Master 2 Internship Project
Report

Statistical identification of protein aggregates in protein crystallography datasets

Submitted by

Martin Garic
Université Paris Cité

Under the guidance of

Martin Lenz &
Lara Koehler

Laboratoire de Physique Théorique et Modèles Statistiques
Laboratoire de Physique Théorique et Modèles Statistiques
Univ Paris-Saclay
Rue André Rivière
91405 Orsay

Internship 2022

Acknowledgment

I would like to thank all the people who contributed to the success of my internship and who helped me during the writing of this report.

First of all I would like to thank Martin Lenz, my internship supervisor, for accepting me in his team for the 6 previous months. Thanks to his guidance throughout this internship I was able to learn a lot. I would also like to thank Lara Koehler who was here to talk or when I needed help.

I would like to thank the whole Soft Biophysics team for exchanging with me and sharing their kindness and experience. Thank you for making my stay at LPTMS so enjoyable.

Contents

Introduction

Proteins are large biomolecules that play an important role in life [1]. They serve multiple functions such as cell migration, DNA replication and catalysis. To fulfill their purposes some proteins assemble into fibres. For example most of the proteins in the cytoskeleton naturally form fibres to give the cell its shape and mechanical properties.

However, there also exist proteins that form fibres in certain diseases. These proteins are usually soluble, but following a mutation, their structures and interactions are affected which can lead to fibre formation. Three of those disease are Sickle-cell anemia [2] or neurodegenerative conditions such as Alzheimer's and Parkinson's. Understanding the physical properties that generally lead to fibre-like aggregation could help understand the cause of such diseases.

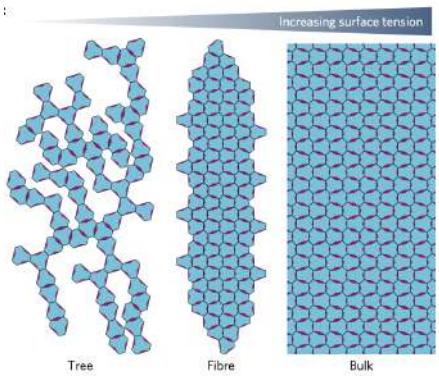

Figure 1: Fibre formation is determined by the surface tension between particles [3]

Our team has recently theorized the process leading to frustrated self-assembly of fibres [3]. According to this theory, general physical principles are responsible for these formations. This would mean that fibres are not only built by disease-inducing proteins but any protein under the right circumstance. As we can see in the Figure 1, surface tension is a parameter that could determine the aggregate geometry for irregular-shaped proteins. Thus, the competition between geometrical frustation and attractive short range interactions would dictate the type of aggregate that is formed. As we do not have data to confirm this theory, we had to search for people who possessed these types of data. X-ray crystallographers study protein structures by making protein crystals, thus they work on a daily basis on protein self-assembly. Their work consists in combining proteins with certain solutions to create crystals. This process is still not fully understood so they have a lot of data where the proteins have not crystallized. If the theory [3] proves itself correct, we should find protein fibres when the proteins don't crystallize. This will be explained more in details in the next chapter. Their data could be interesting for us.

The main objective of this project is to find naturally occurring protein fibres and understanding what causes such aggregations. At this stage of the project we will try to determine if there exist different aggregate types before focusing on fibre-like aggregates. This explanation will be separated in multiple chapters. In the first chapter we will talk about the basics behind X-ray scattering and why the dataset we use is the most suited for this project. In the second chapter we will study the preprocessing of the data to make it more convenient for analysis. Finally, we will see how we can classify this data depending on aggregation levels of the proteins with different methods.

This internship is being held at the Laboratoire de Physique Théorique et Modèles Statistiques or LPTMS under the supervision of Martin Lenz and Lara Koehler. Martin Lenz is a CNRS senior researcher and Lara Koehler is his Ph.D. student. The internship started in January where I worked 2 days a week until March. Then, I became a full-time intern and I will finish in the end of July.

1. Using X-ray crystallography datasets to find protein fibres

In this chapter we will see the basics of photon scattering and how such data will be usefull for our project. In the first section photon scattering will be explained in details. Next, the motivations behind the choice of working on data from X-ray crystallography rather than SAXS will be discussed. Finally we will see what type of experiments we did at Synchrotron SOLEIL.

1.1 Photon scattering helps find the structure of small molecules

Photon scattering is used to study the structure of small objects such as molecules [4]. Depending on the way this method is used, it is possible to study objects at length-scales as big as the length of protein fibers or as small as the distance between two atoms in a molecule (Fig 1.3).

This technique consists in shooting high frequency photons (wavelength between 0.1 and 10 nm) on objects (here proteins) and looking at the way these photons scatter when coming in contact with the molecules' electrons. From the scattering figure (Fig 1.1) it is then possible to infer information on the structure of the studied object.

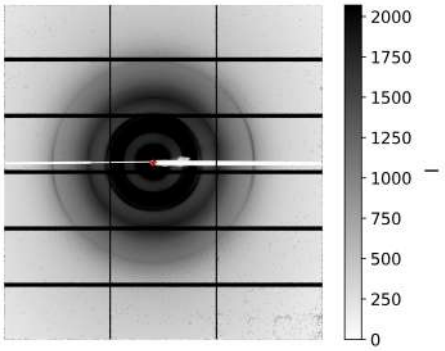

Figure 1.1: Scattering figure of Thaumatin. We can observe the scattered intensity of the photons characterised by rings of different intensities on a photon detector

To understand this method, let $\mathbf{k_i}$ be the wave vector of the incident photon beam and $\mathbf{k_f}$ the wave vector of the scattered photon beam (Fig 1.2). Considering that the scattering is elastic (the energy if the incident photon is the same as the energy of the scattered one), the amplitude of $\mathbf{k}$ is defined by:

$$k = \frac{2\pi}{\lambda} \tag{1.1}$$

with λ the wavelength of the photon. As we want to know the number of photons scattered as a function on the scattering angle, it is possible to define a scattering vector defined by:

$$\mathbf{q} = \mathbf{k_f} - \mathbf{k_i} \tag{1.2}$$

$$\begin{aligned} \|\mathbf{q}\| &= \sqrt{(\mathbf{k_f} - \mathbf{k_i})^2} \\ q &= \frac{4\pi \sin(\theta)}{\lambda} \end{aligned} \tag{1.3}$$

Here, θ is the half angle between the incident vector $\mathbf{k_i}$ and the scattered vector $\mathbf{k_f}$ (Fig 1.2).

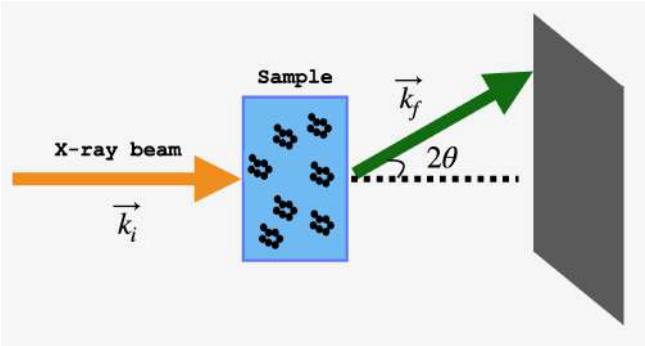

Figure 1.2: The scattering vector $\mathbf{k_i}$ is scattered by electrons in the sample, and the scattered vector $\mathbf{k_f}$ is measured on a screen

It is then possible to know the number of scattered photons as a function of the scattering vector $\mathbf{q}$ [5]. Let this number be denoted as A:

$$A(\mathbf{q}) = \sum_j f_j e^{i\mathbf{q}\mathbf{r}_j} \tag{1.4}$$

where f_j denotes the interaction characteristics of the jth atom and $\mathbf{r}_j$ the position of that atom. The scattering intensity reads:

$$I(\mathbf{q}) = A(\mathbf{q})A^*(\mathbf{q}) \tag{1.5}$$

The scattering intensity is partially determined by the interactions between the different particles because of f_j. Thus, if we consider identical non-interacting particles, the scattering intensity is just N times the scattering intensity of a single particle. When there are interactions, we introduce a Structure Factor $S(\mathbf{q})$ which contains information about the interactions between the particles. If particles do not interact the structure factor is one. Otherwise it varies depending on the types of interactions. The scattered intensity can be written as a product of the structure factor and the form factor $F(\mathbf{q})$ (scattering intensity of a single particle):

$$I(\mathbf{q}) = S(\mathbf{q})F(\mathbf{q}) \tag{1.6}$$

This confirms that there are information on the aggregate type in the scattering signal.

1.2 We use X-ray crystallography data rather than SAXS because lots of data is available

In the previous section, the basics of photon scattering were introduced. In this section we will see that there exist different ranges of scattering wavevectors $\mathbf{q}$ that fulfill different purposes. On the one hand small angle X-ray scattering (SAXS) focuses on the size and shape of macromolecules [6]. On the other hand X-ray crystallography tries to resolve molecule structure with crystallization. Crystals are regular arrays of particles (which can be proteins) that diffract X-rays according to Bragg's law [7]. Indeed, peaks caused by constructive interferences of the scattered beams appear on the detector. The structure of the protein is then inferred using a Fourier transform to get the electron densities. We will now see why we use crystallography data rather than the more intuitive method that is SAXS.

The main parameter here is the scattering wavevector $\mathbf{q}$. This vector's amplitude is related to the inverse of studied object's size as seen in the equation 1.7 (D represents the size of the

observed object). This means that to observe bigger characteristics, a smaller scattering vector is necessary. In this section we will see that there exist two ranges of scattering vectors that fulfill different purposes. As we saw in equation 2.3, the parameter that mainly determines this vector is the angle θ. Thus we can understand the description of SAXS. "Small angle" refers to small scattering vector $\mathbf{q}$ and thus to bigger length-scales [8]. In our case we study protein aggregates so SAXS would be the appropriate method to use. SAXS studies length-scales going from 1 to 100 nm. For example, protein fibres we can find in the cell cytoskeleton have a diameter of the magnitude of 10 nm. As seen in the following equation, this corresponds to a q of $0.1\mathring{A}^{-1}$ if D is the diameter of the fibre.

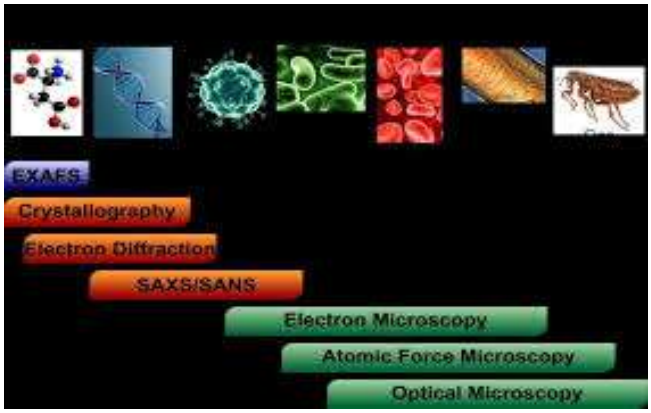

Figure 1.3: By Stanford BioSAXS Workshop 2016

$$q = \frac{2\pi}{D} \tag{1.7}$$

Confirming a theory that states that proteins generically form fibre-like structures under frustration requires a lot of data. Such data is needed because if this principle is true, this fibre-like aggregation should happen to many different proteins. Unfortunately, we do not possess this quantity of SAXS data.

Lucky for us, it is possible to get X-ray crystallography data whose range of q is $[0.06,5]\mathring{A}^{-1}$. This range is chosen because X-ray crystallographers are interested in smaller length-scales than SAXS physicists - they look at the interactions between atoms in proteins. This data comes from Synchrotron Soleil. Despite the fact that this range isn't optimum for finding clues on the aggregate type of the protein (because of the range of q), there might still remain information on these aggregate types. The reason why such data is available is that crystallographers study protein structures by crystallyzing proteins. The perfect conditions to make a protein crystal depend on many factors that are not fully understood by physicists so they prepare multiple samples hoping proteins aggregate into a crystal. These samples are usually composed of water, depletants and most importantly salts such as potassium sodium tartrate or sodium citrate. These molecules act in a way to minimise the energy necessary to cause aggregation. To do so, the salts participate in screening the electrostatic repulsion between the proteins and the depletants create a attractive force between the proteins by excluding themselves from the vicinity of these proteins the for entropic reasons [9]. These preparations are called buffers solutions when there are no proteins inside. When carrying out these preparations, only few of them produce crystals.

In the Figure 2.3 we can see a plate prepared by crystallographers. Each well of the plate has a different preparation. These crystallographers send a beam of electrons on the drops for a short period of time (200 ms) for each well so the data is collected very quickly and in large quantities. In the drops where there are no crystals, the conditions might be such that other aggregate types appear.

A difficulty with this data is that the experimental setup is not adapted for what we want. In fact, as crystallographers search for the structure of proteins, they are mainly interested in Bragg peaks from diffraction at bigger scattering wavevectors (whcich corresponds to length-scales). The scattering signal is composed of the signal of the proteins but also the plate and the buffer solution. The latter two do not make Bragg peaks so crystallographers ignore these signals. Since we care about the signal's shape, it is mandatory for us to remove this "noise" to make the dimensionality of the aggregates appear.

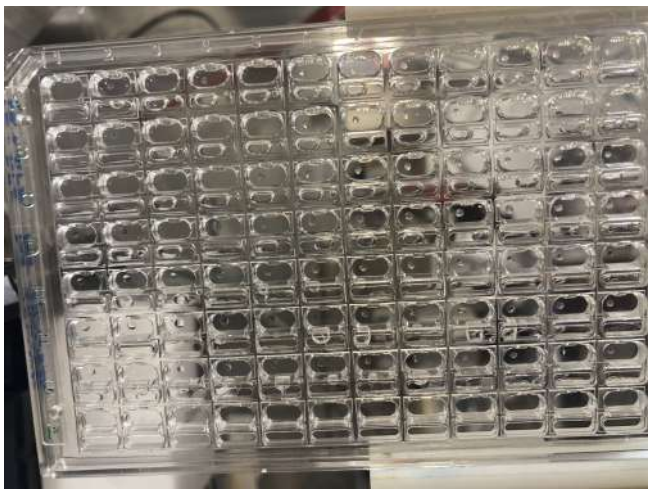

Figure 1.4: Example of a X-ray crystallography plate prepared by crystallographers. Each cell of the plate is called a well. A drop of buffer and protein is put in each well

As explained before, lots of data was necessary so we went to Synchrotron Soleil in the Proxima-2A beamline [10] to perform experiments. The samples were prepared for us in advance but we took care of getting the scattering figures of these samples. To prepare the samples, the proteins are put in a buffer solution and then a device called "Mosquito" places the drop (of constant volume) on the plate. The plate is then sealed and ready to be screened by the X-ray beam. Thanks to a computer program made for this beamline, we could control the position of the plate and chose where we wanted the photon beam to pass through. During three experiment sessions, three different proteins, Thaumatin [11], ISPE [12] and NATA [13] were prepared. As we had no way of knowing the dimension of the aggregate, we had to make guesses with the help of visual markers (Figure 4.1). For this reason, in the rest of the report we will concentrate on determining the dimension of the aggregate rather than if we have fibres or not.

In this chapter we explained the basics of X-ray scattering which allowed us to understand that this type of data was convenient for us because information on the structure of the aggregate. Then we saw the experimental side of the project which allowed us to fully understand how these experiments were held.

2. Isolating the protein signals in crystallography data

In the following chapter we will see that there exist two sources of "noise" (also called "background") that might perturb or analysis of the data. This takes into account the unwanted scattering that arises from sources other than the proteins of interest [14]. This background decreases the signal to noise ration so we will show that is is possible to remove it to make different experiences comparable.

First of all, it is important for us to introduce a new way of expressing $I(\mathbf{q})$:

$$I(\mathbf{q}) = I_{\text{background}}(\mathbf{q}) + I_{\text{proteins}}(\mathbf{q}) \tag{2.1}$$

The reason why this is possible is that in X-ray scattering we consider that the proteins do not interact with the background so their scattering signals are independent [14]. It is possible to write the total scattered signal as a sum of each non-interacting component. Here the two components are the proteins' and the background's signal [14].

It is then possible to decompose the background's signal again in two different signals: the plate's signal and the buffer solution's signal:

$$I_{\text{background}}(\mathbf{q}) = I_{\text{plate}}(\mathbf{q}) + I_{\text{buffer}}(\mathbf{q}) \tag{2.2}$$

The section 2.1 will be dedicated to subtracting the plate's signal. The section 2.2 will tackle the subject of the buffer. As a reminder the buffer is the solution in which the proteins are put to crystallize.

2.1 Removing signal from the plate

This section is dedicated to understanding how the plate's signal should be removed. In droplet in which we find proteins is placed on a plate (Fig 1.4), the plate also scatters high energy photons. This scattering signal represents a large part of the total signal (Fig 2.1). As we want to remove the plate's signal, this would, in principle, require the collection of one scattering signal for each type of plate (plates from different manufacturers have different compositions). Indeed there exist different brands of crystallography plates and from one experiment to another the plates can change.

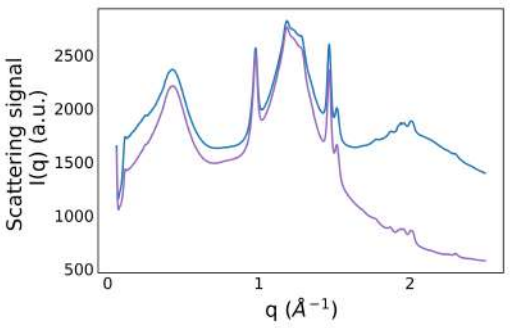

Figure 2.1: Two different scattering signals. Blue: plate + drop; Purple: plate

To verify that a single scattering signal per type of plate was sufficient we collected data for different points on the plate. In the Figure 2.2-b we can see a lattice. This lattice corresponds to the different wells in the plate where we collected

the data. The coordinates of the cells are given by a letter (A to G) and a number (1 to 12). As we can see the mean intensity of the signal varies spatially, allowing us to deduce that the plate's signal is not spatially constant.

Let's now introduce the relative standard deviation (RSD) to verify if these variations are due to different thickness or different compositions of the plate:

$$RSD(\mathbf{q}) = \frac{\sigma(\mathbf{q})}{\mu(\mathbf{q})} \tag{2.3}$$

with σ the standard deviation of the data and μ the mean. If we now look at the RSD as a function of the scattering vector (Fig 2.2-a) we notice peaks. If the RSD was constant, it would have meant that the variations were only due to thickness but it isn't the case. This means that the variations of intensity are due to the thickness of the plate but also variations in composition.

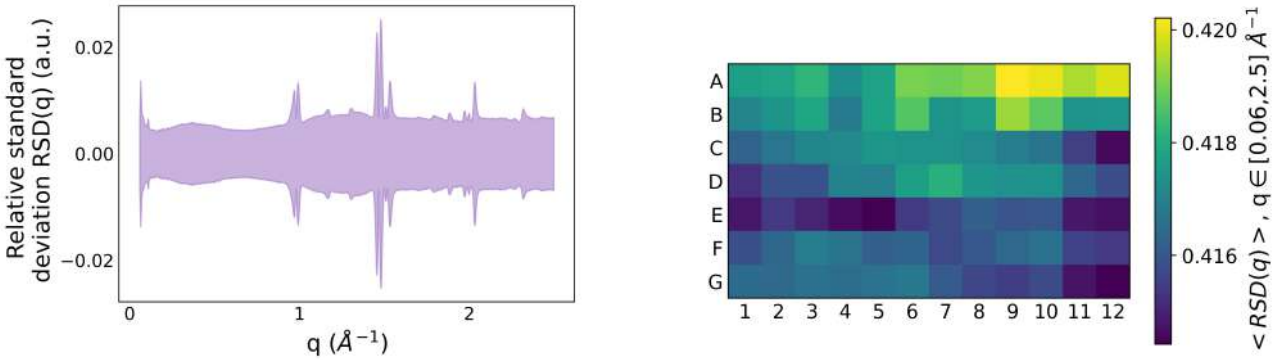

Figure 2.2: (a) Relative standard deviation of the signals taken on different places on a plate. (b) Averaged scattering intensity map of a plate $< RSD(q) >$, q$\in$ $[0.06, 2.5]$Å^{-1}

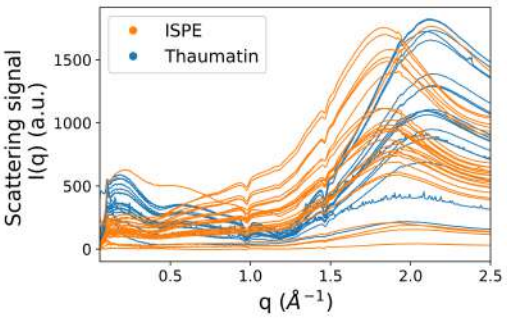

Figure 2.3: Scattering intensities for two proteins in various conditions after removing the plate. There remains the proteins' and the buffer solution's signal

We previously saw that the plate's signal varies spatially. This means that if we only took one signal per plate to subtract, the subtraction would not be accurate. To overcome this problem it is essential to get the scattering intensity of the plate next to each drop so that we subtract the signal. The resulted signals can be seen in the Figure 2.3. In fact, having removed a big block of unwanted signal allows us to see what remains.

Another way to observe the differences between the signals is by using Principal Component Analysis (PCA) [15] [16] which we will introduce now.

Large datasets such as ours are complicated to interpret. Principal component analysis is a technique used to reduce the number of dimensions in the dataset to make classification easier. If we have a dataset of size N where each data is of size M, the data is in the shape of a matrix $T = [N, M]$. The objective of the PCA is to reduce M while minimising the loss of information. To fulfill this goal, a set of m¡M new orthogonal axis are defined in a way to maximise the variance of the data on these new axis. The new axis are called principal components. The new data is then projected along these principal components. Each principal component has a "relative explained variance" which corresponds to the amount of information from the original dataset that is preserved in this

principal component. If u is the set of new vectors composing the new basis, the new dataset T' of shape $[N, m]$ is defined by:

$$T' = T * u \tag{2.4}$$

The Figure 2.4 is a PCA of scattering intensities of two plates from the same manufacturer from two different experiments (purple vs blue). We can see that the points are clustered together by color. This color corresponds to the different experiments. The difference in signals is bigger between two plates rather than in a plate. The Figure 2.4-b is a plot of the two first principal components as a function of the scattering vector. This gives us indications of what points on the scattering signal are important for the PCA when creating these new components.

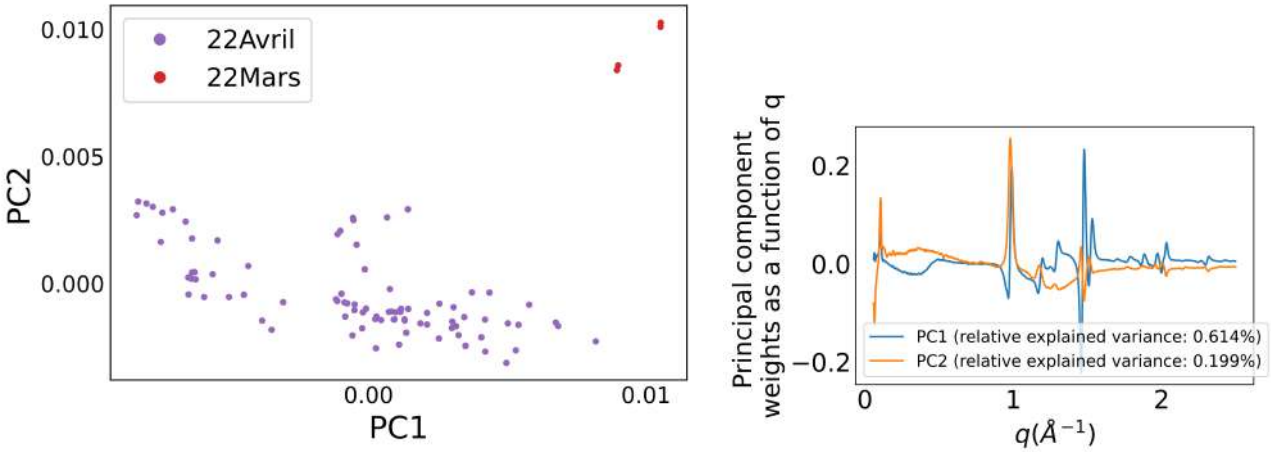

Figure 2.4: (a) Principal Components Analysis on the data of different plates. (b) Plot of the two first components as a function of the scattering vector and their relative explained variances.

Principal component analysis is the main method that we will use to visualize data in the rest of this report. In this section we saw the importance of removing the plate's signal, in the following section we will tackle removing the buffer's signal.

2.2 Removing the buffer's signal

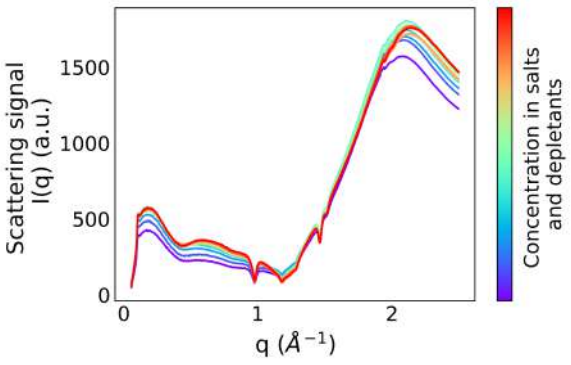

Figure 2.5: Scattering signal of a stock solution composed of rising concentrations of Na/K tartrate and ADA at pH 6.5

In this section we will tackle the subject of removing the buffer's signal. As mentioned earlier, the buffer is the solution in which the proteins are put to form crystals. Thus, it is important to know this solution's signal so we can subtract it. The Figure 2.5 shows the scattering signals of stock solutions with varying concentration of salts and depletants.

The figure 2.5 shows us the scattering signal of buffer solutions for different concentrations of salt and depletants. As we can see the signal varies with concentration. At smaller scattering wavevector (between 0.06 and 1 $Å^{-1}$), the intensity rises with the concentration of the stock solution. This

is due to the fact that if we put more "particles" in a solution, more will have the chance to scatter photons.

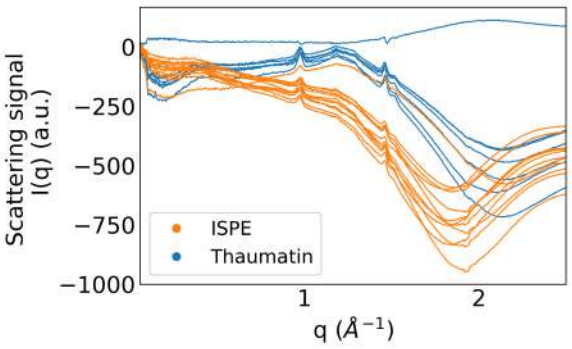

Figure 2.6: Scattering signals of two proteins Thaumatin and ISPE after removing the background signal. Here there only remains the proteins' signals.

After removing both the plate and the buffer's signal, the resulting scattering signals are shown in the Figure 2.6. It is not surprising to see a negative scattering intensity. In fact, when the photon beam passes through the droplet, it is possible for proteins to take the place of other molecules (water). These proteins may scatter less than water in these ranges of scattering vector so when the buffer's signal is subtracted, negative intensities can appear. In this figure we see the signals of two different proteins: Thaumatin and ISPE for various concentrations of buffer solution. We will see in the next chapter that the variations between the signals are due to differents aggregate types.

In this chapter, our work allowed us to remove the unwanted background from our signals. This allows us to only see the proteins' signal without being disrupted by the background for further analysis. We also introduced principal component analysis which will be used in the following chapter to classify the data in the search for different aggregates.

3. Attempting to classify signals with different strategies

We previously saw that in the scattering signal there was information on the structure of the protein aggregates. In this chapter we will first try to extract this information thanks to principal component analysis and classify the signals according to aggregation types in the section 3.1. We will see that there exist limitations to this method. Thereby, in the section 3.2, we will create a machine learning model that will be capable of classifying the signals regardless of these limitations.

3.1 PCA is capable classifying protein signals for a single experiment but has limits when we need to compare different ones

During our experiments, whenever we got a scattering data, we also got a microscopy image of the drop we sent the beam through. On each drop we could see different visual markers (Fig 3.1). These visual markers are information on what is happening in the drop. We had 6 different labels:
OutsideDrop: On the plate
StockSolution: In the buffer solution
Crystal: On protein cystals
InsideDrop: Somewhere clear in the drop
Phases: In the drop where we can observe phase separations
Precipitate: In the drop where we can observe precipitates of different shades of grey

The main objective of the project is to find protein fibres. But as recalled earlier we do not possess data on fibre's signals so we will first concentrate on determining the aggregation type of proteins. Precipitates and phase separations are a proof something is happening to the proteins so a good guess is that they are not monomeric anymore. Our goal is then to determine what signal characterises the different categories. To do so, we will first use PCA. To perform such analysis it is important to preprocess the raw data. Our data is a matrix of shape $[N, M]$ where N is the number of scattering signals and M the number of points per signal. For each column along the M axis, we normalize the data by centering the points around and put the standard deviation at 1.

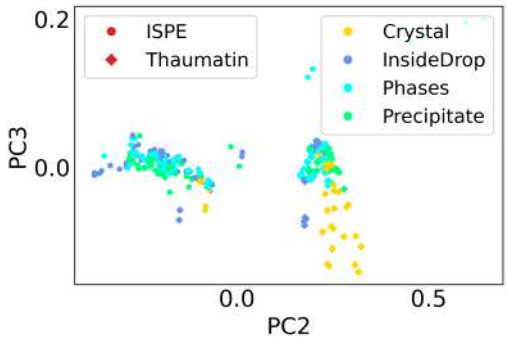

Figure 3.1: PCA of data from a single plate. There are two proteins, ISPE and Thaumatin. The axes are the second and third principle components

The Figure 4.2 shows the PCA for the experiment where we were able to remove the background. The color-code corresponds to the aggregate

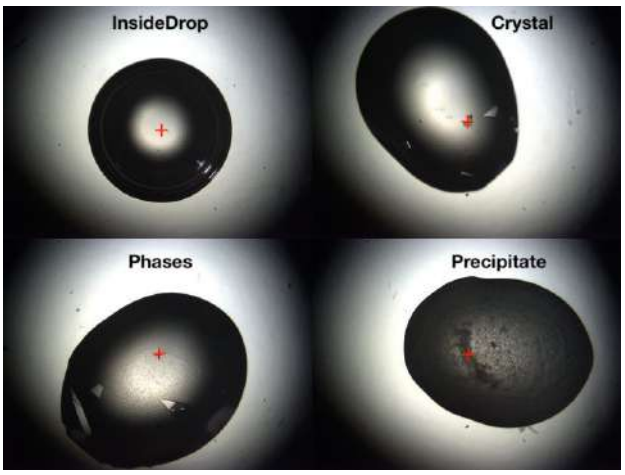

Figure 3.2: Different categories of shots inside a drop

types mentionned earlier. We can see that the two proteins do not group together as their scattering signals are too different. But, the crystals are clearly identifiable, they are segregated in the PCA plane. When it come to precipitate/phases versus clear drops it is more unclear.

We are now going to look at what happens when we compare multiple experiments. We discovered during the course of the internship that the pre-processing mentioned in the previous chapter was important for data visualisation. Thereby, the first experimental data we collected didn't have the buffer solution's signal nor the plate's signal. This means that we can't remove the background's signal if we want to compare the signals. To visualise these differences we will introduce an alternative to PCA called t-distributed stochastic neighbor embedding (t-SNE) [17]. This method first transforms the high-dimension data into a matrix of similarities between the datapoints. Let $p_{j|i}$ be the similarity between a datapoint x_j and x_i. $p_{j|i}$ is then denoted as:

$$p_{j|i} = \frac{exp(-||x_i - x_j||^2/2\sigma_i)}{\sum_{k \neq i} exp(-||x_i - x_k||^2/2\sigma_i)} \tag{3.1}$$

where σ_i is the Gaussian variance centered on x_i which we search for. It is possible to introduce a cost function whose parameters will be the σ_i. The algorithm then carries out a gradient descent to find the optimal set of σ_i to get a correct interpretation of the data. The Figure 3.3 shows the t-SNE plots where there are three different experiment sessions and three different proteins (ISPE, Thaumatin and NATA).

Each subfigure shows the same data but with different color schemes corresponding either to aggregate types, experiment session or protein type. As we see, when we mix multiple experiments, algorithms tend to remark these differences first.

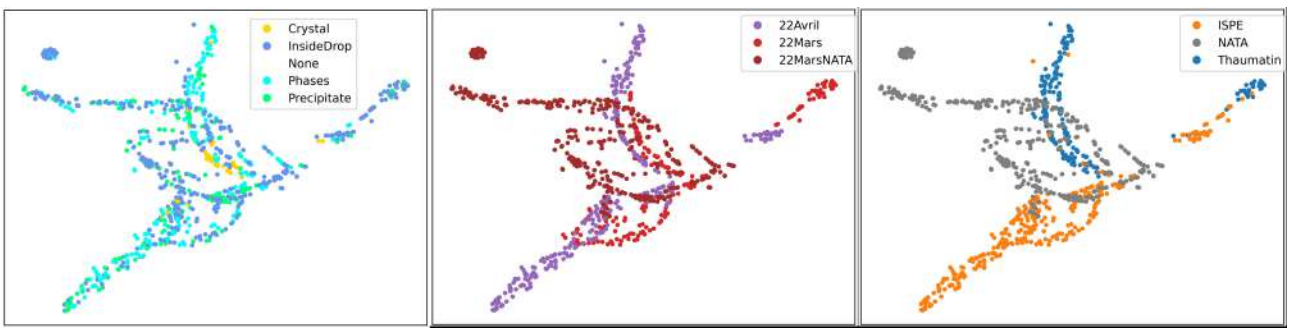

Figure 3.3: t-SNE plot used to visualise what category the algorithm groups together

In the next section we will see that Machine Learning covers what PCA can't do. Indeed, we created a machine learning algorithm able to classify aggregate types.

3.2 Machine learning can compensate for what PCA lacks

There exist different ways to classify large datasets and one of those ways is supervised machine learning [18]. This method where we train an algorithm on a labeled dataset. The algorithm is able to learn the differences between the different classes to predict which data belongs in which class. This method in convenient for us as we have scattering signals labeled by the microscopy photo taken when collecting the data.

Before applying this method to our data it is necessary to perform a preprocessing step. Indeed each signal is composed of 1000 points which are considered as 1000 features for the machine learning algorithm. To reduce this number we can perform a PCA on the normalized dataset and keep the n first principle components in order that 99% of the relative explained variance is kept. This amounts to keeping the $n = 30$ first principle components. This procedure has allowed to reduce the dimensionality of the dataset drastically. The next step was to build "neural network" adapted to our dataset. The neural network is the heart of artificial intelligence and machine learning. A neural network is a set of connected nodes on multiple layers which interact through mathematical functions to collect information on the relevant features necessary for classification. In our case neural network is composed of 4 layers of 25 neurons each and an output layer. The output layer is a vector of size the number of classes we have.

Here we have 5 classes which are (with there respective amounts of data): Stock Solution (154), Precipitate/Phases (445), InsideDrop (441), OutsideDrop (97), Crystal (57). These numbers are quite evenly distributed (except for the crystals as there were not many) on the three experiments. This is important because if there is an uneaven distribution, the algorithm might be biased towards the class that has more data.

On the data-set we used in the Figure 4.2, we trained the algorithm and tested it. The results of the predictions can be seen in the Figure 4.4. The numbers correspond to the number of predictions in each case. Diagonals correspond to correct answers while the rest are mispredictions.

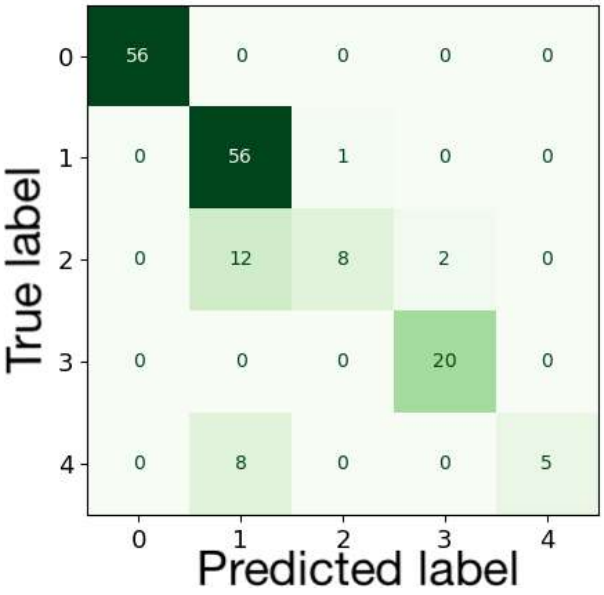

Figure 3.4: Results of the test series after training on the data from a single experiment (April 2022). The proteins are ISPE and Thaumatin. The categories are 0:Stock Solution, 1:Precipitate/Phases, 2:InsideDrop, 3:OutsideDrop, 4:Crystal

We can see that the neural network works with 86% accuracy. We can observe a few mispredictions for InsideDrops and Crystals who are mistaken for Precipitate/Phases. The cause of this is that we label the signals according to the picture that is taken. Thus, if the proteins aggregate, we might not always see it with a microscope and we might mislabel them.

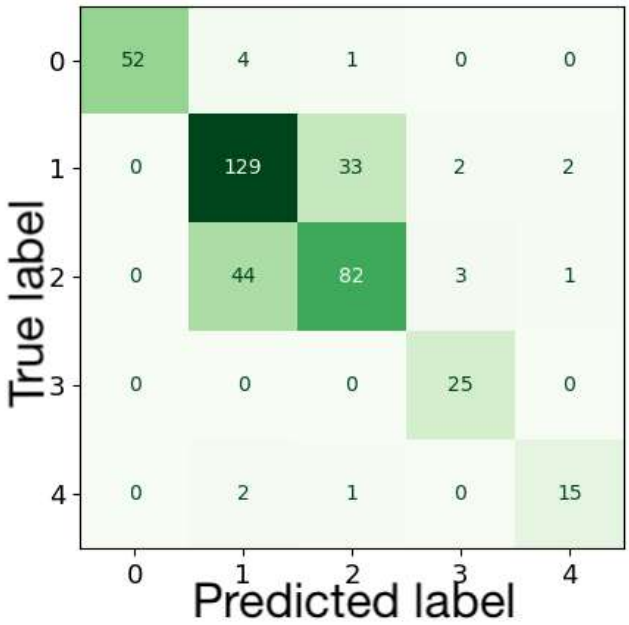

Figure 3.5: Results of the test series after training on the data from every experiment in 2022. The proteins are ISPE, NATA and Thaumatin. The categories are 0:Stock Solution, 1:Precipitate/Phases, 2:InsideDrop, 3:OutsideDrop, 4:Crystal

When adding multiple experiments, the results are quite similar (77% accuracy). Indeed, the main errors come again from a misprediction of InsideDrop and Precipitate/Phases. Once again these errors might be due to the fact that we label the data according to visual markers. Our labels are deduced from what we see with a microscope but we are trying to label aggregate types so it is possible that some drops we labeled as phases or precipitates were actually soluble proteins and vice-versa.

To make this algorithm better (get better accuracy), much more data will be necessary. We know that this limits the current algorithm because when we complexify the neural network we get over-fitting. Over-fitting happens when the network learns the details of the training set negatively impacting the quality of the predictions on a test set.

This use of machine learning has proved that there exist different signals depending on visual markers that should correspond to aggregation states of the proteins of interest. More analysis is needed to understand what these aggregate types are and if fibres can be found.

In the chapter we saw how PCA and machine learning allowed us to determine differences between signals we cannot differentiate by the bare eye. PCA held limitations when it came to comparing data from different experiments but a simple yet effective neural network was able to predict aggregation type quite robustly. Supervised learning appears to be adapted. Even though a bigger database would most likely allow us to get better performances on our algorithm, we were able to confirm that it is possible to predict the type of aggregation happening in crystallography data-sets.

Conclusion: Our work has proved that it is possible to classify protein scattering signals according to arbitrary categories through data analysis and machine learning

During this internship, the main goal was to collect data from X-ray crystallography and attempt to classify these data-sets by aggregate type. Obtaining a large quantity of data and finding different aggregate types for many different proteins is indeed essential to prove the hypothesis that fibres form from common physical laws. We saw that the data we collected at Synchrotron Soleil was not the most optimized - compared to SAXS - to study what we studied but the large amount of data available countered this inconvenient. We then saw some essential preprocessing of the data to increase the signal to noise ratio. To do this the plate and the buffer's signals needed to be retrieved. We then noticed that with Principle Component Analysis some form of classification was happening when we were interested at aggregate types. When tackling the question of analysing multiple experiments together, we needed to use supervised machine learning to compare the signals. Our machine learning algorithm proved itself useful because we were able to confirm that the signals were differentiable.

For a better implementation of the algorithm while keeping a similar architecture, data on more proteins would be necessary, here we only had three proteins: ISPE, NATA and Thaumatin. Another option would be to test the algorithm on proteins that naturally form fibres.

During the following month of the internship I will work on studying the weights of the neurons in the network. This corresponds to studying what parts of the signals the algorithm "looks at" when attempting to classify the data. This study is not trivial as there can be correlations between far away points on the scattering signals. This could give us insight on what is a physical marker of these aggregate types.

References

[1] A. D. Bershadsky and J. M. Vasiliev, "Cytoskeleton," 2012.

[2] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, "The Protein Data Bank," *Nucleic Acids Research*, vol. 28, pp. 235–242, 01 2000.

[3] M. Lenz and T. A. Witten, "Geometrical frustration yields fibre formation in self-assembly," *Nature physics*, vol. 13, no. 11, pp. 1100–1104, 2017.

[4] V. Iveronova and G. Revkevich, "Theory of x-ray scattering," *Moscow Izdatel Moskovskogo Universiteta Pt*, 1978.

[5] D. S. Sivia, *Elementary scattering theory: for X-ray and neutron users*. Oxford University Press, 2011.

[6] A. G. Kikhney and D. I. Svergun, "A practical guide to small angle x-ray scattering (saxs) of flexible and intrinsically disordered proteins," *FEBS letters*, vol. 589, no. 19, pp. 2570–2577, 2015.

[7] C. G. Pope, "X-ray diffraction and the bragg equation," *Journal of chemical education*, vol. 74, no. 1, p. 129, 1997.

[8] C. Giannini, V. Holy, L. De Caro, L. Mino, and C. Lamberti, "Watching nanomaterials with x-ray eyes: Probing different length scales by combining scattering with spectroscopy," *Progress in Materials Science*, vol. 112, p. 100667, 2020.

[9] Y. Mao, M. Cates, and H. Lekkerkerker, "Depletion force in colloidal systems," *Physica A: Statistical Mechanics and its Applications*, vol. 222, no. 1-4, pp. 10–24, 1995.

[10] D. Duran, S. Le Couster, K. Desjardins, A. Delmotte, G. Fox, R. Meijers, T. Moreno, M. Savko, and W. Shepard, "Proxima 2a–a new fully tunable micro-focus beamline for macromolecular crystallography," in *Journal of Physics: Conference Series*, vol. 425, p. 012005, IOP Publishing, 2013.

[11] J.-J. Liu, R. Sturrock, and A. K. Ekramoddoullah, "The superfamily of thaumatin-like proteins: its origin, evolution, and expression towards biological function," *Plant cell reports*, vol. 29, no. 5, pp. 419–436, 2010.

[12] T. Wada, T. Kuzuyama, S. Satoh, S. Kuramitsu, S. Yokoyama, S. Unzai, J. R. Tame, and S.-Y. Park, "Crystal structure of 4-(cytidine 5-diphospho)-2-c-methyl-d-erythritol kinase, an enzyme in the non-mevalonate pathway of isoprenoid synthesis," *Journal of Biological Chemistry*, vol. 278, no. 32, pp. 30022–30027, 2003.

[13] B. Polevoda and F. Sherman, "Composition and function of the eukaryotic n-terminal acetyltransferase subunits," *Biochemical and biophysical research communications*, vol. 308, no. 1, pp. 1–11, 2003.

[14] B. R. Pauw, A. J. Smith, T. Snow, N. J. Terrill, and A. F. Thünemann, "The modular saxs data correction sequence for solids and dispersions," *arXiv preprint arXiv:1706.06769*, 2017.

[15] R. Bro and A. K. Smilde, "Principal component analysis," *Analytical methods*, vol. 6, no. 9, pp. 2812–2831, 2014.

[16] H. Abdi and L. J. Williams, "Principal component analysis," *Wiley interdisciplinary reviews: computational statistics*, vol. 2, no. 4, pp. 433–459, 2010.

[17] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne.," *Journal of machine learning research*, vol. 9, no. 11, 2008.

[18] J. Yosinski, J. Clune, A. Nguyen, T. Fuchs, and H. Lipson, "Understanding neural networks through deep visualization," *arXiv preprint arXiv:1506.06579*, 2015.