Trinity College Dublin

The University of Dublin

Use of Principal Component Analysis for the Rapid Interpretation of Powder X-ray Diffraction Data.

James Ryan

ryanj52@tcd.ie

20335087 - Senior Sophister - Nanoscience

Supervisor: Prof. Peter Dunne, Assistant Professor in Inorganic Energy Materials

Submitted and presented to the School of Chemistry, Trinity College Dublin

September 2023 – January 2024

Declaration

This fourth-year Capstone report has been submitted to the School of Chemistry in Trinity College Dublin and has not been submitted to any other institution or university. The report is an original creation by me, except when stated otherwise.
I have read and I understand the plagiarism provisions in the General Regulations of the of the university calendar for the current year, 2024.

Date 20/01/2024

Signed: James Ryan

Abstract

As technology improves, allowing for more efficient synthesis methods, there comes an increase in size of material science datasets. Conventional methods such as Rietveld Refinement are becoming unfavoured due to time constraints. This makes way of other extraction techniques such as Principal Component Analysis. Principal Component Analysis (PCA) has been applied to X-ray diffraction (XRD) pattern datasets of three different (nano)crystalline systems: cerium oxide, cadmium sulfide, and barium strontium titanate. Each of these systems may exhibit a range of structural and microstructural properties discernible by XRD, such as unit cell parameter, composition, crystallite size, and phase percentage. With larger materials science datasets, conventional analytical approaches for the determination of these properties become less favoured due to time and skills required. PCA is a multivariable statistical technique which dimensionally reduces datasets with large numbers of variables to a smaller number of Principal Components. In isolation this allows ready comparison of how similar or different samples within a set may be, however principal components are not easily related to real physical properties. Here simulated XRD patterns with relevant varying properties for each material have been produced in order for PCA to provide a map where real material properties can be stated and defined. This work shows that each Principal Component (PC) can be linked to specific crystalline properties and hence each parameter can be estimated for each material produced. In the case cerium oxide, PC1 is linked to crystallite size of the particle while PC2 is correlated to unit cell parameter. Cadmium sulfide similarly shows PC1 being related to crystallite size and PC2 to phase percentage, however, presenting PC3 introduces a mathematical artifact which occurs within the covariance matrices in PCA. The same as ceria can be said for the barium strontium titanate dataset, where again PC3 is shown to be an artifact. Comparison of results from PCA to those obtained from Rietveld Refinement show the practically and potential use of PCA for materials discovery.

Acknowledgements

I would like to thank my supervisor Pr. Peter Dunne, as he warmly invited me to his team for my Capstone project. Despite his demanding role in the school of chemistry, Peter has always been exceptionally dedicated, setting aside time to offer advice, guidance, and knowledge throughout the project.

I would also like to thank Annie Regan and Karlijn Hertsig who are two PhD students in the Dunne Group. Although their fields differ to the one in this report, they were able to assist me elsewhere, such as carrying out safe and efficient laboratory practices.

I would like to mention Chloé Flandarin and Andrew Bathe for producing the cerium oxide and cadmium sulfide samples respectively, which were used and analysed in this report.

My sincere thanks go to Pr. Colm Delaney and his group, whom I shared an office during my time working on the project. Colm often went out of his way to answer any question I threw at him. Colm managed to create a great office environment for me to work in.

Finally, I am grateful for my family for financially and my friends for emotionally supporting me throughout this project.

Abbreviations

PC Principal Component

PCA Principal Component Analysis

SEM Scanning Electron Microscopy

TEM Transmission Electron Microscopy

XRD X-Ray Diffraction

Table of Contents

Table of Figures

Table of Tables

1 Introduction

Materials create the basis for everything we use in our lives. With the widespread and rapidly growing use and need for technology, materials science is becoming increasingly important, being involved in a wider range of industries than ever. Materials science is the study of the physical and chemical properties of solid-state materials, stemming from their composition and structure.[1] From their properties, materials can be selected or theorised for various applications, one such method of designing materials is done through computational chemistry. Computational methods are extremely accurate and time efficient, allowing for system optimisation and high throughput screening via automation.[2]

However, this level of automation is difficult to replicate. This is why high throughput synthesis methods for discovery optimisation are becoming popular, such as robotic batch, and continuous flow reactors. Robotic batch systems are capable of navigating about a given chemical space, based on the association between structures and reactivity, while also having the ability to process data from other analytical methods without have knowledge of the properties of the reagents.[3]

The use of continuous hydrothermal flow synthesis, which eases the scalability of chemical processes through adjusting the flow rate through the system, and is simple to adapt for different production levels.[4] With the advances in analytical technology, having the ability to better control reactions and witness processes in real-time eventually leads to the acceleration of data collection.

These synthesis methods are becoming more widespread, giving the opportunity for other areas of chemistry to achieve autonomous synthesis. As automation continues, larger datasets are generated. This leads to the increasing demand for data extraction tools and analysis devices, such as NMR, UV-vis spectroscopy, and X-Ray Diffraction (XRD). XRD is a fundamental technique used throughout chemistry as a characterisation tool for materials, determining crystal structure, crystallite size and other important properties.

1.1 Powder X-Ray Diffraction

Powder X-Ray Diffraction (PXRD) is a primary characteristic technique used to produce a diffraction pattern, which holds certain crystalline properties about the sample.

Unlike single crystal XRD, Powder XRD obtains a diffraction pattern from a powder of the given material and is favoured due the difficulty of acquiring single crystals.

X-Rays are emitted from a source with known frequency and wavelengths towards a sample, where they are scattered by the electrons of the atoms within the solid. For diffraction to occur within a sample, the material must be crystalline and have even and symmetrical spacing between atomic layers. A peak in the pattern occurs when constructive interference between scattered X-rays. The opposite is true for when they are out of phase, no peak is shown on the pattern[5]. This principle is the basis for Bragg's Law.

Bragg's Law is the relation between the angles of incidence and the distances between atomic planes which create an intense reflection of X-rays. For this reflection to occur, multiple incident waves must be in phase and must produce constructive interference. This gives the following relation[6] as seen in equation [1.1]:

$$n\lambda = 2dsin\theta \qquad [1.1]$$

where n is an integer representing the path difference for the X-rays, λ is the wavelength of the X-rays, d is the distance between two lattice planes and θ is the diffraction angle.

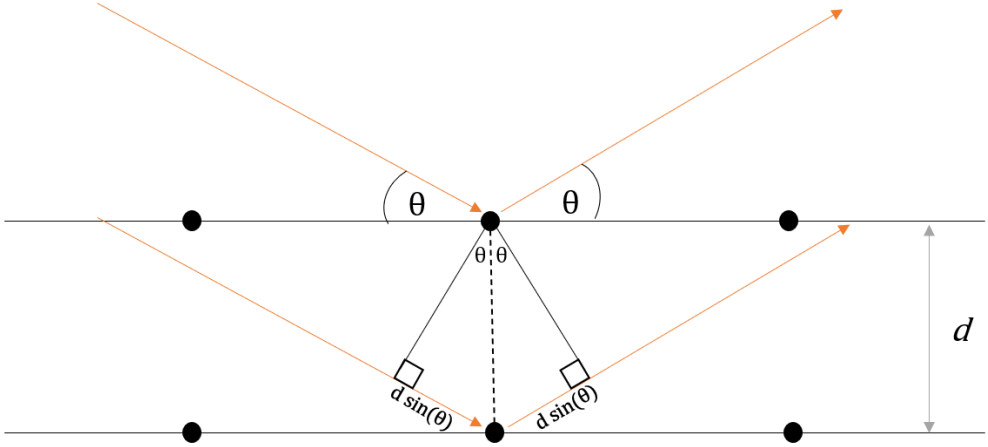


Figure 1.1: Bragg's Law, depiction of interaction between x-rays and crystal planes

Each peak in the pattern represents a set of planes, denoted by the Miller Indices. Miller indices describe the orientation of a set of atomic planes within a crystal lattice and are mathematically defined as the reciprocal intercepts caused by the plane and the crystallographic axes. For the case of the simple cubic unit cells, the lattice spacings, d, can be related to the miller planes, hkl, through the following:[7]

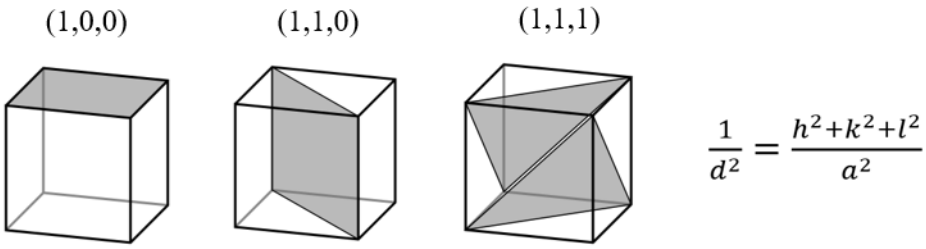


Figure 1.2: Miller Indices (hkl), for a Cubic crystal system and equation for lattice spacings

The crystallite size can also be closely associated with the XRD pattern, this relation is presented through the Scherrer equation [1.2] which states the following:

$$L = \frac{K\lambda}{\beta cos\theta} \qquad [1.2]$$

where L is the crystallite size, λ is the wavelength of the radiation source, β is the measurement of full width at half maximum for a given 2θ peak and K is the shape factor, which can lie anywhere from 0.62-2.08, however is usually taken as about 0.89.[8]

From above, an XRD Pattern can primarily determine the crystallite size and unit cell parameters. Conventionally, obtaining this information is done through software and particular peak analysis and refinement methods. The most common of which is Rietveld Refinement, which implements a least squares approach to refine a theoretical pattern until it matches as much as possible to the measured pattern.[9] This method is quite accurate and can be used on various types of crystal structures. However, with the rise of larger XRD datasets, Rietveld Refinement is becoming less attractive as an analysis tool as it is difficult to carry out and not suitable for researchers or students without specialised training in the field. This leads to the method being time consuming and prone to errors through inexperienced users.

With the increase in larger material science datasets, methods other than Rietveld Refinement are becoming increasingly popular, such methods include multivariable statistical methods particularly Principal Component Analysis (PCA).

1.2 Principal Component Analysis

Principal Component Analysis, PCA, is a dimensional reduction technique which is used to reduce the number of variables in a system by transforming the variables to a hierarchal set of combined variables, known as the principal components, which capture the maximum

covariance between these sets of variables. Essentially, it is used to reduce the number of variables in a system while maintaining as much information as possible. Although the new set of variables comes at the cost of accuracy, the smaller data set allows for better exploration and visualisation of the data. This allows for datasets which contain many variables to be reduced to two or three which can reliably represent a great proportion of the data.[10]

PCA can be carried out using a series of steps; normalisation, calculating the covariance matrix, computing the eigenvalues and eigenvectors to identify the principal components, transforming data into principal component space.

1.2.1 Normalisation

The data is normalised, typically within 0 and 1. This is done to ensure that each variable in the system contributes equally during the analysis, removing bias in the result. This is due to the PCA being sensitive in relation to any variance which may have been in the initial variables, if left unremoved, PCA will pick up these differences and cause unnecessary bias.[11]

1.2.2 Calculating the covariance matrix

The idea of PCA is to understand how each of the variables correlate or vary with one another with respect to the mean, i.e., to identify the relationship between said variables. The covariance, and hence the covariance matrix, is used to obtain these relationships. The covariance matrix is a symmetrical matrix which describes the covariance between pairwise elements in a random vector. The covariance, $\sigma_{(x,y),}$ between two variables is measured by the following equation [1.3].

$$\sigma(x, y, .., n) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})...(\mathrm{n_i} - \bar{\mathrm{n}}) \qquad [1.3]$$

If the covariance of a pair of variables is positive then the variables are correlated, i.e., increasing or decreasing together, If the covariance is negative then the variables are inversely correlated. The covariance matrix, C, can then be obtained by finding the covariances and arranging them into a symmetrical square matrix. Since the matrix is symmetrical the covariance along the diagonal is simply the variance squared of a single variable, as in [1.4].

$$C = \begin{bmatrix} \boldsymbol{\sigma_x^2} & \cdots & \boldsymbol{\sigma}(n,x) \\ \vdots & \ddots & \vdots \\ \boldsymbol{\sigma}(n,x) & \cdots & \boldsymbol{\sigma_n^2} \end{bmatrix} \qquad [1.4]$$

1.2.3 Principal Components from eigenvector and eigenvalue calculations

Principal components (PCs) are a new set of variables which are obtained from linear combinations of the original variables. The PCs are unrelated to one another and are set in a hierarchical order, where PC1 would capture the most information, PC2 captures the second most, and so on. The PCs are what allows for the information to be represented in a reduced dimension while much information is not lost. Graphically, the PCs can be interpreted as the new set of axes which capture the maximum amount of variance in the dataset while keeping the axes unrelated to each other, i.e., perpendicular.

The key element of obtaining the PCs is by computing the eigenvectors and eigenvalues of the covariance matrix. Which can be calculated using the following equations [1.5] and [1.6].

$$CX = \lambda X \qquad [1.5] \qquad\qquad \det|C - \lambda I| = 0 \qquad [1.6]$$

Where C is the covariance matrix, X is the eigenvector and λ is the eigenvalue.

Eigenvectors are vectors which have a constant direction during linear transformations, where it can be scaled by a constant factor during the transformation, this factor is known as the eigenvalue. Each dimension in the system, i.e., variable, has its own pair of eigenvectors and eigenvalues. The eigenvectors of the covariance matrix are the directions of the axes which capture the maximum variance, and the eigenvalues are the scale factors which state the amount of variance within each PC. Each eigenvector is ranked based on the eigenvalue, where the eigenvector with the highest eigenvalue is PC1 and the eigenvector with the second highest eigenvalue is PC2, and so on. The percentage of information in the PCs can be computed by obtaining the ratio of the corresponding eigenvalue over the sum of all the eigenvalues.

Each variable in the system will output an eigenvector and eigenvalue hence a PC. Typically, since most of the information are stored within the first three PCs, the remaining variables are not necessary to analyse as there is little information to be obtained. Although there is an extreme loss of variables, only a small amount of accuracy is lost.

1.2.4 Transforming the original data into Principal Component space.

The original data has still been left untouched, with the exception of normalisation, and need to be translated into PC space to be applicable to be represented by the PCs. The eigenvectors obtained from the covariance matrix are used to reorientate the initial data into

PC space. This is done by multiplying the initial dataset in matrix form by the eigenvectors of the wanted PCs.[12]

PCA for large datasets of XRD patterns can be used to determine key properties of each material. This is done by incidentally targeting known variables such as unit cell parameters, crystallite size or composition. The theory is to create a simulated layer of XRD patterns, of the material in question, with known properties which can be computed against the data of actual materials to determine their own properties, where each principal component will contain information of mostly one single variable.

1.3 Aims

PCA is to be used as a characterisation tool, however, unlike traditional methods will be aimed to perform more efficiency and can process larger materials science datasets. Three different (nano)crystalline systems are examined with PCA in attempt to test the capabilities of the analysis method with respect to various crystalline characteristics. The three crystals are cerium oxide (CeO_2), cadmium sulfide (CdS), and barium strontium titanate ($Ba_xSr_{1-x}TiO_3$). Variables such as unit cell parameters, crystallite size, composition and phase percentage of crystal structure are to be examined and mapped against relative simulated patterns.

1.3.1 Cerium Oxide

Cerium oxide is primarily used as a catalyst due to its unique properties such as having an abundant oxygen vacancy and supportive interactions with metals.[13] Typical industrial applications include polishing machinery and decolourising glass.[14] CeO_2 is shown to exhibit larger unit cell parameters when the crystallite size is decreased.[15] CeO_2 adopts the cubic fluorite structure at room temperature, where cerium occupies the fcc sites and the oxygen atoms occupy the tetrahedral sites. CeO_2 is relatively simple to characterise as no phase transitions occur during synthesis, hence the only properties examined through XRD and PCA are the unit cell parameters and crystallite sizes. A simulated layer of CeO_2 XRD patterns can be produced with crystallite size and unit cell parameters being the varying properties, where the real data obtained can be layer over the simulated to determine values for relative properties.

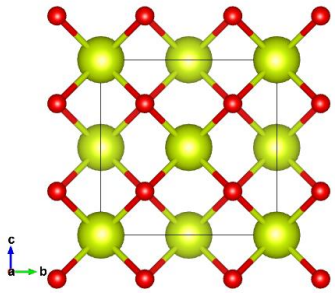


Figure 1.3: Cubic fluorite structure for CeO_2.[16] Obtained through Vesta[17] ,cerium in yellow and Oxygen in red

1.3.2 Cadmium Sulfide

Due to its narrow energy band gap and polymorphic behaviour, cadmium sulfide is an ideal material for quantum dots.[18] It is also one of the most popular sulfides used in photocatalysis. At room temperature, CdS exhibits two main crystal structures, a cubic zinc blende phase which is a face centred cubic structure with S occupying the half the tetrahedral sites and a hexagonal wurtzite phase which is considered to be two hexagonal close packed arrays penetrating one another.[19]

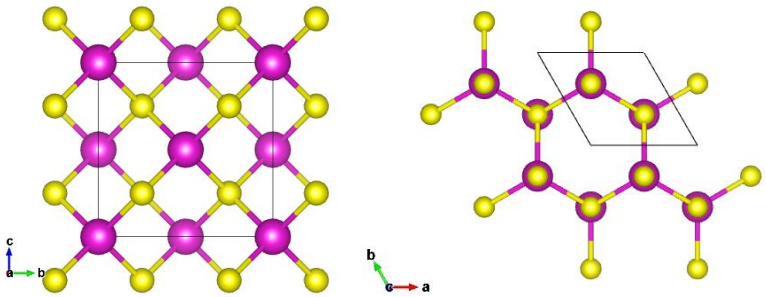


Figure 1.4: [Left] Cubic zinc blende structure and [Right] Hexagonal wurtzite structure[20] for CdS, obtained through Vesta[17], Cadmium in purple and Sulfur in yellow.

With two phases present at room temperature, this makes the phase composition of CdS an important variable to be examined during characterisation. Variables which will be examined include phase composition and crystallite sizes for independent phases. A layer of simulated CdS XRD patterns can be constructed with these targeted variables with an appropriate range to map the physical data overlaying the simulated.

1.3.3 Barium Strontium Titanate

$Ba_xSr_{(1-x)}TiO_3$ is typically ferroelectric and has a high dielectric constant, making the material useful for computer hardware such as DRAM and other applications with high frequency ranges.[21] $Ba_xSr_{(1-x)}TiO_3$ adopts the ABO_3 cubic perovskite structure, which contains

Ba/Sr at the vertices, oxygen at the centre of the faces and a titanium atom at the bodies centre of the unit cell or may alternatively be considered with the unit cell origin on the B (Ti) sites, as shown in Fig. 1.5. A tetragonal structure can be formed depending on the amount of barium present, as between 0 °C and 130 °C the barium titanate structure is tetragonal, and above 130 °C is cubic in shape. Barium and strontium share the primitive sites on the unit cell, where the dominant atom will be determined by weighted composition.[22] With an increasing in strontium content, unit cell parameters decrease, the opposite relationship is true for barium content.[23] Similarly, to CeO_2, unit cell parameters and crystallite size are both examined and characterised during PCA of the XRD patterns.

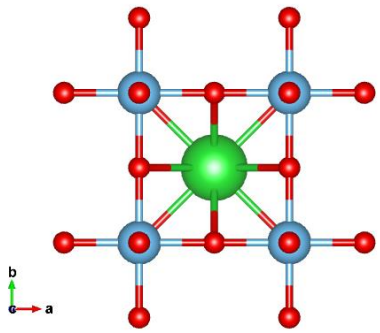


Figure 1.5: Cubic perovskite structure for $Ba/SrTiO_3$[24], obtained through Vesta[17], titanium in blue, oxygen in red and barium/strontium in green

These three materials are used to lay out a foundation of testing the viability of PCA in large material science datasets. Simulated patterns for each material will be plotted against the real physical data obtained. Based on the outcome and accuracy of the results, PCA may become the primary tool for analysis of XRD patterns throughout material science, allowing for a higher rate of materials discovery. Rietveld Refinement will be carried out as it is the traditional method of characterising XRD patterns and will be compared to the results obtained from PCA of the same dataset.

2 Experimental

2.1 Synthesis

Hydrothermal synthesis was the method used to produce all the crystalline materials mentioned above, as the growth of the crystals can (in principle) be precisely controlled, parameters such as temperature and time can be varied, and overall ease of scalability were ideal for the products needed in this experiment. Hydrothermal synthesis of a material can be described as the chemical reaction of precursor materials dissolved in an aqueous solution and placed under high pressure and high temperature conditions.[25] The main apparatus used was an autoclave, which was a Teflon lined stainless steel vessel able to withstand high pressures and temperatures. The autoclave had a volume of 23 cm^3.

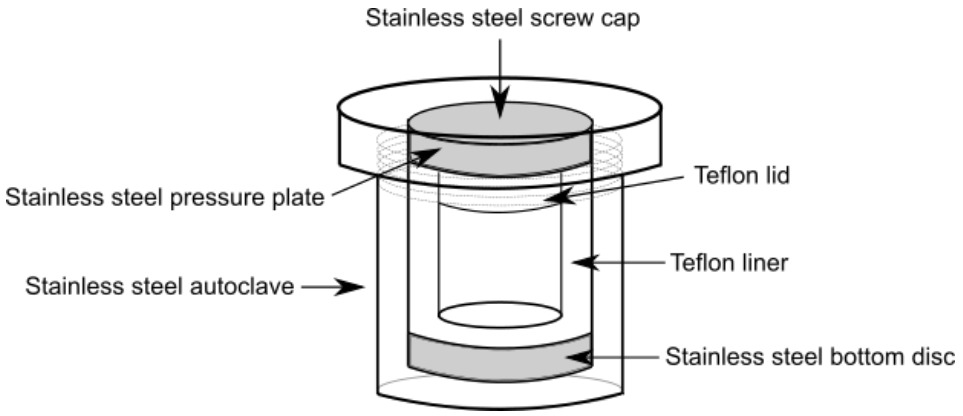


Figure 2.1: Diagram of autoclave used during synthesis.

2.1.1 Synthesis of CeO_2

Synthesis of CeO_2 was carried out by Chloé Flandrin,[26] where batch hydrothermal synthesis was carried out for 48 unique samples. Samples were prepared based on varying cerium precursor, sodium hydroxide ratio, urea content, time, and temperature, see Table 6.1 in appendix.

2.1.2 Synthesis of CdS

Synthesis for CdS was carried by Andrew Bathe[27] where three different synthetic methodologies were applied: conventional batch hydrothermal synthesis as above, and reactor batch and reactor injection hydrothermal synthesis methods, using a custom-designed and in-house built reactor system shown in Fig 2.2. The conventional batch hydrothermal method may be treated as a baseline synthesis method, while in comparison the batch synthesis using the reactor allows for faster heating and higher temperatures, while injection using the reactor

allows for effectively instantaneous heating of the reaction mixture. As the cubic and hexagonal polymorphs are closely related these changes in reaction methods were anticipated to significantly alter nucleation and growth pathways, leading to variations in phase composition, which has been confirmed by the previous analyses carried out by Dr Bathe. A total of 174 samples were collected, each unique, with varying cadmium precursor, sulfur precursor, ratio of Cd/S sources, temperature, time, and synthesis method.

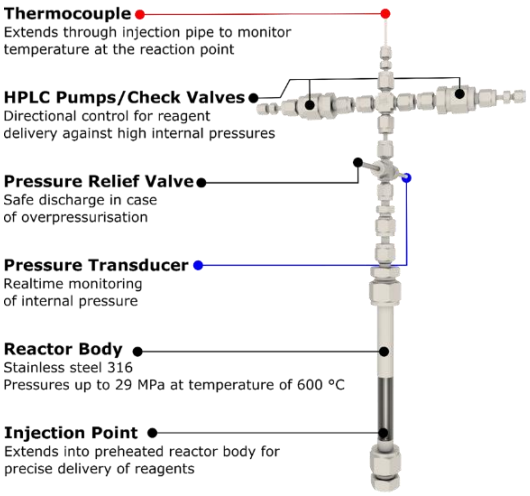


Figure 2.2: Diagram of in-house built reactor for batch and injection synthesis methods.

2.1.3 Synthesis of $Ba_xSr_{1-x}TiO_3$

24 samples of $Ba_xSr_{1-x}TiO_3$ were produced using conventional batch hydrothermal synthesis under three varying conditions, temperature, time, and sample composition. The following values for each variable were carried out.

Table 2.1: Variable ranges used to produce 24 samples of $Ba_xSr_{1-x}TiO_3$

Target composition	Temperature (°C)	Time (hours)
$BaTiO_3$	150	2, 4, 6, 24
	200	2, 4, 6, 24
$Ba_{0.5}Sr_{0.5}TiO_3$	150	2, 4, 6, 24
	200	2, 4, 6, 24
$SrTiO_3$	150	2, 4, 6, 24
	200	2, 4, 6, 24

The autoclave used had a volume of 23 ml, hence each reaction vessel was filled to 15 ml to ensure there was enough product to be examined and that the reaction would not fail. As a barium and strontium source, $Ba(NO_3)_2$ and $Sr(NO_3)_2$ were used. The water-soluble titanium complex titanium(iv) bis(ammonium lactato)dihydroxide, TiBALD, was used as a titanium

source. Standard solutions of $Ba(NO_3)_2$, $Sr(NO_3)_2$ and TiBALD were prepared at 0.5 M each, with a NaOH solution at 5 M. In a typical reaction, using clean and dry glassware 5 ml of the TiBALD solution was pipetted into the Teflon liner, followed by 5 mL of the desired $M(NO_3)_2$ solution (or 2.5 mL of each, when targeting $Ba_{0.5}Sr_{0.5}TiO_3$). 5ml of the 5M NaOH solution were slowly added, while stirring with a magnetic bar. The reaction mixture was further stirred for 60 seconds before sealing and assembling the autoclave. Each vessel was marked and placed into a preheated fan assisted oven, at a given temperature and held for a specific time. Once the reaction was completed the autoclaves were removed from the oven and allowed to cool to room temperature under ambient conditions. Solid samples were then isolated through centrifuging and washings. Samples were centrifuged, for five minutes at 2000 rpm, liquid was decanted and then washed with acetic acid. This process was repeated twice; however water was instead of acetic acid to wash the sample. The obtained products were dried in air overnight at 70 °C and crushed into a fine powder with a pestle and mortar.

2.2 Characterisation

2.2.1 Powder X-ray Diffraction

XRD patterns of CeO_2 and CdS samples were recorded using a Bruker D2 Phaser Diffractometer equipped with a LynxEye detector and a radiation source of Cu-Kα. A range of 15° to 85° 2θ with a step size of 0.01° was used for the CdS and $Ba_xSr_{1-x}TiO_3$ samples, while CeO_2 used a range of 15° to 90° 2θ with a step size of 0.01° using a zero-background Si sample holder. XRD patterns of $Ba_{(1-x)}Sr_xTiO_3$ were recorded using a Bruker D8 Advanced instrument with Cu-Kα radiation from 15° to 85° 2θ with a step size of 0.01°, using PMMA sample holders.

2.3 Simulation

2.3.1 Simulation of CeO_2 XRD patterns

Simulated XRD patterns were generated using GSASII[28] using a CeO_2 CIF obtained from the Crystallography Open Database (COD).[29] which exhibited the fluorite structure.[16] A total of 210 XRD patterns were simulated, with changing unit cell parameters from 5.4 Å to 5.5 Å in steps of 0.005 Å and crystallite size from 2.5 nm to 25 nm in steps of 2.5 nm.

2.3.2 Simulation of CdS XRD patterns

Simulated XRD patterns were generated using GSASII[28] using two CdS CIFs, zinc blende and wurtzite structures.[20] A total of 235 simulated patterns were prepared, varying hexagonal phase percentage from 0% to 100% in steps of 10%, crystallite size of the hexagonal phase from 5 nm to 25 nm in steps of 5 nm and crystallite size of the cubic phase from 5 nm to 25 nm in steps of 5 nm.

2.3.3 Simulation of $Ba_xSr_{1-x}TiO_3$ patterns

Simulated XRD patterns were generated using GSASII[28] using $BaTiO_3$ and $SrTiO_3$ CIFs which exhibited the perovskite structure.[24] A total of 440 XRD patterns were simulated, with changing unit cell parameters from 3.9 Å to 4.1 Å in steps of 0.02 Å, barium composition from 100% to 0% and crystallite size from 2 nm to 20 nm in steps of 2 nm and from 10 nm to 160 nm in steps of 15 nm.

2.4 Analysis

2.4.1 Rietveld refinement

Rietveld refinement was performed on all physical samples collected for $Ba_xSr_{1-x}TiO_3$ and CeO_2 only, through GSASII. Histogram scale factor, background parameters, unit cell parameters and crystallite sizes were refined.

2.4.2 Principal component analysis

OriginLab software was the tool used for data preparation and carrying out PCA with an implemented application named ‘Principal component analysis for spectroscopy’.[30] Due to the inaccuracy of the diffractometer, each step size was not perfect to each 0.01° step size, hence the real patterns were short several hundred points. This caused the real and simulated data not to line-up when grouped. Linear interpolation was performed on the real dataset to 7001 points, identical to that of the simulated dataset. This is true for all materials in this report

3 Results and Discussion

3.1 Cerium Oxide, CeO_2

3.1.1 Simulated CeO_2 XRD Patterns

PCA was performed on 210 simulated CeO_2 XRD patterns with varying unit cell parameter, from 5.4 Å to 5.5 Å and crystallite size from 2.5 nm to 25 nm.

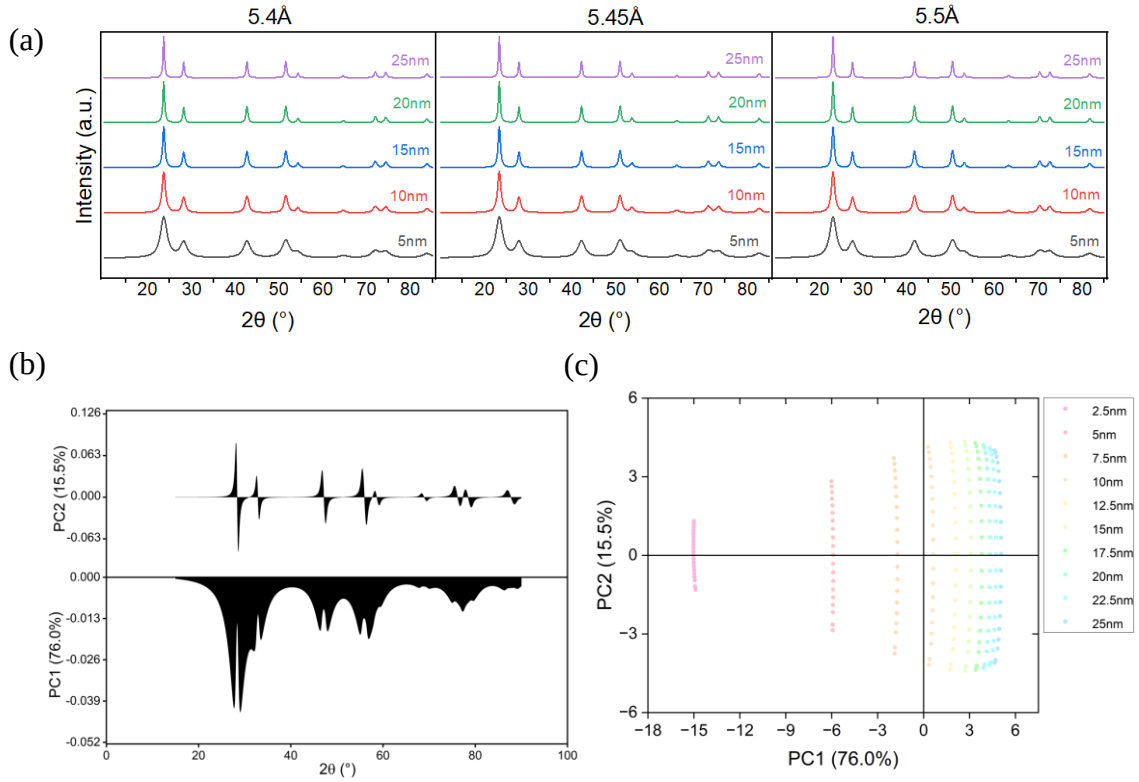


Figure 3.1: (a) Selection of 15 simulated CeO_2 XRD patterns from the 210 produced, 5.4 Å, 5.45 Å, 5.5 Å. Each with five plots for relative crystallite sizes from 5 nm to 25 nm. (b) Loading plot of the full dataset, 210 simulated patterns, representing the first three principal components. (c) A 2D score plot, labelling crystallite sizes, where each simulated pattern is plotted against its PCs.

The two targeted variables, unit cell parameter and crystallite size can be seen in Fig 3.1 (a) above. The broadening of the single peak can be seen for smaller crystals where a shift towards smaller 2θ values is visible for increasing unit cell parameters.

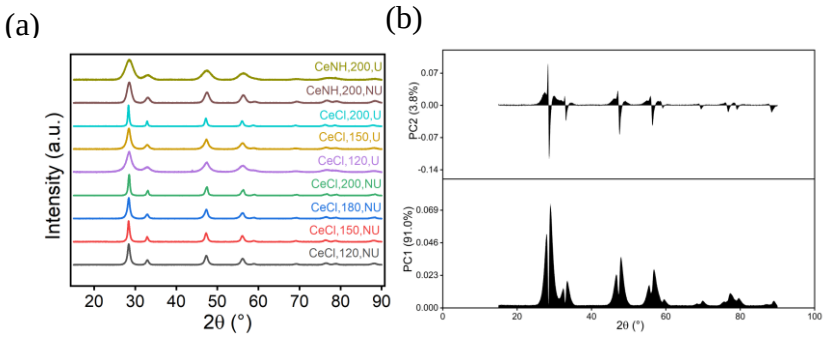
Fig 3.1 (b) displays the loading plot, where PC1 and PC2 represent more than 90% of the data in the system. A loading plot is simply a visual representation of the variables i.e., the CeO_2 patterns, which supplied the most amount of variance in the system, either in the positive or negative direction. PC1 contains broader peaks in the negative space and sharper peaks in the positive. With broadness of peaks being a direct relation to crystallite size, negative PC1

depicts the smaller sizes where the positive space portrays the larger. PC2 illustrates the unit cell parameters, where the negative space shows the larger 2θ values i.e., decreasing unit cell parameter and the positive space displaying the smaller 2θ values i.e., increasing unit cell parameter.

Both claims for PC1 and PC2 can be confirmed when examining the score plot in Fig 3.1 (c). Where positive PC1 depicts larger crystallite sizes and negative PC1 shows smaller sizes, similarly for PC2 where negative PC2 shows smaller unit cell parameters and positive showing larger parameters, see Fig 6.1 in appendix. However, if truly independent, then the plot would present a square figure displaying both properties. This is of course not the case, hence there must be a strong relationship between both variables. The inclusion of simulations at a size of 2.5 nm causes a skew in the dataset in both the loading and score plots. However, small sizes were necessary as CeO_2 produced may reach a size below 5 nm. Exclusion of the 2.5 nm XRD patterns would dramatically decreasing the variance percentage of PC1 as previously shown by Chloé Flandarin.[26] While this may suggest that unit cell parameters and sizes are correlated, it must be borne in mind that these data come from simulated patterns, with each variable independently controlled. Thus, in this case the origin of this deviation is the fact that for extremely small crystallite sizes the diffraction peaks become so broad that peak positions become more difficult to resolve even in this "blind" statistical analysis, thus the values of PC2, which relate predominantly to cell parameters, cluster together for ultrasmall crystallites, giving rise to the apparent correlation seen here in the shape of the PC2 vs PC1 score plot.

3.1.2 Real CeO_2 XRD Patterns

PCA was performed on the diffraction patterns measured from 48 real CeO_2 samples with varying cerium precursor, sodium hydroxide ratio, urea content, time, and temperature as summarised in Table AX.



(c)

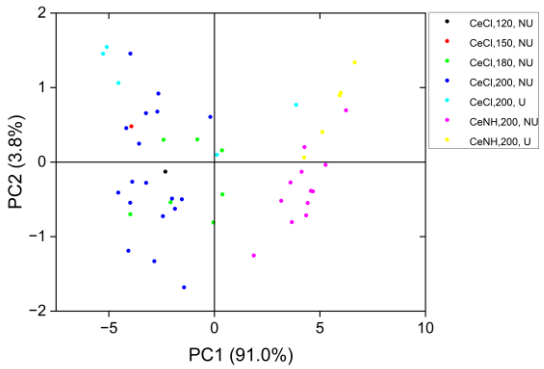


Figure 3.2: (a) Selection of 10 (Ce:OH, 1:40, at 24hr) CeO_2 XRD patterns, coded by Ce precursor (CeCl is $CeCl_3$ *$7H_2O$ and CeNH is $(NH_4)_2Ce(NO_3)_6$), temperatures (°C) and urea. (b) Loading plot of 48 real XRD patterns, representing the first three PCs. (c) A 2D score plot, labelling Ce source, temperature, and urea content, plotted against PCs.

Fig 3.2 (a) shows a selection of XRD patterns, where variances in both unit cell parameter and crystallite sizes are shown. All patterns stemming from the $CeCl_3$ source tend to be larger in size than their $(NH_4)_2Ce(NO_3)_6$ (CeNH) counterpart, as evidenced by the broadness of the $(NH_4)_2Ce(NO_3)_6$ patterns. The addition of urea increases the cerium oxide crystallite size when using $CeCl_3$ as the cerium source, while decreases the size when using $(NH_4)_2Ce(NO_3)_6$. The shifts in the peaks for each pattern only become visually represented when analysing the loading and score plot.

Similarly, to simulation dataset, the loading plot in Fig 3.2 (b) represents crystallite size and unit cell parameter for PC1 and PC2 respectively, however, PC1 holds much more information than PC2, 91% and 3.8%. The shape of the PC2 loading clearly displays some influence from crystallite size, which is expected based on the simulated results. PC1 displays crystallite size with smaller sizes in the positive space and larger sizes in the negative. Again, PC2 represents the unit cell parameters where negative PC2 shows smaller unit cell parameters and positive showing larger unit cell parameters.

These claims can be verified when analysing the score plot in Fig 3.2 (c) above. PC1 contains data on the crystallite size properties where the positive space portrays smaller sizes and the negative space larger. Examining the groups of different cerium precursors, our previous claim upon analysis on the XRD patterns can be validated. The CeNH group can be seen to exist in the positive PC1 space, where you would expect them to be smaller in crystallite size. The same can be said for unit cell parameters when inspecting the CdCl, 200, U group, which is in the positive PC2 space, calling for increasing unit cell parameters. Comparison of

both the score plot and XRD patterns confirm these statements. Once again, only performing PCA on a real dataset proves not as useful as the carrying out PCA for both datasets.

3.1.3 CeO_2 PCA on the simulated and real dataset

PCA was performed on both the 48 real and the 210 simulated CeO_2 XRD patterns. The comparison of both real and simulated datasets aids in describing the properties of the real XRD patterns.

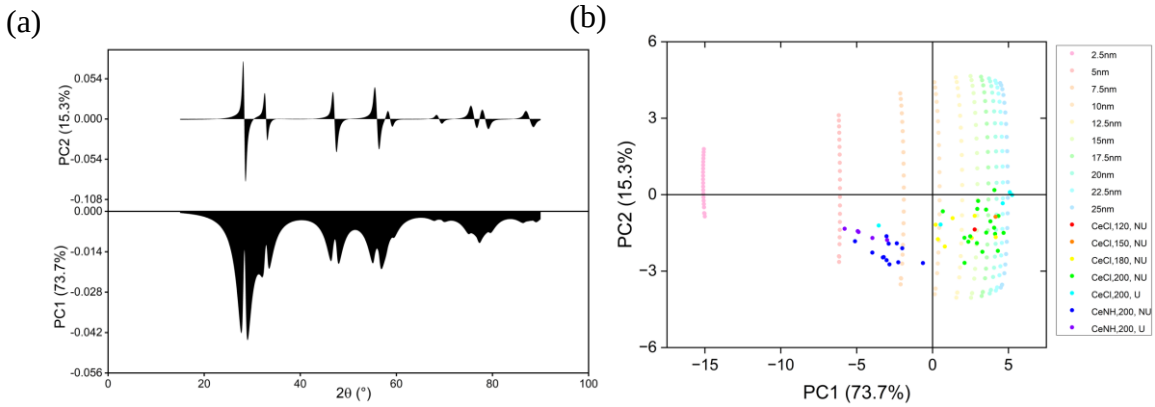


Figure 3.3: (a) Loading plot of the full CeO_2 dataset, both the 48 real and 210 simulated XRD patterns, representing the first three PCs. (b) A 2D score plot where each real pattern is plotted against its PCs.

Expectedly, the simulated dataset has a heavier influence on the combined analysis than the real, hence why the loading plot in Fig 3.3 (a) mimics that of the simulated in Fig 3.1 (b). This is due to PCA picking up greater variance in the simulated patterns as there is simply a wider range of values in both targeted variables. This provides an ideal frame for the real dataset to be examined on.

Analysis of the score plot in Fig 3.3 (b), PC1 is predicted to be related to crystallite size where PC2 is related to unit cell parameter. Upon further examination, a cluster of CeNH source materials can be seen to exist between 5 nm to 10 nm, placing them on the smaller side for crystallite sizes. This supports the analysis carried out in sections 3.2.2 and 3.2.1. From these sections crystals with CeCl precursors were said to be larger in size, which can be seen above. However, with a frame of simulated data the exact sizes for each material can be defined. The same can be said for unit cell parameters and PC2 where a distinct value can be labelled on each real crystal produced without the need for Rietveld refinement, see Fig 6.2 appendix for relevant score plot.

An observation can be noted as some crystals in the real dataset can be seen to leak past the 25 nm range on the simulated map. When setting up a suitable range for the simulated data,

Rietveld refinement was performed on two crystals which were predicted to be the smaller and larger for both parameters, however, one of the CeCl,200,U crystals can be seen to be larger than 25nm which may be due large chlorides particles. Observation of Chloé Flandarin's[26] TEM images, see Fig 6.3 in appendix, for relative CeO_2 can be seen to be consistent with the results found above for crystallite sizes.

(a)

(b)

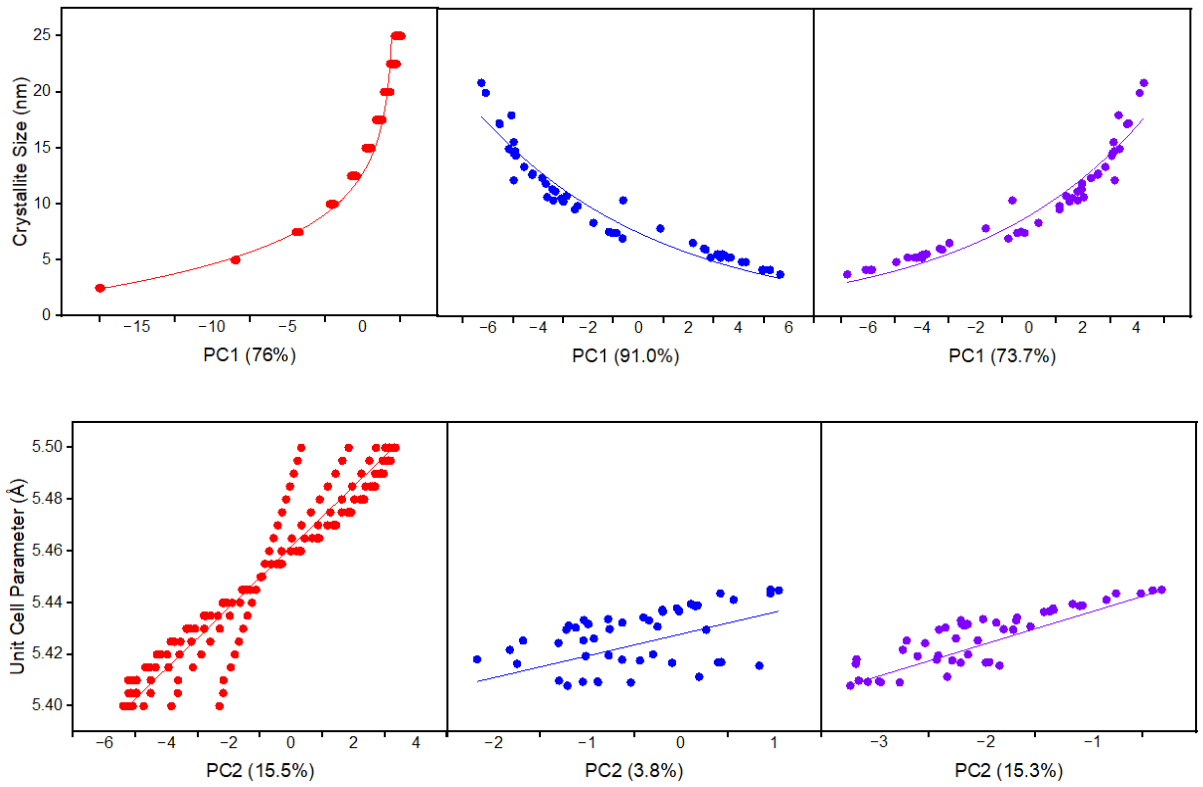


Figure 3.4: (a) Plots of crystallite size vs PC1 for the simulated dataset (red), real dataset (blue) and real PC score values only for the simulated and real dataset (purple), each with an exponential fit, for CeO_2. (b) Plots of unit cell parameter vs PC2 for the simulated dataset (red), real dataset (blue) and real PC score values only for the simulated and real dataset (purple), each with a linear fit, for CeO_2.

Unit cell parameters and crystallite sizes were obtained through Rietveld Refinement, which was conducted by Chloé Flandrin.[26] A method of verifying the usability of PCA is by examining a parameter to their anticipated principal component. Here PC1 is expectedly compared to crystallite size. The simulated plot in Fig 3.4 (a) takes on an exponential form, along with the real and combined datasets. This supports the claim that PC1 is linked to crystallite size, and that the addition of the simulated data to the real has no significant influence of this correlation to the real dataset. Likewise, the same can be called for when examining Fig 3.4 (b) where each plot can be seen to hold a linear relationship between unit cell parameters and PC2. Again, this confirms earlier claims that PC2 is linked to unit cell parameter.

This strongly suggests that in this case the values of principal components obtained from real data analysed in conjunction with simulated datasets can act as a direct proxy for the true physical features of crystallite size and unit cell parameter for cerium oxide. This shows that PCA may indeed serve as a significantly easier and faster method of XRD analysis for screening of reaction conditions for product and/or process optimisation.

3.2 Cadmium Sulfide, CdS

3.2.1 Simulated CdS XRD patterns.

As mentioned previously, 235 simulated patterns were produced with varying hexagonal phase percentage, cubic crystallite size and hexagonal crystallite size. After thorough data transformation, PCA was performed on the prepared data. The figure below shows the XRD patterns, PC plot and respective loading plots for only the simulated CdS dataset.

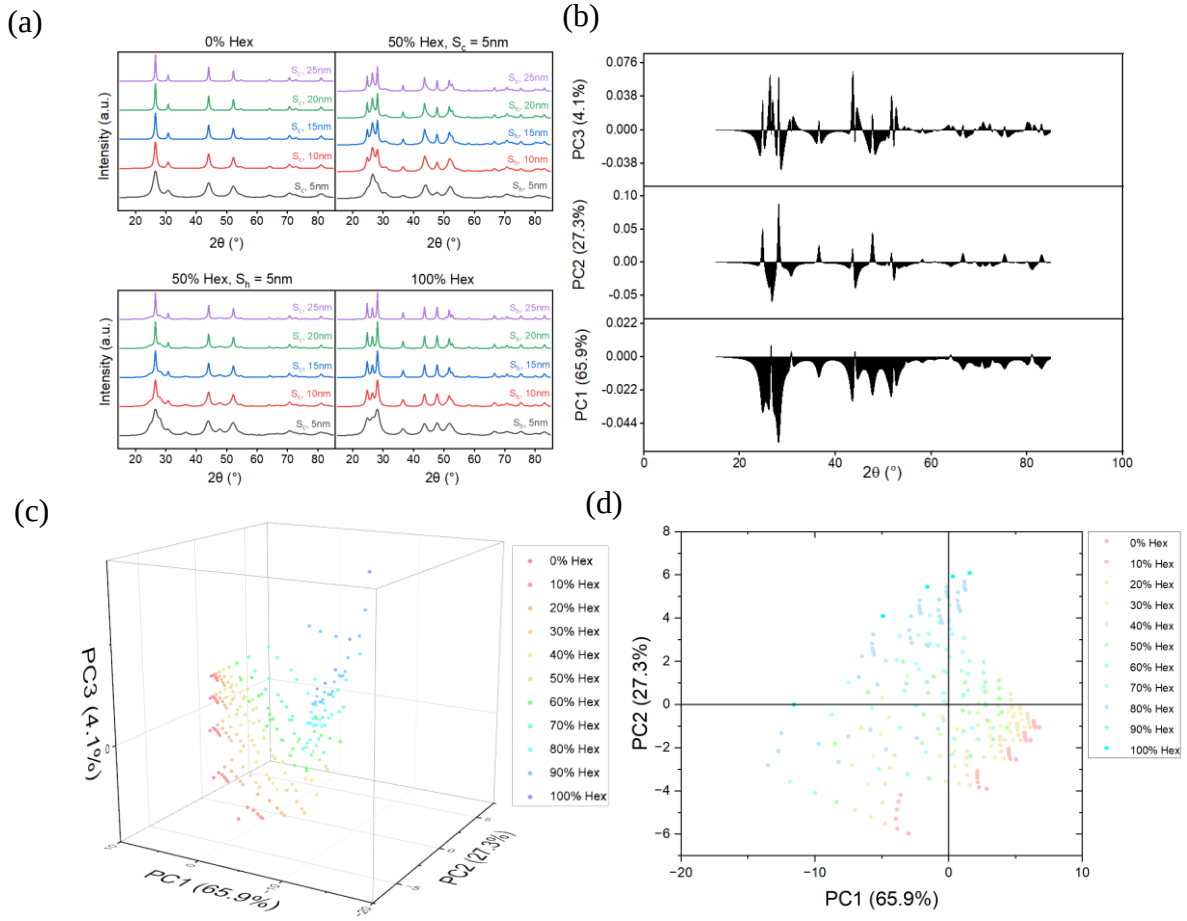


Figure 3.5: (a) Selection of 15 simulated CdS XRD patterns from the 235 produced, all 0% hexagonal, 50% hexagonal and, 100% hexagonal. Each with five plots for relative crystallite sizes from 5 nm to 25 nm. (b) Loading plot of the full CdS dataset, 235 simulated patterns, representing the first three principal components. (c) A 3D score plot, labelling percentage composition, where each simulated pattern is plotted against its PCs. (d) A 2D score plot, labelling percentage composition, where each simulated pattern is plotted against its PCs.

As the cubic and hexagonal phases of CdS are closely related, both being based on close packed sulfide arrays, their diffraction patterns share several features, with the most obvious differences arising in the 20-30 °2θ range. The cubic (zinc blende) CdS XRD pattern involves two peaks, one at (111) at 26.4 ° 2θ and a second at (200) at 30.6 ° 2θ, while the hexagonal (wurtzite) XRD patterns involves a cluster of three peaks from 20° to 30° 2θ, (100), (002), (101) at 24.8 °, 26.4 °, 28.2 ° 2θ, respectively. (Note that the d-spacings of (002) planes of the hexagonal phase and the (111) planes of the cubic phase are the same as these are the close packed layers) As the hexagonal phase percentage increases, the peaks unique to that of the hexagonal structure become more apparent, as can be seen in Fig 3.5 (a) above. The 0% hexagonal patterns expectedly hold a cubic shape, where the 50% hexagonal phase can be seen to hold characteristics of both the wurtzite and zinc blende XRD structures, where the peaks from both groups begin to overlap one another.

These same physical characteristics are carried over to the loading plot represented in Fig 3.5 (b). Here, in the loading of PC1, peak broadening is clearly visible where negative PC1 represents smaller crystallite sizes and positive represents larger sizes. PC2 clearly holds the hexagonal features in the positive direction, where a negative silhouette in the cubic structure is visible, meaning PC2 presents information based on the hexagonal phase percentage of material. Complications begin when examining the loading plot of PC3, as there is no direct correlation of a crystalline property. This is due to a mathematical artifact where the covariance matrix in PCA obtains variances for random variables in the system.[31] The aim here was to inspect whether the crystallite sizes for the cubic and hexagonal phases were independent, however, as can be clearly seen this is not the case as all crystallite size types are captured in PC1. Although difficult to visualise, positive PC1 can be seen to represent the larger cubic and hexagonal crystallite sizes and the negative, the smaller cubic and hexagonal.

The score plot shown in Fig 3.5 (c) and more clearly in Fig 3.5 (d) presents data consistent to that of the loading plots. PC1 has a strong correlation the crystallite size for both the cubic and hexagonal sizes, see Fig 6.4 in appendix. PC2 here can be assigned again to phase composition. The nature of PCA does not mean the data is completely accurate, accuracy is traded with the ease of presenting large datasets. This can be seen in the higher hexagonal composition, where the trend is not consistent and becomes more influenced to PC3. Practicality, it would be more suitable to present this data with only PC1 and PC2 as seen in Fig 3.5 (d).

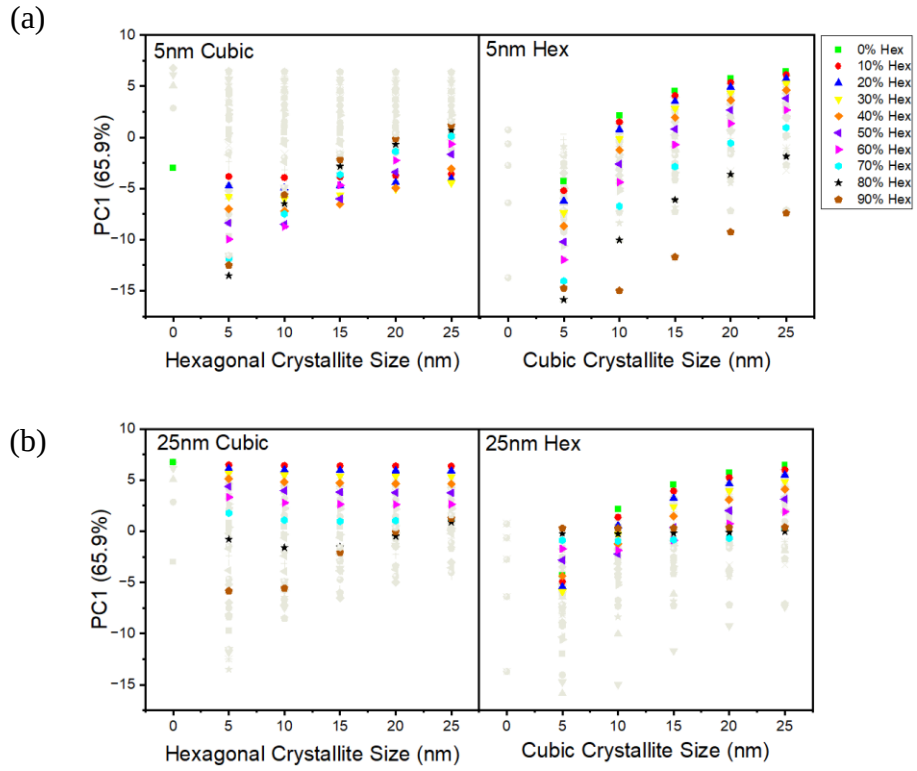


Figure 3.6: (a) Plots of PC1 vs crystallite size for the simulated CdS dataset for varying hexagonal crystallite size and cubic crystallite size at constant 5 nm cubic and 5 nm hexagonal, respectively. (b) Plots of PC1 vs crystallite size for the simulated CdS dataset for varying hexagonal crystallite size and cubic crystallite size at constant 25 nm cubic and 25 nm hexagonal, respectively.

Up to this point to determine which parameter and PC were linked was simply done by examining the loading plots for the simulated dataset and comparing against relevant XRD patterns and score plots. Earlier, PC1 was identified to be correlated to both the size of the crystallite and the form, either hexagonal or cubic. Where negative PC1 was labelled as both smaller crystallite sizes, the opposite exists for positive PC1 being related to larger crystallite sizes. Although both the negative the positive areas of PC1 account for both cubic and hexagonal structures, the negative space seems have a bias relationship towards the hexagonal structure, and the positive space to the cubic shape. Further confirmation is achieved when analysing PC1 score values against crystallite size in Fig 3.6, where smaller in size hexagonal patterns dominate the negative PC1 space over the smaller cubic patterns and larger in size cubic patterns control the positive PC1 space over the larger hexagonal patterns. The same is true when observing similar plots for PC2, see Fig 6.5 in appendix, for the same parameters, where hexagonal phase percentage is heavily linked to PC2. Predictably, there is some relation between PC2 and crystallite size. Analysing the same plot types for PC3 in the appendix (Fig 6.6) can support the previous claim for the interpretation of PC3, it remains a consequence of PCA.

3.2.2 Real CdS XRD patterns

174 CdS products with varying cadmium precursor, sulfur precursor, ratio of Cd/S sources, temperature, time, and synthesis methods, were produced by Andrew Bathe.[27] Samples are denoted by the following labelling scheme: Sample ID: CdXYZ; where X=Cd precursor (N=nitrate, A= acetate), Y=S precursor (Th=thiourea, SO= sodium thiosulfate, S=sodium sulfide) Z= synthesis method (CB=conventional batch, B=batch, I=injection)

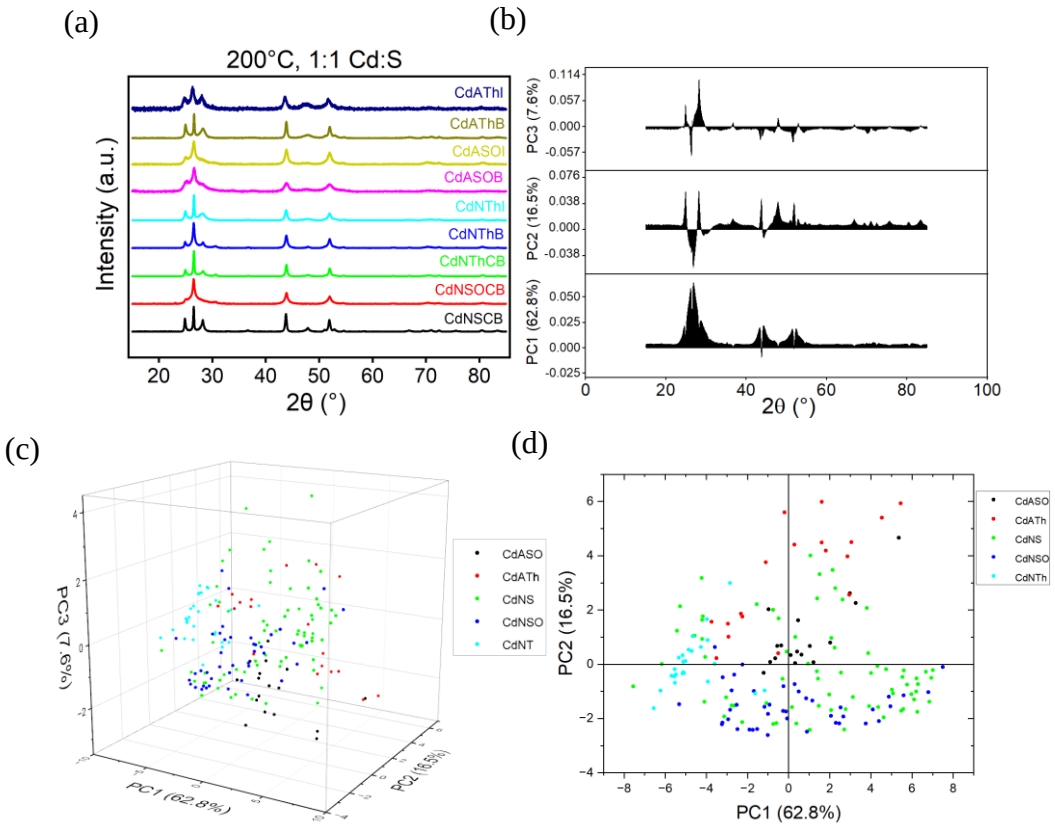


Figure 3.7: (a) Selection of 9 XRD patterns, coded by Cd precursor (CdN is $Cd(NO_3)_2$ and CdA is $Cd(Oac)_2$), S precursor (S is Na_2S, SO is $Na_2S_2O_3$ and, Th is Thiourea) and synthesis method (CB is conventional batch, B is batch, and I is injection). (b) Loading plot of the full dataset, 174 real XRD patterns, representing the first three principal components. (c) A 3D score plot where each real pattern is plotted against its PCs. (d) A 2D score plot where patterns is plotted against its PCs.

The simulated dataset was purposely setup so that PCA would capture wanted crystalline properties, crystallite sizes and hexagonal phase composition, whereas the real was not. Inspection of (a) in Fig 3.7 can be seen to hold the characteristics of both the cubic and hexagonal phases. The CdNSOCB can be seen to take on the most cubic form where CdNSCB seems to be mostly hexagonal. This shows that the sulfur source has a significant effect on the outcome of the structure of CdS produced. The loading plots in Fig 3.7 (b), are heavily related to the one from the simulated dataset. PC1 has an identical broadness figure and sharp

silhouette as seen previously. PC2 holds a hexagonal shape appearance in the positive direction and a cubic form in the negative direction. Again, the artifact is formed in PC3, although visually less complex than in the simulated set, it is still incomprehensible. Simply put, PC1 represents crystallite size and PC2, hexagonal phase composition. This information can be used to depict an estimate of the properties of the materials produced represented in the score plot shown in Fig 3.7 (c). Again, only PC1 and PC2 should be used and represented, as shown in Fig 3.7 (d). It is impractical to only perform and use PCA on the real dataset and it is extremely difficult to identify exact values for the properties of the crystal. For this reason, PCA on the simulated and real dataset are calculated together.

3.2.3 CdS PCA on the Simulated and Real dataset

PCA was performed on both the 174 real and the 235 simulated XRD patterns. As expected, results produced are unchanged from previous calculations. However, the comparison of both real and simulated datasets aids in describing the properties of the real XRD patterns.

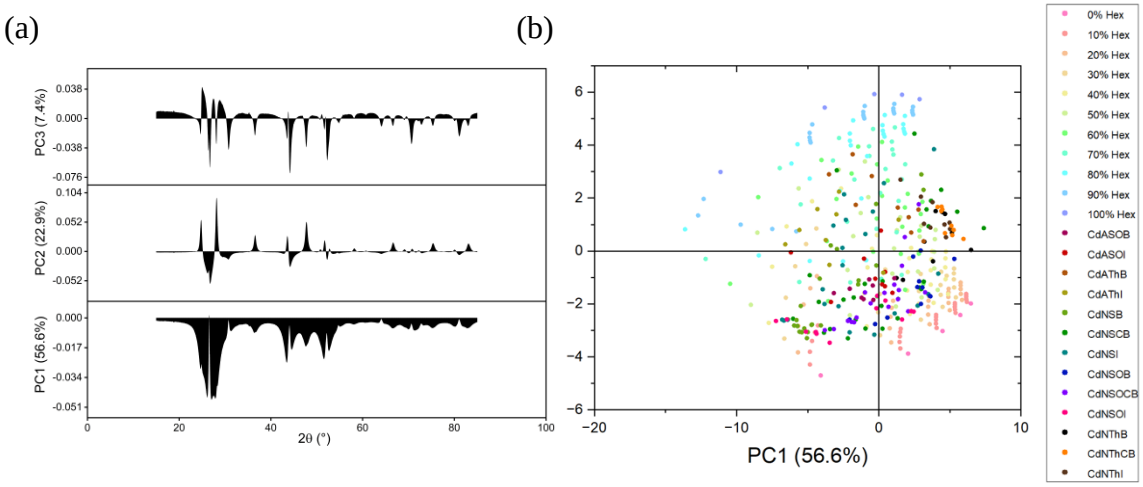


Figure 3.8: (a) Loading plot of the full CdS dataset, both the 174 real and 235 simulated XRD patterns, representing the first three principal components. (b) A 2D score plot where each real pattern to plotted against its PCs.

The loading plot in Fig 3.8 (a) shows near identical results with the ones shown previously in both the simulated and the real case. What is clearly visible is the influence of the real data once calculated together. Where the simulated shows PC1 to contain equal information in both the cubic and hexagonal structures, the real dataset is cubic hence why PC1 in the combined loading plot is more cubic in shape. This can be confirmed when examining the score plot, Fig 3.8 (b). It is possible to misinterpret the shapes of the loading plot peaks as the (002) peak in the hexagonal phase may become overexpressed due to the tendency of hexagonal CdS to grow in rods in (002) direction, which coincides the cubic CdS peak (111)

which may lead to apparent existence of cubic CdS patterns. PC2 again shows the phase composition of the crystal and has remained unmodified when introducing both datasets. PC3 predictably exhibits a structural artifact, as displayed in the 3D score plot in Fig 6.7 in the appendix.

The score plot displays both datasets and fundamentally outputs the properties of the CdS crystals produced in the lab. For example, the CdNThCB (orange) group are in the negative space in PC1 and neutral with respect to PC2. Using the information from the loading plots, the CdNThCB group are small and have a hexagonal phase percentage of 40%-60%. To confirm the sizes, the simulated points in the score plot can be used to label crystallite sizes instead of phase composition, see Fig 6.8 in appendix. One issue at hand, is that the real and simulated datasets do not line up with one another with respect to PC3, although this can be simply ignored as this is an artifact of the mathematics behind covariance matrices. An explanation is that both datasets are different in the fact that the simulated dataset has equal number of patterns with hexagonal phase percentages and crystallite sizes, whereas the real datasets do not produce an equal number of XRD patterns for sizes and compositions, creating differences in the PC3 loading plots for both sets of data. This can be seen in Fig 3.5 (b) and Fig 3.7 (b). Methods such as background correction and linear interpolation of the real dataset initially helped reduce the error found in PC3. Further confirmation is possible if Rietveld Refinements was carried, however, obtaining these calculations is simply not possible due to time constraints and complexity of preforming the refinement on a material with varying cubic and hexagonal phases, as new structural models would need to be constructed for any possibility for refinement

3.3 Barium Strontium Titanate, $Ba_xSr_{1-x}TiO_3$

3.3.1 Simulated $Ba_xSr_{1-x}TiO_3$ XRD Patterns

PCA was performed on 440 unique simulated $Ba_xSr_{1-x}TiO_3$ XRD patterns with varying composition both $BaTiO_3$ and $SrTiO_3$. Unit cell parameters were varied from 3.9 Å to 4.1 Å in steps of 0.02 Å and crystallite size from 2 nm to 20 nm in steps of 2 nm and from 10 nm to 160 nm in steps of 15 nm.

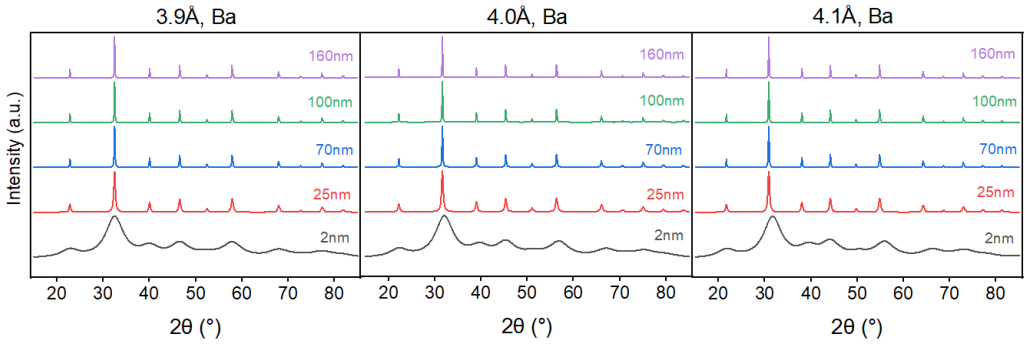


Figure 3.9: Selection of 15 Barium Titanate simulated XRD patterns from the 220 produced, 3.9 Å, 4.0 Å, 4.1 Å. Each with five plots for relative crystallite sizes from 2 nm to 160 nm.

$BaTiO_3$ is shown to have higher relative intensities of the (100) (~23 °2θ) and (210) (~52 °2θ) peaks relative to the (110) (~32 °2θ) peak in comparison to $SrTiO_3$. From observation for both Fig 3.9 and Fig 3.10, this claim is valid. A crystal with a percentage composition of both Barium and Strontium would expectedly exhibit an XRD pattern with a peak intensity being between both pure materials shown.

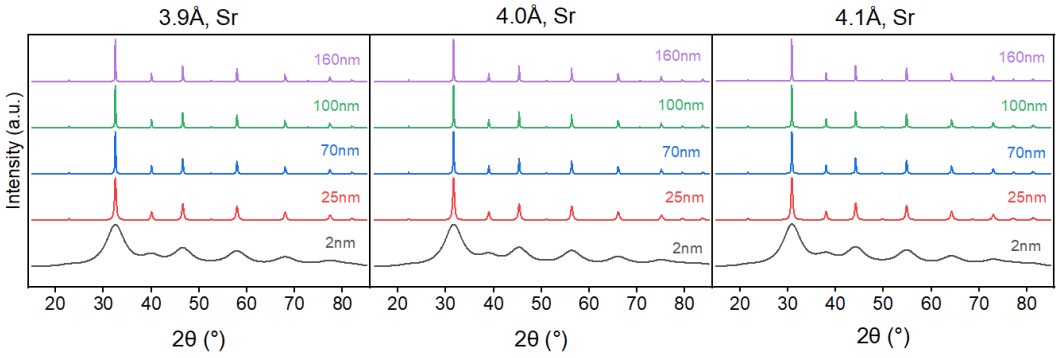


Figure 3.10: Selection of 15 1:1 strontium titanate simulated XRD patterns from the 220 produced, 3.9 Å, 4.0 Å, 4.1 Å. Each with five plots for relative crystallite sizes from 2 nm to 160 nm.

Without surprise, the peak broadness of each pattern changes based on simulated crystallite size. The same can be said for unit cell parameter and peak position, however this becomes more obvious upon analysis of relevant loading and score plots.

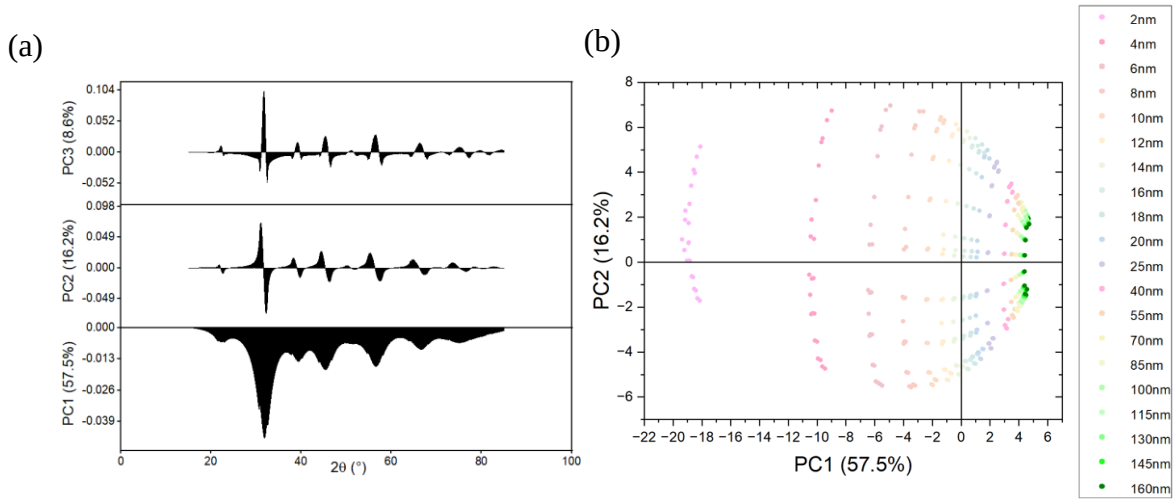


Figure 3.11: (a) Loading plot of the simulated Ba/$SrTiO_3$ dataset, 440 simulated patterns, representing the first three principal components. (b) A 2D score plot, labelling crystallite size, where each simulated pattern is plotted against its PCs.

PC1 in Fig 3.11 (a) can be seen to be related to crystallite size, where negative PC1 displays smaller crystals and positive PC1 shows larger crystals. Again, observation of PC2 in the loading plot implies that positive PC2 is related to larger unit cell parameters, where negative PC2 is correlated to smaller unit cell parameters. The score plot in Fig 3.11 (b), labelled by crystallite sizes, verifies the properties of PC1, where pink dots show smaller crystallite sizes which exist in the negative PC1 space and green dots indicate larger crystallite sizes which can be seen in the positive PC1 space. The same can be said, in relation to PC2, for the score plot labelling unit cell parameters, see Fig 6.9 in appendix.

Ba/$SrTiO_3$ simulations were purposely setup to target three independent variables. However, as can be seen in the figure above, this is not the case. PC1 correctly correlates to crystallite size, PC2 to unit cell parameter, and PC3 is an artifact that is a mixture of all targeted variables, see Fig 6.10 in appendix for 3D score plot. The barium/strontium composition is absent from having its own correlated PC. One may suggest that the relationship may exist in a higher PC, however, analysis of PC4 and PC5 shows no correlation. Labelling the score plot in Fig 3.11 (b) by barium and strontium composition, supports the claim above, see Fig 6.11 in appendix. A reasonable conclusion is that the variance of material composition is spread across all PCs. The artifact in PC3 contains variances for all variables, including composition. Again, this is a consequence of involving functions with a large set of points into PCA.

3.3.2 Real $Ba_xSr_{1-x}TiO_3$ XRD Patterns

PCA was performed on the 24 $Ba_xSr_{1-x}TiO_3$ XRD patterns under three varying conditions, temperature at 150 °C and 200 °C, time at 2, 4, 6 and 24 hours, and material composition of $BaTiO_3$, $Ba_{0.5}Sr_{0.5}TiO_3$ and $SrTiO_3$.

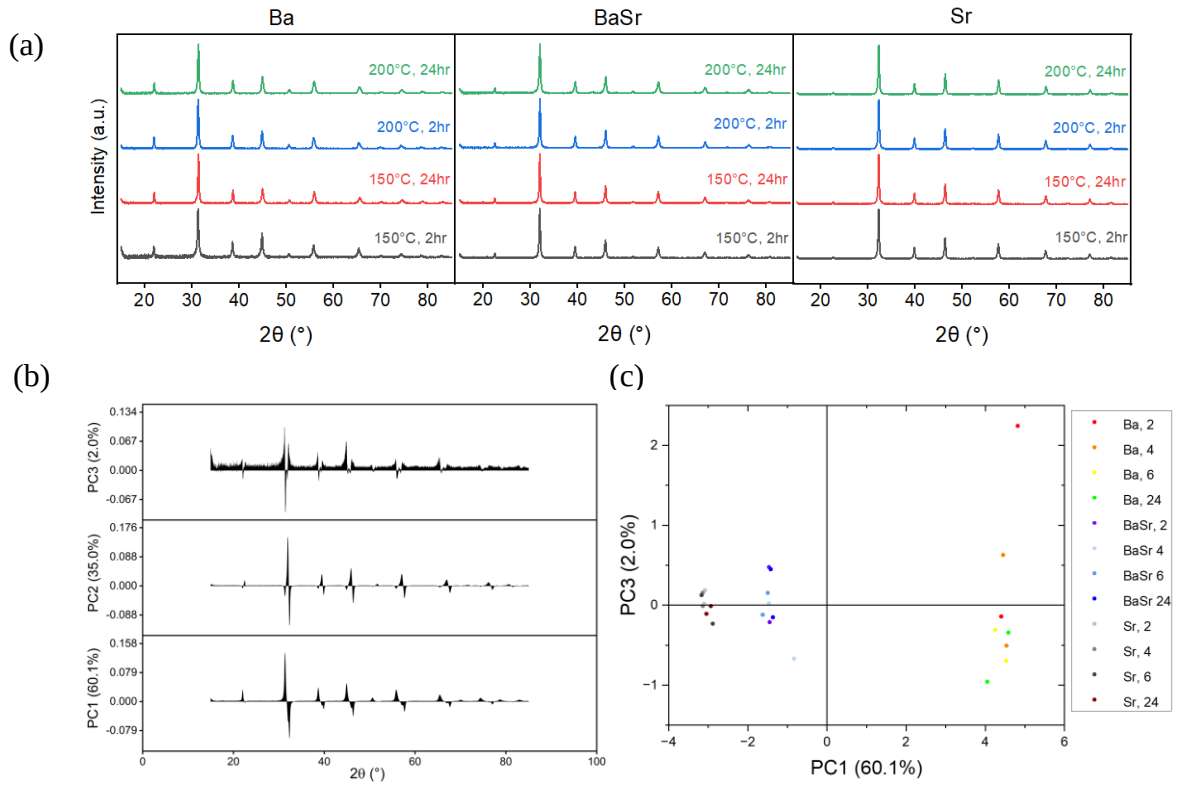


Figure 3.12: (a) Selection of 12 XRD patterns, for each material composition, (Ba is $BaTiO_3$, BaSr is $Ba_{0.5}Sr_{0.5}TiO_3$, Sr is $SrTiO_3$.) and for all temperatures and times 2 and 24 hours. (b) Loading plot of the full dataset, 24 real XRD patterns, representing the first three principal components. (c) A 2D score plot where real patterns is plotted against its PC1 and PC3.

Unlike previous discoveries, the XRD patterns in Fig 3.12 (a) do not have an obvious difference based on their crystallite sizes. This can be translated to the loading plot produced in Fig 3.12 (b), where crystallite size not to correlate to PC1. Crystallite is correlated to PC3 with a low variance of 2%. This means that the crystals produced had a very low variance in crystallite sizes, this can be seen from the Rietveld Refinement results presented below. Increasing the size range can simply be done by performing lower reaction times, under 1 hour. Since the loading PC1 does not take form of crystallite size, it expectedly takes on the form of unit cell parameters, which can be verified through further analysis below. The loading PC2 takes on the same artificial form of PC3 in the simulated dataset.

Table 3.1: Rietveld Refinement results for crystallite sizes and unit cell parameters of $Ba_xSr_{1-x}TiO_3$

		$BaTiO_3$		$Ba_{0.5}Sr_{0.5}TiO_3$		$SrTiO_3$	
Temperature (°C)	**Time (hours)**	**Size (nm)**	**Unit Cell (Å)**	**Size (nm)**	**Unit Cell (Å)**	**Size (nm)**	**Unit Cell (Å)**
150	2	36.5	4.035	29.6	3.951	37.1	3.918
	4	36.4	4.032	34.5	3.964	39.4	3.918
	6	36.7	4.031	31.7	3.952	53.1	3.915
	24	36.3	4.029	38.9	3.954	47.7	3.916
200	2	24.7	4.037	37.0	3.951	42.5	3.916
	4	51.0	4.033	34.3	3.954	43.8	3.916
	6	40.3	4.032	36.3	3.951	41.6	3.915
	24	38.4	4.027	31.8	3.953	45.5	3.914

Performing Rietveld Refinement on the sample data provides a further insight to meaning behind each PC in the loading and score plot in Fig 3.12 above. The crystallite sizes obtained have a range from 24.7 nm to 53.1 nm which is a small enough range for PCA not to pick it up as one of its primary variables. The unit cell parameters are consistent to their composition. Observation of the relevant score plots shows that there is a difference in position in PC1 based on composition and therefore unit cell parameter. This may explain the lack of a PC for solely material composition, as it is heavily related to unit cell parameter.

PC3 was previously stated to correlate to crystallite size, although low in variance. Looking at the score plot above shows a spreading of each group with respect to size ranges. The $BaTiO_3$, samples by Rietveld refinement show a size range of 26.3 nm varying from 24.7 nm to 51 nm, $Ba_{0.5}Sr_{0.5}TiO_3$ a size range of 9.3 nm varying from 29.6 nm to 38.9 nm, and $SrTiO_3$ a size range of 16 nm varying from 37.1 nm to 53.1 nm. These size ranges are clearly reflected in the spread of the data points with respect to PC3. which directly correlate to their observed spread in the score plot in Fig 3.12 (c). The same can be said for corresponding unit cell parameter ranges, meaning that the two variables have a relationship. However, this is expected from PCA as accuracy is traded for generalisation. Again, observing the data here with a simulated map can output more accurate and interpretable results.

3.3.3 $Ba_xSr_{1-x}TiO_3$ PCA on the Simulated and Real dataset

PCA was performed on both the 24 real and the 440 simulated XRD patterns. The comparison of both real and simulated datasets aids in describing the properties of the real XRD patterns.

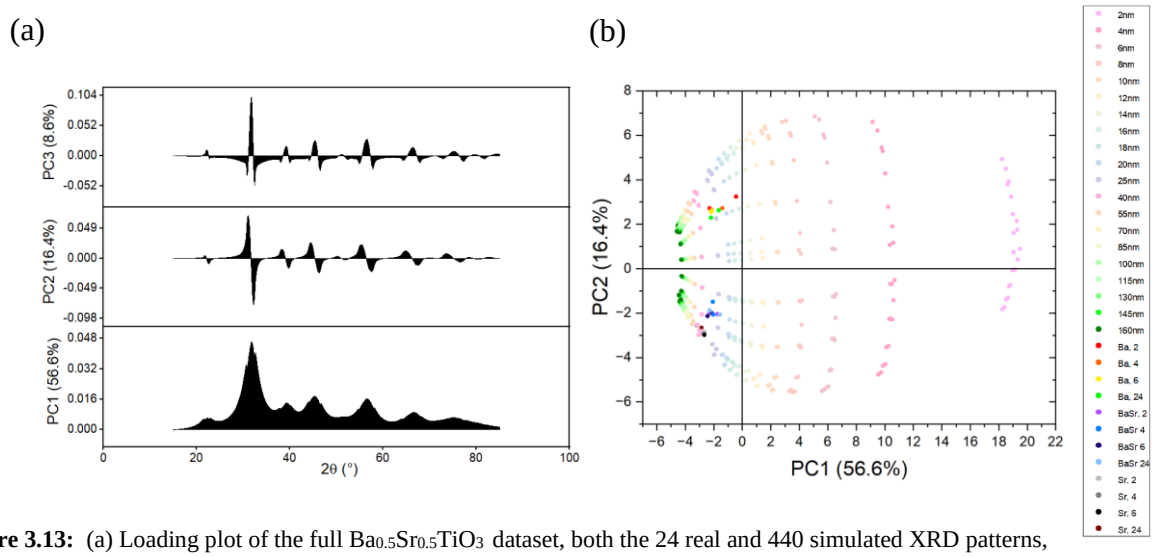


Figure 3.13: (a) Loading plot of the full $Ba_{0.5}Sr_{0.5}TiO_3$ dataset, both the 24 real and 440 simulated XRD patterns, representing the first three principal components. (b) A 2D score plot where each real pattern to plotted against its PCs (Ba is $BaTiO_3$, BaSr is $Ba_{0.5}Sr_{0.5}TiO_3$, Sr is $SrTiO_3$ and 2,4,6,24 is time in hours.).

With the least real samples and the most simulated patterns produced, the loading plot in Fig 3.13 (a) predictably reflects that of the simulated as seen in Fig 3.11 (a). The loading of PC1 relates to crystallite size, where positive PC1 displays smaller crystals and negative PC1 shows larger crystals. Once more, observation of PC2 in the loading plot indicates that positive PC2 is related to larger unit cell parameters, and negative PC2 is correlated to smaller unit cell parameters. PC3, again, displays a blend of variables, i.e., an artifact in the analysis method.

The true practicality of using PCA for XRD datasets comes from examining the score plot in Fig 3.13 (b). The crystallite sizes for each material can be approximated when a simulated frame is implemented below the real data. The red subgroup, Ba 2, can be seen to be placed just above and below the lilac 25 nm indicator from the simulated map, which can be confirmed by the values found from the Rietveld Refinement of 24.7 nm (200 °C) and 36.5 nm (150 °C). Analysing SEM images for each composition further confirms the sizes of these crystals, see Fig 6.13 in appendix. Likewise, labelling the score plot with respect to unit cell allows the unit cell parameter for each real pattern to be determined , see Fig 6.12 in appendix.

The simulated data was setup with an odd crystallite size range from 2 nm to 160 nm, although real data only contains sizes from 24.7 nm to 54.1 nm. This was executed to force PC1 to represent that of crystallite size, not doing so would produce a loading plot like the one seen in Fig 3.12 (b).

(a)

(b)

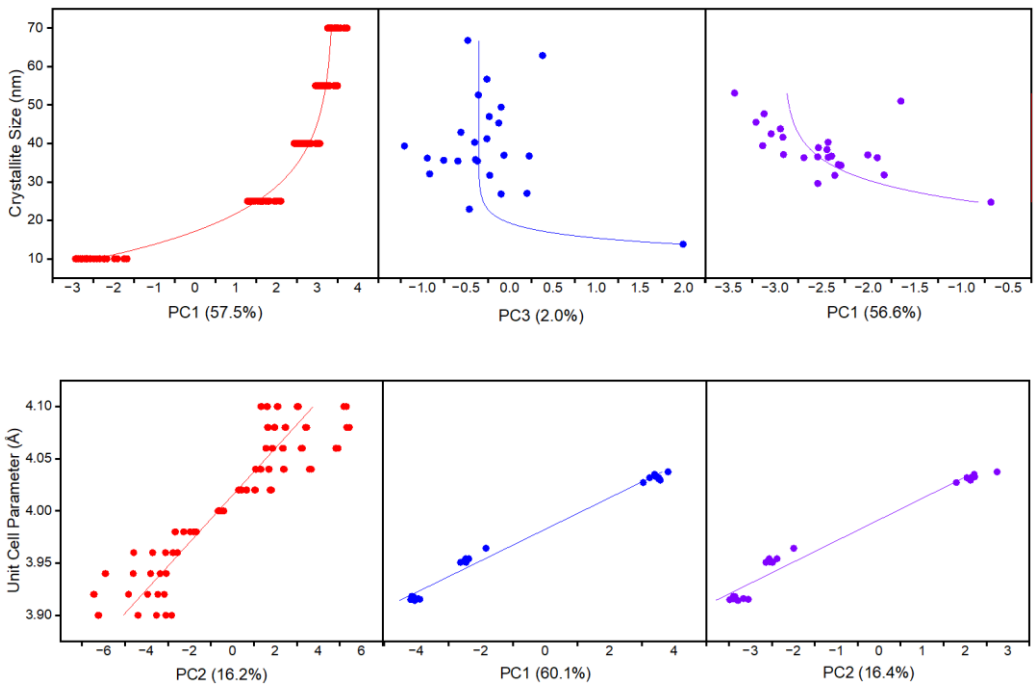


Figure 3.14: (a) Plots of crystallite size vs PC1 for the simulated dataset (red), real dataset (blue), and real PC score values only for the simulated and real dataset (purple), each with an exponential fit, for $Ba_xSr_{1-x}TiO_3$. (b) Plots of unit cell parameter vs PC2 for the simulated dataset (red), real dataset (blue) and real PC score values only for the simulated and real dataset (purple), each with a linear fit, for $Ba_xSr_{1-x}TiO_3$.

A method of verifying the viability of PCA is by examining a crystalline property to their relevant principal component. Here, PC1 for the simulated dataset (red, Fig 3.14 (a)) is linked to crystallite size, as can be seen by the exponential relationship and PC2 (red, Fig 3.14 (b)) is related to unit cell parameter as can be seen by the somewhat linear trend. The real PCA score for the combined datasets (purple) can be seen to display the same relationship to each of their PCs as their corresponding simulated plot. This demonstrates that there is a correlation between PC2 and unit cell parameter, along with PC1 and crystallite sizes.

However, the PCA score values for the real dataset only do not follow the trend. In the case of cerium oxide, crystallite size was not difficult for PCA to distinguish. This is not the case for $Ba_xSr_{1-x}TiO_3$ as can be seen from the results obtained for the PCA on the real dataset, displayed in section 3.3.2, that PCA finds difficulty in distinguishing crystallite size, as it is mainly correlated to PC3 with a low variance of 2%, see Fig 3.14 (a) above. This can be forced by introducing a wider range of sizes in the simulated datasets. However, there is a clear relationship between PC2 and unit cell parameters meaning PCA behaves as expected here. Although PCA performs without complications for cerium oxide, there are limitations and complications to this statistical technique. Although proven to be useful and efficient, PCA should not be solely used to acquire physical (nano)crystalline properties.

4 Conclusions

The purpose of this report is to examine the viability of using Principal Component Analysis, PCA, as a statistical technique to determine various crystalline properties of large material datasets. Due to an increased demand for materials discovery and product/process optimisation, the pursuit for developing methods of rapid data treatment is growing. Conventional methods for determining these crystalline properties, such as Rietveld Refinement, simply require a great of time, expertise, and computational power.

Three materials are used in this project to explore the practically of PCA, cerium oxide (CeO_2), cadmium sulfide (CdS), and barium strontium titanate ($Ba_xSr_{1-x}TiO_3$). Each material exhibits various crystalline characteristics, which have been purposely chosen to test PCA when it comes to a wide range of crystal properties. Simulated XRD patterns for each material were produced with varying crystalline attributes and PCA is performed to create a template to describe the real data. Carrying out combined PCA for both real and simulated datasets allow features such as crystallite size and unit cell parameter of the real dataset to be differentiated and estimated against the simulated map.

The first material, CeO_2, can be described by two fundamental properties, crystallite size and unit cell parameter. A simulated set of 210 XRD pattern have been produced while varying these properties. PCA has been performed on these patterns which display that each PC is linked to a specific crystalline property, PC1 is related to crystallite size and PC2 to unit cell parameter. The same can be said when PCA is carried out on the real and simulated datasets. The validity of PCA can be is confirmed when compared to obtained Rietveld Refinement results, such that there clear relationships between PCs and refined cell parameters and crystallite sizes. Each PC does not fully reflect that of each physical properties, it is simply a correlation, PCA trades accuracy with efficient generalisation.

CdS exhibits two crystal structures, cubic and hexagonal, where both can exist at room temperature. A simulated dataset with three varying properties, hexagonal phase percentage, cubic crystallite size and hexagonal crystallite size was generated to describe the real dataset. However, due to the dependence of both crystallite size structures on one another, the materials properties could only be explained through the first two PCs. The same can be said for the real and simulated dataset. PC3 has been found to depict an artifact, which is a mathematical consequence of PCA. Unit cell parameters may become a suitable variable to introduce into

the system to label each PC to distinct properties. Though, the process of varying unit cell parameters becomes complex when considering the hexagonal structures of CdS.

The third material, $Ba_xSr_{1-x}TiO_3$, had been produced under varying time, temperature, and material composition. Simulated patterns were created with three varying properties, crystallite size, unit cell parameter and composition. Similarly, to CdS, the three properties do not individually correlate to a corresponding PC. Analysis of the simulated dataset shows that crystallite size is linked to PC1 and unit cell parameter to PC2. PC3 exhibits an artifact, a mixture of all targeted variables. PCA for the real dataset shows that crystallite size is not expectedly linked to PC1, but to PC3 with a variance of 2%, which is due to the small size range of the material. This also why an odd range of crystallite sizes had been used for producing the simulated map as without so, PC1 would not be correlated to crystallite size. PCA on the combination of both datasets was proven to be a useful method in determining the unit cell parameters and crystallite sizes for the real materials produced. However, due to complications with crystallite sizes ranges and lack of a distinct PC for material composition, modifications to PCA or additional techniques need to be used.

Overall, PCA has proven to be a suitable method to rapidly determine various physical properties of crystalline materials from large XRD pattern datasets. However, there are limitations of using this multivariable statistical method as a characterisation extraction tool as shown in the case of more complex materials such as $Ba_xSr_{1-x}TiO_3$ and CdS. Although proven to be extremely useful and efficient, PCA should not be solely used to acquire physical (nano)crystalline properties and should be used a tool alongside other techniques to increasing the throughput of materials science discovery.

5 References

1. D. S. Kukich, Patel, C. Kumar N. , McCullough, R.L. , Venables, John D. , Girifalco, Louis A. and Marchant, Roger Eric., 2023.
2. J. George, *Trends in Chemistry*, 2021, **3**, 697-699.
3. D. Caramelli, J. M. Granda, S. H. M. Mehr, D. Cambié, A. B. Henson and L. Cronin, *ACS Central Science*, 2021, **7**, 1821-1830.
4. J. A. Darr, J. Zhang, N. M. Makwana and X. Weng, *Chemical Reviews*, 2017, **117**, 11125-11238.
5. S. E. Dann, *Reactions and Characterization of SOLIDS*, Royal Society of Chemistry, USA, 2002.
6. Britannica, T. Editors of Encyclopaedia (2023, September 22). Bragg law. Encyclopedia Britannica. https://www.britannica.com/science/Bragg-law
7. R. Nix, Miller Indices (hkl), https://chem.libretexts.org/Bookshelves/Physical_and_Theoretical_Chemistry_Textbook_Maps/Surface_Science_(Nix)/01%3A_Structure_of_Solid_Surfaces/1.02%3A_Miller_Indices_(hkl)).
8. M. R. F. Ahmad Monshi, Mohammad Reza Monshi, 2012, DOI: 10.4236/wjnse.2012.23020.
9. A. Hewat, W. I. F. David and L. van Eijck, *Journal of Applied Crystallography*, 2016, **49**, 1394-1395.
10. Turing.com, A Step-By-Step Complete Guide to Principal Component Analysis, https://www.turing.com/kb/guide-to-principal-component-analysis).
11. S. M. Holland, 2008.
12. L. I. Smith, 2002.
13. K. Chang, H. Zhang, M.-j. Cheng and Q. Lu, *ACS Catalysis*, 2020, **10**, 613-631.
14. K. Reinhardt and H. Winkler, in *Ullmann's Encyclopedia of Industrial Chemistry*, DOI: https://doi.org/10.1002/14356007.a06_139.
15. D. Prieur, W. Bonani, K. Popa, O. Walter, K. W. Kriegsman, M. H. Engelhard, X. Guo, R. Eloirdi, T. Gouder, A. Beck, T. Vitova, A. C. Scheinost, K. Kvashnina and P. Martin, *Inorganic Chemistry*, 2020, **59**, 5760-5767.
16. C. Artini, G. A. Costa, M. Pani, A. Lausi and J. Plaisier, *Journal of Solid State Chemistry*, 2012, **190**, 24-28.
17. K. Momma and F. Izumi, *Journal of Applied Crystallography*, 2011, **44**, 1272-1276.
18. J. J. Calvin, A. Ben-Moshe, E. B. Curling, A. S. Brewer, A. B. Sedlak, T. M. Kaufman and A. P. Alivisatos, *The Journal of Physical Chemistry C*, 2022, **126**, 12958-12971.
19. Q. Li, X. Li and J. Yu, in *Interface Science and Technology*, eds. J. Yu, M. Jaroniec and C. Jiang, Elsevier, 2020, vol. 31, pp. 313-348.
20. F. Ulrich and W. Zachariasen, *Zeitschrift für Kristallographie - Crystalline Materials*, 1925, **62**, 260-273.
21. S. Biswas, S. Das, S. Bhattacharya and A. K. Singh, in *Reference Module in Materials Science and Materials Engineering*, Elsevier, 2019, DOI: https://doi.org/10.1016/B978-0-12-803581-8.11581-4.
22. M. Karimi-Jafari, K. Kowal, E. Ul-Haq and S. A. M. Tofail, in *Comprehensive Materials Finishing*, ed. M. S. J. Hashmi, Elsevier, Oxford, 2017, DOI: https://doi.org/10.1016/B978-0-12-803581-8.09203-1, pp. 347-357.
23. A. Yoko, M. Akizuki, N. Umezawa, T. Ohno and Y. Oshima, *RSC Advances*, 2016, **6**, 67525-67533.
24. E. K. Al-Shakarchi and N. B. Mahmood, *Journal of Modern Physics*, 2011, **2**, 1420-1428.
25. R. Nagpal and M. Gusain, in *Graphene, Nanotubes and Quantum Dots-Based Nanotechnology*, ed. Y. Al-Douri, Woodhead Publishing, 2022, DOI: https://doi.org/10.1016/B978-0-323-85457-3.00006-2, pp. 599-630.
26. C. Flandrin, 2023. The Use of Statistical Methods in Characterisation and Analysis of Inorganic Nanomaterials.
27. A. Bathe, 2021. The Phase- and Shape-Controlled Synthesis of Metal Sulfide Nano- and Micromaterials
28. B. H. Toby and R. B. Von Dreele, *Journal of Applied Crystallography*, 2013, **46**, 544-549.
29. A. Vaitkus, A. Merkys, T. Sander, M. Quirós, P. A. Thiessen, E. E. Bolton and S. Gražulis, *Journal of Cheminformatics*, 2023, **15**, 123.
30. Origin. Technical. Support, *Principal Component Analysis*, 2016.
31. M. Björklund, *Evolution*, 2019, **73**, 2151-2158.

6 Appendix

Table 6.1: Summary of experimental conditions for CeO_2, by Chloé Flandarin[26]

Ce source	[OH]/[Ce]	[Urea]/[Ce]	Temperature (°C)	Time (hours)
$CeCl_3*7H_2O$	**4-40**	0	200	24
$CeCl_3*7H_2O$	**4-40**	0	180	24
$CeCl_3*7H_2O$	40	**0-10**	200	24
$CeCl_3*7H_2O$	40	0	**120, 150, 180, 200**	24
$CeCl_3*7H_2O$	40	0	200	**2-24**
$(NH_4)_2Ce(NO_3)_6$	40	**0-10**	200	24
$(NH_4)_2Ce(NO_3)_6$	40	0	200	**2-24**

Table 6.2: A summary of reaction conditions for CdS samples prepared *via* the conventional hydrothermal batch method. Totalling 9 samples, by Andrew Bathe [27]

Cd^{2+} source	**S^{2-} source**	**Cd^{2+}:S^{2-}**	**Temperature (°C)**	**Heating time, t_h (hr)**
Cadmium Nitrate	Thiourea	1:1	200	2 3 4
		1:2	200	2 3 4
		1:6	200	2 3 4

Table 6.3: Summary of the reaction conditions for CdS samples prepared with cadmium nitrate and thiourea *via* the reactor batch and injection methods. Totalling 18 samples, by Andrew Bathe[27]

Cd^{2+} source	**S^{2-} source**	**Method**	**Cd_{2+}:S_{2-}**	**Temperature (ºC)**
Cadmium Nitrate	Thiourea	Reactor batch	1:1	200 250 300
			1:2	200 250 300
			1:6	200 250 300
		Reactor injection	1:1	200 250 300
			1:2	200 250 300
			1:6	200 250 300

Table 6.4: summary of the reaction conditions for CdS samples prepared with cadmium acetate and thiourea *via* the reactor hydrothermal batch and injection methods, by Andrew Bathe[27]

Cd^{2+} source	**S^{2-} source**	**Method**	**Cd_{2+}:S_{2-}**	**Temperature (ºC)**
Cadmium acetate	Thiourea	Reactor batch	1:1	150 200 250
			1:2	150 200 250
			1:6	150 200 250
		Reactor injection	1:1	150 200 250
			1:2	150 200 250
			1:6	150 200 250

Table 6.5: A summary of reaction conditions for CdS samples prepared *via* the hydrothermal reactor batch and injection methods over a series of acidic and basic conditions, by Andrew Bathe[27]

Cd^{2+} source	S^{2-} source	Method	Cd_{2+}:S_{2-}	Temperature (ºC)	Acid/Base solution
Cadmium Nitrate	Thiourea	Reactor batch	1:1	300	0.05 M HNO_3
				300	0.1 M HNO_3
				300	0.15 M HNO_3
		Reactor injection		300	0.05 M HNO_3
				300	0.1 M HNO_3
				300	0.15 M HNO_3

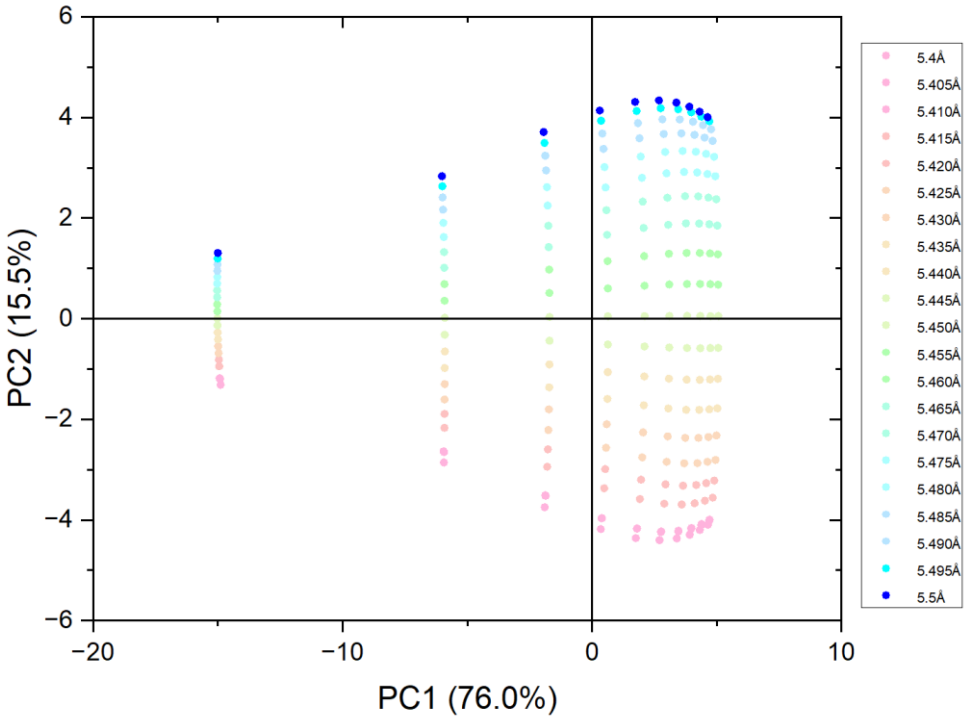


Figure 6.1: A CeO_2 2D simulated dataset score plot, labelling unit cell parameters, where each simulated pattern to plotted against its PCs

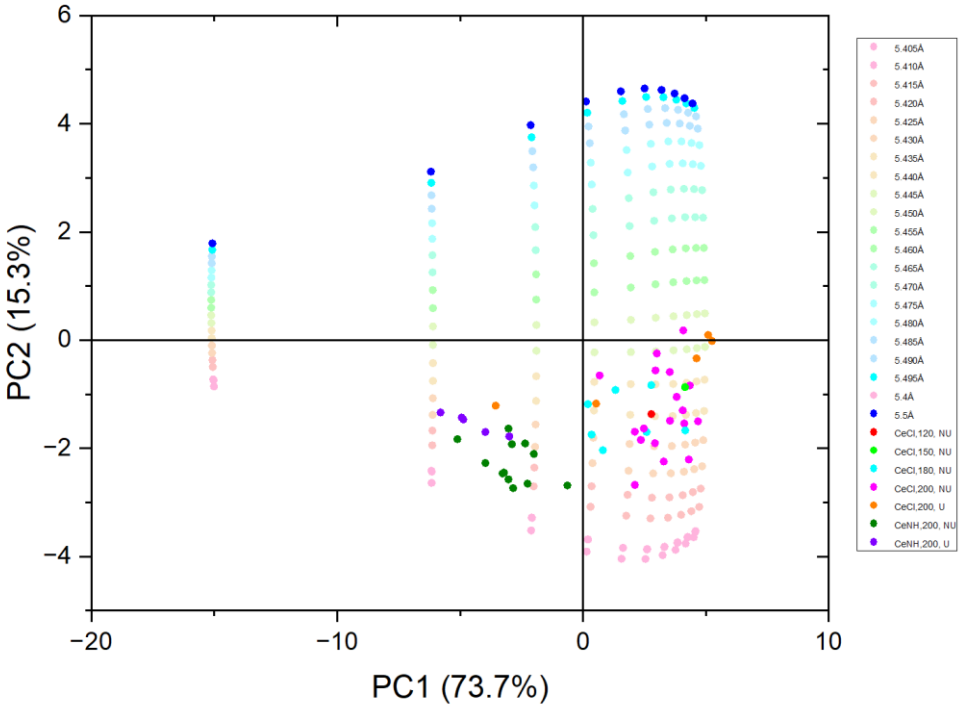


Figure 6.2: A CeO_2 2D dataset score plot, labelling unit cell parameters, where each simulated and real pattern is plotted against its PCs

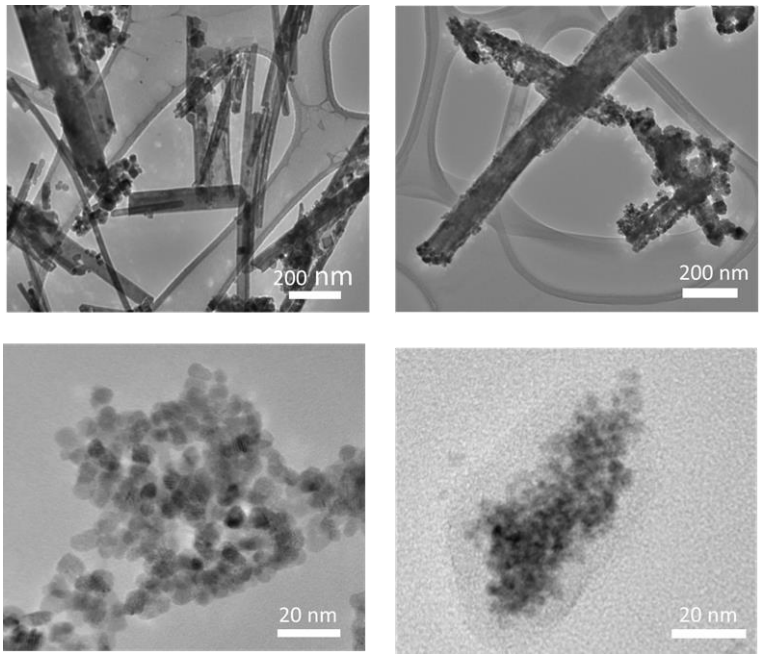


Figure 6.3: TEM pictures of CeO_2 samples synthetised at a temperature of 200 °C, Ce:OH ratio of 1:40 during 24 hours (top left) using CeCl3.7H2O; (top right) with Urea; (bottom left) with (NH4)2Ce(NO3)6 ; (bottom right) with Urea, all taken by Chloé Flandarin

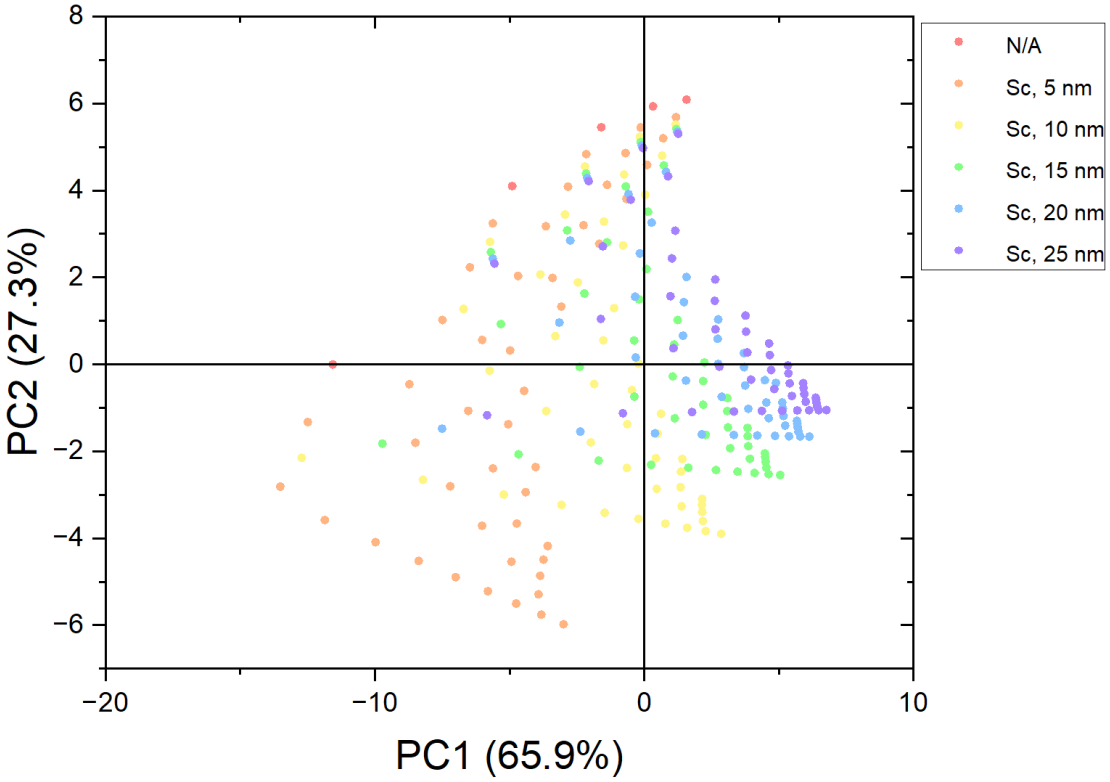


Figure 6.4: A CdS 2D simulated dataset score plot, labelling cubic crystallite size, where each simulated pattern is plotted against its PCs

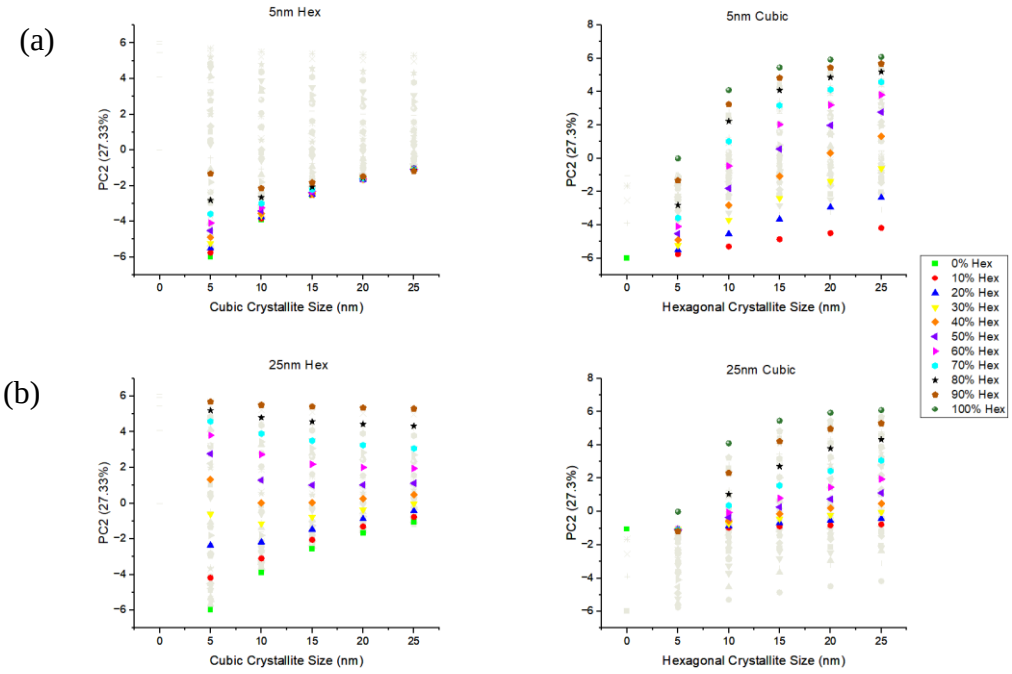


Figure 6.5: (a) Plots of PC2 vs crystallite size for the simulated dataset for varying hexagonal crystallite size and cubic crystallite size at constant 5 nm cubic and 5 nm hexagonal, respectively. (b) Plots of PC2 vs crystallite size for the simulated dataset for varying hexagonal crystallite size and cubic crystallite size at constant 25 nm cubic and 25 nm hexagonal, respectively.

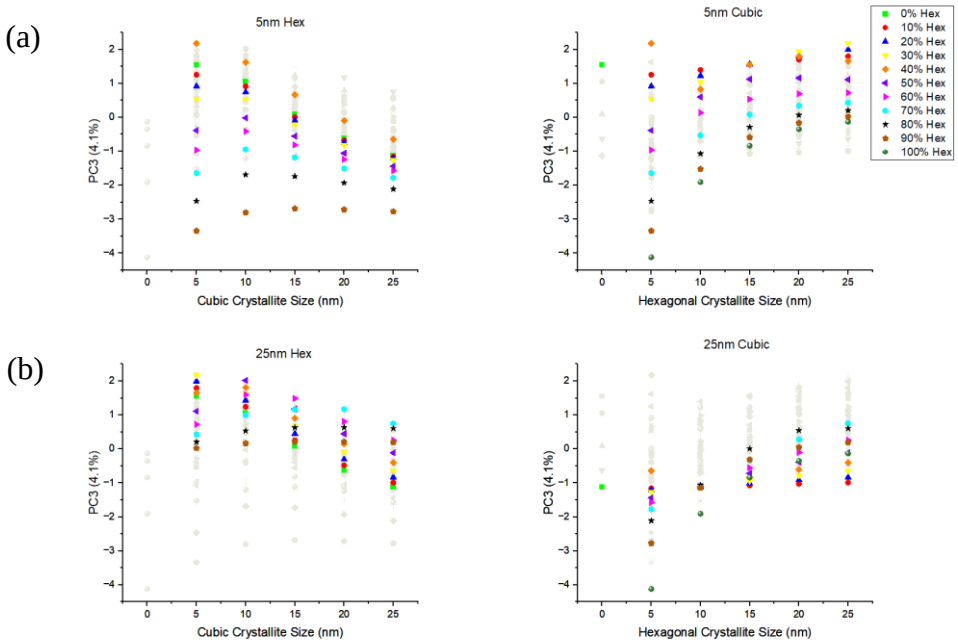


Figure 6.6: (a) Plots of PC3 vs crystallite size for the simulated CdS dataset for varying hexagonal crystallite size and cubic crystallite size at constant 5 nm cubic and 5 nm hexagonal, respectively. (b) Plots of PC3 vs crystallite size for the simulated CdS dataset for varying hexagonal crystallite size and cubic crystallite size at constant 25 nm cubic and 25 nm hexagonal, respectively.

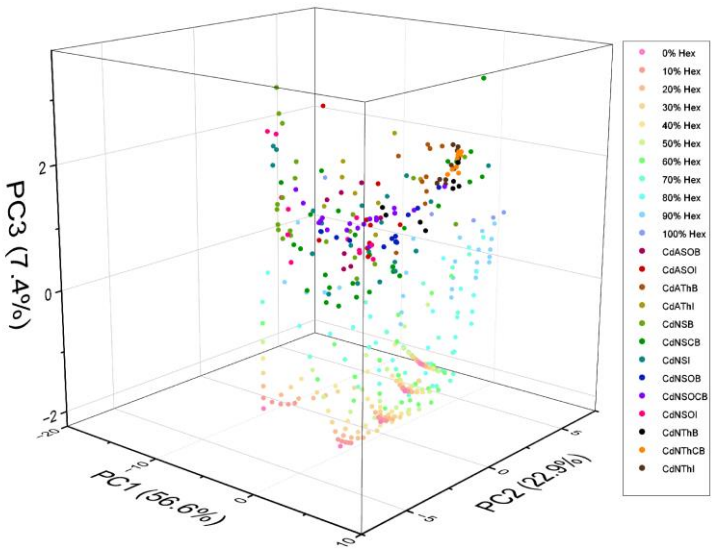


Figure 6.7: A CdS 3D simulated and real dataset score plot, labelling cubic crystallite size, where each simulated and real pattern is plotted against its PCs

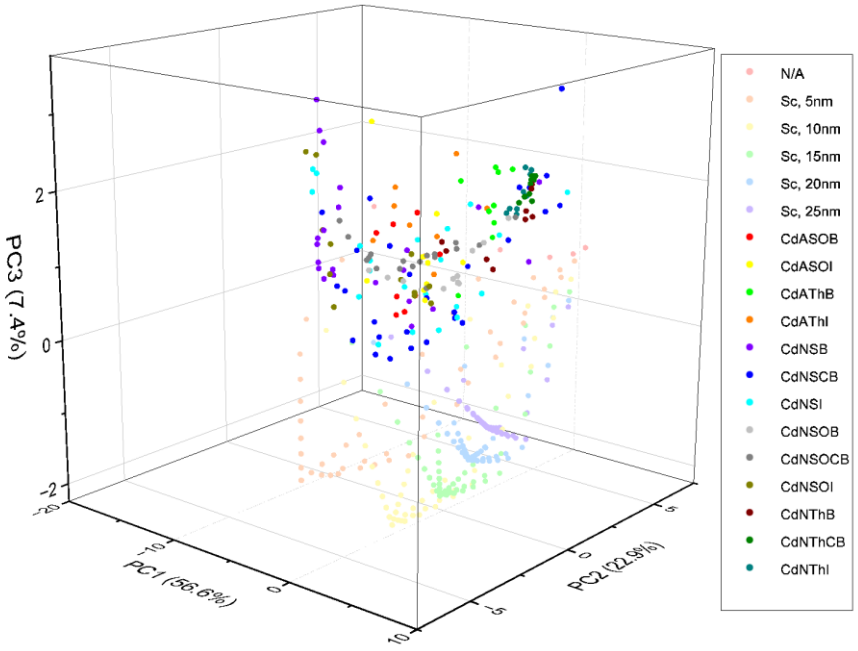


Figure 6.8: A CdS 3D simulated and real dataset score plot, labelling cubic crystallite size, where each simulated and real pattern is plotted against its PCs

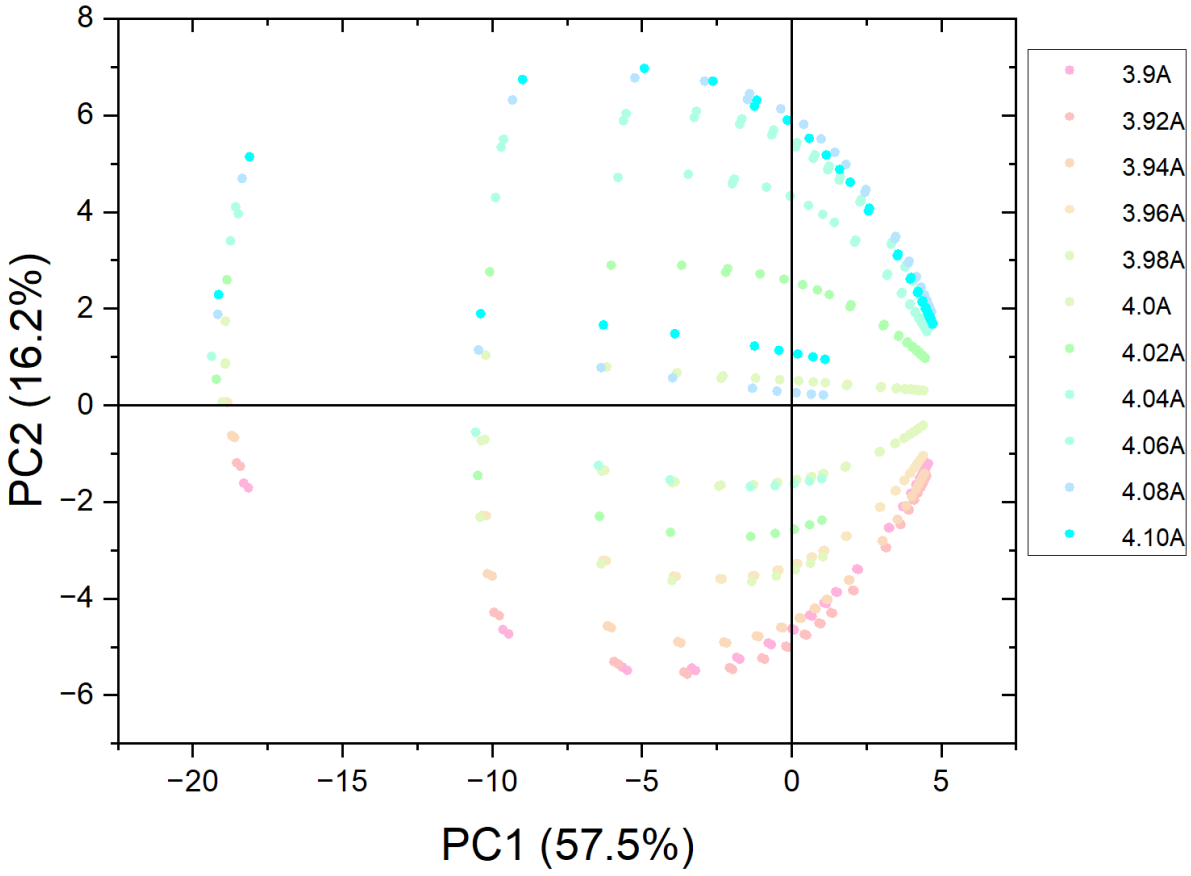


Figure 6.9: A $Ba_xSr_{1-x}TiO_3$ 3D simulated dataset score plot, labelling unit cell parameter, where each simulated pattern is plotted against its PCs

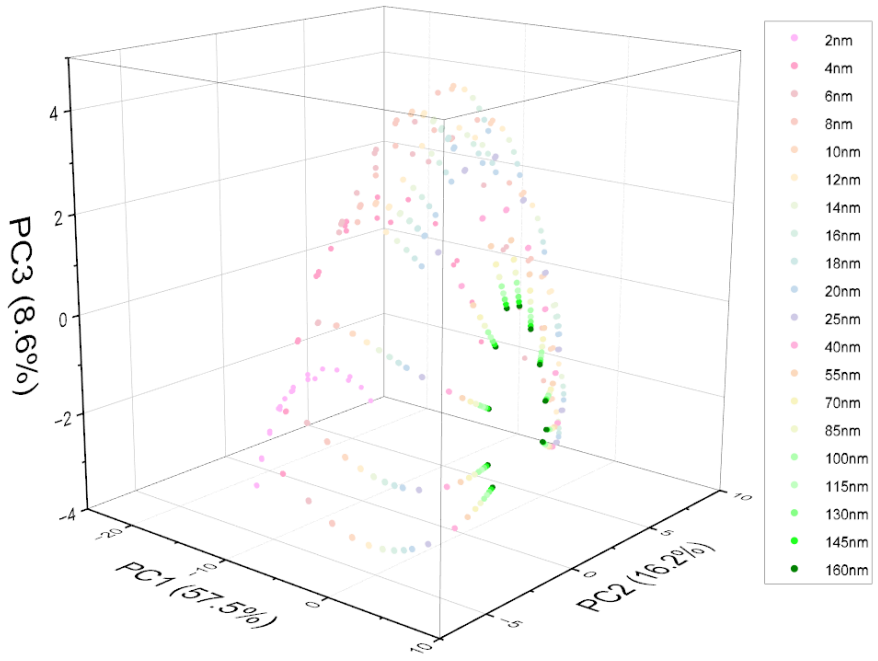


Figure 6.10: A $Ba_xSr_{1-x}TiO_3$ 3D simulated dataset score plot, labelling crystallite size, where each simulated pattern is plotted against its PCs

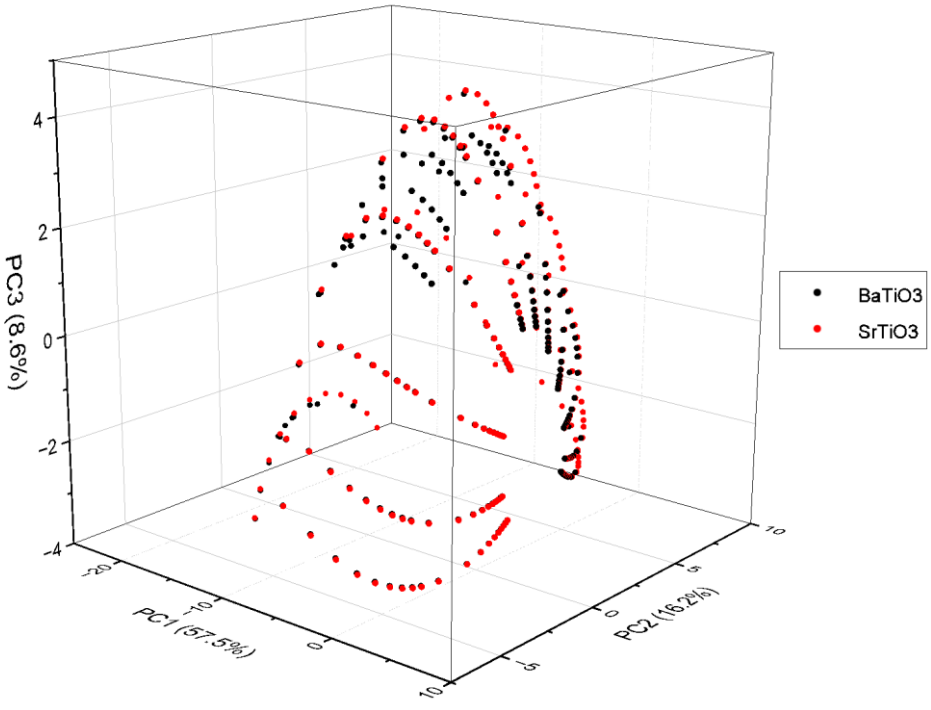


Figure 6.11: A $Ba_xSr_{1-x}TiO_3$ 3D simulated dataset score plot, labelling composition, where each simulated pattern is plotted against its PCs

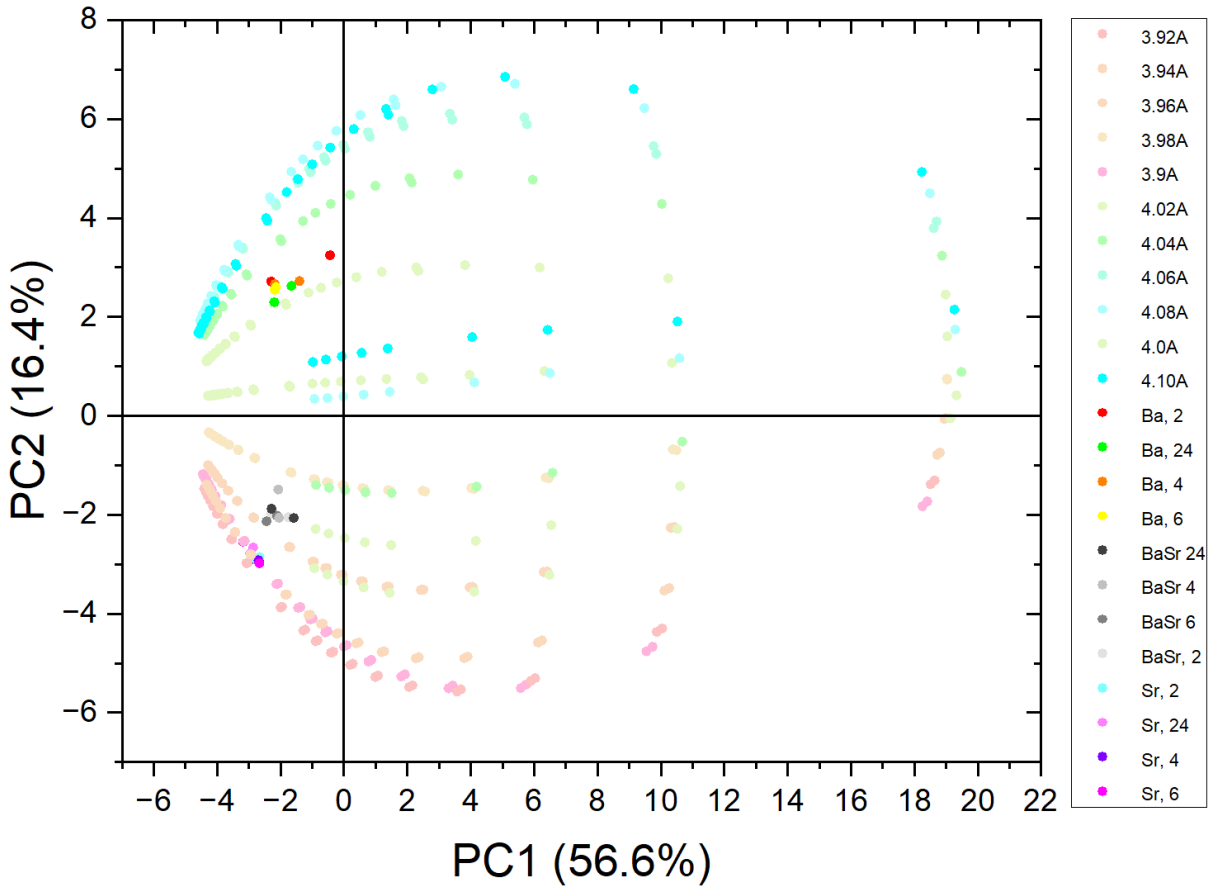


Figure 6.12: A $Ba_xSr_{1-x}TiO_3$ 3D simulated and real dataset score plot, labelling unit cell parameter, where each simulated and real pattern is plotted against its PCs

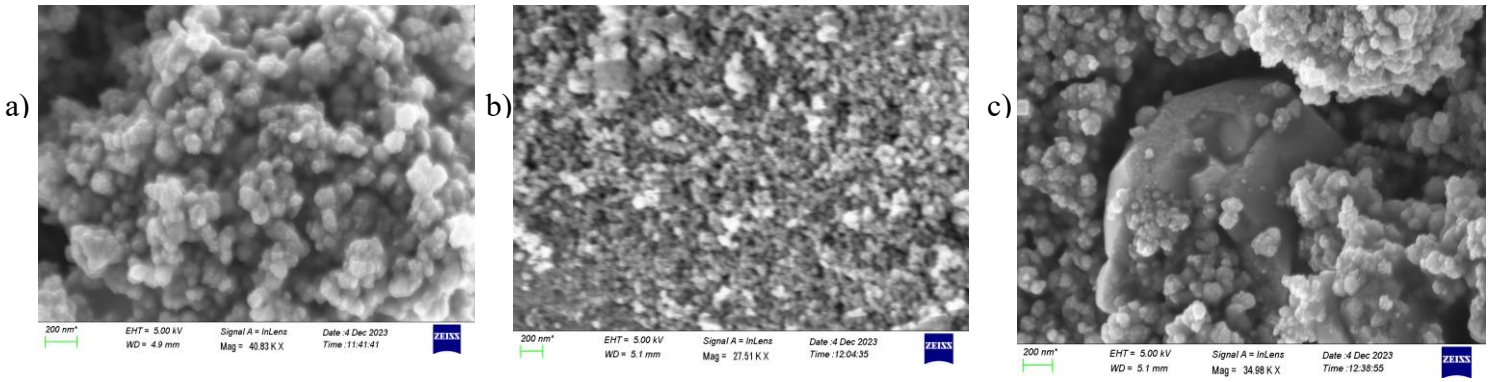


Figure 6.13: SEM pictures of $Ba_xSr_{(1-x)}TiO_3$ samples synthetised at a temperature of 200 °C, for 24 hours (a) is $BaTiO_3$, (b) is $SrTiO_3$, and (c) is $Ba_{0.5}Sr_{(0.5)}TiO_3$