Simulations of single-particle imaging of hydrated proteins with x-ray free-electron lasers

C. Fortmann-Grote
J. Bielecki
Z. Jurek
R. Santra
B. Ziaja-Motyka
A. P. Mancuso
Simulations of single–particle imaging of hydrated proteins with x-ray free-electron lasers

C Fortmann–Grote a, J Bielecki a, Z Jurek b, c, R Santra b, c, d, B Ziaja-Motyka b, c, e, A P Mancuso a

a European XFEL GmbH, Holzkoppel 4, 22869 Schenefeld, Germany
b Center for Free Electron Laser Science, Deutsches Elektronen Synchrotron, Notkestrasse 85, 22607 Hamburg, Germany
c The Hamburg Center for Ultrafast Imaging, Luruper Chaussee 149, 22761 Hamburg, Germany
d Department of Physics, University of Hamburg, Jungiusstrasse 9, 20355 Hamburg, Germany
e Institute of Nuclear Physics, Polish Academy of Sciences, Radzikowskiego 152, 31-342 Krakow, Poland

Abstract. We employ start–to-end simulations to model coherent diffractive imaging of single biomolecules using x-ray free electron lasers. This technique is expected to yield new structural information about biologically relevant macromolecules thanks to the ability to study the isolated sample in its natural environment as opposed to crystallized or cryogenic samples. The effect of the solvent on the diffraction pattern and interpretability of the data is an open question. We present first results of calculations where the solvent is taken into account explicitly. They were performed with a molecular dynamics scheme for a sample consisting of a protein and a hydration layer of varying thickness. Through $R$–factor analysis of the simulated diffraction patterns from hydrated samples, we show that the scattering background from realistic hydration layers of up to 3 Å thickness presents no obstacle for the resolution of molecular structures at the sub–nm level.

Keywords: Start–to–end simulations, single–particle imaging, hydration layer.

* C. Fortmann-Grote carsten.grote@xfel.eu

1 Introduction

X–ray free–electron lasers (XFELs) deliver ultra–short fs pulses of x–ray photons at fluences that exceed those of synchrotron radiation sources by many orders of magnitude. 1 These key characteristics (pulse durations of a few femtoseconds, of the order 10^{12} photons per pulse, focussed to spot sizes of a few 100 nm in diameter) open scientific opportunities in many research fields, 2 including ultra–fast dynamics in atoms, molecules, and clusters, 3–5 condensed matter spectroscopy, 6,7 high energy density science, 8 materials research 9 , and structural biology 10 . In the latter case, the “probe–before–destruction” technique, i.e. taking a diffraction snapshot of a single particle before it disintegrates as a consequence of x–ray induced damage, will allow to reach structure resolution at the Å scale even for molecules that withstand crystallization. This idea, mostly referred to as
single–particle imaging (SPI), was originally envisioned by Breedlove and Trammel,\textsuperscript{11} later supported by simulations,\textsuperscript{12} and finally demonstrated experimentally.\textsuperscript{13} For a summary of SPI results from the Linac Coherent Light Source (LCLS), see the surveys by Barty\textsuperscript{14} and Bostedt et al.\textsuperscript{2}

Forward simulations of coherent diffraction from molecules play a pivotal rôle in determining preferential experimental parameters (e.g. the photon fluence\textsuperscript{15–18}). In two recent papers,\textsuperscript{19, 20} the present authors studied the impact of pulse duration on the quality of measurable data by means of start–to–end simulations of the entire experimental setup. They described the propagation of radiation from the photon source through beamline optics, the photon–matter interaction, scattering from the sample, and photon detection. The pipeline also included the computer assisted reconstruction of the electron density from the simulated diffraction patterns.

In these start–to–end simulations, the original sample structure, measured by protein crystallography, was queried from the protein database (PDB).\textsuperscript{21} The pdb structure typically contains one or several conformations of the protein and incorporated water molecules. In this work, we are interested in diffraction from single particles embedded in water, to mimic the conditions of SPI experiments more closely.

Sample delivery for SPI experiments is an active area of research.\textsuperscript{10, 22} The Gas Dynamic Virtual Nozzle\textsuperscript{23} ejects sub-µm sized droplets containing on average one sample particle (e.g. a virus or a protein) into vacuum, where the solvent partly evaporates from the particle. Molecular Dynamics (MD) simulations of proteins in aqueous environments are able to model this dehydration process.\textsuperscript{24} Results indicate that only a thin layer of solvent remains around the protein, the first hydration shell. In the extended system, i.e. protein plus hydration shell plus bulk water, this first hydration shell is identified through a peak in the protein–solvent radial distribution function between 2 Å and 4 Å.\textsuperscript{25} Variations in the water layer thickness and in the distance to the protein...
result from hydrophobic and hydrophilic sites close to the protein surface. The density in the first hydration shell was found experimentally to be 10% to 15% higher than in bulk water; this result was also confirmed in MD simulations. The higher density was attributed in part to geometric rearrangement of water molecules in presence of the protein and, in part, due to a perturbation of the water structure by the electrostatic field generated by protein surface atoms shortening the O–O distance and increasing the coordination number. Dynamical properties of the hydration layer are discussed in a survey by Halle, see also the recent review by Laage and coworkers.

Simulations suggested and experiments have shown that a hydration layer acts as a tamper, slowing down radiation damage processes by providing an electron reservoir that can compensate for electrons that have been removed from the molecule by photo–ionization or Auger processes. On the other hand, the temporal onset of electrostatic trapping mechanism, leading to enhanced ionization, heating and expansion of the molecule, was found to depend inversely on the size of the system, i.e. the electron trapping sets in faster in larger molecules than in smaller molecules. Finally, the presence of water naturally enhances the scattering background, and it can be concluded that an optimal water layer thickness exists where these counteracting effects are outbalanced.

In this paper, we present first results for start–to–end simulations of a protein imaging experiment at the European XFEL explicitly taking the hydration layer into account. In this first exploratory case study, we treat the effect of the solvent separately from the aforementioned tampering effect, i.e. we are only interested in the additional scattering from the water molecules. To this end, we compare simulated diffraction patterns from the protein with and without hydration layer. We calculate a figure of merit (the $R$ factor) from the diffraction patterns as a function of the
water layer thickness. In this way, we obtain a first estimate for a tolerable water layer thickness which would still allow to interpret the diffraction signal, i.e. to infer structural information at a certain resolution length scale.

In an earlier study of the effect of hydration layers on single particle diffraction data, the water layer thickness was taken in the range 5 Å to 25 Å, i.e. larger than predicted by MD simulations. On the other hand, that paper focussed on the orientational information, i.e. the ability to reconstruct the orientation of the sample in the beam, a prerequisite for the phase retrieval step in the structure determination workflow. Here, we limit our analysis to structural information within averaged diffraction patterns, which is certainly only an intermediate step in this research activity. A comparison to the aforementioned work will follow.

2 Start–to–end simulation

We employ the start–to–end simulation framework simex_platform to simulate a SPI experiment at the European XFEL utilizing 4.96 keV photons in a 9 fs pulse focused to a spotsize of 100 nm. The XFEL pulses are queried from the XFEL X–ray Pulse Database and propagated through the SASE1 beamline and focusing optics of the SPB–SFX scientific instrument. For a detailed analysis of the propagated x–ray pulses, we refer to Ref.19

The sample molecule is the 2–Nitrogenase–Iron Protein (2NIP). The weight of 2NIP is 63.2 kDa and its size is 7 nm in diameter. We consider the situation, where the protein is embedded in a water layer.
Fig 1 Molecular Dynamics simulation box \((27.3 \times 26.1 \times 27.9 \text{nm}^3)\) with 2NIP protein in the box’ centre. Periodic boundary conditions were applied during the MD simulation and the 2NIP atoms were kept at fixed positions.

2.1 Molecular dynamics equilibration

To prepare our sample geometries, we used the MD toolbox VMD\(^{42,43}\) which allows to embed arbitrary molecular structures, obtained from a pdb entry, in a solvent box. Our simulation box of dimensions \(27.3 \times 26.1 \times 27.9 \text{nm}^3\) was chosen in a way that the protein is surrounded by 10 nm of water in all directions. It is shown in Fig. 1. We performed MD simulations using the code NAMD\(^{44}\) to allow the water to find its preferred configuration in the vicinity of the protein. The protein atoms were fixed at their initial positions, which significantly improves the convergence and the hydration layer stability.\(^{45}\) Periodic boundary conditions were applied. We used the TIP3P force field\(^{46}\) for the water–water interactions and the CHARMM22 force field\(^{47,48}\) for the interaction between water atoms and protein atoms. We performed 50000 MD steps of 2 fs stepsize while keeping the temperature at \(T = 293 \text{K}\) using a Langevin thermostat. After \(\approx 10^4\) MD steps, the
total energy reached its equilibrium value as shown in Fig. 2. From the equilibrated part of the trajectory, we picked 125 snapshots as an input to our diffraction simulations.

![Fig 2](image-url) Total energy of the solvated 2NIP system as a function of the MD step. The system is equilibrated after $\approx 1 \times 10^4$ MD steps.

To check the validity of our MD simulations, we also equilibrated the pure solvent box, i.e. containing only water molecules. Fig. 3 displays the pair distribution functions for H–H, H–O, and O–O pairs averaged over the equilibrated part of the run. Our simulation is shown as solid line, while dots denote reference data. Our results are in a good agreement with the reference data, i.e. we reproduce the location and amplitude of maxima in the correlation functions. Finer details, such as the width of the first O–H correlation peak and the discrepancy between simulation and reference data in the O–O structure beyond the first peak deserve further investigation. Here we content ourselves with the achieved agreement, since we do not expect that these sub–Å level deviations have an impact on the structure of the protein at targeted resolution levels of a few Å. Furthermore, the pair distribution functions of the combined solvent plus protein system and identification of the first hydration shell have to be analysed in future work.
**Fig 3** Averaged pair distribution functions from the equilibrated part of the MD trajectory for H–H, H–O, and O–O pairs, respectively. H–H and H–O pair distributions include the intramolecular peak. Reference data from$^{49}$ (dots) without intramolecular peaks are shown for comparison.

2.2 *Scattering simulations*

We calculate the scattering of x–ray photons from the sample, consisting of the protein and the surrounding water molecules, using the atomic positions as given by the MD snapshot. For each atom, we query the atomic form factor from a pre–calculated database. The formfactor database was obtained from first-principle Hartree-Fock-Slater electronic structure calculations using the code XATOM.$^{16,50}$ Here, we employ only the ground state form factors, i.e. excitation and ionization events are neglected. In a forthcoming paper, we will study diffraction from hydrated proteins taking these electronic processes into account, as well as the displacement of atoms during the irradiation. We will employ for that purpose the simulation code suite XRAYPAC.$^{51}$ XRAYPAC implements a microscopic model of radiation damage, containing photo–ionization, secondary processes such as Auger decay, recombination, and electron impact ionization, as well as the real space dynamics of atoms, ions, and free electrons. We have previously applied XRAYPAC to...
model radiation damage processes in pure proteins, i.e. without the hydration layer.\textsuperscript{19, 20}

From the atomic positions $R_i$ and atomic form factors $f_i(q)$, we calculate the scattered intensity at a given detector pixel as a function of transfer wavevector $q$:

$$I(q) = \Omega \frac{d\sigma_{Th}(\theta)}{d\Omega} \langle I_0 \rangle \left| \sum_i f_i(q)e^{i\mathbf{q}\cdot\mathbf{R}_i} \right|^2,$$

where $d\sigma_{Th}/d\Omega$ is the differential Thomson cross section, $\langle I_0 \rangle$ is the average pulse intensity, $\Omega$ is the solid angle spanned by the considered detector pixel. The wavevector $q$ depends on the detector geometry (distance $d$ from the sample) and pixel coordinates $(r_x, r_y)$ in the detector plane assumed to be perpendicular to the beam propagation axis and centered such that the beam axis intersects the detector plane at its origin:

$$q = \frac{2\pi}{\lambda} \begin{pmatrix} \sin(2\theta) \cos(\phi) \\ \sin(2\theta) \cos(\phi) \\ \cos(2\theta) - 1 \end{pmatrix} = \frac{2\pi}{\lambda} \begin{pmatrix} \frac{r_y}{\sqrt{r_x^2 + r_y^2 + d^2}} \\ \frac{r_x}{\sqrt{r_x^2 + r_y^2 + d^2}} \\ \frac{d - \sqrt{r_x^2 + r_y^2 + d^2}}{\sqrt{r_x^2 + r_y^2 + d^2}} \end{pmatrix}.$$

Here, $2\theta = \arctan \left( \frac{r_x^2 + r_y^2}{d} \right)$ is the scattering angle and $\phi = \arctan \left( \frac{r_y}{r_x} \right)$. In this notation, the pixel solid angle becomes $\Omega = 4 \arcsin \left( a^2 / \left[ 4(r_x^2 + r_y^2 + d^2) + a^2 \right] \right)$, with the pixel width $a$.

In our simulations, $d = 13 \text{ cm}$. The detector is a $512 \times 512$ array of pixels of side length $a = 200 \mu\text{m}$ corresponding to one quadrant of a 1 megapixel AGIPD detector.\textsuperscript{52}

3 Results

As a first test of our simulation pipeline, we calculate diffraction patterns from the pure solvent box, i.e. without the protein. At $24 \text{ keV}$ photon energy, the water ring shows up as the dominant
feature in the diffraction patterns at $\approx 2 \text{ Å}^{-1}$. An average over 100 simulated water–only patterns is shown in Fig. 4a. Fig. 4b shows the azimuthally integrated signal as function of $q$ and compared to Wide–Angle X–ray Scattering (WAXS) reference data$^{53}$ taken at $16 \text{ keV}$. As before, in the case of the pair distribution functions (see Fig. 3), we find relatively good agreement between our simulations and the reference data. The main features, i.e. the location of the principal maximum at $q \approx 2.1 \text{ Å}^{-1}$ and the shoulder at $q \approx 4.8 \text{ Å}^{-1}$ are reproduced.

![Diffraction patterns](image)

**Fig 4** (a) Averaged diffraction pattern from 125 MD snapshots showing the water ring at $q \approx 2 \text{ Å}^{-1}$. (b) Azimuthal integral over diffraction pattern in (a), normalized to maximum intensity and compared to reference diffraction data from$^{53}$

We now turn to the question how the water environment, and, in particular, the thickness of the solvent layer, influences the quality and interpretability of the measurable diffraction patterns. We picked 125 snapshots, separated by $\approx 0.6 \text{ ps}$ in time, from the equilibrated part of the MD trajectory. From each snapshot we generated a series of samples by extracting the protein plus a water layer of varying thickness. The thickness values are 0 Å, 1 Å, 3 Å, 6 Å, 10 Å, 15 Å and 20 Å.

We can now study characteristics of the diffraction from our solvated protein samples as a function of the water layer thickness. From each sample, we calculated 142 diffraction patterns after...
applying a randomly chosen rotation to all sample atoms. In this way, we emulate the unknown orientation of the sample in the experimental geometry.

Figs 5 and 6 visualize the sample (left column) for all six water layer thicknesses and one chosen simulated diffraction pattern (right column). In all six cases, the pattern corresponds to the same MD snapshot in the same orientation with respect to the beam to allow a visual comparison of the results.

The clearly visible increase in intensity of the central speckle reflects the increasing number of scatterers as the water layer thickness is incremented. Secondly, as the sample size increases, the speckle size decreases as follows from the Fourier transform relationship between real space structure and diffraction signal. We also find, that the patterns from the largest samples start to resemble concentric fringes that one would expect from a spherical object.

These three main characteristics (increasing total intensity, increasing background and increasing centricity) are observed again when looking at the azimuthally integrated averaged patterns, shown in Fig. 7. In particular, in the $q$ range $0.1 \, \text{Å}$ to $0.5 \, \text{Å}$, the appearance and increasing sharpness of fringes can be observed, which can be understood as a consequence of decreasing surface roughness of the sample as the water layer thickness increases, i.e. the water smoothes out the surface of the protein and the object becomes more spherical.

A well established figure of merit to quantify the quality of diffraction patterns in the presence of non–ideal contributions, is the so–called $R$ factor,

$$R(D) = \int_{q \leq 2\pi/D} d\Omega \left| \frac{\sqrt{n(\Omega)}}{N(D)} - \frac{\sqrt{n_{\text{ref}}(\Omega)}}{N_{\text{ref}}(D)} \right|.$$  \hfill (3)

Here, $D$ is the considered resolution length scale determined by the detector geometry, $n(\Omega)$ is the...
number of simulated photons per pixel and \( N(D) = \int_{q \leq 2\pi/D} d\Omega \sqrt{n(q, \Omega)} \) is the normalization. The index ref denotes the reference system, i.e. an ideal diffraction signal. In crystallography, one method to determine the effective resolution scale of measure structures is to determine the resolution at which \( R \) becomes larger than 0.2.

The \( R \) factor was previously used to analyze the effect of radiation damage on single particle diffraction data,\(^{12,16,54}\) employing diffraction from an undamaged sample as the reference system. Here, we use the \( R \) factor analysis to quantify the effect of the water layer, hence the reference system is the pure protein with all water molecules removed.

Fig. 8 shows the results of our \( R \) factor analysis: The \( R \) factor increases with the water layer thickness as was to be expected from the earlier qualitative discussion. In the range of resolution scales \( D < 10 \text{ Å} \), which is the range of interest for SPI, only the \( R \) factor curves for water layer thicknesses \( \leq 3 \text{ Å} \) remain below the \( R \leq 0.2 \) criterion. For water layer thicknesses above 6 Å, the criterion limits the resolution to \( D \gtrsim 30 \text{ Å} \). The precise value for the water layer thickness at which \( R \) crosses the 0.2 critical value at or below \( D = 10 \text{ Å} \) has not been identified, here. It will require a more fine grained scan between 3 Å and 6 Å thicknesses and will be carried out at a later stage.

4 Conclusions and outlook

In conclusion, we have found that the resolution length scale for SPI of proteins embedded in a hydration layer can well reach sub–nm levels if the water layer thickness does not exceed values of \( \approx 3 \text{ Å} \). This values is in agreement with predictions for the hydration layer thickness from MD simulations and experiments.
In our study, we have concentrated on the pure effect of the thickness of the water layer. Radiation damage, i.e. ionization, atomic displacement, plasma formation, and hydrodynamic expansion have been neglected. More realistic simulations would have to include these effects, including the tampering effect.

The small molecule studied here represents a limiting case. More realistic, larger molecules will be studied in future. We will then also perform the orientation and phasing of the diffraction patterns from solvated proteins to investigate the effect of the presence of a water layer on the results and performance of these algorithms. We would then also explore the interpretability of diffraction patterns, taking into account the shot–to–shot fluctuations of the XFEL beam as done in our previous work.

Disclosures

The authors declare no conflict of interests.

Acknowledgments

CFG acknowledges support from the European Cluster of Advanced Laser Light Sources (EU-CALL) project which has received funding from the European Unions Horizon 2020 research and innovation programme under grant agreement No 654220.

References


43 “The VMD Molecular Graphis Viewer.”


List of Figures

1 Molecular Dynamics simulation box (27.3 $\times$ 26.1 $\times$ 27.9 nm$^3$) with 2NIP protein in the box’ centre. Periodic boundary conditions were applied during the MD simulation and the 2NIP atoms were kept at fixed positions.

2 Total energy of the solvated 2NIP system as a function of the MD step. The system is equilibrated after $\approx 1 \times 10^4$ MD steps.

3 Averaged pair distribution functions from the equilibrated part of the MD trajectory for H–H, H–O, and O–O pairs, respectively. H–H and H–O pair distributions include the intramolecular peak. Reference data from$^{49}$ (dots) without intramolecular peaks are shown for comparison.

4 (a) Averaged diffraction pattern from 125 MD snapshots showing the water ring at $q \approx 2$ Å$^{-1}$. (b) Azimuthal integral over diffraction pattern in (a), normalized to maximum intensity and compared to reference diffraction data from$^{53}$.
2NIP sample molecule in water layers of increasing thicknesses 0 Å to 3 Å (left column) and corresponding diffraction intensities (right column).

2NIP sample molecule in water layers of increasing thicknesses 6 Å to 20 Å (left column) and corresponding diffraction intensities.

Azimuthal integral over averaged diffraction patterns as function of $q$ for various water layer thicknesses.

R factor as function of resolution scale $D = 2\pi/q$ for various water layer thicknesses.

List of Tables
Fig 5 2NIP sample molecule in water layers of increasing thicknesses 0 Å to 3 Å (left column) and corresponding diffraction intensities (right column).
Fig 6 2NIP sample molecule in water layers of increasing thicknesses 6 Å to 20 Å (left column) and corresponding diffraction intensities.
**Fig 7** Azimuthal integral over averaged diffraction patterns as function of $q$ for various water layer thicknesses.

**Fig 8** R factor as function of resolution scale $D = 2\pi/q$ for various water layer thicknesses.