# Classifying Petabytes of varying structural motifs in supercooled water

**Ervin S.H. Chia** [1] **Filipe R.N.C. Maia** [2] **Carl Caleman** [3] **Nicusor Timneanu** [3] **Jonas A. Sellberg** [4] **Andrew V. Martin** [5] **N. Duane Loh** [1]

[1]*National University of Singapore, Department of Physics* [2]*Uppsala University, Department of Cell and Molecular Biology* [3]*Uppsala University, Department of Physics and Astronomy* [4]*KTH Royal Institute of Technology, Department of Applied Physics* [5]*RMIT University, School of Science, STEM College*. Correspondence to: N. Duane Loh duaneloh@nus.edu.sg.

## 1. Introduction

Water's transition to ice impacts life across many scales, yet the structural motifs governing ice nucleation remain poorly understood. These motifs are hypothesised to have short lifetimes and span across multiple length scales, from Angstroms to micrometres. Using MHz X-ray Free Electron Laser (XFEL) serial imaging, supercooled water droplets were probed near instantaneously, obtaining time-resolved information from snapshots effectively frozen in time. Up to 580 million femtosecond-resolved diffraction patterns (2.3 PB imaging data) across different experimental configurations were collected to identify and study these motifs: the largest dataset collected thus far, to our knowledge, to study structural motifs in water.

Inferring the unknown structural motifs in supercooled water is challenging for several reasons. First, despite the bright XFEL pulses, each measurement is still noisy because of water's small scattering cross-section. Second, each droplet will contain multiple structurally distinct motifs.

Third, each of these motifs have random latent parameters(e.g. crystal orientation, size). These latent parameters confound the classification of many potential motifs by only their diffraction features (i.e. two different features can arise either from distinct motifs or the same motif presented differently).

Classifying these diverse structures is not only computationally expensive but also especially challenging because they are neither large crystals with well-defined structures nor structurally homogeneous particles, for which extremely robust methods have been developed[1, 2].

Crystalline ice and liquid water have their own distinct diffraction features arising from their structure. These signatures are highlighted in an exemplary diffraction pattern in Figure 1. However, many patterns exhibit a collection of expanded and extruded blobs that represent small and repeating molecular motifs, but are difficult to identify and classify.

In total, 77% of the dataset was automatically classified in this manner, with only validation checks needed on cluster averages per run.

## 2. Approach

Classifying these structural motifs requires a hierarchical approach where we first classify the diffraction patterns that contain them into two categories:

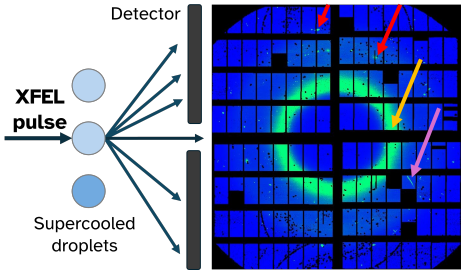


Fig. 1: Schematic of the experimental setup. The scattered beam arrives at the detector, resulting in the diffraction pattern observed. Different structures form different diffraction features. Red arrows indicate Bragg scattering, arising from crystalline ice. The diffuse ring indicated in orange arises from the isotropic random distribution of water molecules in the liquid state. Smears and blobs, such as the one highlighted in pink, points to more exotic arrangements of molecules.

those that contain structural motifs (hits), and those that do not (blanks). However, this pattern classification comprises three challenges: (1) the high dimensionality of each pattern; (2) a large number of patterns; (3) the myriad of possible and unknown motifs in each pattern with latent parameters. To overcome the first two challenges, the dimensionality of each pattern has to be reduced.

We found that a simple coarse-graining of scattering signal in reciprocal (i.e., momentum) space was effective in preserving key diffraction features of structural motifs of individual droplets. Two different levels of coarse graining lead to either a 64-fold ($X_{\text{med}} \in \mathcal{R}^{16384}$) or 2048-fold reduction ($X_{\text{low}} \in \mathcal{R}^{256}$) of the dimensionality of each raw measurement (left columns of Figure 2).

The third challenge, further classifying hits according to their structural motifs, is trickier. Standard diffraction peak-finding algorithms used in crystallography fail here because each of our measurement contains a large variety of complex structures with latent parameters. Manual labelling of vector quantizing angular averages of individual patterns was aambiguous,as these averages were insensitive to weak and diffuse signals from the important small crystals, leading to high false negative rates.

These coarse-grained feature vectors $X_{\text{low}}$ were now small enough for efficient Principal Component Analysis (PCA) to reveal which features were statisti-

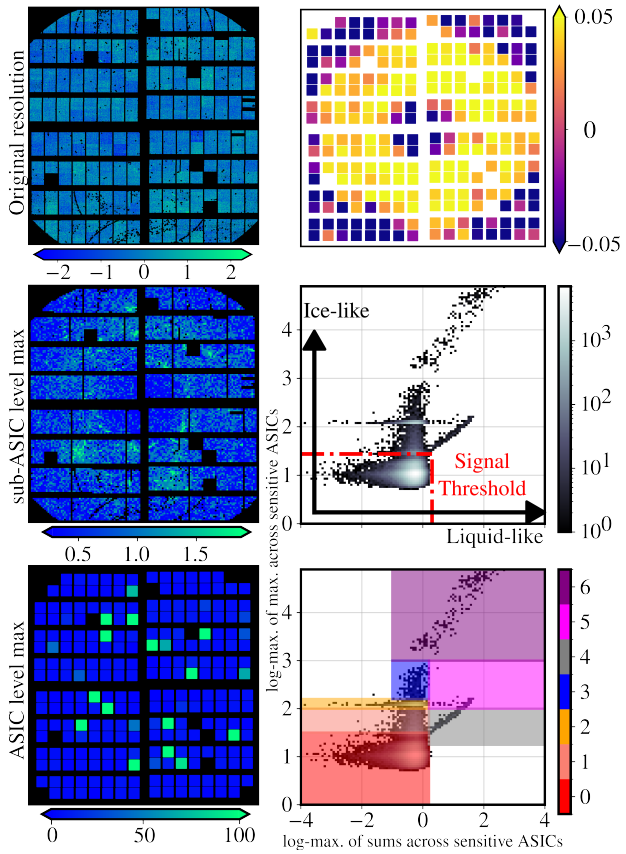


Fig. 2: Left column: Diffraction patterns at increasing levels of coarse-graining. Top right: First PCA basis (99.7% explained variance) of the entire dataset at the maximum coarse-graining, showing co-fluctuating detector regions. Middle right: Distribution of patterns in one run, re-parametrised by PCA-informed detector maximum and sums. Bottom right: Initial decision boundaries based on structures in this re-parametrised space.

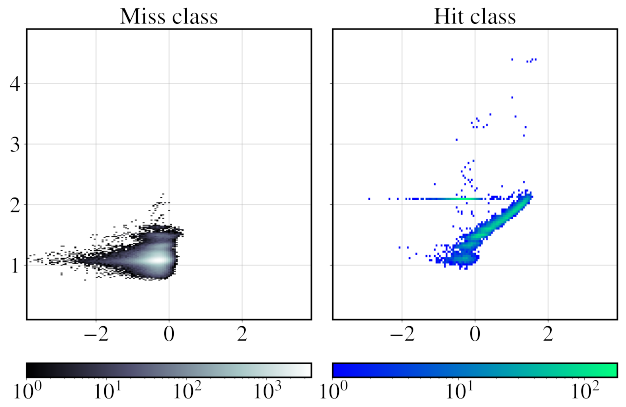


Fig. 3: Binary classification by MLP, trained on the clusters found in UMAP space. The high AMI score indicates a sufficiently trained MLP that maps the binary distinction on the UMAP space to coarse-grained diffraction patterns. Overlapping clusters in this 2D space highlights the actual decision boundary being larger than 2 dimensions.

cally significant (top right panel in Figure 2). These features were further summarized to reveal two non-linear feature vectors ($y_{\text{ice-like}}$ and $x_{\text{liquid-like}}$). Interpreting the PCA basis as an attention map for important ASICs, the maximum of sums and maximum of maximum features across these identified ASICs were found to be sensitive to the categories of structural motifs (i.e., blanks, liquid water, ice, or ice and liquid water) present in each measurement (right middle and bottom panels in Figure 2).

## 3. Validation

These labels were first physically validated. In the liquid-water label, the average radial profile for different droplet temperatures were computed with post-corrections. Patterns with this label correctly reproduced the trend of signature $q$-peaks in supercooled liquid water [3](Figure A1), observed in earlier experiments.

However, the classification does not generalise well to different runs. Differences in experimental parameters shift the distribution of points in the parametrised space, limiting the suitability of fixed decision boundaries. A better classification pipeline with more information is necessary to account for these variations. To encapsulate more information compactly, a non-linear dimensionality reduction method, UMAP[4], was used. By embedding the compact 256-dimensional coarse-grained features into 2 dimensions, clearer class boundaries can be identified, as shown in A2.

A binary classification, obtained by a density-based clustering algorithm DBSCAN [5] on the UMAP embedding, was used to train a multi-layer perceptron (MLP) to refine the hit-blank decision boundary. The adjusted Mutual Information between the DBSCAN clusters and MLP results was 0.9198. When visualised, the new decision boundary was revealed to be indistinguishable in 2 dimensions (Figure 3).

## 4. Conclusion

While coarse-grained features worked well for dimensionality reduction and early classification, they were unable to resolve fine details of structural motifs. However, it allows us to focus our attention on a subset of the data that is expected to contain the motifs of interest. A complementary method to extract and classify motifs from their complex diffraction signatures will follow.

This work describes a coarse-graining framework for classifying large numbers of diffraction patterns. Although neural networks can classify pre-defined categories of diffraction patterns[6], generalising for unseen patterns is a challenge. The right approach to further classifying motifs within diffraction hits remains an open challenge.

## References


[1] H. Olof Jönsson, Carl Caleman, Jakob Andreasson, and Nicuşor Tîmneanu. Hit detection in serial femtosecond crystallography using X-ray spectroscopy of plasma emission. *IUCrJ*, 4(6):778–784, Nov 2017.

[2] K. Ayyer, T.-Y. Lan, V. Elser, and N. D. Loh. Dragonfly: an implementation of the expand-maximize-compress algorithm for single-particle imaging. *Journal of Applied Crystallography*, 49(4):1320–1335, Aug 2016.

[3] Niloofar Esmaeildoost, Harshad Pathak, Alexander Späh, Thomas J. Lane, Kyung Hwan Kim, Cheolhee Yang, Katrin Amann-Winkel, Marjorie Ladd-Parada, Fivos Perakis, Jayanath Koliyadu, Alexander R. Oggenfuss, Philip J. M. Johnson, Yunpei Deng, Serhane Zerdane, Roman Mankowsky, Paul Beaud, Henrik T. Lemke, Anders Nilsson, and Jonas A. Sellberg. Anomalous temperature dependence of the experimental x-ray structure factor of supercooled water. *J. Chem. Phys.*, 155(21), December 2021.

[4] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction, 2020.

[5] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, KDD'96, page 226–231. AAAI Press, 1996.

[6] D. Assalauova, A. Ignatenko, F. Isensee, D. Trofimova, and I. A. Vartanyants. Classification of diffraction patterns using a convolutional neural network in single-particle-imaging experiments performed at X-ray free-electron lasers. *Journal of Applied Crystallography*, 55(Pt 3):444, 4 2022.


## Appendix A. Validating and refining the decision boundaries

The radial distribution of the scattering ring due to liquid structures, in this case water, encodes inter-molecular distance information. These intermolecular distances are expected to vary with temperature, as the molecules pack closer or stay further apart. In supercooled water, the shifting of two peaks, indicative of the most likely nearest-neighbour and next-nearest-neighbour distance, was the most significant observation in earlier experiments. This trend was successfully recreated here when the appropriate detector corrections were applied on frames with the coarse label 4 (grey), which were images that corresponded to the strongest liquid signature and weakest ice signature. By stratifying the statistics by the nozzle

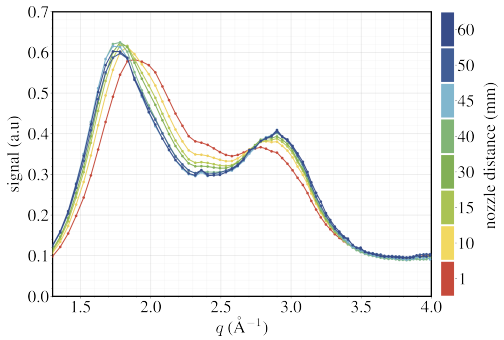


Fig. A1: Radial profiles of average liquid water scattering $I(q)$, stratified by nozzle distance, which controls droplet temperatures. The peak shifts agree with earlier experiments.

distance, which in turn controls the temperature of the droplet, this trend was re-created, validating our physical interpretations.

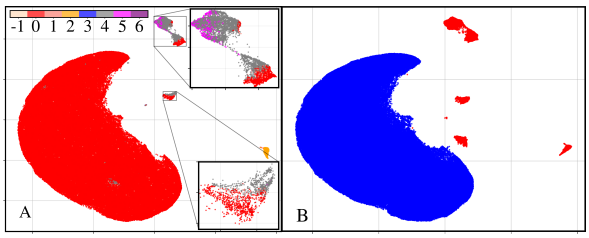


Fig. A2: UMAP embeddings of reduced representations, coloured by the preliminary classification labels obtained separately. **A**: All images from a single run. **B**: UMAP distribution re-clustered with DBSCAN. The bulk of the distribution were set re-classified as blanks (blue), while the rest were considered hits (red).

To refine the decision boundaries that were manually drawn on the coarse-grained representation space (liquid-like and ice-like), each diffraction pattern, on account of its high compressibility from PCA, was represented as PCA weights from both the sum and maximum coarse-graining. This highly compact representation ($D = 10$) enabled UMAP across many measurements. This method was suitable for images within a single run $N \simeq 10^6$, but not for the entire dataset.

To refine the decision boundaries, we exploited the knowledge of the low hit-rates to identify the cluster with the largest population as blanks with DBSCAN (Figure A2 B). With a new set of binary classifications, a 5-layered MLP was trained to map images for hit classification. In this manner, a more nuanced decision boundary can be obtained.


## Acknowledgments

EC thanks the EuXFEL SPB beamline and IT & Data Management groups, and the NUS Centre of Bio-Imaging Sciences IT for their expert support and computational resources, and is grateful for funding from the NUS Graduate Tutor Scheme.