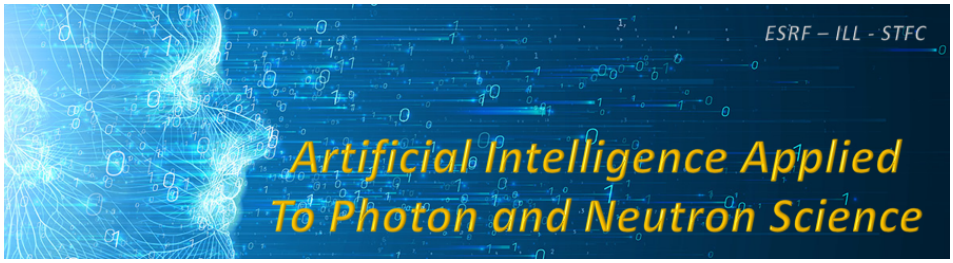

Contribution ID: **41** Type: **not specified**

Machine learning applications for Small Angle X-ray Scattering data collection and analysis at EMBL-Hamburg

Tuesday, 12 November 2019 15:15 (25 minutes)

In recent years, machine learning and artificial intelligence rapidly gained popularity in many fields of industry and research, particularly as a tool capable of extracting information from amounts of data often too large to analyze manually. Small Angle X-Ray Scattering (SAXS) of biological macromolecules in solution is routinely being used to evaluate the structural parameters and low resolution shapes of the specimen under study. Here, the application of machine-learning methods seems to be a natural extension, maybe even evolution, of established analysis methods.

At EMBL in Hamburg, two applications of machine learning methods have previously been developed: firstly, based on 450.000 scattering patterns predicted from geometrical objects with uniform density (BODIES; Konarev et al., 2003) and 550.000 scattering patterns from random chains (EOM; Tria et al., 2015), a k-Nearest-Neighbor (kNN) learner is utilized to reliably distinguish between compact objects with or without cavities, extended and flat objects, random chains, as well as “unrecognizable data”, i.e. anything not suitable for the other categories. Secondly, based on 150.000 atomic structures from the PDB (Berman et al. 2000) and their calculated scattering patterns (CRYSOL; Svergun et al., 1995), a similar kNN learner has been to evaluate the maximum dimension (Dmax) and Molecular Weight (MW) from the input data. In both cases the same data transformation into a reduced feature space is applied: the theoretical scattering data, which may be provided on any scale and with any angular spacing is transformed to dimensionless Kratky scale (Durand et al., 2010) and subsequently integrated up to sRg = 3, 4 and 5, respectively. The results of this integration are used as input features for learning. To evaluate the performance, the available data has been randomly split into training and cross-validation data sets. Performance of shape classification is rated at 99% for both F1-score an Matthews Correlation Coefficient (Matthews, 1975), with all categories exceeding 90% of one-vs-all classification accuracy. Further, about 90% of continuous Dmax and MW estimates are within 10% of the expected value obtained from the corresponding databank-entry (Franke, 2018).

In addition, recent work aims to employ classical neural networks and/or deep-learning to more challenging tasks. In particular, extensions of the previous work to determine the radius of gyration (Rg, Dmax and MW) from the experimental data directly. Even further, early work suggests that it is possible to develop deep-learning networks to retrieve feasible approximations of the inverse Fourier Transform of the experimental data. A preview version of this has been implemented as a public web, providing the possibility to inspect and download the results (https://dara.embl-hamburg.de/gnnom.php).

This work was supported by the Bundesministerium für Bildung und Forschung project BIOSCAT, Grant 05K12YE1, and by the European Commission FP7, BioStruct-X grant 283570 and iNext grant 653706.

Presenter: Dr FRANKE, Daniel (EMBL)

Session Classification: Afternoon 1