Supporting Information: Controlling Protein Orientation in Vacuum Using Electric Fields

Erik G. Marklund,[*,†,‡] Tomas Ekeberg,[¶] Mathieu Moog,[§] Justin L.P. Benesch,[‡] and Carl Caleman[*,§,¶]

†*Department of Chemistry – BMC, Uppsala University, Box 576, SE-751 23 Uppsala, Sweden*

‡*Physical & Theoretical Chemistry Laboratory, Department of Chemistry, University of Oxford, South Parks Road, Oxford, GB-OX1 3QZ, United Kingdom*

¶*Center for Free-Electron Laser Research, Deutsches Elektronen Synchrotron, DE-22607, Hamburg, Germany*

§*Department of Physics and Astronomy, Uppsala University, Box 516, SE-751 20 Uppsala, Sweden*

E-mail: erik.marklund@kemi.uu.se; carl.caleman@physics.uu.se

Supporting Methods

MD simulations

Using the GROMACS simulation package, we performed vacuum MD simulations of a set of proteins, depicted in Fig. 1 (b), that have been investigated in earlier gas-phase studies: Trp-cage (1L2Y), a C-terminal fragment ('Ctf') from the bacterial ribosomal protein L7/L12 (1CTF), ubiquitin (1UBQ), and lysozyme (1AKI).[1–7] The four proteins are known to be stable in vacuum in a seemingly force-field independent manner.[2,3] Starting structures were taken from the final frames of earlier gas-phase simulations at 250 K, carrying net charges corresponding to dominant peaks in electrospray MS experiments.[3] The details of their protonation states in vacuum, which have been determined from theory and experiments (see ref[2]), were kept fixed throughout the simulations. A certain heterogeneity in terms of the exact location of protons can be expected for any protein, which adds variations in the dipole moment, and consequently also in how the proteins orient. The extent of such heterogeneity is a bit uncertain, but references in ref[2] suggest minor variations and one dominant configuration. The simulations employed the OPLS-AA/L force field.[8] To capture the long range of electrostatic forces in the absence of solvent, no cutoffs for non-bonded interactions were used. The Leap-Frog integration scheme was used with a 0.5-fs time step to propagate the equations of motion.[9] Neither pressure coupling nor periodic boundary conditions were applied, mimicking prefect vacuum. After energy minimisation followed by 10 ps equilibration with a Berendsen thermostat set to 300 K, 10-ns simulations were run for each protein without temperature coupling (some extended to 50 ns). The proteins were exposed to static external electric fields ranging from 0 to 30 000 kV cm^{-1}.[10] Each combination of protein and field strength was simulated 10 times, each time at a different random starting orientation. A schematic illustration of the simulation workflow is given in Fig. 1(c). The center-of-mass-motion was removed in these simulations, but the proteins were free to tumble. The former should not affect the simulations, except that the integration

will accumulate fewer numerical errors, whereas the latter is essential for this setup.

Root mean square deviations (RMSDs) of the C_α atoms in the protein backbone are around 0.2–0.45 nm for these proteins in vacuum at 300 K, relative to their solution structures, as are RMSDs between replicate simulations at 325 K.[2,3] We thus consider RMSDs above and below 0.5 nm with respect to starting structure to denote unfolded and structurally intact proteins, respectively.

Simulated XFEL diffraction

Diffraction patterns for a lysozyme structure (PDB code 1AKI) in random orientations were simulated using the Condor software [Fig. 1(a)], and the orientation information was saved for later use.[7,11] We imposed experimental parameters similar to those of the Coherent X-ray Imaging end-station at the Linac Coherent Light Source including 100 nm focus size and a wavelength of 8 keV. The pulse energy was however set to 40 mJ, which is a factor of 10 higher than specifications. This was necessary as lysozyme is much smaller than the particles that are currently being studied with this technique. The detector arrangement was placed such that the edge of the detector corresponds to a resolution of 2.8 Å.

Orientation recovery

Orientation recovery was done using the Expand, Maximize and Compress (EMC) algorithm, which we modified to take the orientation and associated error into account [Fig. 1(a)].[12] For each pattern, EMC calculates the probability that it represents each of the sampled orientations given the Fourier model from the previous iteration. These probabilities are then used as weights when assembling the patterns to the Fourier model of the next iteration. In our enhanced EMC (EEMC), the probabilities are multiplied with the corresponding probabilities from the dipole orientation. These are modelled as a Gaussian centered around the applied field and with no restriction for in-plane rotation. The Gaussian is given a standard deviation of 7°, which corresponds to the degree of orientation achieved for lysozyme. The

field is in turn modelled being offset from the true orientation with an angle from the same Gaussian distribution. As such, our procedure regards both the error in the orientation and the corresponding uncertainty when using that information. See Fig. 2(a).

Orientation recovery was attempted for a varying number of diffraction patterns (1000, 3000, and 10 000) with missing data in the center, representing a beamstop diameter of 1.4 Shannon Pixels (SP). The beam stop is used to protect the detector from the direct X-ray beam and appears as a shadow in the diffracted image. Orientation recovery was also carried out for a range of different beamstop diameters (1.4, 2.8, 4.2, 5.6, and 7.0 SP), using a data set of 10 000 diffraction patterns. EMC and our EEMC were used separately for each case.

To evaluate the success of the orientation recovery we compare the recovered orientation for each diffraction pattern to the true orientation that was used for simulating the respective pattern. To do this comparison, we first have to recover the overall orientation that relates the true fourier-space to the recovered one. This was done by calculating the average relative orientation between the true and recovered orientation for all diffraction patterns:

$$c = \frac{\Sigma_{i=0}^{N} b_i \cdot a_i^{-1}}{\left|\Sigma_{i=0}^{N} b_i \cdot a_i^{-1}\right|} \tag{1}$$

Where a_i and b_i are respectively the true and recovered orientation of the ith pattern, c is the overall orientation and N is the number of diffraction patterns. The sum is a element-wise addition of the orientations represented as quaternions.

The result is presented as the average angular difference, ϕ, between the true rotations and the recovered ones.

$$\phi = \Sigma_{i=0}^{N} \text{angle}\left(b_i^{-1} \cdot c \cdot a_i\right) \tag{2}$$

Here the function $\text{angle}\,(q)$ gives the angle of the rotation q given by

$$\text{angle}\,(q) = 2\arccos w \tag{3}$$

where w is the first element of the quaternion representing the rotation q.

Supporting Figures

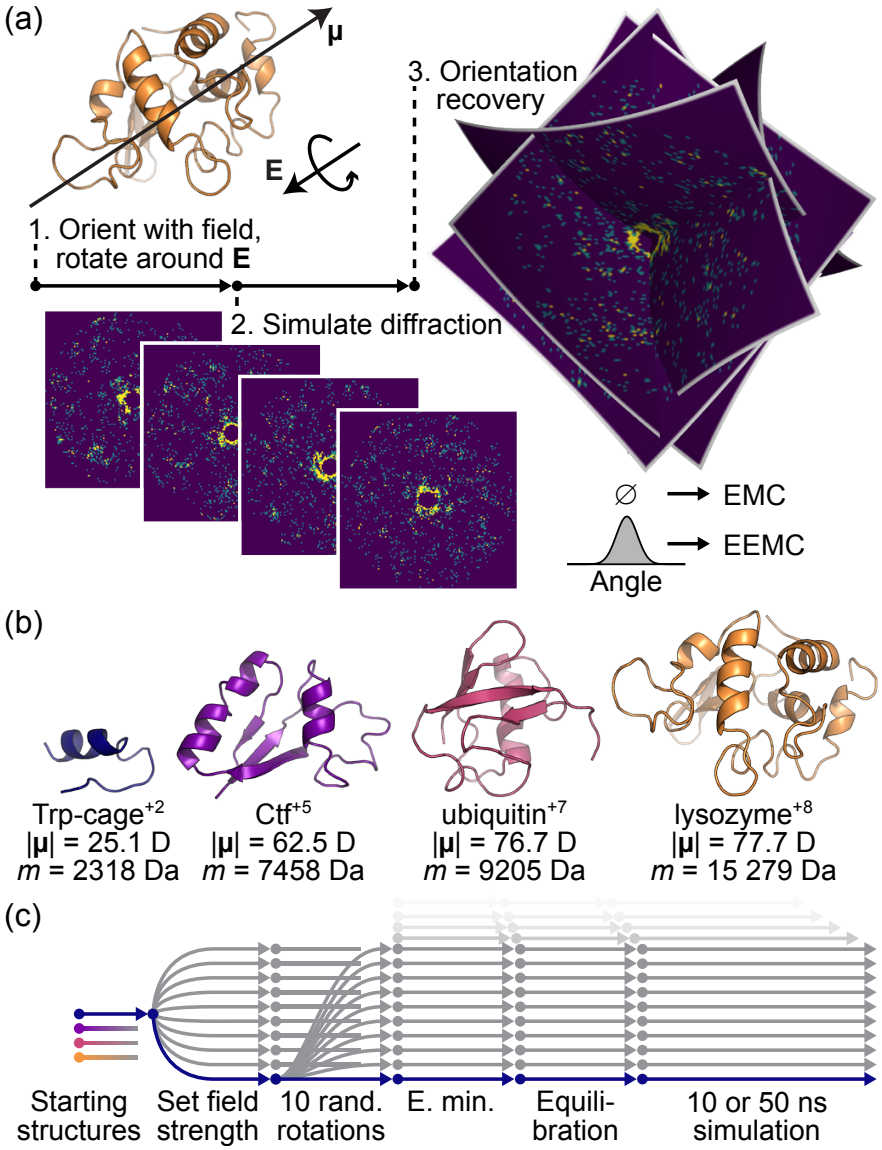

Figure 1: Methods outline. (c) Outline of how diffraction patterns were generated and combined to assess the impact of the added orientation information. (b) The four proteins used as model systems in this study, their respective charge, dipole moment, and mass. (c) Outline of the MD simulation procedure for the four proteins.

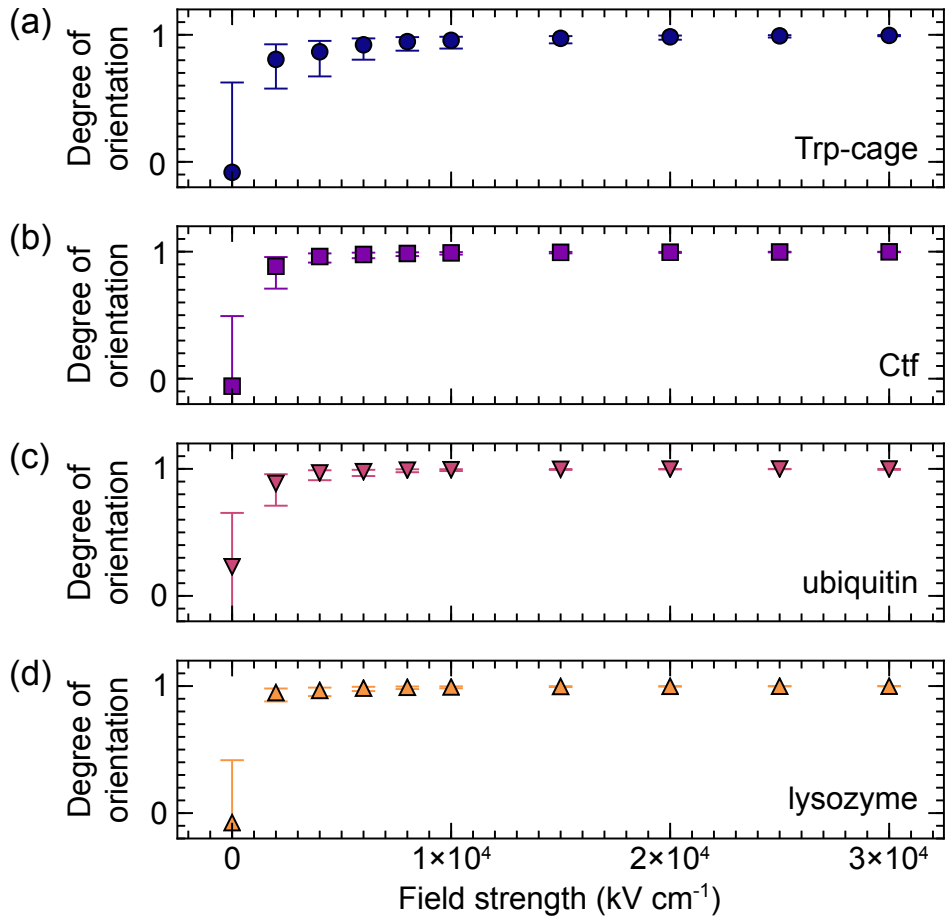

Figure 2: Degree of orientation, 5–10 ns. This is the data seen in Fig. 2(b) in the main text, but with the four proteins displayed separately, and with error bars representing the standard deviation on each side of the mean. (a)–(d) Trp-cage, Ctf, ubiquitin, and lysozyme all attain an orientation in the field direction within the 10 ns of simulations for all non-zero fields investigated in the first round of simulations. At zero field strength the orientation is on average close to zero, and the standard deviation is on the order of $\sqrt{1/2}$, which is expected for random rotations.

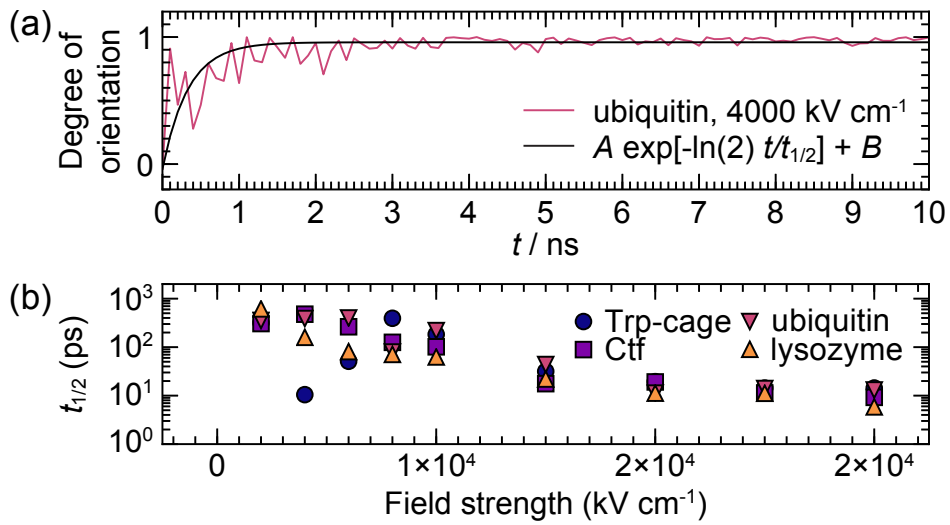

Figure 3: Orientation characteristics in the 10-ns simulations. (a) The degree of orientation over time for a selected ubiquitin trajectory at $4000\,\mathrm{kV\,cm^{-1}}$. Fig. 2(a) and the top part of Fig. 2(e) are also based on this particular simulation. A simple model, $A\exp[-\ln(2)\,t/t_{1/2}]+B$, constrained to intersect the orientation at $t=0$, has been fitted to the data. This allows for the long-term orientation (B) and the characteristic timescale of orientation ($t_{1/2}$) to be estimated. (b) Average $t_{1/2}$, for the four proteins as function of field strength. $t_{1/2}$ could not always be robustly determined at lower field strengths.

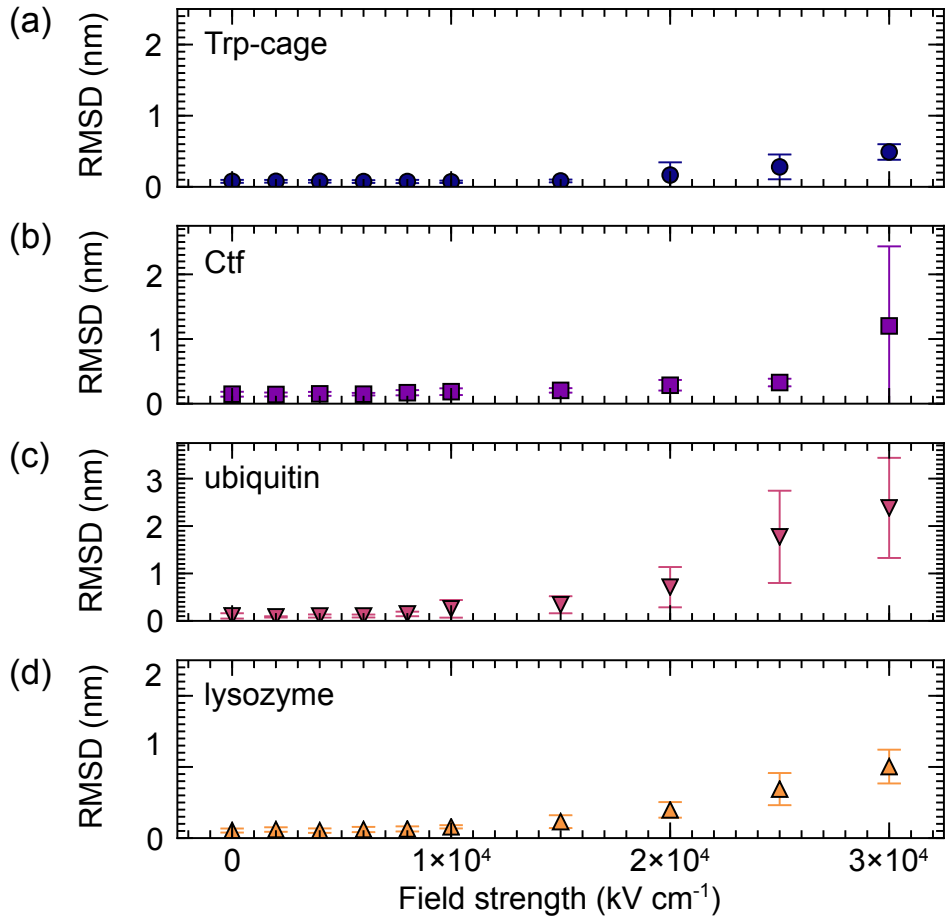

Figure 4: Average RMSD of the proteins' C_α atoms with respect to the starting configuration, 5–10 ns. This is the data seen in Fig. 2(c) in the main text, but with the four proteins displayed separately, and with error bars representing the standard deviation around the mean. (a)–(d) Trp-cage, Ctf, ubiquitin, and lysozyme all remain nativelike in fields up to 15 000 kV cm^{-1}. Only a slight increase can be seen for Trp-cage and Ctf up to 25 000 kV cm^{-1}. Ubiquitin and, to some degree, lysozyme start to lose their structures at 20 000 kV cm^{-1}. All proteins have heavily distorted structures when exposed to fields of 30 000 kV cm^{-1}.

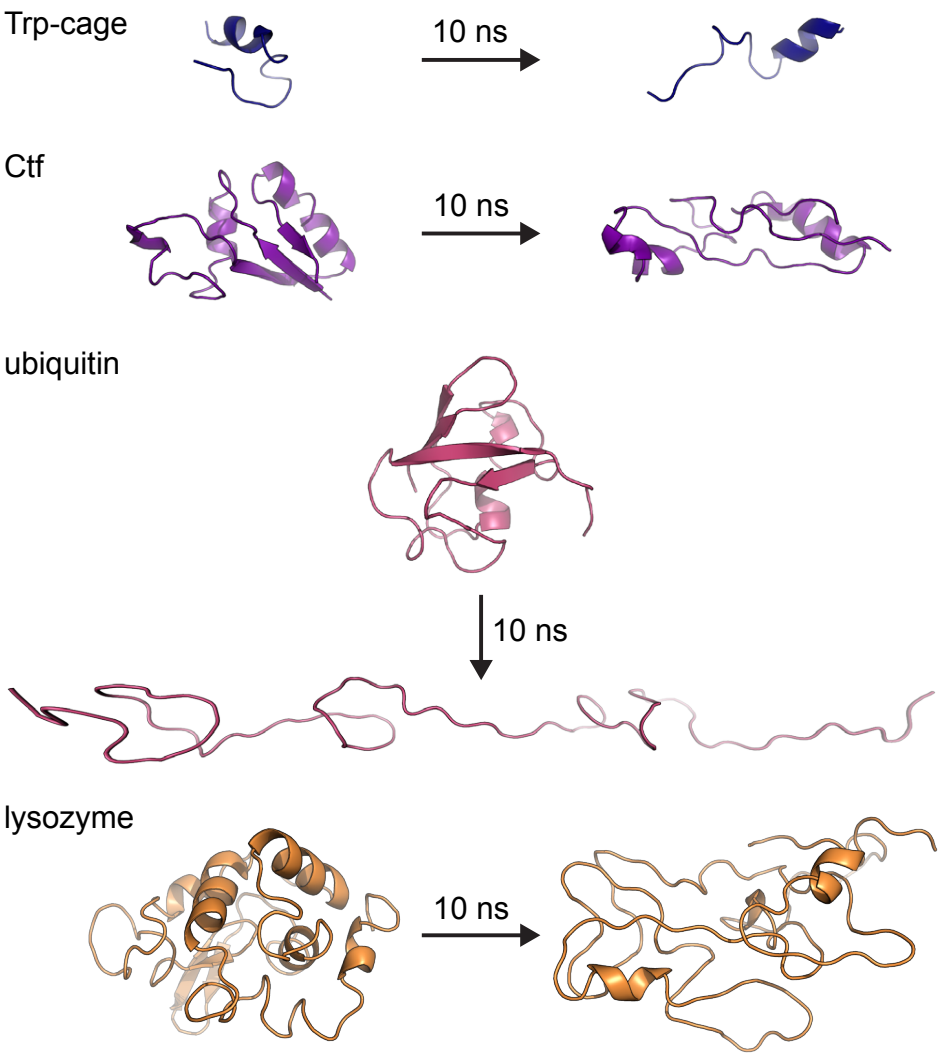

Figure 5: Final structures after 10 ns of simulation in 30 000 kV cm^{-1} from selected simulation trajectories. Most simulations above the unfolding threshold ended with a globularlike distorted structure, but near-linear structures were produced in a few cases.

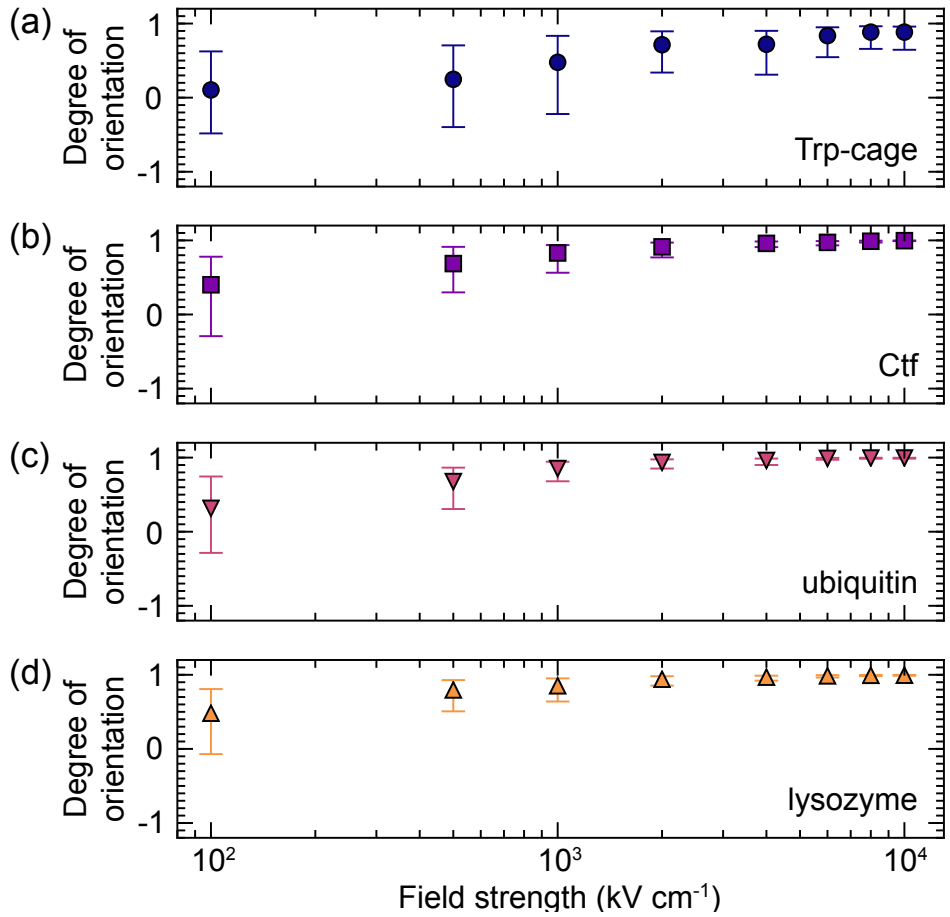

Figure 6: Degree of orientation, 45–50 ns. (a)–(d) shows the data from Fig. 2(f), but with the four proteins displayed separately, and with error bars representing the standard deviation on each side of the mean. Note the logarithmic field axis. Trp-cage, Ctf, ubiquitin, and lysozyme are only weakly oriented by fields at $100\,\mathrm{kV\,cm^{-1}}$, gradually becoming more oriented as the field strength increases. At $1000\,\mathrm{kV\,cm^{-1}}$ Ctf, ubiquitin, and lysozyme are close to perfectly oriented, whereas orientation of Trp-cage require higher fields. With Trp-cage being the smallest protein, this hints that there is a size dependence to the orientation phenomenon in addition to the dependence on exposure time.

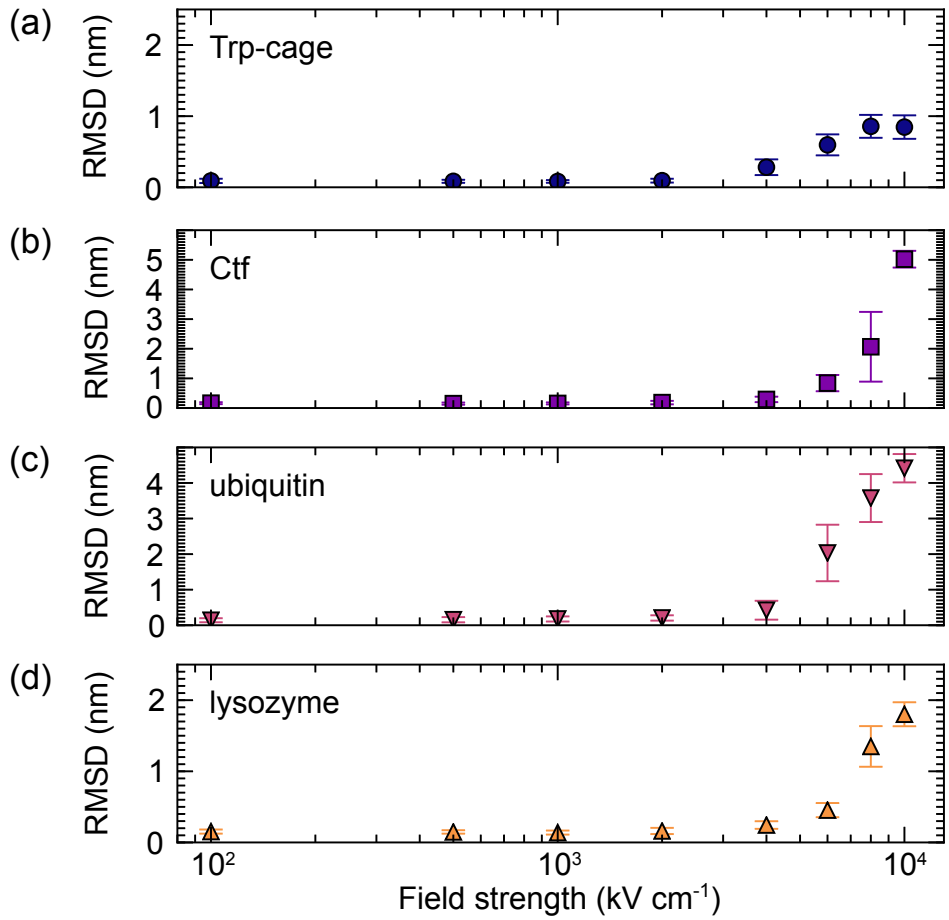

Figure 7: Average RMSD of the proteins' C_α atoms with respect to the starting configurations, 45–50 ns. (a)–(d) shows the data from Fig. 2(g), but with the four proteins displayed separately, and with error bars representing the standard deviation around the mean. Note the logarithmic field axis. On this timescale Trp-cage, Ctf, ubiquitin, and lysozyme all remain nativelike in fields up to about 4000 kV cm^{-1}, beyond which unfolding starts.

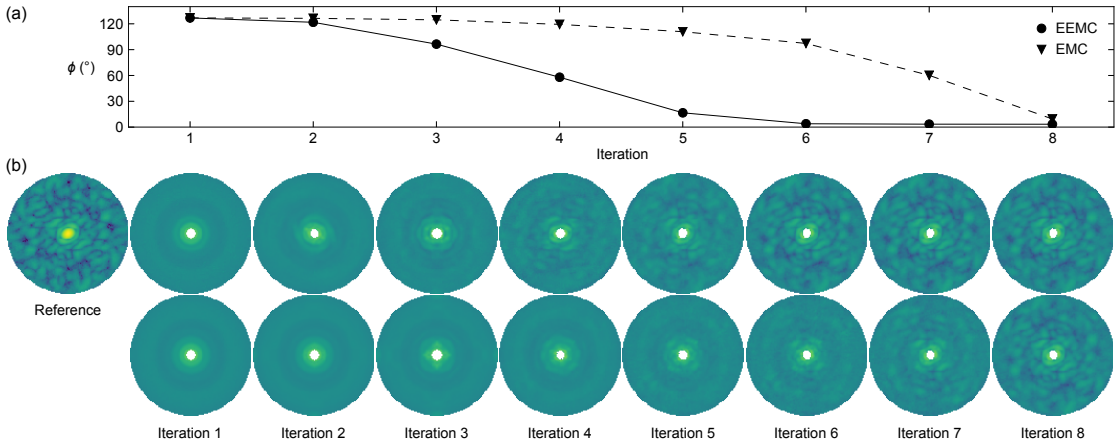

Figure 8: Speed of convergence. (a) Evolution of the average angular error for EMC and EEMC over iterations. Both runs were provided 10 000 diffraction patterns and correspond to the rightmost column in Supplementary Fig. 9. (b) The top row shows the behaviour of EEMC by the dipole orientation and the bottom row shows the behaviour of EMC at the first 8 iterations (out of a total of 10 iterations). EEMC had converged already at iteration 6, whereas plain EMC converged at iteration 8.

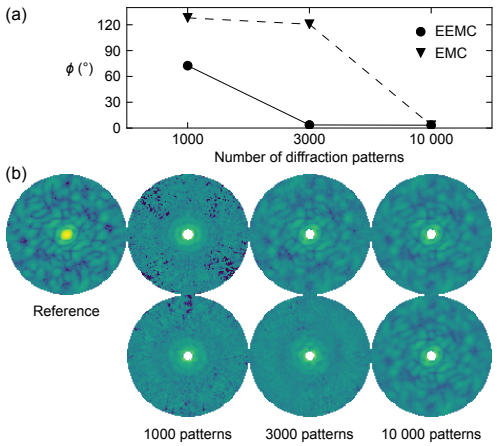

Figure 9: Error and data volume. (a) Average angular error for EMC and EEMC applied to a range of differently sized data sets with a beamstop of 1.4 SP. (b) The top row shows the output from EEMC and the bottom row shows the output from EMC, corresponding to the data points in (a). Both methods failed at low pattern counts and succeed at high counts. For 3000 available patterns however only the EEMC converged.

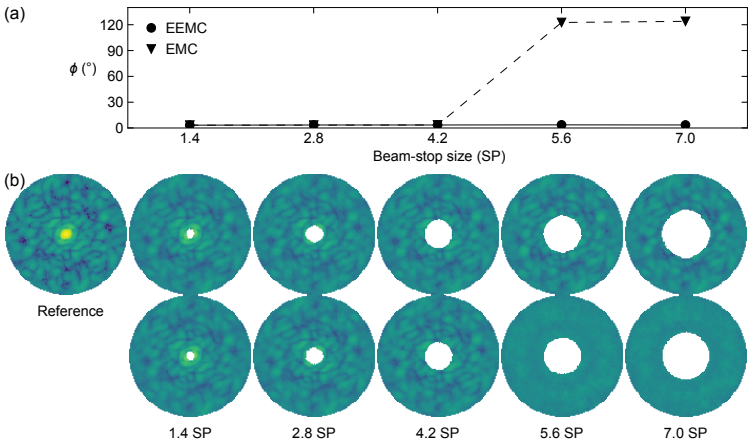

Figure 10: Error and beamstop size. (a) Average angular error for EMC and EEMC applied to data sets comprising 10 000 diffraction patterns but with differently sized beamstops. (b) The top row shows the output from EEMC and the bottom row shows the output from EMC, corresponding to the data points in (a). For large beamstops the EMC failed but EEMC still converged.

Supplemental Videos

Refer to the online version for Supplemental Videos S1 and S2.

References

(1) Hess, B.; Kutzner, C.; van der Spoel, D.; Lindahl, E. GROMACS 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation. *J. Chem. Theory Comput.* **2008**, *4*, 435–447.

(2) Patriksson, A.; Marklund, E.; van der Spoel, D. Protein Structures under Electrospray Conditions. *Biochemistry* **2007**, *46*, 933–945.

(3) Marklund, E. G.; Larsson, D. S. D.; van der Spoel, D.; Patriksson, A.; Caleman, C. Structural stability of electrosprayed proteins: Temperature and hydration effects. *Phys. Chem. Chem. Phys.* **2009**, *11*, 8069.

(4) Neidigh, J. W.; Fesinmeyer, R. M.; Andersen, N. H. Designing a 20-residue protein. *Nat. Struct. Mol. Biol.* **2002**, *9*, 425 – 430.

(5) Leijonmarck, M.; Liljas, A. Structure of the C-terminal domain of the ribosomal protein L7/L12 from Escherichia Coli at 1.7 Ångströms. *J. Mol. Biol.* **1987**, *195*, 555–580.

(6) Vijay-Kumar, S.; Bugg, C. E.; Cook, W. J. Structure of Ubiquitin Refined at 1.8 Å Resolution. *J. Mol. Biol.* **1987**, *194*, 531–544.

(7) Artymiuk, P. J.; Blake, C. F.; Rice, D. W.; Wilson, K. S. The structures of the monoclinic and orthorombic forms of hen egg-white lysozyme at 6 Ångstroms resolution. *Acta Crystallogr. Sect. B* **1982**, *38*, 778–783.

(8) Kaminski, G. A.; Friesner, R. A.; Tirado-Rives, J.; Jorgensen, W. L. Evaluation and Reparametrization of the OPLS-AA Force Field for Proteins via Comparison with Accurate Quantum Chemical Calculations on Peptides. *J. Phys. Chem. B* **2001**, *105*, 6474–6487.

(9) van Gunsteren, W. F.; Berendsen, H. J. C. A Leap-Frog Algorithm for Stochastic Dynamics. *Mol. Simul.* **1988**, *1*, 173–185.

(10) Caleman, C.; van der Spoel, D. Picosecond melting of ice by an infrared laser pulse: A simulation study. *Angew. Chem. Intl. Ed.* **2008**, *47*, 1417–1420.

(11) Hantke, M. F.; Ekeberg, T.; Maia, F. R. N. C. Condor: A simulation tool for flash X-ray imaging. *J. App. Cryst.* **2016**, *49*, 1356–1362.

(12) Loh, N.-T. D.; Elser, V. Reconstruction algorithm for single-particle diffraction imaging experiments. *Phys. Rev. E Stat. Nonlin. Soft Matt. Phys.* **2009**, *80*, 026705.

(13) Seibert, M. M.; Ekeberg, T.; Maia, F. R. N. C.; Svenda, M.; Andreasson, J.; Jonsson, O.; Odic, D.; Iwan, B.; Rocker, A.; Westphal, D. et al. Single mimivirus particles intercepted and imaged with an X-ray laser. *Nature* **2011**, *470*, 78–81.

(14) Ekeberg, T. E.; Svenda, M.; Abergel, C.; Maia, F. R. N. C.; Seltzer, V.; Claverie, J.-M.; Hantke, M.; Joensson, O.; Nettelblad, C.; van der Schot, G. et al. Three-Dimensional

Reconstruction of the Giant Mimivirus Particle with an X-Ray Free-Electron Laser. *Phys. Rev. Lett.* **2015**, *114*, 098102.

(15) Loh, N. D.; Bogan, M. J.; Elser, V.; Barty, A.; Boutet, S.; Bajt, S.; Hajdu, J.; Ekeberg, T.; Maia, F. R. N. C.; Schulz, J. et al. Cryptotomography: reconstructing 3D Fourier intensities from randomly oriented single-shot diffraction patterns. *Phys. Rev. Lett.* **2010**, *104*, 225501.

(16) Aquila, A.; Barty, A.; Bostedt, C.; Boutet, S.; Carini, G.; dePonte, D.; Drell, P.; Doniach, S.; Downing, K. H.; Earnest, T. et al. The linac coherent light source single particle imaging road map. *Struct. Dyn.* **2015**, *2*, 041701.