.24}$$

This is different to the coherent single-molecule scattering case, as the atomic form-factors are not blurred by the vibrational Gaussian term, $e^{-\gamma_{ij}\mathbf{q}^2}$, as can be seen by comparing Eqs. (2.21) and (2.24). Only the molecular intensity term is blurred, not the atomic term, i.e. atomic terms do not scatter with themselves (there is no superposition), and it is an entirely classical situation.

2.2 Fundamental ideas of *ab initio* theory

A brief discussion of the fundamental ideas of *ab initio* theory follows, which is used to calculate molecular wave functions and electron densities in the context of this thesis. *Ab initio* is Latin for 'from the beginning'. It is fairly synonymous with the term 'from first principles'. It means using minimal approximations and deriving equations based on the postulates of quantum mechanics. Computational chemistry ranges from *ab initio* quantum chemistry methods such as Hartree-Fock (HF) and multiconfigurational self-consistent field (MCSCF), for modelling small molecules quantum mechanically; to classical molecular dynamics (MD) simulations of relatively large systems or ensembles of molecules. The 1998 Nobel Prize in Chemistry was awarded to Kohn and Pople for their development of computational methods in quantum chemistry [85]. More recently, Warshel, Karplus, and Levitt were awarded the Nobel Prize in Chemistry in 2013 for progress in the field of computational chemistry [86], particularly for developing a multiscale method, i.e. an efficient combination of quantum and classical techniques for modelling complex chemical systems.

The computational chemistry methods used throughout this thesis are predominantly wave function based, such as HF for molecules in their electronic ground state, and MCSCF for ground state and excited state molecules. These techniques are ways of computationally solving the time-independent Schrödinger equation for molecules, employing the Born-Oppenheimer approximation, whereby the molecular wave function only depends parametrically on the nuclear geometry. Basis sets are used; in particular, Gaussian-type orbitals (GTOs), which speed-up the so-called Coulomb and exchange integrals tremendously, which are crucial to the linear algebraic solutions to the HF equations. Quantum mechanical treatments are restricted computationally as the processing power necessary scales exponentially with the number of basis functions. This generally restricts these methods to small molecules (e.g. <100 atoms). In the following some fundamental

ideas underlying *ab initio* methodology are discussed [87].

2.2.1 The Born-Oppenheimer approximation

The Born-Oppenheimer or adiabatic approximation [88] is central to quantum chemistry. It separates the nuclear and electronic degrees of freedom, making solving the time-independent Schrödinger equation tractable. It states that since nuclei are much heavier than electrons, they move much slower; and within a molecule, the nuclei can be considered stationary with the electrons moving in the fixed nuclear field. The full N-electron and M-nuclei Hamiltonian in atomic units is,

$$\begin{aligned} \hat{H} &= \hat{T}_N + \hat{T}_e + V_{eN} + V_{ee} + V_{NN} \qquad (2.25) \\ &= -\sum_A^M \frac{1}{2m_A}\nabla_A^2 - \frac{1}{2}\sum_i^N \nabla_i^2 - \sum_A^M \sum_i^N \frac{Z_A}{r_{iA}} + \sum_{i>j}^N \frac{1}{r_{ij}} + \sum_{A>B}^M \frac{Z_A Z_B}{R_{AB}}, \end{aligned}$$

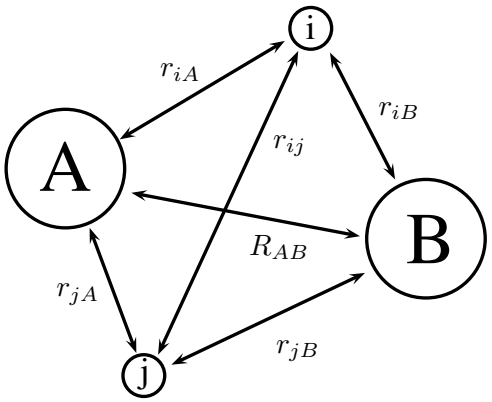


Figure 2.3: Distances within a diatomic molecule with nuclei A and B, and two electrons i and j.

where $\hat{T}_N$ and $\hat{T}_e$ are the nuclear and electronic kinetic energy operators; V_{eN}, V_{NN}, and V_{ee} are the potential energies for electron-nuclei, nuclei-nuclei, and electron-electron pairs; m_A is the mass of the nucleus with charge Z_A; the distance between nucleus A and electron i is r_{iA}; and the distance between nuclei A and B is R_{AB}. The distances between each particle within a molecule are shown in Fig. 2.3. Application of the Born Oppenheimer approximation means the Laplacian operator, ∇_A^2, acting on the wave function, Ψ, equals zero, as the coordinates $\{\mathbf{R}_A\} = (\mathbf{R}_1, \mathbf{R}_2, \ldots, \mathbf{R}_M)$ are fixed. The distances R_{AB} are also fixed, thus, the nuclear repulsion term is constant, and any constant added to an operator

does not affect its eigenfunction, is only added to its eigenvalue. Therefore, an electronic Hamiltonian can be defined by omitting the terms, $\hat{T}_N$ and V_{NN},

$$\hat{H}_{\text{el}} = -\frac{1}{2}\sum_{i}^{N}\nabla_i^2 - \sum_{A}^{M}\sum_{i}^{N}\frac{Z_A}{r_{iA}} + \sum_{i>j}^{N}\frac{1}{r_{ij}}, \tag{2.26}$$

which has a corresponding wave equation called the electronic Schrödinger equation,

$$\hat{H}_{\text{el}}\Psi_{\text{el}} = E_{\text{el}}\Psi_{\text{el}}, \tag{2.27}$$

with the electronic wave function and electronic energy as its solutions. The electronic wave function depends explicitly on the electronic coordinates but depends only parametrically on the nuclear coordinates,

$$\Psi_{\text{el}} = \Psi_{\text{el}}(\{\mathbf{r}_i\};\{\mathbf{R}_A\}), \tag{2.28}$$

as does the electronic energy,

$$E_{\text{el}} = E_{\text{el}}(\{\mathbf{R}_A\}), \tag{2.29}$$

meaning that for difference nuclear arrangements Ψ_{el} is a different function of the electronic coordinates and the function $E_{\text{el}}(\{\mathbf{R}_A\})$ defines the potential energy surface for the molecule. The total energy for fixed nuclei must also include the constant nuclear repulsion,

$$E_{\text{tot}} = E_{\text{el}} + \sum_{A>B}^{M}\frac{Z_A Z_B}{R_{AB}}. \tag{2.30}$$

It is now possible to solve for the nuclear motion under the Born-Oppenheimer approximation. As the nuclei are fixed with the electrons moving extremely quickly, it is reasonable to assume the nuclei move in the average field of the

electrons. From Eq. (2.25),

$$\hat{H}_{\text{nuc}} = -\sum_{A}^{M} \frac{1}{2m_A}\nabla_A^2 + \left\langle -\frac{1}{2}\sum_{i}^{N}\nabla_i^2 - \sum_{A}^{M}\sum_{i}^{N}\frac{Z_A}{r_{iA}} + \sum_{i>j}^{N}\frac{1}{r_{ij}} \right\rangle \quad (2.31)$$

$$+ \sum_{A>B}^{M}\frac{Z_A Z_B}{R_{AB}}$$

$$= -\sum_{A}^{M}\frac{1}{2m_A}\nabla_A^2 + E_{\text{tot}}. \quad (2.32)$$

Using this Hamiltonian the nuclear Schrödinger equation is,

$$\hat{H}_{\text{nuc}}\Psi_{\text{nuc}} = E\Psi_{\text{nuc}}, \quad (2.33)$$

whose solutions describe the vibrational, rotational, and translational motions of a molecule,

$$\Psi_{\text{nuc}} = \Psi_{\text{nuc}}(\{\mathbf{R}_A\}), \quad (2.34)$$

and E is the Born-Oppenheimer approximated total energy of the molecule, which includes electronic, vibrational, rotational, and translational energies. The corresponding Born-Oppenheimer wave function is,

$$\Psi(\{\mathbf{r}_i\};\{\mathbf{R}_A\}) = \Psi_{\text{el}}(\{\mathbf{r}_i\};\{\mathbf{R}_A\})\Psi_{\text{nuc}}(\{\mathbf{R}_A\}). \quad (2.35)$$

2.2.2 Hartree products

In this section only the electronic wave function is discussed, which will be termed simply as the wave function, and its subscript *el* will be dropped. The *independent particle model* can now be invoked. This assumption allows the N-electron Hamiltonian to be written as a sum of N one-electron Hamiltonians,

$$\hat{H}_{\text{el}} = \sum_{i}^{N}\hat{h}(i), \quad (2.36)$$

where,

$$\hat{h}(i) = -\frac{1}{2}\nabla_i^2 - \sum_{A}^{M}\frac{Z_A}{r_{iA}}. \quad (2.37)$$

This is an oversimplification, as the electron-electron repulsion term r_{ij}^{-1} has been omitted, meaning there is no correlation of electron motion. The one-electron Hamiltonians are termed *core-Hamiltonians*, since the only interactions included are those between the electrons and the nuclei. A consequence of using such an electronic Hamiltonian is that the wave function can be written as a product of n single particle wave functions, also known as a Hartree product,

$$\Psi = \chi_1(\mathbf{x}_1)\chi_2(\mathbf{x}_2)\dots\chi_N(\mathbf{x}_N), \tag{2.38}$$

where $\{\chi_i(\mathbf{x}_i)\}$ are called spin orbitals, which are the product of spatial orbitals $\{\psi_i(\mathbf{r}_i)\}$ and spin functions $\alpha(\omega)$ or $\beta(\omega)$, which denote up and down spin. The coordinates $\{\mathbf{x}_i\}$ take into account the electrons spatial and spin coordinates, $\mathbf{x} = \{\mathbf{r}, \omega\}$.

2.2.3 Slater determinants

Electrons are fermions and must obey the Pauli exclusion principle, which leads to the requirement of an overall antisymmetric wave function. Therefore, no two electrons can occupy the same spin orbital, and swapping two electrons must change the sign of the wave function,

$$\Psi(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N) = -\Psi(\mathbf{x}_2, \mathbf{x}_1, \dots, \mathbf{x}_N). \tag{2.39}$$

The use of a Slater determinant enforces this,

$$\Psi(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N) = (N!)^{-1/2} \begin{vmatrix} \chi_1(\mathbf{x}_1) & \chi_2(\mathbf{x}_1) & \dots & \chi_N(\mathbf{x}_1) \\ \chi_1(\mathbf{x}_2) & \chi_2(\mathbf{x}_2) & \dots & \chi_N(\mathbf{x}_2) \\ \vdots & \vdots & \ddots & \vdots \\ \chi_1(\mathbf{x}_N) & \chi_2(\mathbf{x}_N) & \dots & \chi_N(\mathbf{x}_N) \end{vmatrix}, \tag{2.40}$$

which for short-hand is written as,

$$|\Psi\rangle = |\chi_1\chi_2\dots\chi_N\rangle. \tag{2.41}$$

Note that the factor $(N!)^{-1/2}$ in Eq. (2.40) is a normalisation factor. Determinants have various mathematical properties. Most importantly in this case, the interchanging of any pair of columns or rows of a matrix multiplies its determinant by -1, which enforces an antisymmetric wave function, as swapping any electronic

coordinate is the same as swapping a row within the matrix; and whenever two rows of a matrix are identical, its determinant is 0, which avoids any electrons occupying the same spin orbital i.e. having the same spatial and spin coordinate simultaneously. The dip to zero when any two coordinates are equal, $\mathbf{x}_i = \mathbf{x}_j$, is referred to as a *Fermi hole*, and is a consequence of the antisymmetric wave function. This repulsive effect is, at least partially, a correlation of two electrons with equal spin. Electrons with opposite spin are not correlated, as such, single determinant wave functions are described as *uncorrelated*.

In general multiple Slater determinants are necessary to best approximate the true molecular wave function. However single determinant methods are conceptually helpful and are often used as a starting point for more accurate methods. One such single determinant method is Hartree-Fock theory.

2.2.4 Hartree-Fock theory

The essence of the Hartree-Fock (HF) approximation is to replace the complicated many-electron Schrödinger equation by a one-electron problem in which electron-electron repulsion is treated in an average way, for this reason it is known as a mean-field theory. To achieve this it assumes that only one Slater determinant can be used to approximate the wave function, and that the independent particle model is appropriate, thereby reducing the problem to N single particle equations. Generally, these are poor assumptions but they can be reasonable for molecules in their ground state, and as a starting point for more accurate theoretical treatments.

In practice, *ab initio* electronic structure methods make use of an important theorem in quantum mechanics called the variational theorem. It states that for any trial wave function, $|\Psi_{\text{trial}}\rangle$, which satisfies the correct boundary conditions, the expectation value of the energy, E, obtained with this wave function never lies below the true or exact energy, E_{exact}, of the state. Expressed in mathematical terms,

$$E = \frac{\langle \Psi_{\text{trial}} | \hat{H} | \Psi_{\text{trial}} \rangle}{\langle \Psi_{\text{trial}} | \Psi_{\text{trial}} \rangle} \geq E_{\text{exact}}. \tag{2.42}$$

In HF theory, trial wave functions are chosen by assigning the N electrons to occupy a set of spin orbitals,

$$|\Psi_{\text{trial}}\rangle = |\chi_1 \chi_2 \dots \chi_N\rangle. \tag{2.43}$$

There is an infinite number of spin orbitals to choose from, but generally a basis set of K spatial orbitals is selected, which gives rise to $2K$ spin orbitals, restricting the choice of spin orbitals to $\binom{2K}{N}$ combinations, where N is the number of electrons, or equivalently, spin orbitals to be occupied. The choice of spin orbitals that minimises the energy E is the variationally best choice, and is called the Hartree-Fock ground state, $|\Psi_0\rangle$. To find this set of spin orbitals the Hartree-Fock equations are used, which are the N one-electron eigenvalue equations,

$$\hat{f}(i)\chi(\mathbf{x}_i) = \epsilon_i \chi_i(\mathbf{x}_i), \tag{2.44}$$

where $\hat{f}(i)$ is an effective one-electron operator called the Fock operator,

$$\hat{f}(i) = -\frac{1}{2}\nabla_i^2 - \sum_{A=1}^{M} \frac{Z_A}{r_{iA}} + \nu^{HF}(i), \tag{2.45}$$

where $\nu^{HF}(i)$ is the average potential experienced by the ith electron due to the presence of other electrons. This depends on the choice of spin orbitals, i.e. the Fock operator, $\hat{f}(i)$, which depends on its eigenfunctions, $\{\chi_k\}$. Therefore, the HF equations must be solved iteratively, via a method called the self-consistent field (SCF) method. The trial set of orbitals, $\{\chi_k\}$, are used to calculate $\nu^{HF}(i)$, thus, $\hat{f}(i)$, then, solving the HF equations (2.44) gives a new set of orbitals, $\{\chi'_k\}$, and a set of energy eigenvalues, $\{\epsilon_k\}$, which can then be used to repeat the process until 'self-consistency' is achieved, that is, until the new orbital set equals the previous one, $\{\chi'_k\} = \{\chi_k\}$, and the total energy, $\sum_k \epsilon_k$, is minimised.

The exact form of $\nu^{HF}(i)$ involves the so-called Coulomb and exchange integrals. For a detailed evaluation of these integrals and discussion of Hartree-Fock theory and the SCF method, the reader is referred to Szabo-Oslund [87].

2.3 Multi-determinant methods

The Hartree-Fock method often results in inadequate descriptions of excited states, chemical reactions, and locations of saddle points. A well known error of HF is that it wrongly predicts molecular dissociation limits. The reason for these problems are essentially that HF theory is an uncorrelated theory, that is the motion of electron 1 is independent of the motion of electron 2. Scenarios where electrons are very close together are therefore given too high weighting. So-called

correlated methods fix this problem by introducing electron correlation. They mainly do this by utilising more than one Slater determinant to construct the total wave function. There is a wide range of such methods, including Configuration Interaction (CI), and Multi-Configurational Self-Consistent Field (MCSCF). The general properties of excited determinants are described in the next Section, followed by a closer look at CI and MCSCF.

2.3.1 Excited determinants

If a basis of K spatial orbitals are selected, there are $2K$ spin orbitals, with N occupied, and $2K - N$ unoccupied. The Slater determinant,

$$|\Psi_0\rangle = |\chi_1\chi_2 \cdots \chi_a\chi_b \cdots \chi_N\rangle, \tag{2.46}$$

is one of $\binom{2K}{N}$ possible determinants, or *configurations*, that could be formed from the $2K > N$ spin orbitals. If $|\Psi_0\rangle$ is the Hartree-Fock ground state wave function, it is convenient to define other 'excited' determinants with respect to it. A singly excited determinant is one in which an electron has been promoted from an occupied orbital χ_a in $|\Psi_0\rangle$, to a virtual orbital χ_r, this can be written as,

$$|\Psi_a^r\rangle = |\chi_1\chi_2 \cdots \chi_r\chi_b \cdots \chi_N\rangle. \tag{2.47}$$

A doubly excited determinant is one in which two electrons from orbitals χ_a and χ_b have been promoted to virtuals χ_r and χ_s. Using previous notation,

$$|\Psi_{ab}^{rs}\rangle = |\chi_1\chi_2 \cdots \chi_r\chi_s \cdots \chi_N\rangle. \tag{2.48}$$

All $\binom{2K}{N}$ determinants can be classified as N-tuply excited determinants with respect to the Hartree-Fock ground state determinant. These excited determinants are not accurate representations of the true excited states of the system, but they are important as N-electron basis functions for an expansion of the exact N-electron states of the system. In general, the importance of excited determinants drops off with their excitation number, i.e. an expansion involving the ground state determinant and singly and doubly excited determinants is usually not greatly improved upon by adding triply excited determinants and so on.

2.3.2 Configuration interaction

Using the ground state determinant and all possible excited Slater determinants, as allowed within the chosen basis set, i.e. using all possible configurations, to form a basis for an improved wave function, gives the configuration interaction (CI) wave function,

$$|\Psi_{\mathrm{CI}}\rangle = c_0|\Psi_0\rangle + \sum_{ra} c_a^r|\Psi_a^r\rangle + \sum_{a<b,r<s} c_{ab}^{rs}|\Psi_{ab}^{rs}\rangle + \dots . \tag{2.49}$$

The infinite set of N-electron determinants $\{|\Psi_i\rangle\} = \{|\Psi_0\rangle, |\Psi_a^r\rangle, |\Psi_{ab}^{rs}\rangle, \dots\}$ is a complete set for the expansion of any N-electron wave function. Since every $|\Psi_i\rangle$ can be defined by specifying a configuration of spin orbitals, the method is known as configuration interaction. In practice, a finite basis of $2K$ spin orbitals is necessary. Then, the set of N-electron determinants is finite. Using all $\binom{2K}{N}$ possible determinants in the wave function expansion, constitutes the best possible wave function within the basis set constraints, this is called a *full* CI. It quickly becomes impossible to calculate, as for increasing N, larger sets of $2K$ spin orbitals become necessary and the value of $\binom{2K}{N}$ rises extremely rapidly. It can be appropriate to omit all doubly excited or higher determinants, which is called CIS for CI 'singles'. Omission of all triply excited or higher determinants is called CISD for CI 'singles and doubles', and so on.

2.3.3 Multiconfigurational self-consistent field

Multiconfigurational self-consistent field (MCSCF) is closely related to CI theory, and can be regarded as a truncated CI expansion where both the expansion coefficients and the orbitals are optimised variationally. It is often used in studies of chemical reactions, where prediction of the correct dissociation limits is necessary. It is also effective at modelling excited states as long as it is appropriately used.

It utilises a pre-selected *active space* where electrons can adapt. The active space should be chosen to contain the orbitals where main changes in occupancy occur during a reaction. Table 2.1 is an illustration of the active space in which electron occupancies can be non-integer in the range [0,2], whereas the secondary and inactive orbitals have fixed integer occupancies of 0 and 2 respectively. A shorthand description of an active space consisting of six electrons within four orbitals is written as (6,4). This often follows the method, e.g. MCSCF(6,4).

Secondary
$a = 0$
Active
$0 < a < 2$
Inactive
$a = 2$

Table 2.1: MCSCF orbital distribution showing occupancies a for each class of orbitals.

This means detailed pre-knowledge of a system is often necessary as the method strongly depends on the choice of active space. This drawback is overcome by the fact that it is a powerful method which allows study of much larger systems compared to full CI.

2.4 Basis sets

The spin orbitals which make up the Slater determinant (Eq. (2.40)) consist of a spatial orbital and spin function,

$$\chi(\mathbf{r}, \omega) = \phi(\mathbf{r})\alpha(\omega), \tag{2.50}$$

where α is a spin function which can either be up or down depending on the spin coordinate, ω. The spatial orbitals can be constructed from a basis set, making the problem of solving the Schrödinger equation tractable, as the infinite number of possible spin orbitals becomes a finite set of $\binom{2K}{N}$ orbitals (that have a reasonable likelihood to be occupied).

2.4.1 Slater-type and Gaussian-type orbitals

In 1930, J. C. Slater proposed approximate analytical wave functions for all the atoms [89], in any ionisation state. They have radial form,

$$\phi = r^{n^*-1} \exp\Big[- \frac{(Z-s)}{n^*} r \Big], \tag{2.51}$$

with Z the charge on the nucleus; s a screening constant; the effective quantum number, n^*, is assigned via a direct mapping from the principle quantum number n; and s assigned by simple rules. Which makes a complete set of one-electron wave functions. They can be rewritten as general Slater-type orbitals,

$$\phi^{\mathrm{STO}} = r^{n^*-1}e^{-\eta r}, \tag{2.52}$$

where $r = |\mathbf{r} - \mathbf{R}_A|$, for an electron positioned at $\mathbf{r} = (x, y, z)$, centred at nuclear coordinates $\mathbf{R}_A = (X_A, Y_A, Z_A)$, with tabulated values for η.

Later in 1950, S. F. Boys proposed an computationally practical approximation to the one-electron wave functions in the form of Gaussian functions [90],

$$\phi^{\mathrm{GTO}} = C^G e^{-\alpha r^2} \tag{2.53}$$

The main functional differences between Gaussian and Slater-type functions are at $r = 0$ and large r. At $r = 0$, $[\frac{\mathrm{d}}{\mathrm{d}r}e^{-\eta r}]_{r=0} \neq 0$, and $[\frac{\mathrm{d}}{\mathrm{d}r}e^{-\alpha r^2}]_{r=0} = 0$; and as $r \to \infty$, the Gaussian function decays toward zero, $e^{-\alpha r^2} \to 0$, at a faster rate than the Slater function. In terms of computing one and two electron integrals, Gaussian functions are much more efficient. However, they do not approximate the true one-electron wave functions to the same level of accuracy as Slater-type functions. This can be mitigated by using a sum of multiple Gaussian functions, or *primitives*, to approximate a Slater function, still with net gain in the computational effort in solving integrals. This is known as a Gaussian *contraction*, and replaces Slater-type orbitals as basis functions,

$$\phi^{\mathrm{BF}} = \sum_{s=1}^{n_k} \mu_s \phi_s^{\mathrm{GTO}}, \tag{2.54}$$

with fixed basis coefficients $\{\mu_i\}$. The contraction size n_k and coefficients $\{\mu_i\}$ are pre-optimised variables, and are contained within well-established basis sets; such as Pople basis sets, 6-31G, ..., 6-311++G**; and Dunning basis sets, VDZ, ..., aug-cc-pV6Z. See Table 2.2 for a comparison of the number of primitives per basis set for HF calculations on the H_2 molecule. Each spatial orbital can be expressed as a linear combination of basis functions,

$$\psi(\mathbf{r}) = \sum_{k=1}^{N_{\mathrm{BF}}} M_k \phi_k^{\mathrm{BF}}, \tag{2.55}$$

weighted by $\{M_j\}$. In *ab initio* electronic structure theory, the Gaussian primi-

Basis set	Primitives
STO-3G	12
6-31G	16
6-31G**	28
6-311++G**	36
aug-cc-pVDZ	60
aug-cc-pVTZ	148
aug-cc-pVQZ	300

Table 2.2: Comparison of basis set size, by total number of primitive functions, for Hartree-Fock calculations on H_2.

tives, $\{\phi_k^{\mathrm{GTO}}\}$, can either be Cartesian Gaussians or spherical Gaussians. Spherical Gaussians, centred at distance r_0 from the origin, have the form,

$$\tilde{g}(\mathbf{r}) = \tilde{N} Y_{m'}^{l'}(\theta, \phi)(r - r_0)^{n'} e^{-\gamma(r-r_0)^2}, \tag{2.56}$$

where θ and ϕ are the polar and azimuthal angles, $Y_{m'}^{l'}(\theta, \phi)$ is a spherical harmonic, and l', m', and n' are the respective polar, azimuthal, and radial quantum numbers. Transformation between Cartesian and spherical harmonic Gaussians is described in Ref. [91]. From here on, only Cartesian Gaussian functions are used, for their simplicity in the algebra necessary in the following. A general Cartesian GTO centred at coordinates $\mathbf{r}_0 = (x_0, y_0, z_0)$ has the form,

$$g(\mathbf{r}) = \mathcal{N}(x - x_0)^l (y - y_0)^m (z - z_0)^n e^{-\gamma(\mathbf{r}-\mathbf{r}_0)^2}, \tag{2.57}$$

with exponent γ and Cartesian orbital angular momentum $L = l + m + n$. Each individual GTO must be normalised using,

$$\mathcal{N}^2 \int_{-\infty}^{\infty} |g(\mathbf{r})|^2 \mathrm{d}\mathbf{r} = 1, \tag{2.58}$$

where the normalisation constant $\mathcal{N}$ is dependent on the Gaussian exponent, γ, and orbital angular momentum numbers, l, m, and n. A general GTO normalisation constant can be derived [92],

$$\mathcal{N} = \left[\left(\frac{2}{\pi} \right)^{3/4} \frac{2^{(l+m+n)} \gamma^{(2l+2m+2n+3)/4}}{[(2l-1)!!(2m-1)!!(2n-1)!!]^{1/2}} \right], \tag{2.59}$$

where $(2l-1)!! = 1 \cdot 3 \cdot 5 \cdots (2l-1)$.

2.5 Electron density

The electron density is a physical observable of the wave function. As discussed earlier, see e.g. Eq. (2.7), it is experimentally probed by X-ray diffraction. As such, it is an important property throughout this thesis. It is the real space analogue of the X-ray molecular form-factor (which is in reciprocal space). For a N-electron wave function the electron density is the probability of an electron being found at a specific location. The electron density operator is a one electron operator, defined as,

$$\hat{\rho}(\mathbf{r}) = \sum_{i}^{N} \hat{\rho}_i(\mathbf{r}_i) = \sum_{i}^{N} \delta(\mathbf{r}_i - \mathbf{r}). \tag{2.60}$$

This is important when applying it to wave functions consisting of multiple Slater determinants, such as in MCSCF theory, which will be shown in Chapter 3 when calculating the multi-determinant electron density (and following from that deriving the X-ray molecular form-factor based on multiple determinants). Its expectation value with the wave function can be used to calculate the electron density,

$$\begin{aligned} \rho(\mathbf{r}) = \langle\Psi| \sum_{i}^{N} \delta(\mathbf{r}_i - \mathbf{r}) |\Psi\rangle &= \int |\Psi(\mathbf{r}, \mathbf{r}_2, \ldots, \mathbf{r}_N)|^2 \mathrm{d}\mathbf{r}_2 \mathrm{d}\mathbf{r}_3 \ldots \mathrm{d}\mathbf{r}_N \\ &+ \int |\Psi(\mathbf{r}_1, \mathbf{r}, \ldots, \mathbf{r}_N)|^2 \mathrm{d}\mathbf{r}_1 \mathrm{d}\mathbf{r}_3 \ldots \mathrm{d}\mathbf{r}_N \\ &+ \ldots \\ &+ \int |\Psi(\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r})|^2 \mathrm{d}\mathbf{r}_1 \mathrm{d}\mathbf{r}_2 \ldots \mathrm{d}\mathbf{r}_{N-1}. \end{aligned} \tag{2.61}$$

These integrals are identical because swapping electron coordinates only alters the sign of the wave function, $\Psi(\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_N) = -\Psi(\mathbf{r}_2, \mathbf{r}_1, \ldots, \mathbf{r}_N)$, thus it can be written as N multiples of the $N-1$ dimensional integral,

$$\rho(\mathbf{r}) = \langle\Psi| \sum_{i}^{N} \delta(\mathbf{r}_i - \mathbf{r}) |\Psi\rangle = N \int |\Psi(\mathbf{r}, \mathbf{r}_2, \ldots, \mathbf{r}_N)|^2 \mathrm{d}\mathbf{r}_2 \ldots \mathrm{d}\mathbf{r}_N. \tag{2.62}$$

The exact form of the wave function is highly dependent on the model chemistry used. In this thesis *ab initio* electronic structure theory is used, such as HF or

MCSCF theory, with various Gaussian basis sets. The wave function depends on $3N$ spatial coordinates and N spin coordinates, e.g. as in Eq. (2.40). The electron density is reduced to a three dimensional function, because $3(N-1)$ spatial coordinates and the spin coordinates are integrated out. It's worth noting that this is main strength of density functional theory (DFT), as it uses the electron density as the main variable rather than the wave function. This reduces the high dimensional problem to only a three dimensional one, making it much more computationally tractable. However, a considerable focus in the next Chapter is on X-ray diffraction from electronically excited states, therefore MCSCF was chosen as the central theoretical method, with HF calculations for quick verification purposes.

Chapter 3

Ab Initio Molecular X-ray Diffraction

3.1 Introduction

This chapter outlines how the elastic X-ray molecular form-factor can be calculated efficiently from single determinant wave functions, such as from Hartree-Fock theory, and multi-determinant (multiconfigurational) wave functions, such as obtained by MCSCF computations. This method is coined *ab initio* X-ray diffraction (AIXRD) [1]. This expands upon previous work by Debnarova and Techert [80, 93, 94] by deriving the form-factor from a multiconfigurational wave function; as well as obtaining a full analytical expression for the Fourier transform of two Gaussian-type orbitals (GTOs), which is fundamental to a computationally efficient solution. In addition, a fast Fourier transform (FFT) method is investigated and compared to the analytical method in terms of accuracy and computational time. Time-resolved X-ray diffraction experiments are capable of imaging photochemical molecular dynamics [5, 6], which involves electronic state crossings and molecular geometry changes. Ultrafast diffraction experiments generally compliment time-resolved spectroscopy studies such as those pioneered by Zewail et al. [12, 95]. Multiconfigurational methods are necessary to accurately model electronically excited states in molecules [87]. Thus, this method is an important connection between modern quantum chemistry methods and state-of-the-art X-ray diffraction pump-probe experiments at XFELs and synchrotrons.

The effect of the molecular geometry, the *ab initio* electronic structure method

(HF or MCSCF), and the basis set (STO-3G, 6-31G, ...), on the predicted diffraction signal is examined, as well as the significance of using a wave function method rather than the straight-forward independent atom model (IAM). A key advantage of a wave function method is the capacity to calculate diffraction from electronically excited states in molecules, and that valence and bonding electrons are fully accounted for.

3.2 Theory

3.2.1 Multiconfigurational electron density

Multiconfigurational methods will be used in this chapter to model the electronic wave function for molecules in their ground and excited states. For this reason and because the X-ray scattering depends on the electron density, a derivation follows of the electron density based on a multi-determinant wave function. Such a wave function is a sum over configurations,

$$|\Psi_{\text{multi}}\rangle = \sum_{I}^{N_{\text{conf}}} c_I |\Psi_I\rangle, \tag{3.1}$$

where $\sum_I |c_I|^2 = 1$, and $|\Psi_I\rangle = |\chi_1\chi_2\ldots\chi_n\rangle$ is short-hand for a Slater determinant, with configurations defined by the ordering of spin orbitals $|1, 2, \ldots, n\rangle$. Applying the electron density operator to Eq. (3.1) gives the electron density,

$$\begin{aligned} \rho(\mathbf{r}) &= \langle\Psi_{\text{multi}}|\hat{\rho}(\mathbf{r})|\Psi_{\text{multi}}\rangle \\ &= \sum_{I}^{N_{\text{conf}}} c_I^2 \langle\Psi_I|\hat{\rho}(\mathbf{r})|\Psi_I\rangle + \sum_{I\neq J} c_I c_J \langle\Psi_I|\hat{\rho}(\mathbf{r})|\Psi_J\rangle, \end{aligned} \tag{3.2}$$

where $c_I^* = c_I$. There are two distinct integrals here, one where each determinant is the same and one where they differ. Using the electron density operator, $\hat{\rho}(\mathbf{r}) = \delta(\mathbf{r}_i - \mathbf{r})$, the first case is,

$$\begin{aligned} \langle\Psi_I|\hat{\rho}(\mathbf{r})|\Psi_I\rangle &= \sum_{i}^{N} \langle\chi_i(i)|\delta(\mathbf{r}_i - \mathbf{r})|\chi_i(i)\rangle \\ &= \sum_{i}^{N_{\text{occ}}} b_i \langle\phi_i(i)|\delta(\mathbf{r}_i - \mathbf{r})|\phi_i(i)\rangle = \sum_{i}^{N_{\text{occ}}} b_i |\phi_i(\mathbf{r})|^2, \end{aligned} \tag{3.3}$$

where there are N_{occ} occupied spatial orbitals with occupancies $b_i = \{1, 2\}$, i stands for the ith-electron coordinate. Note that for a single determinant wave function, such as the ground state Hartree-Fock wave function $|\Psi_0\rangle$, this is the total electron density expression, as $\rho(\mathbf{r}) = \langle\Psi_0|\hat{\rho}(\mathbf{r})|\Psi_0\rangle$. The second integral between different determinants I and J is only non-zero when they differ by only one spin orbital [87], that is,

$$\langle\Psi_I|\hat{\rho}(\mathbf{r})|\Psi_J\rangle = \langle\chi_a(j)|\hat{\rho}_j(j)|\chi_r(j)\rangle, \tag{3.4}$$

where $\chi_a(j)$ and $\chi_r(j)$ are the orbitals that differ. In other words, determinant I has its jth-electron in spin orbital a; determinant J has its jth-electron in spin orbital r; and all other electrons are in the same spin orbitals in both determinants. Terms involving spin orbitals with opposite spin evaluate as zero due to the Pauli exclusion principle, i.e. $\langle\alpha|\alpha\rangle = \langle\beta|\beta\rangle = 1$ and $\langle\alpha|\beta\rangle = 0$, therefore,

$$\langle\chi_a(j)|\hat{\rho}_j(j)|\chi_r(j)\rangle = \begin{cases} \langle\phi_a(j)|\hat{\rho}_j(j)|\phi_r(j)\rangle, & \text{if spins are parallel.} \\ 0, & \text{otherwise.} \end{cases} \tag{3.5}$$

Applying the density operator 'sifts out' the coordinate j, because $\int \delta(\mathbf{r}_j - \mathbf{r}) f(\mathbf{r}_j) \mathrm{d}\mathbf{r} = f(\mathbf{r})$ for an arbitrary function f, and the only ket orbitals which give non-zero integrands are ones with the same coordinates,

$$\langle\phi_a(j)|\hat{\rho}_j(j)|\phi_r(j)\rangle = \langle\phi_a(j)|\delta(\mathbf{r}_j - \mathbf{r})|\phi_r(j)\rangle = \phi_a(\mathbf{r})\phi_r(\mathbf{r}), \tag{3.6}$$

and the spatial orbitals are all real-valued, i.e. $\phi_i^* = \phi_i$. Thus, the total electron density for a multi-determinant wave function is,

$$\rho(\mathbf{r}) = \sum_I^{N_{\text{conf}}} c_I^2 \sum_i^{N_{\text{occ}}} b_{iI} |\phi_{iI}(\mathbf{r})|^2 + \sum_{I \neq J} c_I c_J \phi_{aI}(\mathbf{r}) \phi_{rJ}(\mathbf{r}). \tag{3.7}$$

The first term can be rewritten as,

$$\sum_j^{N_{\text{occ}}} a_j |\phi_j(\mathbf{r})|^2, \tag{3.8}$$

where, $0 \leq a_j = \sum_I c_I^2 b_{jI} \leq 2$, are partial occupancies, which take into account configuration weightings. The sum of configuration weightings equals one, $\sum_I c_I^2 = 1$, thus, the sum of partial occupancies equals the sum of configuration occupancies for each configuration, and equals the number of electrons,

$\sum_j a_j = \sum_i b_{iI} \forall I = N$. Therefore, the integral over all $\mathbf{r}$ for the first term in the total density (Eq. (3.8)) equals the number of electrons,

$$\sum_{i}^{N_{\text{occ}}} a_i \int |\phi_{iI}(\mathbf{r})|^2 \mathrm{d}\mathbf{r} = N. \tag{3.9}$$

The integral over total electron density equals N by definition, that is, $\int \rho(\mathbf{r})\mathrm{d}\mathbf{r} \equiv N$, which means the integral over the second term equals zero,

$$\sum_{I \neq J} c_I c_J \int \phi_{aI}(\mathbf{r})\phi_{rJ}(\mathbf{r})\mathrm{d}\mathbf{r} = 0, \tag{3.10}$$

and does not contribute to the net electron density.

3.2.2 *Ab initio* X-ray diffraction

As described in Section 2.1, the X-ray molecular form-factor is a Fourier transform of the electron density,

$$f_\alpha(\mathbf{q};\mathbf{R}) = \int \rho_\alpha(\mathbf{r};\mathbf{R}) e^{i\mathbf{q}\mathbf{r}} d\mathbf{r}, \tag{3.11}$$

which depends on the electronic state α and parametrically on the nuclear geometry $\mathbf{R}$. The electronic coordinate is $\mathbf{r}$, and the expression $\rho_\alpha(\mathbf{r};\mathbf{R})d\mathbf{r}$ is the probability of finding an electron in the infinitesimal volume element $d\mathbf{r}$. Single and multiconfigurational *ab initio* methods are used to calculate $\rho_\alpha(\mathbf{r};\mathbf{R})$ throughout this thesis. To solve Eq. (3.11) and determine the molecular form-factor for elastic X-ray scattering, an analytical or a numerical approach can be used. The analytical approach described in Section 3.2.3 is referred to as such because it involves analytical solutions to the Fourier transform of the electron density, despite being evaluated computationally. In contrast, the 'pure' numerical approach is based on a Fast Fourier Transform (FFT) of the electron density, which is described in Section 3.2.4. A strong focus on the analytical approach follows with a brief discussion of the numerical solution towards the end of Section 3.2.

The electron density in terms of a contracted Gaussian basis set is,

$$\rho(\mathbf{r}) = \sum_{j=1}^{N_{\text{occ}}} a_j \Big| \sum_{k=1}^{N_{\text{BF}}} M_k^j \sum_{s=1}^{n_k} \mu_s^k g_s^k(\mathbf{r}) \Big|^2, \tag{3.12}$$

where there are N_{occ} occupied orbitals with partial occupancies $0 < a_j \leq 2$, N_{BF} basis functions or *Gaussian contractions* with orbital coefficients M_k^j, n_k primitives per k^{th} basis function with fixed basis coefficients μ_s^k, and the Cartesian Gaussian primitives are centred at coordinates $\mathbf{r}_0 = (x_0, y_0, z_0)$, and are defined as,

$$g(\mathbf{r}) = \mathcal{N}(x - x_0)^l (y - y_0)^m (z - z_0)^n e^{-\gamma(\mathbf{r}-\mathbf{r}_0)^2}, \tag{3.13}$$

with exponent γ and Cartesian orbital angular momentum $L = l + m + n$, where $\{l, m, n\} \in \mathbb{N}$, and normalisation constant $\mathcal{N}$, defined as,

$$\mathcal{N} = \left(\frac{2}{\pi}\right)^{3/4} \frac{2^{(l+m+n)} \gamma^{(2l+2m+2n+3)/4}}{[(2l-1)!!(2m-1)!!(2n-1)!!]^{1/2}}, \tag{3.14}$$

where $(2l-1)!! = 1 \cdot 3 \cdot 5 \cdots (2l-1)$.

3.2.3 Fourier transform of GTO products

To analytically solve Eq. (3.11) the Fourier transform operator $\hat{F}_{\mathbf{r}}$, for transformation of the electronic coordinates $\mathbf{r}$ to their reciprocal $\mathbf{q}$, is applied to Eq. (3.12) giving,

$$\begin{aligned} f_\alpha(\mathbf{q}; \mathbf{R}) &= \sum_{i=1}^{N_{\text{MO}}} a_i \hat{F}_{\mathbf{r}} \Big[\Big| \sum_{j=1}^{N_{\text{BF}}} M_j^i \sum_{s=1}^{n_k} \mu_s^k g_s^k(\mathbf{r}) \Big|^2 \Big](\mathbf{q}) \\ &= \sum_{i=1}^{N_{\text{MO}}} a_i \sum_{j,k=1}^{N_{\text{BF}}} M_j^i M_k^i \sum_{s=1}^{n_j} \sum_{t=1}^{n_k} \mu_s^j \mu_t^k K_{st}^{jk} \hat{F}_{\mathbf{r}} \Big[G_{st}^{jk}(\mathbf{r}) \Big](\mathbf{q}), \end{aligned} \tag{3.15}$$

where $G_{st}^{jk}(\mathbf{r}) = g_s^j(\mathbf{r})g_t^k(\mathbf{r})$ is a new Gaussian defined by the Gaussian product theorem [92], with $K_{st}^{jk} = \exp[-\gamma_s^j\gamma_t^k(\mathbf{r}_s^j - \mathbf{r}_t^k)^2/(\gamma_s^j + \gamma_t^k)]$, centred at $\mathbf{r}_G = (\gamma_s^j\mathbf{r}_s^j + \gamma_t^k\mathbf{r}_t^k)/(\gamma_s^j + \gamma_t^k)$, and new exponent $\gamma_G = \gamma_s^j + \gamma_t^k$. It is a product of three 1D Gaussians in Cartesian coordinates,

$$G_{st}(\mathbf{r}) = \prod_{r'=x,y,z} (r' - r'_s)^{p_s^{r'}} (r' - r'_t)^{p_t^{r'}} e^{-\gamma_G(r' - r'_G)^2}, \tag{3.16}$$

where r'_s and r'_t define the central coordinates of Gaussians s and t and the superscripts j and k (which refer to basis functions j and k) are dropped for brevity; and $p_s^x = l_s$, $p_s^y = m_s$, and $p_s^z = n_s$ are the orbital angular momentum numbers. Thus, the Fourier integral in Eq. (3.15), omitting super- and subscripts, can be simplified to three 1D integrals,

$$\hat{F}_{\mathbf{r}}\Big[G(\mathbf{r})\Big](\mathbf{q}) \equiv \int G(\mathbf{r})e^{i\mathbf{qr}}\mathrm{d}\mathbf{r} = \prod_{r'=x,y,z} \int G(r')e^{iq_{r'}r'}\mathrm{d}r'. \tag{3.17}$$

This amounts to Fourier transforming each combination of 1D Gaussian products, defined by $\{p_i^{r'} = 0, 1, 2, \ldots; p_j^{r'} = 0, 1, 2, \ldots\}$. The orbital angular momentum numbers $0, 1, 2, 3, \ldots$ correspond to $s, p, d, f, \ldots$ GTOs. In practice, GTOs with $p > 3$ (higher than f-orbitals) are rare, and only used for atoms with many electrons or where high accuracy is necessary. Thus, only the combinations of GTO products up to and including $p = 3$ are practically needed. However, the derivation for all combinations $(\forall p_i^{r'}, p_j^{r'})$ follows, i.e. the completely general mathematical solution to the problem.

To solve the 1D Fourier transform from coordinate x to reciprocal q_x of a general Gaussian product centred at x_G with exponent α, where each individual Gaussian is centred at x_i and x_j with angular momentum numbers l_i and l_j, the binomial theorem is used (twice),

$$\begin{aligned} \int G(x)e^{iq_x x}\mathrm{d}x &= \int (x - x_i)^{l_i}(x - x_j)^{l_j} e^{-\alpha(x - x_G)^2} e^{iq_x x}\mathrm{d}x \qquad (3.18) \\ &= \sum_{m=0}^{l_i}\sum_{n=0}^{l_j} \binom{l_i}{m}\binom{l_j}{n}(-x_i)^{l_i - m}(-x_j)^{l_j - n} \int x^{m+n} e^{-\alpha(x - x_G)^2} e^{iq_x x}\mathrm{d}x. \end{aligned}$$

Completing the square, $-\alpha(x-x_G)^2+iq_xx=-\alpha(x-\sigma)^2+\tau-\frac{q_x^2}{4\alpha}$, gives, $\sigma=x_G+\frac{iq_x}{2\alpha}$, $\tau=iq_xx_G$, and,

$$\begin{aligned}&\int G(x)e^{iq_xx}\mathrm{d}x \qquad (3.19)\\ =\ &\sum_{m=0}^{l_i}\sum_{n=0}^{l_j}\binom{l_i}{m}\binom{l_j}{n}(-x_i)^{l_i-m}(-x_j)^{l_j-n}e^{\tau-\frac{q_x^2}{4\alpha}}\int x^{m+n}e^{-\alpha(x-\sigma)^2}\mathrm{d}x.\end{aligned}$$

Changing coordinates to $u=x-\sigma$, using the binomial theorem again, and defining $\nu=m+n$ gives,

$$\begin{aligned}&\int G(x)e^{iq_xx}\mathrm{d}x \qquad (3.20)\\ =\ &\sum_{m=0}^{l_i}\sum_{n=0}^{l_j}\binom{l_i}{m}\binom{l_j}{n}(-x_i)^{l_i-m}(-x_j)^{l_j-n}e^{\tau-\frac{q_x^2}{4\alpha}}\int(u+\sigma)^{m+n}e^{-\alpha u^2}\mathrm{d}u\\ =\ &\sum_{m=0}^{l_i}\sum_{n=0}^{l_j}\sum_{p=0}^{\nu}\binom{l_i}{m}\binom{l_j}{n}\binom{\nu}{p}(-x_i)^{l_i-m}(-x_j)^{l_j-n}\sigma^{\nu-p}e^{\tau-\frac{q_x^2}{4\alpha}}\int u^pe^{-\alpha u^2}\mathrm{d}u.\end{aligned}$$

The integral $\int u^pe^{-\alpha u^2}\mathrm{d}u$ can be solved with some useful tricks. Consider the analogous integral,

$$H(l,\alpha):=\int x^le^{-\alpha x^2}\mathrm{d}x. \qquad (3.21)$$

If l is not a multiple of 2, the function is odd and the integral is zero. In the even case,

$$\int x^{2l}e^{-\alpha x^2}\mathrm{d}x\ =\ \frac{2\alpha}{2l+1}\int x^{2l+2}e^{-\alpha x^2}\mathrm{d}x \qquad (3.22)$$

Replacing l with $l-1$, the recurrence relation is,

$$\int x^{2l}e^{-\alpha x^2}\mathrm{d}x\ =\ \frac{2l-1}{2\alpha}\int x^{2l-2}e^{-\alpha x^2}\mathrm{d}x=\frac{(2l-1)!!}{(2\alpha)^l}\sqrt{\frac{\pi}{\alpha}}. \qquad (3.23)$$

Hence,

$$H(l,\alpha)=\begin{cases}\frac{(l-1)!!}{(2\alpha)^{l/2}}\sqrt{\frac{\pi}{\alpha}}, & \text{if } l\in 2\mathbb{N}.\\ 0, & \text{otherwise.}\end{cases} \qquad (3.24)$$

This allows Eq. (3.20) to be solved, giving the general analytical solution for the

integral of any two GTO products,

$$\int G(x)e^{iq_x x}\mathrm{d}x \tag{3.25}$$
$$= \sum_{m=0}^{l_i}\sum_{n=0}^{l_j}\sum_{p=0}^{\nu}\binom{l_i}{m}\binom{l_j}{n}\binom{\nu}{p}(-x_i)^{l_i-m}(-x_j)^{l_j-n}\sigma^{\nu-p}e^{\tau-\frac{q_x^2}{4\alpha}}H(p,\alpha),$$

which is a polynomial in q_x multiplied by a Gaussian, as expected for a Gaussian Fourier transform problem. Table 3.1 shows analytical Fourier integrands for GTO combinations up to $l_i + l_j = 4$. The numerical implementation of AIXRD therefore involves reading all required wave function coefficients from an *ab initio* electronic structure calculation, followed by the numerical evaluation of the above analytical expressions.

$l_i\ l_j$	$\hat{F}_x[G(x;l_i,l_j)](q)$
0 0	$G_{00}(q) = \sqrt{\frac{\pi}{\alpha}}e^{-\frac{\gamma_i\gamma_j}{\alpha}(x_i-x_j)^2}e^{iqx_G}e^{-s_x^2/4\alpha}$
1 0	$(a - x_i)\,G_{00}(q)$
2 0	$\left(a^2 + \frac{1}{2\alpha} - 2x_i a + x_i^2\right)G_{00}(q)$
1 1	$\left(a^2 + \frac{1}{2\alpha} - (x_i + x_j)a + x_i x_j\right)G_{00}(q)$
3 0	$\left\{a^3 - 3x_i a^2 + \left(3x_i^2 + \frac{3}{2\alpha}\right)a\right.$ $\left.-\frac{3x_i}{2\alpha} - x_i^3\right\}G_{00}(q)$
2 1	$\left\{a^3 - (2x_i + x_j)a^2 + (x_i^2 + 2x_i x_j\right.$ $\left.+\frac{3}{2\alpha}\right)a - \frac{2x_i+x_j}{2\alpha} - x_i^2 x_j\Big\}G_{00}(q)$
4 0	$\left\{a^4 - 4x_i a^3 + \left(6x_i^2 + \frac{3}{\alpha}\right)a^2\right.$ $\left.-\left(4x_i^3 + \frac{6x_i}{\alpha}\right)a + \frac{3x_i^2}{\alpha} + x_i^4 + \frac{3}{4\alpha^2}\right\}G_{00}(q)$
3 1	$\left\{a^4 - (3x_i + x_j)a^3 + 3\left(x_i^2 + x_i x_j\right.\right.$ $\left.+\frac{1}{\alpha}\right)a^2 - \left(x_i^3 + 3x_i^2 x_j + \frac{9x_i+3x_j}{2\alpha}\right)a$ $\left.+\frac{3x_i x_j+3x_i^2}{2\alpha} + x_i^3 x_j + \frac{3}{4\alpha^2}\right\}G_{00}(q)$
2 2	$\left\{a^4 - 2(x_i + x_j)a^3 + \left(\frac{3}{\alpha} + x_i^2 + x_j^2\right.\right.$ $\left.+4x_i x_j\right)a^2 - \left(2x_i^2 x_j + 2x_i x_j^2 + \frac{3(x_i+x_j)}{\alpha}\right)a$ $\left.+\frac{x_i^2+x_j^2+4x_i x_j}{2\alpha} + x_i^2 x_j^2 + \frac{3}{4\alpha^2}\right\}G_{00}(q)$

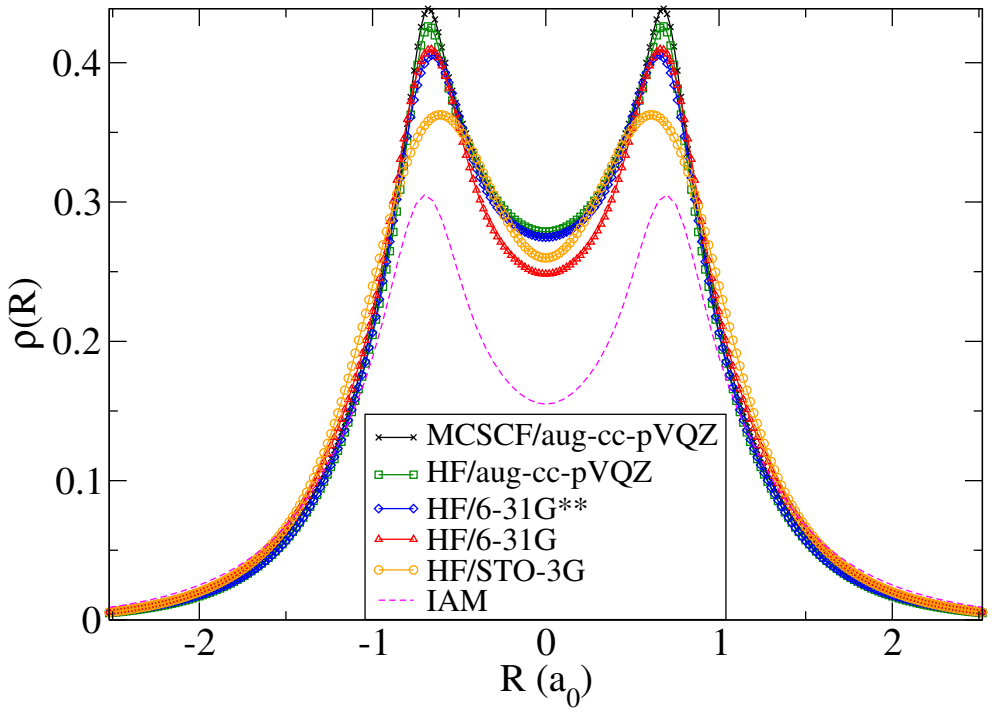
Table 3.1: Fourier transforms of the product of two 1D Cartesian Gaussian functions, $\hat{F}_x[G(x;l_i,l_j)](q)$, where $G(x;l_i,l_j) = (x-x_i)^{l_i}(x-x_j)^{l_j}e^{-\alpha(x-x_G)^2}$, with $\alpha = \gamma_i+\gamma_j$ and $x_G = (\gamma_i x_i + \gamma_j x_j)/\alpha$. The factor a in the table is $a = (iq/2\alpha) + x_G$.

3.2.4 Numerical approach

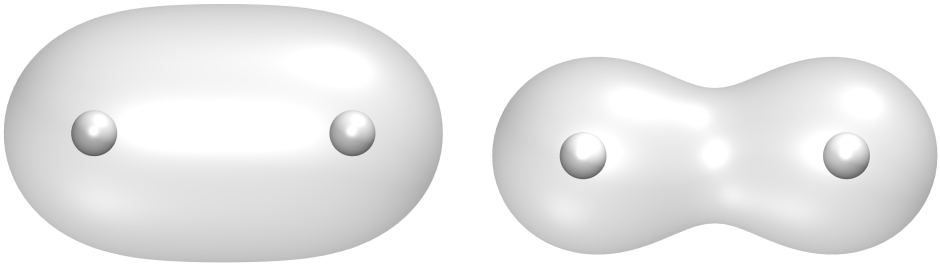
An alternative to the analytical approach detailed above is based on the direct Fourier transform of the electron density represented on a grid. In this case solving Eq. (3.11) requires that the electron density, Eq. (3.12), is computed on a 3D grid in $\mathbf{r}$, followed by use of a FFT algorithm. The choice of grid spacing and the size of the grid depends on the system being studied. In general, it is preferable to use as small a grid spacing Δr as computationally feasible (e.g. $\Delta r = 0.1\ a_0$), and a grid size such that as much of the electron density as possible (e.g. $> 99\%$) is contained within it, and the integral $\int \rho(\mathbf{r})\mathrm{d}\mathbf{r} \approx N_{\mathrm{elec}}$. This often means a very large grid with $N_r^3 \approx 200^3$ points, which introduces a costly overhead calculation in the overall computation. Note that the analytical solution described in Section 3.2.3 does not involve explicitly computing the electron density, effectively skipping this step, and directly evaluating Eq. (3.11) on a grid in reciprocal space $\mathbf{q}$ with the help of Eq. (3.25). The $\mathbf{q}$-grid can be much smaller than the $\mathbf{r}$-grid, such as $N_q^3 \approx 50^3$. The reason for this is that it always gives the correct number of electrons, as $\int \rho(\mathbf{r})\mathrm{d}\mathbf{r} = f(\mathbf{q}=0) \equiv N_{\mathrm{elec}}$, and at every point on the $\mathbf{q}$-grid the molecular form-factor is calculated exactly, whereas the FFT method has associated numerical error. A comparison of the accuracy and computational scaling for both methods is shown in Section 3.3.6. Numerically solving the Fourier transform of the electron density,

$$f_\alpha(\mathbf{q};\mathbf{R}) = \hat{F}_{\mathbf{r}}\Big[\rho_\alpha(\mathbf{r};\mathbf{R})\Big](\mathbf{q}) = \hat{F}_{\mathbf{r}}\Big[\sum_{j=1}^{N_{\mathrm{MO}}} a_j|\phi_j(\mathbf{r})|^2\Big](\mathbf{q}), \tag{3.26}$$

by FFT avoids summation over Gaussian products, as each spatial orbital $\phi_j(\mathbf{r})$ is evaluated on the grid, then each grid point is squared. That is, $\rho_\alpha(\mathbf{r};\mathbf{R})$ is a sum of N_{MO} grid-evaluated then squared sums of $N_{\mathrm{BF}} \times n_k$ Gaussian primitives. This compares with a scaling of $N_{\mathrm{MO}}(N_{\mathrm{BF}} \times n_k)^2$ for the analytical method. As such, the FFT method can be computationally faster under certain conditions. This is discussed further in Section 3.3.6.



(a)



(b) *Ab initio*.

(c) IAM.

Figure 3.1: (a) Electron density slices $\rho(R)$ for $H_2(X^1\Sigma_g^+)$ along its bond-axis, using Hartree-Fock (HF) with various basis sets and one multiconfigurational self-consistent field (MCSCF) method; the HF/aug-cc-pVQZ optimised geometry with bond-length 1.38658524 a_0 was used for all calculations. The MCSCF method used a (2,2) active space. (b,c) Electron density isosurfaces containing 87% of $\rho(\mathbf{r})$ for H_2 at its ground state geometry; calculated with (b) MCSCF(2,7)/aug-cc-pVTZ and (c) the independent atom model (IAM).

3.3 Results and discussion

3.3.1 Electron density

In *ab initio* electronic structure calculations, the use of large basis sets is important for obtaining the best wave functions and lowest energies. The more accurate

the wave function, the more accurate the electron density, and hence the predicted elastic scattering. Fig. 3.1 shows slices through the electron density for H_2 along its bond-axis. Each electron density was calculated using single-point energy HF theory using Molpro [96] with different basis sets at the HF/aug-cc-pVQZ optimised geometry. For comparison a reference electron density slice was calculated using MCSCF(2,2)/aug-cc-pVQZ. It is clear the use of the smallest basis set STO-3G achieves a very different electron density profile than the reference. The modest 6-31G basis electron density is similar to the reference close to the nuclei, but differs the most (aside from IAM) at the centre of the bond, $R = 0$. The calculations using the 6-31G** basis set and the HF/aug-cc-pVQZ are very similar to the reference density. However, preference is given to the 6-31G** basis as it gives the best trade-off between computational efficiency and accuracy in this instance. The IAM method is also shown for comparison. The electron density implicit in the IAM has been reverse engineered using isolated atomic densities calculated at the HF/6-31G level. Unsurprisingly, its $\rho(R)$ electron density distribution is the furthest from the reference. Figs. 3.1(b) and 3.1(c) show electron density isosurfaces containing 87% of the total electron density for the reference *ab initio* calculation and the IAM. The IAM only has radially symmetric density centred at each nucleus, whereas the *ab initio* density has much more bonding characteristics as it takes into account the bonding valence electrons. This is a fundamental difference between AIXRD and IAM, and the impact this has on the scattering will depend on the ratio of valence to core electrons. In the case of H_2, both its electrons are valence so taking them into account correctly becomes crucial. This effect diminishes for molecules containing atoms with larger atomic number Z, as will be shown later.

3.3.2 Atomic form-factors

Atomic form-factors are assumed to be radially-symmetric (corresponding to an isotropic electron density), and thus depend only on $q = |\mathbf{q}|$. They are well approximated by an analytical fit of a sum of N_g Gaussians of the form,

$$f(q) = \sum_{i=1}^{N_g} a_i \exp\left(- b_i \Big(\frac{q}{4\pi} \Big)^2 \right) + c, \tag{3.27}$$

where there are fitted, $\{a_i\}$, $\{b_i\}$, and c, for each atom in the international tables of crystallography [62]. Fig. 3.2 compares these tabulated form-factors to AIXRD based on HF/STO-3G and HF/6-31G wave functions. The atoms H, He, Li, Be, and Ne, are spherically symmetric as they either consist entirely of s-orbitals or are closed-shell. For atoms such as B, C, N, O, and F, there is slight angular anisotropy in the calculated electron density. For this reason, the AIXRD result $f_0(\mathbf{q})$ has been rotationally-averaged,

$$f_0(q) = \frac{1}{4\pi} \int_0^{2\pi} \int_0^{\pi} f_0(\mathbf{q}) \sin\theta_q \mathrm{d}\theta \mathrm{d}\phi_q, \tag{3.28}$$

as would be physically observable, where $\mathbf{q} = (q, \theta_q, \phi_q)$, and $q = |\mathbf{q}|$. The 6-31G results agree exceptionally with the reference curves, but the STO-3G results do not. This is a manifestation of the STO-3G basis failing to accurately approximate the atomic Slater-type orbitals. The larger 6-31G basis set is very similar to the International Table of Crystallography results [62], which is derived from Dirac-Slater relativistic wave functions [97, 98].

3.3.3 Molecular structure in diffraction

The focus of this section is on the interplay between molecular structure and the scattering signal in the gas-phase. The main aim is to show the utility and strengths of the AIXRD method in this context, and to compare it to the simple IAM method. The assumption of the IAM model is that the electron density for a molecule is the sum of its atomic densities. The scattering equivalent is that the molecular form-factor is the sum of its atomic form-factors. This fails to take into account the redistribution of valence electrons due to chemical bonding. Thus, it is a more accurate approximation for molecules containing heavier atoms because the ratio of valence electrons to total number of electrons is small. This effect is examined in Fig. 3.3, which compares diffraction intensities calculated by the IAM method and the AIXRD method for various molecules (H_2, CO, H_2O, O_3, C_2H_4, and C_4H_4). Fig. 3.3a shows side-by-side curves $I_0(q) = I(q)/I(q=0)$ for the IAM and AIXRD methods, and Fig. 3.3b compares the two in terms of their percent difference, defined as $\%\Delta I(q) = 100(I_{AI} - I_{IAM})/I_{IAM}$. Table 3.2 shows the mean and median values of the absolute percent difference $|\%\Delta I(q)|$ for each

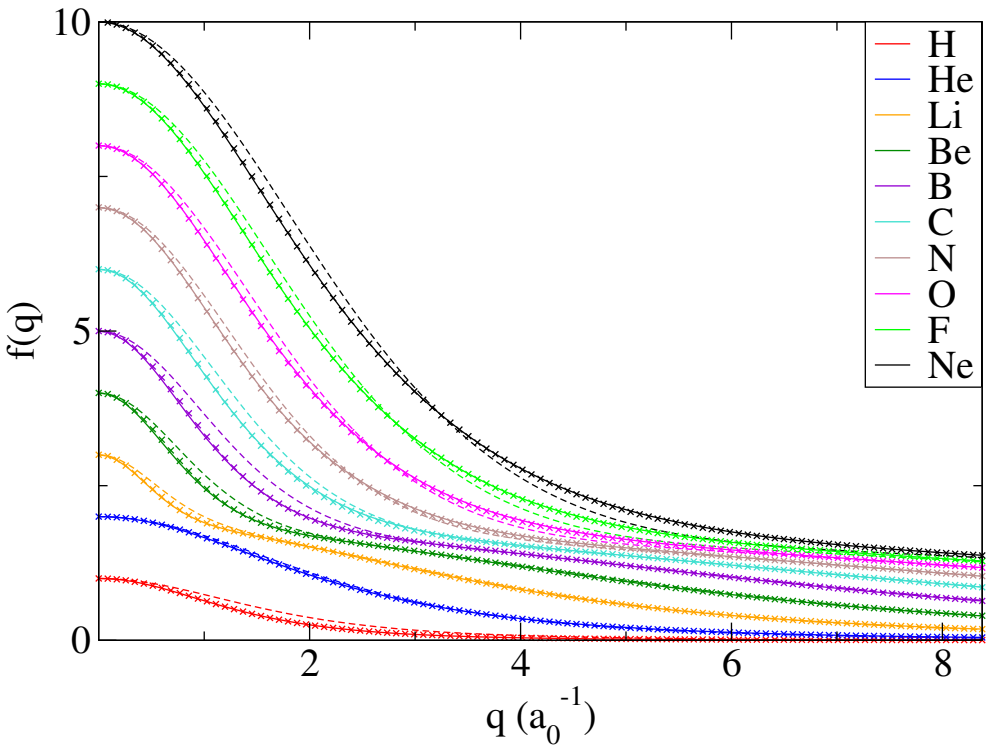
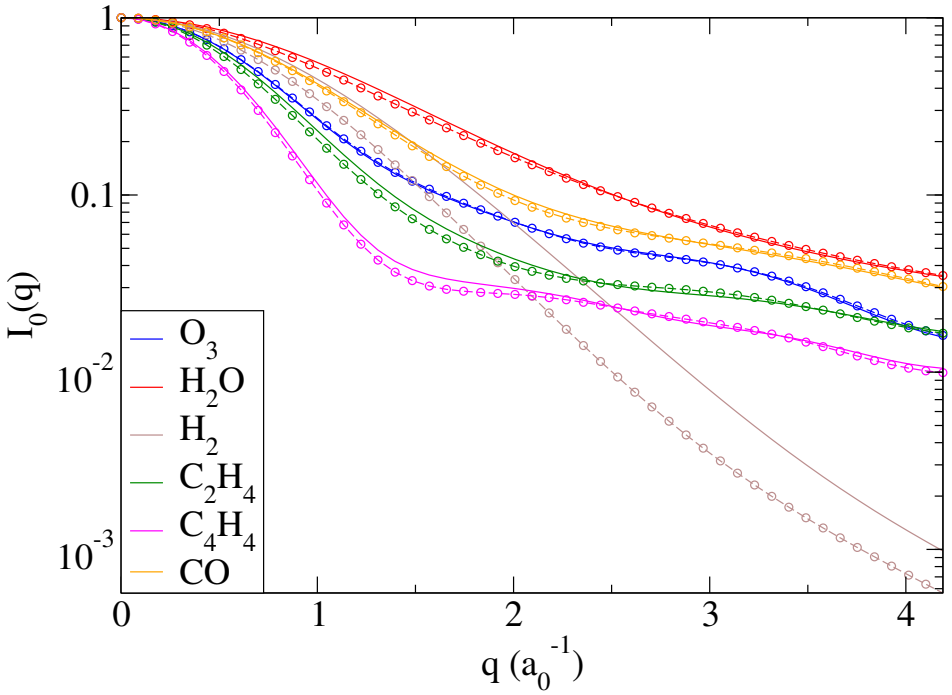


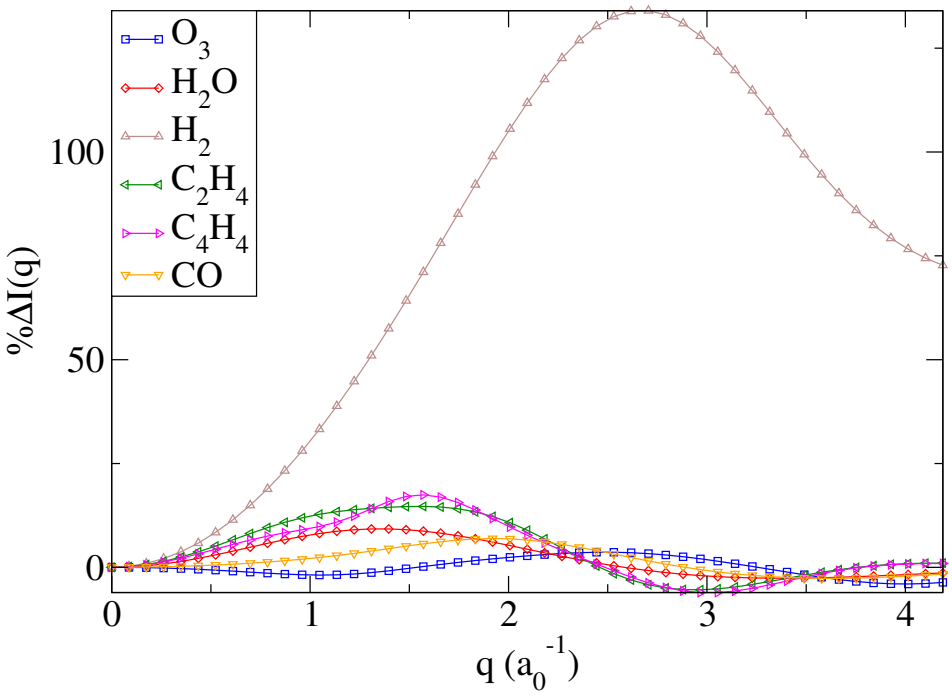
Figure 3.2: Atomic form-factors from the International Table of Crystallography [62] (lines), AIXRD using HF/6-31G (×'s), and AIXRD using HF/STO-3G (dashed). The kinematic X-ray form-factor at $q = 0$ is $f(q = 0) = \int \rho(\mathbf{r})\mathrm{d}\mathbf{r} \equiv N_{\text{elec}}$, where N_{elec} is the number of electrons. Note the exceptional similarity between the tabulated form-factors (lines) and the AIXRD using HF/6-31G calculated form-factors (×'s).

molecule. The clear outlier is the H_2 molecule, which has a mean $|\%\Delta I(q)|$ of 74.9%, whereas this value is <6.2% for the other chosen molecules. This is because the independent atom model is a poor approximation for H_2 since it contains only valence bonding electrons and the IAM approximation assumes the electrons are centred about the nuclei (as in Fig. 3.1c). The other mean values of $|\%\Delta I(q)| = (1.9, 6.1)\%$ are still significant and represent the percent error caused by not taking into account the redistribution of electrons by chemical bonding. Such errors are increasingly important with the rise of gas-phase X-ray diffraction experiments at XFELs. For example the time-resolved X-ray diffraction experiment of the photochemical ring-opening reaction of the CHD molecule [5,6] was possible based on being able to experimentally resolve pump-probe differences $I(q,t) - I(q,t=0)$ on the order of 1%.

In general, each nucleus within a molecule has many strongly associated core electrons, which means gas-phase X-ray diffraction signals are highly determined by the molecular geometry, i.e. the atomic positions. This is the underlying approximation of the IAM, where the diffraction only depends on the relative



(a)



(b)

Figure 3.3: (a) Comparison between orientationally-averaged AIXRD method (solid lines) and IAM approximation (dashed lines with circles), with $I_0(q) = I(q)/I(q=0)$. MCSCF optimised geometries and wave functions were used with the 6-311++G** basis. (b) Percent difference between AIXRD and IAM, $\%\Delta I(q) = 100(I_{\mathsf{AI}} - I_{\mathsf{IAM}})/I_{\mathsf{IAM}}$.

atomic positions within the molecule. An illustration of this strong dependence is shown in Fig. 3.4. X-ray diffraction patterns are calculated from two

Molecule	Mean $\lvert\%\Delta I(q)\rvert$	Median $\lvert\%\Delta I(q)\rvert$
CO	2.76	2.37
C_4H_4	5.94	4.87
C_2H_4	6.11	4.72
H_2	74.87	82.37
H_2O	3.74	2.53
O_3	1.89	1.77

Table 3.2: Complimentary table for Fig. 3.3. Mean and median values of $\lvert\%\Delta I(q)\rvert$ are shown; where $\%\Delta I(q) = 100(I_A - I_0)/I_0$.

stereoisomers of perfectly aligned butadiene (C_4H_6), cis- and transbutadiene, as well as cyclobutadiene (C_4H_4), which is the Jahn-Teller distorted closed-ring version of butadiene [99]. The molecular geometries were optimised via the multiconfigurational self-consistent field (MCSCF) method [100] using Molpro [96] at the MCSCF(4,4)/6-31G** model chemistry, with an active space consisting of four electrons within four π-orbitals. The energies for each geometry optimised butadiene form are $E_{cis} = -154.980896$, $E_{trans} = -154.985878$, and $E_{cyclo} = -154.716941$ Hartree. The same *ab initio* wave functions were used to analytically calculate the corresponding diffraction patterns using AIXRD. From here on the maximum elastic scattering angle $\theta = \pi$ corresponding to $q = 9.67$ Å^{-1}, where $q = (4\pi/\lambda)\sin(\theta/2)$, which is equivalent to a 1.3 Å X-ray wavelength. The molecular point groups for trans-, cis-, and cyclobutadiene are C_{2h}, C_{2v}, and D_{2h} respectively. This is apparent in their diffraction patterns, which in general share symmetry elements as long as the molecules are assumed to be perfectly aligned, as they are here.

Considering the strong influence of the molecular geometry on the diffraction pattern, the question arises, is it worth going beyond the IAM? Fig. 3.5a shows the difference between geometry optimised MCSCF(6,6)/6-31G** level of theory, with six electrons in six π-orbitals, and keeping the geometry fixed, the IAM diffraction for the benzene molecule. For additional perspective, diffraction at this level of theory and geometry is compared to AIXRD using Hartree-Fock (HF) and the STO-3G basis set. The MCSCF/6-31G** energy is $E_{\text{multi}} = -230.786833$ Hartree. The C-C bond lengths are 1.396 Å. The maximum difference between MCSCF and IAM diffraction is approximately 45% in Fig. 3.5a, calculated by the equation $100(I_{\text{MCSCF}} - I_{\text{HF}})/I_{\text{HF}}$. Most of the difference pattern is approximately

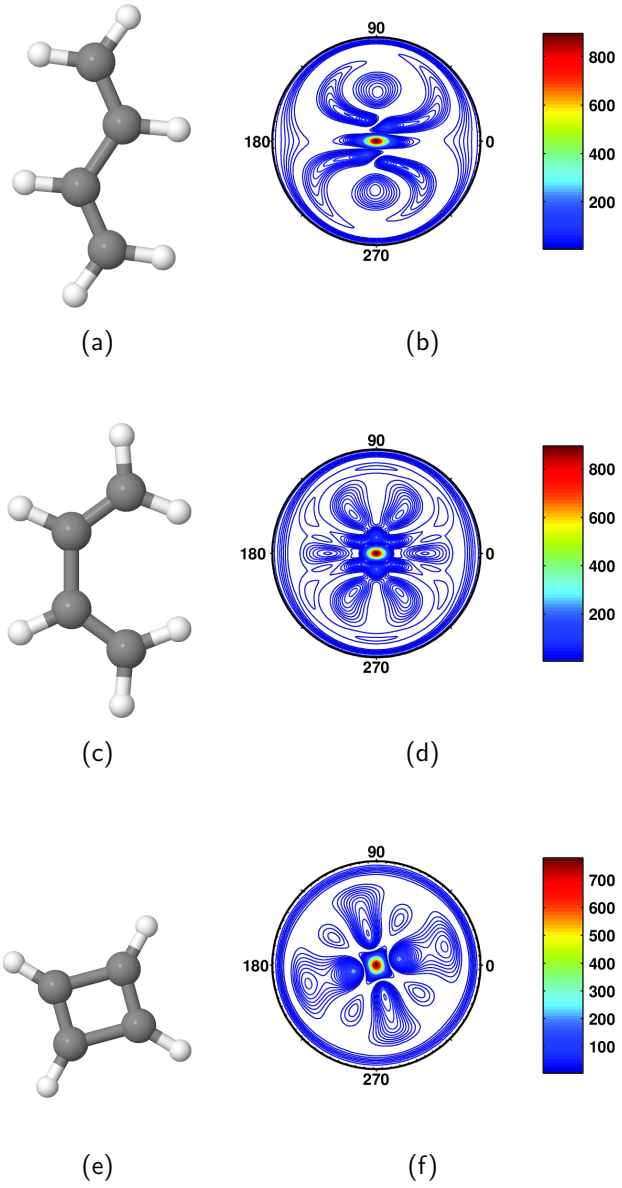
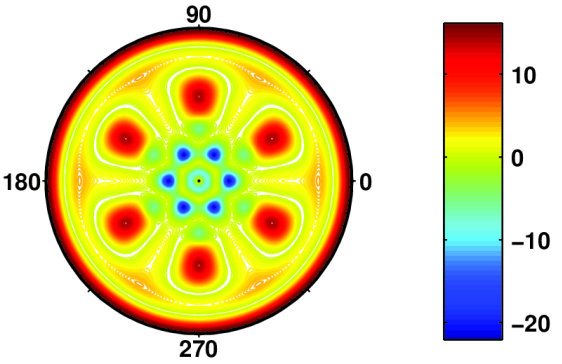


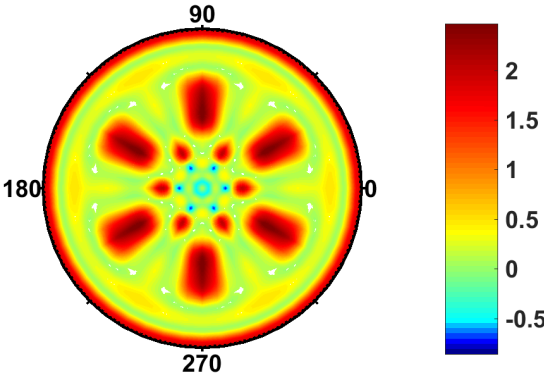
Figure 3.4: X-ray (1.3 Å) diffraction from multiconfigurational *ab initio* using 6-31G(d,p) basis for (a,b) trans-, (c,d) cis-, and (e,f) cyclo-, butadiene forms. Each pattern has the same point group as the projection of the molecule onto a plane perpendicular to incoming X-rays. Exceptionally, the cis-butadiene pattern (d) has 2 planes of mirror symmetry. This is because mirror images of this molecule give the same diffraction patterns. The central peak in all patterns has intensity n^2_{elec}, as $|\int \rho(\mathbf{r}) e^{i\mathbf{qr}} \mathrm{d}\mathbf{r}|^2$ has maximum value $|\int \rho(\mathbf{r}) \mathrm{d}\mathbf{r}|^2 = n^2_{\text{elec}}$. This is not the case in experiment as the central peak, usually behind beam-stopper, has intensity $I_{\text{X-ray}}$.

0 except for particular regions in the pattern (i.e. at the detector).

The maximum difference between the MCSCF(6,6)/6-31G** and HF/STO-3G diffraction is significantly smaller, approximately 9%, in Fig. 3.5b. As this dif-



(a)



(b)

Figure 3.5: Benzene, (a) MCSCF(6,6) 6-31G(d,p) compared to IAM difference pattern, $|f_{\text{6-31G(d,p)}}|^2 - |f_{\text{IAM}}|^2$, which corresponds to $\approx 45\%$ maximum difference, and (b) benzene MCSCF(6,6) 6-31G(d,p) compared to HF STO-3G difference pattern, $|f_{\text{6-31G(d,p)}}|^2 - |f_{\text{STO-3G}}|^2$, which corresponds to $\approx 9\%$ maximum difference.

ference is much lower than the previous comparison, it implies that taking into account the valence electrons is more significant than enhancing the model chemistry. The main difference in fact manifests itself in the basis set difference, not so much in the difference between the electronic structure method (HF or MCSCF). This is because the STO-3G basis set is generally inadequate and should only be used for quick or benchmark calculations at a significant cost to accuracy. Also the difference between MCSCF and HF lies in the active space, which is rather small in comparison to the total number of electrons, and the difference it causes is therefore minimal. In other words, allowing six electrons to move slightly into other orbitals is insignificant compared to the background signal from the total 42 electrons.

The computational time required for an *ab initio* X-ray diffraction calculation is proportional to the number of molecular orbitals which construct the total wave function, and the number of Gaussian primitives (GTOs) per MO squared, $n_g^2 N_{\mathrm{MO}}$, where $n_g = n_k N_{\mathrm{BF}}$. MCSCF methods include additional MOs for electrons to occupy compared to Hartree-Fock (i.e. the active space). Therefore, extra basis functions must be used to take account of these, and an AIXRD calculation based on MCSCF requires a linear multiple of the time taken for one based on HF. In addition, each basis set can contain different numbers of GTOs. For example, STO-3G uses 3 GTOs per atomic orbital (AO) to describe an atom, whereas 6-31G* uses six GTOs per core AO, and contractions containing 3 and 1 GTOs for valence AOs. This means there are significantly more Gaussian products per MO. For diffraction from molecules in their electronic ground state, HF/STO-3G may be useful as a quick calculation that includes valence electrons. However, this computational speed is lost when multiconfigurational methods are needed. This is because often they need large basis sets to converge to an accurate wave function. This is especially true for electronically excited states. Whether or not it may be possible to determine the electronic state of a molecule via elastic X-ray diffraction is discussed in the next section.

3.3.4 Electronic structure in diffraction

AIXRD can be used to predict molecular form-factors, thus diffraction patterns, for molecules in electronically excited states, and from individual molecular orbitals. The latter is shown in Fig. 3.6. Singly occupied σ and π-orbitals are shown, with corresponding diffraction patterns, for O_2 in its electronic ground state calculated at the MCSCF(8,8)/6-31G** level, with an active space consisting of eight electrons in four π and four σ-orbitals. The geometry optimised energy is $E_{O_2} = -149.663091$ Hartree, with a bond-length of 1.245 Å. There are characteristic diffraction patterns for each MO type, such as σ, $\sigma*$, π, and $\pi*$, which can be determined by AIXRD.

Theoretically, X-ray diffraction can detect electronic state changes, as the elastic X-ray form-factor is the Fourier transform of the electron density (Eq. (3.11)), which in turn depends on the electronic state. Experimental verification of this is challenging for several reasons. A high signal to noise ratio is necessary, since the changes in electron density are small. Often, only a small number of valence electrons are reconfigured during excitation to a different electronic state. In

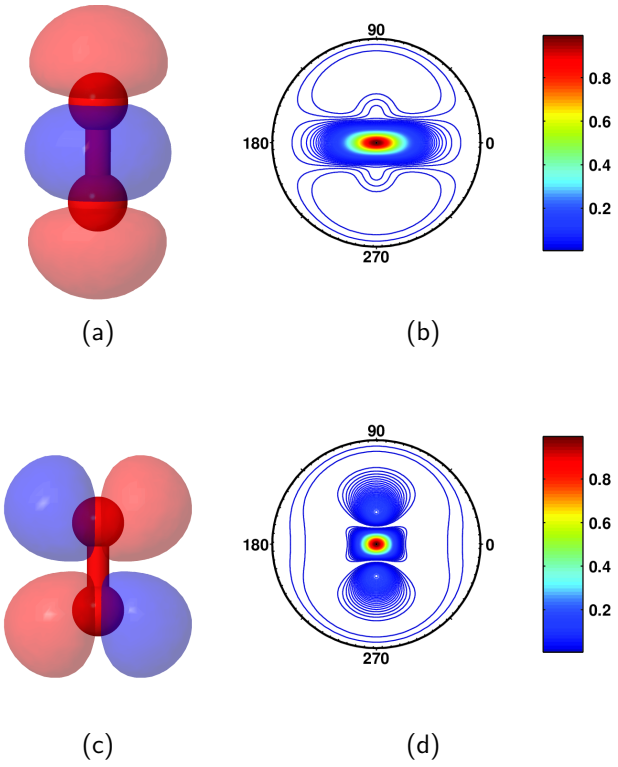
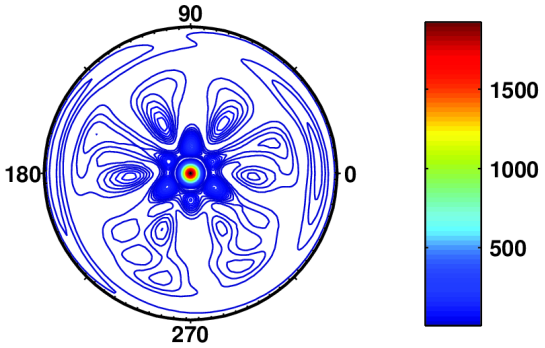


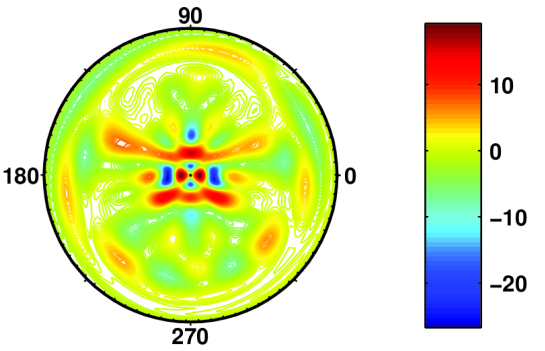
Figure 3.6: O_2 various molecular orbitals, (a) σ-bond, red and blue represent different phases, (b) corresponding normalised diffraction pattern, (c) π^*-bond (π-antibond), and (d) corresponding normalised diffraction pattern.

practice, excitation usually comes with associated nuclear geometry changes, and the two are not easily distinguishable. Furthermore, a pump-laser only excites a fraction of target molecules, leaving a fractionally diminished difference signal after subtraction [5, 6, 76]. Finally, electronic states have finite lifetimes, which may make inelastic effects more important [101, 102].

The diffraction pattern for the first singlet excited state, S_1, vertically excited from the ground state, S_0, of perfectly aligned 1,3-cyclohexadiene (CHD) is shown in Fig. 3.7a. The S_0 geometry optimisation was performed at the MCSCF(6,6)/6-31G** level of theory, with an active space consisting of six electrons within four π-orbitals and 2 σ-orbitals, and the S_0 and S_1 wave functions were calculated at this geometry. They have energies of $E_{S_0} = -231.916756$ Hartree and $E_{S_1} = -231.668618$ Hartree, respectively. Fig. 3.7b shows the difference pattern between S_1 and S_0 states. Although near-zero for most detector regions, at certain angles this difference is up to 40%, as calculated by the equation $100(|f_1|^2 - |f_0|^2)/|f_0|^2$. This example is analogous to a 'perfect experiment' with 100% vertical excitation from S_0 to S_1 state, and perfect alignment i.e. no blur-



(a)



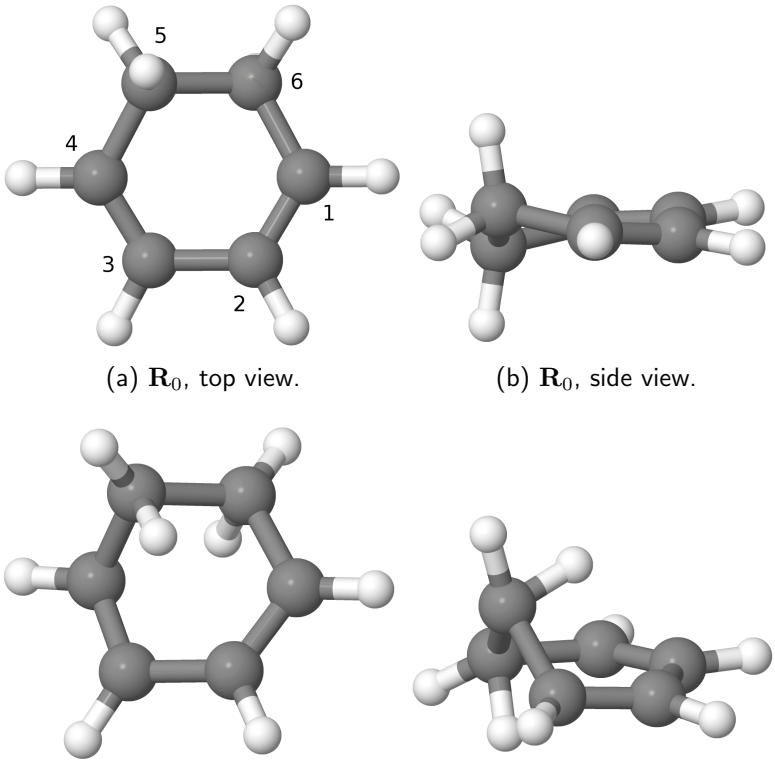
(b)

Figure 3.7: 1,3-Cyclohexadiene vertical excitation, (a) 1st singlet excited state, $|f_1|^2$, and (b) difference pattern $|f_1|^2 - |f_0|^2$, red areas are characteristic of the excited state and blue areas are ground state remnants.

ring due to vibrational and rotational motion. Signature diffraction differences for identifying an electronic state may remain with reasonable and even non-existent alignment. This hypothesis will be explored in detail in Chapters 5 and 6.

3.3.5 Fitted rotationally-averaged form-factors for solvent molecules

The molecular form-factor is a key variable in X-ray diffraction experiments. Its absolute square gives the intensity distribution of scattered photons as a function of the momentum transfer, $\mathbf{q}$. Crystallographic data often contains significant signal from solvent molecules, which are dispersed throughout the sample. The



(a) $\mathbf{R}_0$, top view. (b) $\mathbf{R}_0$, side view.

(c) $\mathbf{R}_1$, top view. (d) $\mathbf{R}_1$, side view.

Figure 3.8: CHD geometries top and side view schematics, (a) top view of global minimum geometry, $\mathbf{R}_0$, with numbered carbons; (b) corresponding side view; (c,d) puckered ring geometry, $\mathbf{R}_1$, top and side views respectively.

use of a form-factor derived for the entire rotationally-averaged molecule could therefore be useful to subtract out or otherwise infer extra structural information about the crystal sample. The molecule 1,3-cyclohexadiene (CHD) is used here as an example solvent molecule.

Using Molpro [96], MCSCF(4,4)/6-31G** geometry optimisation was performed to find two ground state stationary points for CHD, the global minimum, $\mathbf{R}_0$, ($E = -231.90032678$ a.u.) and a local 'puckered ring' minimum, $\mathbf{R}_1$, ($E = -231.71508274$ a.u.). A schematic of each is shown in Fig. 3.8 and their geometric parameters are shown and compared to experimental electron diffraction values [103] in Table 3.3. AIXRD was used at these geometries with the corresponding MCSCF(4,4)/6-31G** wave functions to compute the rotationally-averaged elastic X-ray diffraction curves for the range $q = [0, 5.8860]$ Å^{-1},

$$I(q) = \frac{1}{4\pi} \int_0^{2\pi} \int_0^{\pi} |f(q, \theta, \phi)|^2 \sin\theta \mathrm{d}\theta \mathrm{d}\phi. \tag{3.29}$$

Parameter	Experiment	$\mathbf{R}_0$	$\mathbf{R}_1$
C_2C_3 (Å)	1.468 ± 0.014	1.471	1.483
C_2C_1 (Å)	1.350 ± 0.004	1.330	1.361
C_1C_6 (Å)	1.523 ± 0.016	1.512	1.535
C_6C_5 (Å)	1.534 ± 0.020	1.538	1.599
$<C_3C_2C_1$	120.13 ± 0.6	120.65	114.40
$<C_2C_1C_6$	120.14 ± 0.5	120.77	110.05
$<C_1C_6C_5$	110.7	111.85	92.31

Table 3.3: Experimental and theoretical geometric parameters for CHD. The experimental values are from [103], and the theory used is MSCSF(4,4)/6-31G** geometry optimisation using Molpro [96].

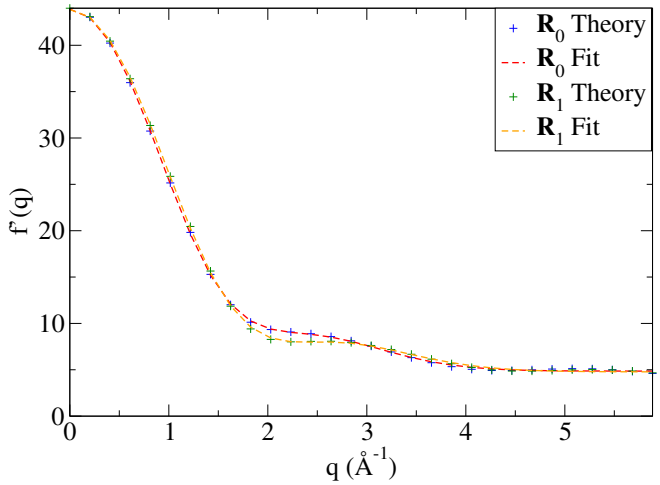


Figure 3.9: Rotationally-averaged theoretical form-factors for the CHD molecule, and corresponding fitted curves for the global minimum geometry $\mathbf{R}_0$ and a local 'puckered-ring' minimum geometry $\mathbf{R}_1$.

The square-root was taken to approximate a real-valued molecular form-factor, $f'(q) = \sqrt{I(q)}$. The fitting equation for atomic form-factors (Eq. (3.27)) was used, and by variation of the coefficients $\{a_i\}, \{b_i\}$, and c and minimising the mean of the difference,

$$\langle \Delta f \rangle = \frac{1}{N} \sum_{i=1}^{N} |f'(q_i) - f(q_i)|, \tag{3.30}$$

where N is the number of points computed in q, an optimised fitted molecular form-factor $f(q)$ is obtained. The theoretical and corresponding fitted curves are

Coefficient	$\mathbf{R}_0$	$\mathbf{R}_1$
a_1	570.1466	625.5978
b_1	108.7478	97.5134
a_2	241.0022	219.6332
b_2	57.4738	50.6245
a_3	-554.7766	-546.6454
b_3	75.6884	67.6380
a_4	-217.3425	-259.4986
b_4	139.7514	122.9526
c	4.8595	4.7768

Table 3.4: Curve fitting parameters for CHD stationary point geometries $\mathbf{R}_0$ and $\mathbf{R}_1$ using Eq. (3.27).

shown in Fig. 3.9 for the $\mathbf{R}_0$ and $\mathbf{R}_1$ molecular geometries, and the optimised fitting coefficients are shown in Table 3.4.

The utilisation of pre-fitted, rotationally-averaged form-factors such as for CHD in the present example is currently being investigated in collaboration with Prof. Simon Parsons (University of Edinburgh) in the refinement of X-ray scattering data from metal-organic frameworks (MOFs), where solvent molecules are known to disperse inside the cavities in a quasi-structured manner.

3.3.6 Numerical calculations using FFT

As briefly mentioned in the theory section, an alternative fast Fourier transform (FFT) method can be used to calculate the elastic molecular form-factor. It involves representing the electron density on an appropriate grid with regular spacing Δr, then, numerically Fourier transforming via an FFT algorithm. This approach benefits from the efficiency of the FFT algorithm, and from conceptual simplicity. To converge to the analytical solution, a small grid spacing and large grid size are crucial. This means significant computational effort is required to produce accurate results.

Vital to the accuracy of the FFT algorithm is that the signal to be Fourier transformed is within the Nyquist critical frequency, $q_c = 1/2\Delta r$. Equivalently,

$N_r = N_q$	N_{grid}	n_g^{crit}
50	512	500
50	1024	1700
50	2048	5500
100	512	200
100	1024	600
100	2048	2000

Table 3.5: Timing comparison between analytical and FFT methods for calculation of the molecular form-factor, as Eq. (3.11). In the FFT method, the electron density is represented on a N_r^3 sized grid with equal spacing Δr. The FFT algorithm requires extra 'padding' of trailing zeros to make a grid of size, $N_{\text{grid}}^3 = (2^m)^3$, where $m \in \{0, 1, 2, \dots\}$ and $N_{\text{grid}} > N_r$. It has scaling of $n_g N_{\text{MO}} N_r^3 + (N_{\text{grid}} \ln N_{\text{grid}})^3$, for N_{MO} MOs and $n_g = n_k N_{\text{BF}}$ GTOs per MO. Grid sizes for powers of 2^m where $m = \{9, 10, 11\}$ are shown, which are generally appropriate for such calculations, and the larger the value the more accurate the FFT. The analytical method directly calculates the elastic molecular form-factor from an *ab initio* wave function on a N_q^3-sized grid in $\mathbf{q}$, and has scaling $n_g^2 N_{\text{MO}} N_q^3$. Shown are the critical numbers of Gaussian primitives (GTOs) per MO, n_g^{crit}, rounded to the nearest 100, for each calculation method to take equal computational time, below which the analytical calculation is faster.

this means the electron density must be represented on a grid with sufficiently small spacing, Δr, to produce accurate results. Often the Nyquist frequency is greater than $2|\mathbf{k}_0|$, so a large number of grid points are needed to accurately resolve the experimentally relevant region, $0 \leq |\mathbf{q}| \leq 2|\mathbf{k}_0|$. This is done by enlarging the electron density grid with trailing zeros, while ensuring each spatial dimension has length equal to a power of 2 for optimal FFT performance [104]. The 3D FFT algorithm scales as, $(N_{\text{grid}} \ln N_{\text{grid}})^3$. Thus, including the initial calculation of the N_r^3 sized electron density grid, the overall scaling of the FFT method to calculate elastic molecular form-factors is, $n_g N_{\text{MO}} N_r^3 + (N_{\text{grid}} \ln N_{\text{grid}})^3$. Comparing to the 3D analytical method scaling, $n_g^2 N_{\text{MO}} N_q^3$, shows that the FFT method is preferable for n_g above a certain critical value, n_g^{crit}, with appropriately chosen constants, N_{grid}, N_r, and N_q. For $N_{\text{grid}} = 2048$, and $N_r = N_q = 100$, n_g must be greater than about 2000 for the FFT to be worthwhile in terms of efficiency. For small molecules with reasonably sized basis sets, $n_g < 2000$, and the analytical method is preferable. The analytical route also has the benefit of guaranteeing exact solutions at each chosen value of $\mathbf{q}$. Values of n_g at which the FFT and analytical method take equal computational time are shown in Table 3.5 for other values of N_{grid}, N_r, and N_q. Note that in practice, $N_q < N_r$, because

the grid in **r** must be fine and large enough to accurately encompass most of the electron density (e.g. $> 99\%$), whereas the grid in **q** does not suffer from this problem. It only has to be calculated precisely at the desired points in **q**, and extend as far as an experimental setup can reach. Both grids are set to be equal in size here, which is a fairly extreme example. This means that n_g^{crit} is actually a lot larger than shown in Table 3.5 and the analytical method is computationally faster for a far more extensive range of n_g values. It is also always more accurate because it does not incur the numerical errors inherent in FFT. Thus, even at slightly greater than n_g^{crit} GTOs per MO, the analytical method may still be preferable.

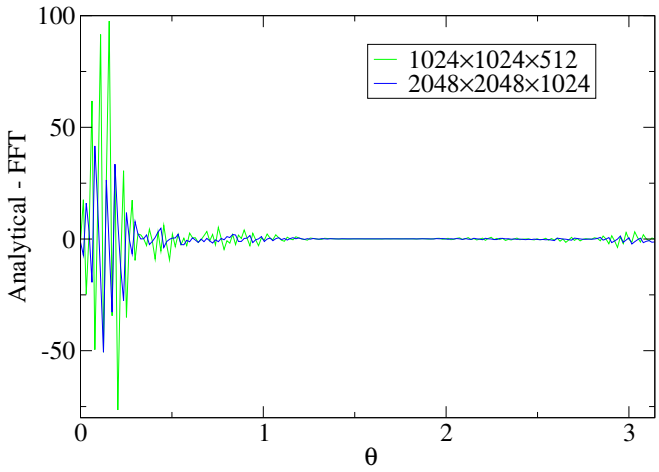


Figure 3.10: Convergence of numerical calculations using FFT. The difference between the diffraction calculated by numerical FFT and the analytical *ab initio* approach. The comparison is made for the ground state benzene molecule as a function of the radial angle θ with azimuthal angle $\phi = 0$. The FFT has been calculated on two different sized electron density grids, $1024 \times 1024 \times 512$, and $2048 \times 2048 \times 1024$. The error in the FFT calculation, relative to the analytical calculation, is on the order of 2.7% for the larger grid.

The convergence of the FFT towards the analytic result, as a function of grid size, is shown in Fig. 3.10. The largest grid in Fig. 3.10 reproduces the analytic result with approximately 2.7% average difference. Calculations of the X-ray diffraction signal were performed for ground state benzene with the MCSCF(6,6)/6-31G** model chemistry. Fast-Fourier transform methods suffer from 'artificial interference', due to periodicity of discrete Fourier transforms. Deviations from correct values appear close to edge regions because of this. This is mitigated by increasing the number of trailing zeros, and the initial grid density. However, there is a practical limit to the amount this can be done, due to the $(N_{\text{grid}} \ln N_{\text{grid}})^3$ scaling.

3.4 Conclusion

The diffraction pattern for molecules in various electronic states can be efficiently and directly calculated from *ab initio* multiconfigurational wave functions expressed by Gaussian basis sets. Single determinant wave functions, such as those calculated from HF theory, can be used in conjunction with AIXRD for modelling the X-ray diffraction from molecules in their ground state. The calculation is based on the derived general analytical formula for the Fourier transform of any two Cartesian GTO products. In general, analytical solutions are sought out over their numerical counterparts for many scientific problems, due to their efficient and exact nature.

Diffraction patterns calculated from perfectly aligned butadiene isomers show large variations in signal at all scattering angles >0, demonstrating the potential of X-ray diffraction from aligned molecules to track structural changes. A comparison is made between the *ab initio* X-ray diffraction (AIXRD) method and the independent atom model (IAM) for benzene in its ground state. As expected, the main difference stems from accounting for the delocalised valence electrons, and a maximum difference of 45% is seen at specific scattering angles. The reference X-ray diffraction pattern was calculated from a MCSCF(6,6)/6-31G** wave function, which was compared to a HF/STO-3G calculated diffraction pattern revealing a maximum difference of 9%, and about a four times decrease in computational effort. Thus, for electronic ground states IAM can be improved at low computational cost using AIXRD with a small basis set and single determinant wave function, and enhanced accuracy can be achieved straight-forwardly by using larger basis sets.

Multiconfigurational electronic structure methods are necessary to accurately obtain excited state wave functions and energies. These methods require large basis sets to converge, thus large computational effort to calculate their corresponding AIXRD diffraction patterns. The diffraction patterns for 1,3-cyclohexadiene in its first singlet excited state and its ground state are compared. This shows that changes in electronic structure, e.g. during a photochemical reaction, are in principle visible with X-ray diffraction, at least for aligned molecules. The interpretation of experimental data remains complex, especially in terms of separating contributions from nuclear motion and electronic structure changes. Theoretical calculations such as AIXRD can help to elucidate these distinctions.

A hierarchy of approaches to calculating elastic X-ray molecular form-factors has been outlined. The fastest by far is the independent atom model which merely scales with the number of atoms. However, it can only be used for the electronic ground state and does not take account of delocalised valence electrons. Single determinant *ab initio* X-ray diffraction with a small basis set such as HF/STO-3G can be used at modest computational investment. It is also essentially constrained to the electronic ground state but includes delocalised and bonding valence electrons. To model the X-ray diffraction from electronically excited states and improve the overall *ab initio* wave function, multiconfigurational methods with larger basis sets are needed. In this case, one of two AIXRD methods are viable: the numerical FFT; or the analytical approach. The analytical method is more accurate, and computationally faster except when a very large number of GTOs have been used to construct the total wave function.

Chapter 4

Fragment-Based *Ab Initio* X-ray Diffraction from Biomolecules

4.1 Introduction

Since Max von Laue's discovery of X-ray diffraction from crystals in 1914 [8] there have been many Nobel prizes related to X-ray crystallography. Today, it remains one of the foremost techniques for determining the structure of molecules, and its success has greatly advanced the understanding of molecular structure and function. However, there are many biomolecules which are difficult or impossible to crystallise, and therefore difficult to study using X-ray crystallography [34,105]. In addition, the growth of large well-ordered protein crystals is a major bottleneck in structure determination by X-ray crystallography.

X-ray free-electron laser (XFEL) sources are so bright that they overcome the need for conventional crystals. Serial femtosecond crystallography [35, 36, 106], where a liquid jet containing 100,000s of micro-crystals intersects the femtosecond X-ray source, circumvents the need to grow large crystals. This gives many snapshots of microcrystals in different orientations, and a computer algorithm must be used to extract meaningful information from the data to determine the molecular structure. An example of this is the *de novo* determination of the structure of lysozyme with a 2.1 Å resolution [35]. In addition, the problem of X-ray radiation damage and degradation of the sample is avoided because of diffraction before destruction [30]. That is the XFEL pulses outrun the destruction of the micro-crystals, giving damage-free diffraction snapshots. This is in contrast to

conventional crystallography, where the protein crystal accumulates damage over several hours of data collection time.

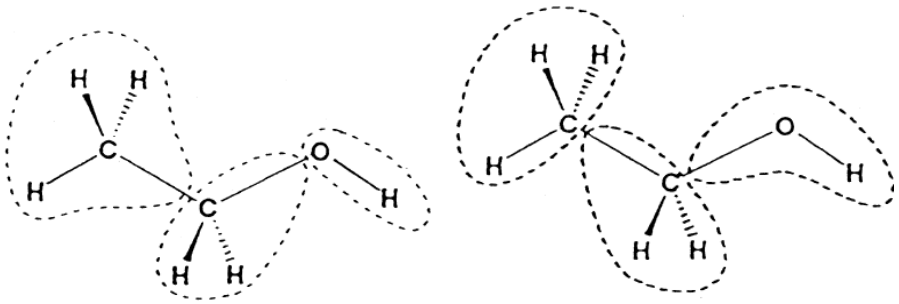
Fragment-based electronic structure methods have opened the way for quantum mechanical treatment of large molecules, such as biomolecules [107]. There are many fragment-based electronic structure methods [108], such as fragment molecular orbital (FMO) theory [109], subsystem DFT [110], effective FMO [111, 112], 3-body FMO [113], and multiconfigurational FMO [114]. Many popular fragment-based methods are implemented in the electronic structure program GAMESS [113]. Fragment-based methods make possible nearly linear scaling of calculations of large molecular systems [115], such as water clusters, proteins, and DNA. In addition to this, the capacity of *ab initio* electronic structure methods to be used on larger molecules is improving alongside the growth in available computational resources. Recently, full *ab initio* geometry optimisation was performed on 55 protein structures [116].

It is intuitive to combine a fragment-based electronic structure method with X-ray diffraction theory. In 1916, Debye hypothesised that a sum of atomic form-factors, with phase terms dependent on the positions of the nuclei, is a good approximation for the total molecular form-factor [38]. In the current context, his method can be thought of as a simple fragment-based method with isolated atomic fragments, where X-rays scatter from radially-symmetric distributions of electron density centred at each atom. Combining *ab initio* X-ray diffraction (AIXRD), as outlined in the previous Chapter, with a fragment-based method allows the calculation of fragment form-factors which are based on electronic structure theory, and summing these fragment form-factors is an approximate way to calculate the total molecular form-factor of a large molecule. Structure determination, and TRXRD experiments could benefit from molecular form-factors based on *ab initio* electronic structure theory. For example, serial crystallography can theoretically achieve very high resolution structure determination due to the mitigation of radiation damage and the unprecedented peak intensity of XFEL pulses. X-rays probe the electron distribution rather than the atomic positions, therefore it may be possible to probe subtleties in electronic structure, such as those from delocalised electrons [1, 117]. In this new age of X-ray structure determination, it is worthwhile thinking about going beyond the IAM approximation in the calculation of the X-ray diffraction signal.

4.2 Theory

The following theory briefly describes fragment molecular orbital (FMO) theory [109], which is an inspiration for this Chapter. The idea that follows involves combination of a fragment-based *ab initio* electronic structure method with the AIXRD method described in Chapter 3. Ideally, the true FMO electron density would be Fourier transformed to obtain the total molecular form-factor. However, a similar yet less sophisticated model is used in this Chapter, where entirely isolated fragment form-factors are used. The method here is analogous to the IAM method, except instead of isolated atomic fragments, larger fragments are used. For example, each amino acid residue in the protein lysozyme is defined as a fragment in Section 4.3.4. The problem with this is that each fragment is defined by severing a bond, and as such, is a radical. There are several ways to avoid the radicalisation of each fragment, e.g. by simply introducing fragment pairs or *dimers* which include the unsevered bond. This is briefly explored in the following discussions. The method arrived upon in this Chapter is an intermediate method between the IAM and AIXRD methods. It is called the Independent Fragment Model (IFM).

4.2.1 Fragment molecular orbital theory



(a) IFM1. (b) IFM2.

Figure 4.1: Ethanol divided into two different sets of $N = 3$ fragments, denoted IFM1 and IFM2. From Mulliken population analysis using HF/6-31G, the OH group a net negative charge of $-0.35388e$, and the CH_3 group a net positive charge of $+0.45377e$. Images taken from Kitaura et al. [109].

In FMO theory [109], a molecule is divided into fragments. See Fig. 4.1 for

reference. The fragment Hamiltonian is defined as,

$$H_I = \sum_i^{n_I} \Big\{ -\frac{1}{2}\nabla_i^2 - \sum_s^{N_{\rm at}} \frac{Z_s}{|\mathbf{r}_i - \mathbf{R}_s|} + \sum_{J\neq I}^{N} \int \mathrm{d}\mathbf{r}' \frac{\rho_J(\mathbf{r}')}{|\mathbf{r}_i - \mathbf{r}'|} \Big\} + \sum_{i>j}^{n_I} \frac{1}{|\mathbf{r}_i - \mathbf{r}_j|}, \tag{4.1}$$

where n_I is the number of electrons in fragment I, $N_{\rm at}$ is the total number of atoms in the molecule, Z_s is the nuclear charge on atom s, $\{\mathbf{r}_i\}$ are the n_I electronic coordinates within fragment I, $\{\mathbf{R}_s\}$ are the $N_{\rm at}$ nuclear coordinates within the molecule, N is the number of fragments, and $\rho_J(\mathbf{r}')$ is the electron density of fragment J. The electron densities for each fragment are calculated in an iterative manner. Dimers are defined as two fragments (or monomers) together. The dimer Hamiltonian is analogous to the fragment Hamiltonian,

$$H_{IJ} = \sum_i^{n_I+n_J} \Big\{ -\frac{1}{2}\nabla_i^2 - \sum_s^{N_{\rm at}} \frac{Z_s}{|\mathbf{r}_i - \mathbf{R}_s|} + \sum_{K\neq I,J}^{N} \int \mathrm{d}\mathbf{r}' \frac{\rho_K(\mathbf{r}')}{|\mathbf{r}_i - \mathbf{r}'|} \Big\} + \sum_{i>j}^{n_I+n_J} \frac{1}{|\mathbf{r}_i - \mathbf{r}_j|}. \tag{4.2}$$

The electron densities are not recalculated for the dimers, the fragment densities are simply used again. Trimers can be subsequently defined as sets of three fragments, as in 3-body FMO [113]. Here only 2-body FMO is considered. The Hamiltonians give rise to two sets of time-independent Schrödinger equations; one for the individual fragments, and one for the dimers,

$$H_I \Psi_I = E_I \Psi_I, \tag{4.3}$$

$$H_{IJ} \Psi_{IJ} = E_{IJ} \Psi_{IJ}. \tag{4.4}$$

Solving these with an appropriate level of theory such as Hartree-Fock(HF)/6-31G gives the energies $\{E_I, E_{IJ}\}$ and wave functions $\{\Psi_I, \Psi_{IJ}\}$. The total energy in 2-body FMO is,

$$E = \sum_{I>J} E_{IJ} - (N-2) \sum_I^N E_I + \sum_{s>t}^{N_{\rm at}} \frac{Z_s Z_t}{|\mathbf{R}_s - \mathbf{R}_t|}, \tag{4.5}$$

if all possible dimers are calculated. The last term is the nuclear repulsion energy. In practice, only fragments which are within an appropriate cut-off radius of each other are defined as dimers. This is called the electrostatic dimer approximation.

A dimer consisting of far separated fragments is very well approximated by the fragments alone, which is already accounted for in the fragment sum. More generally, the electronic energy is,

$$E' = \sum_{I>J}(E_{IJ} - E_I - E_J) + \sum_{I}^{N} E_I, \tag{4.6}$$

where the double sum is over the chosen set of dimers only. The total (2-body FMO) electron density is analogously,

$$\rho(\mathbf{r}) = \sum_{I>J}\Big[\rho_{IJ}(\mathbf{r}) - \rho_I(\mathbf{r}) - \rho_J(\mathbf{r})\Big] + \sum_{I}^{N}\rho_I(\mathbf{r}), \tag{4.7}$$

where the individual monomer and dimer densities are as in conventional *ab initio* electronic structure theory, and are previously derived in Section 2.5.

4.2.2 Independent and pairwise fragment models

A fragment-based improvement to the IAM scattering approximation is considered here, where the sum over the atomic form-factors is upgraded to a sum over fragment form-factors. This introduces an *ab initio* electronic structure theory aspect to the calculation of elastic X-ray molecular form-factors for large molecules, while still maintaining a modest computational resource cost. Unfortunately, it also introduces the breaking of covalent bonds, causing radicals, which is problematic, as the electronic structure of the molecule is no longer correct. However, there are ways to somewhat fix this issue. The FMO method involves self-consistent iteration of the fragment energies within the average field of the other fragments. In this way, the radical electrons are accounted for in the total energy. Additionally, dimers are defined as pairs of fragments which include the unsevered bond. Inclusion of dimer terms is another way of correcting the electronic structure. For the purposes of improving upon the IAM method, it turns out, a simple sum of radical fragment form-factors (as calculated with AIXRD) is a reasonable approach. The reason for this is that the IAM method is itself a sum of radical atomic form-factors, because not all atoms are fundamentally closed shell.

In any fragment-based method, a first approximation to the electron density is a

simple sum of fragment densities,

$$\rho(\mathbf{r}) = \sum_{I}^{N} \rho_I(\mathbf{r}), \tag{4.8}$$

which is the case in subsystem DFT [110]. The elastic X-ray molecular form-factor, is the Fourier transform (FT) of the electron density (as in Section 2.1), which is,

$$f(\mathbf{q}) = \sum_{I}^{N} \int \rho_I(\mathbf{r}) e^{i\mathbf{qr}} \mathrm{d}\mathbf{r} = \sum_{I}^{N} f_I(\mathbf{q}). \tag{4.9}$$

This is analogous to the IAM scattering approximation, which is $f(\mathbf{q}) = \sum_{j}^{N_{\mathrm{at}}} f_j(\mathbf{q})$, with $\{f_j(\mathbf{q})\}$ a set of atomic form-factors, except here the sum is over fragments consisting of >1 atom and $\{f_I(\mathbf{q})\}$ is a set of molecular (radical) form-factors. This is reasonable in a scattering context as an improvement over the IAM approximation. From an *ab initio* perspective, the wave function, and therefore the electron density, can be improved upon by explicitly taking into account all interactions, e.g. by using the FMO Hamiltonians in Eqs. (4.1) and (4.2). However, the approximation used in this Chapter uses isolated fragments and fragment-pairs (dimers). This implies effective Hamiltonians which omit fragment-fragment interactions,

$$H'_I = \sum_{i}^{n_I} \Big\{ -\frac{1}{2}\nabla_i^2 - \sum_{s}^{N_I} \frac{Z_s}{|\mathbf{r}_i - \mathbf{R}_s|} \Big\} + \sum_{i>j}^{n_I} \frac{1}{|\mathbf{r}_i - \mathbf{r}_j|}, \tag{4.10}$$

$$H'_{IJ} = \sum_{i}^{n_I+n_J} \Big\{ -\frac{1}{2}\nabla_i^2 - \sum_{s}^{N_{IJ}} \frac{Z_s}{|\mathbf{r}_i - \mathbf{R}_s|} \Big\} + \sum_{i>j}^{n_I+n_J} \frac{1}{|\mathbf{r}_i - \mathbf{r}_j|}, \tag{4.11}$$

where N_I and N_{IJ} are the number of atoms in fragment I and dimer IJ respectively. Solving the resulting Schrödinger equations for Eqs. (4.10) and (4.11) gives energies E'_I and E'_{IJ}; and wave functions ψ'_I and ψ'_{IJ}. The corresponding Kohn-Sham or Hartree-Fock electron densities, from Eq. (3.8) are,

$$\rho'_I(\mathbf{r}) = \sum_{i}^{n_I} b_i |\phi'_i(\mathbf{r})|^2 \tag{4.12}$$

$$\rho'_{IJ}(\mathbf{r}) = \sum_{j}^{n_{IJ}} b_j |\phi'_j(\mathbf{r})|^2, \tag{4.13}$$

where n_I and n_{IJ} are the number of occupied orbitals in the fragment and dimers respectively, $b_i, b_j \in \{0, 1, 2\}$ are their occupancies, and $\{\phi'_i\}$ and $\{\phi'_j\}$ are the molecular orbitals which construct ψ'_I and ψ'_{IJ}. Similar to Eq. (4.7), a pairwise sum can be constructed,

$$\rho'(\mathbf{r}) = \sum_{I>J} \Big[\rho'_{IJ}(\mathbf{r}) - \rho'_I(\mathbf{r}) - \rho'_J(\mathbf{r})\Big] + \sum_{I}^{N} \rho'_I(\mathbf{r}), \tag{4.14}$$

the FT of which is,

$$f'(\mathbf{q}) = \sum_{I>J} (f'_{IJ} - f'_I - f'_J) + \sum_{I}^{N} f'_I, \tag{4.15}$$

a molecular form-factor based on a pairwise sum of isolated fragment and dimer densities, and takes into account fragment-fragment interactions, albeit in an approximate way.

4.3 Results and discussion

The molecular form-factor is generally a 3-dimensional function, i.e. $f = f(q, \theta, \phi)$, where $q = |\mathbf{q}|$, and θ and ϕ are the polar and azimuthal scattering angles. The molecular form-factor squared is the diffraction intensity, $I(\mathbf{q}) = |f(\mathbf{q})|^2$. The rotationally-averaged diffraction intensity is the molecular form-factor squared taking into account all possible orientations of the molecule,

$$I(q) = \frac{1}{4\pi} \int_0^{2\pi} \int_0^{\pi} |f(q, \theta, \phi)|^2 \sin\theta \mathrm{d}\theta \mathrm{d}\phi, \tag{4.16}$$

which is observed in unaligned gas or liquid-phase X-ray diffraction experiments, where the X-rays scatter from the randomly-orientated ensemble-average of sample molecules.

Throughout this Chapter comparisons are made between theoretical methods for determining $I(\mathbf{q})$ or $I(q)$. The percent difference is used for this, which is defined as,

$$\%\Delta I = 100\Big(\frac{I_A - I_B}{I_B}\Big), \tag{4.17}$$

where diffraction intensity of method A is compared to reference intensity I_B.

4.3.1 Small molecules

Ethanol was chosen as a prototype small molecule to test the effect of various fragmentation methods on the resulting AIXRD results. It is used as an example case to determine the effect of very small fragments on the X-ray diffraction compared to full ($N = 1$ fragment) AIXRD. In later Sections where fragment-based AIXRD is used on peptides and a protein molecule, fragments will be at least as large as an entire ethanol molecule. Ethanol was geometry optimised at the Hartree-Fock(HF)/6-31G level, and a Mulliken population analysis was performed using Molpro [96]. A low quality electronic structure theory was used, and the effect of increasing the basis set size is compared to the choice of fragments at the end of this section (see Fig. 4.4 and corresponding text). The partial charges on the oxygen atom and the end hydrogen atom were $-0.75401e$ and $+0.40013e$, giving the OH group a net negative charge of $-0.35388e$. The middle group (CH_2) has a net charge of $-0.09988e$, and the end group (CH_3) has a net charge of $+0.45377e$. For this reason, ethanol was divided into two different sets of $N = 3$ fragments, CH_3-CH_2-OH, and CH_3^+-CH_2-OH^-, which are denoted IFM1 and IFM2 respectively. See Fig. 4.1 for a representation of each fragment definition. HF/6-31G was used on each fragment, all $N(N-1)/2 = 3$ unique dimers, each individual atom ($N = N_{\mathrm{at}}$); denoted the independent atom model (IAM), and the full ($N = 1$) molecule. The full *ab initio* calculation can be thought of as a trimer calculation using any definition of $N = 3$ fragments. It is therefore more accurate than any pairwise or simple sum approximation of $N = 3$ fragments.

The energies of each fragment were summed and pairwise summed and compared with the full HF/6-31G energy. See Table 4.1 for the energies of each method. Aside from the IAM energy, the IFM2 energy is the furthest from the full energy, followed by the IFM1 energy which is 0.42 Hartree closer and 0.27 Hartree larger than the $N = 1$ energy. The $\mathrm{IFM1_{pw}}$ and $\mathrm{IFM2_{pw}}$ energies are similar by 0.017 Hartree and $0.06 - 0.08$ larger than the $N = 1$ energy. The IAM energy is the furthest from the full HF/6-31G energy as it is constructed by completely isolated atomic energies. Comparing the simple $N = 3$ fragment sum energies, IFM1 is a much better choice than IFM2. For the pairwise sum energies, $\mathrm{IFM2_{pw}}$ is only a slightly better choice than $\mathrm{IFM1_{pw}}$.

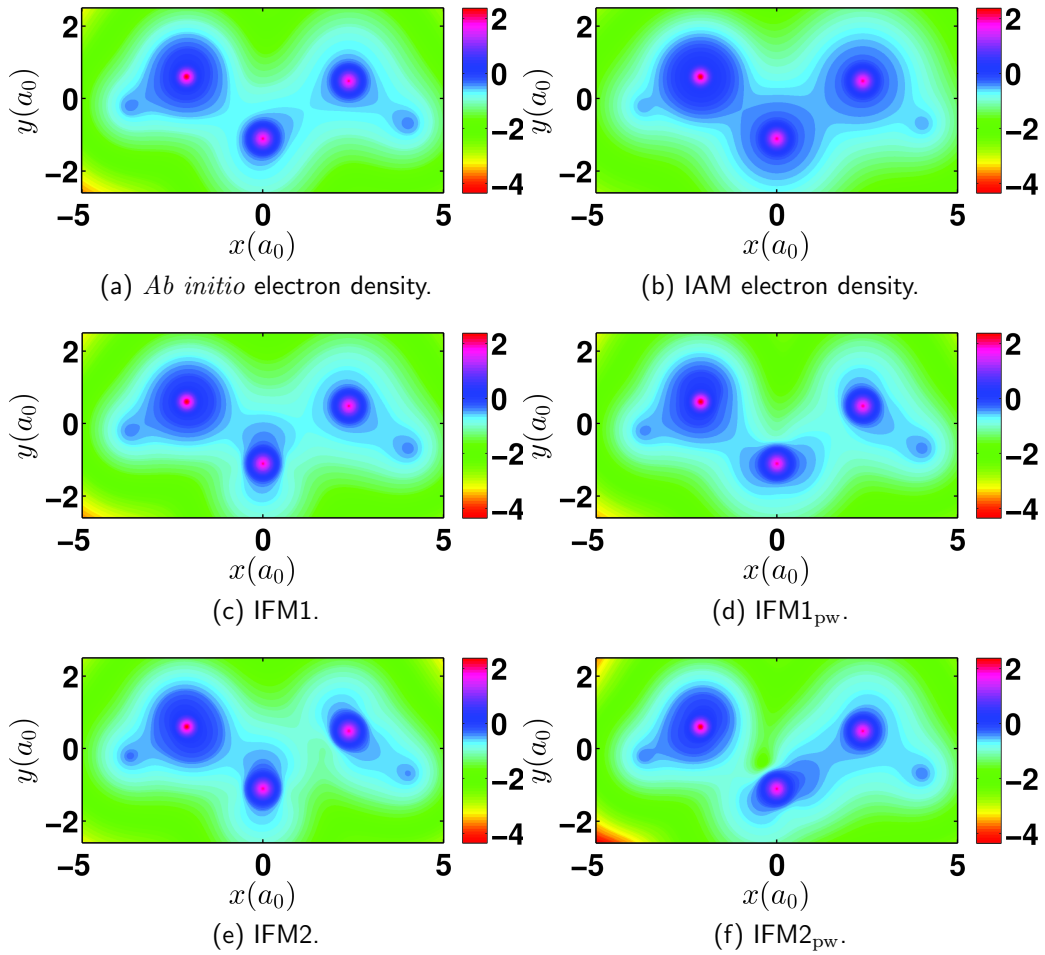
The electron densities, $\rho_A(\mathbf{r})$, for each method A were calculated with Eqs. (4.8) and (4.14) on a $(12\ a_0)^3$ grid with spacing $0.1\ a_0$. Slices through the xy-plane

Method	E (Hartree)	ΔE (Hartree)
Full	-154.01322721	0
IFM1	-153.74573447	0.26749274
IFM2	-153.32481489	0.68841232
IFM1_{pw}	-153.93283357	0.08039364
IFM2_{pw}	-153.94940080	0.06382641
IAM	-152.82161838	1.19160883

Table 4.1: Energy extensivity for ethanol divided into $N = 3$ fragments, CH_3-CH_2-OH, denoted IFM1; and CH_3^+-CH_2-OH^-, denoted IFM2. The simple and pairwise sums (pw) of fragment energies and full HF/6-31G energy are shown. HF/6-31G was used for each fragment and unique dimer. The independent atom model (IAM) energy is shown, which is a sum of isolated atomic energies. It is equivalent to dividing the molecule into $N = N_A$ fragments.

of the electron densities, $\rho(x, y, z = 0)$, as well as the IAM electron density, are shown in Fig. 4.2 where the atomic centre of the end hydrogens, both carbons, and the oxygen lie on the xy-plane. The IAM density was calculated by summing the individual HF/6-31G atomic electron densities. By qualitatively examining the electron density slices for each method shown in Fig. 4.2, it appears that the IFM1 slice better represents the oxygen and end-carbon distributions than IFM2. The IFM2_{pw} slice has a slightly closer end carbon distribution to the full *ab initio* electron density compared to IFM1_{pw}, and a potentially better representation of the oxygen distribution. An absolute difference $|\Delta\rho| = |\rho_A - \rho_{\text{full}}|$ analysis between each method A and the full *ab initio* calculation is shown in Table 4.2. It shows that the IFM1 sum of fragment densities is actually better that the pairwise summations in terms of mean and median $|\Delta\rho|$, and IFM2 is the worst choice apart from IAM. As expected, the IAM method has the largest maximum, mean and median $|\Delta\rho|$. The IFM1 method has a significantly lower mean and median $|\Delta\rho|$ than IFM2, i.e. the IFM1 choice of fragments gives a result closer to the full *ab initio* method, showing strong dependence on the choice of fragments even if they only differ by the location of one electron. The IFM1_{pw} and IFM2_{pw} have slightly lower maximum values of $|\Delta\rho|$ than their non-pairwise counterparts.

Ab initio X-ray diffraction (AIXRD) was performed on each of the IFM1 and IFM2 fragments and dimers to give fragment-based AIXRD results via Eqs. (4.9) and (4.15), which correspond to the fragment-based electron densities. Fig. 4.3 shows orientationally-averaged X-ray diffraction intensities, Eq. (4.16), in



(a) *Ab initio* electron density. (b) IAM electron density.

(c) IFM1. (d) IFM1_{pw}.

(e) IFM2. (f) IFM2_{pw}.

Figure 4.2: Ethanol electron density slices $\rho(x, y, z = 0)$, with $\rho(\mathbf{r})$ calculated on a $(12\ a_0)^3$ cubic grid with $0.1\ a_0$ spacing. Calculated via: (a) full ($N = 1$) HF/6-31G; (b) independent atom model (IAM), a sum of atomic densities each calculated with HF/6-31G; (c) fragment-based *ab initio*, a sum of $N = 3$ fragment-densities defined as CH_3-CH_2-OH, denoted IFM1, each calculated with HF/6-31G; (d) the pairwise sum of IFM1 fragments, denoted IFM_{pw}, which includes a sum of all unique dimer-densities; (e,f) same as (c,d) but with fragments defined as CH_3^+-CH_2-OH^-, denoted IFM2 and IFM2_{pw} respectively. The positions of the atoms were defined by a HF/6-31G geometry optimisation. The colour-scale is a base 10 log-scale e.g. the green region has a value of $\rho(x, y, z = 0) = 10^{-2}$.

the range $q = \{0, 4.2\}\ a_0^{-1}$, for each method; and the percent difference curves $\%\Delta I(q)$, Eq. (4.17), comparing each method to full AIXRD reference method. Table 4.3 shows the maximum, mean, and median of $|\%\Delta I(q)|$ for each method. The IAM method is the furthest from full AIXRD with the largest maximum, mean, and median $|\%\Delta I(q)|$. Each other method is comparatively close to the full *ab initio* method. That is, approximating the *ab initio* molecular form factor with a fragment-based method is worthwhile and a significant improvement over the IAM. IFM1_{pw} has the smallest maximum, mean, and median $|\%\Delta I(q)|$,

Method	Max. $\|\Delta\rho\|$	Mean $\|\Delta\rho\|$	Median $\|\Delta\rho\|$
IAM	7.40	2.8×10^{-3}	2.3×10^{-4}
IFM1	0.82	1.1×10^{-3}	0.1×10^{-4}
IFM2	0.82	1.7×10^{-3}	0.5×10^{-4}
IFM1_{pw}	0.62	1.3×10^{-3}	0.4×10^{-4}
IFM2_{pw}	0.37	1.4×10^{-3}	0.4×10^{-4}

Table 4.2: Electron density difference $|\Delta\rho| = |\rho_A - \rho_{\text{full}}|$ analysis, where method ρ_A is either the sum of 3 fragment densities, $\rho_{\text{IFM1}} = \rho(\text{CH}_3) + \rho(\text{CH}_2) + \rho(\text{OH})$, denoted IFM1; $\rho_{\text{IFM2}} = \rho(\text{CH}_3^+) + \rho(\text{CH}_2) + \rho(\text{OH}^-)$, denoted IFM2; the pairwise density sums for IFM1 and IFM2, denoted IFM1_{pw} and IFM2_{pw} respectively; or the independent atom model (IAM) density. The full electron density, ρ_{full}; each fragment; and dimer were calculated using HF/6-31G at the HF/6-31G optimised geometry for the full molecule. The IAM density was calculated by summing the individual densities of the atoms in the molecule. The positions of the atoms were defined by a HF/6-31G geometry optimisation. Each electron density was computed on a $(12\ a_0)^3$ cubic grid with $0.1\ a_0$ spacing.

Method	Mean $\|\%\Delta I(q)\|$	Median $\|\%\Delta I(q)\|$	Max. $\|\%\Delta I(q)\|$
IAM	3.98	2.90	9.25
IFM1	1.68	1.57	4.28
IFM2	1.54	1.00	4.10
IFM1_{pw}	0.70	0.60	1.72
IFM2_{pw}	1.16	0.96	2.92

Table 4.3: Mean, median and maximum values of $|\%\Delta I(q)|$ for ethanol using the independent atom model and fragment-based approximations compared to the full AIXRD method.

followed by IFM2_{pw}, IFM2, and IFM1. This implies the best method after full AIXRD is IFM1_{pw}, and a pairwise method is more accurate than a non-pairwise one.

To examine the effect of larger basis set on the fragment-based method, full AIXRD and AIXRD with IFM1 fragments were performed with the 6-311++G** basis set. These curves were compared to full AIXRD and IFM1 AIXRD as before with the 6-31G basis set. Fig. 4.4 shows $\%\Delta I(q)$ using full AIXRD with 6-311++G** as the reference signal. The 6-31G full AIXRD curve is the closest to having a 0% difference with a mean $|\%\Delta I(q)|$ of 0.63%. All three curves have a similar trend at $q > 2.5$ Å^{-1}, corresponding to real space distances of < 2.51 Å.

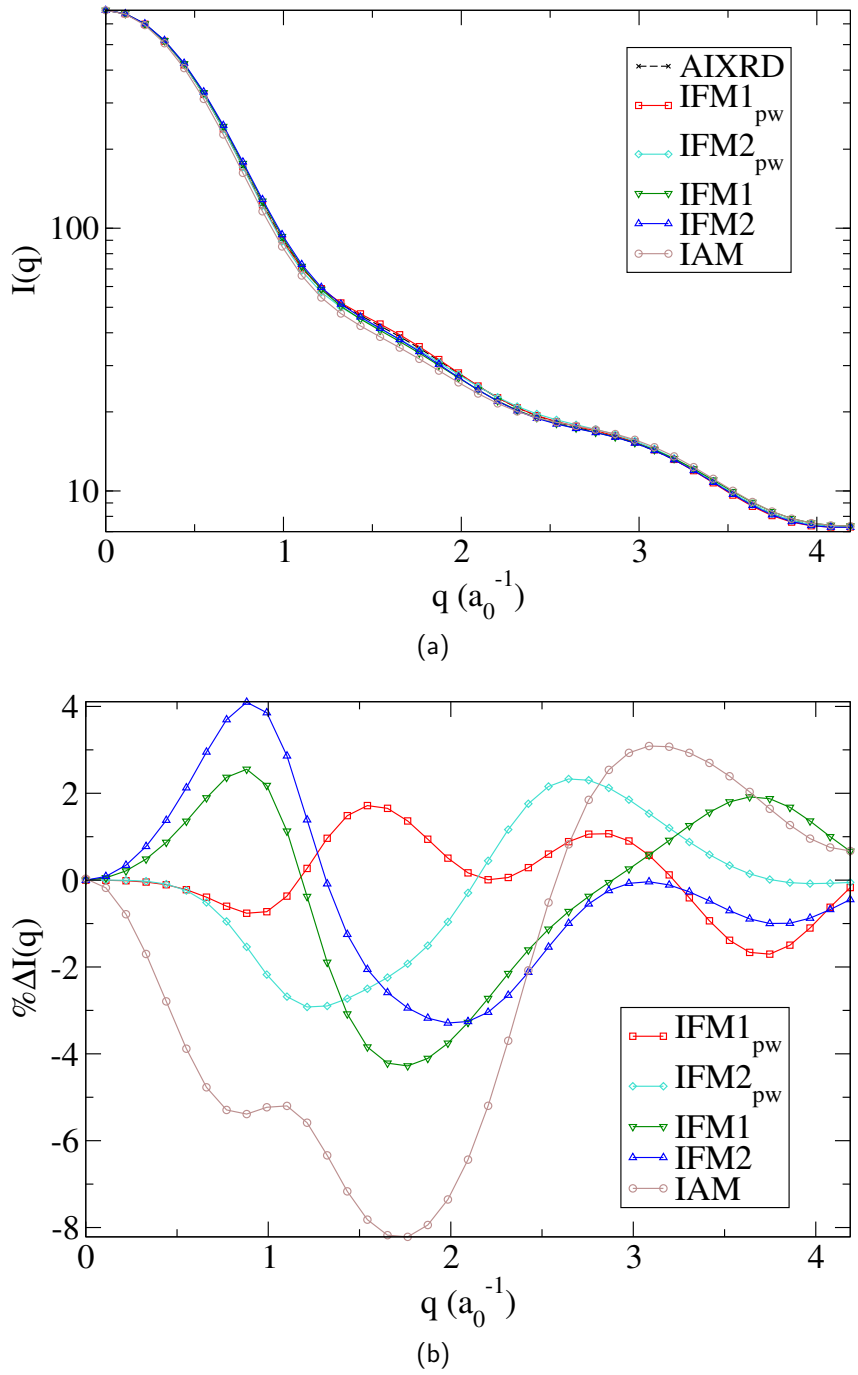


Figure 4.3: Ethanol orientationally-averaged elastic X-ray diffraction curve comparison for each method. Where the methods are, full *ab initio* X-ray diffraction (AIXRD), IFM1, IFM2, IFM1_{pw}, and IFM2_{pw}, and the independent atom model (IAM). (a) The log of the diffraction intensity $I_A(q) = |f_A(q)|^2$ for each method. (b) The percent difference, $\%\Delta I(q) = 100(I_A - I_{\text{full}})/I_{\text{full}}$, between each method and the full AIXRD reference curve.

Both the $N = 3$ fragment-based methods have similar curves with mean $|\%\Delta I(q)|$ values of 1.61% for 6-311++G** and 1.97% for 6-31G despite the difference in basis set size. This shows that increasing the basis set size cannot outperform or

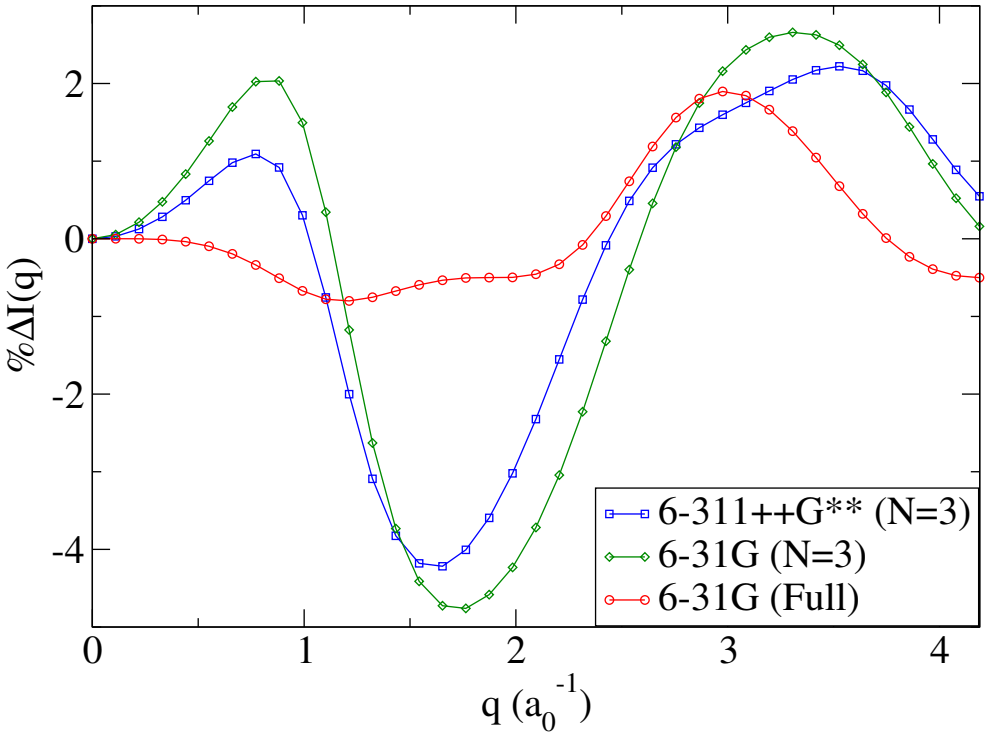


Figure 4.4: Comparison for ethanol between the effect of basis set and fragmentation on the percent difference $\%\Delta I(q)$, as in Eq. (4.17), with a reference signal of full ($N = 1$) AIXRD HF/6-311++G**. Increasing the basis set does not circumvent the effect of fragmentation. The full HF/6-31G calculation is more accurate than the $N = 3$ HF/6-311++G** calculation, which is not much of an improvement over the $N = 3$ HF/6-31G calculation

circumvent raising the fragment size. In this example going from $N = 3$ fragments each with size $\approx 1/3$ of the molecule to a molecule sized $N = 1$ fragment gave a much more accurate result than using $N = 3$ fragments and going from 6-31G to 6-311++G**.

4.3.2 Amino acids

Keeping in mind a later Section will involve calculating the elastic X-ray molecular form-factor for a protein, amino acids are examined, which are biologically important organic molecules containing amine (-NH_2) and carboxylic acid (-COOH) functional groups. Proteinogenic ("protein-building") amino acids can construct peptide chains and form the building blocks of proteins. Nine proteinogenic amino acids are defined as "essential" for human diet because the body cannot produce them from other compounds. In solution with moderate pH of about $2.2 - 9.4$ the α-carboxylic acid group is deprotonated becoming -COO^- and the α-amine

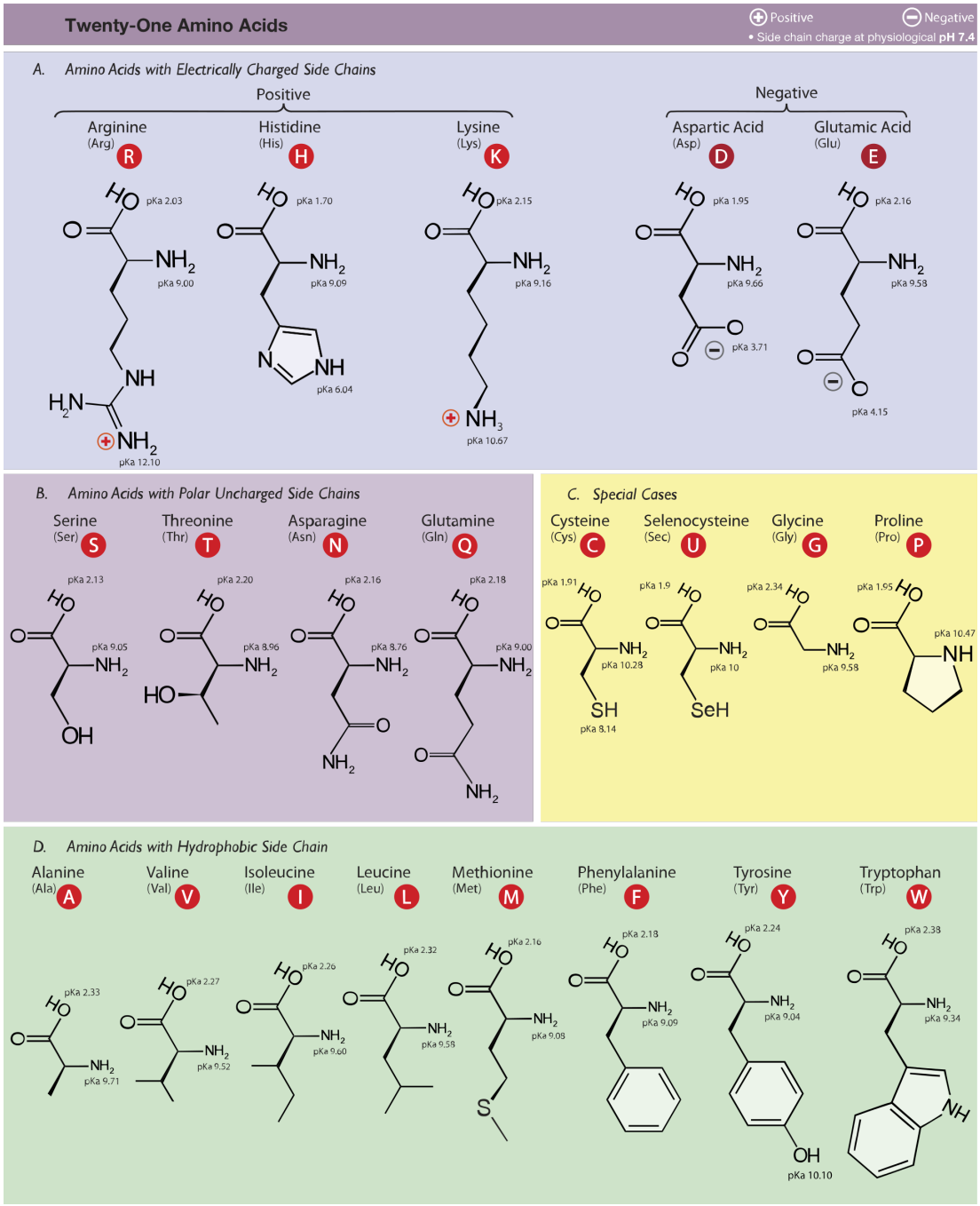
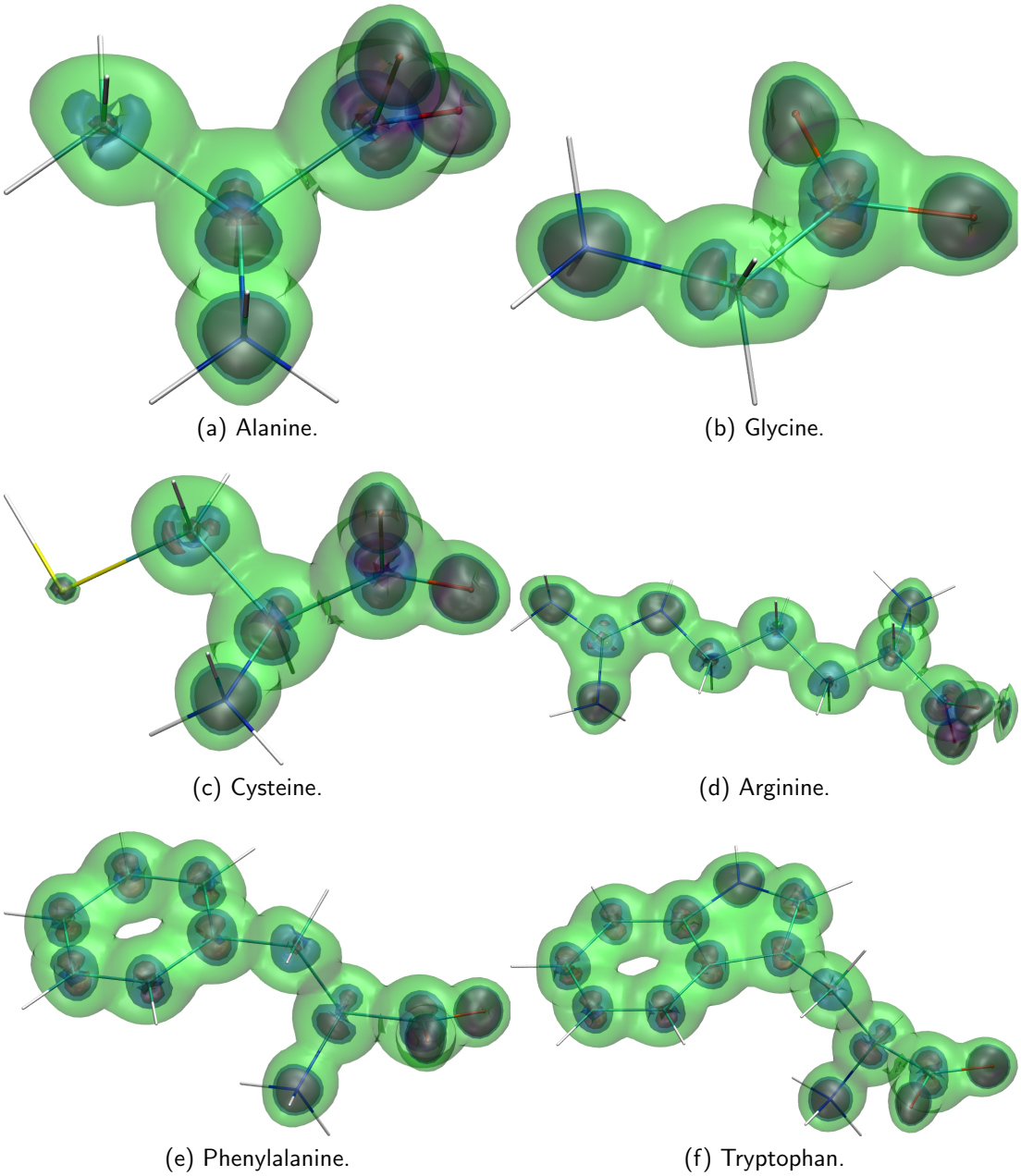


Figure 4.5: The 21 proteinogenic amino acids.

group is protonated becoming -NH_3^+. These are the zwitterionic forms of amino acids. See Fig. 4.5 for graphical representation of 21 proteinogenic amino acids.

A qualitative comparison between the *ab initio* and IAM electron densities for six amino acids is shown in Fig. 4.6, and a comparison between IAM and AIXRD intensities for 20 (selenocysteine not included) zwitterionic proteinogenic amino acids is shown in Fig. 4.7. The electron density difference $|\Delta\rho|$ is shown in Fig.



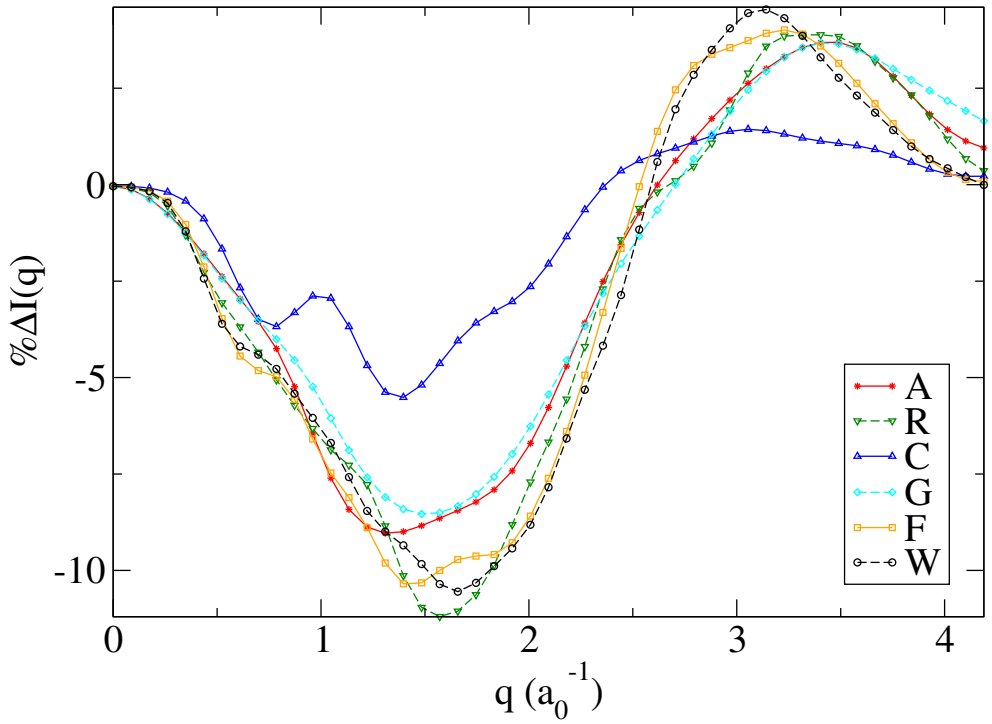
(a) Alanine. (b) Glycine.

(c) Cysteine. (d) Arginine.

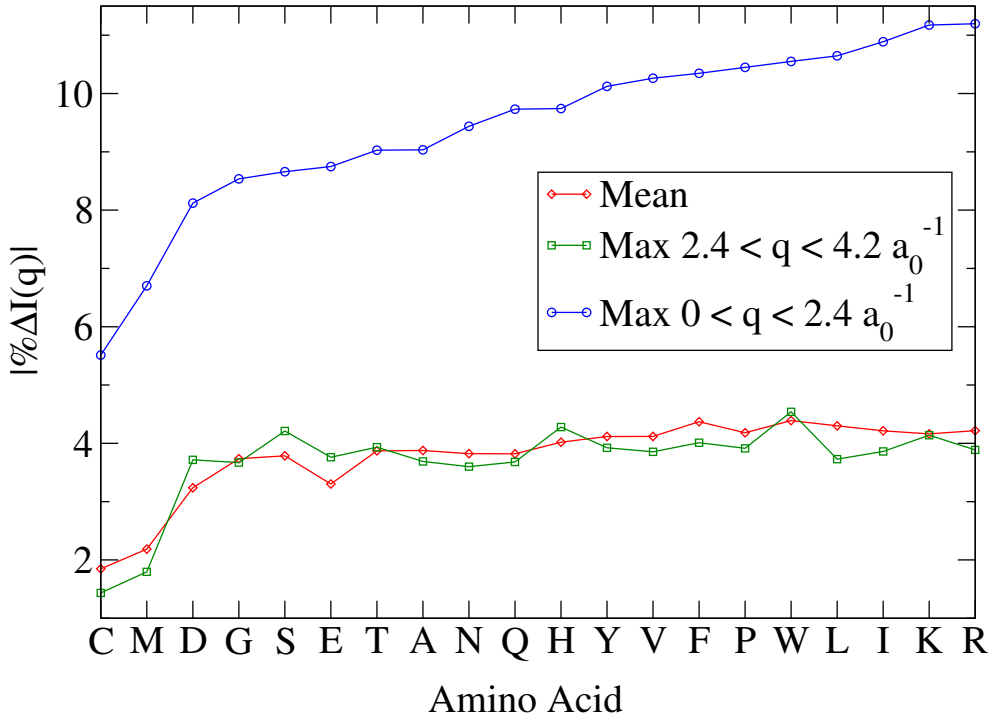
(e) Phenylalanine. (f) Tryptophan.

Figure 4.6: Isosurfaces of difference between IAM and *ab initio* electron densities for various amino acids. $\Delta\rho(\mathbf{r}) = |\rho_{IAM}(\mathbf{r}) - \rho_{AI}(\mathbf{r})| = \{0.23, 0.2, 0.1\}$ (arb. units) for red, blue, and green respectively. Images generated with VMD [118].

4.6 for three isosurfaces of $|\Delta\rho| = \{0.23, 0.20, 0.10\}$ (arb. units), these values are arbitrary and only serve to visualise the qualitative difference between IAM and *ab initio* in terms of electron densities. Non-spherically symmetric distributions are clearly visible about atomic centres, such as p-orbitals, especially around phenyl-rings e.g. Fig. 4.6e and 4.6f. There is less difference at the sulphur atom in the cysteine residue (Fig. 4.6c) because higher Z atoms have a lower valence/-total electron ratio, and they have predominantly spherically symmetric electron



(a) Percent difference $\%\Delta I(q)$ for 6 amino acids.



(b) Mean and maximum absolute percent difference $|\%\Delta I(q)|$ for 20 amino acids, labelled by their single letter codes.

Figure 4.7: Amino acids, percent difference between IAM and AIXRD.

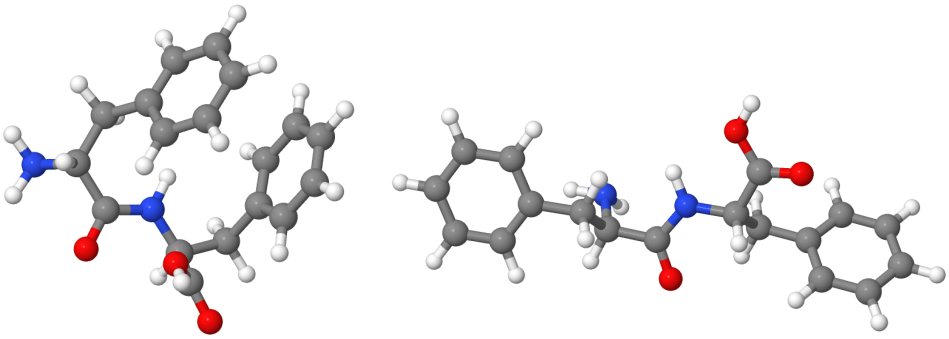
distributions. It is clear that there are important effects taken into account in the *ab initio* electron distribution compared to IAM.

The AIXRD results in Fig. 4.7 were calculated from HF/6-31G wave functions from PDB file geometries contained in the Facio molecular visualisation program [119]. The IAM results use tabulated atomic X-ray form-factors from The International Table of Crystallography [62]. AIXRD improves upon the IAM approximation for molecular form-factors primarily by taking into account valence electrons [1]. As seen in Chapter 3, the higher the valence/total electron ratio, the greater the difference between IAM and AIXRD. Residues containing groups with high valence/total electron ratio e.g. phenyl-groups, such as phenylalanine (F) and tryptophan (W) have some of the largest differences in percent difference terms, $\%\Delta I(q)$; (Fig. 4.7a), and in average absolute percent difference, $\int |\%\Delta I(q)| \mathrm{d}q$, (Fig. 4.7b). Sulphur-containing residues cysteine (C) and methionine (M) have the least difference between IAM and AIXRD, due to their relatively small size and the large Z sulphur atom, giving a low overall valence/total electron ratio.

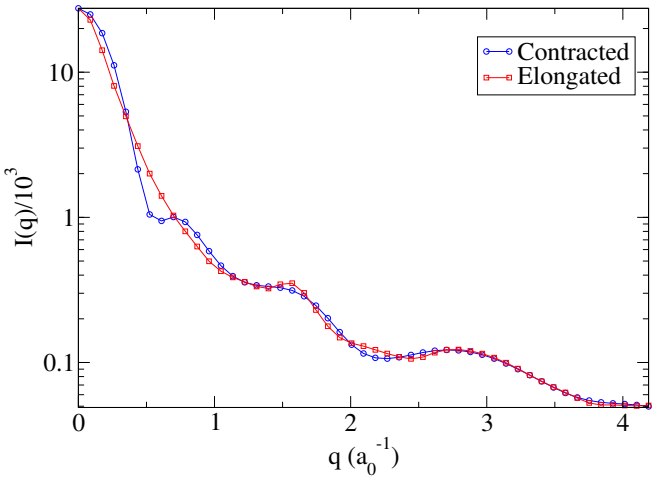
4.3.3 Peptides

As an example of a peptide, diphenylalanine (FF) is next examined. It is a common naturally-occurring dipeptide consisting of two phenylalanine (F) residues. Known from molecular dynamics simulations, it has three distinct conformers in water [120]: contracted (Fig. 4.8a); intermediate; and elongated (Fig. 4.8b), and spends the most time in its contracted conformer. Compared to other amino acid residues, its constituent, phenylalanine (F) has one of the largest average absolute percent differences, $\int |\%\Delta I(q)| \mathrm{d}q$, between IAM and AIXRD, predominantly because of the delocalised electrons within the phenyl-group. Moving beyond individual amino acids and to compare residue-size fragments and smaller fragment sizes, two HF/6-31G calculations were performed for FF in its contracted and elongated forms, from geometries obtained via steepest descent using the universal force-field (UFF) [121]. Molecular mechanics geometry optimisation is capable of achieving physically relevant geometries at fractions of the computational time of *ab initio* methods. Where accurate biomolecular geometries are a priority, the force-fields AMBER [122], CHARMM [123], and many others [124] are readily available. For an X-ray diffraction theory it makes sense to utilise this benefit, and becomes more necessary when moving towards protein structures consisting of > 100 residues. The HF/6-31G energies for each optimised conformer are $E_{\mathrm{contracted}} = -1026.34853945$ Hartree and $E_{\mathrm{elongated}} = -1026.35699991$ Hartree, with $\Delta E = |E_{\mathrm{contracted}} - E_{\mathrm{elongated}}| = 0.00846046$ Hartree. Rotationally-averaged diffraction intensities were determined from AIXRD molecular form-factors, see

Fig. 4.8c. There are distinct diffraction signals for each conformer.



(a) Contracted. (b) Elongated.



(c) Orientationally-averaged diffraction intensity $I(q)$.

Figure 4.8: Diphenylalanine (FF), representation of contracted and elongated conformers (a,b); and corresponding orientationally-averaged elastic X-ray diffraction intensity curves (c), calculated with AIXRD.

Then, FF was split into $N = 2$ fragments, each consisting of one F residue. The total energy was constructed by $E_{N=2} = E_{\mathrm{F1}} + E_{\mathrm{F2}}$. AIXRD form-factors were calculated for each fragment from two separate electronic structure wave functions, and then added to approximate the total molecular form-factor, $f_{\mathrm{FF}}(\mathbf{q}) = f_{\mathrm{F1}}(\mathbf{q}) + f_{\mathrm{F2}}(\mathbf{q})$. The same was done except using $N = 5$ fragments (Phenyl-$C_2H_3NH_2$-HNCO-C_2H_3COOH-Phenyl). All $N(N-1)/2 = 10$ dimer form-factors were calculated and a pairwise sum of $N = 5$ fragment form-factors was made to better approximate the molecular form-factor. The total energy E was constructed again by a sum of fragment energies, and a pairwise sum of fragment and dimer energies. These are shown and compared to full *ab initio* in Table 4.4. Each fragment-based method of calculating the total molecular form-factor are compared to the full ($N = 1$) AIXRD method, for both the elongated and contracted FF conformers, see Fig. 4.9 and corresponding Table 4.5. The percent difference

Method	$E_{contracted}$	ΔE	$E_{elongated}$ (Hartree)	ΔE
$N = 5$	-1025.61372264	0.7348	-1025.78545524	0.5715
$N = 5$ (pairwise)	-1025.98608101	0.3625	-1025.86825528	0.4887
$N = 2$	-1026.08604548	0.2625	-1026.03736460	0.3196
Full	-1026.34853945	0	-1026.35699991	0

Table 4.4: FF contracted and elongated conformer energies for each method. Hartree-Fock with 6-31G basis set was used for each fragment and for the full calculation. Note that an $N = 2$ pairwise calculation is equivalent to an $N = 1$ or full calculation.

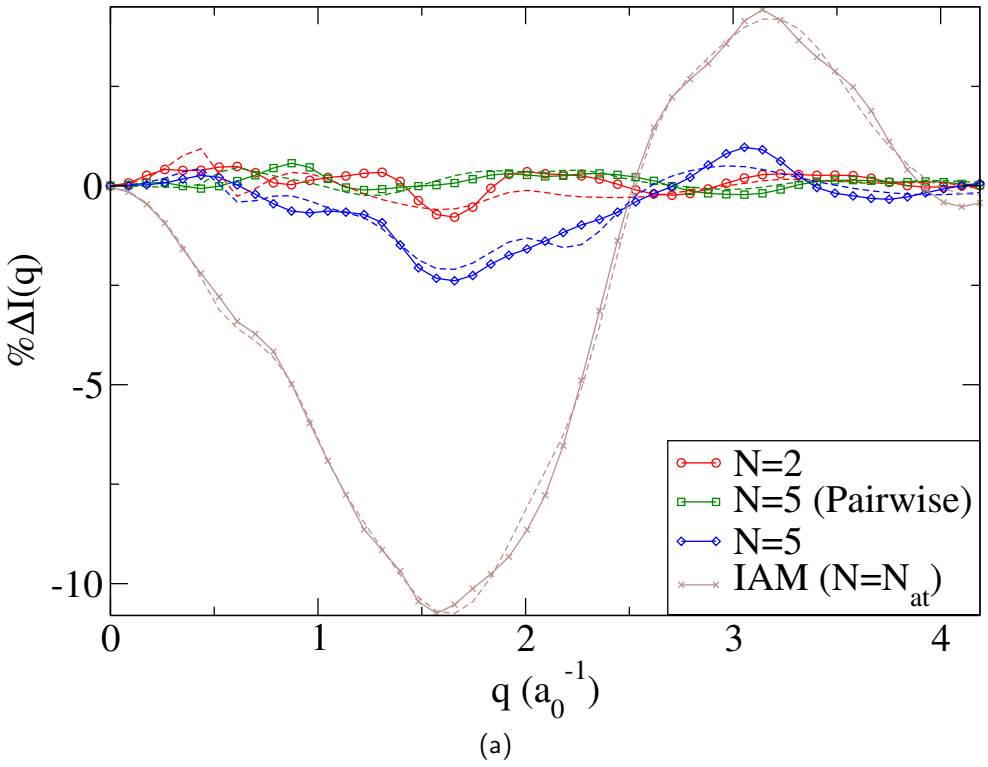


Figure 4.9: Diphenylalanine (FF) percent difference $\%\Delta I(q)$ between fragment-based methods and full AIXRD. HF/6-31G was used for each calculation. The IAM calculation is equivalent to a $N = N_{at}$ fragment-based calculation. Solid lines are for elongated FF and dashed are for contracted FF.

is greatest for IAM. The closest approximations to the full AIXRD calculation are the $N = 2$ sum and the $N = 5$ pairwise sum. The pairwise sum can be thought to mitigate the error caused by the choice of small fragments, which are an intuitively worse approximation than larger residue-sized fragments. There is a minor difference in $\%\Delta I(q)$ in all cases between contracted and elongated FF. This implies that fragment-fragment interactions have a low dependence on the molecular geometry and there is no need to alter the fragment definition due to

Conformer	Method	Mean	Median	Maximum
Elongated	$N = 2$	0.24	0.25	0.79
	$N = 5$ (pairwise)	0.15	0.12	0.57
	$N = 5$	0.68	0.45	2.39
	IAM	4.27	3.41	10.74
Closed	$N = 2$	0.23	0.17	0.93
	$N = 5$ (pairwise)	0.16	0.12	0.46
	$N = 5$	0.63	0.38	2.10
	IAM	4.24	3.58	10.75

Table 4.5: Diphenylalanine, analysis table for absolute percent difference $|\%\Delta I(q)|$ between fragment-based AIXRD and full AIXRD.

Method	N fragment-based AIXRD scaling
$N = 1$	$t_1 \approx NM^2 = M^2$
$N = 2$	$t_2 \approx N(\frac{M}{N})^2 = \frac{M^2}{N} \approx 0.5t_1$
$N = 5$	$t_5 \approx \frac{M^2}{N} \approx 0.2t_1$
$N = 5$ (pairwise)	$t_{5\text{pw}} \approx \frac{N(N-1)}{2}(\frac{2M}{N})^2 + N(\frac{M}{N})^2 = \frac{(2N-1)M^2}{N} \approx 1.8t_1$

Table 4.6: Fragment-based AIXRD scaling, where AIXRD scales with M^2, the total number of primitive Gaussians squared. For N fragments there is an average of M/N primitives per fragment, therefore a $t_N = N(M/N)^2$ scaling for a sum of fragment-based AIXRD form-factors. An average of $2M/N$ primitives per dimer and including all $N(N-1)/2$ unique dimers, gives a pairwise scaling of $t_{N\text{pw}} = \frac{N}{2}(N-1)(2M/N)^2 + N(M/N)^2 = (2N-1)M^2/N$, which is $\geq M^2\ \forall N$. In this case the $N = 2$ pairwise calculation is equivilent to the $N = 1$ calculation, therefore $t_{2\text{pw}} \equiv t_1$.

molecular geometry in this case. The scaling factors for computational calculation time for each method are shown in Table 4.6. For fragment-based AIXRD, simple sums of fragment form-factors are linearly faster to compute than the full AIXRD method. The more fragments there are the quicker the calculation takes at the cost of accuracy. The pairwise summation for any value of N is slower than full AIXRD. Therefore non-pairwise fragment-based AIXRD is preferable, it is just a matter of choosing appropriately large fragments with reasonable fragment-fragment interface. Note on defining fragments: Residues make good fragments due to reasonable 'break-region' near each peptide bond, such as a C-C single bond. The Facio program [119] has a graphical interface for defining fragments, and can automatically detect appropriate break-regions for peptides and proteins.

A bad fragmentation would be one which results in a large alteration to the local electronic structure, such as splitting a benzene ring in two. For these reasons, the $N = 2$ is the optimal choice for FF and individual residues appear to be a good choice for fragments [115].

4.3.4 Proteins

This final results Section is based on the calculation of the elastic X-ray molecular form-factor from a protein, specifically lysozyme, it compares the selected fragment-based AIXRD method to the conventional IAM method, with temporal-integration over a molecular dynamics trajectory.

Before he discovered penicillin, Alexander Fleming discovered lysozyme [125] by chance during a search for medical antibiotics. It is a small stable enzyme which is abundant in hen egg whites and part of our innate immune system, giving protection from bacterial infection. In 1965 it became the second protein structure and the first enzyme structure to be solved via X-ray crystallography [126]. It consists of 128 amino acid residues and contains all 20 common amino acids which are encoded by the universal genetic code.

There are many available protein data bank (PDB) files for lysozyme. Its structure has recently been determined (de novo) with diffraction imaging at a XFEL source [35] via femtosecond protein serial nanocrystallography. The resulting 4N5R PDB file has 992 atoms in the protein, with 480 side-chain atoms and 512 backbone atoms. After adding hydrogens with a tool in the computational biochemistry program Maestro [127] it has 1941 atoms in the protein, with 1172 side-chain atoms and 769 backbone atoms. There are $N_{\text{elec}} = 7559$ electrons, which is a very large amount in the context of electronic structure theory, and shows the strength of fragment-based methods which make such large calculations tractable. Fig. 4.10(a,c,e) shows representations of the 4N5R PDB file coordinates with added hydrogens from three different directions (along each axis). The 128 amino acid residues are shown in different colours by residue type. The 4N5R PDB geometry with added hydrogens was used, each residue was defined as a fragment, and AIXRD HF/6-31G was performed on each fragment, then added as $f(\mathbf{q}) = \sum_I^N f_I(\mathbf{q})$. As a side note, instead of using the PDB file geometry, molecular dynamics could be used to search for stable geometries in aqueous solution. This would be logical for unsolved structures, such as membrane pro-

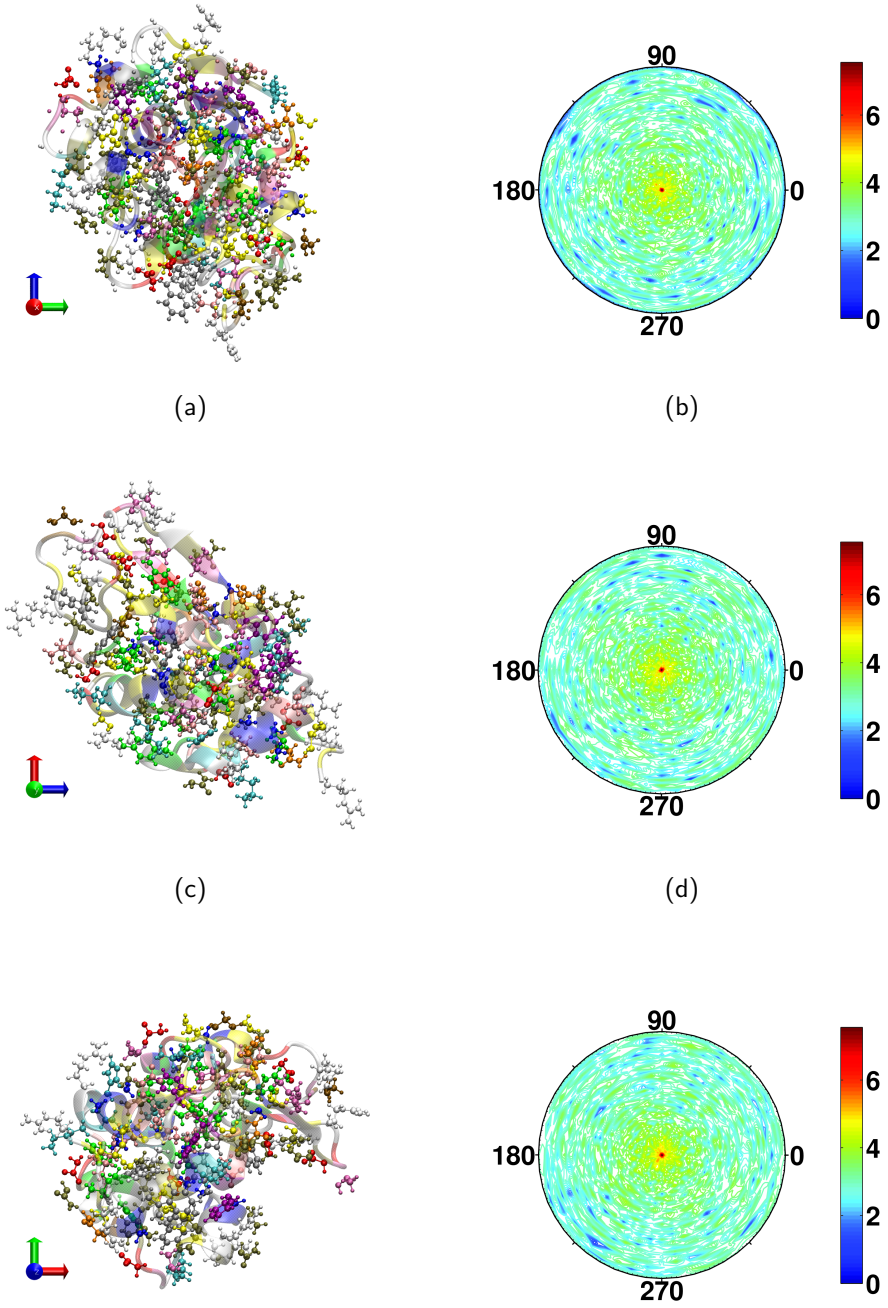


Figure 4.10: Lysozyme from three perspectives (a) x-perspective, (c) y-perspective, and (e) z-perspective with each amino acid type distinctly coloured, side-chains shown as ball and stick, and backbone as cartoon (made with VMD [118]), with corresponding diffraction patterns $I(\theta, \phi)$, $q = q(\theta, \phi)$ (b,d,f), calculated with residue-sized fragment-based AIXRD using HF/6-31G, for incident X-rays along each axis (perpendicular to plane of paper), with 0 uncertainty in alignment. A log-scale is used for the colour-axis.

teins, and would give a theoretical X-ray form-factor based on MD geometries and fragment-based AIXRD to aid structure determination experiments, and po-

tentially structural dynamics at a XFEL source for example.

The diffraction intensities, $|f(\mathbf{q})|^2$, are shown in Fig. 4.10(b,d,f) for the three different orientations (x, y, z) of the protein. The polar plot images shown are $|f(\mathbf{q}(\theta, \phi))|^2$ with $|f(\mathbf{q} = 0)|^2 \equiv N_{\text{elec}}^2$ at the centre, going out from here to the edge of the circle is the polar scattering angle $0 \leq \theta \leq \pi$, and going counter-clockwise around the circle is the azimuthal scattering angle $0 \leq \phi \leq 2\pi$ (shown in degrees). A log-scale is used as the central peaks are orders of magnitude greater in intensity than the larger θ scattering directions. There are distinct peaks in the diffraction intensity which depend parametrically on the protein orientation.

$\lvert\%\Delta I(\mathbf{q})\rvert$	$\langle x, y, z\rangle$	x	y	z
mean	4.23	4.46	4.13	4.10
median	2.22	2.28	2.34	2.05
fraction $> 5\%$	0.23	0.23	0.24	0.22
fraction $> 10\%$	0.09	0.09	0.09	0.10
fraction $> 25\%$	0.01	0.02	0.01	0.01

Table 4.7: Analysis table for absolute percent difference $|\%\Delta I(\mathbf{q})|$ between IAM and residue-sized fragment-based AIXRD for lysozyme, for each direction x, y, z of the incident X-rays.

The absolute percent difference between the IAM and fragment-based AIXRD diffraction intensities, $|\%\Delta I(\mathbf{q})|$ is shown in Fig. 4.11(a,c,e). The complementary Table 4.7 shows the mean, median, and fractions $> 5, 10, 25\%$. Considerable mean $|\%\Delta I(\mathbf{q})|$ values of 4.10-4.46 and median values of 2.05-3.34 for each orientation of the protein reveals the importance of taking into account valence electrons with an *ab initio* based method. In addition, over 1/5 of $|\%\Delta I(\mathbf{q})|$ values are $> 5\%$, close to 1/10 are $> 10\%$, and there are particular peaks which have $> 80\%$ difference for each orientation of the protein. This is because individual amino acids give rise to maximum percent differences between IAM and AIXRD of $5.6 - 11.2\%$ (Section 4.3.2), and a lot of them scattering together will amplify this effect at certain angles. This shows that valence electrons give rise to large differences at particular scattering angles. Fig. 4.11(b,d,f) shows normalised counts divided amongst 200 bins in the range $|\%\Delta I(\mathbf{q})| \in [0, 25]\%$ for each orientation of the protein. The counts for the ith bin, C_i, are defined as,

$$C_i = \text{n}(x \in [x_i, x_{i+1})), \tag{4.18}$$

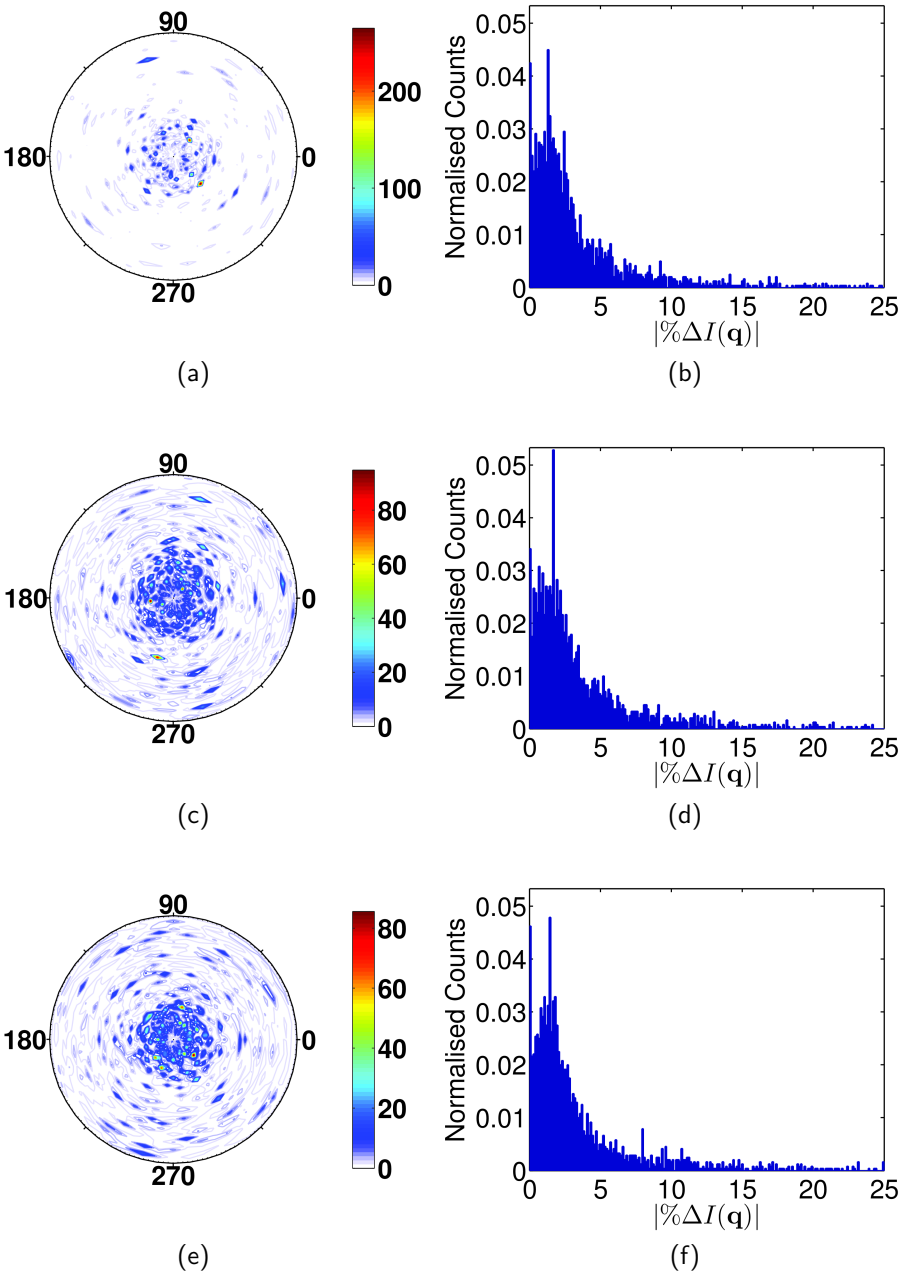


Figure 4.11: (a,c,e) Absolute percent difference $|\%\Delta I(\mathbf{q}(\theta,\phi))|$, and (b,d,f) absolute percent difference normalised counts between 0 and 25% counted into 200 bins, between IAM and residue-sized fragment-based AIXRD for the 128 residue protein; lysozyme, for incident X-rays in the (a,b) x-direction, (c,d) y-direction, and (e,f) z-direction.

where $\mathrm{n}(x)$ denotes the number of elements in set x, and $x \in [x_i, x_{i+1})$ denotes the elements of x in the interval from x_i to x_{i+1} inclusive of x_i but exclusive of x_{i+1}. Here, $x = |\%\Delta I(\mathbf{q})|$. The counts are normalised by dividing by the sum of

all N_{bins} counts,

$$C_i' = \frac{C_i}{\sum_i^{N_{\text{bins}}} C_i}. \tag{4.19}$$

This makes the sum of normalised counts equal unity, i.e. $\sum_i^{N_{\text{bins}}} C_i' = 1$. The normalised counts show the relative amount of errors for each value of $|\%\Delta I(\mathbf{q})|$. There are clear peaks in the range 0-5% with counts decreasing at larger values of $|\%\Delta I(\mathbf{q})|$, for each orientation of the protein. The distributions do not appear to be significantly influenced by molecular orientation.

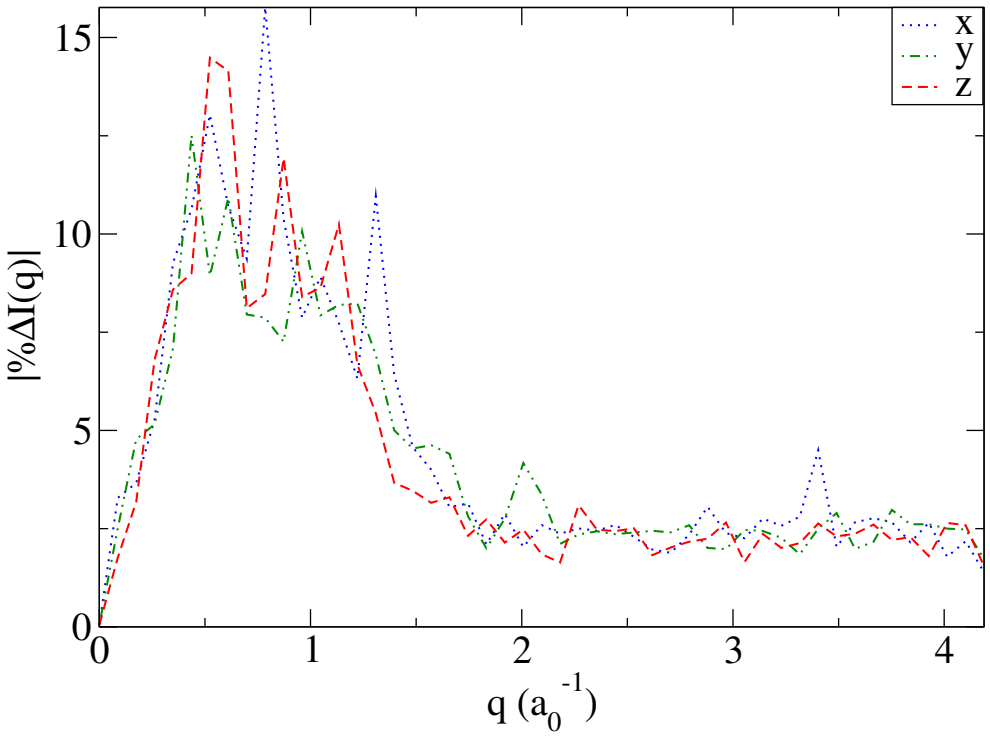


Figure 4.12: Percent difference between fragment-based AIXRD and IAM elastic X-ray intensities, integrated over the azimuthal scattering angle ϕ, $|\%\Delta I(q(\theta))| = \frac{1}{2\pi}\int |\%\Delta I(\mathbf{q}(\theta,\phi))|\mathrm{d}\phi$, for X-rays incident along the x, y, and z axes.

Fig. 4.12 shows the results from Fig. 4.11(a,c,e) integrated over the azimuthal scattering angle ϕ,

$$|\%\Delta I(q)| = \frac{1}{2\pi}\int_0^{2\pi} |\%\Delta I(\mathbf{q}(\theta,\phi))|\mathrm{d}\phi, \tag{4.20}$$

where $q = |\mathbf{q}| = 4\pi\sin(\theta)/\lambda$. This reveals the location of errors as a function of q only. The curves are similar in shape for each direction (x, y, z) of incident X-ray, again showing that the error is not strongly dependent on molecular orientation, i.e. there is not a particular X-ray direction which is solely responsible for large

outlier errors. The main peak spans $q \approx [0.2, 1.5]\ a_0^{-1}$, which corresponds to real space distances of 2.2-16.6 Å. At larger values of q ($>1.2\ a_0^{-1}$) the error is non-zero ($\approx 2.5\%$), thus, the entire electron distribution is affected by the inaccuracy of the IAM approximation.

4.3.5 Temporal integration

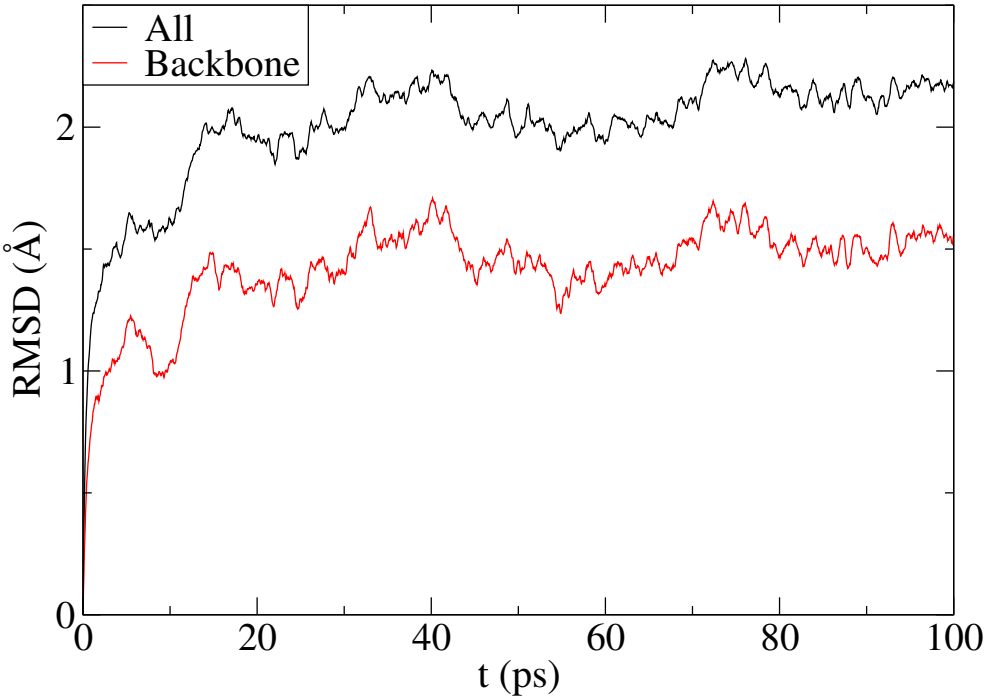


Figure 4.13: Lysozyme RMSD (all and backbone) compared to starting geometry for a 100 ps NVT 298 K molecular dynamics (MD) trajecotry in vacuum. Starting geometry was obtained via steepest descent energy minimisation of the 4N5R PDB [35] with added hydrogens geometry. MD trajectory calculated with Gromacs [128] and the AMBER99SB-ILDN force-field [129].

A 100 ps molecular dynamics (MD) simulation was performed on a single lysozyme molecule in vacuum. The starting geometry was obtained via steepest descent energy minimisation of the 4N5R PDB [35] geometry with added hydrogens. The MD was performed using an NVT ensemble at 298 K with a 0.2 fs timestep using Gromacs [128] and the AMBER99SB-ILDN force-field [129]. Fig. 4.13 shows the root-mean-squared deviation (RMSD) from the starting geometry for the whole protein and the backbone only. This simulation was done to model the effect of different molecular geometries during an X-ray diffraction experiment. The recent experiment [35] used a jet of lysozyme micro-crystals with passed through ultrabright femtosecond-order X-ray pulses from an XFEL source. In this study,

the difference between the conventional IAM method and the fragment-based AIXRD method is quantified, including temporal effects. These temporal effects manifest themselves in the data due to averaging over many X-ray diffraction snapshots to obtain statistically meaningful data. Thus, although the XFEL pulses are of femtosecond-order, it is appropriate to allow for picosecond order molecular motion. Keeping this in mind, starting at $t = 20$ ps, X-ray diffraction calculations were performed every 0.4 ps using IAM and fragment-based AIXRD (as before with each residue defined as a fragment) for incident X-rays along the z axis. The temporal-averaged absolute error is defined as,

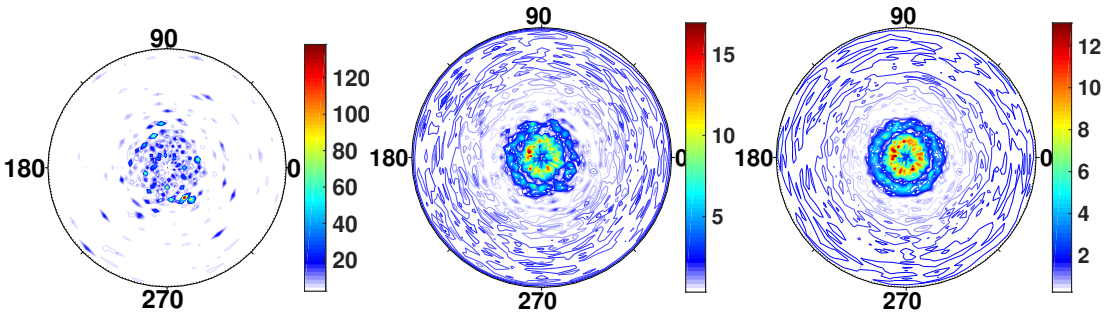
$$|\%\Delta I(\mathbf{q}, \Delta t)| = 100 \left| \frac{\langle I_{\mathrm{AI}} \rangle_{\Delta t} - \langle I_{\mathrm{IAM}} \rangle_{\Delta t}}{\langle I_{\mathrm{IAM}} \rangle_{\Delta t}} \right|, \tag{4.21}$$

where,

$$\langle I_{\mathrm{AI}} \rangle_{\Delta t} = \int_0^{\Delta t} I_{\mathrm{AI}}(\mathbf{q}, t) \mathrm{d}t, \tag{4.22}$$

and,

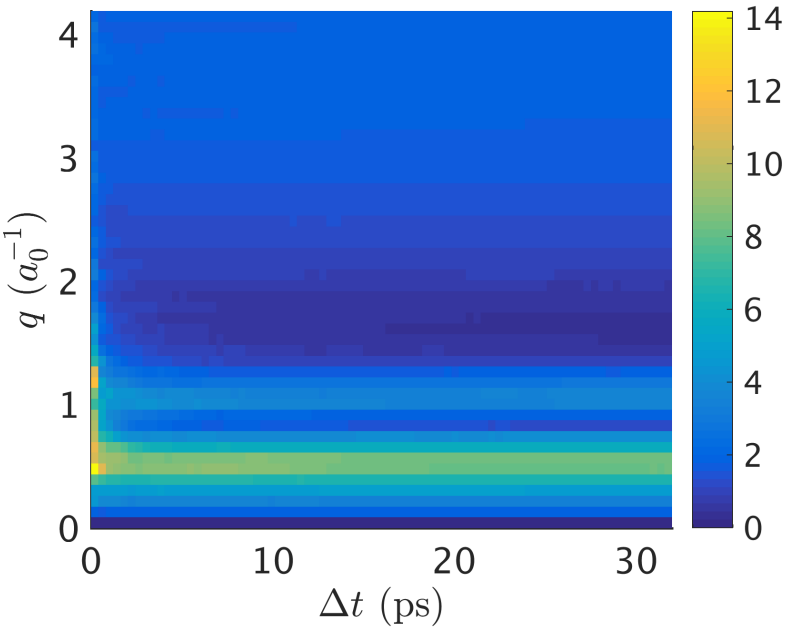
$$\langle I_{\mathrm{IAM}} \rangle_{\Delta t} = \int_0^{\Delta t} I_{\mathrm{IAM}}(\mathbf{q}, t) \mathrm{d}t, \tag{4.23}$$



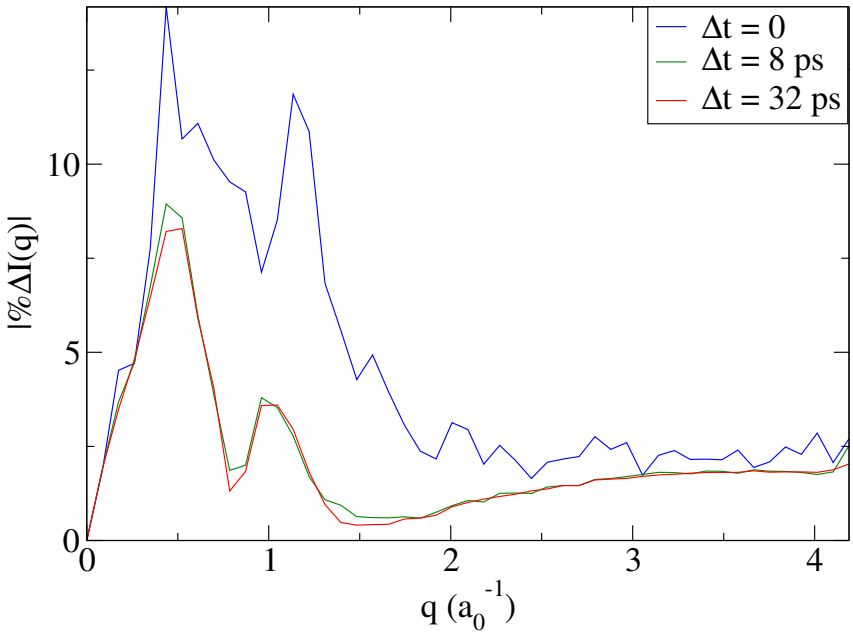
(a) $\Delta t = 0$. (b) $\Delta t = 8$ ps. (c) $\Delta t = 32$ ps.

Figure 4.14: Percent error, Eq. (4.21), detector images between time-integrated fragment-based AIXRD, Eq. (4.22), and time-integrated IAM, Eq. (4.23), for ($\lambda = 3\ a_0$) X-rays incident along the z-axis; and for time-intervals $\Delta t = 0, 8, 32$ ps, corresponding to $1, 20$, and 80 snapshots, each 0.4 ps apart.

and is shown in Fig. 4.14 at the detector for $\Delta t = 0, 8, 32$ ps, (corresponding to 1, 20, and 80 snapshots,) for X-rays incident along the z-axis. For $\Delta t = 0$ (single snapshot), a large error of $> 120\%$ appears at a particular scattering direction, with relatively large ($> 20\%$) errors being more common. For larger Δt such as $\Delta t = 8, 32$ ps (20 and 80 snapshots respectively), such large errors are averaged



(a) Percent error, $|\%\Delta I(q, \Delta t)|$.



(b) Corresponding slices through $|\%\Delta I(q, \Delta t)|$ at time-intervals $\Delta t = 0, 8, 32$ ps.

Figure 4.15: Lysozyme X-ray diffraction percent error $|\%\Delta I(q, \Delta t)|$, Eq. (4.21), between the fragment-based AIXRD and IAM methods, for absolute value of $q = |\mathbf{q}| = 4\pi \sin(\theta)/\lambda$ (azimuthal scattering angle, ϕ, integrated out), with incident X-rays along the z-axis, after temporal-averaging over time-interval Δt during a MD trajectory.

out by the motion of the molecule, e.g. for $\Delta t = 8$ ps the maximum error is 16%. Distinct rings of error points appear at particular radial angles, θ, which correspond to $q \approx 0.5\ a_0^{-1}$ and $q \approx 1.0\ a_0^{-1}$, or real space distances of 3.3 Å

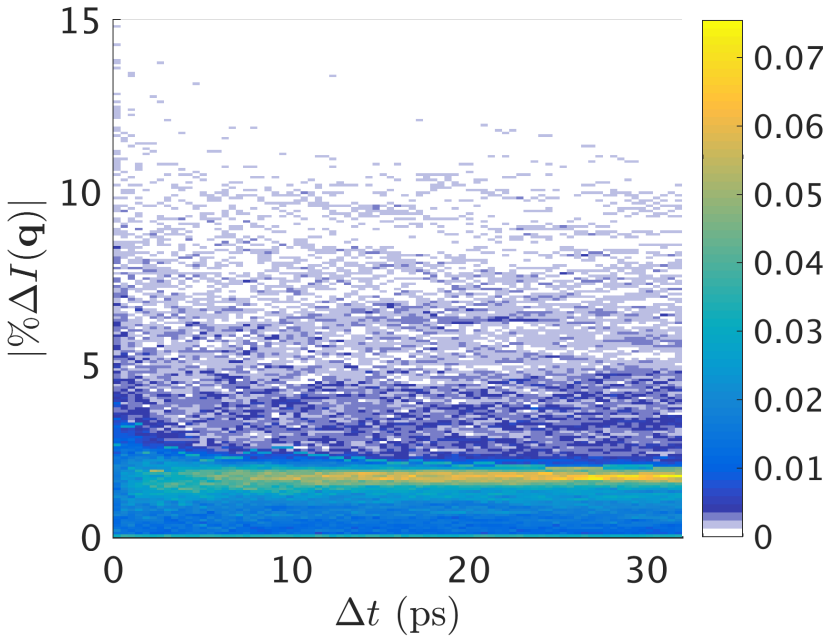
and 6.6 Å. In addition, the error dependence on the azimuthal scattering angle ϕ decreases as Δt increases. This makes sense as over time the molecule finds more of its conformational space. The full set of data is shown in Fig. 4.15, except the data for each Δt has been integrated over ϕ, as in Eq. (4.20). It is notable that the $\Delta t = 0$ (1 snapshot) result in Fig. 4.15b agrees well with the static results in Fig. 4.12. The $\Delta t = 8$ ps matches the $\Delta t = 32$ ps curve, showing that the integrals in Eq. (4.21) converge with increasing Δt. The distinct peaks at $q \approx 0.5\ a_0^{-1}$ and $q \approx 1.0\ a_0^{-1}$ are more visible in these curves, and have maxima of 8.1% and 3.7% error respectively. There is also a low error (<1%) region in the range $q = [1.4, 1.9]\ a_0^{-1}$, corresponding to real space distances of 3.3-4.5 Å, suggesting that the IAM is a good approximation only in this region.

The data is also shown in terms of normalised counts, Eq. (4.19), of $|\%\Delta I(\mathbf{q}, \Delta t)|$ at each value of Δt in Fig. 4.16. Similar to the static results in Fig. 4.11, counts C_i' were assigned to 200 bins in the range $|\%\Delta I(\mathbf{q}, \Delta t)| = [0, 15]\%$, where the total counts equal 1 (i.e. are normalised) for each value of Δt. A similar temporal convergence is seen as $C_i'(\Delta t)$ at Δt =8 ps agrees well with the Δt =32 ps slice. There is a distinct peak at $|\%\Delta I(\mathbf{q}, \Delta t > 8 \text{ ps})| \approx 2\%$, meaning most errors are of this magnitude. Larger errors diminish with increasing Δt, but a non-zero amount remain even at Δt =32 ps.

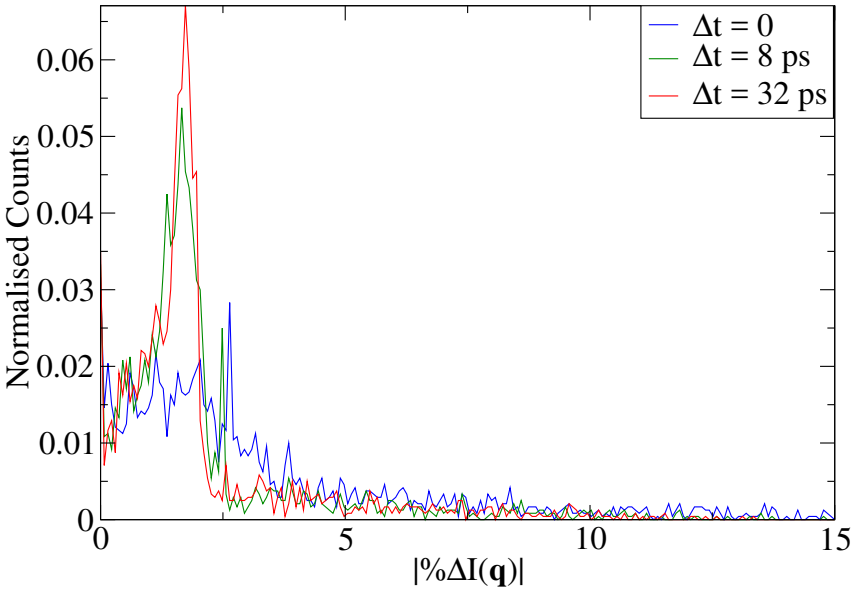
These results show that the difference between the IAM and the AIXRD approximations remain even after temporal-averaging, as would be physically observable in a X-ray diffraction experiment. To clarify these results further, the same analysis was done for incident X-rays along the x and y axes. Then the average of $|\%\Delta I(\mathbf{q}, \Delta t)|$ was taken over each incident direction (x, y, z), to take into account different molecular orientations, that is,

$$\langle |\%\Delta I(\mathbf{q}, \Delta t)| \rangle_{xyz} = \frac{\sum_{j=x,y,z} |\%\Delta I(\mathbf{q}, \Delta t)|_j}{3}. \tag{4.24}$$

The corresponding results from Fig. 4.15 are shown in Fig. 4.17, and the corresponding results from Fig. 4.16 are shown in Fig. 4.18. The new results averaged over the three incident X-ray directions are very similar to the results from incident X-rays along the z axis only. Specifically, for Δt =32 ps there is still $\approx 8\%$ error at $q \approx 0.5\ a_0^{-1}$ and $\approx 3.5\%$ error at $q \approx 1.0\ a_0^{-1}$; and most errors are of $\approx 2\%$ magnitude in the entire range $q \in [0, 4.2]\ a_0^{-1}$. The main difference is that there are slightly less large magnitude ($> 5\%$) errors due to the orientational averaging (compare Fig. 4.16(b) and Fig. 4.18(b)). This implies little molecular orientation



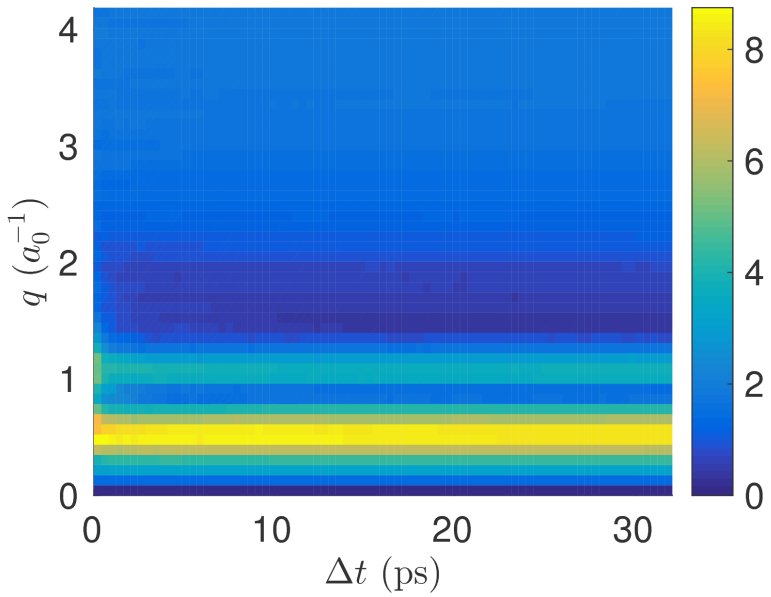
(a) Normalised counts, Eq. (4.19), as a function of percent error $|\%\Delta I(\mathbf{q})|$ and time-interval Δt.



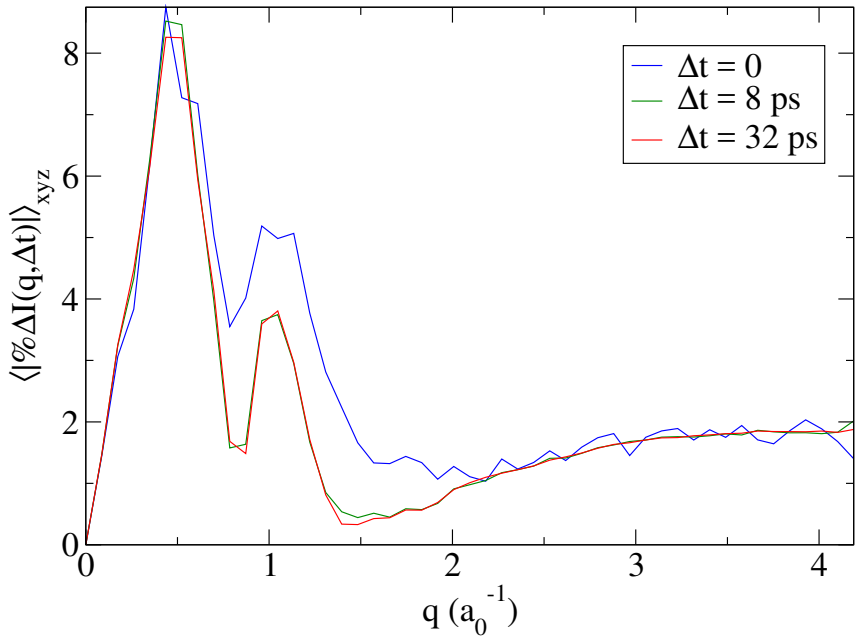
(b) Corresponding slices through (a) at time-intervals $\Delta t = 0, 8, 32$ ps.

Figure 4.16: (a) Normalised counts, Eq. (4.19), for 200 bins in the range $[0, 15]\%$ percent error, Eq. (4.21), between the time-integrated fragment-based AIXRD, and time-integrated IAM, for ($\lambda = 3\ a_0$) X-rays incident along the z-axis; and (b) for time-intervals $\Delta t = 0, 8, 32$ ps, corresponding to 1, 20, and 80 snapshots, each 0.4 ps apart.

dependence on the errors, even when averaging over different orientations, as would likely occur in a serial crystallography experiment. Thus, fragment-based AIXRD is an improvement over the IAM, with consideration to averaging over



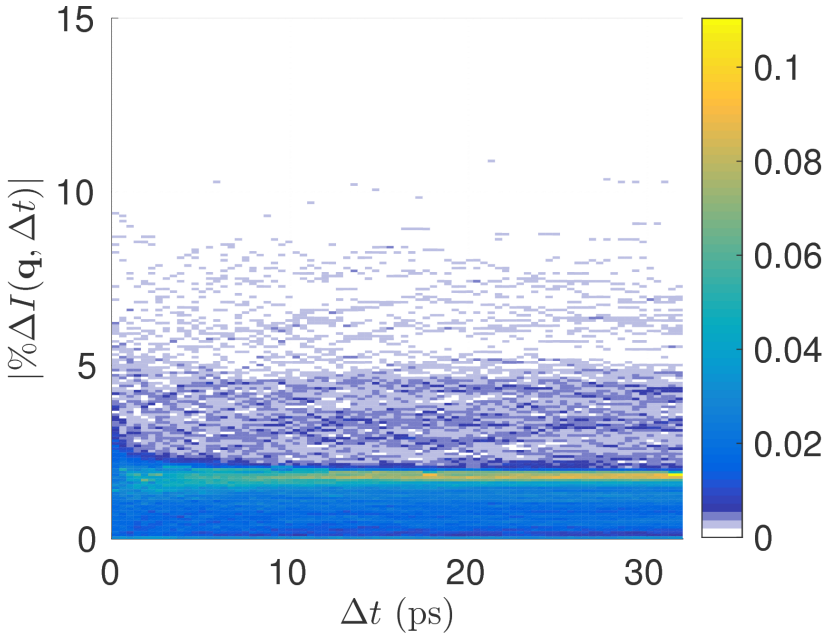
(a) Percent error, $\langle|\%\Delta I(\mathbf{q},\Delta t)|\rangle_{xyz}$.



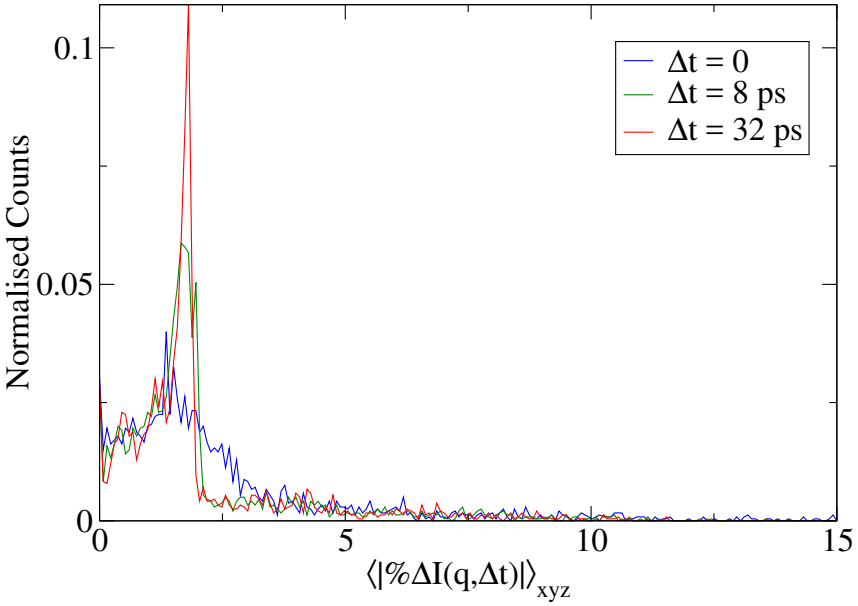
(b) Corresponding slices through $\langle|\%\Delta I(\mathbf{q},\Delta t)|\rangle_{xyz}$ at $\Delta t = 0, 8, 32$ ps.

Figure 4.17: Lysozyme X-ray diffraction percent error $\langle|\%\Delta I(q,\Delta t)|\rangle_{xyz}$, Eq. (4.24), between the fragment-based AIXRD and IAM methods, for absolute value of $q = |\mathbf{q}| = 4\pi\sin(\theta)/\lambda$ (azimuthal scattering angle, ϕ, integrated out), after temporal-averaging over time-interval Δt during a MD trajectory.

different snapshots of proteins in various molecular geometries and orientations.



(a) Normalised counts, Eq. (4.19), as a function of percent error $\langle|\%\Delta I(\mathbf{q})|\rangle_{xyz}$ and time-interval Δt.



(b) Corresponding slices through (a) at time-intervals $\Delta t = 0, 8, 32$ ps.

Figure 4.18: (a) Normalised counts, Eq. (4.19), for 200 bins in the range $\langle|\%\Delta I(\mathbf{q}, \Delta t)|\rangle_{xyz} \in [0, 15]\%$ percent error, Eq. (4.24), between the time-integrated fragment-based AIXRD, and time-integrated IAM, for ($\lambda = 3\ a_0$) X-rays; and (b) for $\langle|\%\Delta I(\mathbf{q}, \Delta t)|\rangle_{xyz}$ at time-intervals $\Delta t = 0, 8, 32$ ps, corresponding to 1, 20, and 80 snapshots, each 0.4 ps apart.

4.4 Conclusion

The method described in this Chapter is a conceptually simple improvement to the independent atom model (IAM) for elastic X-ray diffraction, whereby, a

molecule is divided into fragments larger than its individual atoms. Then, *ab initio* electronic structure theory is used on each fragment, followed by *ab initio* X-ray diffraction (AIXRD), as described in Chapter 3. In this way, the effect of electron interactions from atoms within the same fragment are taken into account in the X-ray molecular form-factor, as well as non-spherically-symmetric electron distributions about atomic-centres such as p-orbitals and delocalised electrons within phenyl rings. To further improve on this, dimers are defined as fragment pairs and also calculated with AIXRD, which takes into account fragment-fragment electron interactions.

Fragment-based AIXRD makes an *ab initio* quantum chemistry treatment of the electron distribution feasible in the theoretical calculation of the X-ray molecular form-factor for large molecules such as proteins and other biomolecules. In terms of AIXRD timing a simple sum approach is linearly faster with the number of fragments compared to a computation based on a full (single fragment) *ab initio* wave function. The pairwise approach is slower than full AIXRD but the maximum computational time taken is only larger by a factor of two. The real limitation that is overcome by a fragment-based approach is the *ab initio* quantum chemistry calculation e.g. Hartree-Fock which scales approximately with the 4th power of the total number of basis functions. This becomes very costly for large molecules such as proteins. A fragment-based approach gives a drastic speed up, bringing it into the realm of routine calculation possibility.

Comparisons to the IAM method were performed for the ethanol molecule, 20 common amino acid residues, diphenylalanine (FF), and a 128 residue protein; lysozyme. In all cases, the use of fragments larger than individual atoms are a considerable improvement in terms of electronic energy, electron distribution, and X-ray molecular form-factor. Introducing dimers and using a pairwise summation approach is a comparatively small improvement which can be made if needed. However, residue-sized fragments with a simple summation of fragment-terms gives results which are close to those from full AIXRD.

Although not calculated by full AIXRD, the moderately sized protein lysozyme was split into 128 residue-sized fragments and a simple sum of residue form-factors was used to approximate the full molecular form-factor. This gives a median of 2.23% and a mean of 4.23% difference from the IAM method, with particular scattering angles showing $> 80\%$ difference. The quantitative results for how close this method is to a full *ab initio* method for lysozyme are left for a future calculation when 100+ residue proteins are routinely treated with

electronic structure methods. It can be inferred from the $N = 2$ FF results that residue-sized fragment-based AIXRD has an approximate mean diffraction intensity percent difference to full *ab initio* of 0.23% with a median of 0.17-0.25%.

Temporal integration of the X-ray diffraction results calculated with fragment-based AIXRD and IAM for a short (100 ps) MD simulation of a single lysozyme molecule show that the resulting errors of the IAM approximation persist and are not averaged-out over time, Δt. This remains true for the same results averaged over three incident X-ray directions (x, y, z), which approximates an orientational-averaging of the protein. These results are relevant to X-ray structure determination and pump-probe diffraction experiments for proteins and other biomolecules, as there are significant errors (2-8%) introduced by not taking into account the true electronic structure. The fragment-based AIXRD approach described in this Chapter is a possible solution to this problem.

Chapter 5

State selective diffraction measurements: H_2

5.1 Introduction

Since their discovery by Wilhelm Conrad Rontgen in 1895, X-rays have been established as an invaluable probe of the structure of matter. X-ray scattering has had exceptional impact on our understanding of the structure and properties of atoms to proteins, crystals, and solids. This chapter examines elastic scattering from gas-phase H_2 pumped to specific electronic, vibrational, and rotational states, i.e. state-selective diffraction imaging. This work is motivated by recent experimental preparations of molecules in non-thermal distributions, such as H_2 [73], N_2 [71], and polyatomics such as acetylene [72]; and the unprecedented peak intensities of X-rays available at newly constructed X-ray free-electron lasers (XFEL) [130,131]. The *ab initio* X-ray diffraction (AIXRD) method, as described in Ch. 3 and in reference [1], is used for the theoretical predictions, which are benchmarked against earlier calculations [132] and recent experiments [50].

The idea to use elastic scattering to image molecular states, whether electronic [1,65,66], vibrational [77], or rotational [78], as well as aligned gas-phase molecules [79, 80] has been around for quite a while. Experimental realisations are rare and limited to a few unique cases such as partially aligned and somewhat state-selected molecules [48, 75, 81, 82], and indirect detection of metastable excited electronic states via changes in geometry [67, 68]. Among the references above, [1, 48, 65–68, 80] pertain to X-ray scattering and [75, 77–79, 81, 82] to electron

scattering. Although electron scattering has been thought as a preferable technique for gas-phase scattering due to greater scattering cross-sections [133], the arrival of XFELs, such as the LCLS [22], the European XFEL [24], SACLA [23], and FERMI [20], has renewed interest in gas-phase X-ray scattering [6, 48], partially driven by the prospect of crystal-free biomolecular structure determination [14, 34–36].

High-resolution synchrotron experiments of X-ray differential cross-sections have been recently performed, and are capable of distinguishing between elastic [50] and inelastic [51–53] scattering contributions. In one such experiment, the elastic X-ray scattering from gas-phase H_2 in its electronic ground state was measured [50]. This is a strong motivating factor for this chapter and the theoretical work within. In addition, the experimental rise of molecular orientation (i.e. $+M$ magnetic sublevels have different population to $-M$ sublevels) and alignment (i.e. only $|M|$ is controlled) gives the motivation for calculating these states and their corresponding X-ray scattering factors (specifically for H_2 in this Chapter). Bartlett et al. [73] reports retention of initial polarisation for >100 ns for H_2 making it a suitable candidate for future scattering experiments. Due to large retention times, it is more feasible to consider imaging the oriented and aligned states ($v = 1; J = 2; M = 0, 2$). In addition, an oriented or aligned molecule gives more structural information via a second diffraction coordinate, i.e. $\mathbf{q} = \mathbf{q}(\theta, \phi)$ for polar and azimuthal scattering angles θ and ϕ.

5.2 Theory

5.2.1 X-ray scattering from diatomic molecules

A Born-Oppenheimer wave-function for a diatomic molecule has separate nuclear (rotational and vibrational) and electronic parts,

$$\Psi = Y_M^J(\Theta, \Phi)\psi_v(R)\psi_\alpha(\mathbf{r}; R), \tag{5.1}$$

where v is vibrational quantum number, J and M are rotational quantum numbers, α is the electronic state, R is the diatomic bond-length, and Θ and Φ are molecular frame rotation angles for a linear molecule. The rotational wave

function is a spherical harmonic, defined as,

$$Y_M^J(\Theta, \Phi) = ne^{iM\Phi} P_M^J(\cos\Theta), \tag{5.2}$$

with associated Legendre polynomials P_M^J and normalisation constant n. The absolute value squared, $|Y_M^J(\Theta, \Phi)|^2$; real value squared, $\Re(Y_M^J(\Theta, \Phi))^2$; and imaginary value squared, $\Im(Y_M^J(\Theta, \Phi))^2$ are shown in Fig. 5.1 for $J = 1, M = 0, 1$ and $J = 2, M = 2$. The electron density is,

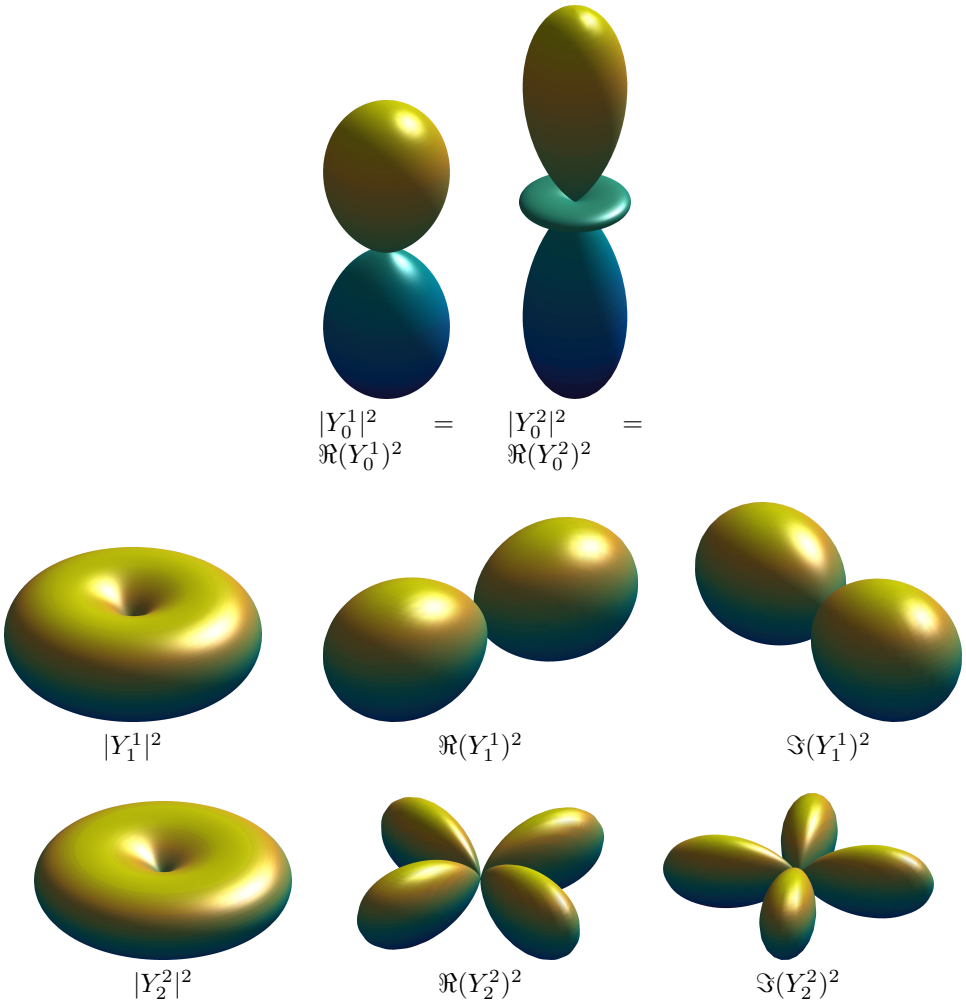


Figure 5.1: Spherical harmonics: absolute value squared, $|Y_M^J(\Theta, \Phi)|^2$; real value squared, $\Re(Y_M^J(\Theta, \Phi))^2$; and imaginary value squared, $\Im(Y_M^J(\Theta, \Phi))^2$. The polar and azimuthal angles Θ and Φ are defined from the vertical (z) axis, which is also the C_∞ principle rotation axis, and the origin is at the point of inversion symmetry for all Y_M^J. The colour and lighting are for aesthetic purpose only.

$$\rho(\mathbf{r}) = \langle\Psi|\hat{\rho}(\mathbf{r})|\Psi\rangle = \int_0^{2\pi}\int_0^{\pi} |\psi_{JM}(\Theta,\Phi)|^2 \int_{R_{\min}}^{R_{\max}} |\psi_v(R)|^2 \rho_\alpha(\mathbf{r};R,\Theta,\Phi)\mathrm{d}R\sin\Theta\mathrm{d}\Theta\mathrm{d}\Phi, \tag{5.3}$$

with N electron density operator $\hat{\rho}(\mathbf{r}) = \sum_i^N \delta(\mathbf{r}_i - \mathbf{r})$, and ρ_α is a multi-determinant density for electronic state α. The elastic X-ray molecular form-factor is the Fourier transform of the electron density,

$$f(\mathbf{q}) = \int_0^{2\pi}\int_0^{\pi} |Y_M^J(\Theta,\Phi)|^2 \int_{R_{\min}}^{R_{\max}} |\psi_v(R)|^2 f_\alpha(\mathbf{q};R,\Theta,\Phi)\mathrm{d}R\sin\Theta\mathrm{d}\Theta\mathrm{d}\Phi, \tag{5.4}$$

where $f_\alpha(\mathbf{q};R,\Theta,\Phi) = \hat{F}_{\mathbf{r}}[\rho(\mathbf{r};R,\Theta,\Phi)](\mathbf{q})$, and its evaluation is outlined in Section 3.2 for a multi-configurational electronic structure calculation using a Gaussian basis set. As this equation involves triple integration over both rotation angles and the bond-length, the evaluation of $f_\alpha(\mathbf{q};R,\Theta,\Phi)$ must be computationally fast. This is one of the reasons H_2 was chosen for this chapter.

5.3 Results and discussion

5.3.1 Potential energy curves and vibrational wave functions

Energy (Hartree)	MCSCF(2,7)	Reference[a]	$\lvert\Delta U\rvert$
$U_X(R_0)$	-1.17089198	-1.17447589	0.00358391
$U_{EF}(R_E)$	-0.71554080	-0.71815456	0.00261376
$U_{EF}(R_F)$	-0.70762693	-0.71452167	0.00689474

Table 5.1: H_2 energies calculated by MSCSF(2,7)/aug-cc-pVQZ compared to reference values at the X minimum $U_X(R_0)$; and the inner and outer EF minima $U_{EF}(R_E)$ and $U_{EF}(R_F)$ respectively. The absolute difference $|\Delta U|$ between the values from both methods are also shown. Note that the *ab initio* minimum bond lengths differ slightly from reference values, with maximum difference of $0.016a_0$ at the outer EF minimum.
[a] See Refs [134, 135].

Accurate potential energy curves (PECs) for H_2 in its ground state and first excited $^1\Sigma_g^+$ state, denoted X and EF respectively, have been calculated using essentially the exact H_2 wave function [134, 135]. They can be closely reproduced

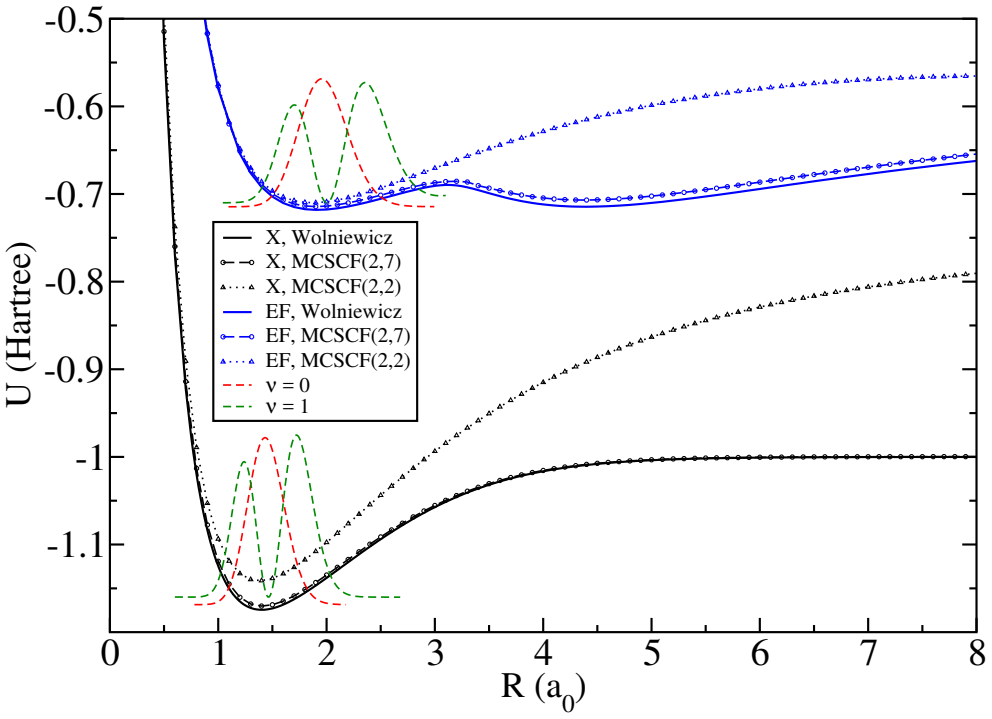
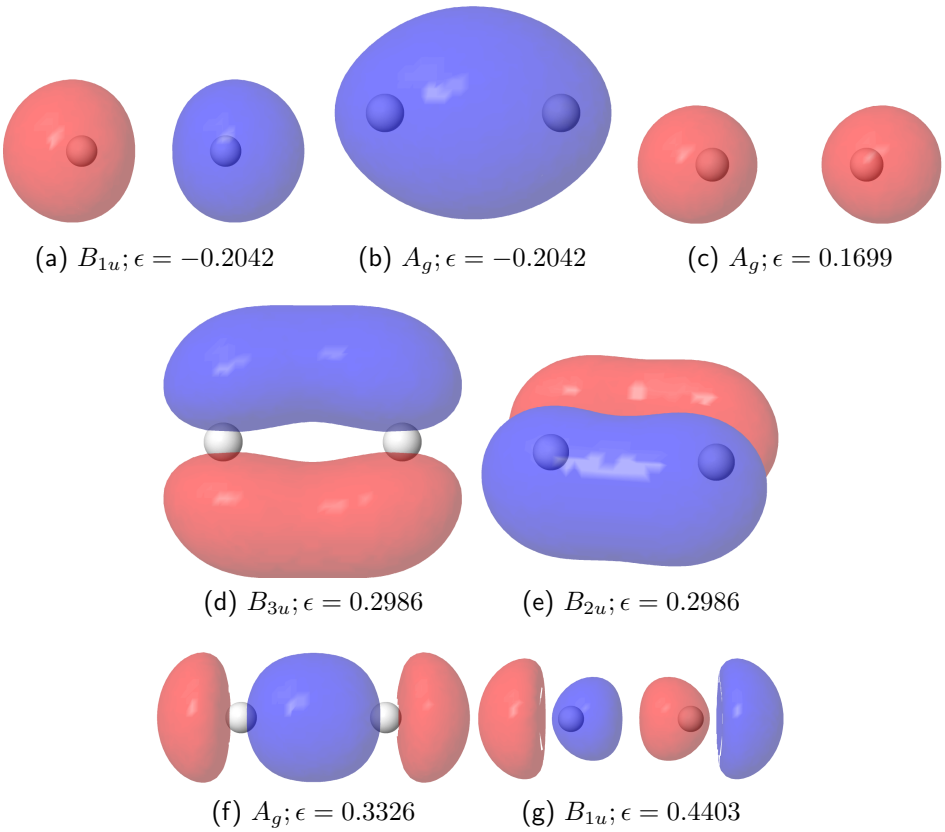


Figure 5.2: H_2 PECs for the X and EF $^1\Sigma_g^+$ states calculated with MCSCF(2,7) and MCSCF(2,2) with the aug-cc-pVQZ basis set, compared to highly accurate PECs [134, 135]. Bond-length probabilities for $v = 0$ and $v = 1$ vibrational distributions are shown on each PEC with ascetically adjusted heights.

with the multi-configurational self-consistent field (MCSCF) method using an appropriate active space. A comparison between these accurate PECs and those from MCSCF with two different active spaces, both using the aug-cc-pVQZ basis set, is shown in Fig. 5.2. The (2,7) active space calculation gives good agreement with the reference PECs, but the (2,2) calculation fails at large R. It also correctly predicts the second minimum on the EF PEC, whereas the (2,2) active space does not. The seven molecular orbitals (MOs) in the (2,7) active space are shown in Fig. 5.3 at the equilibrium bond-distance of the second EF minimum, $R_F = 4.37475793\ a_0$. To predict the second stationary point, it is necessary to include additional bonding-MOs such as the third A_g MO, and the B_{2u} and B_{3u} MOs. These orbitals and the B_{1u} MOs, which are σ^* orbitals, also improve the X PEC, especially at large bond-length. See Table 5.1 for MCSCF(2,7)/aug-cc-pVQZ energies of each minima compared to the Wolniewicz values.



(a) $B_{1u}; \epsilon = -0.2042$ (b) $A_g; \epsilon = -0.2042$ (c) $A_g; \epsilon = 0.1699$

(d) $B_{3u}; \epsilon = 0.2986$ (e) $B_{2u}; \epsilon = 0.2986$

(f) $A_g; \epsilon = 0.3326$ (g) $B_{1u}; \epsilon = 0.4403$

Figure 5.3: H_2 MCSCF(2,7)/aug-cc-pVQZ active space MOs at the second EF minima ($R_F = 4.37475793\ a_0$) with D_{2h} point group symmetry labels. From left to right along each row MOs are in order of energy, $U_{\mathrm{MO}}(R_F)$ (eigenvalues, ϵ, are shown).

5.3.2 Model chemistry for X-ray scattering calculations

For consistency, it is desirable to use the same model chemistry as the calculation of the PECs in subsequent *ab initio* X-ray diffraction (AIXRD) calculations. It is also useful to compare basis sets to see convergence in the X-ray curves. This is shown in Fig. 5.4 via percent difference, defined as,

$$\%\Delta I(q) = 100\frac{I_{\mathrm{method}} - I_0}{I_0}, \tag{5.5}$$

where I_0 and I_{method} are rotationally-averaged (thermal-averaging) X-ray intensities, and I_0 is the reference signal. In this case, the reference uses the aug-cc-pVQZ basis set. Each curve was calculated at the MSCSF(2,7)/aug-cc-pVQZ ground state minimum bond-length, R_0, using AIXRD MSCSF(2,7) for all basis sets except STO-3G and 6-31G where MSCSF(2,2) was used. This is because there are not enough basis functions in the STO-3G and 6-31G basis sets to perform a MSCSF(2,7) calculation. The independent atom model (IAM) curve is also

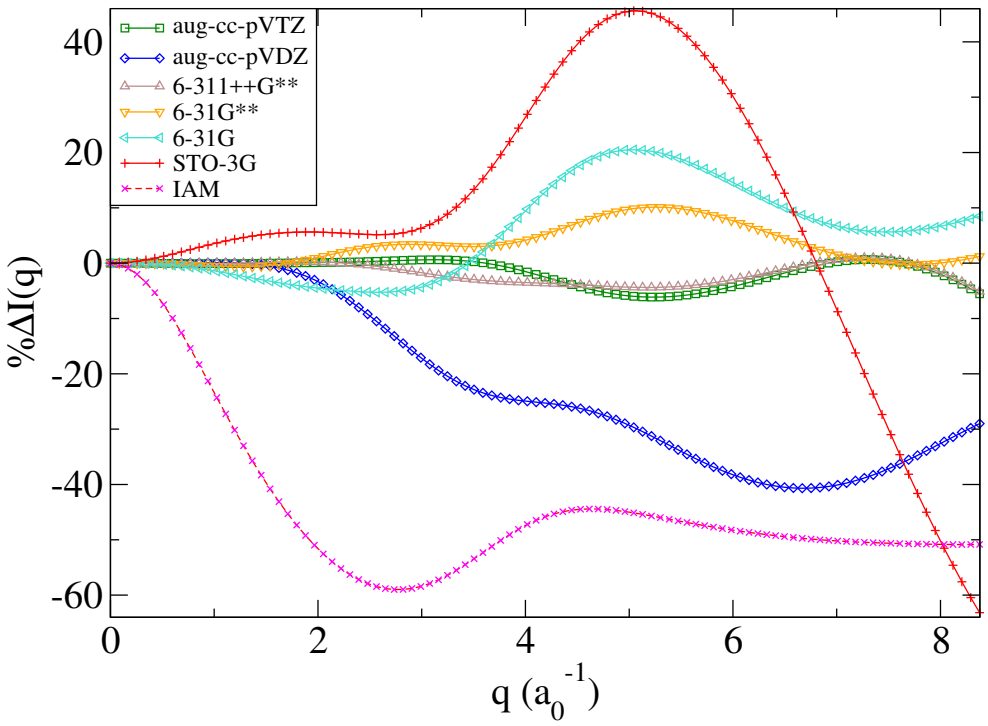


Figure 5.4: H_2 basis set convergence for orientationally-averaged elastic X-ray diffraction intensity calculated with *ab initio* X-ray diffraction (AIXRD). Percent difference, $|\%\Delta I(q)| = 100(I_A - I_0)/I_0$, is shown; where I_A is the diffraction signal from method A, and the reference signal I_0 is from an MCSCF(2,7)/aug-cc-pVQZ AIXRD calculation. The other curves were calculated with MCSCF(2,7) with different basis sets, except STO-3G and 6-31G used MCSCF(2,2). All curves are at the MCSCF(2,7)/aug-cc-pVQZ optimised geometry. Also shown is the independent atom model (IAM) curve, calculated via the square of a sum of tabulated atomic X-ray form-factors from [62].

shown for comparison. Table 5.2 gives the mean and maximum absolute errors. The mean absolute error is defined as,

$$\langle|\%\Delta I(q)|\rangle = \frac{1}{\Delta q}\int_{q_{\min}}^{q_{\max}} |\%\Delta I(q)|\mathrm{d}q, \tag{5.6}$$

with integration over the interval $[q_{\min}, q_{\max}] = [0, 8.3]\ a_0^{-1}$ and $\Delta q = q_{\max} - q_{\min}$.

In terms of the AIXRD method, larger basis sets naturally improve results. In fact, the energy difference $|E - E_{\text{ref}}|$ shown in Table 5.2 predicts quite well the quality of the AIXRD calculations. This energy gap correctly ranks the top four performing methods, and the three poorest performing ones. The closest method to the reference uses the aug-cc-pVTZ basis set with $\langle|\%\Delta I(q)|\rangle = 1.7\%$, closely followed by the 6-311++G** basis set with $\langle|\%\Delta I(q)|\rangle = 1.8\%$. The IAM method is furthest from the reference with $\langle|\%\Delta I(q)|\rangle = 44.1\%$. It is important

METHOD	Error (%)		$\lvert E - E_{\text{ref}}\rvert$	Primitives		Speed-
	Mean	Max	$(10^{-2}\ E_h)$	N_{g}	N_{gp}	up
IAM	44.1	59.0	-	-		826.5
STO-3G	20.3	63.1	3.723	12	30	110.2
6-31G	7.9	20.5	2.819	16	56	59.0
6-31G**	3.4	10.1	1.011	28	90	36.7
6-311++G**	1.8	4.4	0.692	36	182	18.2
aug-cc-pVDZ	21.7	40.7	1.065	60	240	13.8
aug-cc-pVTZ	1.7	6.1	0.427	148	1122	3.0
aug-cc-pVQZ[1]	0.0	0.0	0.358	300	3306	1.0

Table 5.2: Comparison of accuracy and speed for various levels of theory for scattering from the $X^1\Sigma_g^+$ ground state at $R_0 = 1.40496862\ a_0$. The mean and maximum errors shown correspond to the errors in Fig. 5.4, with the scattering from MSCSF$(2,7)$/aug-cc-pVQZ taken as reference, see Eq. (5.5). The gap $|E - E_{\text{ref}}|$ between calculated energies and Wolniewicz reference value [134] is used as a proxy for *ab-initio* convergence. The speed of the scattering calculations scales linearly with the number of non-zero unique Gaussian products per molecular orbital, N_{gp}, with the number of Gaussian primitives given by N_{g}. The IAM calculation uses tabulated atomic form-factors [62].

to minimise computational time, keeping in mind evaluation of the triple integral in Eq. (5.4) which will be used in Section 5.3.4. The AIXRD scaling for H_2 with different basis sets is shown in Table 5.2. It scales non-linearly with the number of primitive Gaussians per basis set, N_{g}. The scaling is proportional to the number of unique non-zero Gaussian products per MO, which in this case is $N_{\text{gp}} \approx N_{\text{g}}^{1.42}$. Taking this and the diffraction convergence into account, an ideal basis set is aug-cc-pVTZ as it comes very close to the quadruple-zeta basis set X-ray diffraction result while being ≈3 times faster computationally. As H_2 is the smallest molecule, this basis set is affordable. For larger molecules, the 6-311++G** basis set would be advisable as it is ≈18 times faster than aug-cc-pVQZ while still maintaining a low $\langle|\%\Delta I(q)|\rangle$.

5.3.3 Electronic structure in diffraction

After selecting an active space and basis set for accurately describing the H_2 X and EF $^1\Sigma_g^+$ PECs while considering the AIXRD calculation, orientationally-averaged

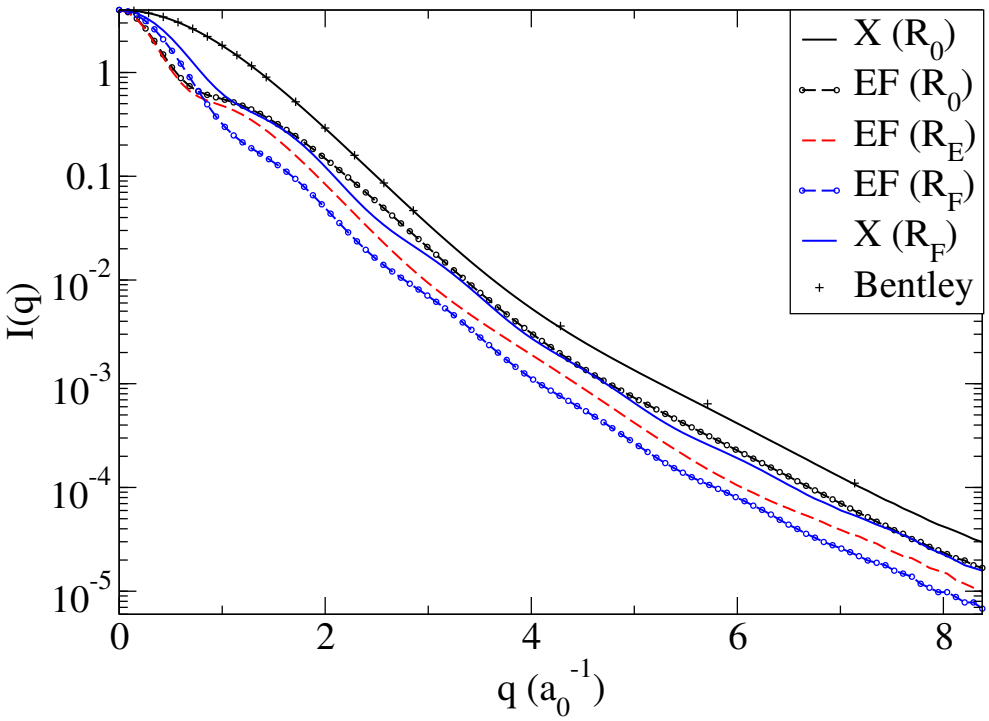


Figure 5.5: H_2 X and EF $^1\Sigma_g^+$ orientationally-averaged X-ray diffraction intensities: At the X minimum geometry on the X and EF PECs, X(R_0) and EF(R_0) respectively; the inner and outer EF minima geometries (labelled E and F respectively) on the EF PEC, EF(R_E) and EF(R_F); and the outer EF minimum geometry on the X PEC, X(R_F). AIXRD with MCSCF(2,7)/aug-cc-pVTZ was used for each calculation which strongly matches theory by Bentley and Stewart [132, 136].

diffraction intensities are computed for points of interest on each PEC, which are shown in Fig. 5.5. A previous theory curve from Bentley and Stewart [132, 136] is also shown, which is based on a Davidson-Jones expansion of essentially the exact wave function for the H_2 ground state at equilibrium geometry. The points of interest chosen for Fig. 5.5 are: on the X PEC at its equilibrium bond-length, X(R_0); on the EF PEC at the X equilibrium bond-length, EF(R_0); at the inner and outer EF minima, EF(R_E) and EF(R_F) respectively (as defined by the double minima on the solid blue curve in Fig. 5.2); and on the X PEC at the outer EF minimum bond-length, X(R_F). This essentially follows an imaginary excitation from the X minimum vertically to the EF PEC, relaxation to the inner EF minimum, barrier-crossing (e.g. from thermal energy) to the outer EF minimum, finally returning to the X curve at an increased bond-length.

Each X-ray diffraction curve is distinct, showing that not only changes in bond-length alter the diffraction but also changes in electronic structure. The low-q region $q \in [0, 1]$ a_0^{-1} is strongly affected by the overall size of the electron distribution. The vertical excitation from the X state to the EF state (X→EF) causes

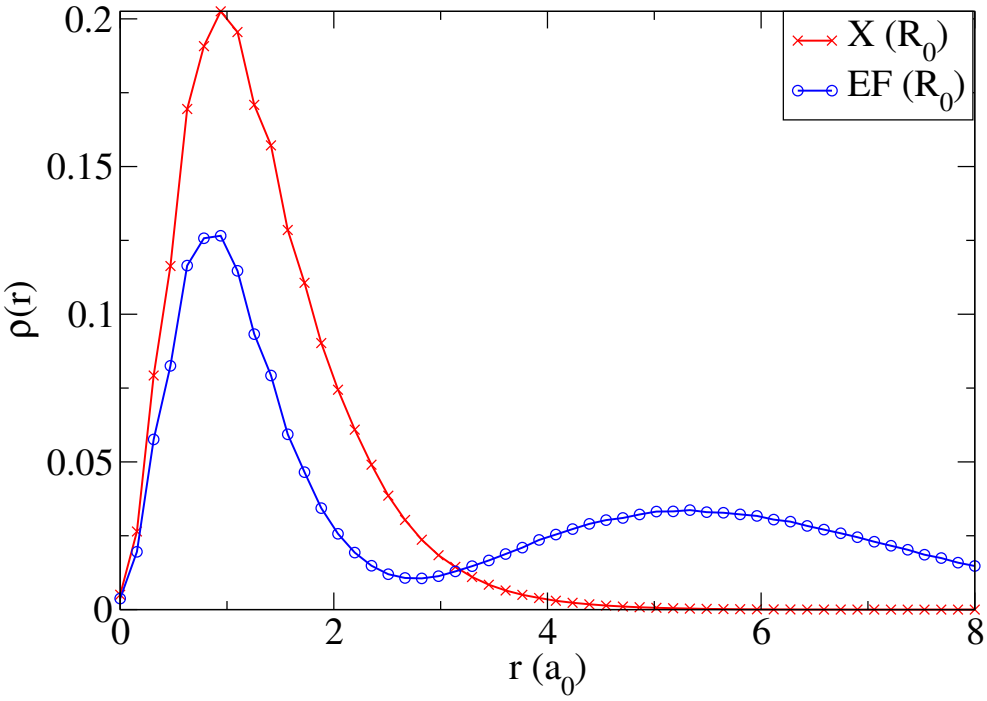


Figure 5.6: H_2 radial electron density calculated with MCSCF(2,7)/aug-cc-pVTZ for X and EF states at the ground state equilibrium bond-length, centred at centre of mass.

an outward radial shift in the electron density, shown in Fig. 5.6. This causes a drop in signal in the low-q region, consistent with the general principles of Fourier transforms. Also notable is the appearance of a large low-q ($\approx 1.5\ a_0^{-1}$) peak after the vertical excitation X→EF. Using the reciprocal relation $q = 2\pi/d$, where d is the corresponding real space distance, gives $d \approx 4.2\ a_0$. This is approximately the distance between the inner and outer peaks in the EF radial electron density at the X equilibrium bond-length, shown in Fig. 5.6. The appearance of the second peak in mainly caused by a change of occupancy in the A_g orbitals. In the ground state, the first A_g MO (Fig. 5.3b) is highly occupied ($a_{\text{occ}} = 1.9634$). After excitation the second A_g MO (Fig. 5.3c) gains occupancy ($a_{\text{occ}} = 0.8183$). The X (R_0) radial electron distribution is a smooth single-peaked curve, and its corresponding scattering curve is almost a Gaussian, which is characteristic of no second peak in the electron distribution.

In general, it is hard to separate the effect of bond-length from the effect of the electronic structure on the scattering. In fact, the differences between the X (R_0) and X (R_F), and the EF (R_E) and EF (R_F) curves in Fig. 5.5 are caused by a combination of bond-length and electronic structure changes. For example, on the EF PEC at the inner EF minimum, R_E, the first two A_g orbitals have significant electron occupancy ($a_{\text{occ}} = 1.1730$ and $a_{\text{occ}} = 0.8175$, respectively), and have

single-centred diffuse shape. At the outer EF minimum, R_F, the B_{1u} orbital which has a two-centre shape (Fig. 5.3a) becomes the most occupied ($a_{\rm occ} = 1.0251$ from $a_{\rm occ} = 0.0075$ at R_E). At the same time the second A_g orbital decreases in occupancy to $a_{\rm occ} = 0.1094$ and changes from a one-centre to a two-centre shape (Fig. 5.3c) while the first A_g orbital retains its one-centre shape (Fig. 5.3b) and a relatively large occupancy ($a_{\rm occ} = 0.8324$). In addition, the more diffuse orbitals, B_{2u} and B_{3u}, gain about 1% of total occupancy. This shows that the change in bond-length from R_E to R_F causes significant electronic structure changes, making it difficult to decouple the effects of electronic structure and bond-length. Another outward radial shift is apparent when considering the change in $I(q)$ based on a small increase in bond-length such as the change between EF (R_0) and EF (R_E) in Fig. 5.5, which have similar shaped curves but are shifted in q.

5.3.4 Accounting for vibrations and rotations

To investigate the effect of molecular vibrations on the X-ray scattering, vibrational wave functions for $v = 0, 1, 2, 3$ on the X surface and for $v = 0, 1$ on the EF surface were calculated on the Wolniewicz PECs. The vibrational distributions for X and EF $v = 0, 1$ are shown in Fig. 5.2. The 1st EF minimum has lower gradient $\mathrm{d}U/\mathrm{d}R$ than the X minimum, thus has more diffuse vibrational distributions. Not shown are the $v = 0, 1$ vibrational wave functions for the second EF minimum and the $v = 2$ EF wave function which spreads across both EF minima. Orientationally-averaged X-ray diffraction curves were calculated for the X $v = 0, 1, 2, 3$ and EF $v = 0, 1$ states using AIXRD with MCSCF(2,7)/aug-cc-pVQZ wave functions. They are shown in Fig. 5.7, which also shows the equilibrium bond-length curves for both electronic states.

It is clear that the electronic state change X→EF is more apparent in the diffraction than vibrational transitions. A slight change in the bond-length distribution, thus in $\rho(\mathbf{r}; R)$, caused by vibrational excitation gives greater changes at high-q ($> 1\ a_0^{-1}$). The $v = 0$ state scattering curves are very close to the equilibrium bond-length (R_0) curves, only differing slightly at q ($> 4\ a_0^{-1}$). Their vibrational distributions are almost symmetric Gaussians about R_0 (see Fig. 5.2), therefore have average bond-lengths of $\approx R_0$ and barely affect the diffraction. Higher vibrational states increase the average bond-length, thus have similar shaped curves to those of the same electronic state, but have lower intensity as q increases.

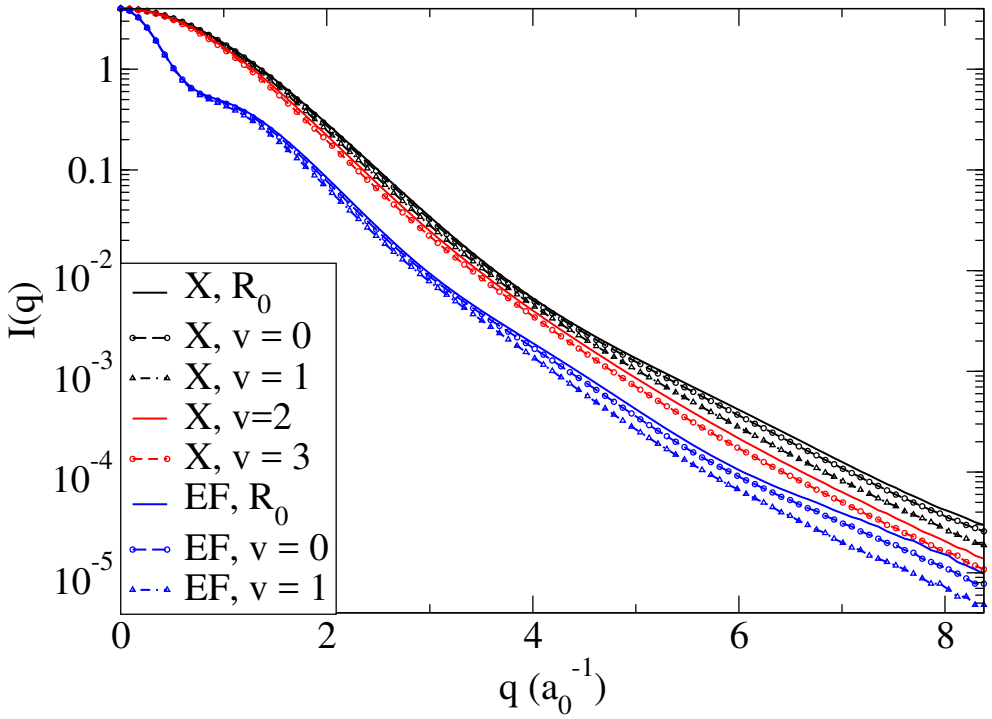
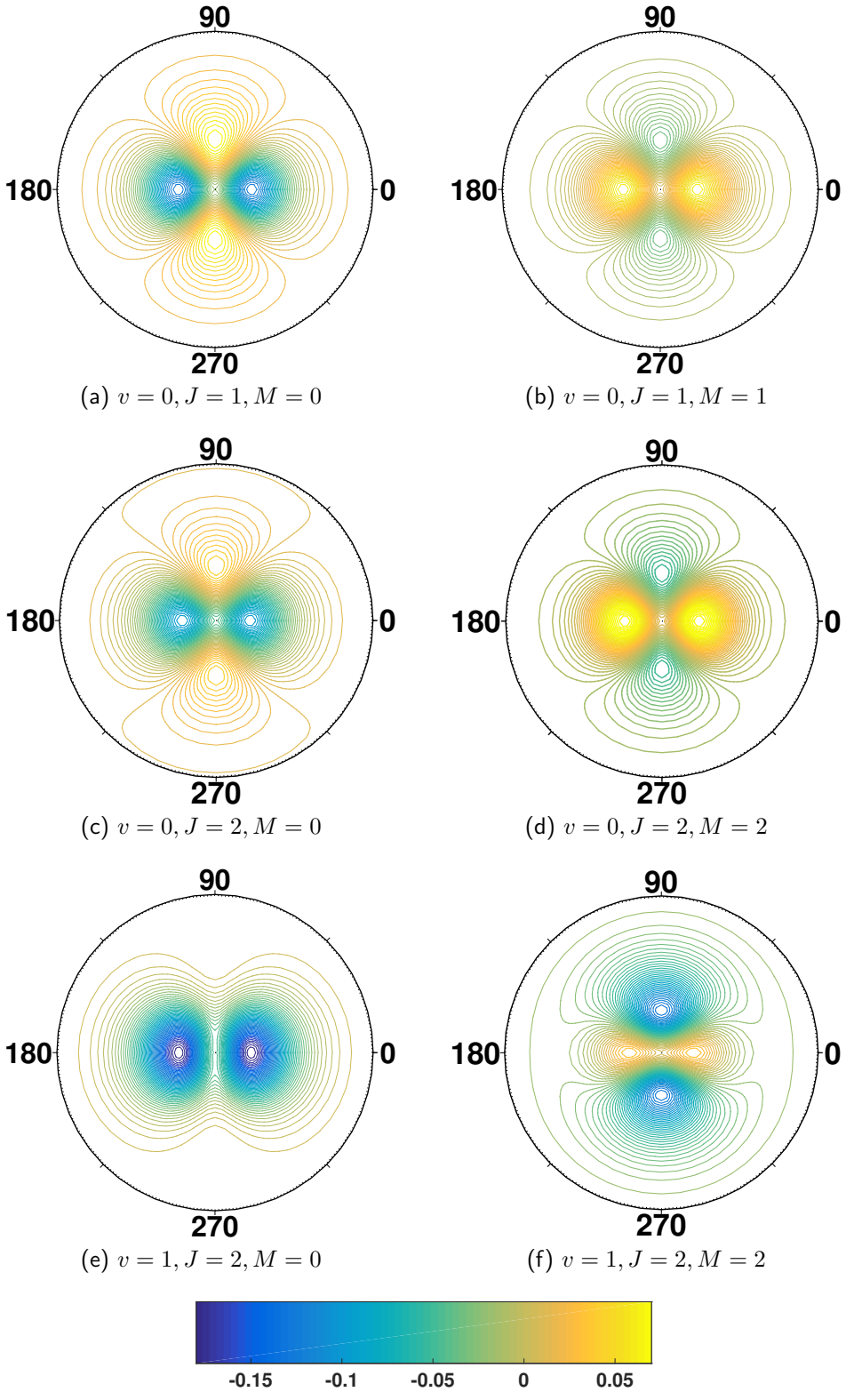


Figure 5.7: H_2 X $v = 0, 1, 2, 3$ and EF $v = 0, 1$ $^1\Sigma_g^+$ orientationally-averaged X-ray diffraction intensities, also showing the curves for the equilibrium bond-length R_0 for each electronic state. AIXRD with MCSCF(2,7)/aug-cc-pVTZ used at each iteration of bond-length in the vibrational distributions, which were calculated by the shooting method on the Wolniewicz [134, 135] X and EF PECs.

Next, scattering from specific electronic, vibrational, and rotational states is considered. The addition of non-ground state rotational wave functions breaks the rotational symmetry of the scattering patterns, and gives more information at the detector. Here the focus is on the states reported by Bartlett et al. [73] that retain polarisation for > 100 ns, the X $^1\Sigma_g^+$ $v = 1, J = 2, M = 0, 2$ states.

Diffraction difference images between theoretically pumped states and the ground state ($v = 0, J = 0, M = 0$), $\Delta I = I_{vJM} - I_{000}$, are shown in Fig. 5.8 for 2.60 keV ($\lambda = 9\ a_0 = 4.96$ Å) X-rays incident along the y-axis, and in Fig. 5.9 for X-rays incident along the z-axis; for $v = 0$ states with $J = 1, M = 0, 1$; $J = 2, M = 0, 2$; and $v = 1$ states with $J = 2, M = 0, 2$. Their rotational distributions are the absolute square of spherical harmonics $|Y_M^J|^2$, which have C_∞ principle axes of rotation, denoted the z-axis. X-rays incident perpendicular to this axis (as in Fig. 5.8) see a side view of either a dumbbell-shape (Figs 5.1a and 5.1b), or a disc shape (Figs 5.1c and 5.1f), whereas X-rays incident parallel to this axis (as in Fig. 5.9) see a top view which is centrosymmetric for all L and M. These shapes are blurred by the vibrational distribution and the diffuse electron density, *cf* Eq. (5.4). Nonetheless, specific magnetic rotational

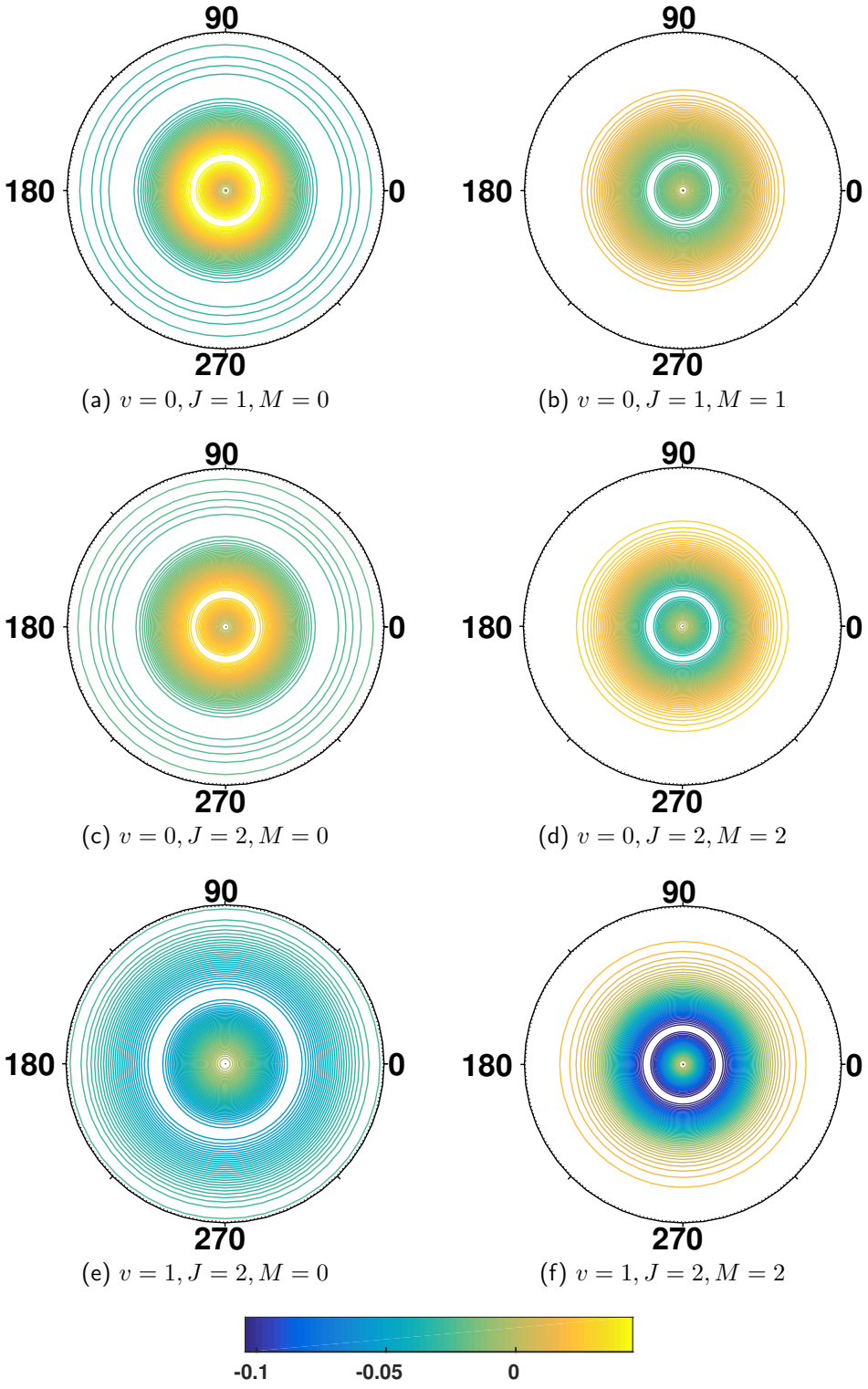


(a) $v = 0, J = 1, M = 0$ (b) $v = 0, J = 1, M = 1$

(c) $v = 0, J = 2, M = 0$ (d) $v = 0, J = 2, M = 2$

(e) $v = 1, J = 2, M = 0$ (f) $v = 1, J = 2, M = 2$

Figure 5.8: H_2 detector diffraction difference signals, $\Delta I(q) = I_{vJM}(q) - I_{000}(q)$, between various vibrational/rotational states and the ground state ($v = 0, J = 0, M = 0$). AIXRD was used with MCSCF(2,7)/aug-cc-pVTZ. X-rays incident along the y-axis.



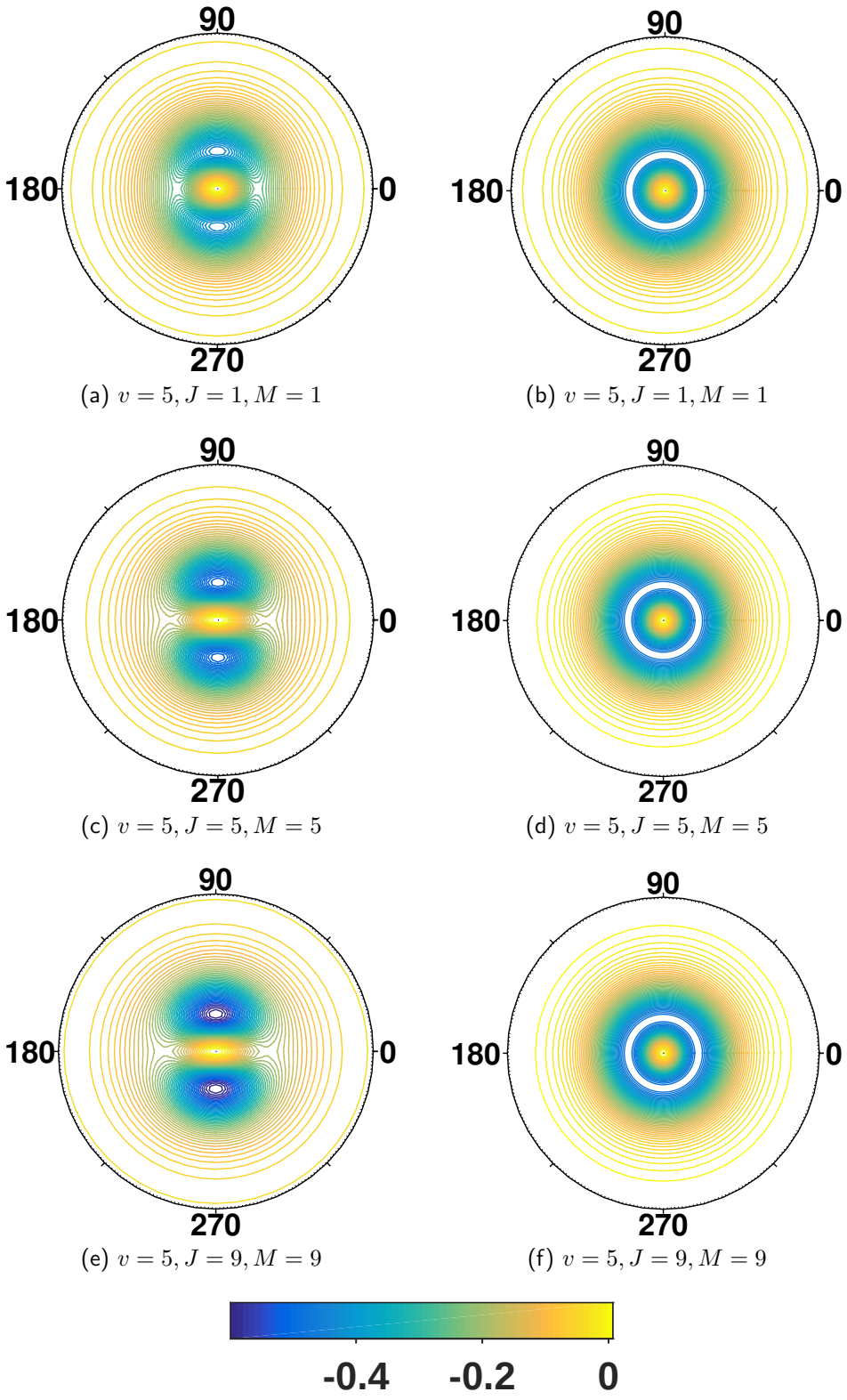
(a) $v = 0, J = 1, M = 0$ (b) $v = 0, J = 1, M = 1$

(c) $v = 0, J = 2, M = 0$ (d) $v = 0, J = 2, M = 2$

(e) $v = 1, J = 2, M = 0$ (f) $v = 1, J = 2, M = 2$

Figure 5.9: H_2 detector diffraction difference signals, $\Delta I(q) = I_{vJM}(q) - I_{000}(q)$, between various vibrational/rotational states and the ground state ($v = 0, J = 0, M = 0$). AIXRD was used with MCSCF(2,7)/aug-cc-pVTZ. X-rays incident along the z-axis.

states are distinguishable in the X-ray diffraction, i.e. the left and right images of each row in Figs 5.8 and 5.9, which only differ by magnetic quantum number



(a) $v = 5, J = 1, M = 1$ (b) $v = 5, J = 1, M = 1$

(c) $v = 5, J = 5, M = 5$ (d) $v = 5, J = 5, M = 5$

(e) $v = 5, J = 9, M = 9$ (f) $v = 5, J = 9, M = 9$

Figure 5.10: H_2 detector diffraction difference signals, $\Delta I(q) = I_{vJM}(q) - I_{000}(q)$, between various vibrational/rotational states and the ground state $(v = 0, J = 0, M = 0)$. AIXRD was used with MCSCF(2,7)/aug-cc-pVTZ. Left column is for X-rays incident perpendicular to z-axis, right column is for X-rays incident parallel to z-axis.

M, are strongly distinct, and in a qualitative way are almost the inverse of each other. This is because they correspond to dumbbell-shaped (density along the z-axis) and horizontal disc-shaped (density on the xy-plane) rotational distributions respectively. Comparing the 1st two rows of these figures represents a change $J = 1 \rightarrow 2$ (and $M = 1 \rightarrow 2$ for the right column) and does not strongly alter the X-ray difference patterns ΔI. This is because increasing the rotational quantum number L does not dramatically alter the shape of $|Y_M^J|^2$, for example compare Figs 5.1a and 5.1b, which are both dumbbell-shaped rotational distributions; and Figs 5.1c and 5.1f, which are both disc-shaped rotational distributions. Not shown however is an intermediate state between these extremes, such as $|Y_1^2|^2$, which has an hourglass shape. Its corresponding scattering signal would simply look like an intermediate between the $M = 2$ and $M = 0$ scattering signals. X-rays incident parallel to the z-axis only see the top-view of the dumbbell and disc-shaped rotational distributions, which is a radially-symmetric view-point. Therefore, the detector difference images ΔI in Fig. 5.9 have no dependence on the azimuthal scattering angle ϕ.

Comparing the second and third rows of Figs 5.8 and 5.9 involves a change in vibrational quantum number $v = 0 \rightarrow 1$. This gives a relatively large change in ΔI as the $v = 1$ state has a larger average bond-length $\langle R \rangle$ than the $v = 0$ state. It can also be thought of as the X-rays distinguishing a vibrational state change in the third row images, whereas the second row images correspond to the $v = 0$ state, i.e. no vibrational state change. This effect shows the power of exciting a rotational state and increasing the information available on the detector (from a function of θ to a function of θ and ϕ), allowing greater distinction between not only rotational states but vibrational states as well.

Fig. 5.10 shows X-ray diffraction difference signals, ΔI, for the $v = 5$ vibrational state, and for progressively higher energy rotational states ($J = M = \{1, 5, 9\}$) down each column. The left and right columns represent incident X-rays perpendicular and parallel to the z-axis respectively. There is a clear progression of increasing signal down the left column, as the disc-shaped rotational distribution becomes more and more planar, i.e. Figs 5.10a, 5.10c, and 5.10e. This is not true for the signals from incident X-rays parallel to the z-axis (Figs 5.10b, 5.10d, and 5.10f), as they only see the top-view of the disc-shaped distribution which does not change much with increasing $J = M$. Based on these results the information at the detector has a strong dependence on the incident X-ray direction. The columns in Fig. 5.10 represent the most and least information extremities.

The left column gives the largest variance with the azimuthal scattering angle ϕ, whereas the right column gives no variance with ϕ and the advantage of having aligned or oriented molecules is mostly lost.

5.3.5 Comparison to experiment

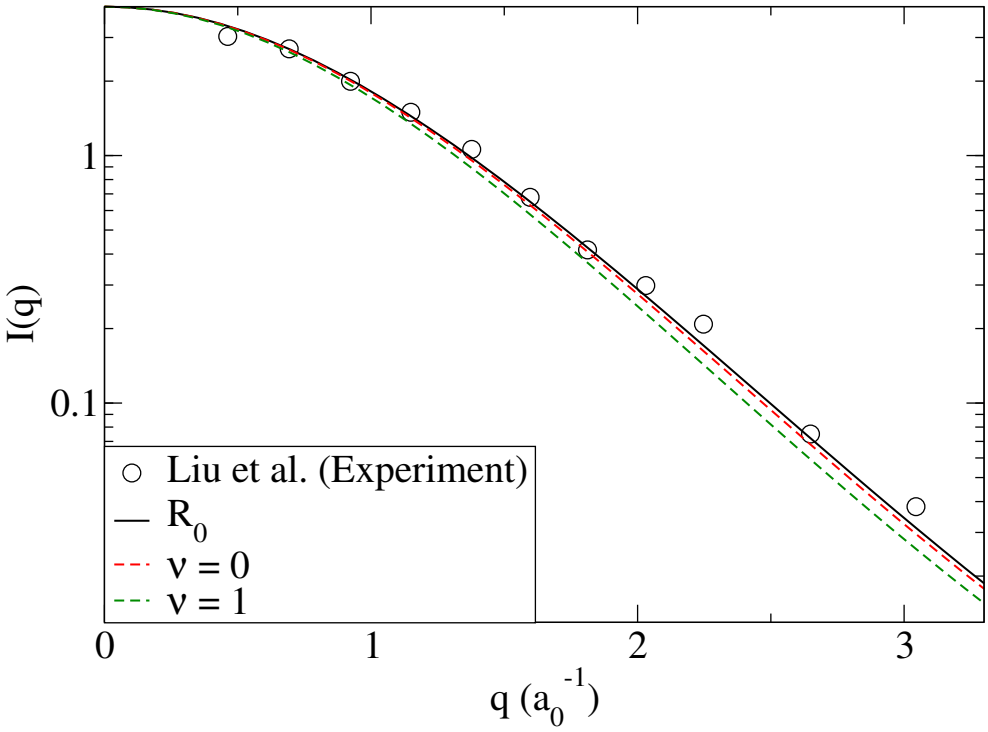


Figure 5.11: H_2 comparison to a gas-phase X-ray diffraction experiment performed at a synchrotron [50]. AIXRD was used at the equilibrium bond-length, R_0, to calculate $I(q)$. The shooting method on the ground state Wolniewicz X $^1\Sigma_g^+$ PEC [134,135] was used to compute the $v = 0$ and $v = 1$ vibrational distributions, and AIXRD was used at each integration step. MCSCF(2,7)/aug-cc-pVTZ was used for geometry optimisation and with AIXRD to calculate each curve.

In a recent experiment by Liu et al. X-ray diffraction at the Taiwan Beamline BL12XU at SPring-8, a third generation synchrotron source, was used to determine the molecular form-factor for H_2 in the gas phase with incident photon energy of about 9889 eV at 8.94 atm and 298 K [50]. A spectrometer was used to isolate the pure elastic signal, at it would otherwise contain inelastic and ionization channels. A Boltzmann distribution can be used to determine the ratio between the number of molecules in the $v = 0$ and the $v = 1$ state, N_0/N_1. At $T = 298$ K, $N_0/N_1 = e^{-(E_0-E_1)/kT} \approx 1.6 \times 10^9$, with $E_v = hc\tilde{\nu}(v + \frac{1}{2})$, and the H_2 MCSCF(2,7)/aug-cc-pVTZ normal mode wavenumber is $\tilde{\nu} = 4382.87$ cm^{-1} (the

experimental reference value is 4401.21 cm^{-1} [137]) and $c = 2.99 \times 10^{10}$ cm s^{-1}. That is, at 298 K the vibrational ground state is extremely dominant.

Comparison between the AIXRD equilibrium bond-length R_0, $v = 0$ and $v = 1$ diffraction curves is shown in Fig. 5.11. There is good agreement between the Liu et al. experimental curve and the R_0 and $v = 0$ curves. The similarity between R_0 and $v = 0$ results shows it is unnecessary to account for the ground state vibrational distribution in this experimental comparison. It would require very high experimental resolution to detect a difference of the magnitude of the R_0 and $v = 0$ theoretical difference. The $v = 1$ curve is plotted despite its extremely low state population. It shows that for non-negligible $v > 0$ population it becomes necessary to consider the vibrational state in the X-ray diffraction. The agreement with experiment validates this method predicting molecular form-factors based on *ab initio* electronic structure calculations.

5.4 Conclusion

The results in this chapter demonstrate that electronic structure gives a strong signature in the elastic X-ray scattering signal, in agreement with the results of Ch. 3 and Ref. [1], with scattering from Rydberg states [66, 117] as an extreme example. The importance of accounting for electronic structure, hence electron density, beyond the independent atom model is validated, especially for electronically excited states, and even vibrationally excited states ($v > 0$).

In addition to electronic states, even vibrational states in a thermal distribution of molecules are distinct, although with lower magnitude differences in this case. Rotational states grant partial alignment and information at the detector becomes dependent on the azimuthal scattering angle, ϕ, although only for X-rays non-parallel to the principle rotation axis of the spherical harmonic rotational distributions. In fact, the most scattering information is available for X-rays exactly perpendicular to this axis. These results show that a pump-probe X-ray diffraction experiment, with sufficient pulse intensity and duration, can theoretically image a chemical reaction pathway including different state changes. More specifically, and because the electronic ground state of H_2 has been imaged via X-ray diffraction [50], gives plausibility for state selective diffractive imaging of H_2. Specific states could be pumped e.g. as in Ref. [73], then imaged e.g. via femtosecond X-ray diffraction at a XFEL source.

The AIXRD method described in Ch. 3 and in Ref. [1] is upgraded in this chapter to take into account vibrational and rotational states, although only for diatomics. It is relatively expensive to accurately compute the effect of these states as triple integration over rotational angles and the vibrational coordinate is necessary. For larger molecules this scales quickly towards unfeasibility. The theoretical method in this chapter is successfully benchmarked against previous theory [132] and experiment [50].

Chapter 6

State selective diffraction measurements: Polyatomics

6.1 Introduction

Continuing from the previous chapter, the effect of vibrations on the elastic X-ray scattering from larger molecules ($N_{\mathrm{at}} > 2$) is studied. This is a demanding problem as the exact vibrational wave function depends on the $3N_{\mathrm{at}} - 6$ dimensional nuclear potential energy surface (PES). It would require a costly mapping of the PES at appropriately large deviations centred at the equilibrium geometry. The aim of this short chapter is to show a simple model of the X-ray diffraction from molecules undergoing motion via their vibrational states. As such, a highly accurate nuclear wave function is not necessary, and the harmonic approximation is reasonable for lowly excited vibrational states, where the nuclear motion only encompasses a local vicinity of the PES, about the equilibrium molecular geometry, which is approximately harmonic-well shaped. X-ray diffraction from small molecules (BF_3, NH_3, and H_2O) is calculated directly from Born-Oppenheimer wave functions, where the electronic part is calculated from *ab initio* electronic structure theory and the nuclear (vibrational) part is calculated by the harmonic approximation. The rotational wave functions are left out in this work, but can be included. They are molecule specific, and as shown in the previous Chapter, for diatomics they are spherical harmonics. Molecules such as BF_3 are so-called symmetric tops, and have corresponding rotational wave functions. The effect of rotational wave functions on polyatomic molecules is specifically looked at in an

upcoming paper [4]. Here, only the (thermal) orientational-average of the molecular form-factor is considered, which is the case for an ensemble of randomly orientated gas-phase molecules probed by X-ray diffraction. The electronic term in the X-ray molecular form-factor is calculated by AIXRD. The convergence of this term with basis set (STO-3G, 6-31G,. . . ,aug-cc-pVQZ) is quantified for BF_3 and NH_3. The X-ray diffraction signals from H_2O in various vibrationally excited states are calculated and compared to the vibrational ground state by percent difference in X-ray diffraction intensities.

The motivation for this chapter is analogous to that of Chapter 5. The preparation of molecules in specific quantum states is increasingly possible [70, 72, 138]. In addition, gas-phase X-ray diffraction has been spurred on by the brilliance of XFEL sources, such as LCLS and other facilities [22–25]. Diffraction from aligned molecules [75, 139–142] has been experimentally successful, will continue, and likely be improved upon in the future.

6.2 Theory

The following theory pertains to the calculation of the X-ray molecular form-factor based on a Born-Oppenheimer wave function, where the nuclear wave function is split into separate rotational and vibrational parts. The electronic part is calculated by AIXRD as per the previous chapters, which is integrated over the vibrational distribution, calculated with the harmonic approximation as shown in the following subsection. The implicit rotational wave functions are omitted here and will be shown in an upcoming paper [4]. The work here assumes fully orientationally-averaged diffraction intensities, as would be physically observable in an X-ray diffraction experiment on randomly orientated molecules in a gas-phase ensemble.

6.2.1 Harmonic vibrations

To determine the nuclear vibrational wave function for a molecule, a convenient starting point is the N_{at}-atom nuclear Hamiltonian, which is defined as,

$$\hat{H}_{\text{nuc}} = -\frac{1}{2}\sum_{i=1}^{N_{\text{at}}} \frac{1}{M_i}\frac{\mathrm{d}^2}{\mathrm{d}\mathbf{R}_i^2} + V(\mathbf{R}_1, \mathbf{R}_2, \ldots, \mathbf{R}_{N_{\text{at}}}), \tag{6.1}$$

where $\{M_i\}$ are the atomic masses within the molecule, $\{\mathbf{R}_i\}$ are the N_{at} nuclear coordinates, and V is the $3N_{\text{at}}$ dimensional potential energy function. It is convenient to use mass-weighted coordinates $q_i = \sqrt{M_i}d_i$ for displacements $d_i = R_i - a_i$, where $\{a_i\}$ is the set of $3N_{\text{at}}$ 1D equilibrium coordinates and the subscript is now $i = 1, 2, \ldots, 3N_{\text{at}}$. The kinetic energy operator is then,

$$\hat{T} = -\frac{1}{2}\sum_{i}^{3N_{\text{at}}} \frac{\partial^2}{\partial q_i^2}, \tag{6.2}$$

and the potential energy expressed as a Taylor expansion about the stationary point is,

$$V = V_0 + \sum_{i}^{3N_{\text{at}}} \Big(\frac{\partial V}{\partial q_i}\Big)_0 q_i + \frac{1}{2}\sum_{i,j}^{3N_{\text{at}}} \Big(\frac{\partial^2 V}{\partial q_i \partial q_j}\Big)_0 q_i q_j + \ldots, \tag{6.3}$$

where V_0 is a constant and for here on can be omitted for brevity. At equilibrium the gradient is zero, $(\partial V/\partial q_i)_0 = 0\ \forall i$, and omitting higher than quadratic terms gives,

$$V \approx \frac{1}{2}\sum_{i,j}^{3N_{\text{at}}} \Big(\frac{\partial^2 V}{\partial q_i \partial q_j}\Big)_0 q_i q_j = \frac{1}{2}\sum_{i,j}^{3N_{\text{at}}} f_{ij} q_i q_j, \tag{6.4}$$

where the $3N_{\text{at}} \times 3N_{\text{at}}$ matrix elements $\{f_{ij}\}$ make up the mass-weighted Hessian matrix,

$$\mathbf{f} = \begin{pmatrix} f_{1,1} & \cdots & f_{1,3N_{\text{at}}} \\ \vdots & \ddots & \vdots \\ f_{3N_{\text{at}},1} & \cdots & f_{3N_{\text{at}},3N_{\text{at}}} \end{pmatrix}. \tag{6.5}$$

Now the Hamiltonian can be rewritten in matrix form as,

$$\mathbf{H}_{\text{nuc}} = \frac{1}{2}\mathbf{p}^{\text{T}}\mathbf{p} + \frac{1}{2}\mathbf{q}^{\text{T}}\mathbf{f}\mathbf{q}, \tag{6.6}$$

where $\mathbf{q}$ and $\mathbf{p}$ are column matrices of length $3N_{\text{at}}$ containing the coordinates $\{q_i\}$ and momenta $\{p_i\}$ respectively. The coordinates can be transformed to become the *normal coordinates* of the molecule,

$$\mathbf{Q} = \mathbf{L}^{-1}\mathbf{q}, \qquad \mathbf{q} = \mathbf{L}\mathbf{Q}, \tag{6.7}$$

$$\mathbf{P} = \mathbf{L}^{\text{T}}\mathbf{p}, \qquad \mathbf{p} = (\mathbf{L}^{\text{T}})^{-1}\mathbf{P}, \tag{6.8}$$

where $\mathbf{L}$ is a real-valued orthogonal transformation matrix, with $\mathbf{L}^{\mathrm{T}}\mathbf{L} = \mathbf{E}$ the identity matrix, and $\mathbf{L}^{\mathrm{T}} = \mathbf{L}^{-1}$. If $\mathbf{L}^{\mathrm{T}}\mathbf{f}\mathbf{L} = \mathbf{\Lambda}$, where the matrix $\mathbf{\Lambda}$ is diagonal and has the values $\{\lambda_i\}$ along its diagonal, then pre-multiplication by $\mathbf{L}$ gives a matrix of eigenvalue equations, $\mathbf{f}\mathbf{L} = \mathbf{L}\mathbf{\Lambda}$. Which can be solved by diagonalising the mass-weighted Hessian, $\mathbf{f}$, to find $\mathbf{L}$; and the Hamiltonian becomes,

$$\hat{H}_{\mathrm{nuc}} = \frac{1}{2}\mathbf{P}^{\mathrm{T}}\mathbf{P} + \frac{1}{2}\mathbf{Q}^{\mathrm{T}}\mathbf{\Lambda}\mathbf{Q} = \frac{1}{2}\sum_{i}^{3N_{\mathrm{at}}}[P_i^2 + \lambda_i Q_i^2]. \tag{6.9}$$

In other words, the transformation $q \to Q$, $p \to P$ gives $3N_{\mathrm{at}}$ uncoupled harmonic oscillators with frequencies $\omega_i = \sqrt{\lambda_i}$, with well-known solutions,

$$E_i = \omega_i(v_i + \frac{1}{2}), \quad v_i = 0, 1, 2, \ldots, \tag{6.10}$$

$$\psi_i(Q_i) = n_i H_{v_i}(\sqrt{\omega_i}Q_i)\exp\Big[-\frac{1}{2}\omega_i Q_i^2\Big], \tag{6.11}$$

where $\{E_i\}$ are the energies of each normal mode, $\{v_i\} \in \mathbb{N}$ are their vibrational quantum numbers, $n_i = (\omega_i/\pi)^{1/4}/\sqrt{2^{v_i}v_i!}$ is the normalisation constant, and H_{v_i} is a Hermite polynomial. It is worth noting that for non-linear molecules there are three translational and three rotational degrees of freedom, which have $\omega_i = 0$. Thus, in general there are $3N_{\mathrm{at}} - 6$ vibrational degrees of freedom or *modes*.

6.2.2 X-ray scattering

The X-ray scattering calculation used in this Chapter involves integration over all $3N_{\mathrm{at}} - 6$ polyatomic vibrational modes, using the vibrational wave functions obtained from the harmonic approximation (Eq. (6.11)). This is an eigenstate averaging, that is, the molecule is in a (vibrational) superposition which scatters in-phase. Thus, to take into account the phase terms, integration is done with the molecular form-factor,

$$\langle f(\mathbf{q};\mathbf{Q})\rangle_{\mathrm{vib}} = \int_{Q_1^{\mathrm{min}}}^{Q_1^{\mathrm{max}}}\int_{Q_2^{\mathrm{min}}}^{Q_2^{\mathrm{max}}}\cdots\int_{Q_{3N_{\mathrm{at}}-6}^{\mathrm{min}}}^{Q_{3N_{\mathrm{at}}-6}^{\mathrm{max}}}\Big|\prod_{i}^{3N_{\mathrm{at}}-6}\psi_{\nu_i}(Q_i)\Big|^2 \tag{6.12}$$
$$f_{\nu_1,\nu_2,\ldots,\nu_{3N_{\mathrm{at}}-6}}(\mathbf{q};Q_1,Q_2,\ldots,Q_{3N_{\mathrm{at}}-6})\mathrm{d}Q_1\mathrm{d}Q_2\ldots\mathrm{d}Q_{3N_{\mathrm{at}}-6}.$$

Then, the diffraction intensity is the absolute square of this, and its rotational-average (thermal averaging) is the gas-phase experimental observable,

$$I(q) = \frac{1}{4\pi} \int_0^{2\pi} \int_0^{\pi} |\langle f(\mathbf{q};\mathbf{Q})\rangle_{\text{vib}}|^2 \sin\theta \mathrm{d}\theta \mathrm{d}\phi, \tag{6.13}$$

where $\mathbf{q} = (q, \theta, \phi)$, $q = |\mathbf{q}| = 4\pi \sin(\theta)/\lambda$, and θ and ϕ are reciprocal space spherical coordinate angles.

6.3 Results and discussion

6.3.1 Basis set convergence (BF_3, NH_3)

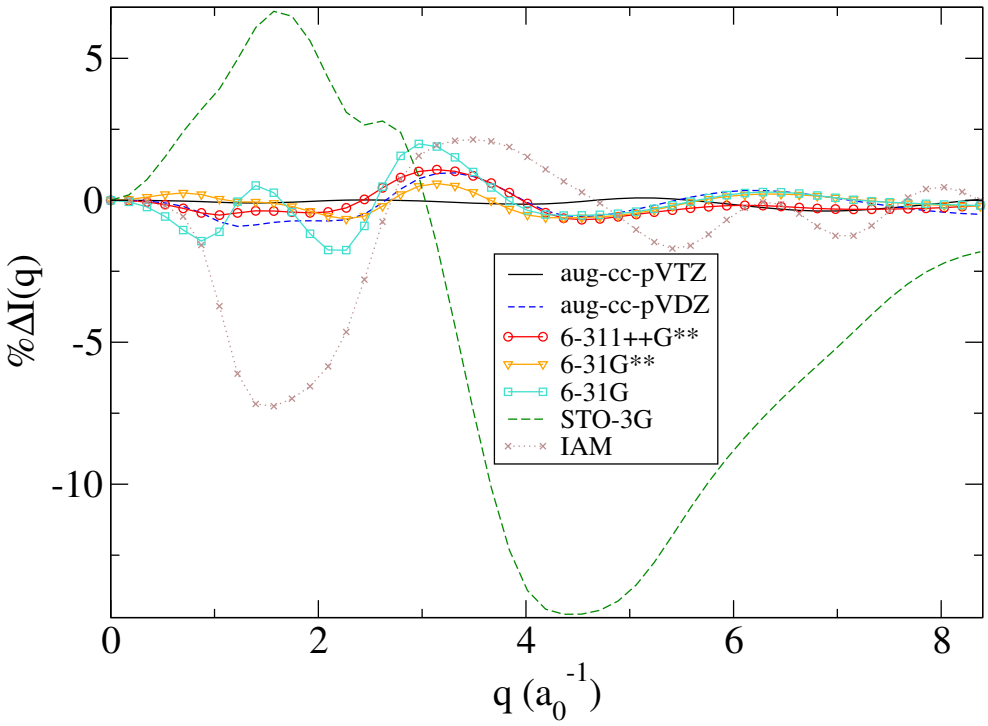


Figure 6.1: BF_3 basis set percent error for AIXRD calculations based on Hartree-Fock (HF) wave functions at the HF/aug-cc-pVQZ optimised geometry, compared to the reference intensity calculated with the aug-cc-pVQZ basis set. The IAM percent error is shown for comparison.

The trigonal planar molecule BF_3 was geometry optimised at the Hartree-Fock (HF)/aug-cc-pVQZ level using Molpro [96], resulting in BF bond-lengths, $R_{\text{BF}} = 2.44432190\ a_0$. AIXRD was used at this geometry using HF with various basis sets. The aug-cc-pVQZ basis set was used for the reference signal I_0 and the

METHOD	Error (%)		$\lvert E - E_{\text{ref}}\rvert$	Primitives		Speed-
	Mean	Max	(E_h)	N_{g}	N_{gp}	up
IAM	1.75	7.26	-	-		4.1×10^4
STO-3G	6.36	14.58	4.694	960	9697	67.6
6-31G	0.54	1.99	0.274	1408	21319	30.7
6-31G**	0.26	0.66	0.164	1792	32105	20.4
6-311++G**	0.39	1.08	0.070	2304	51222	12.8
aug-cc-pVDZ	0.41	0.96	0.133	4288	104984	6.2
aug-cc-pVTZ	0.11	0.38	0.024	8128	258515	2.6
aug-cc-pVQZ[1]	0.0	0.0	0.0	14080	655366	1.0

Table 6.1: Comparison of accuracy and speed for various basis sets for scattering from the Hartree-Fock (HF) electronic ground state of BF_3 at the HF/aug-cc-pVQZ optimised geometry ($R_{\text{BF}} = 2.44432190\ a_0$). The mean and maximum errors shown correspond to the errors in Fig. 6.1, with the scattering from HF/aug-cc-pVQZ taken as reference, see Eq. (5.5). The gap $|E - E_{\text{ref}}|$ between calculated energies and the reference value is used to show *ab-initio* convergence. The speed of the scattering calculations scales linearly with the number of non-zero unique Gaussian products per molecular orbital, N_{gp}, with the number of Gaussian primitives given by N_{g}. In this case, $N_{\text{gp}} \approx N_{\text{g}}^{1.38}$. The IAM calculation uses tabulated atomic form-factors [62].

percent error, $\%\Delta I(q) = 100(I_{\text{method}} - I_0)/I_0$, curves are shown in Fig. 6.1. Table 6.1 shows the mean and maximum absolute percent errors for each basis, the energy gaps $|E - E_{\text{ref}}|$ between each basis set and the reference, the number of primitive Gaussian functions per calculation N_{g} and the number of non-zero unique Gaussian products per MO N_{gp} (AIXRD scales with N_{gp}), and a speed up factor relative to the reference calculation. Somewhat surprisingly, IAM isn't the furthest from the reference in this case. The minimal basis set (STO-3G) calculation is however, with mean error $\langle|\%\Delta I(q)|\rangle = 6.36\%$ and maximum error 14.58%, while the IAM only has $\langle|\%\Delta I(q)|\rangle = 1.75\%$ and maximum error is 7.26%. For this reason, it may be beneficial to simply use IAM for a molecule with a low valence to core electron ratio, such as BF_3, as it can be computed $\sim10^4$ times faster than AIXRD while maintaining a reasonably low $\langle|\%\Delta I(q)|\rangle$. This analysis of BF_3 is a benchmark to assess the quality of each basis set for use in AIXRD for further molecules. The aug-cc-pVTZ basis set is closest with $\langle|\%\Delta I(q)|\rangle = 0.11\%$ and has the lowest energy gap, $|E - E_{\text{ref}}| = 0.024$ Hartree. This isn't surprising as it is the second largest basis set after the reference basis set with about $0.6N_g^{\text{ref}}$, where N_g^{ref} is the number of primitive Gaussians in the

reference calculation. Oddly, it is closely followed by the relatively small basis set 6-31G** which has about $0.1N_g^{\text{ref}}$, and its AIXRD calculation is 20.4 times faster. In fact, all the basis sets have $\langle|\%\Delta I(q)|\rangle < 0.6\%$ except STO-3G, showing that unless supreme accuracy is desired, any medium sized basis set is a reasonable choice. The 6-31G** appears to lie in a sweet spot of low percent error and large computational efficiency, but the 6-31G basis set could also be used for an even larger speed-up.

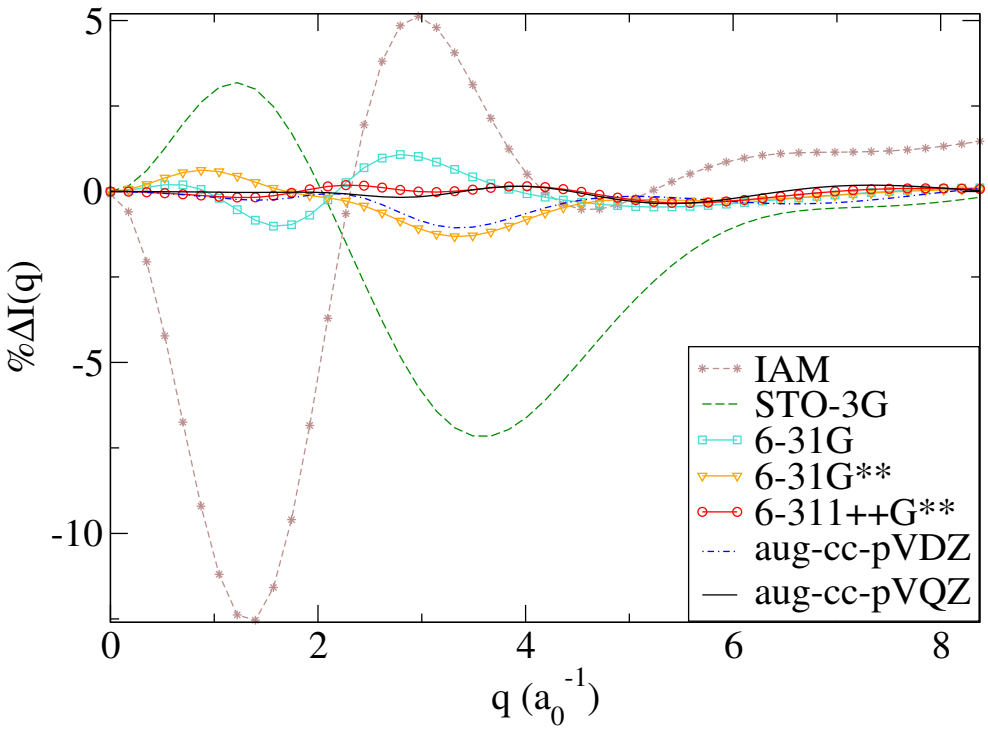


Figure 6.2: NH_3 basis set percent error for AIXRD calculations based on Hartree-Fock (HF) wave functions at the HF/aug-cc-pVQZ optimised geometry, compared to the reference intensity calculated with the aug-cc-pVQZ basis set. The IAM percent error is shown for comparison.

The IAM method works quite well for molecules containing larger Z atoms because they have a low valence/total election ratio. The use of a more accurate method which takes into account the redistribution of valence electrons due to bonding (such as AIXRD) is less needed in these cases. In addition, STO-3G is a poor choice for modelling the Slater-function asymptotic form of the wave function near the nuclei, which is important for such molecules. To test this hypothesis, the analogous calculations as were done for BF_3 were done on a small molecule containing hydrogen atoms, specifically NH_3. Fig. 6.2 show $\%\Delta I(q)$ for AIXRD using HF wave functions with various basis sets. Like before, the reference basis set is aug-cc-pVQZ. Table 6.2 is the equivalent to Table 6.1 but for the NH_3 results. This time, the IAM method is the furthest from the reference with

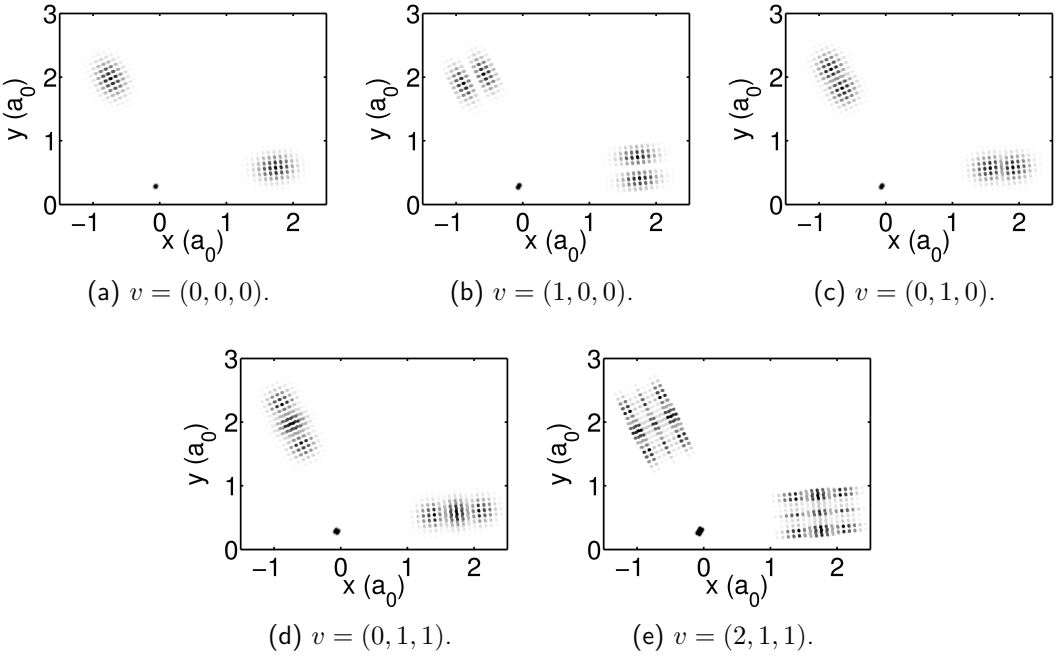
METHOD	Error (%)		$\lvert E - E_{\text{ref}}\rvert$	Primitives		Speed-
	Mean	Max	(10^{-1} E_h)	N_{g}	N_{gp}	up
IAM	2.94	12.54	-	-		1.6×10^4
STO-3G	2.54	7.16	7.729	120	1500	173.3
6-31G	0.35	1.08	0.621	170	2975	87.4
6-31G**	0.38	1.32	0.292	245	6125	42.5
6-311++G**	0.12	0.34	0.098	315	10080	25.8
aug-cc-pVDZ	0.31	1.06	0.190	560	18705	13.9
aug-cc-pVTZ	0.11	0.36	0.037	1190	74390	3.5
aug-cc-pVQZ[2]	0.0	0.0	0.0	2225	260015	1.0

Table 6.2: Comparison of accuracy and speed for various basis sets for scattering from the Hartree-Fock (HF) electronic ground state of NH_3 at the HF/aug-cc-pVQZ optimised geometry ($R_{\text{NH}} = 1.88579802\ a_0$). The mean and maximum errors shown correspond to the errors in Fig. 6.2, with the scattering from HF/aug-cc-pVQZ taken as reference, see Eq. (5.5). The gap $|E - E_{\text{ref}}|$ between calculated energies and the reference value is used to show *ab-initio* convergence. The speed of the scattering calculations scales linearly with the number of non-zero unique Gaussian products per molecular orbital, N_{gp}, with the number of Gaussian primitives given by N_{g}. In this case, $N_{\text{gp}} \approx N_{\text{g}}^{1.58}$. The IAM calculation uses tabulated atomic form-factors [62].

$\langle|\%\Delta I(q)|\rangle = 2.94\%$, closely followed by STO-3G with $\langle|\%\Delta I(q)|\rangle = 2.54\%$. All other basis sets have $\langle|\%\Delta I(q)|\rangle \leq 0.38\%$. These results agree with the hypothesis, and based on the near necessity to account for the valence electrons in H_2, it seems that for calculating the X-ray molecular form-factor for small molecules containing light atoms (low Z) such as hydrogen the IAM is a poor choice and another method such as AIXRD is increasingly useful. Also STO-3G is a poor choice of basis set in general as there is a leap of improvement in percent error going from STO-3G to 6-31G, then a convergence of error mostly inline with the size of basis set (i.e. number of primitive functions N_g).

6.3.2 Triatomics (H_2O)

Finally, to examine the effect of nuclear vibrations, H_2O was chosen, as it is a small molecule containing hydrogens, where a method such as AIXRD is more valid. The other main reason it was chosen is that it is a triatomic, and therefore only has three vibrational modes to integrate over via Eq. (6.12). Larger



(a) $v = (0, 0, 0)$. (b) $v = (1, 0, 0)$. (c) $v = (0, 1, 0)$.

(d) $v = (0, 1, 1)$. (e) $v = (2, 1, 1)$.

Figure 6.3: H_2O atom positions for various vibrational states with $N_Q^3 = 9^3$ integration steps.

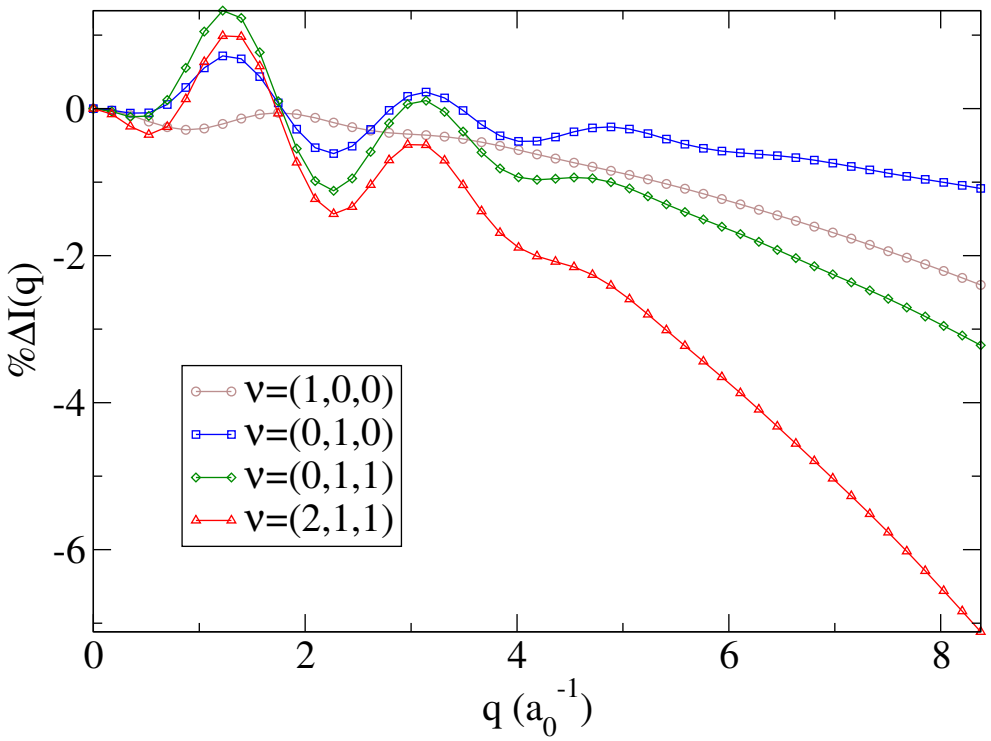


Figure 6.4: H_2O rotationally-averaged percent difference signals for various vibrationally excited states, with the $v = (0, 0, 0)$ curve as reference signal. AIXRD with MCSCF(8,6)/6-31G was used for the X-ray scattering, and the harmonic approximation was used to calculate the vibrational distributions.

molecules become unfeasible computationally. In this case, $N_Q = 9$ integration steps per vibrational mode were used, giving $N_Q^3 = 729$ total integration steps. For a HF/6-31G calculation on the similarly sized (it also has ten electrons) NH_3 molecule, the number of unique Gaussian products per MO is $N_{\text{gp}} = 2975$. There are $N_{\text{el}}/2 = 5$ closed shell MOs, therefore 595 Gaussian products per MO. In this case, a MSCSF(8,6)/6-31G calculation was used for H_2O, with four electrons in closed shells, and the remaining six in the eight orbital active space. This gives ten total MOs, thus $N_{\text{gp}} = 5950$. In other words, the active space in the MCSCF wave function up-scales the AIXRD calculation by a factor of 2 compared to the corresponding HF wave function (in this particular case). A MSCSF(8,6)/6-31G geometry optimisation followed by a frequencies calculation for H_2O using Molpro [96] gives normal mode wavenumbers of 1702.9, 3779.49, 3897.56 cm^{-1}. Then, the normal mode wavenumbers were used in the vibrational wave function Eq. (6.11), and Eq. (6.12) was solved using AIXRD with MSCSF(8,6)/6-31G at each of the N_Q^3 integration steps. Each step takes about 5 seconds on a single 3.1 GHz core. The nuclear coordinates were saved and plotted in the plane of the HOH angle for various vibrational states, they are shown in Fig. 6.3. Note that the nuclear motions are almost entirely the hydrogen atoms, and the oxygen atom has negligible vibrational movement. Fig. 6.4 shows the percent difference X-ray diffraction intensities for the various vibrational states, $I(q) = 100(I_{\text{method}} - I_0)/I_0$, with the ground vibrational state, $v = (0, 0, 0)$, as the reference calculation. Similar to H_2 in the previous chapter, the higher the vibrational state the further the X-ray diffraction curve diverges from the ground state curve, especially at high values of q.

6.4 Conclusion

An attempt was made to continue from diatomic calculations shown in Chapter 5 to polyatomic molecules. However, using the harmonic approximation and performing the full integration over all $3N_{\text{at}} - 6$ vibrational modes scales extremely quickly towards computational unfeasibility. Even with $N_Q = 9$, which is as small as possible without severe degradation of the quality of the numerical integration, there are $N_Q^{3N_{\text{at}}-6}$ total integration steps with each step taking about 5 seconds for H_2O. This can be reduced by choosing a smaller basis set or only using HF theory with AIXRD, but these reductions are only linear, and do not counteract the exponential scaling with N_{at}. Similar results have been achieved as in

previous Chapters, such as more validation that the IAM is a poor choice for small molecules containing hydrogen atoms, and in these cases a method such as AIXRD has greater validity. In addition, just like H_2, increasing the vibrational excitation of H_2O causes the X-ray diffraction difference signal from the ground vibrational state to diverge. This is an intuitive and expected result however. An upcoming paper addresses much more valid experimental examples, and includes calculations of the X-ray scattering from rotational wave packets for polyatomic molecules [4].

Chapter 7

Conclusion

This thesis contains a description of the *ab initio* X-ray diffraction (AIXRD) method for calculating the elastic X-ray scattering factor based on *ab initio* electronic structure theory [1]. It requires that the wave functions are expressed by Gaussian basis sets, such as Pople-type (6-31G, ..., 6-311++G**), or Dunning-type (VDZ, ..., V6Z). The wave function can be expressed in a single or multiconfigurational form, and calculations based on both Hartree-Fock (HF) and multiconfigurational self-consistent field wave functions are included. The proposed method goes beyond the conventional independent atom model (IAM) because it takes into account the redistribution of valence electrons due to bonding, and it can calculate X-ray diffraction from electronically excited states.

This approach was expanded for use with large molecules by introducing a fragment-based method, loosely based on fragment-molecular orbital theory [109]. The molecule, for instance a protein, was divided into fragments, then HF with the 6-31G basis set was used on each fragment, followed by AIXRD. A good compromise of accuracy and efficiency was achieved by using 6-31G, while STO-3G was unsurprisingly too small a basis set to produce reliable results. A sum of isolated fragment scattering factors gave an approximate total molecular scattering factor, analogous to the IAM method. A pairwise summation of fragment and dimer contributions allowed fragment-fragment interactions to be taken into account. Significant improvements of 4-8% compared to the IAM were achieved for lysozyme, even after temporal-averaging to allow for slight variation of molecular geometry, as would be observed in serial crystallography [105]. This type of approach has potential for use in structural refinement algorithms because it gives a more physically accurate structure factor compared to the IAM [143].

Finally, state-specific scattering from molecules was examined. Calculations of elastic X-ray diffraction from H_2 in specific quantum states [2] were motivated by experimental progress in state selection of various molecules in the gas-phase such as H_2 and HD [73, 83], N_2 [71], and other polyatomic molecules [72]; as well as gas-phase diffraction experiments of oriented and aligned molecules [75, 81, 82] and indirect detection of metastable electronically excited states via changes in geometry [67,68]. A comparison is also made to an experimental X-ray scattering measurement of the rotationally-averaged diffraction intensity of H_2 in its ground state [50].

7.1 Outlook

Continuation of the work described in this thesis falls into three categories. Firstly, focus could be applied to modelling the elastic X-ray diffraction signal during a photochemical reaction, involving quantum dynamical treatment of the molecular motion and electronic state changes. By using the multiconfigurational Ehrenfest method [144], for example, followed by AIXRD calculations based on the same MCSCF wave functions as obtained every n timesteps, would give a time-dependent and time-evolving electronic state dependent molecular form-factor. Expanding further upon this would involve temporal integration over the X-ray pulse duration, as would occur in experiments. Progress has been made in this direction but only the IAM method was used to calculate the time-dependent scattering [145]. It would be interesting to improve on this by using AIXRD, and to compare to a gas-phase time-resolved X-ray diffraction experiment at an XFEL source, such as in Ref. [6], especially if the photochemical reaction involved substantial electronic state changes which could be visible in the diffraction signal.

Secondly, improvements could be made to the fragment-based AIXRD method described in Ch. 4. This is relevant to protein structure determination, and the future possibility of single-molecule X-ray diffraction experiments (the natural progression from the imaging of a single mimivirus particle at an XFEL source [36]). In fact, improvements over the conventional method could be welcomed by scientists for usage alongside experiments. This is because there are troublesome sources of error in structure determination [143], followed with refinement by comparison to an IAM structure factor. The improvements of the fragment-

based AIXRD method are promising in this regard. The method outlined in Ch. 4 has room for improvement however, as the full wave function obtained from a Fragment Molecular Orbital (FMO) calculation could be used rather than the isolated fragment and dimer summation method described.

Finally, AIXRD is ripe for combination with a wide variety of electronic structure methods. As the only limitation is for a Gaussian basis set to be used, there is no intrinsic reason why it could not be combined with orbital-based density function theory (DFT) [93, 146], and full configuration interaction, for two examples.

Bibliography

[1] T. Northey, N. Zotev, and A. Kirrander. *Ab Initio* calculation of molecular diffraction. *Journal of Chemical Theory and Computation*, 10(11):4911–4920, 2014.

[2] T. Northey and A. Kirrander. Fragment-based *ab initio* X-ray diffraction from biomolecules. *(manuscript)*, 2017.

[3] T. Northey, A. M. Carrascosa, S. Schäfer, and A. Kirrander. Elastic X-ray scattering from state-selected molecules. *The Journal of Chemical Physics*, 145(15):154304, 2016.

[4] A. Moreno Carrascosa, T. Northey, and A. Kirrander. Imaging rotations and vibrations in polyatomic molecules with X-ray scattering. *(manuscript)*, 2016.

[5] M. P. Minitti, J. M. Budarz, A. Kirrander, J. S. Robinson, D. Ratner, T. J. Lane, D. Zhu, J. M. Glownia, M. Kozina, H. T. Lemke, M. Sikorski, Y. Feng, S. Nelson, K. Saita, B. Stankus, T. Northey, J. B. Hastings, and P. M. Weber. Imaging Molecular Motion: Femtosecond X-Ray Scattering of an Electrocyclic Chemical Reaction. *Physical Review Letters*, 114:255501, Jun 2015.

[6] M. P. Minitti, J. M. Budarz, A. Kirrander, J. Robinson, T. J. Lane, D. Ratner, K. Saita, T. Northey, B. Stankus, V. Cofer-Shabica, J. Hastings, and P. M. Weber. Toward structural femtosecond chemical dynamics: imaging chemistry in space and time. *Faraday Discussions*, 171:81–91, 2014.

[7] W. L. Bragg. The structure of some crystals as indicated by their diffraction of X-rays. In *Proceedings of the Royal Society of London A: Mathematical,*

Physical and Engineering Sciences, volume 89, pages 248–277. The Royal Society, 1913.

[8] M. von Laue. Concerning the detection of X-ray interferences. *Nobel Lecture*, 1915.

[9] R. L. Weber. *Pioneers of science: Nobel Prize winners in Physics*. Taylor & Francis, 1988.

[10] T. L. Sourkes and L. G. Stevenson. *Nobel Prize winners in Medicine and Physiology, 1901-1965*. Number 45. Abelard-Schuman, 1967.

[11] J. K. Laylin. *Nobel Laureates in Chemistry, 1901-1992*, volume 1. Chemical Heritage Foundation, 1993.

[12] A. H. Zewail, F. C. De Schryver, S. De Feyter, and G. Schweitzer. *Femtochemistry: with the Nobel lecture of A. Zewail*. Wiley-VCH, 2001.

[13] Z. Huang et al. Brightness and coherence of synchrotron radiation and FELs. *MOYCB101, Proceedings of IPAC2013, Shanghai, China*, 2013.

[14] E. F. Garman. Developments in X-ray crystallographic structure determination of biological macromolecules. *Science*, 343(6175):1102–1108, 2014.

[15] J. M. J. Madey. Stimulated emission of bremsstrahlung in a periodic magnetic field. *Journal of Applied Physics*, 42(5):1906–1913, 1971.

[16] Y. S. Derbenev, A. Kondratenko, and E. Saldin. On the possibility of using a free electron laser for polarization of electrons in storage rings. *Nuclear Instruments and Methods in Physics Research*, 193(3):415–421, 1982.

[17] J. B. Murphy and C. Pellegrini. Free electron lasers for the XUV spectral region. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 237(1-2):159–167, 1985.

[18] C. Milne, T. Penfold, and M. Chergui. Recent experimental and theoretical developments in time-resolved X-ray spectroscopies. *Coordination Chemistry Reviews*, 277:44–68, 2014.

[19] V. Ayvazyan, N. Baboi, J. Bähr, V. Balandin, B. Beutner, A. Brandt, I. Bohnet, A. Bolzmann, R. Brinkmann, O. Brovko, et al. First operation

of a free-electron laser generating gw power radiation at 32 nm wavelength. *The European Physical Journal D-Atomic, Molecular, Optical and Plasma Physics*, 37(2):297–303, 2006.

[20] V. Lyamayev, Y. Ovcharenko, R. Katzy, M. Devetta, L. Bruder, A. LaForge, M. Mudrich, U. Person, F. Stienkemeier, M. Krikunova, et al. A modular end-station for atomic, molecular, and cluster science at the low density matter beamline of FERMI@ Elettra. *Journal of Physics B: Atomic, Molecular and Optical Physics*, 46(16):164007, 2013.

[21] E. Springate, C. Froud, I. Turcu, S. Spurdle, D. Wolff, S. Hook, B. Landowski, J. Underwood, A. Cavalleri, S. Dhesi, et al. Artemis: Synchronised XUV and laser sources for ultrafast time-resolved science. *CLF Annual Report*, 9:221–224, 2008.

[22] P. Emma, R. Akre, J. Arthur, R. Bionta, C. Bostedt, J. Bozek, A. Brachmann, P. Bucksbaum, R. Coffee, F. J. Decker, Y. Ding, D. Dowell, S. Edstrom, A. Fisher, J. Frisch, S. Gilevich, J. Hastings, G. Hays, HeringPh, Z. Huang, R. Iverson, H. Loos, M. Messerschmidt, A. Miahnahri, S. Moeller, H. D. Nuhn, G. Pile, D. Ratner, J. Rzepiela, D. Schultz, T. Smith, P. Stefan, H. Tompkins, J. Turner, J. Welch, W. White, J. Wu, G. Yocky, and J. Galayda. First lasing and operation of an angstrom-wavelength free-electron laser. *Nature Photonics*, 4(9):641–647, 2010.

[23] M. Yabashi, H. Tanaka, T. Tanaka, H. Tomizawa, T. Togashi, M. Nagasono, T. Ishikawa, J. R. Harries, Y. Hikosaka, A. Hishikawa, K. Nagaya, N. Saito, E. Shigemasa, K. Yamanouchi, and K. Ueda. Compact XFEL and AMO sciences: SACLA and SCSS. *Journal of Physics B: Atomic, Molecular and Optical Physics*, 46(16):164001, 2013.

[24] J. Feldhaus, M. Krikunova, M. Meyer, T. Möller, R. Moshammer, A. Rudenko, T. Tschentscher, and J. Ullrich. AMO science at the FLASH and European XFEL free-electron laser facilities. *Journal of Physics B: Atomic, Molecular and Optical Physics*, 46(16):164002, 2013.

[25] E.-S. Kim and M. Yoon. Beam dynamics in a 10-GeV linear accelerator for the X-Ray free electron laser at PAL. *Nuclear Science, IEEE Transactions on*, 56(6):3597–3606, 2009.

[26] R. Ganter. SwissFEL-Conceptual design report. Technical report, Paul Scherrer Institute (PSI), 2010.

[27] C. Bostedt, J. Bozek, P. Bucksbaum, R. Coffee, J. Hastings, Z. Huang, R. Lee, S. Schorb, J. Corlett, P. Denes, et al. Ultra-fast and ultra-intense x-ray sciences: first results from the linac coherent light source free-electron laser. *Journal of Physics B: Atomic, Molecular and Optical Physics*, 46(16):164003, 2013.

[28] N. Berrah and P. H. Bucksbaum. The ultimate X-ray machine. *Scientific American*, 310(1):64–71, 2014.

[29] J. Als-Nielsen and D. McMorrow. *Elements of modern X-ray physics*. John Wiley & Sons, 2011.

[30] R. Neutze, R. Wouts, S. Techert, J. Davidsson, M. Kocsis, A. Kirrander, F. Schotte, and M. Wulff. Visualizing photochemical dynamics in solution through picosecond X-ray scattering. *Physical Review Letters*, 87(19):195508, 2001.

[31] E. Curwood, H. Quiney, and K. Nugent. Determining electronic damage to biomolecular structures in x-ray free-electron-laser imaging experiments. *Physical Review A*, 87(5):053407, 2013.

[32] O. Y. Gorobtsov, U. Lorenz, N. Kabachnik, and I. Vartanyants. Theoretical study of electronic damage in single-particle imaging experiments at x-ray free-electron lasers for pulse durations from 0.1 to 10 fs. *Physical Review E*, 91(6):062712, 2015.

[33] L. Redecke, K. Nass, D. P. DePonte, T. A. White, D. Rehders, A. Barty, F. Stellato, M. Liang, T. R. Barends, S. Boutet, et al. Natively inhibited Trypanosoma Brucei Cathepsin B structure determined by using an X-ray laser. *Science*, 339(6116):227–230, 2013.

[34] J. R. Helliwell. How to solve protein structures with an X-ray laser. *Science*, 339(6116):146–147, 2013.

[35] T. R. M. Barends, L. Foucar, S. Botha, R. B. Doak, R. L. Shoeman, K. Nass, J. E. Koglin, G. J. Williams, S. Boutet, M. Messerschmidt, et al. De novo protein crystal structure determination from X-ray free-electron laser data. *Nature*, 505(7482):244–247, 2014.

[36] M. M. Seibert, T. Ekeberg, F. R. Maia, M. Svenda, J. Andreasson, O. Jönsson, D. Odić, B. Iwan, A. Rocker, D. Westphal, et al. Single mimivirus particles intercepted and imaged with an X-ray laser. *Nature*, 470(7332):78–81, 2011.

[37] P. Cohen. Protein kinasesthe major drug targets of the twenty-first century? *Nature Reviews Drug Discovery*, 1(4):309–315, 2002.

[38] P. Debye. X-ray dispersal. *Annalen Der Physik*, 46(6):809–823, 1915.

[39] I. Hargittai and M. Hargittai. *Stereochemical Applications of Gas-Phase Electron Diffraction: Part A The Electron Diffraction Technique*. VCH, First edition, 1988.

[40] P. Debye, L. Bewilogua, and F. Ehrhardt. Zerstreuung von Röntgenstrahlen an einzelnen Molekülen (vorläufige Mitteilung). *Physikalische Zeitschrift*, 30:84, 1929.

[41] M. H. Pirenne. *The Diffraction of X-Rays and Electrons by Free Molecules*. Cambridge University Press: London, 1946.

[42] C. S. Barrett. The Scattering of X-Rays from Gases. *Physical Review*, 32(1):22–29, 1928.

[43] E. O. Wollan. Scattering of X-Rays from Gases. *Physical Review*, 37(8):862, 1931.

[44] D. R. Chipman and L. D. Jennings. Measurement of the Atomic Scattering Factor of Ne, Ar, Kr, and Xe. *Physical Review*, 132(2):728, 1963.

[45] H. Mark and R. Wierl. Die Ermittlung von Molekülstrukturen durch Beugung von Elektronen an einem Dampfstrahl. *Zeitschrift für Elektrochemie und Angewandte Physikalische Chemie*, 36(9):675–676, 1930.

[46] A. H. Zewail. Four-Dimensional Electron Microscopy. *Science*, 328(5975):187, 2010.

[47] G. Sciaini and R. J. D. Miller. Femtosecond electron diffraction: heralding the era of atomically resolved dynamics. *Reports on Progress in Physics*, 74:096101, 2011.

[48] J. Küpper, S. Stern, L. Holmegaard, F. Filsinger, A. Rouzée, A. Rudenko, P. Johnsson, A. V. Martin, M. Adolph, A. Aquila, et al. X-ray diffraction from isolated and strongly aligned gas-phase molecules with a free-electron laser. *Physical Review Letters*, 112(8):083002, 2014.

[49] J. M. Budarz, M. P. Minitti, D. V. Cofer-Shabica, B. Stankus, A. Kirrander, J. B. Hastings, and P. M. Weber. Observation of femtosecond molecular dynamics via pump-probe gas phase X-ray scattering. *Journal of Physics B: Atomic, Molecular and Optical Physics*, 49(3):034001, 2016.

[50] Y.-W. Liu, X.-X. Mei, X. Kang, K. Yang, W.-Q. Xu, Y.-G. Peng, N. Hiraoka, K.-D. Tsuei, P.-F. Zhang, and L.-F. Zhu. Determination of the electronic structure of atoms and molecules in the ground state: Measurement of molecular hydrogen by high-resolution X-ray scattering. *Physical Review A*, 89(1):014502, 2014.

[51] B. Xie, L. Zhu, K. Yang, B. Zhou, N. Hiraoka, Y. Cai, Y. Yao, C. Wu, E. Wang, and D. Feng. Inelastic X-ray scattering study of the state-resolved differential cross section of Compton excitations in Helium atoms. *Physical Review A*, 82(3):032501, 2010.

[52] L. Zhu, L. Wang, B. Xie, K. Yang, N. Hiraoka, Y. Cai, and D. Feng. Inelastic x-ray scattering study on the single excitations of helium. *Journal of Physics B: Atomic, Molecular and Optical Physics*, 44(2):025203, 2011.

[53] Y.-G. Peng, X. Kang, K. Yang, X.-L. Zhao, Y.-W. Liu, X.-X. Mei, W.-Q. Xu, N. Hiraoka, K.-D. Tsuei, and L.-F. Zhu. Squared form factors of vibronic excitations in 12-13.3 eV of nitrogen studied by high-resolution inelastic X-ray scattering. *Physical Review A*, 89(3):032512, 2014.

[54] S. Techert, F. Schotte, and M. Wulff. Picosecond X-ray diffraction probed transient structural changes in organic solids. *Physical Review Letters*, 86(10):2030, 2001.

[55] H. Ihee, M. Lorenc, T. K. Kim, Q. Y. Kong, M. Cammarata, J. H. Lee, S. Bratos, and M. Wulff. Ultrafast X-ray diffraction of transient molecular structures in solution. *Science*, 309(5738):1223–1227, 2005.

[56] Q. Kong, J. H. Lee, M. Lo Russo, T. K. Kim, M. Lorenc, M. Cammarata, S. Bratos, T. Buslaps, V. Honkimaki, H. Ihee, and M. Wulff. Photolysis

of Br_2 in CCl_4 studied by time-resolved X-ray scattering. *Acta Crystallographica Section A: Foundations of Crystallography*, 66(2):252–260, 2010.

[57] J. H. Lee, M. Wulff, S. Bratos, J. Petersen, L. Guerin, J.-C. Leicknam, M. Cammarata, Q. Kong, J. Kim, K. B. Mller, and H. Ihee. Filming the birth of molecules and accompanying solvent rearrangement. *Journal of the American Chemical Society*, 135(8):3255–3261, 2013.

[58] K. H. Kim, J. Kim, K. Y. Oang, J. H. Lee, D. Grolimund, C. J. Milne, T. J. Penfold, S. L. Johnson, A. Galler, T. W. Kim, et al. Identifying the major intermediate species by combining time-resolved x-ray solution scattering and x-ray absorption spectroscopy. *Physical Chemistry Chemical Physics*, 17(36):23298–23302, 2015.

[59] M. Levantino, G. Schirò, H. T. Lemke, G. Cottone, J. M. Glownia, D. Zhu, M. Chollet, H. Ihee, A. Cupane, and M. Cammarata. Ultrafast myoglobin structural dynamics observed with an X-ray free-electron laser. *Nature Communications*, 6(6772), 2015.

[60] K. H. Kim, J. G. Kim, S. Nozawa, T. Sato, K. Y. Oang, T. W. Kim, H. Ki, J. Jo, S. Park, C. Song, et al. Direct observation of bond formation in solution with femtosecond X-ray scattering. *Nature*, 518(7539):385–389, 2015.

[61] P. Emma, J. Frisch, Z. Huang, A. Marinelli, T. Maxwell, H. Loos, Y. Nosochkov, T. Raubenheimer, J. Welch, L. Wang, et al. Linear accelerator design for the LCLS-II FEL facility. In *Proceedings of 36th Int. Free-Electron Laser Conf., Basel, Switzerland*, 2014.

[62] P. J. Brown, A. G. Fox, E. N. Maslen, M. A. O'Keefe, and B. T. M. Willis. Intensity of diffracted intensities. In *International Tables for Crystallography Volume C: Mathematical, physical and chemical tables*, pages 554–595. Springer, 2006.

[63] D. T. Cromer and J. B. Mann. X-ray scattering factors computed from numerical Hartree-Fock wave functions. *Acta Crystallographica Section A: Crystal Physics, Diffraction, Theoretical and General Crystallography*, 24(2):321–324, 1968.

[64] C. T. Chantler. Detailed tabulation of atomic form factors, photoelectric absorption and scattering cross section, and mass attenuation coefficients in the vicinity of absorption edges in the soft X-ray (Z=30–36, Z=60–89, E=0.1 keV–10 keV), addressing convergence issues of earlier work. *Journal of Physical and Chemical Reference Data*, 29(4):597–1056, 2000.

[65] M. Ben-Nun, T. J. Martínez, P. M. Weber, and K. R. Wilson. Direct imaging of excited electronic states using diffraction techniques: theoretical considerations. *Chemical Physics Letters*, 262(3):405–414, 1996.

[66] A. Kirrander. X-ray diffraction assisted spectroscopy of Rydberg states. *The Journal of Chemical Physics*, 137(15):154310, 2012.

[67] M. R. Pressprich, M. A. White, and P. Coppens. Single-crystal X-ray analysis of an electronic excited state: the structure determination of a metastable state of sodium nitroprusside. *Journal of the American Chemical Society*, 115(14):6444–6445, 1993.

[68] M. R. Pressprich, M. A. White, Y. Vekhter, and P. Coppens. Analysis of a metastable electronic excited state of sodium nitroprusside by x-ray crystallography. *Journal of the American Chemical Society*, 116(12):5233–5238, 1994.

[69] C. Vallance. Generation, Characterisation, and Applications of Atomic and Molecular Alignment and Orientation. *Physical Chemistry Chemical Physics*, 13(32):14427, 2011.

[70] H. Stapelfeldt and T. Seideman. Colloquium: Aligning molecules with strong laser pulses. *Reviews of Modern Physics*, 75(2):543, 2003.

[71] G. O. Sitz and R. L. Farrow. Preparation and decay of alignment in N_2 ($v = 1$). *The Journal of Chemical Physics*, 101(6):4682–4687, 1994.

[72] A. D. Rudert, J. Martin, W.-B. Gao, J. B. Halpern, and H. Zacharias. Collisional effects on angular momentum orientation in acetylene $\tilde{X}\ ^1\Sigma_g^+(v_2'' = 1, j'')$. I. Preparation, detection and conservation in single collisions. *Journal of Chemical Physics*, 111(21):9549–9559, 1999.

[73] N. C. M. Bartlett, D. J. Miller, R. N. Zare, D. Sofikitis, T. P. Rakitzis, and A. J. Alexander. Preparation of oriented and aligned H_2 and HD by

stimulated Raman pumping. *The Journal of Chemical Physics*, 129:084312, 2008.

[74] N. Mukherjee and R. N. Zare. Stark-induced adiabatic Raman passage for preparing polarized molecules. *The Journal of Chemical Physics*, 135(2), 2011.

[75] C. J. Hensley, J. Yang, and M. Centurion. Imaging of isolated molecules with ultrafast electron pulses. *Physical Review Letters*, 109(13):133202, 2012.

[76] K. B. Møller and N. E. Henriksen. Time-resolved x-ray diffraction: The dynamics of the chemical bond. In D. M. P. Mingos, P. Day, and J. P. Dahl, editors, *Molecular Electronic Structures of Transition Metal Complexes I*, volume 142 of *Structure and Bonding*, pages 185–211. Springer Berlin Heidelberg, 2012.

[77] J. D. Geiser and P. M. Weber. Pump-probe diffraction imaging of vibrational wave functions. *The Journal of Chemical Physics*, 108(19):8004–8011, 1998.

[78] D. Kohl and E. Shipsey. Elastic electron scattering from state-selected molecules. *Zeitschrift für Physik D Atoms, Molecules and Clusters*, 24(1):39–45, 1992.

[79] S. Ryu, R. M. Stratt, and P. M. Weber. Diffraction signals of aligned molecules in the gas phase: Tetrazine in intense laser fields. *The Journal of Physical Chemistry A*, 107(34):6622–6629, 2003.

[80] A. Debnarova, S. Techert, and S. Schmatz. Computational studies of the X-ray scattering properties of laser aligned stilbene. *The Journal of Chemical Physics*, 134(5):054302, 2011.

[81] N. Böwering, M. Volkmer, C. Meier, J. Lieschke, and M. Fink. Electron diffraction from oriented molecules: e–CH_3Cl. *Zeitschrift für Physik D Atoms, Molecules and Clusters*, 30(2):177–182, 1994.

[82] N. Böwering, M. Volkmer, C. Meier, J. Lieschke, and R. Dreier. Electron diffraction from oriented molecules and implications for molecular structure analysis. *Journal of Molecular Structure*, 348:49–52, 1995.

[83] W. Dong, N. Mukherjee, and R. N. Zare. Optical preparation of H_2 rovibrational levels with almost complete population transfer. *The Journal of Chemical Physics*, 139(7):074204, 2013.

[84] R. Santra. Concepts in X-ray physics. *Journal of Physics B: Atomic, Molecular and Optical Physics*, 42(2):023001, 2008.

[85] W. Kohn. Nobel Lecture: Electronic structure of matterwave functions and density functionals. *Reviews of Modern Physics*, 71(5):1253, 1999.

[86] W. Thiel and G. Hummer. Nobel 2013 chemistry: Methods for computational chemistry. *Nature*, 504(7478):96–97, 2013.

[87] N. S. Ostlund and A. Szabo. *Modern Quantum Chemistry: Introduction to Advanced Electronic Structure Theory*. Dover Publications Inc, 1996.

[88] M. Born and R. Oppenheimer. Zur quantentheorie der molekeln. *Annalen der Physik*, 389(20):457–484, 1927.

[89] J. C. Slater. Atomic shielding constants. *Physical Review*, 36(1):57, 1930.

[90] S. F. Boys. Electronic wave functions. I. A general method of calculation for the stationary states of any molecular system. In *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, volume 200, pages 542–554. The Royal Society, 1950.

[91] H. B. Schlegel and M. J. Frisch. Transformation between Cartesian and pure spherical harmonic Gaussians. *International Journal of Quantum Chemistry*, 54(2):83–87, 1995.

[92] E. Besalú and R. Carbó-Dorca. The general gaussian product theorem. *Journal of Mathematical Chemistry*, 49:1769–1784, 2011.

[93] A. Debnarova, S. Techert, and S. Schmatz. *Ab initio* treatment of time-resolved X-ray scattering: Application to the photoisomerization of stilbene. *The Journal of Chemical Physics*, 125(22):224101, 2006.

[94] A. Debnarova, S. Techert, and S. Schmatz. *Ab Initio* Studies of Ultrafast X-ray Scattering of the Photodissociation of Iodine. *The Journal of Chemical Physics*, 133(12), 2010.

[95] A. H. Zewail. 4d ultrafast electron diffraction, crystallography, and microscopy. *Annu. Rev. Phys. Chem.*, 57:65–103, 2006.

[96] H.-J. Werner, P. J. Knowles, G. Knizia, F. R. Manby, and M. Schtz. Molpro: a general-purpose quantum chemistry program package. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 2(2):242–253, 2012.

[97] D. T. Cromer and D. Liberman. Relativistic calculation of anomalous scattering factors for x rays. *The Journal of Chemical Physics*, 53(5):1891–1898, 1970.

[98] D. Liberman, J. T. Waber, and D. T. Cromer. Self-consistent-field dirac-slater wave functions for atoms and ions. i. comparison with previous calculations. *Physical Review*, 137(1A):A27, 1965.

[99] L. C. Snyder. Jahn-Teller distortions in cyclobutadiene, cyclopentadienyl radical, and benzene positive and negative ions. *The Journal of Chemical Physics*, 33(2):619–621, 1960.

[100] A. C. Wahl and G. Das. The multiconfiguration self-consistent field method. In *Methods of electronic structure theory*, pages 51–78. Springer, 1977.

[101] N. E. Henriksen and K. B. Møller. On the theory of time-resolved X-ray diffraction. *The Journal of Physical Chemistry B*, 112(2):558–567, 2008.

[102] U. Lorenz, K. B. Møller, and N. E. Henriksen. Theory of time-resolved inelastic X-ray diffraction. *Physical Review A*, 81(2):023422, 2010.

[103] H. Oberhammer and S. H. Bauer. Structures and conformations of the cyclohexadienes. *Journal of the American Chemical Society*, 91(1):10–16, 1969.

[104] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. *Numerical recipes in FORTRAN*. Cambridge University Press, 1992.

[105] S. Boutet, L. Lomb, G. J. Williams, T. R. Barends, A. Aquila, R. B. Doak, U. Weierstall, D. P. DePonte, J. Steinbrener, R. L. Shoeman, et al. High-resolution protein structure determination by serial femtosecond crystallography. *Science*, 337(6092):362–364, 2012.

[106] H. N. Chapman, P. Fromme, A. Barty, T. A. White, R. A. Kirian, A. Aquila, M. S. Hunter, J. Schulz, D. P. DePonte, U. Weierstall, et al. Femtosecond x-ray protein nanocrystallography. *Nature*, 470(7332):73–77, 2011.

[107] S. Tsuneyuki, T. Kobori, K. Akagi, K. Sodeyama, K. Terakura, and H. Fukuyama. Molecular orbital calculation of biomolecules with fragment molecular orbitals. *Chemical Physics Letters*, 476(1):104–108, 2009.

[108] R. Zalesny, M. G. Papadopoulos, P. G. Mezey, and J. Leszczynski. *Linear-scaling techniques in computational chemistry and physics: Methods and applications*. Springer Science+ Business Media BV, 2011.

[109] K. Kitaura, E. Ikeo, T. Asada, T. Nakano, and M. Uebayasi. Fragment molecular orbital method: An approximate computational method for large molecules. *Chemical Physics Letters*, 313(3):701–706, 1999.

[110] C. R. Jacob and J. Neugebauer. Subsystem density-functional theory. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 4(4):325–362, 2014.

[111] C. Steinmann, D. G. Fedorov, and J. H. Jensen. Effective Fragment Molecular Orbital Method: A Merger of the Effective Fragment Potential and Fragment Molecular Orbital Methods. *The Journal of Physical Chemistry A*, 114(33):8705–8712, 2010.

[112] C. Steinmann, D. G. Fedorov, and J. H. Jensen. The effective fragment molecular orbital method for fragments connected by covalent bonds. *PLoS One*, 7(7):e41117, 2012.

[113] D. G. Fedorov and K. Kitaura. The importance of three-body terms in the fragment molecular orbital method. *The Journal of Chemical Physics*, 120(15):6832–6840, 2004.

[114] D. G. Fedorov and K. Kitaura. Multiconfiguration self-consistent-field theory based upon the fragment molecular orbital method. *The Journal of Chemical Physics*, 122(5):054108, 2005.

[115] D. G. Fedorov, T. Nagata, and K. Kitaura. Exploring chemistry with the fragment molecular orbital method. *Physical Chemistry Chemical Physics*, 14(21):7562–7577, 2012.

[116] H. J. Kulik, N. Luehr, I. S. Ufimtsev, and T. J. Martinez. *Ab initio* quantum chemistry for protein structures. *The Journal of Physical Chemistry B*, 116(41):12501–12509, 2012.

[117] H. J. Suominen and A. Kirrander. How to observe coherent electron dynamics directly. *Physical Review Letters*, 112(4):043002, 2014.

[118] W. Humphrey, A. Dalke, and K. Schulten. VMD: visual molecular dynamics. *Journal of Molecular Graphics*, 14(1):33–38, 1996.

[119] M. Suenaga. Development of a GUI for GAMESS/FMO calculation. *Journal of Computational Chemistry Japan*, 7:33–53, 2008.

[120] C. M. Kelly, T. Northey, K. Ryan, B. R. Brooks, A. L. Kholkin, B. J. Rodriguez, and N.-V. Buchete. Conformational dynamics and aggregation behavior of piezoelectric diphenylalanine peptides in an external electric field. *Biophysical Chemistry*, 196:16–24, 2015.

[121] A. K. Rappé, C. J. Casewit, K. Colwell, W. Goddard Iii, and W. Skiff. UFF, a full periodic table force field for molecular mechanics and molecular dynamics simulations. *Journal of the American Chemical Society*, 114(25):10024–10035, 1992.

[122] J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman, and D. A. Case. Development and testing of a general amber force field. *Journal of computational chemistry*, 25(9):1157–1174, 2004.

[123] A. D. MacKerell, N. Banavali, and N. Foloppe. Development and current status of the charmm force field for nucleic acids. *Biopolymers*, 56(4):257–265, 2000.

[124] W. Wang, O. Donini, C. M. Reyes, and P. A. Kollman. Biomolecular simulations: recent developments in force fields, simulations of enzyme catalysis, protein-ligand, protein-protein, and protein-nucleic acid noncovalent interactions. *Annual review of biophysics and biomolecular structure*, 30(1):211–243, 2001.

[125] A. Fleming. On a remarkable bacteriolytic element found in tissues and secretions. *Proceedings of the Royal Society of London B: Biological Sciences*, 93(653):306–317, 1922.

[126] C. C. F. Blake, D. F. Koenig, G. A. Mair, A. C. T. North, D. C. Phillips, and V. R. Sarma. Structure of hen egg-white lysozyme: a three-dimensional Fourier synthesis at 2 Å resolution. *Nature*, 206(4986):757–761, 1965.

[127] S. Release. 4: Maestro, version 10.4, 2015. *Schrödinger, LLC, New York, NY*, 2015.

[128] S. Pronk, S. Páll, R. Schulz, P. Larsson, P. Bjelkmar, R. Apostolov, M. R. Shirts, J. C. Smith, P. M. Kasson, D. van der Spoel, et al. GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics*, 29(7):845–854, 2013.

[129] K. Lindorff-Larsen, S. Piana, K. Palmo, P. Maragakis, J. L. Klepeis, R. O. Dror, and D. E. Shaw. Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins: Structure, Function, and Bioinformatics*, 78(8):1950–1958, 2010.

[130] J. N. Galayda, J. Arthur, D. F. Ratner, and W. E. White. X-ray free-electron laserspresent and future capabilities. *Journal of the Optical Society of America B*, 27(11):B106–B118, 2010.

[131] A. Barty, J. Küpper, and H. N. Chapman. Molecular imaging using X-ray free-electron lasers. *Annual Review of Physical Chemistry*, 64:415–435, 2013.

[132] J. Bentley and R. F. Stewart. Two-centre calculations for x-ray scattering. *Journal of Computational Physics*, 11:127–145, 1973.

[133] I. Hargittai and M. Hargittai. *Stereochemical applications of gas-phase electron diffraction*. VCH, 1988.

[134] L. Wolniewicz, T. Orlikowski, and G. Staszewska. $^1\Sigma_u$ and $^1\Pi_u$ states of the hydrogen molecule: Nonadiabatic couplings and vibrational levels. *Journal of Molecular Spectroscopy*, 238(1):118–126, 2006.

[135] T. Orlikowski, G. Staszewska, and L. Wolniewicz. Long range adiabatic potentials and scattering lengths for the EF, e and h states of the hydrogen molecule. *Molecular Physics*, 96(10):1445–1448, 1999.

[136] R. F. Stewart, J. Bentley, and B. Goodman. Generalized X-ray scattering factors in diatomic molecules. *The Journal of Chemical Physics*, 63(9):3786–3793, 1975.

[137] K. K. Irikura. Experimental vibrational zero-point energies: Diatomic molecules. *Journal of Physical and Chemical Reference Data*, 36(2):389–397, 2007.

[138] X. Ren, V. Makhija, and V. Kumarappan. Multipulse three-dimensional alignment of asymmetric top molecules. *Physical Review Letters*, 112(17):173602, 2014.

[139] P. Ho, D. Starodub, D. Saldin, V. Shneerson, A. Ourmazd, and R. Santra. Molecular structure determination from X-ray scattering patterns of laser-aligned symmetric-top molecules. *The Journal of Chemical Physics*, 131(13):131101, 2009.

[140] P. Reckenthaeler, M. Centurion, W. Fuß, S. A. Trushin, F. Krausz, and E. E. Fill. Time-resolved electron diffraction from selectively aligned molecules. *Physical Review Letters*, 102(21):213001, 2009.

[141] J. Yang, V. Makhija, V. Kumarappan, and M. Centurion. Reconstruction of three-dimensional molecular structure from diffraction of laser-aligned molecules. *Structural Dynamics*, 1(4):044101, 2014.

[142] C. Yu, H. Wei, X. Wang, A.-T. Le, R. Lu, and C. Lin. Reconstruction of two-dimensional molecular structure with laser-induced electron diffraction from laser-aligned polyatomic molecules. *Scientific Reports*, 5(15753), 2015.

[143] C.-I. Bränd'en and T. Alwyn Jones. Between objectivity and subjectivity. *Nature*, 343:687–689, 1990.

[144] K. Saita and D. V. Shalashilin. On-the-Fly *Ab Initio* Molecular Dynamics with Multiconfigurational Ehrenfest Method. *The Journal of Chemical Physics*, 137(22), 2012.

[145] A. Kirrander, K. Saita, and D. V. Shalashilin. Ultrafast X-ray Scattering from Molecules. *Journal of Chemical Theory and Computation*, 12(3):957–967, 2016. PMID: 26717255.

[146] A. Debnarova, S. Techert, and S. Schmatz. *Ab initio* studies of ultrafast X-ray scattering of the photodissociation of iodine. *The Journal of Chemical Physics*, 133(12):124309, 2010.