

**Citation for published version:**

Emmanuel Oluwatobi Salawu, Evelyn Hesse, Chris Stopford, Neil Davey, and Yi Sun, 'Applying machine learning methods for characterization of hexagonal prisms from their 2D scattering patterns – an investigation using modelled scattering data', *Journal of Quantitative Spectroscopy and Radiative Transfer*, Vol. 201, pp. 115-127, Nov 2017.

**DOI:**

<https://doi.org/10.1016/j.jqsrt.2017.07.001>

**Document Version:**

This is the Accepted Manuscript version.

The version in the University of Hertfordshire Research Archive may differ from the final published version.

**Copyright and Reuse:**

© 2017 Elsevier Ltd. All rights reserved.

This Manuscript version is distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported (CC BY-NC-NC 3.0) license

<https://creativecommons.org/licenses/by-nc-nd/3.0/>

**Enquiries**

If you believe this document infringes copyright, please contact the Research & Scholarly Communications Team at [rsc@herts.ac.uk](mailto:rsc@herts.ac.uk)

Applying machine learning methods for characterization of hexagonal prisms from their  
2D scattering patterns – an investigation using modelled scattering data

Emmanuel Oluwatobi Salawu <sup>1</sup>, Evelyn Hesse<sup>2</sup>, Chris Stopford<sup>2</sup>, Neil Davey<sup>3</sup>, Yi Sun<sup>3</sup>

<sup>1</sup> TIGP Bioinformatics Program, Academia Sinica, Taipei, Taiwan; Institute of  
Bioinformatics and Structural Biology, National Tsing Hua University, Hsinchu, Taiwan;  
School of Computer Science, University of Hertfordshire, Hatfield, Herts, UK  
(emmanuel@gapp.nthu.edu.tw)

<sup>2</sup> The School of Physics Astronomy Maths, University of Hertfordshire, Hatfield, Herts,  
UK ({e.hesse, c.stopford}@herts.ac.uk)

<sup>3</sup> The School of Computer Science, University of Hertfordshire, Hatfield, Herts, UK  
({n.davey, comrys}@herts.ac.uk)

## Abstract

Better understanding and characterization of cloud particles, whose properties and distributions affect climate and weather, are essential for the understanding of present climate and climate change. Since imaging cloud probes have limitations of optical resolution, especially for small particles (with diameter  $< 25 \mu\text{m}$ ), instruments like the Small Ice Detector (SID) probes, which capture high-resolution spatial light scattering patterns from individual particles down to  $1 \mu\text{m}$  in size, have been developed. In this work, we have proposed a method using Machine Learning techniques to estimate simulated particles' orientation-averaged projected sizes (PAD) and aspect ratio from their 2D scattering patterns. The two-dimensional light scattering patterns (2DLSP) of hexagonal prisms are computed using the Ray Tracing with Diffraction on Facets (RTDF) model. The 2DLSP cover the same angular range as the SID probes. We generated 2DLSP for 162 hexagonal prisms at 133 orientations for each. In a first step, the 2DLSP were transformed into rotation-invariant Zernike moments (ZMs), which are particularly suitable for analyses of pattern symmetry. Then we used ZMs, summed intensities, and root mean square contrast as inputs to the advanced Machine Learning methods. We created one random forests classifier for predicting prism orientation, 133 orientation-specific (OS) support vector classification models for predicting the prism aspect-ratios, 133 OS support vector regression models for estimating prism sizes, and another 133 OS Support Vector Regression (SVR) models for estimating the size PADs. We have achieved a high accuracy of 0.99 in predicting prism aspect ratios, and a low value of normalized mean square error of 0.004 for estimating the particle's size and size PADs.

**Keywords:** Machine learning, scattering pattern, hexagonal prisms, ice crystals, size, aspect ratio, Ray Tracing with Diffraction on Facets, Zernike moments.

## 1.0 Introduction

Cloud feedbacks are a large source of uncertainty in climate models [1]. In particular, there are uncertainties about the radiative forcing of clouds containing ice crystals, especially cirrus. Whether cirrus clouds warm or cool the Earth's surface depends on ice crystal morphology [2]. Reducing this uncertainty requires detailed *in situ* characterization of cloud particles. Cloud probes based on imaging techniques, such as the Cloud Particle Imager (CPI) have limitations of optical resolution when dealing with particles smaller than about 25  $\mu\text{m}$  [3]. These limitations do not apply to instruments detecting light scattering patterns of cloud particles, like the Small Ice Detector (SID) developed at the University of Hertfordshire [4].

However, to obtain particle information from the scattering patterns requires solving the inverse scattering problem. This is often facilitated by previous knowledge of how scattering properties of small particles depend on the particle size parameter, morphology, relative refractive index and orientation. This knowledge may be the cumulative result of investigating many specific cases combined with interpolation or extrapolation of particle characteristics not covered by existing theoretical or experimental data [5]. In order for such databases to be created, either computational models which can determine the intensity of light scattered by a known particle into a given angular range, or experimental data are required. Scattering by spheres can be described by Lorenz-Mie theory. For computation of non-spherical particles, numerically exact methods like T-matrix [6,7] or Discrete Dipole Approximation (DDA) [8, 9] can be applied, however they are currently computationally very expensive and are therefore restricted to small particle sizes. This leaves a size parameter range that is covered neither by exact methods nor by geometric optics (GO). Approximate methods, such as the geometric optics approximation or physical optics, have to be used for scatterers much larger than the wavelength. In the classical geometric optics approximation, scattered light is divided into two parts, firstly light reflected or transmitted by the scatterer, and secondly externally diffracted light [e.g. 10]. Diffraction of reflected and refracted light is neglected, resulting in singular intensity peaks. Therefore, this method is not applicable for interpreting measured 2D scattering patterns. Improved methods including diffraction of the ray-tracing component have been presented e.g. in [11-15], and a volume integral method in [15,16]. Ulanowski et al. [17] retrieved the size of particles with rough and complex surfaces from two-dimensional scattering patterns by investigating the median surface area of intensity peaks. The RTDF model [14] has been used to generate a database of 2D light scattering patterns (2DLSPs) [18] investigated here.

Machine Learning is ideally suited for estimating a range of properties of scattering particles from linked known parameters. Baran and Newman [19] have demonstrated the application of principal component analysis to estimate cloud ice water content and environmental temperature from bulk integral optical properties of ice cloud particles. Here, we wish to derive single particle properties from their 2DLSP, which are in general quite complex. Radial Basis Function (RBF) neural networks have been applied to solve the inverse light scattering problem for spheres [20]. In order to discriminate potentially hazardous respirable fibres, such as asbestos, Kaye et al. [21] have used experimental data to train RBF neural networks. Genuer et al. [22] applied machine learning on scatterograms of microcolonies for discriminating bacteria and yeasts at an early stage of growth. For this, they projected the patterns on either the Zernike orthogonal basis or a Fourier-Bessel function basis. The radial function of the latter was found to be more useful for analysis of the patterns, which consisted largely of concentric rings.

The work presented here aims to investigate the applicability of advanced machine learning methods to solve the inverse problem for hexagonal prisms. They are a useful test case, since virtually all the ice on Earth's surface and in its atmosphere is of a hexagonal crystalline structure. Due to their symmetry, scattering patterns of hexagonal prisms are quite different from the largely concentric patterns discussed in [22]. At this initial stage we disregard complex

crystals and any surface roughness. Since roughness has found to be important [2], this will restrict the direct applicability of the results obtained in this study, but we aim to demonstrate a proof of concept here.

For any particle except a sphere the 2DLSP will depend on its orientation with respect to the incident laser beam. However, for a highly symmetric particle like a hexagonal prism there will be a continuous range of particle orientations which result in 2DLSP's differing only by a rotation around the direction of incident light (see section 2). Such orientations should be attributed to *one single representative orientation*. This means that the representations of 2DLSPs which are the inputs of each computational model should be rotationally invariant. Zernike polynomials, which were originally derived to assist the characterization of imperfections of concave mirrors by analyzing their diffraction pattern [23] and have been used for other surface reflectance applications [24,25] seem a suitable approach to achieve this. Building on Zernike polynomials and the general theory of moments, Teague [26] derived the Zernike moments, in which the Zernike polynomials have been used as the basis functions for the moments, and applied them to visual pattern recognition [27]. This method has been widely applied in pattern recognition [28–31].

The aim of this work is two-fold: 1) to investigate if machine learning methods can be applied for estimating characteristics of small particles from 2DLSP using the example of 2DLSP of hexagonal prisms computed using RTDF; 2) to investigate if Zernike moments are suitable for representing 2DLSPs, and to know the range of the moments that should be used.

We employ machine learning methods, such as Random Forests Classification (RFCs) [33], Support Vector Classification (SVC) and the Support Vector Regression (SVR) methods, to estimate characteristics of small particles based on sets of representations including Zernike moments, summed intensities and root mean square contrasts from each 2DLSP image. The hexagonal prism properties we wish to estimate from a 2DLSP are aspect ratio, projected size and orientation averaged projected size (generally, particle projected size is orientation dependent). By projected size we mean the diameter of a circle with the same area as the projected cross section of the particle. The orientation-averaged projected size is described by the diameter of a circle with the same area as the orientation averaged projected cross section of the particle (PAD denotes average projected area diameter). In our previous work [34], we have developed a method combining a Feed Forward Multi-Layer Perceptron neural network with Bayesian regularization back-propagation and rotation invariance with Fast Fourier Transform. However, the model could not deal with orientation averaged projected size, and the model could not predict the size of very small particles (between 3 and 10  $\mu\text{m}$  size) with the same precision as it did for the larger particles.

The rest of paper is organized as follows. In Section 2, we introduce the dataset used in this work. In Section 3, we describe our method and give details on approaches we have applied in this work. We evaluate our approach and show experimental results in Section 4. We conclude our paper in Section 5 by discussing the potential use of our approach.

## 2.0 Dataset

Atmospheric ice crystals are typically of an intermediate size range that can neither be covered by exact electromagnetic methods like T-matrix, because they are currently computationally very expensive, nor by classical geometric optics. Here, we used the Ray Tracing with Diffraction on Facets (RTDF) model [14], which is a hybrid method combining ray-tracing with a physical optics approximation for diffraction, for computing the 2DLSP of hexagonal prisms. The 2DLSP are azimuthally resolved phase functions  $P_{11}$  [32], i.e. the intensity has been normalised to  $4\pi$  over the complete scattering sphere. In general, incident light can be described by its Stokes vector  $[I_i Q_i U_i V_i]^T$ . As a result of the scattering event, it is transformed into the

Stokes vector  $[I_s Q_s U_s V_s]^T$  of the scattered field. This corresponds to multiplying  $[I_i Q_i U_i V_i]^T$  with the 4x4 phase matrix,  $\mathbf{P}(\theta, \varphi)$  of the particle (eq.1).

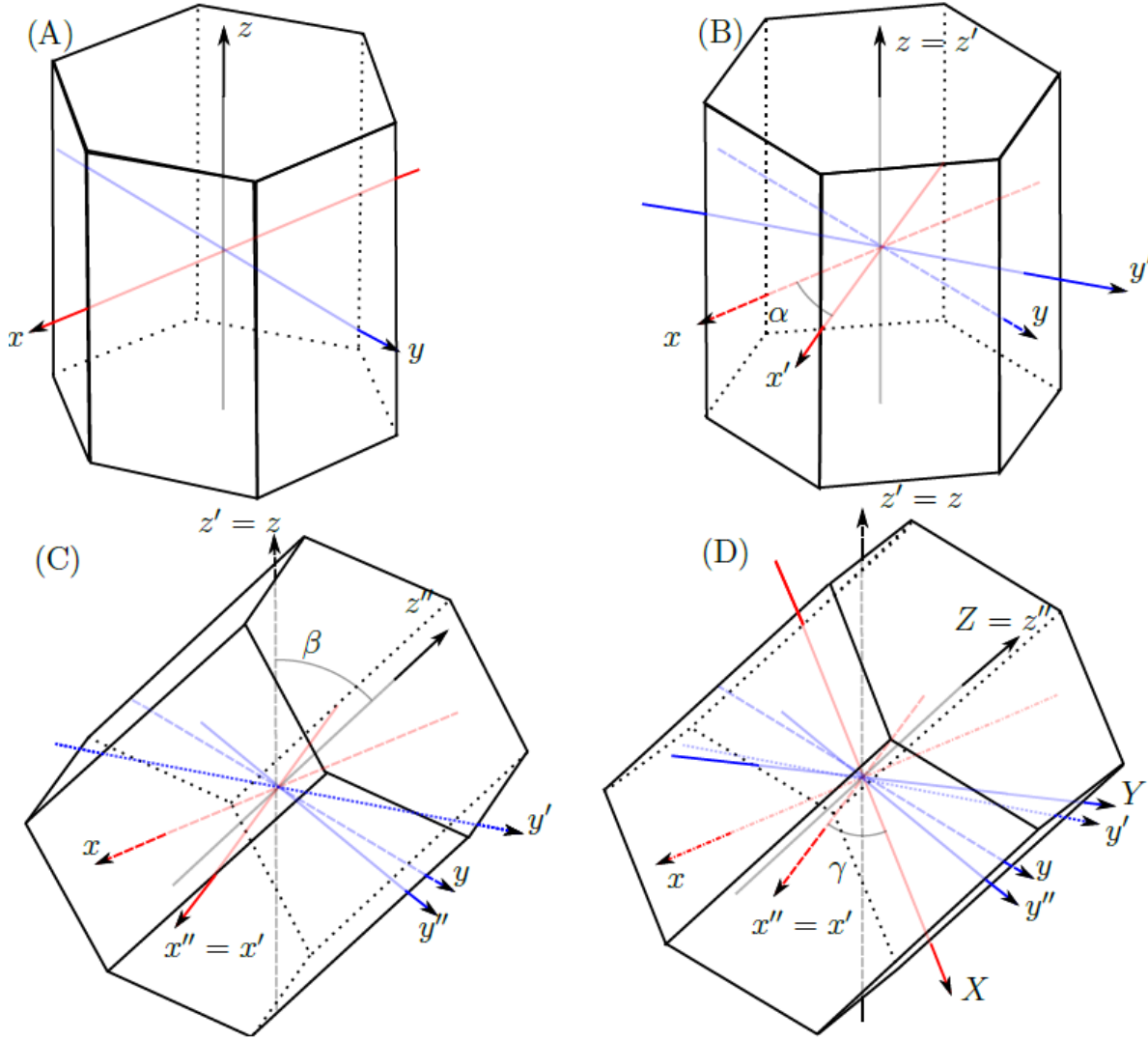
$$\begin{pmatrix} I_s \\ Q_s \\ U_s \\ V_s \end{pmatrix} = \begin{pmatrix} P_{11} & P_{12} & P_{13} & P_{14} \\ P_{21} & P_{22} & P_{23} & P_{24} \\ P_{31} & P_{32} & P_{33} & P_{34} \\ P_{41} & P_{42} & P_{43} & P_{44} \end{pmatrix} \begin{pmatrix} I_i \\ Q_i \\ U_i \\ V_i \end{pmatrix} \quad (1)$$

Unpolarised light of unit intensity has the Stokes vector  $[1 \ 0 \ 0 \ 0]^T$ . This results in an intensity  $I_s = P_{11}$  of the scattered light. The laser in the SID instrument emits circularly polarised light to minimize polarization-dependent variations in the captured particle scattering patterns [5]. Circularly polarised light of unit intensity has the Stokes vector  $[1 \ 0 \ 0 \ \pm 1]^T$ . The sign of the  $V$  parameter will be positive for right-handed and negative for left-handed light. This results in an intensity  $I_s = P_{11} \pm P_{14}$  of the scattered light. The phase matrix element  $P_{14}$ , which is also called circular intensity differential scattering, is very small for particles of non-chiral materials. Therefore,  $I_s \cong P_{11}$ .

For our analysis we chose the wavelength of 532nm, which is emitted by the laser in the SID instrument. The refractive index of water ice at this wavelength is  $n = 1.31 + 1.48 \cdot 10^{-9} i$  [35,36]. Hexagonal prisms in ice clouds vary in shape from long columns to thin plates. Observations obtained from cloud campaigns in the tropics, Antarctic and mid-latitudes were analysed in [37]. We define the aspect ratio of the prisms as length divided by diameter, where the diameter is twice the edge length of the basal facet. This means that columns and plates have aspect ratios larger and smaller than one, respectively. For the database we chose an interval  $[1/8, 8]$  containing the following nine aspect ratios:  $[1/8, 1/5.5, 1/3, 1/2, 1, 2, 3, 5.5, 8]$ . This covers and, for columns, extends beyond the aspect ratio space reported in [37] and should be a suitable range for testing the method. For each aspect ratio we chose 18 different particle sizes defined by their PAD, ranging from 6  $\mu\text{m}$  to 82  $\mu\text{m}$  (the upper limit is related to the range of the SID instrument [18, 38]).

Therefore, we have  $9 \times 18 = 162$  particles in the dataset. Particle orientation is defined by the Euler angles  $\alpha$ ,  $\beta$  and  $\gamma$ , which describe the required rotations of the particle coordinate system with respect to the laboratory system: At the start, the prism axis is aligned along the  $z$ -axis. Two prism edges are in the  $xz$ -plane (Fig. 1A). First, the particle system is rotated by an angle  $\alpha$  around the  $z$ -axis, resulting in the new coordinate system  $[x', y', z']$  (Fig. 1B). In a next step, the particle system is rotated by  $\beta$  about the  $x'$  axis resulting in the new coordinate system  $[x'', y'', z'']$  (Fig. 1C). Finally, the particle system is rotated by  $\gamma$  about the  $z''$  axis (Fig. 1D). The final particle system is denoted  $[X, Y, Z]$ . For the database we chose 133 unique orientations described by their Euler angles  $[\alpha, \beta, \gamma]$ : Due to the hexagonal crystal symmetry, the  $\alpha$  and  $\gamma$  rotations are not independent. Therefore,  $\alpha$  was set to  $0^\circ$  and  $\gamma$  was varied between  $0^\circ$  and  $30^\circ$  in  $5^\circ$  steps. The Euler angle  $\beta$  was varied between  $2^\circ$  and  $88^\circ$ , avoiding the extreme values  $\beta=0^\circ$  and  $\beta=90^\circ$ . The step size between  $\beta=5^\circ$  and  $\beta=85^\circ$  was  $5^\circ$ . This results in a dataset of  $18 \times 9 \times 133 = 21,546$  2DLSP. The 2DLSPs contain only the angular region covered by the SID instrument: Elevation and azimuthal angle ranges are  $6^\circ$  to  $25^\circ$  and  $0^\circ$  to  $360^\circ$ , respectively. Step size is  $1^\circ$ . This results in  $20 \times 361 = 7220$  data points.

The database was generated automatically using the following algorithm: In the first step, a desired aspect ratio was selected. Next, for this aspect ratio and each desired size PAD a best guess for length and diameter was made and the PAD calculated. Finally, length and diameter were adjusted, until the appropriate size PAD was reached.

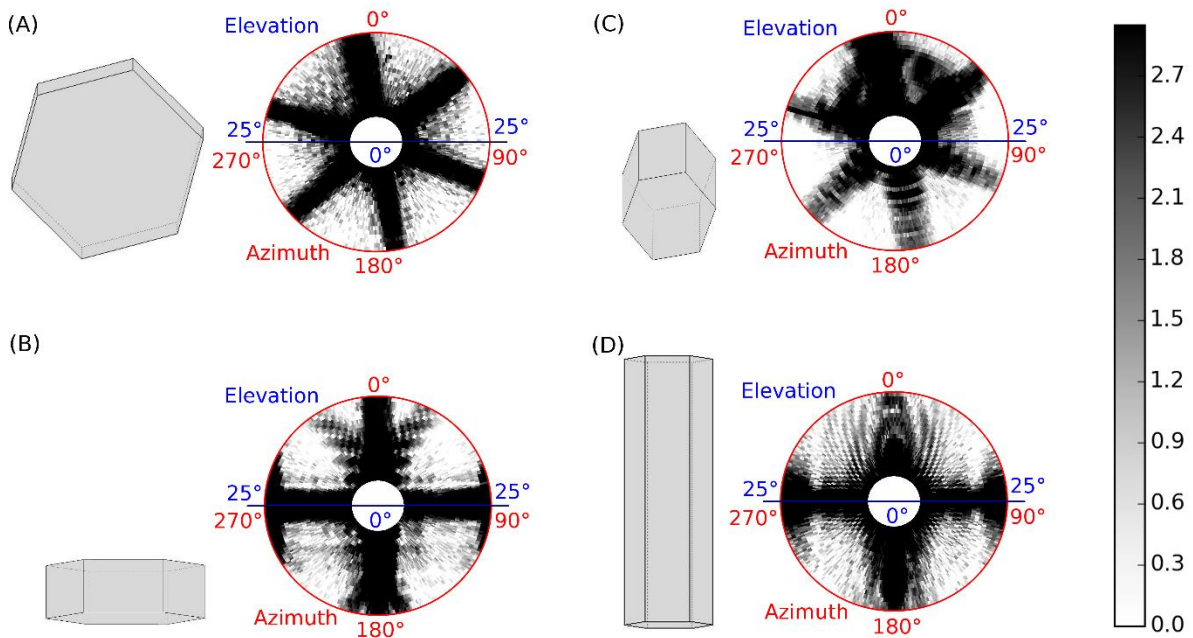


**Figure 1. Particles orientations defined by  $\alpha$ ,  $\beta$ , and  $\gamma$ .** Rotation of a hexagonal column using the Euler angles. (A) is the original crystal orientation, (B) shows the first rotation by  $\alpha$  about the  $z$  axis, (C) shows the second rotation by  $\beta$  about the  $x'$  axis, (D) shows the third rotation by  $\gamma$  about the  $z'' = Z$  axis. The rotated axes are denoted by  $X, Y, Z$ . (This figure is a copy from [39].)

Figure 2 shows examples of 2DLSP. They are natural-log plots of mean-truncated intensities. To achieve mean-truncation, the intensities higher than the mean intensity are set to the mean intensity. The mean-truncation was done to reduce the contrast only for the purpose of better visualization and not for subsequent computations. Panels A and B are for the same hexagonal plate of length  $8.9 \mu\text{m}$  and of diameter  $26.6 \mu\text{m}$  but for different orientations:  $[\alpha = 0.0^\circ, \beta = 10.0^\circ, \gamma = 10.0^\circ]$  and  $[\alpha = 0.0^\circ, \beta = 85.0^\circ, \gamma = 2.0^\circ]$ , respectively. Panels C and D are for a column (length  $36.0 \mu\text{m}$ , diameter  $12.0 \mu\text{m}$ ) with the same orientations as panels A and B, respectively. Whereas the patterns in panels A and C are dominated by a star-like feature with six arms due to diffraction at the hexagonal facet facing the incident beam (*see inset*), the patterns in panels B and D are dominated by cross-shaped features due to diffraction at the close to rectangular contour of the projected area of the crystal. For the column, which has smaller basal facets than the plate, diffraction arms in panels (A) and (C) are wider than for the plate.

Furthermore, the height of the plate is much smaller than the column height, therefore the vertical diffraction arc in panel (B) is much stronger than the one in panel (D).

Table 1 shows the basic statistics of the dataset. The intensity has been normalised to  $4\pi$  over the complete scattering sphere, i.e. the phase function [32] is presented. The intensities in the angular region of interest have a mean of 3.62 and a standard deviation (sd) of 28.75. The distribution of the intensities is highly skewed, but log-normal (mean: -0.89, sd: 2.10). Based on the discontinuous nature of the aspect ratio (nine possible values in the current study) we treated particles' aspect ratios as a binary categorical variable (rather than a continuous variable) such that a particle having an aspect ratio less than and up to 1 is labelled "0", a particle with an aspect ratio higher than 1 is labelled "1". Figure 3 shows the distribution of aspect ratios as well as the distributions of size PAD and size in the complete data set, the training set and the test set.

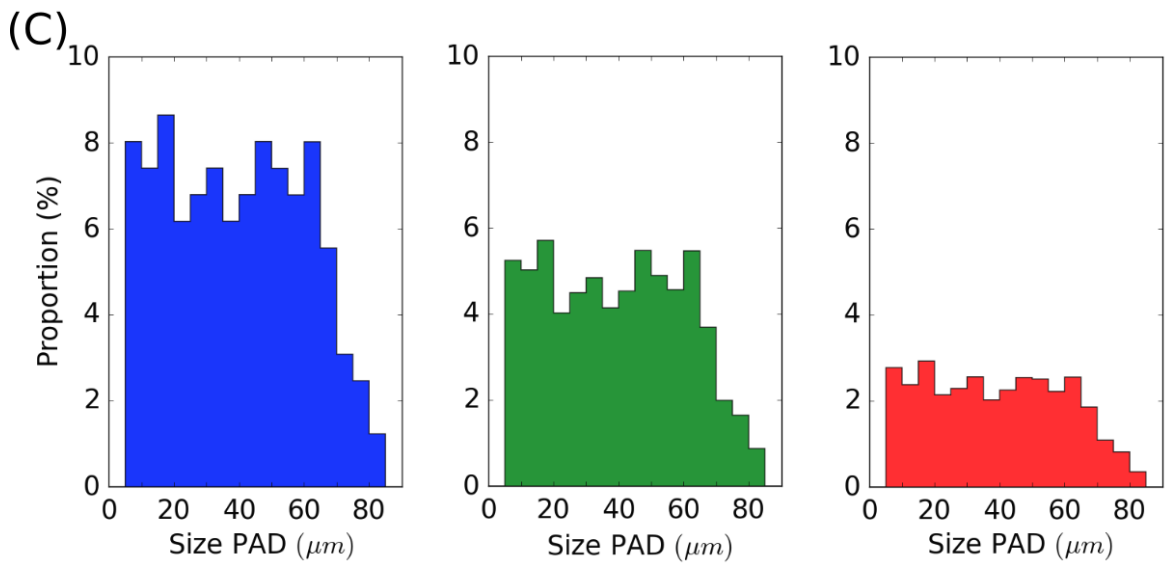
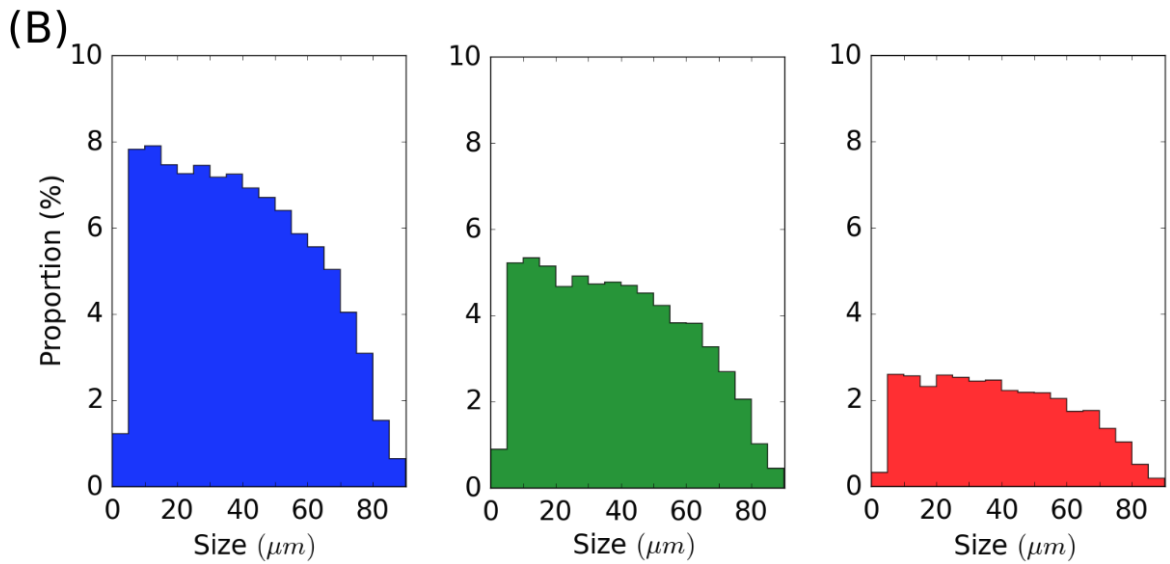
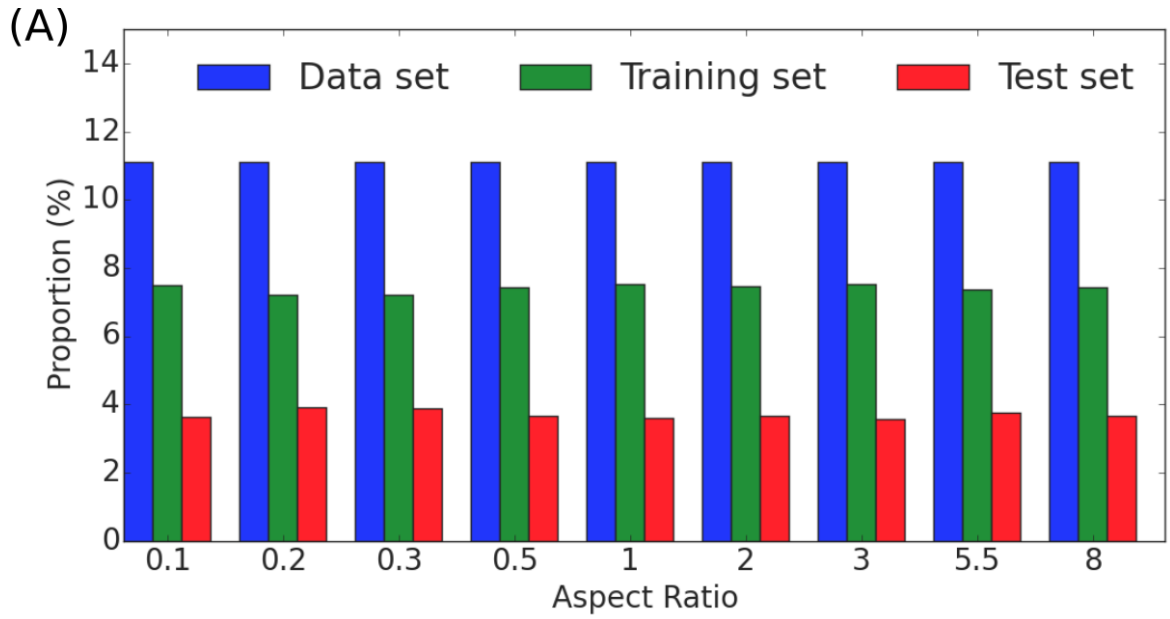


**Figure 2. Two out of 133 orientations of 2DLSP for each small particle.** *A and B are for a hexagonal plate of length  $8.9 \mu\text{m}$  and of diameter  $26.6 \mu\text{m}$  (aspect ratio  $\approx 0.3$ ), while C and D are for a hexagonal column of comparable volume but with a different aspect ratio (3.0) of length  $36.0 \mu\text{m}$  and diameter  $12.0 \mu\text{m}$ . Panel A (and panel C) is for the orientation where  $\alpha = 0.0^\circ$ ,  $\beta = 10.0^\circ$ , and  $\gamma = 10.0^\circ$ ; while B (and D) is for the orientation where  $\alpha = 0.0^\circ$ ,  $\beta = 85.0^\circ$ , and  $\gamma = 2.0^\circ$ . The data points here are limited to the intensities within the elevation range ( $6^\circ \leq \text{elevation} \leq 25^\circ$ ) that the SID detector captures.*



**Table 1.** Descriptive statistics of the dataset

	Intensity	ln(Intensity)	Size	Size PAD	Aspect Ratio
Mean	3.62	-0.89	39.11	39.36	2.29
Standard deviation	28.75	2.10	21.86	20.69	2.61
Median	0.39	-0.95	37.46	38.82	1.0
Coefficient of variation	7.94	2.36	0.56	0.53	2.29
Minimum (m)	1.62e-11	-24.84	2.13	5.04	0.12
Maximum (M)	7580.19	8.93	96.81	82.31	8.04
Range (i.e. M – m)	7580.19	33.78	94.68	77.27	7.92
Number of data points	155,562,120	155,562,120	21,546	21,546	21,546



**Figure 3.** Distributions of (A) aspect ratios (B) size, and (C) size PAD in the complete data set, training set and test set.

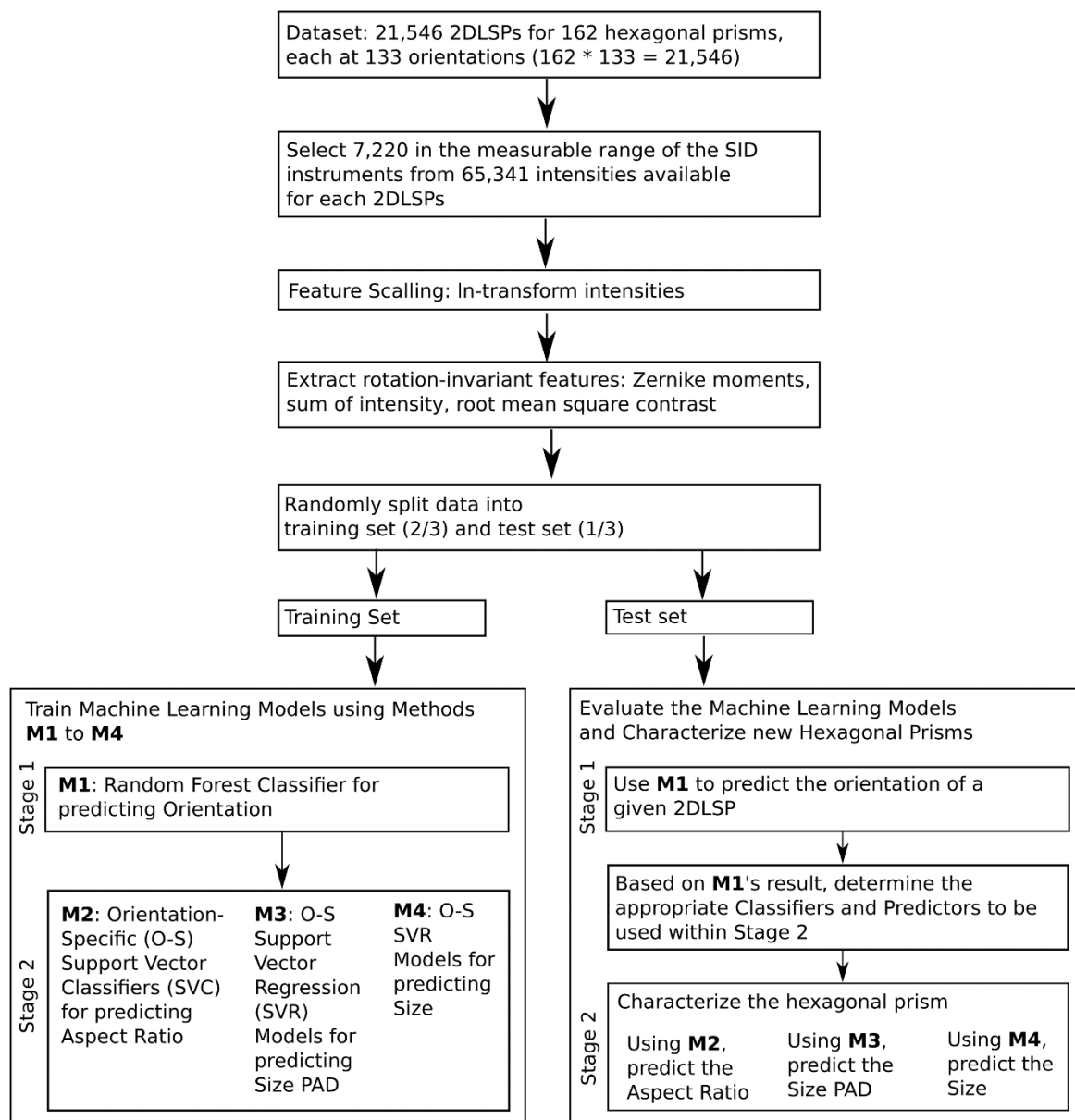
### 3.0 Methods

To accurately characterize the hexagonal prisms from their 2DLSPs, we explored various feature (i.e. intensity) scaling methods and found natural-log-transformation appropriate as the intensities are log-normal ( $\ln$ ). To reduce the dimensionality while retaining the useful information contained in the 2DLSPs, and to make the 2DLSPs rotation-invariant (i.e. resistant to rotation), we calculated Zernike moments (ZMs) of the intensities and used the ZMs as the main features for the training of the machine-learning models.

After producing representations of 2DLSPs, we used ZMs together with summed intensities and root mean square contrast of the intensities as the features for machine-learning models. We then trained a random forests classifier for predicting the orientation for a given 2DLSP. Accurately predicting the orientation of the hexagonal prism forms the first stage of the characterization, and the second stage of the characterization builds on this. For the second stage, we trained 133 orientation-specific support vector regression models for predicting size of hexagonal prisms, 133 orientation-specific support vector regression models for predicting size PAD of hexagonal prisms, and 133 orientation-specific support vector classifiers for predicting aspect ratio of hexagonal prisms. As we have discussed in Section 2, we consider aspect ratio as a binary classification problem (aspect ratio  $\geq 1$  or  $<1$ ).

When a new 2DLSP (or a new test example) was received, we first produced its representation, then we used the trained random forests classifier to predict its orientation thereby completing its “stage 1” characterization. The predicted orientation was then used to select the appropriate model for predicting its size, size PAD, and aspect ratio (for its stage-2 characterization). Figure 4 shows a schematic representation of the nature of our datasets and our proposed modelling methods.

We introduce the computational models used in this work in the following sub-sections.

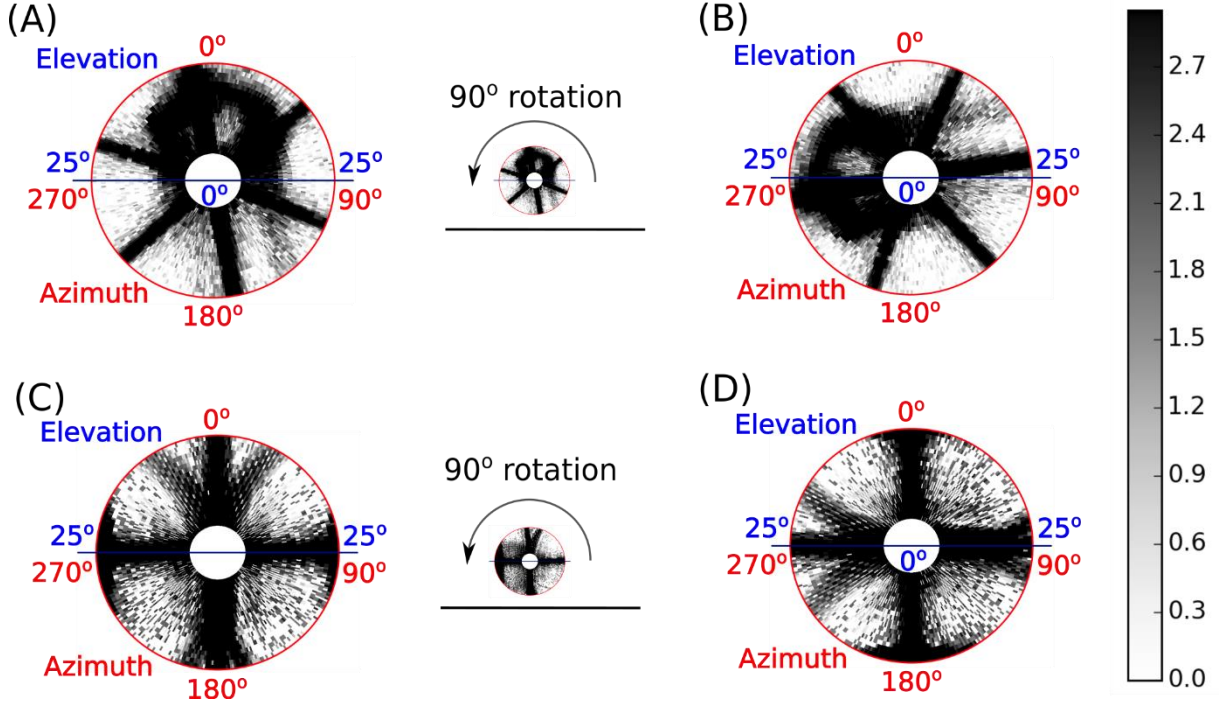


**Figure 4.** Schematic representation of the dataset and the methods used

### 3.1 Representations of 2DLSP

#### 3.1.1 Zernike moments in image processing

Figure 5 shows two example images (A and C) of one particle (length  $89.1 \mu\text{m}$  and diameter  $29.7 \mu\text{m}$ ) in two different orientations ( $\alpha = 0.0^\circ$ ,  $\beta = 10.0^\circ$ , and  $\gamma = 10.0^\circ$ ; and  $\alpha = 0^\circ$ ,  $\beta = 10.0^\circ$ , and  $\gamma = 20.0^\circ$ ). Images B and D are produced by rotating A and C anticlockwise respectively. The model must be robust enough to know that A and B (or C and D) are from exactly the same particle with the same particle orientation (defined by the Euler angles) and contain the same information (and that the rotation through an arbitrary angle does not change the characteristics of the particle whose 2DLSPs is being considered).



**Figure 5.** Depiction of the need for rotation-invariant models using images from one particle (length 89.1  $\mu\text{m}$  and diameter 29.7  $\mu\text{m}$ ; A:  $\alpha = 0.0^\circ$ ,  $\beta = 10.0^\circ$ , and  $\gamma = 10.0^\circ$ ; and C:  $\alpha = 0^\circ$ ,  $\beta = 85.0^\circ$ , and  $\gamma = 2.0^\circ$ ). A (or C) has been rotated anticlockwise through  $90^\circ$  to produce B (or D). The rotation of A (or C) to B (or D) should not change its features.

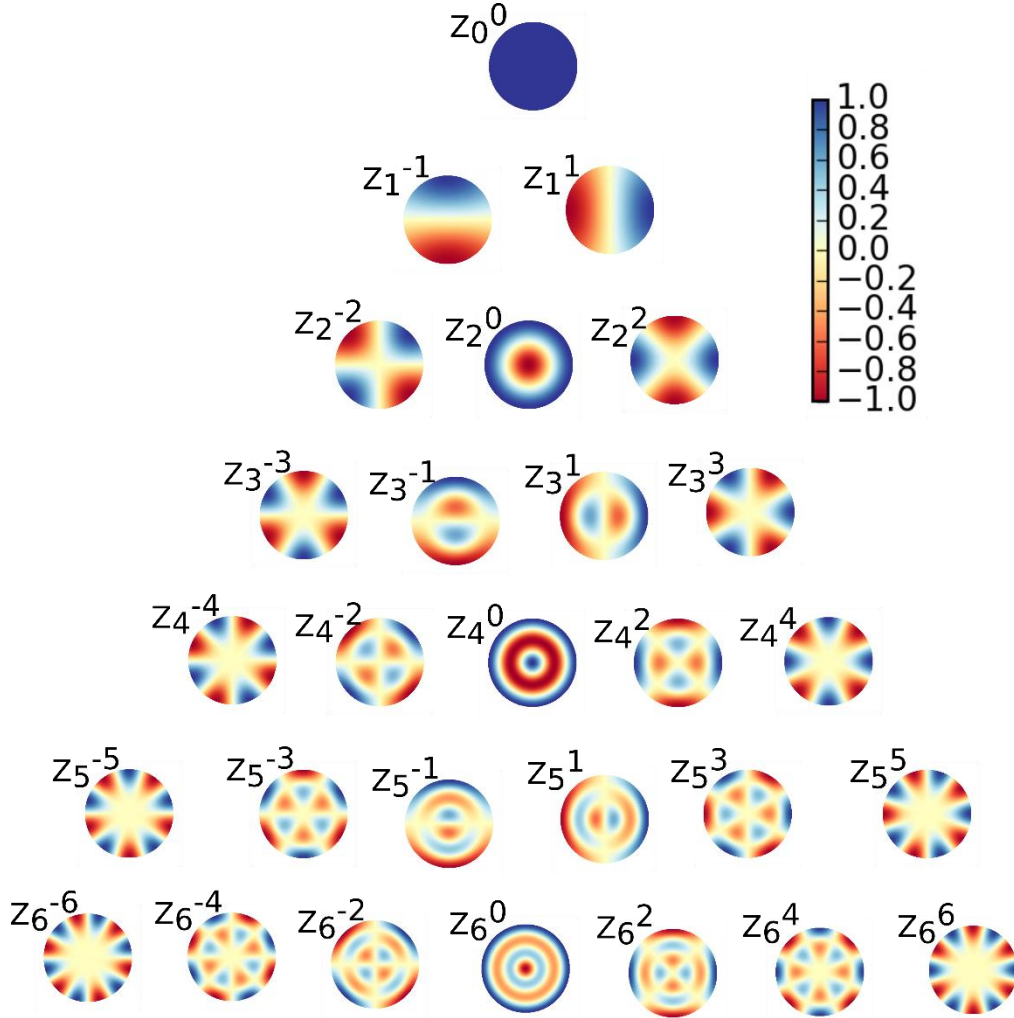
To create rotation-invariant computational models for the smooth hexagonal prisms such that a mere rotation of the 2DLSP of a smooth hexagonal prism through an arbitrary angle does not change its features, Zernike moments (ZMs) based on Zernike polynomials [23] have been applied in this work. Zernike polynomials were introduced by Frits Zernike [23]. ZMs can be obtained by projecting an image function  $f(x, y)$  onto a set of orthogonal basis functions, called Zernike polynomials. These Zernike polynomials are a set of complex polynomials which are orthogonal over a unit circle, that is,  $x^2 + y^2 = 1$ , and are defined as follows [28]:

$$Z_n^m(x, y) = R_n^m(\rho)e^{jm\theta} \quad (2)$$

where  $n$  denotes the order (or power) of the radial polynomial ( $R_n^m(\rho)$ ) and  $m$  denotes the azimuthal order.  $m$  can be either positive or negative integers subject to constraints that  $|n - m|$  should be an even number, and  $n \geq |m| \geq 0$ .  $\rho$  is the length of the vector from the origin to the  $(x, y)$  pixel, and  $\theta$  is the angle between vector  $\rho$  and  $x$  axis in counterclockwise direction, i.e.  $x = \rho \cos\theta$  and  $y = \rho \sin\theta$ . The radial polynomial,  $R_n^m(\rho)$ , is defined as follows.

$$R_n^m(\rho) = \begin{cases} \sum_{l=0}^{(n-m)/2} \frac{(-1)^l (n-l)!}{l! \left[\frac{1}{2}(n+m)-l\right]! \left[\frac{1}{2}(n-m)-l\right]!} \rho^{n-2l} & \text{for } (n-m) \text{ even} \\ 0 & \text{for } (n-m) \text{ odd} \end{cases} \quad (3)$$

Figure 6 shows the first 28 Zernike polynomials. As can be seen, for each order, that is each row in the figure, there are a number of polynomials satisfying the constraints, that are:  $n - |m|$  is an even number, and  $n \geq |m|$ .



**Figure 6. The first 28 Zernike polynomials.** Zernike polynomials form the basis for Zernike moments.

Furthermore, Zernike moments (ZMs) for a digital image are defined as follows [28]:

$$ZM_n^m = \frac{n+1}{\pi} \sum_x \sum_y f(x, y) Z_n^{m*}(\rho, \theta) \quad (4)$$

where  $Z_n^{m*}(\rho, \theta)$  is the conjugate of  $Z_n^m(\rho, \theta)$

To reconstruct an image ( $\hat{f}(\rho, \theta)$ ) from Zernike moments, we can obtain the following by the orthogonality of Zernike polynomials as previously shown in [28]:

$$\hat{f}(\rho, \theta) = \sum_{n=0}^{n_{max}} \sum_m ZM_n^m Z_n^m(\rho, \theta) \quad (5)$$

The number of ZMs, i.e. dimension of the array, **ZM**, formed by them, depends on the order  $n_{max}$  of radial Zernike polynomials considered. It can be seen in Table 2 that the number of azimuthal Zernike polynomials increases dramatically when the radial order of Zernike polynomials (DZM), which is equal to  $n_{max}$ , increases from 0 to 100. It is worthy of note that different DZMs produce arrays of ZMs with different numbers of elements. In turn, the useful information contained in the ZMs depends on their dimension. However, the amount of useful information contained in ZM does not essentially monotonically increase with increasing DZM [28,40–42]. We used image reconstruction from ZMs to evaluate each order of ZM in section 4.1.1.

**Table 2.** DZM and number of elements in (i.e. dimension of) the corresponding ZM array

DZM	Number of elements in ZM's array, <b>ZM</b> , for ( $m \geq 0$ )*
0	1
1	2
2	4
3	6
4	9
6	16
8	25
10	36
15	72
20	121
30	256
60	961
90	2116
95	2352
100	2601

\*Only the ZMs for  $m \geq 0$  are considered because  $|ZM_n^{-m}| \sim |ZM_n^m|$  for any given pair of  $n$  and  $m$ .

### 3.1.2 Features for the Characterization of a Particle

In addition to the ZMs, two other features – summed intensity, and root mean square contrast (RMS) – derived from untransformed intensities, are found to be important and are used in the training of the machine learning models.

Unlike the ZMs, which were calculated from ln-transformed intensities, summed intensities and RMS contrasts were calculated directly from the original intensities of the 2DLSPs. The summed intensity for a given 2DLSP is the summation of the intensities at all its pixels. RMS Contrast was calculated as square root of the sum over all pixels of the squares of the deviations of pixel intensity,  $I_n$ , from the mean intensity,  $\bar{I}$ , divided by the total number of pixels,  $N$  (where  $N \gg 1$ ), in the 2DLSP of interest, as shown in equation (6).

$$RMS \text{ Contrast} = \sqrt{\frac{1}{N} \sum_{n=1}^N (I_n - \bar{I})^2} \quad (6)$$

### 3.2 Random forests classification (RFC)

Random forests are an ensemble classifier that consists of multiple decision trees. One type of widely used decision trees is called CART [43]. Ho [44] first proposed random forests. In [44] each tree is built in randomly selected subspaces of the feature space. This means one can use a randomly selected feature vector when constructing a tree. Breiman [43] further developed the algorithm by combining the idea of randomly selecting features and the technique of bootstrap aggregating to the training sets. This approach has been widely adopted [45,46].

The RFC algorithm can be summarized as follows:

For each tree:

1. Draw a bootstrap sample from the data. Those not in the bootstrap sample are called the “out-of-bag” data.
2. Construct a tree, where at each node, the best split is chosen among a number of randomly selected features. The tree is grown to the maximum size and not pruned back.
3. Use the tree to predict out-of-bag data, which is used to estimate the classification error.

To predict the class label for a new pattern, the pattern is pushed down the tree and is assigned the label of the training sample in the terminal node it ends up in. This procedure is iterated over all trees, and the modal vote of all the trees is reported as the random forests classification’s result.

### 3.3 Support Vector Machine

Support Vector Machine (SVM) has been widely used in the machine learning field over the past two decades. SVM can be applied in both classification and regression problem domains.

#### 3.3.1 Support vector classifier (SVC)

The basic idea of SVC is to find the maximal-margin decision hyperplane, such that the distance of the closest data points to the hyperplane is maximized. Given a training dataset  $\{\mathbf{x}_n, t_n\}_{n=1, \dots, N}$ , where  $t_n$  is the target for the corresponding input  $\mathbf{x}_n$ , one would be interested in creating an accurate model that can assign each new input  $\mathbf{x}$  to one of the classes,  $c_i$ , where  $i = 1, \dots, C$ . If  $y$  denote the predictor of each input, then the very general form SVC decision function [47] is

$$y(\mathbf{x}) = \sum_{n=1}^N \alpha_n t_n k(\mathbf{x}, \mathbf{x}_n) + b \quad , \quad (7)$$

subject to constraints  $\sum_{n=1}^N \alpha_n t_n = 0$  and  $0 \leq \alpha_n \leq A$ , where  $b$  is a threshold, the  $\alpha_n$  are Lagrange multipliers in the *constrained optimisation problem*, and  $A$  is a constant for determining the trade-off between the minimization of the training error and the maximization of the margin. The kernel function  $k(\mathbf{x}, \mathbf{x}_n)$  defines a similarity measure for  $\mathbf{x}$  and  $\mathbf{x}_n$ , and implicitly maps the data points into a higher-dimensional feature space, and takes the inner-product in that feature space. The use of a kernel function often allows the data to be linearly separable in the feature space. We used a Radial Basis kernel Function (RBF) in our experiments. During the training, only a few  $\alpha_n$  are non-zero and the entries with non-zero  $\alpha_n$  are called the *support vectors*.



### 3.3.2 Support vector regression (SVR)

The  $\varepsilon$ -SVR [28] is used in this study. Given a training dataset,  $\{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)\}$ ,  $\varepsilon$ -SVR aims to fit a function,  $f_{\mathbf{x}}$ , based on the available training dataset such that the difference between the target value,  $y_i$ , and the estimated value,  $\hat{y}_i$ , is no larger than  $\varepsilon$  for the  $i^{\text{th}}$  training example. This can be expressed as follows.

$$\min_{\mathbf{w}, b, \xi, \xi^*} \tau(\mathbf{w}, \xi, \xi^*, b) = \frac{1}{2} \mathbf{w}^T \mathbf{w} + C \sum_{i=1}^n \xi_i + C \sum_{i=1}^n \xi_i^* \quad (8)$$

$$\text{subject } \begin{cases} \mathbf{w}^T \Phi(\mathbf{x}_i) + b - y_i \leq \varepsilon + \xi_i, \\ y_i - \mathbf{w}^T \Phi(\mathbf{x}_i) - b \leq \varepsilon + \xi_i^*, \\ \xi_i, \xi_i^* \geq 0, i = 1, \dots, n. \end{cases}$$

where  $\mathbf{w}$  is a weight vector,  $b$  is the bias and  $\Phi(\mathbf{x}_i)$  maps the input  $\mathbf{x}_i$  to the feature space.  $C$  is a constant (and  $C \geq 0$  in all cases) named the regularization parameter.  $C$  determines the trade-off between the soft margins described by the constraints and the amount up to which differences larger than  $\varepsilon$  are tolerated.  $\xi_i$  and  $\xi_i^*$  are slack variables which are used to relax the constraints slightly to allow for bad estimations. Lagrange multipliers  $\alpha$  are applied to produce predictions for data points. The solution for the estimation at each new point,  $\mathbf{x}_*$ , is determined by:

$$\hat{y}_* = \sum_{i=1}^n (-\alpha_i + \alpha_i^*) \Phi(\mathbf{x}_i)^T \Phi(\mathbf{x}_*) + b \quad (9)$$

In practice, the solution for the estimation at the new point,  $\mathbf{x}_*$ , is given by:

$$\hat{y}_* = \sum_{i=1}^n (-\alpha_i + \alpha_i^*) k(\mathbf{x}_i, \mathbf{x}_*) + b \quad (10)$$

where

$$k(\mathbf{x}_i, \mathbf{x}_j) = \Phi(\mathbf{x}_i)^T \Phi(\mathbf{x}_j) \quad (11)$$

In this work, radial basis function (RBF) kernel is applied to the data. Further information on the use of this approach, and the calculation of  $b$ , can be found in [48].

### 3.4 Performance measurements

Each of the machine-learning classifiers or predictors was trained with two-thirds of the dataset (the training set), and evaluated with the remaining one-third of the dataset (test set). For the machine-learning classifiers, accuracy, and F1 score (eq. 11) were used as performance metric. Accuracy as well as F1 score for a given machine-learning classifier can have values between zero (worst possible performance) and one (best possible performance). Coefficient of

determination ( $R^2$ ) and normalized mean square error (NMSE) were used as performance metric for evaluating the machine-learning predictors.  $R^2$  for a given machine-learning predictor can have values between zero (worst possible performance) and one (best possible performance). NMSE also have values that range from zero to one, but with a value of zero corresponding to best performance and one to worst performance.

$$F1score = \frac{2 * Recall * Precision}{Recall + Precision} \quad (12)$$

where

$$Recall = \frac{TruePositive}{(TruePositive) + (FalseNegative)}$$

and

$$Precision = \frac{TruePositive}{(TruePositive) + (FalsePositive)}$$

## 4. Experimental Results and Discussion

To investigate whether the proposed framework works properly, we have carried out the following three experiments. Firstly, we validate the reliability of Zernike moments. Secondly, we apply RFC to the representations of 2DLSPs to predict the orientation for each given test pattern. Finally, we created 133 orientation-specific support vector regression models for predicting size of hexagonal prisms, 133 orientation-specific support vector regression models for predicting size PAD of hexagonal prisms, and 133 orientation-specific support vector classifiers for predicting aspect ratio of hexagonal prisms.

Most of the computations have been done using the Python programming language and Scikit-learn [26] – a well-documented Python machine-learning toolbox. With multidimensional grid search in Scikit-learn [26], parameter spaces have been explored using a 10-fold cross-validation. Parameters that produced best models have been retained as the optimal model parameters.

### 4.1 Studies on Zernike Moments

#### 4.1.1 Reliability of the Zernike Moments

First, we randomly selected 17 particles (10% of the 162 particles, at  $\beta=10$ ,  $\gamma=10$ ) and rotated them through 50, 100, 150, 200, 250, 300 and 350 degrees. Table 3 lists the means, the standard deviations, and the coefficient of variations for the ZMs of the particles, for DZMs of 2, 4, 5, 20, 35, 50, 65, 80, and 95. The coefficient of variation is obtained by dividing the standard deviation by the corresponding mean value. It can be seen that the coefficients of variation are reasonably small when the order of ZMs not greater than 65, especially for those orders less than 20, the corresponding coefficient of variation are consistently smaller than 0.1. When the order of ZM equals to 80 or 95, the coefficient of variation is greater than 0.2. However, we notice that their ZM values are also very small since the value of ZMs is dramatically decreased as the order of ZMs increases. Overall, the ZMs from the original un-rotated images are (within a very narrow margin of error) similar to those from the rotated images (Table 3).

Furthermore, we reconstructed each image from its ZMs and checked how similar an image reconstructed from ZMs is to the original image by calculating the Pearson correlation between the intensities of the original image and the intensities of the reconstructed image. The order of ZMs we have used is from 1 to 105. This is an intuitive way of assessing the reliability of the information retained in the ZMs. We show examples of such image reconstruction from ZMs alongside the Pearson correlation coefficients between the original images and the reconstructed images in Table 4.

**Table 3.** Comparison of the ZM of unrotated and rotated images

Rotation Angle	ZM <sub>2</sub> <sup>0</sup>	ZM <sub>2</sub> <sup>2</sup>	ZM <sub>3</sub> <sup>1</sup>	ZM <sub>3</sub> <sup>3</sup>	ZM <sub>4</sub> <sup>4</sup>	ZM <sub>5</sub> <sup>5</sup>	ZM <sub>20</sub> <sup>20</sup>	ZM <sub>35</sub> <sup>35</sup>	ZM <sub>50</sub> <sup>50</sup>	ZM <sub>65</sub> <sup>65</sup>	ZM <sub>80</sub> <sup>80</sup>	ZM <sub>95</sub> <sup>95</sup>
0°	3.49e+03	2.62e+01	4.12e+02	6.09e+00	3.94e+00	2.81e+00	2.20e-06	1.29e-11	9.95e-18	4.96e-23	5.66e-29	1.47e-34
50°	3.48e+03	2.48e+01	4.11e+02	6.57e+00	3.69e+00	2.83e+00	2.52e-06	1.30e-11	1.08e-17	4.36e-23	4.35e-29	1.61e-34
100°	3.48e+03	2.72e+01	4.07e+02	6.12e+00	3.84e+00	2.79e+00	2.74e-06	1.36e-11	1.16e-17	4.87e-23	4.43e-29	1.58e-34
150°	3.47e+03	2.67e+01	4.05e+02	6.49e+00	3.92e+00	2.71e+00	2.52e-06	1.26e-11	1.29e-17	4.57e-23	5.23e-29	1.50e-34
200°	3.47e+03	2.54e+01	4.07e+02	6.21e+00	3.79e+00	2.70e+00	2.40e-06	1.25e-11	1.05e-17	6.00e-23	6.72e-29	2.99e-34
250°	3.47e+03	2.54e+01	4.14e+02	6.01e+00	3.68e+00	2.67e+00	2.56e-06	1.23e-11	9.04e-18	5.50e-23	6.13e-29	2.45e-34
300°	3.46e+03	2.90e+01	4.20e+02	7.22e+00	4.00e+00	2.86e+00	3.36e-06	1.15e-11	1.46e-17	5.35e-23	8.04e-29	2.37e-34
350°	3.45e+03	2.71e+01	4.25e+02	5.88e+00	4.11e+00	3.04e+00	2.76e-06	1.26e-11	1.01e-17	5.17e-23	4.96e-29	1.61e-34
Mean	3.47e+03	2.64e+01	4.13e+02	6.32e+00	3.87e+00	2.80e+00	2.63e-06	1.26e-11	1.12e-17	5.10e-23	5.69e-29	1.95e-34
Std*	1.25e+01	1.25e+00	6.35e+00	4.03e-01	1.41e-01	1.10e-01	3.21e-07	5.75e-13	1.67e-18	4.93e-24	1.17e-29	5.37e-35
CoV**	0.004	0.047	0.015	0.064	0.036	0.039	0.122	0.046	0.149	0.097	0.206	0.276

\* Std=Standard deviation; \*\* CoV = Coefficient of Variation = std/mean

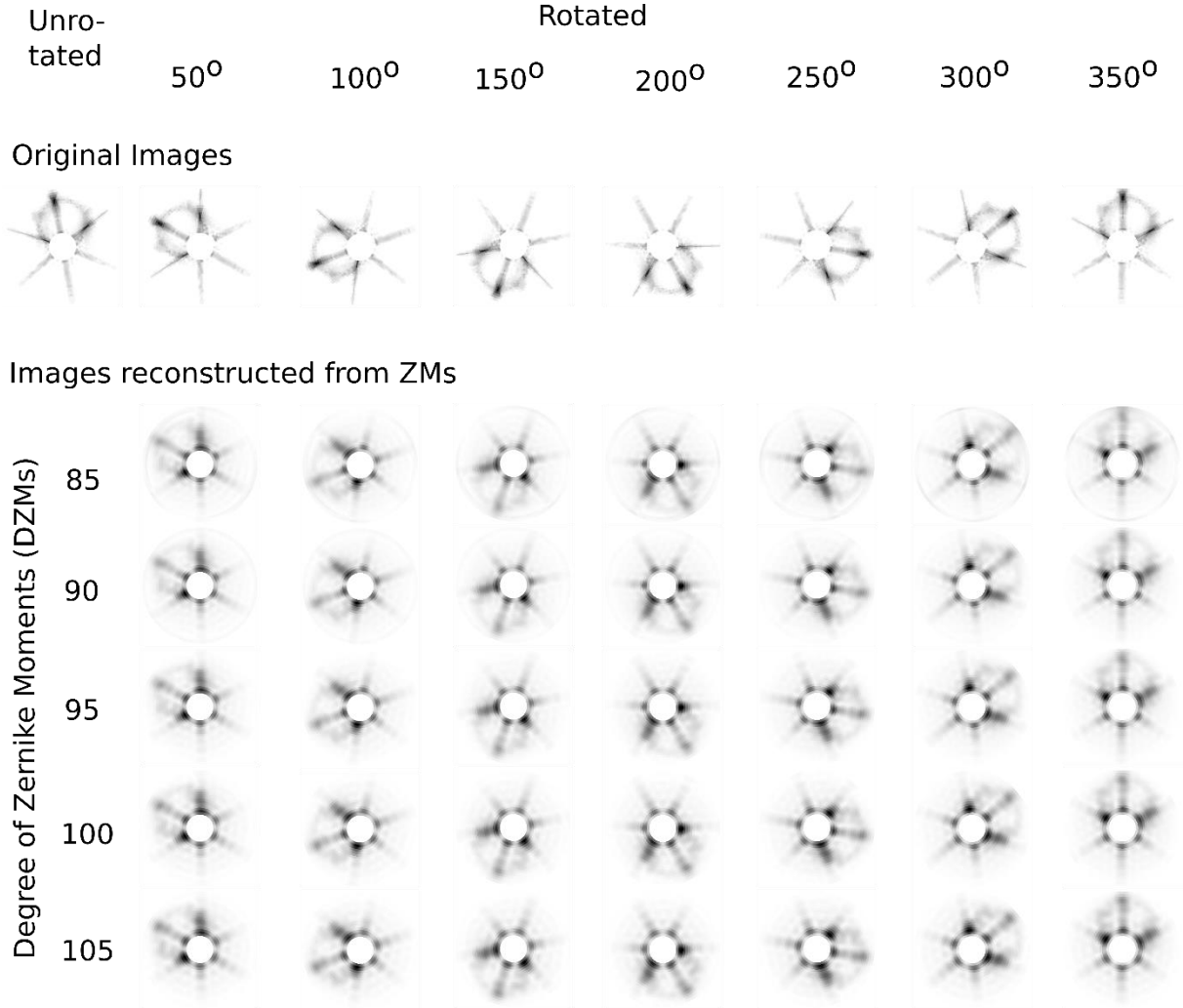
We observe a correlation of up to 0.76 between the original image and the images reconstructed from the ZMs with order of 95 across all different rotations. The data shown in Table 4 is averaged over 17 (~10% of the 162) particles. An example of image reconstruction from ZMs before and after rotation is shown in Fig. 7 for a hexagonal prism with a length of 108.9  $\mu\text{m}$ , and a diameter of 36.3  $\mu\text{m}$  at orientation  $\beta=10^\circ$ ,  $\gamma=10^\circ$ . Table 5 shows correlation coefficients for this particle only. It can be seen that the highest correlation value is obtained with the order of either 95 or 100 for this particle.

Table 4. Correlation between intensities of original and reconstructed images

DZM	Un-rotated	Rotated 50°	Rotated 100°	Rotated 150°	Rotated 200°	Rotated 250°	Rotated 300°	Rotated 350°
1	0.126	0.126	0.125	0.124	0.124	0.126	0.129	0.132
2	0.427	0.420	0.412	0.404	0.399	0.391	0.382	0.370
4	0.443	0.435	0.426	0.418	0.412	0.404	0.394	0.382
8	0.477	0.467	0.456	0.446	0.438	0.428	0.418	0.403
16	0.513	0.497	0.483	0.470	0.460	0.448	0.437	0.423
20	0.515	0.501	0.488	0.477	0.470	0.462	0.455	0.447
40	0.638	0.630	0.624	0.619	0.615	0.612	0.610	0.609
60	0.702	0.699	0.695	0.693	0.692	0.691	0.692	0.692
85	0.751	0.745	0.741	0.739	0.738	0.738	0.741	0.744
90	0.755	0.751	0.748	0.746	0.746	0.748	0.750	0.754
<b>95</b>	<b>0.760</b>	<b>0.756</b>	<b>0.754</b>	<b>0.752</b>	<b>0.753</b>	<b>0.754</b>	<b>0.756</b>	<b>0.760</b>
100	0.739	0.742	0.745	0.748	0.753	0.757	0.760	0.766
105	0.534	0.518	0.550	0.557	0.546	0.537	0.452	0.439

\* DZM: Order of Zernike Moment

From above results, we have seen that as the order of ZMs increases from 1 to 95, the Pearson correlation between the original image and the constructed image increases from 0.12 to 0.76. Although values of ZMs with larger orders are small, they contribute significantly to the reconstructed images. It suggests either 95 or 100 can be used as a suitable value for the order of ZMs in this application. However, the bigger the order of ZMs, the larger the size of the ZM array. Therefore, we choose 95 for the order of ZMs in the following experiments.



**Figure 7.** Construction of rotated images from ZMs of original un-rotated images. ZMs obtained from unrotated image (2DLSPs in this case) is able to reliably reconstruct rotated images. The example shown here is for a hexagonal prism with a length of  $108.9 \mu\text{m}$ , and a diameter of  $36.3 \mu\text{m}$  at orientation  $\beta=10^\circ$ ,  $\gamma=10^\circ$ .

**Table 5.** Correlation between intensities of original and reconstructed images shown in Fig. 7

DZM	Rotated $50^\circ$	Rotated $100^\circ$	Rotated $150^\circ$	Rotated $200^\circ$	Rotated $250^\circ$	Rotated $300^\circ$	Rotated $350^\circ$
85	0.739	0.738	0.736	0.735	0.739	0.739	0.739
90	0.741	0.741	0.741	0.741	0.746	0.747	0.748
<b>95</b>	<b>0.744</b>	<b>0.745</b>	0.746	0.747	0.751	0.751	0.752
100	0.737	0.744	<b>0.748</b>	<b>0.752</b>	<b>0.758</b>	<b>0.757</b>	<b>0.758</b>
105	0.595	0.651	0.658	0.639	0.629	0.582	0.559

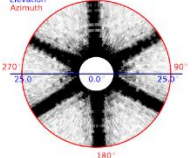
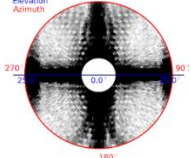






\* DZM: Order of Zernike Moment

#### 4.1.2 Relationship between the Overall Symmetry/Patterns of 2DLSP and ZMs

From our careful observation of Zernike polynomials (Table 6), it became apparent that some of the ZMs have high values due to the symmetric nature of a given 2DLSP. Therefore, we investigated  $Z_n^m$  that satisfies  $m = n$ . Our experiments show that indeed  $Z_n^m$  that satisfies  $m = n$  can describe the 6/6, and 4/4 symmetry nature of our dataset. We present the results in

Table 6. It is observed that values of ZM are generally higher whenever the symmetry nature of the image whose ZM is being computed has similar symmetry pattern as the given Zernike polynomial (because the Zernike polynomials form the basis for the ZMs). These can explain why the 4/4-symmetric 2DLSP has higher values for  $ZM_2^2$  and  $ZM_4^4$  compared to the 6/6-symmetric 2DLSP, and 6/6-symmetric 2DLSP has higher values for  $ZM_3^3$  and  $ZM_6^6$  compared to the 4/4-symmetric 2DLSP (Table 6).

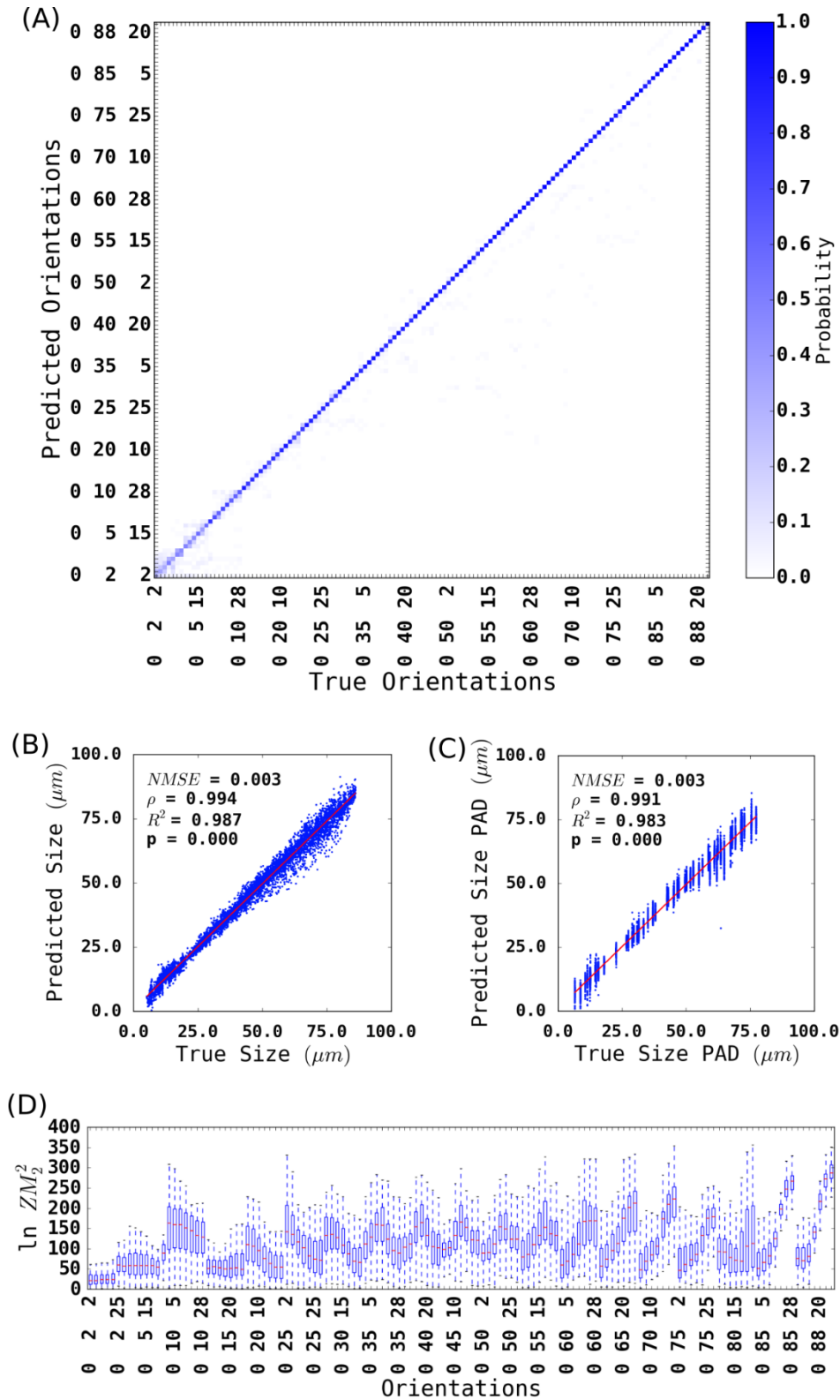
Table 6. ZMs for selected 6/6, 4/4, and 2/2 symmetry particles

DZM	Zernike Polynomial	6/6 Symmetry	4/4 Symmetry
		Orientation: 0.0 2.0 2.0  l=196.9, d=24.6, $\beta=2$ , $\gamma=2$	Orientation: 0.0 88.0 2.0  l=19.2, d=19.2, $\beta=88$ , $\gamma=2$
$ZM_1^1$		13.87	144.25
$ZM_2^2$		34.72	94.8
$ZM_3^3$		15.09	7.2
$ZM_4^4$		1.66	123.03
$ZM_5^5$		6.10	5.05
$ZM_6^6$		20.32	7.16
$ZM_8^8$		6.93e-02	6.29e-01
$ZM_{10}^{10}$		2.25e-02	6.67e-02
$ZM_{12}^{12}$		5.12e-02	3.63e-02
$ZM_{16}^{16}$		1.07e-04	1.05e-03
$ZM_{20}^{20}$		6.45e-07	5.73e-07
$ZM_{24}^{24}$		2.88e-05	2.30e-05
$ZM_{35}^{35}$		6.42e-12	9.78e-12
$ZM_{36}^{36}$		5.71e-11	6.67e-12
$ZM_{48}^{48}$		5.87e-16	2.40e-16
$ZM_{50}^{50}$		1.13e-16	1.09e-16
$ZM_{65}^{65}$		4.77e-22	7.37e-23
$ZM_{80}^{80}$		3.57e-28	4.62e-28
$ZM_{95}^{95}$		2.62e-33	1.35e-33
$ZM_{96}^{96}$		8.78e-34	1.04e-33

## 4.2 Performances of the Machine-Learning Models

Our computational characterization of the particles is two-staged. The first stage involves prediction of particle orientation. The predicted orientation is then used to select the appropriate models for predicting the particle's size, size PAD, and aspect ratio in the second stage of the characterization. The random forests classifier that had been trained with 14,362 training examples was tested with a test set consisting of 7,182 randomly pre-selected 2DLSP (which were not part of with the training set). Our model achieves an accuracy of 0.84 (F1-score = 0.85). The 133 by 133 matrix showing the true versus the predicted orientations is represented as a heatmap in Fig. 8A. It can be seen that some orientations with lower values of  $\beta$  are difficult to discriminate. Further investigations into this shows that the ZMs for the 2DLSPs for particles at orientations with lower values of  $\beta$  (e.g. for  $\beta=2$ ) have relatively lower variance across  $\gamma$  and across the particles (e.g. Fig. 8D, for  $ZM_2^2$ ), hence the difficulty in predicting their orientations.

The SVR model for estimating particle size PAD achieves a coefficient of determination,  $R^2$ , of 0.983, normalized mean square error of 0.003, and correlation of 0.991 between the true and the estimated (predicted) size PAD (Fig. 8B). The SVR model for estimating particle size achieves a coefficient of determination,  $R^2$ , of 0.987, normalized mean square error of 0.003, and correlation of 0.994 between the true and the estimated (predicted) size (Fig. 8C).



**Figure 8. Performances of the machine-learning models.** (a) The 133 by 133 matrix showing the true versus the predicted orientations. The 133 orientations are defined by the Euler angles  $\alpha$ ,  $\beta$ , and  $\gamma$  (in degree) as described in Fig. 1. (b) Performance of SVR model for estimating size. (c) Performance of SVR model for estimating size PAD. (d) Boxplot showing the distribution of the  $\ln(ZM^2_2)$  for all the 162 particles and across the 133 orientations defined by the Euler angles  $\alpha$ ,  $\beta$ , and  $\gamma$  as described in Fig. 1.



Our SVC model for predicting aspect ratio achieves an accuracy of 0.99 (F1-score = 0.99). We show the associated correct and in-correct classifications in Table 7.

Table 7. Performance of the support vector classifier in the aspect-ratio-based classification problem

		Predicted label		
		Aspect ratio $\leq 1$	Aspect ratio $> 1$	Marginal total
True label	Aspect ratio $\leq 1$	3813	44	3857
	Aspect ratio $> 1$	44	3281	3325
	Marginal total	3857	3325	7182

## 5.0 Conclusions

A machine learning based computational method has been proposed in this work for estimating a range of properties for hexagonal prisms from their 2DLSP (elevation and azimuthal angle ranges are  $6^\circ$  to  $25^\circ$  and  $0^\circ$  to  $360^\circ$ , respectively, as covered by the SID instrument). The method has been validated on a dataset including 162 particles with 133 orientations, generated using the Ray Tracing with Diffraction on Facets model.

Our experimental results show that the use of the combination of the random forests classification and the support vector method can provide good estimations for both the particle size PAD and the particle size with a normalized mean square error of 0.003, and correlation coefficients of 0.99, although we found a slightly weaker accuracy rate for identifying particle's orientation, especially for particles having a lower value of  $\beta$ . The method can also give good predictions for aspect ratio with an accuracy rate of 0.99. These results suggest that advanced machine learning methods can be potentially used for estimating properties of single particles from their 2D scattering patterns.

The use of Zernike moments is a key stage in the proposed work, since they are particularly suitable for the analysis of scattering patterns of particles with high symmetry, such as ice crystals. The highest correlation between the original images and the reconstructed images is obtained when we used the degree of Zernike moments of 95.

Overall, this work contributes to the existing body of knowledge in interpreting light scattering patterns, i.e. solving the 'inverse problem' of deriving particle geometry from its scattering pattern. Here we have concentrated on the specific case of hexagonal columns. This led to the creation of computational models with which sizes, size PAD, and aspect ratios of atmospheric particles can be estimated for a given 2DLSP. The Python Programming Language codes developed in this work can also serve as the basis for others who may be interested in computing Zernike moments using Python Programming Language and for those who seek to develop additional computational models for studying other characteristics (such as surface roughness, shape, etc.) of hexagonal prisms and other atmospheric particles from their two-dimensional light scattering patterns and similar data. Such an extended and refined method would need to be validated against light scattering data measured at particles which are well characterized with respect to the parameters under investigation.

## Acknowledgement

E O Salawu acknowledges financial supports from Taiwan International Graduate Program, Academia Sinica, Taipei, Taiwan, and from National Tsing Hua University, Hsinchu, Taiwan. E Hesse acknowledges support by the Natural Environment Research Council (NERC), United Kingdom, grant NE/I020067/1.

## References

- [1] IPCC. Climate Change 2013 - The Physical Science Basis. *Clim Chang* 2014 Synth Rep 2014;151. doi:10.1017/CBO9781107415324.
- [2] Baran AJ. From the single-scattering properties of ice crystals to climate prediction: A way forward. *Atmos Res* 2012;112:45–69. doi:10.1016/j.atmosres.2012.04.010.
- [3] Lawson RP, Korolev AV, Cober SG, Huang T, Strapp JW, Isaac, GA. Improved measurements of the drop size distribution of a freezing drizzle event. *Atmos Res* 1998; 47–48:181–191.
- [4] Kaye PH, Hirst E, Greenaway RS, Ulanowski Z, Hesse E, Demott PJ, et al. Classifying atmospheric ice crystals by spatial light scattering. *Opt Lett* 2008;33:1545–7. doi:10.1364/OL.33.001545.
- [5] Mishchenko MI, Travis LD, Lacis AA. *Scattering, absorption, and emission of light by small particles*. Cambridge university press; 2004.
- [6] Waterman PC. Matrix formulation of electromagnetic scattering. *Proc IEEE* 1965;53:805–12. doi:10.1109/PROC.1965.4058.
- [7] Mishchenko MI, Zakharova NT, Khlebtsov NG, Wriedt T, Videen G. Comprehensive thematic T-matrix reference database: A 2013-2014 update. *J Quant Spectrosc Radiat Transf* 2014;146, 349-354.
- [8] DeVoe H. Optical Properties of Molecular Aggregates. I. Classical Model of Electronic Absorption and Refraction. *J Chem Phys* 1964;41:393–400.
- [9] Yurkin MA, Maltsev VP, Hoekstra AG. The discrete dipole approximation: An overview and recent developments. *J Quant Spectrosc Radiat Transf* 2007;106:558–89.
- [10] Macke A, Mueller J, Raschke E. Single scattering properties of atmospheric ice crystals. *Journal of the Atmospheric Sciences* 1996;53:2813–25.
- [11] Muinonen K. Scattering of light by crystals: a modified Kirchhoff approximation. *Applied Optics* 1989;28:3044–50.
- [12] Yang P, Liou KN. Geometric-optics-integral equation method for light scattering by nonspherical ice crystals. *Applied Optics* 1996;35:6568–84.
- [13] Borovoi, AG, and Grishin, IA, Scattering matrices for large ice crystal particles, *J. Opt. Soc. Am. A* 20(2003), 2071-2080.
- [14] Hesse E, Mc Call DS, Ulanowski Z, Stopford C, Kaye PH. Application of RTDF to particles with curved surfaces. *J Quant Spectrosc Radiat Transf* 2009;110:1599–603. doi:10.1016/j.jqsrt.2009.01.011.
- [15] Bi L, Yang P, Kattawar GW, Hu Y, Baum BA. Scattering and absorption of light by ice particles: solution by a new physical–geometric optics hybrid method. *J Quant Spectrosc Radiat Transf* 2011:1492–508.
- [16] Yang, P and Liou, KN. Light scattering by hexagonal ice crystals: solutions by a ray-by-ray integration algorithm. *J. Opt. Soc. Am. A* 1997;14:2278-2289.
- [17] Ulanowski Z, Hirst E, Kaye PH, Greenaway R. Retrieving the size of particles with rough and complex surfaces from two-dimensional scattering patterns. *J Quant Spectrosc Radiat Transf* 2012;113:2457–2464.
- [18] Stopford C. Ice crystal classification using two dimensional light scattering patterns [PhD Thesis]. Hatfield, UK: University of Hertfordshire; 2010 Available at:

[http://researchprofiles.herts.ac.uk/portal/en/publications/ice-crystal-classification-using-two-dimensional-light-scattering-patterns\(f93115e5-8539-4d57-ae5c-30805c066dc6\).html](http://researchprofiles.herts.ac.uk/portal/en/publications/ice-crystal-classification-using-two-dimensional-light-scattering-patterns(f93115e5-8539-4d57-ae5c-30805c066dc6).html)

- [19] Baran AJ, Newman, SM. On the application of principal component analysis to the calculation of the bulk integral optical properties for radiation parameterizations in climate models. *Optics Letters* 2017;42:983-986.
- [20] Wang Z, Ulanowski Z, Kaye PH. On Solving the Inverse Scattering Problem with RBF Neural Networks: Noise-Free Case. *Neural Comput Appl* 2014;8:177–86. doi:10.1007/s005210050019.
- [21] P. Kaye, E. Hirst, and Z. Wang-Thomas. Neural-network-based spatial light-scattering instrument for hazardous airborne fiber detection. *Appl. Opt.*1997;36:6149–6156.
- [22] Genuer V, Gal O, Méteau J, Marcoux P, Schultz E, Lacot É, Maurin M, Dinten J-M. *Proc. SPIE* 9698. Optical elastic scattering for early label-free identification of clinical pathogens. *Advanced Biomedical and Clinical Diagnostic and Surgical Guidance Systems XIV*, 96980A (March 7, 2016); doi:10.1117/12.2211021
- [23] Zernike F. Beugungstheorie des Schneidenverfahrens und seiner verbesserten Form, der Phasenkontrastmethode. *Physica* 1934;1:689–704. doi:10.1016/S0031-8914(34)80259-5.
- [24] Koenderink JJ, van Doorn AJ. Phenomenological description of bidirectional surface reflection. *J Opt Soc Am A* 1998;15:2903. doi:10.1364/JOSAA.15.002903.
- [25] Koenderink JJ, Van Doorn AJ, Stavridi M. Bidirectional reflection distribution function expressed in terms of surface scattering modes. *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 1065, 1996, p. 28–39. doi:10.1007/3-540-61123-1\_125.
- [26] Teague, MR. Image analysis via the general theory of moments. *Journal of the Optical Society of America* 1980;70:920-930.
- [27] Belkasim SO, Shridhar M, Ahmadi M. Pattern recognition with moment invariants: a comparative study and new results. *Pattern Recognit* 1991;24:1117–38.
- [28] Khotanzad A, Hong YH. Invariant image recognition by Zernike moments. *Patt Anal Mach Intel, IEEE Trans* 1990;12:489–97.
- [29] Khotanzad A, Lu J-H. Classification of Invariant Image Representations Using a Neural Network. *T-Assp* 1990;38:1028–38.
- [30] Perantonis SJ, Lisboa PJG. Translation, Rotation, and Scale Invariant Pattern Recognition by Higher-Order Neural Networks and Moment Classifiers. *T-Nn* 1992;3:241–51.
- [31] Wang Y, Zhao Y, Chen Y. Texture classification using rotation invariant models on integrated local binary pattern and Zernike moments. *EURASIP J Adv Signal Process Springer* 2014;2014:1–12.
- [32] Mishchenko MI, Hovenier JW, Travis LD, editors. *Light scattering by nonspherical particles*. New York: Academic Press;2000.p.19
- [33] Breiman L. Random forests. *Mach Learn* 2001;45:5–32. doi:10.1023/A:1010933404324.
- [34] Priori D, de Sousa G, Roisenberg M, Stopford C, Hesse E, Salawu E, et al. Using Machine Learning Techniques to Recover Prismatic Cirrus Ice Crystal Size from 2-Dimensional Light Scattering Patterns. *Int. Conf. Artif. Neural Networks*, 2016, p. 372–9.
- [35] Warren SG, Optical constants of ice from the ultraviolet to the microwave, *Applied Optics* 1984;23:1206–1225.
- [36] Warren SG, Brandt RE, Thomas CG. Visible and near-ultraviolet absorption spectrum of ice from transmission of solar radiation into snow. *Applied Optics* 2006;45:5320-5334.
- [37] Um J, McFarquhar GM, Hong YP, Lee S-S, Jung CH, Lawson RP, and Mo Q. Dimensions and aspect ratios of natural ice crystals. *Atmos. Chem. Phys.* 2015;15:3933–3956.

- [38] Hirst E, Kaye PH, Greenaway RS, Field P, Johnson DW. Discrimination of micrometre-sized ice and super-cooled droplets in mixed-phase cloud. *Atmospheric Environment* 2001; 35: 33-47.
- [39] Taylor L. A beam tracing model for electromagnetic scattering by atmospheric ice crystals [PhD Thesis]. Hatfield, UK: University of Hertfordshire; 2016 Available at: <http://uhra.herts.ac.uk/handle/2299/17645>.
- [40] Liao SX, Pawlak M. On the accuracy of zernike moments for image analysis. *IEEE Trans Pattern Anal Mach Intell* 1998;20:1358–64. doi:10.1109/34.735809.
- [41] Papakostas GA, Boutalis YS, Papaodysseus CN, Fragoulis DK. Numerical error analysis in Zernike moments computation. *Image Vis Comput* 2006;43:960–9. doi:10.1016/j.imavis.2006.02.015.
- [42] Li B, Zhang G, Fu B. Accurate computation and error analysis of pseudo-Zernike moments. *ICETC 2010 - 2010 2nd Int. Conf. Educ. Technol. Comput.*, vol. 4, 2010, p. V4--85. doi:10.1109/ICETC.2010.5529727.
- [43] Breiman L. *Algorithm CART. Classif Regres Trees Calif Wadsworth Int Group, Belmont, Calif* 1984.
- [44] Tin Kam Ho. Random decision forests. *Proc. 3rd Int. Conf. Doc. Anal. Recognit.*, vol. 1, 1995, p. 278–82. doi:10.1109/ICDAR.1995.598994.
- [45] Sofiyanti N, Fitmawati DI, Roza AA. *Stenochlaena Riauensis (Blechnaceae), A new fern species from riau, Indonesia. Bangladesh J Plant Taxon* 2015;22:137–41. doi:10.1007/s13398-014-0173-7.2.
- [46] Salawu EO. RaFoSA: Random forests secondary structure assignment for coarse-grained and all-atom protein systems. *Cogent Biol* 2016;2:1214061.
- [47] Schölkopf B, Smola AJ. *Learning with Kernels: Support Vector Machines, Regularization.* MIT press; 2001.
- [48] Fan R, Chan P, Lin C. LIBSVM: A library for support vector machines. *ACM Trans Intell Syst Technol* 2005;2:27.