## Assessing the severity of ice contamination in processed data sets with a combination of statistical tools and machine learning in AUSPEX

Yunyun Gao[1], Kristopher Knolte[1], Andrea Thorn[1]

[1]Institut für Nanostruktur und Festkörperphysik, Universität Hamburg, Luruper Chaussee 149 (Bldg. 610 - HARBOR), Germany

Ice ring contamination is present in roughly 21% of crystallographic data sets in the world-wide protein data bank, posing a serious problem for the quality and determination of macromolecular structures [1]. The automatic identification of these Debye–Scherrer rings after data processing and merging is difficult, hence we explore two automatic approaches: statistical testing and machine learning.

The statistical method is the Auspex:IceFinderScore [1], which analyses the behaviour of the standardized mean intensity in resolution bins, essentially measuring the departure from a uniform intensity distribution. It can be loosely interpreted as a Z-score, which is sensitive to resolution ranges that exhibit sudden sharp changes in the intensity distribution, and gives the severity of the deviation from the expected distribution. However, under the null hypothesis, the threshold set before performing a test is not absolutely objective and hence undermines both the accuracy and the specificity of the score.

On the other hand, data-driven machine learning research based on convolutional neural networks enables Auspex:Helcaraxe [2]. Helcaraxe outperforms all previous methods in binary classification (ie, ice ring or no ice ring); in our tests, the accuracy of the ice crystal artefact detection has reached an unprecedented 96% compared to humans analysing intensity distribution plots visually. While the binary classification guarantees that the output of the neural network is explicitly interpretable, the drawback is the Helcaraxe network prediction has no clear indication how severe the ice contamination is.

Both methods will be soon implemented in AUSEPX, which is available in ISPyB from 2022, on www.auspex.de and as part of the CCP4 software suite [3].

In order to combine the strengths of both IceFinderScore and Helcaraxe, the Helcaraxe prediction array on the full dataset is used as a guide to dynamically adjust the significance level of the local IceFinderScores. The new assessment shows quantitatively, at the potential ice ring ranges, how severe the intensity observations are affected by the presence of ice rings (Fig. 1), guiding the automatic processing and enabling users to get the best out of their data.

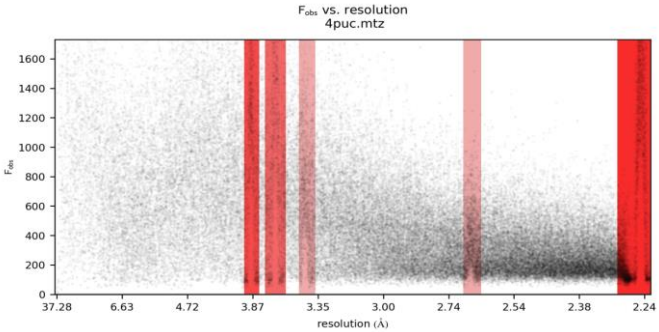


Fig 1. An example of the AUSPEX plot. The ice artefacts are flagged in red shades. The occupancy of the shades suggests the severity of the local ice ring contamination.

[1] Thorn, A., Parkhurst, J., Emsley, P., Nicholls, R. A., Vollmar, M., Evans, G. & Murshudov, G. N. (2017). Acta.Cryst. D, 73, 729–737

[2] Nolte K., Gao, Y., Emsley, Staeb, S., Kollmansberger P., & Thorn, A. (2022). Acta. Cryst. D. [in print]

[3] M. D. Winn et al. (2011) Acta. Cryst. D, 67, 235-242