HZDR
HELMHOLTZ ZENTRUM
DRESDEN ROSSENDORF

Lossless and Lossy Compression for Photon Science

Felicita Purnama Dewi Gernhardt, Peter Steinbach[1]

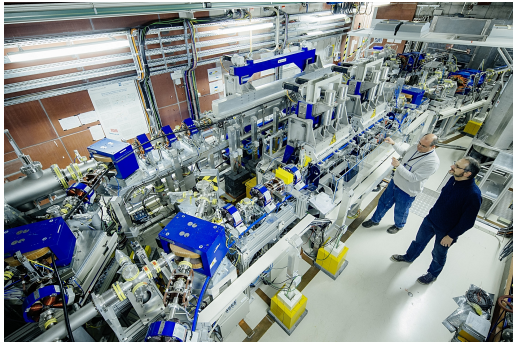
[1] *Helmholtz-Zentrum Dresden-Rossendorf, Department of Information Services and Computing*
mailto:p.steinbach@hzdr.de, https://www.hzdr.de/fwcc

September 21, 2023

CC-BY 4.0 Licensing

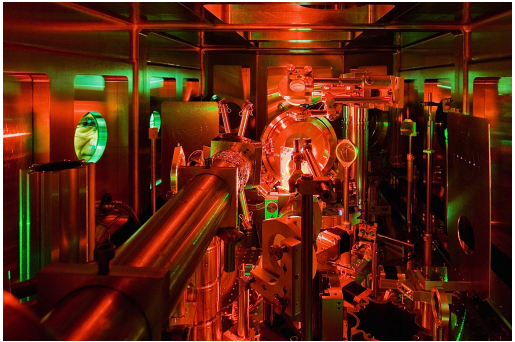
Introduction: Photon Science

Free Electron Laser at ELBE



(Copyright HZDR/Oliver Killig)

Experimental Chamber of DRACO laser



(Copyright HZDR/Jürgen Lösel)

high fidelity experiments, vast topic reach (physics, materials, life sciences, chemistry), diverse science community

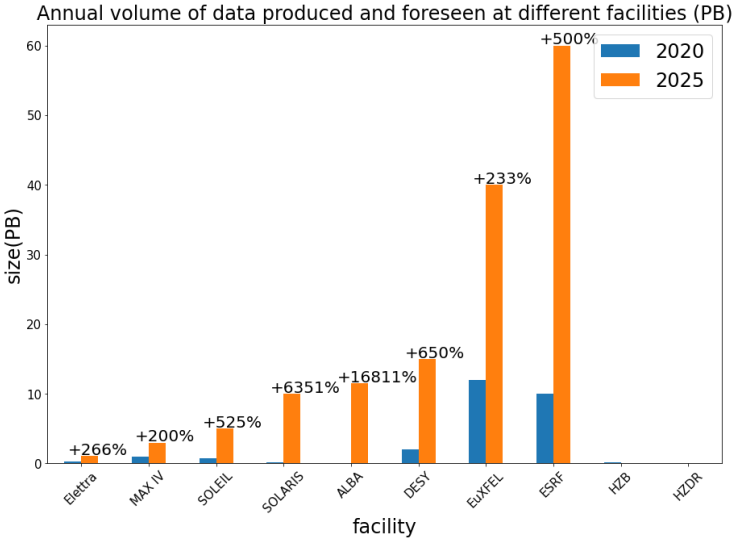
Photon Science in Europe

- 356 beamlines in 24 facilities (see wayforlight.eu)
- mode of operation at beamlines:

1. scientists apply for (limited) beamtime
2. plan/prototype for experiments (custom, standardized)
3. intense research during beamtime (24/7)
4. analysis period afterwards

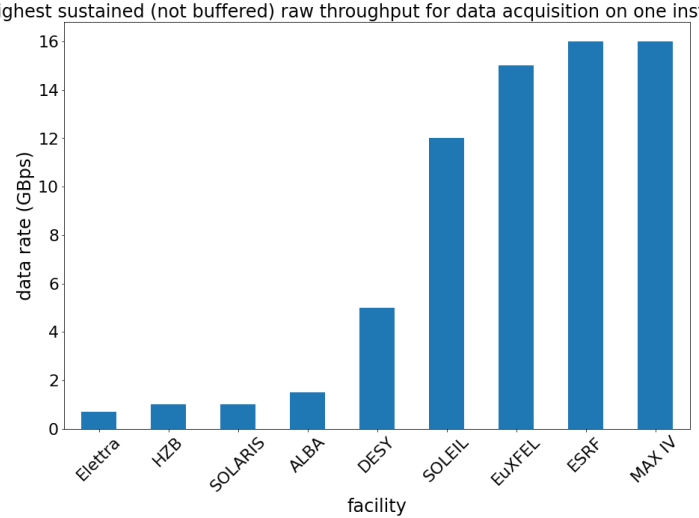
SESAME LENS Canadian Light Source Centre canadien de rayonnement synchrotron ESUO EUROPEAN SYNCHROTRON AND FEL USER ORGANISATION

Data Volumes in Photon Science



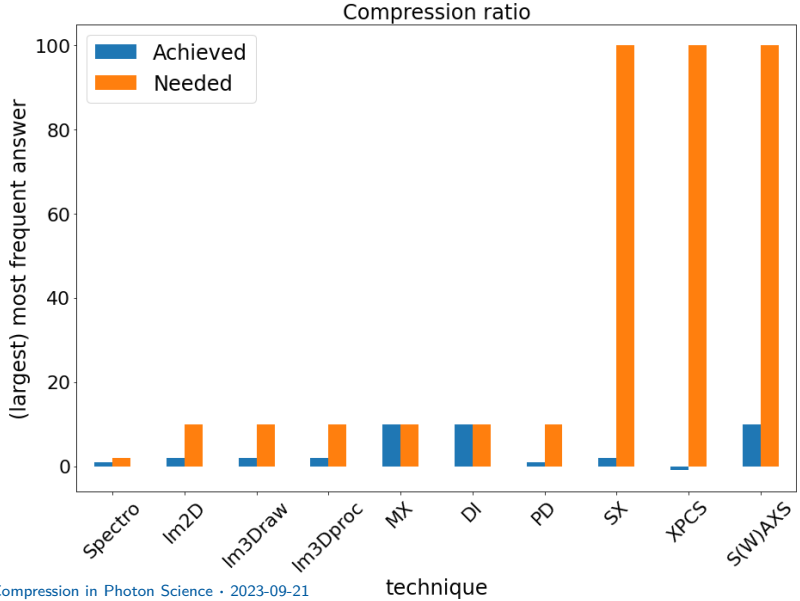
(source: LEAPS-INNOV WP7.2 report, survey among participating centers, 2020)

Data Rates in Photon Science



(source: LEAPS-INNOV WP7.2 report, survey among participating centers, 2020)

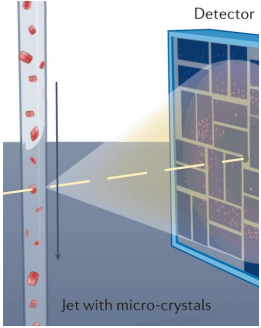
Present and Future



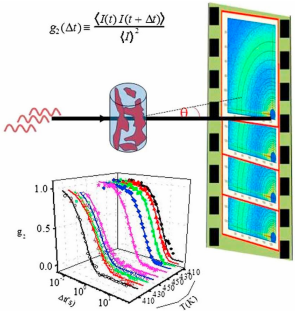
(source: LEAPS-INNOV WP7.2 report, survey among participating centers, 2020)

The Diabolic Three

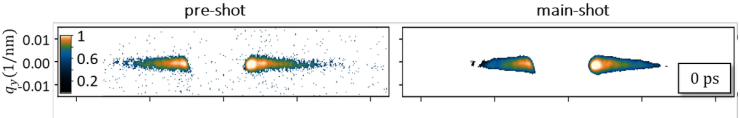
Serial Crystolography
(SX, [Barends et al., 2022])



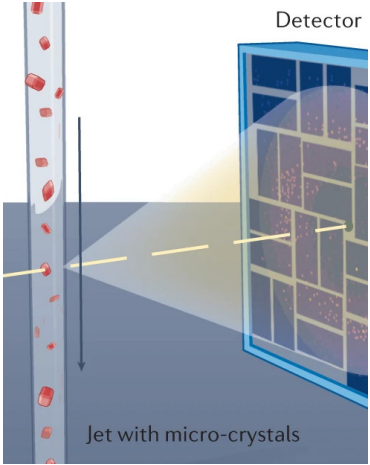
X Ray Photon Correlation Spectroscopy
(XPCS, [Nogales and Fluerasu, 2016])



Small-Angle X-Ray Scattering
(SAXS, [Kluge et al., 2023])

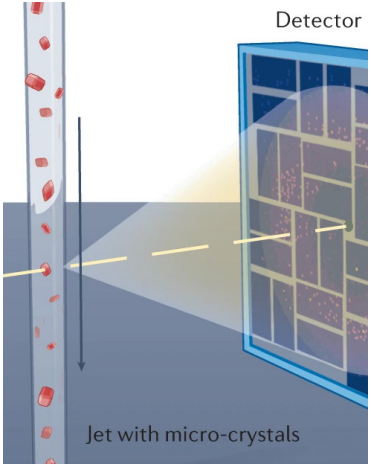


Common Observations



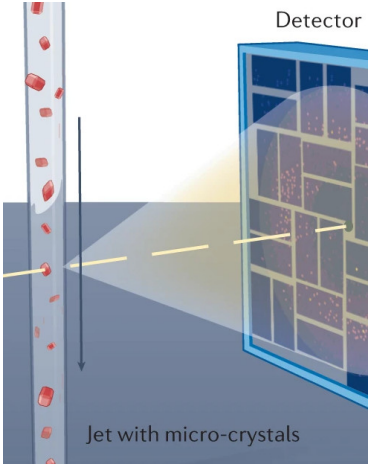
- data rates are ever increasing ($10^{14-15} b/s$ per experiment)

Common Observations



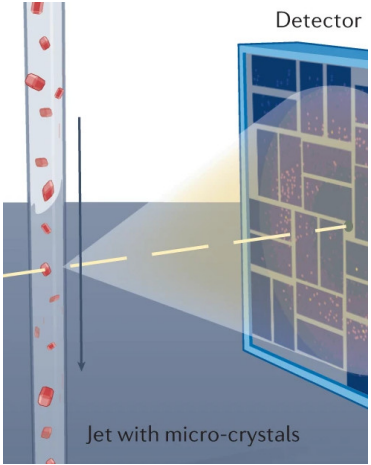
- data rates are ever increasing ($10^{14-15} b/s$ per experiment)
- data needs to be processed by scientists (integration into analysis software)

Common Observations



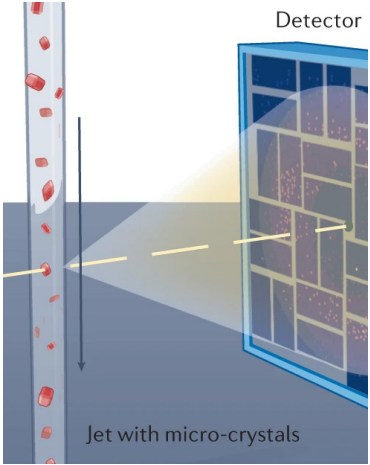
- data rates are ever increasing ($10^{14-15} b/s$ per experiment)
- data needs to be processed by scientists (integration into analysis software)
- most data:
  2D (image-like), 2D+t (video-like)
  3D, 3D+t

Common Observations



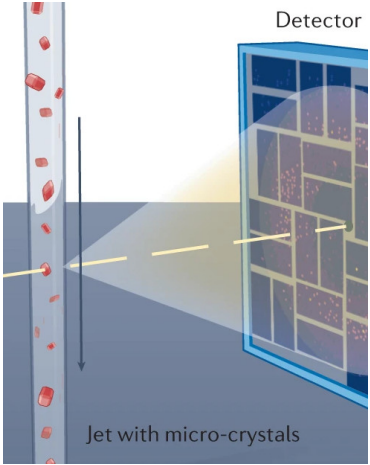
- data rates are ever increasing ($10^{14-15} b/s$ per experiment)
- data needs to be processed by scientists (integration into analysis software)
- most data:
  2D (image-like), 2D+t (video-like)
  3D, 3D+t
- pipelines are experiment specific:

Common Observations



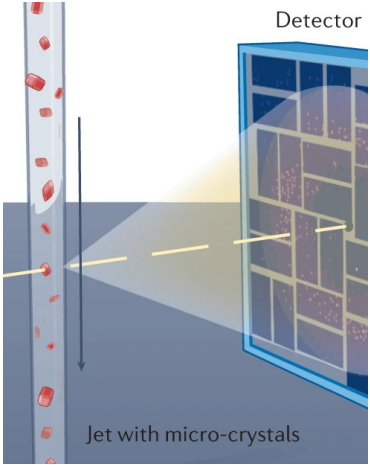
- data rates are ever increasing ($10^{14-15} b/s$ per experiment)
- data needs to be processed by scientists (integration into analysis software)
- most data:
  2D (image-like), 2D+t (video-like)
  3D, 3D+t
- pipelines are experiment specific:
  - veto as early as possible

Common Observations



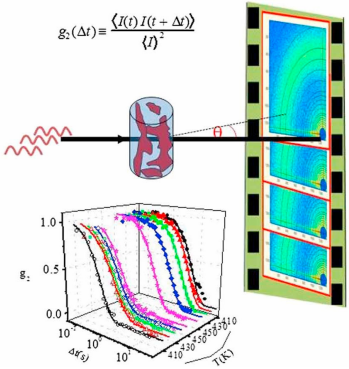
- data rates are ever increasing ($10^{14-15} b/s$ per experiment)
- data needs to be processed by scientists (integration into analysis software)
- most data:
  2D (image-like), 2D+t (video-like)
  3D, 3D+t
- pipelines are experiment specific:
  - veto as early as possible
  - denoise + compress

Common Observations



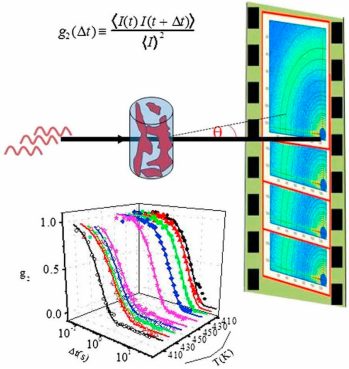
- data rates are ever increasing ($10^{14-15} b/s$ per experiment)
- data needs to be processed by scientists (integration into analysis software)
- most data:
  2D (image-like), 2D+t (video-like)
  3D, 3D+t
- pipelines are experiment specific:
  - veto as early as possible
  - denoise + compress
  - reconstruct as early as possible

Uncommon Observations



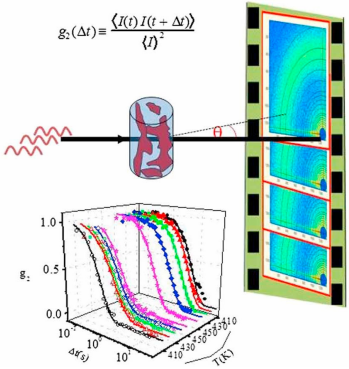
- signal processing before compression can become key
  (denoising with AI, quantisation, blocking, etc)

Uncommon Observations



- signal processing before compression can become key
  (denoising with AI, quantisation, blocking, etc)
- lossless compression has limits $c_r = 2 - 2.5$
  (with btune possibly $c_r = 3 - 4$)

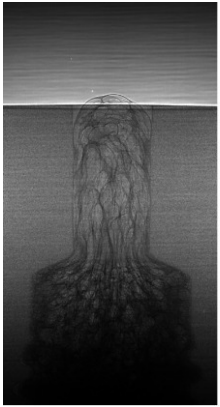
Uncommon Observations

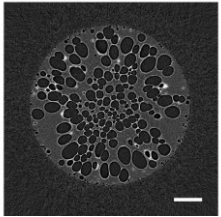


- signal processing before compression can become key
  (denoising with AI, quantisation, blocking, etc)
- lossless compression has limits $c_r = 2 - 2.5$
  (with btune possibly $c_r = 3 - 4$)
- sometimes simple filters help
  (dictionary coding when only unique values are stored, currently missing from hdf5 plugins)

Lossy Compression: Are metrics enough?

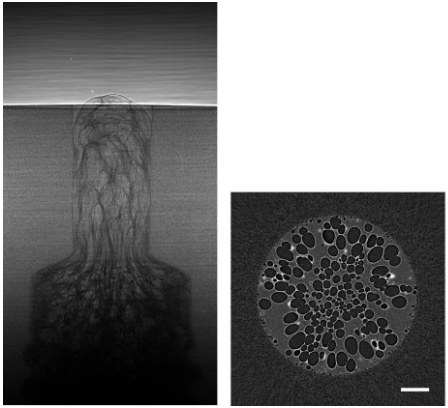
X-Ray Tomography dataset on evolving magma
[Pistone et al., 2021]



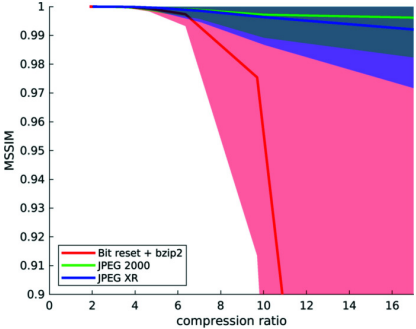


Lossy Compression: Are metrics enough?

X-Ray Tomography dataset on evolving magma
[Pistone et al., 2021]

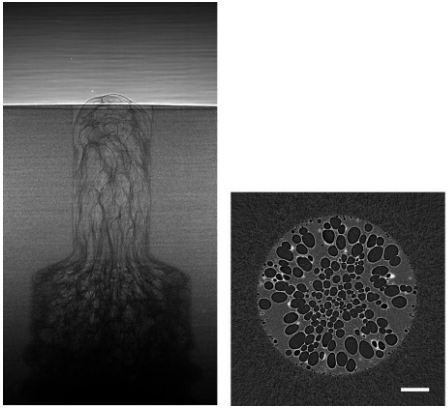


Compression Working Point Scan
[Marone et al., 2020]

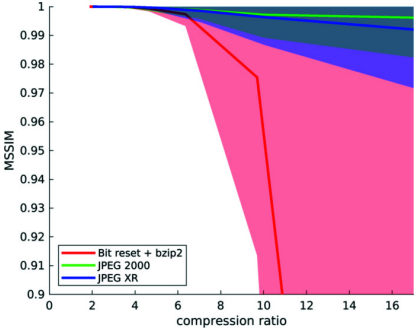


Lossy Compression: Are metrics enough?

X-Ray Tomography dataset on evolving magma
[Pistone et al., 2021]

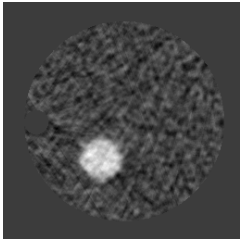


Compression Working Point Scan
[Marone et al., 2020]



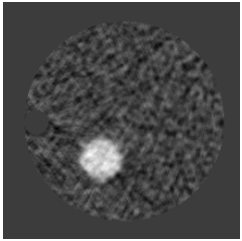
Which working point to choose? What is the impact on (downstream) science?

End-to-End Lossy Compression Workflow

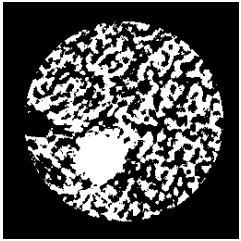


X-Ray Tomography ROFEX,
1 timepoit $= 15000 \times 256 \times 256$,
`float32`, 3750 MB

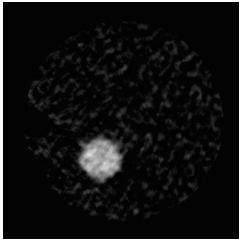
End-to-End Lossy Compression Workflow



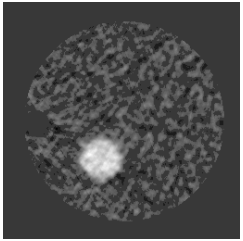
X-Ray Tomography ROFEX,
1 timepoit $= 15000 \times 256 \times 256$,
`float32`, 3750 MB



Otsu-Thresholding

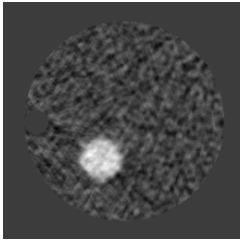


Quantise `float32` to `uint8`

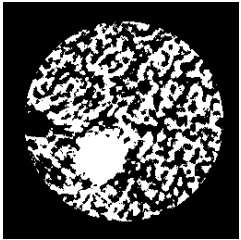


zip compress, 214 MB, $c_r \approx 17.5$

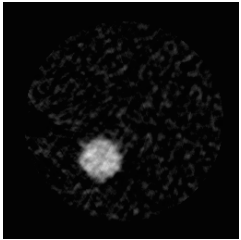
End-to-End Lossy Compression Workflow



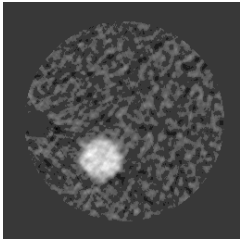
X-Ray Tomography ROFEX, 1 timepoit $= 15000 \times 256 \times 256$, `float32`, 3750 MB



Otsu-Thresholding



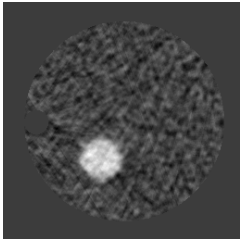
Quantise `float32` to `uint8`



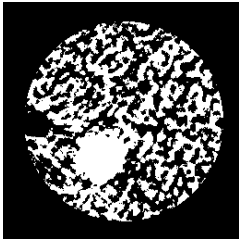
zip compress, 214 MB, $c_r \approx 17.5$

Does lossy compression impact science output?

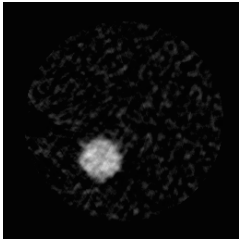
End-to-End Lossy Compression Workflow



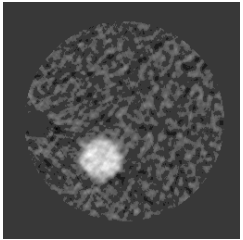
X-Ray Tomography ROFEX, 1 timepoit $= 15000 \times 256 \times 256$, `float32`, 3750 MB



Otsu-Thresholding

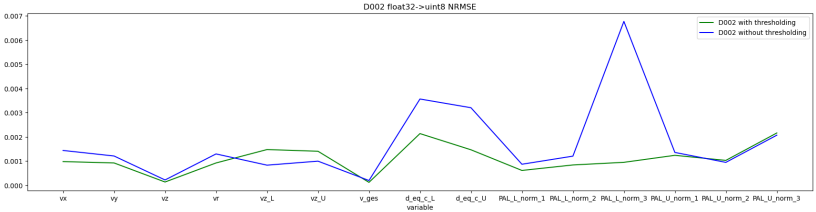


Quantise `float32` to `uint8`



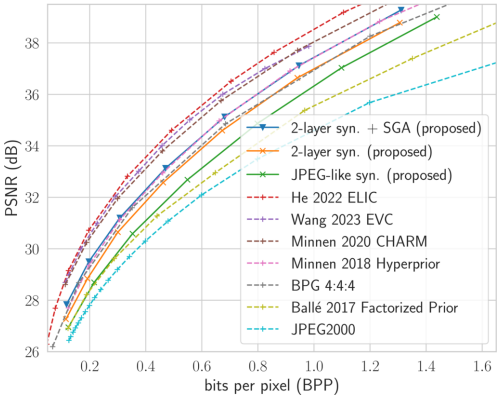
zip compress, 214 MB, $c_r \approx 17.5$

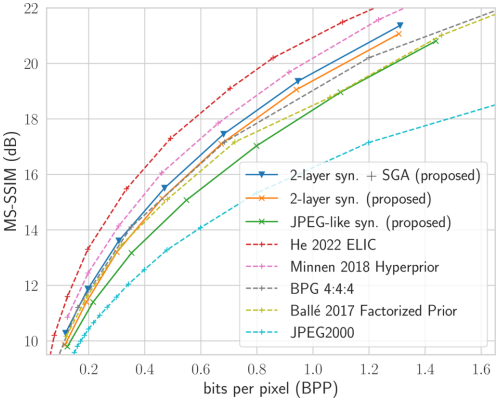
Does lossy compression impact science output?



New Kids on the Block: Neural Compressors

Rate-Distortion performance on Kodak [Yang and Mandt, 2023]





New challenges: Store model, sampling distribution and encoded data!

Summary

- high bandwidth detectors and simulations proliferate
- photon science is a diverse field with respect to data generation bandwidths
- lossless compression remains a first solid choice
  (albeit with humble compression ratios)
- scalable and reproducbile signal processing pipelines required for lossy compression
  (important: end-to-end quality assurance to sustain scientific outputs)

Thank you for your attention!
Happy to hear your thoughts, feedback, questions and concerns!
Or reach out by , , !

References (I)

Thomas RM Barends, Benjamin Stauch, Vadim Cherezov, and Ilme Schlichting. Serial femtosecond crystallography. *Nature Reviews Methods Primers*, 2(1):59, 2022. doi: https://doi.org/10.1038/s43586-022-00141-7.

Thomas Kluge, Michael Bussmann, Eric Galtier, Siegfried Glenzer, Jörg Grenzer, Christian Gutt, Nicholas J. Hartley, Lingen Huang, Alejandro Laso Garcia, Hae Ja Lee, Emma E. McBride, Josefine Metzkes-Ng, Motoaki Nakatsutsumi, Inhyuk Nam, Alexander Pelka, Irene Prencipe, Lisa Randolph, Martin Rehwald, Christian Rödel, Melanie Rödel, Toma Toncian, Long Yang, Karl Zeil, Ulrich Schramm, and Thomas E. Cowan. Probing the dynamics of solid density micro-wire targets after ultra-intense laser irradiation using a free-electron laser, 2023.

Federica Marone, Jakob Vogel, and Marco Stampanoni. Impact of lossy compression of X-ray projections onto reconstructed tomographic slices. *Journal of Synchrotron Radiation*, 27 (5):1326–1338, Sep 2020. doi: 10.1107/S1600577520007353. URL https://doi.org/10.1107/S1600577520007353.

References (II)

Aurora Nogales and Andrei Fluerasu. X ray photon correlation spectroscopy for the study of polymer dynamics. *European Polymer Journal*, 81:494–504, 2016. ISSN 0014-3057. doi: https://doi.org/10.1016/j.eurpolymj.2016.03.032. URL https://www.sciencedirect.com/science/article/pii/S001430571630146X.

Mattia Pistone, Julie L. Fife, Nicola Tisato, Luca Caricchi, Eric Reusser, Peter Ulmer, Kevin Mader, and Federica Marone. Seismic attenuation during magma vesiculation: A combination of laboratory constraints and modeling. *Geophysical Research Letters*, 48(8): e2020GL092315, 2021. doi: https://doi.org/10.1029/2020GL092315. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2020GL092315. e2020GL092315 2020GL092315.

Yibo Yang and Stephan Mandt. Asymmetrically-powered neural image compression with shallow decoders, 2023.