Supervised classification methods for flash X-ray single particle diffraction imaging

JING LIU,1,2 GIJS VAN DER SCHOT,3 AND STEFAN ENGBLOM1,*

1Division of Scientific Computing, Department of Information Technology, Uppsala University, SE-751 05 Uppsala, Sweden
2Laboratory of Molecular Biophysics, Department of Cell and Molecular Biology, Uppsala University, SE-751 24 Uppsala, Sweden
3Cryo-Electron Microscopy, Bijvoet Center for Biomolecular Research, Utrecht University, 3584 CH Utrecht, The Netherlands
*stefane@it.uu.se

Abstract: Current Flash X-ray single-particle diffraction Imaging (FXI) experiments, which operate on modern X-ray Free Electron Lasers (XFELs), can record millions of interpretable diffraction patterns from individual biomolecules per day. Due to the practical limitations with the FXI technology, those patterns will to a varying degree include scatterings from contaminated samples. Also, the heterogeneity of the sample biomolecules is unavoidable and complicates data processing. Reducing the data volumes and selecting high-quality single-molecule patterns are therefore critical steps in the experimental setup. In this paper, we present two supervised template-based learning methods for classifying FXI patterns. Our Eigen-Image and Log-Likelihood classifier can find the best-matched template for a single-molecule pattern within a few milliseconds. It is also straightforward to parallelize them so as to match the XFEL repetition rate fully, thereby enabling processing at site. The methods perform in a stable way on various kinds of synthetic data. As a practical example we tested our methods on a real mimivirus dataset, obtaining a convincing classification accuracy of 0.9.

© 2019 Optical Society of America under the terms of the OSA Open Access Publishing Agreement

1. Introduction

Modern X-ray Free Electron Laser (XFEL) technology has provided the opportunity for exploring biological structures from individual biological particles, rather than relying on crystallization-based technologies. It is therefore potentially possible to investigate biomolecules or biological processes that are intrinsically dynamic. XFELs produce X-ray pulses shorter than 50 femtosecond (fs), which are 10^9 times more brilliant than the radiation produced in conventional synchrotrons. The ultra-short and extremely bright X-ray pulses outrun the radiation damage and allow the recording of sufficiently strong and interpretable 2-dimensional (2D) diffraction patterns from single biological particles [1, 2]. This principle is called diffract-and-destroy and has been shown to be successful for particles as large as small cells, and down to viruses smaller than 50 nanometers (nm) [3–7].

Another feature of XFELs is their high repetition rates. The Linac Coherent Light Source (LCLS) [8] operates at 120 Hz and can produce over 400,000 diffraction patterns per hour, i.e., more than 1.6 TB per hour or 38 TB per day. The massive volume of data makes manual classification of diffraction patterns impractical. The challenge is much more severe in the newest facility — the European XFEL [9], which operates at up to 27,000 Hz and can store more than 12.6 million images per hour [10]. Ideally, all these images would originate from one single biomolecule per exposure. However, the detector also records diffracted signals from multiple scatterers such as particle clusters, buffer impurities, and contaminant materials as discussed in [5, 7].

In order to assemble the 2D diffraction patterns into 3D structures, it is essential that data...
frames are classified and that diffraction patterns originating from contaminants and multiple molecules are sorted out. In 2014, a real-time rejection method [11] was proposed to select diffraction patterns by thresholding and using Time-of-Flight spectroscopy. Other previous sorting algorithms were based on feature vectors and spectral clustering techniques [12, 13]. Diffusion maps and manifold embedding [14] have also been proposed, and have successfully classified icosahedral shaped viruses [15–18]. However, all of these methods work best after a substantial amount of data has been retrieved, and/or are in the need of a solid understanding of the raw data. This makes them less ideal in a streaming context.

In this paper, we develop two template-based classification methods for particle selection — the Eigen-Image (EI) and the Log-Likelihood (LL) method. Both methods assess the similarity between template diffraction patterns and incoming patterns by analyzing eigenvector projections and log-likelihood functions, respectively. By relying on templates none of these methods are dependent on access to the full dataset, and consequently they are suitable for real-time processing at site. With our methods, we thus aim to select single-particle diffraction patterns of homogeneous quality. Such datasets may hopefully help 3D assembling algorithms [19] to converge more quickly and improve on the final 3D resolution.

In §2, we briefly describe a typical Flash X-ray single-particle diffraction Imaging (FXI) experiment. Next, we introduce the EI and the LL method for classification in §3. Following data descriptions in §4, we perform numerical experiments to evaluate the sharpness of our classification methods in §5. A concluding discussion is found in §6.

2. Flash X-ray single particle diffraction Imaging (FXI)

For a typical FXI experiment, the diffraction data acquired is depicted graphically in Fig. 1. A stream of biological molecules is injected into the X-ray interaction region, where sample particles interact with incoming coherent X-ray pulses, resulting in a collection of diffraction patterns on the detector. This procedure is a stochastic process as the interactions between particles and X-ray pulses occur at random. Firstly, the number of particles at the interaction point is unobserved, i.e., we may obtain blank frames with only background noises, single-particle patterns, multiple-particles patterns, and frames with signals from contaminants. Secondly, the current FXI technology cannot monitor the orientations of particles, and therefore extra steps are necessary to recover the 3D structure from single-particle frames. Last but not least, the strengths of the diffraction signals vary a lot, mainly due to the stochastic nature of the XFELs and the different locations of particles in the interaction region, respectively. The relative strength of the diffraction signal is referred to as photon fluence, and we denote it by $\phi$. Typical FXI setups use digital detectors made up of individual pixels, and therefore the captured frames are pixelized. Further, some pixel counts near the center are inaccessible or overflow as a result of physical limitations and arrangements of the detector.

3. Classification Methods

Template-based methods for classifying patterns allow for identifying the class of an unlabelled pattern by searching for its best-matched template. For such methods, the collection of templates is referred to as the training dataset, an unlabelled pattern is called a testing image, and the classification procedure is referred to as the classifier. In this section, we discuss two classifiers — the Eigen-Image (EI) [20–25] and the Log-Likelihood (LL) classifier [26–28], to classify a testing diffraction pattern relying on a training dataset.

3.1. Eigen-Image (EI) Classifier

The EI method has two steps — the training and the classification step. In the training step, we train our EI classifier by projecting the training dataset to its eigenvectors. In the classification
Fig. 1. (a): A typical setup of an FXI experiment. [(b)–(e)]: Four diffraction patterns from a mimivirus taken from [4]. (b) was a blank frame which contains only background scattering. [(c) and (d)] were frames from multiple particles or with contaminants. (e) was a single-particle frame from an icosahedral-shape virus with a relatively strong signal. This is the most interesting pattern and can be used for assembling a 3D structure in later steps. All diffraction patterns are displayed in logarithmic scale.

step, we label a testing image by minimizing the distance between the eigenvector projections of the testing image and the training dataset.

Let i.i.d. template diffraction patterns $T = (T_k)_{k=1}^{M_{\text{dat}}}$ be the training dataset, consisting of $M_{\text{dat}}$ frames. Since the detector is pixelized, we denote the $k$th pattern by $T_k = (T_{ik})_{i=1}^{M_{\text{pix}}}$. To train an EI classifier, we first transfer the training dataset $T$ into the image space $A$ by the shift

$$ A = (A_k)_{k=1}^{M_{\text{dat}}} = (T_k - \bar{T})_k, \quad (1) $$

where $\bar{T}$ is the pixel average of the training dataset,

$$ \bar{T} = \frac{1}{M_{\text{dat}}} \sum_{k=1}^{M_{\text{dat}}} T_k. \quad (2) $$

Practically, the covariance matrix of $A$ (namely $AA^T$) is too large to decompose into eigenvectors (size $M_{\text{pix}}$), and therefore we factorize the matrix $A^T A$ instead (size $M_{\text{dat}}$, where usually $M_{\text{dat}} \ll M_{\text{pix}}$). Hence,

$$ A^T A = V \Lambda V^T, \quad (3) $$
where $\Lambda$ is the main diagonal matrix, whose diagonal elements are the corresponding eigenvalues, and $V$ is the matrix of eigenvectors of $A^T A$. We can now compute the eigenvectors of the covariance matrix $AA^T$ by

$$U = AV, \quad (4)$$

and $U$ is sometimes also referred to as eigenfaces [20, 25].

The eigenvector projection matrix of the image space $A$ is defined as follows:

$$\Omega = U^T A. \quad (5)$$

Using $U$ and $\Omega$, we can now classify a testing diffraction pattern $P = (P_i)_{i=1}^{M_{\text{pix}}}$, by minimizing the Euclidean distance (i.e. the $L_2$-norm) between its eigenvector projection matrix $W$ and $\Omega$,

$$\arg_k \min ||W_k - \Omega_k||_{L_2}, \quad (6)$$

where

$$W = U^T (P - \bar{T}). \quad (7)$$

### 3.2. Log-Likelihood (LL) Classifier

The LL Classifier attempts to classify a testing image by maximizing the log-likelihood function of a given probability density function. Since the photon counting procedure is assumed to obey the Poisson distribution, we can write the joint likelihood function as follows:

$$\prod_{i=1}^{M_{\text{pix}}} \mathcal{P}(P_i | T_{ik}, \phi_k) = \prod_{i=1}^{M_{\text{pix}}} \frac{(\phi_k T_{ik})^{P_i} e^{-\phi_k T_{ik}}}{P_i!} =: Q_{ik}, \quad (8)$$

where $\phi$ is the photon fluence (relative signal strength), and can be estimated by

$$\phi_k = \frac{\sum_{i=1}^{M_{\text{pix}}} P_i}{\sum_{i=1}^{M_{\text{pix}}} T_{ik}}. \quad (9)$$

The joint log-likelihood function $\mathcal{L}$ for the LL classifier is therefore

$$\log(Q_{ik}) \propto \sum_{i=1}^{M_{\text{pix}}} P_i \log(T_{ik}) + P_i \log(\phi_k) - \phi_k T_{ik} =: \mathcal{L}_k, \quad (10)$$

We can now classify the testing image $P$ by simply maximizing the joint log-likelihood function in [Eq. (10)]:

$$\arg_k \max \mathcal{L}_k = \arg_k \max \sum_{i=1}^{M_{\text{pix}}} P_i \log(T_{ik}) + P_i \log(\phi_k) - \phi_k T_{ik}. \quad (11)$$

For classifying multiple testing images, the EI method computes $U$ and $\Omega$ only once at the beginning, and we then compute [Eq. (6)] and [Eq. (7)] for each testing image. The computations needed by EI is smaller than the LL classifier, since the computational complexity of [Eq. (7)] is $O(M_{\text{pix}} \times M_{\text{pix}})$ arithmetic operations, which is the complexity of the last term alone in [Eq. (11)].
4. Data Description

The raw FXI data frames differ to varying degrees. Typically they may be sorted into hierarchical categories. On the top level, the classes can be single-particle patterns and non-single-particle patterns. Further, single-particle patterns, which typically are the most interested ones, may be classified by the particle rotations, sizes, shapes, etc. Making templates for all possible categories can be hard or even impossible, and consequently we train and evaluate our classifiers with single-particle diffraction patterns.

Most viruses have either a helical or an icosahedral capsid structure [29, 30], and icosahedral viruses are of great interests [15, 17, 31]. Therefore, we focus on FXI experiments of icosahedral particles. To illustrate our method, we used regular uniform-density icosahedrons to generate diffraction patterns via Condor [32]. For our simulations, we used a setup similar to the beam profile of the FXI mimivirus experiment [4]. More specifically, we used X-ray pulses with 1 mJ peak energy and 1 nm wavelength. We also assumed that the X-ray pulses had a circular focus of 10 µm in diameter. Further, the distance between the detector and the interaction region was 0.74 meters, and the detector itself was 960 x 960 pixels, with the size of each pixel 75 x 75 µm². Finally, a circular missing-data area of 80 pixels in diameter was set to zero.

To assess our classifiers systematically, we gradually increased the complexity of the testing dataset. With five synthetic testing datasets, we mimicked diffraction patterns of particles with noise, different fluences, and of various sizes and shapes. We also evaluated our methods for the actual mimivirus FXI data [31]. Fig. 2 illustrates two noisy icosahedral diffraction patterns at particle sizes 180 nm and 200 nm in the same particle orientation, and one spheroid diffraction pattern at size 180 nm.

(a) (b) (c)

Fig. 2. (a): A noisy diffraction pattern from a 180 nm icosahedron. (b): A noisy pattern from a 200 nm icosahedron in the same particle orientation. (c): A noiseless pattern from a 180 nm spheroid.

4.1. Homogeneous Datasets

We first simulated diffraction patterns from a regular icosahedron of diameter 180 nm. The training dataset $T$ had 290 frames, and the Euclidean distances between two arbitrary patterns were larger than or equal to 220. The first testing dataset $D$ was a noiseless homogeneous dataset, which contained $M_{\text{test}} = 1000$ noiseless icosahedral diffraction patterns. The first 290 frames were from the training dataset $T$ and were used as benchmarks. The rest 810 frames were random-orientation patterns from the same icosahedron.

Since the photon counting procedure is assumed to follow the Poisson distribution, we added Poissonian noise to $D$ for our noisy dataset $P$,

$$P_k \sim \text{Poisson}(D_k), \ k = 1, 2, \ldots, M_{\text{test}}.$$ (12)
By scaling $P$ with different fluences, we obtained our last homogeneous testing dataset — the scaled noisy dataset $F$ by

$$F_k \sim \text{Poisson}(\Phi_k D_k),$$

(13)

where $\Phi_k$ was uniformly and randomly chosen between 0.01 to 1.1,

$$\Phi_k \sim \mathcal{U}\{0.01, 1.1\}. \quad (14)$$

4.2. Heterogeneous Particle Sizes

Considering the potential size variation of viruses, we generated our testing dataset $S$ ($M_{\text{tot}} = 2000$) from uniform-density icosahedrons with randomly and uniformly chosen diameters between 150 nm and 210 nm ($\sim \mathcal{U}\{150, 210\}$). Similar to $F$, all patterns in $S$ were Poissonian with random fluences according to [Eq. (14)].

4.3. Heterogeneous Particle Shapes

To mimic heterogeneous particle shapes, the synthetic testing dataset $X$ contained diffraction patterns from both icosahedrons and spheroids. The diameters of the objects varied from 150 nm to 210 nm, with changing fluences $\Phi_k \sim \mathcal{U}\{0.01, 1.1\}$. Further, the shapes of the spheroids were also changing, as the aspect ratios of the spheroids (the ratio of the length of the minor axis to the length of the major one) were varying between 0.6 and 1. In total, the dataset $X$ contained $M_{\text{tot}} = 1200$ frames — 200 spheroidal patterns and 1000 icosahedral patterns randomly selected from $S$.

4.4. Mimivirus Dataset

To be relevant to real FXI experiments, we also classified the mimivirus dataset [4, 31], which contained 198 single mimivirus patterns. To classify this dataset, we generated a new training dataset ($T_2$) with the corresponding experimental beam profile [4], consisting of 1000 random-orientation frames of a 490 nm icosahedron. Later, we moved forward to the raw data frames from the FXI mimivirus experiments [4]. To capture non-single particles, and keep the training dataset ($T_3$) as small as possible, we only added ten spherical patterns, with particle sizes varying
from 100 nm to 1000 nm. Table 1 lists the primary parameters of all datasets.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Size (nm)</th>
<th>( M_{\text{dat}} )</th>
<th>Noise</th>
<th>Fluence ( \Psi )</th>
<th># Photons ( \times 10^5 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( T^a ) (Training)</td>
<td>180</td>
<td>290</td>
<td>N/A</td>
<td>1</td>
<td>( \approx 10 )</td>
</tr>
<tr>
<td>( D^a ) (Noiseless)</td>
<td>180</td>
<td>1000</td>
<td>N/A</td>
<td>1</td>
<td>( \approx 10 )</td>
</tr>
<tr>
<td>( P^b ) (Shot-noise)</td>
<td>180</td>
<td>1000</td>
<td>Poisson</td>
<td>1</td>
<td>( \approx 10 )</td>
</tr>
<tr>
<td>( F^b ) (Fluence-scaled)</td>
<td>180</td>
<td>1000</td>
<td>Poisson</td>
<td>( \mathcal{U}(0.01, 1.1) )</td>
<td>( [0.1, 11] )</td>
</tr>
<tr>
<td>( S^c ) (Varied-size)</td>
<td>( \mathcal{U}(150, 210) )</td>
<td>2000</td>
<td>Poisson</td>
<td>( \mathcal{U}(0.01, 1.1) )</td>
<td>( [0.1, 11] )</td>
</tr>
<tr>
<td>( X^d ) (Spheroid)</td>
<td>( \mathcal{U}(150, 210) )</td>
<td>1200</td>
<td>Poisson</td>
<td>( \mathcal{U}(0.01, 1.1) )</td>
<td>( [0.1, 11] )</td>
</tr>
<tr>
<td>( T^1_i ) (Training 1)</td>
<td>180</td>
<td>290+1</td>
<td>N/A</td>
<td>1</td>
<td>( \approx 10 )</td>
</tr>
<tr>
<td>Mimivirus( ^e )</td>
<td>( \approx 490 )</td>
<td>198</td>
<td>N/A</td>
<td>N/A</td>
<td>( [4.5, 34] )</td>
</tr>
<tr>
<td>( T^2_i ) (Training 2)</td>
<td>( \approx 490 )</td>
<td>1000</td>
<td>N/A</td>
<td>1</td>
<td>( \approx 10 )</td>
</tr>
<tr>
<td>Raw-mimi( ^b )</td>
<td>( \approx 490 )</td>
<td>50712</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>( T^3_i ) (Training 3)</td>
<td>490</td>
<td>1000+10</td>
<td>N/A</td>
<td>1</td>
<td>( \approx 10 )</td>
</tr>
</tbody>
</table>

\( ^a \) The training dataset for classifying synthetic testing datasets.
\( ^b \) Only synthetic regular icosahedral patterns were included in these datasets.
\( ^c \) \( \mathcal{U} \) is the uniform distribution.
\( ^d \) \( X \) contained 1000 random icosahedral frames from \( S \) and 200 spheroidal patterns with aspect ratio \( \sim \mathcal{U}(0.6, 1) \).
\( ^i \) \( T^1_i \) is the training dataset for classifying the dataset \( X \), which contains \( T \) and a single sphere diffraction pattern of size 180 nm. \( T^2_i \) and \( T^3_i \) are training datasets for classifying the mimivirus and the raw-mimi datasets, respectively. \( T^1_i \) contains 1000 randomly-oriented diffraction patterns of a 490 nm uniform icosahedron. \( T^3_i \) includes the same icosahedral patterns, and ten sphere patterns at the sizes \( [100, 200, \ldots, 1000] \) nm.
\( ^e \) Patterns were as used in [4,31] and of icosahedral shape.
\( ^f \) The number of patterns.
\( ^g \) The total number photons of one diffraction pattern.
\( ^h \) Part of run 92 and run 93 from the FXI mimivirus experiment. See more details about experimental runs from Table 2 of [4].

5. Experiments

We now perform numerical experiments to investigate the efficiency and the accuracy of our EI and LL classifiers. For saving memory and execution time without losing much accuracy in the classification, only the central \( 480 \times 480 \) pixels were used in the computations, and they were binned into \( 120 \times 120 \) pixels, i.e., every \( 4 \times 4 \) pixels were averaged into one pixel.

5.1. Metrics

Since the proposed methods are sensitive to the particle rotations, the best-matched pattern from the training dataset should have the closest particle rotation to the testing image. Therefore, we compare the best-matched template with the testing image taking also the particle size into consideration. Let \( \Gamma_k = (\Gamma_{ik})_{i=1}^{M_{\text{pix}}} \) be the \( k \)-th frame of the testing dataset \( \Gamma \). Let \( R = (R_{i})_{i=1}^{M_{\text{pix}}} \) be the best-matched pattern of \( \Gamma_k \) from the training dataset. The pattern distance between \( \Gamma_k \) and \( R \) is now defined as:

\[
C_k(\Gamma_k, R) = \arg \min_{s, \hat{\Phi}_k} \frac{\sum_{i=1}^{M_{\text{pix}}} (\hat{\Phi}_k U_{ik}(R, s) - \Gamma_{ik})^2}{\sum_{i=1}^{M_{\text{pix}}} (\hat{\Phi}_k U_{ik}(R, s))^2},
\]  

(15)
where $\hat{\Phi}_k$ is the estimated fluence,

$$\Phi_k = \frac{\sum_{i=1}^{M} \Gamma_{ik}}{\sum_{i=1}^{M} U_{ik}(R, s)}.$$  \hspace{1cm} (16)

Further, $U_k(R, s)$ is an interpolation (or extrapolation) method that resizes the pattern $R$ $s$ times and returns a scaled image at the same size as $\Gamma_k$. Note that, $U(R, s) = R$ for our homogeneous testing datasets $D$, $P$ and $F$, and $\Phi_k = 1$ for the first two. It is natural to set 0.5 as a threshold to determine if the testing image $\Gamma_k$ can be accepted by $R$.

Similarly, we define the fluence distance by

$$E_k = \frac{(\Phi_k - \hat{\Phi}_k)^2}{\Phi_k^2},$$ \hspace{1cm} (17)

where $\Phi_k$ is the true fluence used to generate $\Gamma_k$.

### 5.2. Homogeneous Patterns

We first tested our classifiers on the homogeneous synthetic datasets — the noiseless dataset $D$, the noisy dataset $P$, and the scaled-Poisson dataset $F$, as listed in Table 1. Since the first 290 images of the three testing datasets were modifications of the training dataset, we used them as benchmarks, and compared their average pattern distance with the distances from the remaining patterns in the datasets. As listed in Table 2, we observed that both classifiers matched all benchmark frames successfully with pattern distance around 0, 0.03 and 0.04 on datasets $D$, $P$ and $F$, respectively. The classification accuracy was 100% for the benchmark frames. However, for the remaining patterns, the EI classifier performed slightly better than the LL classifier, obtaining about 1% less pattern distance.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Benchmark</th>
<th>Remaining</th>
<th>Benchmark</th>
<th>Remaining</th>
</tr>
</thead>
<tbody>
<tr>
<td>$D$ (Noiseless)</td>
<td>0</td>
<td>0.036</td>
<td>0</td>
<td>0.041</td>
</tr>
<tr>
<td>$P$ (Shot-noise)</td>
<td>0.031</td>
<td>0.053</td>
<td>0.031</td>
<td>0.063</td>
</tr>
<tr>
<td>$F$ (Fluence-scaled)</td>
<td>0.042</td>
<td>0.062</td>
<td>0.043</td>
<td>0.074</td>
</tr>
</tbody>
</table>

The fluence distance, as defined in [Eq. (17)], of the dataset $F$ from the EI classifier was 0.035, comparing with 0.049 from the LL classifier, excluding the benchmark patterns. Further, the EI classifier was more efficient and took only 3.7 ms per image in our Matlab implementation, nearly 15 times faster than the LL classifier.

### 5.3. Heterogeneous Particle Sizes

We next evaluated the dataset $S$, containing patterns from icosahedrons of diameters between 150 nm to 210 nm. Fig. 3 illustrate the average pattern and fluence distance for the EI and the LL classifiers. As expected, both classifiers obtained the smallest distances when the particle size of the testing pattern was similar to the template size (180 nm), and the LL classifier had slightly larger distances on the average. Furthermore, the EI classifier was better at estimating particle sizes, as shown in Fig. 3(c), and this also implied that the EI classifier was more accurate in searching for the best-matched template than the LL classifier was.
The size and the fluence estimation procedures together took around 80 ms for each image, i.e., approximately 1.5 times longer than the LL classifier or 22 times longer than the EI classifier. In other words, with size estimation, the EI classifier can handle 12 images per minute and the LL classifier can perform 8 images per minute using our straightforward Matlab implementation. With Matlab Parallel Computing Toolbox and Distributed Computing Server, it is therefore possible to speed up both classifiers to the LCLS repetition rate (120 Hz). Since both methods can be parallelized, the European XFEL detector read-out rate of 3,520 Hz [10] is within reach for a compiled-language implementation.

5.4. Heterogeneous Shapes

In this section, we investigate the performances of the LL and EI classifier for the dataset \( X \), which contained particles with heterogeneous shapes and sizes. For identifying the spheroids in \( X \), we added a 180 nm sphere diffraction pattern into the training dataset \( T \), see \( T_1 \) (the training dataset 1) and \( X \) for more details in Table 1. Both classifiers distinguished the icosahedral and spheroidal diffraction patterns successfully, as listed in Table 3. All icosahedral diffraction patterns were classified as icosahedron with small pattern distances \((< 0.25)\). With a pattern-distance threshold of 0.5, the EI classifier rejected 78 elongated spheroidal patterns and identified 114 spheroidal frames as spheroids successfully. However, 8 (4\%) frames were misclassified as an icosahedron, and their pattern distances were between 0.42 and 0.5, see Fig. 4(b). The LL classifier gave a
similar but slightly worse result as it misclassified 9 spheroidal frames as icosahedrons. Since our desired patterns were indeed icosahedral, we could also describe the testing dataset by two classes – the icosahedron and the non-icosahedron. Both classifiers gave 100% recall (true positive rate), while the EI classifier gave a classification accuracy of 99.33% that was slightly better than the LL classifier (99.25%), see more details on the definitions of accuracy and recall in [33].

Table 3. Classification results of dataset $X$. The threshold of the pattern distance for rejection was set to 0.5. For the cases where the results from the LL classifier were different from the EI classifier, the values from the LL classifier are shown in parentheses.

<table>
<thead>
<tr>
<th></th>
<th>Classified as</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Icosahedron</td>
</tr>
<tr>
<td>Data: icosahedron</td>
<td>1000</td>
</tr>
<tr>
<td>Data: spheroid</td>
<td>8 (9)</td>
</tr>
</tbody>
</table>

We visually illustrate the classification results in Fig. 4. The rejected frames from both classifiers were elongated frames and the aspect ratios for most of them were smaller than 0.75, see Figs. 4(b) and 4(c). Further, we observed that the pattern distances decreased with increasing aspect ratio for both classifiers.

![Fig. 4](image-url)

Fig. 4. (a): The pattern distances of dataset $X$. For both classifiers, all icosahedral patterns were located in the perfectly matched region. For the EI classifier, all elongated spheroidal patterns (78 patterns) were rejected. For the remaining 122 accepted spheroidal patterns, 114 were successfully classified as spheroids, and eight frames or 4% of the spheroidal patterns were misclassified. The LL classifier gave one more misclassified spheroidal pattern. [(b) and (c)]: The relationship between the pattern distances and the aspect ratios for the spheroidal patterns from the EI (b) and the LL classifier (c). The red stars were misclassified patterns. The aspect ratio was the ratio of the length of the minor axis to the length of the main axis of the spheroidal particle. [(d)–(h)]: Five combination images, corresponding to the five data points (red circles) in (a) of the EI classifier. The left half of each image was from the testing dataset $X$, and the right half was the best-matched patterns from the training dataset. The number in each figure was the pattern distance. All figures are drawn in logarithmic scale.

5.5. Mimivirus diffraction patterns

We also tested our classifiers on the mimivirus FXI dataset, which has been used previously for a 3D mimivirus reconstruction [31]. To compensate detector saturation at the image center and low signal-to-noise ratio at the edges of the patterns, we used the central part of the diffraction patterns for classification, see Fig. 5. The training dataset for the mimivirus dataset ($T_2$) contained...
1000 randomly-oriented icosahedral patterns of 490 nm in diameter. Furthermore, we binned 4 × 4 pixels into one pixel in the calculations.

![Fig. 5. A mimivirus diffraction pattern (a) and its central region used for classification (b). The region shown in (b) was the region between two circles in (a).](image)

As expected, we obtained larger pattern distances from both the EI and the LL classifier, comparing with the synthetic dataset, see Fig. 6, and this is due to the heterogeneity in size and shape of the mimiviruses. Again, a pattern-distance threshold of 0.5 was used to detect irregular patterns. In total, both classifiers rejected 9.1% of patterns (18 patterns). To quantify the quality of our selected patterns, we fitted the completed dataset and the two selected datasets into a 3D intensity by the Expansion Maximization Compression (EMC) method [34], and looked at the correlation between the pattern distances and the sum of the largest 0.035% (or the largest 30), rotational probabilities of each diffraction pattern in Figs. 6(b) and 6(c).

As expected, the sum of the rotational probabilities increased with decreasing pattern distance. However, we did not get a linear correlation, most likely due to the fact that the mimiviruses samples were not regular uniform-density icosahedrons, and had different particle sizes. For example, in Figs. 6(d)–6(f), the templates and the mimivirus patterns matched quite well, however, the particle in Fig. 6(g) was slightly elongated, and the particle size of Fig. 6(h) was 457 nm, which was 33 nm smaller than the template size (490 nm). Further, our classifiers also improved the 3D reconstruction results, i.e., the average (minimum) of the sum of the rotational probabilities were improved to 0.207 (0.034) for the EI classifier, and 0.204 (0.032) for the LL classifier, compared with 0.174 (0.029) from the completed dataset. However, the 3D intensities reconstructed from these three sets were quite similar, mainly due to the quite limited number of patterns.

5.6. Raw mimivirus dataset

The purpose of our final experiments is to investigate the method’s behaviour when confronted with a raw mimivirus dataset. These data frames differ considerably in between, and hence it is hard to simulate all possibilities in order to obtain good templates. For classifying the raw mimivirus dataset [31], we therefore used 1000 randomly-oriented patterns and 10 sphere patterns in different sizes as the training dataset, see Table 1 for more details.

From 50,712 raw diffraction patterns in run 92 and run 93 (see more details about the experimental runs in Table 1), we found 578 hits by using the methods described in [35]. We centred all 578 hits (patterns with more than one particles) and set the pattern-distance threshold to 0.5. Further, we also limited the estimated sizes of selected patterns within [450, 540] nm. Finally, we also rejected all patterns which were accepted by the sphere templates. With these three criteria the EI classifier obtained 108 single-particle frames, and among them, 75 frames were manually classified as single-particle patterns, giving a precision (Positive Predictive Value, or PPV) of 0.69, see Table 4. Further, 14 manual singles were misclassified, giving a recall (True
Fig. 6. (a): The pattern distances of the mimivirus dataset from the EI and the LL Classifier. 18 of 198 patterns (9.1%) were rejected with a threshold of 0.5 from both classifiers. (b): The relationship between the pattern distance and the sum of the largest 0.035% (the largest 30) rotational probabilities of each diffraction pattern. [(d)–(h)]: Five combination images at the data points from the EI classifier (red circles) in (a). The left part of each image was from the mimivirus dataset, and the right part was the corresponding template scaled by the recovered fluence. (g) was a slightly elongated pattern and the particle size of (h) was smaller than the template size.

Positive Rate, TPR) of 0.84. The LL classifier gave similar but slightly worse results, see Table 5. Both classifiers obtained high accuracies (> 0.9). However, 33 (38) accepted frames were non-single hits for the EI (LL) classifier, giving a relatively low precision value of 0.69 (0.65). The often quoted F1 score (the harmonic mean of precision and recall) was 0.76 (0.72).

It is worth to point out that the mimivirus dataset had a low single-hit rate, and we may get a better precision if we have more single-particles in the testing dataset. Further, most rejected single-particle frames were sorted out by the size threshold. By loosening the size constraint, we could reduce the number of rejected single-particle patterns. However, the number of accepted non-single-particle patterns may then instead increase. In Fig. 7, we also visually illustrate five diffraction patterns together with their classification results from the EI classifier.

Table 4. Classification results of the raw mimivirus dataset [31] from the EI classifier, see text for more details on the classification measures (following [33, 36] closely). In the table, ACC, TPR, PPV, and F1, respectively, abbreviates Accuracy, True Positive Rate, Positive Predictive Value, and F1 score. TPR and PPV are sometimes referred to as the recall and the precision.
### Table 5. Classification results of the raw mimivirus dataset [4] from the LL classifier.

<table>
<thead>
<tr>
<th></th>
<th>Run 92</th>
<th></th>
<th>Run 93</th>
<th></th>
<th>Run 92 &amp; 93</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Single</td>
<td>Other</td>
<td>Single</td>
<td>Other</td>
<td>Single</td>
</tr>
<tr>
<td>Accepted</td>
<td>43</td>
<td>23</td>
<td>28</td>
<td>15</td>
<td>71</td>
</tr>
<tr>
<td>Rejected</td>
<td>8</td>
<td>196</td>
<td>10</td>
<td>255</td>
<td>18</td>
</tr>
<tr>
<td>ACC</td>
<td>0.89</td>
<td></td>
<td>0.92</td>
<td></td>
<td>0.90</td>
</tr>
<tr>
<td>TPR</td>
<td></td>
<td></td>
<td>0.84</td>
<td></td>
<td>0.74</td>
</tr>
<tr>
<td>PPV</td>
<td>0.65</td>
<td></td>
<td>0.65</td>
<td></td>
<td>0.65</td>
</tr>
<tr>
<td>F1</td>
<td>0.74</td>
<td></td>
<td>0.69</td>
<td></td>
<td>0.72</td>
</tr>
</tbody>
</table>

Fig. 7. Five combination images from the EI classifier for the raw mimivirus data in [31]. The left part of each image was from the mimivirus dataset, and the right part was the corresponding template scaled by the recovered fluence. The color scale is logarithmic and ranges from 0 to 1000 photons per pixel. (a): One of the best accepted diffraction patterns. (b): One selected diffraction pattern with low photon counts: the total number of photons was around $2.4 \times 10^5$. (c): A potential water droplet, which was rejected since it was accepted by a sphere template. (d) and (e): Patterns rejected by the size threshold.

### 6. Conclusions

The FXI technique holds the promise of obtaining biomolecule structures from single particles. It operates at a high repetition rate and records thousands of millions of diffraction data every day. The stochastic nature of XFELs and the heterogeneity of the sample molecules make the recorded dataset too complex and massive to classify manually. By using our knowledge of the sample molecules, such as sizes and shapes, we can use template-based methods to reduce the complexity of the classification problem and select more homogeneous diffraction patterns. In consequence, the next step of FXI data analysis – the 3D orientation determination procedure, will hopefully need less computations, and be faster to converge, leading to a 3D model with an improved resolution. Both proposed methods obtained a high classification accuracy, and most non-single-particle patterns were sorted out. Some non-single-particle patterns were still selected and since the testing raw dataset had a low single-particle hit rate, we may had selected less non-single-particle patterns if we had a higher single-particle hit rate. Improvements in the quality of the raw data frames would thus still remain very beneficial for the end-result resolution.

In our straightforward Matlab implementations, both methods can classify a testing pattern in a few milliseconds, and they certainly can be accelerated to the XFEL repetition rates, albeit using considerable resources. We also observed that the rotational probabilities, from the 3D orientation determination procedure, increased with decreasing pattern distances. Further, the selected patterns from our classifiers fit better into a 3D Fourier intensity, resulting in a potentially better resolution of the 3D electron density of the sample molecules.

Newer facilities, such as the European XFEL, operate at high repetition rates and will create massive volumes of FXI diffraction data with heterogeneities to varying degrees. With our methods, we can use most of our knowledge of the sample molecules to reduce data storage and automatically select homogeneous single-particle patterns. We also foresee that an on-site FXI...
analysis pipeline, which connects our classifier to the 3D reconstruction procedure, can solve the 3D structure with sub-nanometer resolution in the near future.

**Funding**

This work was financially supported by the Swedish Research Council within the UPMARC Linnaeus center of Excellence (S. Engblom, J. Liu) and by the Swedish Research Council, the Röntgen-Ångström Cluster, the Knut och Alice Wallenbergs Stiftelse, the European Research Council (J. Liu, G. Schot).

**References**


