Characterizing the dimension of protein aggregates through simulated X-ray scattering data

Marianne Billoir

May 10^{th} - July 30^{th}, 2021

Supervisor:

Martin Lenz

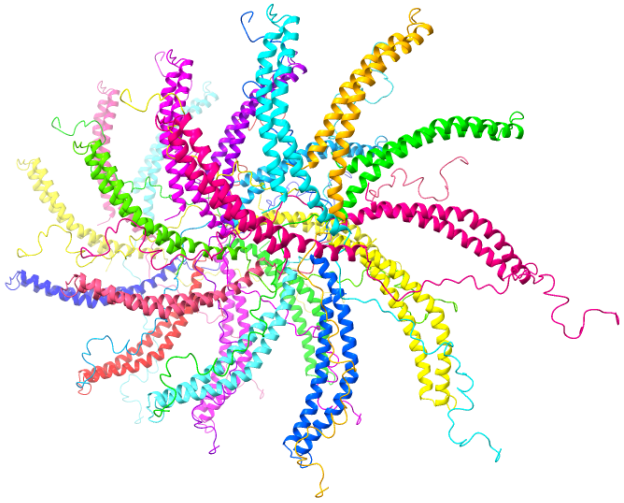

Figure 1: Picture of an aggregate of α-synuclein built numerically with the software ChimeraX

LPTMS
Paris-Saclay University
Orsay

Second year of the Magistère of
fundamental Physics of Orsay
Paris-Saclay University
Orsay

Laboratoire de Physique Théorique et Modèles Statistiques

université PARIS-SACLAY

Acknowledgements

I warmly thank my supervisor, Martin Lenz, for welcoming me as an intern within his team as well as for all his precious advice and lessons. I also especially thank Lara Koelher for being a great guide to me throughout my internship, for sharing her experience with me and finally for her kindness and patience towards me. I also thank all the Soft biophysics team of the laboratory for their warm welcome into the team. Finally, I thank Claudine Le Vaou for making all the administrative details easy.

Abstract

It sometimes happens that proteins normally soluble under conditions where they fulfill their function irreversibly aggregate under the shape of fibres, leading to pathologies. If theories on the mechanisms driving such aggregations already exist, we want to test the hypothesis that the formation of fibres in general - pathological as well as physiological - arises from generic physics principles. To that aim, we wish to use the large amount of protein aggregation crystallography data already available. However, these data are not well suited to retrieve the type of aggregate that has formed - hence the need to first determine the possibility to use such data.

To do so, we chose to numerically build our own database of protein aggregates starting from the 3D-structure of single proteins. Then we simulated their scattering intensities and trained a machine learning algorithm on it. From this, we have been able to characterize the type of aggregate associated with a given scattering profile with a high accuracy. We now need to use our trained algorithm on experimental data to confirm our preliminary results.

Contents

1 Introduction

Proteins are biological macromolecules at the foundation of every living system. They thus have numerous functions. They can play a structural role such as components of our cell membranes like the macromolecular actors of muscle contraction. They are also directly involved in the chemical processes essential for life[1].

Some proteins aggregate into fibres to fulfill their role such as in the case of muscle contraction where filaments of actin and myosin slide over each other to contract or relax the muscle. However, proteins that are normally soluble under conditions where they fulfill their function may in some cases irreversibly aggregate under the shape of fibres. This phenomenon is at the origin of numerous medical conditions such as Parkison's or Alzheimer's diseases. Therefore, it is crucial to understand what is the mechanism driving this kind of aggregation. Moreover, very diverse molecular mechanisms actually lead to fibre formation suggesting the possibility of a more general origin of the phenomenon.

Thus one recent physical theory proposes to explain the origin of fibre formation through general physics principles. This theory predicts in particular that fibres will form not only in disease-inducing proteins but for a large fraction of all proteins under mild aggregation conditions. However, we lack experimental data to test this prediction.

Acquiring such data is a long and expensive work. Indeed, the samples have to be prepared in very strict conditions with precise values of concentration of elements favoring protein aggregation like salt in order to yield the desired aggregates. Once the samples have been well-prepared, large structures such as synchrotrons are used to analyze them but they are expensive and screening dozen to hundreds of proteins in them is challenging notably regarding their availability.

However, there actually already exists a large amount of data of scattering experiments on proteins in solution made by crystallographers. If the latter are interested into the samples in which proteins have crystallized, they also store the scattering intensity of samples that have not crystallized but may have formed other types of aggregates. These scattering experiments interest us because it can actually help us trace back which type of aggregates may have formed. More particularly we want to know if fibres have formed compared to other types of aggregates such as bulks or sheets or even to the case where no aggregate has formed. This differentiation actually corresponds to the characteristic called the dimension of our aggregate which is the number of directions in which our aggregate is theoretically "not limited" in size. For instance, a fibre has a morphology particularly elongated in one direction - hence its dimension is of 1. As for the others, dispersed monomers in solution have a dimension equal to zero - the aggregate is finite in all directions -, sheets have a dimension equal to two and crystals have a dimension equal to three.

Yet, the issue is that crystallographers study the structure of the proteins on an atomic scale. Hence, crystallography data predominantly yields data on these small scales rather than on the characteristic scale of aggregates which are larger. Nevertheless, it might be possible that their experiments contain a little part of exploitable data. If other scattering experiments might then appear as more appropriate we will detail in the section 1 what still led us to try to use crystallography data.

Thus, the question we want to answer is : could we use these unused X-ray crystallography data to determine/characterize the dimension of experimental protein aggregates ?

First, we will detail in the section 2 the stakes of the subject - that is to obtain experimental data on protein aggregation to test a recent theory on fibre formation. Then, in the section 3, we will explain that to further test the exploitability of crystallography data, we numerically created a database of different kind of protein aggregates. Finally, we will describe in the section 4 how we simulated their X-ray scattering intensity and we will show that data mining methods of these simulated data yield promising results.

This internship has been realized at the Laboratory of Theoretical Physics and Statistical Models (LPTMS) - a joint research unit of CNRS and University Paris-Saclay - aiming at modeling, qualitatively understand, and quantitatively describe complex phenomena arising in diverse areas of physics and disciplines related to statistical physics.[2]. I worked under the supervision of Martin Lenz, a CNRS researcher managing the Soft biophysics group of the laboratory. The group is composed of eight people - mainly working on self-assembly. I particularly worked with Lara Koehler, a Ph.D. student working on the use of renormalization as a tool to better understand fibre formation.

2 Being able to determine the dimensionality of protein aggregates would help us test physical theories on frustrated auto-assembly

In the following, we will present in the subsection 2.1 a theoretical model of self-assembly developed to understand fibre formation. We will stress out the difficulty to relate their model which uses strong simplifications on the type of objects that assemble to real protein fibres, leading to the necessity of having experimental data to compare it with. Thus, in the subsection 2.2, we will explain that X-ray scattering experiments are the suited experiments to obtain these needed experimental data notably by explaining the theory behind such experiments. Then, in the subsection 2.3, we will compare two techniques of scattering - Small Angle X-ray Scattering (SAXS) and crystallography - and what led us to finally use X-ray crystallography data despite their not adequate use to study the shape of protein aggregates. Finally, in the subsection 2.4, we will show experimental results suggesting that we can differentiate fibres from a monomeric solution of given proteins from their X-ray crystallography scattering profiles.

2.1 Fibre formation is a result of geometrical frustration of self-assembled matter in simplified models

Self-assembly is defined as the spontaneous organization of multiple subunits into ordered multi-unit structures[3]. It is a ubiquitous process in living organisms. Protein fibres are a good example of self-assembled matter since it is the spontaneous organization of monomers under the shape of a well-defined structure that are fibres. Yet, what drives this kind of aggregation is not well determined yet in many cases of formation of protein fibres because of the complexity of the objects that are proteins.

We will present the model developed by Martin Lenz and Thomas A. Witten [6] that proposes to explain generically fibre formation with general physics principles by simplifying the shapes of the objects that assemble - the point being to demonstrate that fibre formation

can occur for any particle with an irregular shape under mild aggregation conditions. However, an experimental study is needed in order to relate this theory to the more general and complex subject of protein aggregation.

Theoretical models of protein aggregation are hardly generic because of the complex biological nature of proteins. Indeed, proteins are biological macromolecules present in every living cell. They are composed of molecular blocks called amino acids which are organic compounds with both a carboxyl group –COOH and an amine group. In the case of proteins, these amino acids form by their association into what is called polypeptide chains. Proteins are the arrangement of one polypeptide chain.[5] Thus, they have very complex shapes (see example below). Moreover, they also interact with each other with very various and specific chemical interactions. Hence, if investigations has been led in many cases of fibre formation and understood the molecular details of the specific fibre formation studied, it does not yield any generic result. However, could there be generic trends in the process of fibre formation that go beyond the specifities of individual systems ? To test this hypothesis, some simplifications need to be made in order to get a generic model.

Figure 2: 3D-shape of the actin protein - molecular actor of the muscle contraction - taken from the Protein Data Bank, PDB file=2hf4 [6]

Thus, a theory - published in the journal *Nature Physics* in 2017 by Martin Lenz and Thomas A. Witten[4] - suggests that fibres - rather than being a simple coincidental convergence of molecular mechanisms - would result from generic physical principles[6]. More particularly, they test the hypothesis that fibre formation arises from the competition between short-range interactions and the geometrical constraints yielded by such aggregation for particles with irregular shapes. This incompatibility of local interactions with global geometric constraints is called geometrical frustration[7].

This model - which we will from now on designate as the irregular hexagon model -is not restricted only to proteins and applies to any particle with a shape that cannot nicely fill the space by its aggregation without any deformation. To give an example, a square particle would have no problem to tile the plan by its aggregation but if we now consider irregular

hexagons, these cannot tile the plan by simply sticking together. They have to deform in order to aggregate. Thus, the attraction of such particles is hindered by this inevitable deformation which is energetically costly - hence the name of frustration. In protein aggregates, this frustration can have different origins such as deformed or partially denatured protein domains, the juxtaposition of residues with unfavourable interactions, or sterically hindered hydrogen bonding [7].

Hence they choose to model the situation by two-dimensional deformable hexagonal particles driven to aggregate by attractive interactions that have no range and whom magnitude is parametrized by a surface tension. They then study the morphologies yielded by these irregular hexagons depending on the value of the surface tension. The value of this surface tension directly affects the morphology of the aggregates. Indeed, with a low surface tension, particles are less likely to stick together because the geometrical constraints yielded would be much more costly energetically. Conversely, a high value of the surface tension leads to space-filling aggregates in which all particles are greatly deformed - the geometrical constraints being this time less costly than not being stuck together.

Thus, the authors demonstrate that fibres form at intermediate values of the surface tension, where the energetic costs of both phenomenons are comparable. An example of the resulting aggregates is presented in the Figure 2 in function of the surface tension.

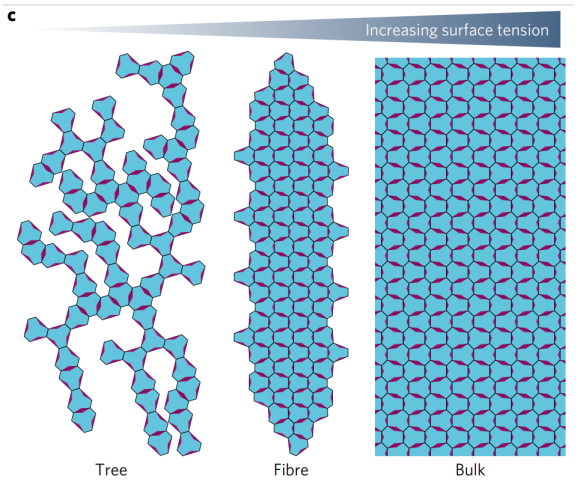

Figure 3: Figure taken from the article entitled "Geometrical frustration yields fibre formation in self-assembly" published in 2017 in the journal *Nature Physics* (vol. 13, pages 1100–1104) wrote by Martin Lenz and Thomas A. Witten[4]. It represents the resulting aggregates in function of the surface tension for irregular hexagons. We can see that fibres are a compromise between the two extremes which are a tree-like aggregate and a bulk.

To summarize the main points made by the authors in the article, the latter demonstrated in a minimal model that fibre formation upon aggregation is yielded by geometrical frustration - and therefore is a generic phenomenon. Moreover, for various shapes and aggregation

conditions with an intermediate value of surface tension they still obtained fibres, stressing out the robustness of the phenomenon.

These results could have beneficial consequences. Indeed, apart from the fact that understanding fibre formation might help us prevent certain diseases, this could also help us build highly complex artificial nano-structures by taking advantage of these efficient biological processes of self-assembly.

However, this irregular hexagons model is only in two dimensions and the shapes of the particles are still much more regular than the ones we can encounter in nature. If these simplifications suggests a generic nature of the phenomenon, to now test this prediction it is indispensable to compare these results to aggregates we can observe in nature. One way to proceed is to perform X-ray scattering experiments.

2.2 X-ray scattering experiments are well suited for the structural study of biological objects

We will study here why X-ray scattering experiments are a dedicated tool to study the structure of objects in solution - proteins in our case. X-ray radiations are scattered by the electrons of heavy objects, the scattering intensity producing then a pattern of the electron density of the objects considered. Then, this scattering intensity can be analysed to determine characteristics of the objects. However, this intensity depends a lot on the range of scattering vectors that are measured on the detector that receive the scattered light. Thus we will describe how the objects in the sample scatter the light and we will explain how information is retrieved from the scattered intensity measured in experiments.

To describe the system, we consider an incident plane wave characterized by the incident wave vector $\mathbf{k_i}$, scattered in the direction of observation with the scattered wave vector $\mathbf{k_s}$. We assume elastic diffusion, namely $|\mathbf{k_s}|=|\mathbf{k_i}|=k$. We define the scattering vector $\mathbf{q}$ as the difference between the scattered wave vector $\mathbf{k_s}$ and the incident wave vector $\mathbf{k_i}$:

$$\mathbf{q} = \mathbf{k_s} - \mathbf{k_i} \tag{1}$$

The situation is represented on the following diagram where 2θ is the scattering angle and $\lambda = \frac{2\pi}{k}$ is the wavelength of the incident radiation.

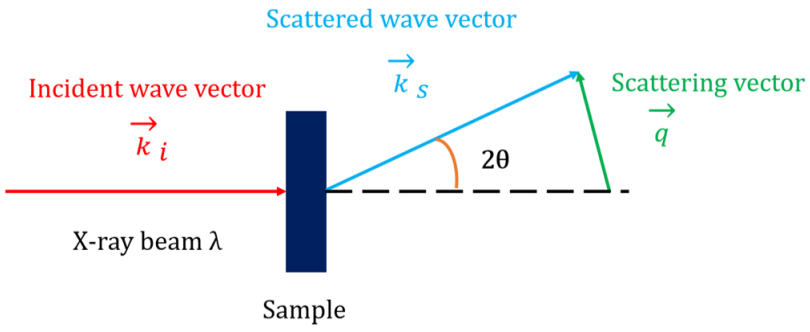

Figure 4: Diagram representing the incident wave vector, the scattered wave vector and the associated scattering vector of an X-ray scattering experiment

The magnitude of the scattering vector can then be easily deduced through :

$$\mathbf{q}^2 = {\mathbf{k_s}}^2 + {\mathbf{k_i}}^2 - 2\mathbf{k_s} \cdot \mathbf{k_i} \tag{2}$$

that is (because of elastic diffusion)

$$\mathbf{q}^2 = 2k^2[1 - \cos(2\theta)] \tag{3}$$

Using $1 = \cos(\theta)^2 + \sin(\theta)^2$ and $\cos(2\theta) = \cos(\theta)^2 - \sin(\theta)^2$, we get

$$\mathbf{q}^2 = 4k^2 \sin(\theta)^2 \tag{4}$$

Hence finally, since $k = \frac{2\pi}{\lambda}$, we have :

$$q = \frac{4\pi \sin\theta}{\lambda} \tag{5}$$

The waves scattered by two scattering objects - here electrons - separated from the position vector $\mathbf{r}$ are out of phase by $\mathbf{q} \cdot \mathbf{r}$. Hence, by generalizing to a continuous distribution of scattering objects described here by the electron density ρ_e, the magnitude of the scattered wave for the scattering vector A($\mathbf{q}$) is given by :

$$A(\mathbf{q}) = \int \exp(-i\mathbf{q}.\mathbf{r})\rho_e(\mathbf{r})d\mathbf{r}, \tag{6}$$

where the exponential factor translates the phase difference between the different scattering objects - that is $\rho_e(\mathbf{r})d\mathbf{r}$. A($\mathbf{q}$) appears as the Fourier transform of the electron density. Indeed, to describe the scattering phenomenon, we resort to the Fourier space of the real space, the space of vectors $\mathbf{k}$, called the reciprocal space.

Finally, the intensity of the scattered wave is the modulus to the square of the scattering amplitude averaged over all solvent and proteins degrees of freedom - that is an average over the solid angle of q such that $|\mathbf{q}| = q$ - and averaged in time over the duration of exposure which is much longer than the observation time. Hence, we get :

$$I(q) = \langle |A(\mathbf{q})|^2 \rangle \tag{7}$$

Now, we will explain how information is found in the scattered intensity measured in experiments. In practice, we have a sample that contains a solution of proteins. We send on it a beam of X-rays and all the electrons contained in the solution then scatter the X-ray radiation which is retrieved by a detector. This produces a two-dimensional pattern characterized by rings of various intensities. These rings actually result from the random orientations that the proteins in the solution may take - i.e. that the detected intensity at a given solid angle is independent of the rotation according to the axis of rotation of the beam light[8]. According to our previous explanation, the intensity of these rings are a consequence of the disposition of the proteins in the solution. Hence it can be used to obtain low-resolution structural information.

These 2D-pattern are then converted into one-dimensional intensity curves through an average for the scattering intensity on the azimuthal angle from the images retrieved by the detector - the aim being to obtain the scattering intensity of the sample $I(q)$ as a function of the magnitude of the scattering vector. We show below an example of a 2D-pattern obtained by X-ray scattering and of its conversion into a 1D-plot of the scattering intensity.

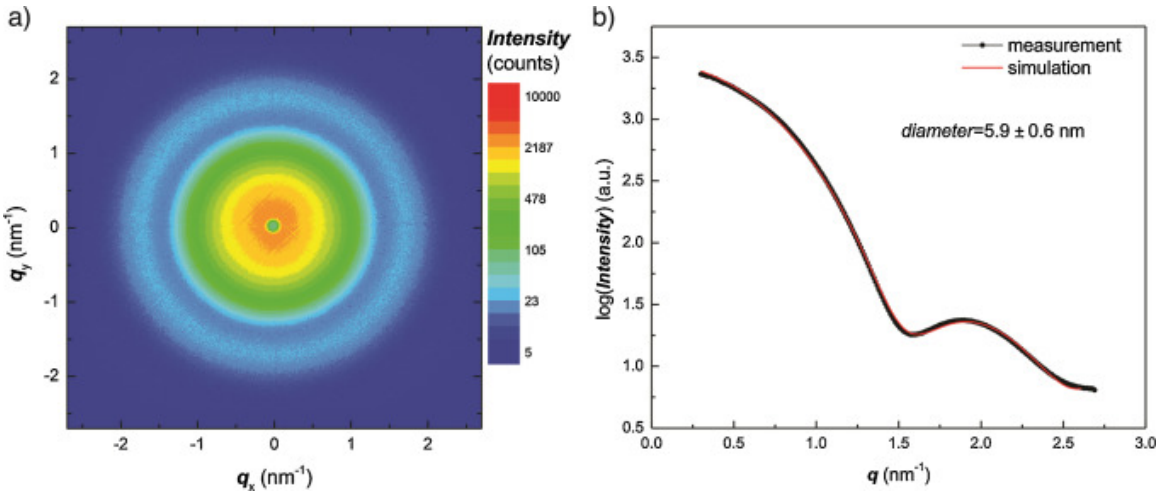

Figure 5: Example of the 2D-pattern with characteristic rings obtained by the detector for an X-ray scattering experiment and the curve obtained for its scattering intensity [9]

Then, by studying the curve obtained - and depending on the range of q we are studying - we can infer information on the structure formed by the proteins in the solution or directly on the structure of the proteins, notably the radius of gyration which is one of the characteristic size of the object in solution that we consider here. This radius of gyration can be obtain in the low-q region - which depends on the object considered - of the scattering curve by the Guinier plot which is tracing $\ln(I(q))$ in function of q^2[10]. Indeed, the theory gives that on this range of q :

$$\ln\left(I(q)\right) = \ln\left(I_0\right) - \frac{1}{3}q^2R_g^2 \tag{8}$$

where I_0 denotes the scattering intensity for $q \longrightarrow 0$ and R_g the radius of gyration. Hence R_g is given by the slope of the Guinier plot[10]. However, this works for ideal solutions in which the particles studied have no interactions. If this is not the case, then it is directly visible on the Guinier plot, notably if the particles attract each other as in an aggregate[11].

Another analysis to exploit the scattering data is for instance the Krakty analysis which consist in plotting $I(q)q^2$ in function of q[11]. The shape of the curves yielded gives information on the morphology of the protein, especially if the protein is unfolded or not - that is if the 3D-arrangement of its polypeptide chain has been "untied". For instance, the scattering intensity of a globular protein has a Gaussian-like shape at small q whereas in the case of an unfolded protein, the Kratky plot also present a plateau over a specific range of q which depends on the protein considered. We show on the figures 6 and 7 an example of these two typical cases.

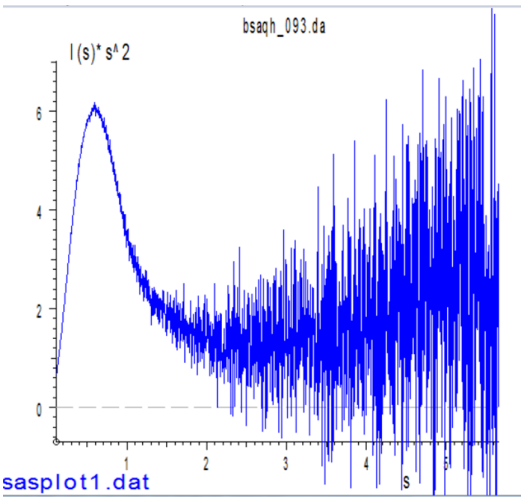

Figure 6: Krakty plot of the folded globular protein [11]

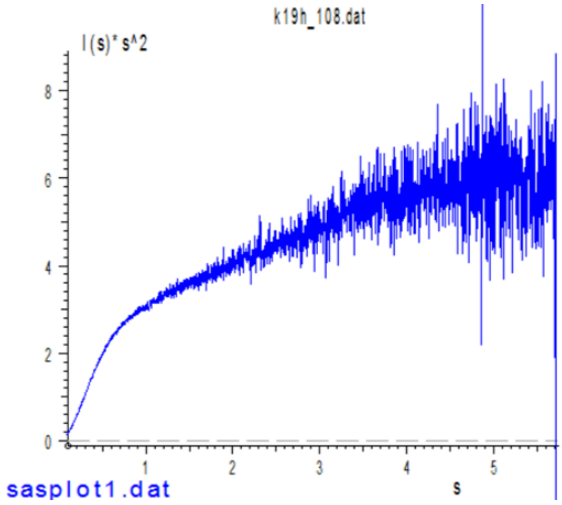

Figure 7: Krakty plot of a completely undolded protein [11]

Finally, another major analysis of these data is the Porod analysis which consist into plotting $q^4 I(q)$ in function of q[11]. The slope of the asymptotic behavior observed corresponds to a constant K depending of the surface of the scattering protein. Then, by calculating the quantity named the Porod invariant Q which is given by :

$$Q = \int_{q=0}^{\infty} q^2 I(q) \mathrm{d}q, \tag{9}$$

we can retrieve the molecular weight through the quantity $\frac{K}{Q}$ proportional to the surface over the volume of the protein.

The Guinier approximation is valid for small q values i.e. large interparticle distances whereas the Porod and Krakty analysis are valid at high q regime - the scattering occuring then from the internal structure[11]. Finally, several other analysis exist to infer structural information on proteins. Notably, in crystallography experiments, by studying the periodicity of pics obtained in the scattering curve we can trace back the 3D-shape of the protein.

Thus, scattering experiments are well suited to study biomolecules. Hence, any available scattering data on protein aggregation is of interest for us. However, the kind of information that we can retrieve from these experiments depends actually on the scattering angle.

2.3 SAXS and X-ray crystallography are two scattering techniques that fulfill different purposes and from which we do not have the same amount of data on protein aggregation

We will present two different X-ray scattering regimes - SAXS and crystallography - stressing out their differences and our interest in them. More particularly, we will justify our choice to focus on crystallography experiments even if these are not the most suited experiments for the study of aggregates dimension.

First, let us detail why the information we can retrieve from scattering profile actually depends on the scattering angle. The scattering vector $\mathbf{q}$ is a vector of the reciprocal space so - by the definition of the reciprocal space given previously - its magnitude is related to the real space by an inverse proportionality relation. More precisely, if we denote by the letter L the characteristic size of the object we want to study, we have :

$$q \propto \frac{2\pi}{L} \tag{10}$$

In our case, we study protein fibres. Their characteristic size - their diameter here - depends on the protein considered. For instance, an actin filament has a diameter around 7 nm [12] whereas microtubules - see description in section 2.4 - have a diameter around 25 nm[13]. Thus, although this size varies, its magnitude is usually of a few nanometers. If we take 2 nm as an example - which is already quite small for a fibre -then, by the previous formula, it means we will need the scattering intensity for q around $\frac{2\pi}{20} \simeq 0.3$ Å^{-1}. This correspond to a small scattering angle - hence the name of Small-Angle X-ray Scattering (SAXS) for such scattering experiments. This experimental method offers us structural information at length scales greater than about 10 Å- that means q on the range $[0;0.6]\text{Å}^{-1}$ - which is really adequate to estimate macromolecular shapes - exactly what we are looking for.

However, we do not possess actual SAXS data on protein aggregates - SAXS experiments being more employed to study the structure of one specific protein or protein complex. Yet, we would like to have a database as wide as possible since we want to be generic - that means a database with a lot of different proteins. Hence, if SAXS experiments would provide us the data we need, we do not have the time nor the money to realize such a broad set of experiments.

Nevertheless, structural biologists studying the X-ray crystallography of proteins actually posses a large amount of scattering data on protein aggregates coming from all their failed attempts to make proteins aggregate under the shape of a crystal. Indeed, to optimize their procedure to obtain crystals, they put their various solution of protein samples in plates presenting several wells - the number of wells depending on the type of plate - that are directly passed under X-ray by a robot. An example is showed in the Figure 8 below. This way they do not have to examine themselves in which well proteins have crystallized, but that also means that they collect X-ray scattering data for samples that have not crystallized.

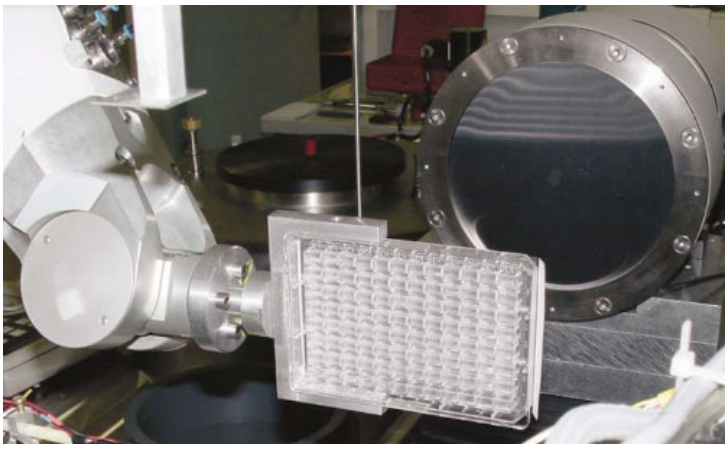

Figure 8: Photograph of a plate of crystallography being passed under X-ray by an automate[14]

Crystallography experiments differ from SAXS experiments by the fact that they study the atomic structure of the proteins. Hence by the inverse proportionality relation we introduced above, it means it gives us the scattering intensity at a larger q since it study smaller objects. This results into experimental constraints, especially the fact that crystallography experiments usually give q only on the range $[0.1;5.5]\text{Å}^{-1}$ which is higher than the range given by SAXS experiments. However, there still exists an overlap between both range of q of these scattering experiments. Hence, X-ray crystallography might give us exploitable data on this shared scale of structural resolution.

Thus, considering the large amount of already available X-ray scattering data from crystallography and the costly and long process to plan SAXS experiments, it is more advantageous for us to study these X-ray crystallography experiments and try to retrieve the dimension of our protein aggregates from it.

2.4 Preliminary experimental analysis suggest data on the shape of aggregates can be extracted from X-ray crystallography experiments

To preliminary test the possibility to retrieve the dimension of protein aggregates from crystallography experiments, Martin and Lara organized a series of experiments with the cooperation of the research teams directed by Monica Spano and William Shepard at synchrotron SOLEIL. The aim was to test specific samples containing either filaments or simple monomers where they knew beforehand which samples contained the filaments and which ones contained only dispersed monomers - the idea being to observe if their respective scattering intensities have significant differences. We will show that their results suggest that we can indeed differentiate a solution containing fibres from a monomeric solution thanks to their respective crystallography data. However, the differences observed remain difficult to interprete.

The samples tested contained either actin, α -synuclein, tau or tubulin proteins. Indeed, these proteins are known to form fibres :

- The actin protein has structural and dynamics functions[15]. In particular, it plays a role in the muscle contraction where it polymerizes under the shape of filaments.
- The α-synuclein protein is mainly present in brain cells. Although its function is not well-determined yet, its fibrillous form has clearly been identified as responsible of Parkinson's disease.[16]
- The tubulin proteins are the molecular blocks of microtubules - fibers constitutive of the cytoskeleton which hold an essential role in the cell division , motility and cellular transport [17].
- The tau protein has the role to stabilize the microtubules and promote their self-assembly from tubulin subunits. They can be found under an abnormal aggregated form which is one of the major hallmarks of Alzheimer's disease[18].

We show below the scattering intensities obtained with actin. Six types of sample have been collected - two containing actin monomers, two containing actin filaments and two containing a mix of actin filaments and fascin drug which makes bundles - in each case with different concentrations between both samples. On the left we show the whole scattering signal whereas on the right we show the scattering signal zoomed in on the range of q that interests us.

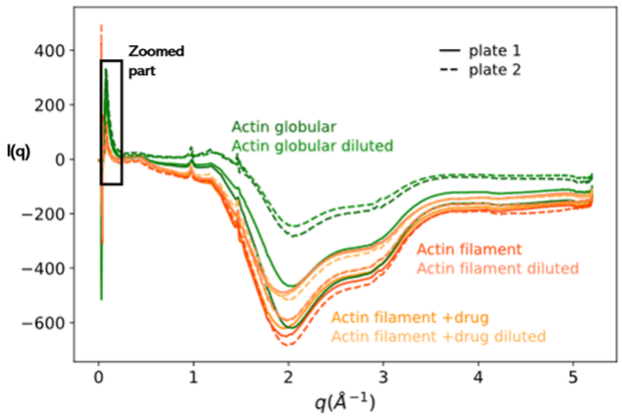

Figure 9: Scattering intensities of samples of actin with different concentrations and morphologies measured at the SOLEIL synchrotron on the range $q \in [0.03;5.5]\text{Å}^{-1}$.

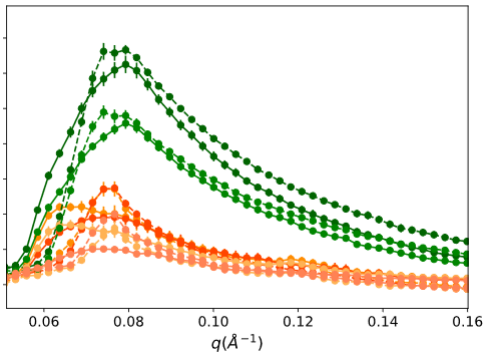

Figure 10: Scattering intensity of the samples of actin in function of the magnitude of the scattering vector zoomed in on the range $q \in [0.05;0.16]\text{Å}^{-1}$

The scattering profiles observed are actually the scattering profiles of the solutions to which the scattering profile of the buffer - corresponding to a sample containing only the solvent - has been subtracted. Indeed, the point is to only study the radiation scattered by the proteins and not the one scattered by the solvent.

As we can see on this example, even with different concentrations, the scattering profiles of the filaments are very distinguishable from the scattering profiles of the monomers. However,

it was not that clear for other proteins, notably with the tau protein. Yet our goal is to be as generic as possible - that is determining if the proteins have formed or not a fibre in the sample regardless which protein we are looking at. Thus, to compare all the data recovered from the different experiments, Lara performed a two components Principal Component Analysis. This method of analysis makes it possible to directly visualize the two main dimensions of variation of our data - which is especially practical when we possess numerous data.

We show below the results yielded by this analysis on the range $q \in [0.04;0.30] \text{Å}^{-1}$.

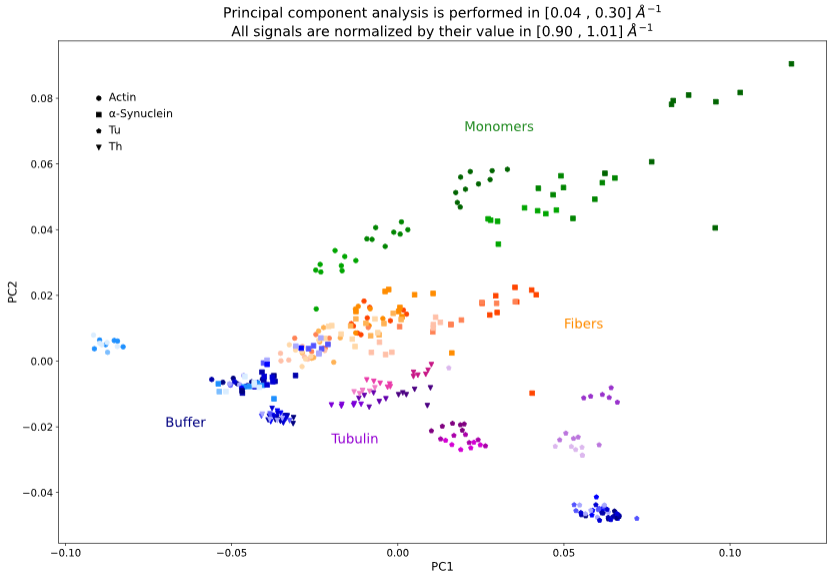

Figure 11: Principal Component Analysis realized from the SOLEIL experiments

As we can see, monomers and fibres are clearly distinguishable. However, the different kind of fibres are not similar enough to form a unique and well-determined cluster. Nevertheless, these results suggest that it could actually be possible to determine if we are dealing with a fibre or not from X-ray crystallography experiments.

However, to achieve such a goal requires a much more exhaustive study. This is the reason that led us to create our own database of fictitious protein aggregates from which we simulated the scattering intensity associated.

3 We numerically create a database of different kind of aggregates

We want to create a numerical database of scattering profiles from various protein aggregates. Yet, as we explained in the section 2.3, this requires to have the electron density map of our aggregates. Fortunately for us, there actually exists an online database of files - named PDB files - containing structural information on numerous protein monomers, notably their atomic

coordinates from which the electron density map can be easily retrieved. Hence we decided to build our aggregates from these files. For that we used the software ChimeraX that enables us to modify these PDB files and more interestingly to arrange them such as to form well-defined morphologies.

Thus, in the subsection 3.1, we will explain which proteins we chose as building blocks of our aggregates and the basic contents of PDB files. In the subsection 3.2, we will describe how we can actually arrange the proteins used with ChimeraX and how we hence constructed aggregates characteristic of each dimension that is random distributions (of dimension 0), fibres (of dimension 1), sheets (of dimension 2) and finally crystals (of dimension 3). Then, in the subsection 3.3, we will justify the fact that we do not take into account the local contact between proteins as building our aggregates. Furthermore, We will show in the subsection 3.4 that we can shuffle the orientations of the monomers within our aggregates, still in the aim to only retrieve the characteristic dimension of our aggregates. Finally, in the subsection 3.5, we will sum-up the content of our database.

3.1 As building blocks of our aggregates, we used proteins of different shapes and sizes from the Protein Data Bank

We built our aggregates starting from protein monomers as building blocks. In order to be as generic as possible we needed to create a lot of different shapes of aggregates and this from various proteins with different shapes and sizes. However, the point being in the end to simulate the scattering intensity of these aggregates - which can take a lot of computational time - we restrained ourselves to rather small proteins.

Thus, we chose to focus on the actin, lysozyme, α-synuclein and the insulin proteins :

- The actin protein has already been introduced in the section 2.4. It is composed of 374 amino acids and its characteristic size is around 6 nm. It is a globular protein - that is a protein with a rather spherical shape.
- The lysozyme protein is involved in the defense against bacterial infections. It is present in numerous species[6]. It is composed of 129 amino acids and it is also a globular protein. Its characteristic size is around 40 Å.
- The α-synuclein has also already been introduced in the section 2.4. It is composed of 140 amino acids. It has an elongated shape whom characteristic length is around 14 nm.
- The insulin is an hormone which plays an essential role in the carbohydrate metabolism[6]. It is composed of only 51 amino acids and its characteristic size is around 15 Å.

The structural information of these proteins is actually available at the Protein Data Bank website which is an international database for the 3D shapes of proteins and nucleic acids[5]. These data are usually obtained from several experimental methods such as X-ray crystallography but also NMR spectroscopy and cryo-electron miscroscopy and are thereby accessible to everyone. This information is saved under the form of Protein Data Bank (PDB) files which can be opened by several software to visualize and manipulate the protein or even perform simulations of scattering experiments such as SAXS.

We chose to visualize the proteins with the software ChimeraX[19]. We show below a picture of the 3D-structural arrangement visualized with ChimeraX for all the proteins introduced before.

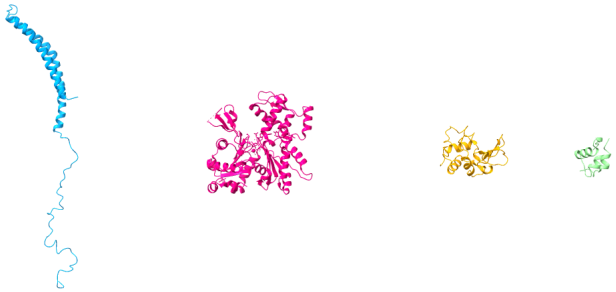

Figure 12: From left to right : (1) α-synuclein, PDB file=1XQ8 ; (2) Actin, PDB file=2HF4; (3) Lysozyme, PDB file=6LYZ ; (4) Insulin, PDB file=2C8R

To sum-up, it is possible to numerically retrieve the experimentally determined 3D-shape of proteins, visualize them and even manipulate them. More particularly, ChimeraX enables us to open several times the same file of one protein and then move and turn the proteins relatively to each other. Therefore, because of this ability, we have been able to build different kinds of aggregates.

3.2 Iteratively displacing and rotating proteins in ChimeraX enables to build any shape of aggregate

Using PDB files through the software ChimeraX, we can arrange the monomers as we want and even automate the procedure using Python scripts. Thus, we have been able to produce various shapes of protein aggregates.

The software ChimeraX possesses many specific commands that we can enter into a command bar, in particular a move and a turn command that enable us to move and rotate a PDB File that has been opened with the command open. Each opened PDB file is called a model and is numbered regarding to the order of opening. Therefore, by using their number and the key words " models number" we can decide which one we want to move and rotate.

Initially, all the opened files are opened at the same position so they are superimposed. To use the command move, we need to specify of how many frames - a frame corresponding to one Angstrom - we want to move the model in each of the directions which are by default in cartesian coordinates. However, the origin of the coordinate system is not well defined. Hence, to be more precise we can indicate which model we want to take as a reference for the translation. For instance, the following command translate the protein number 2 from

a distance of 20 Åin the x-direction from the protein number 1 : " move 20,0,0 models 2 coordinateSystem 1"

The turn command works similarly. We indicate the axis of rotation in the coordinate system and then the angle of rotation as well as the model we want to rotate. We can also specify the center of rotation we want to use, notably because this one is not well-defined by the software like the coordinate system used which can reveal critical for building some shapes of aggregates as we explain later. Here is an example of a command line that rotate the protein number 2 of an angle of 30° around the y-axis using its cartesian coordinates as the center of rotation in the coordinate system given by the position of the protein number 1 : "turn 0,1,0 30 models 2 center 20,0,0 coordinateSystem 1".

Finally, all these commands can be written in a command file from a Python program that can afterwards be read by ChimeraX which we directly launch from the Python program thanks to the module subprocess. Moreover, ChimeraX enables us to save the aggregate formed in the end as a PDB file.

Thus, we can automate the construction of different aggregates such as the ones introduced at the beginning of this section - that is random distributions, fibres, sheets and crystals - all characteristic of a different dimension. We will further detail how we built these aggregates.

- **Random distributions**

  A random distribution is characteristic of the fact that no aggregation occurred in the solution.

  To build random distributions, we randomly pick a value in a volume chosen in input for the coordinates x,y,z given afterwards to the move command. Hence, depending of the volume given, the proteins can be well dispersed or condensed. We also randomly choose an axis and angle of rotation given to the turn command so that the proteins are not all in the same direction. All these values are picked using a uniform distribution.

  Finally we also created random distributions of dimers to see if we were still able to associate these with the zero-dimension. To do so, we just modified a little the algorithm used in the case of monomers. We actually added a function that previously creates dimers then saved them under new "homemade" PDB files. Afterwards what we applied our first algorithm but at the difference that this time we opened these new PDB files instead of the PDB file corresponding to the sole protein. To create the dimers, we simply used the same algorithm as the one to create fibres with random configurations which is described in the subsection 3.4.

  We show below an example of a random distribution of monomers (on the left) and dimers (on the right) with the lysozyme protein.

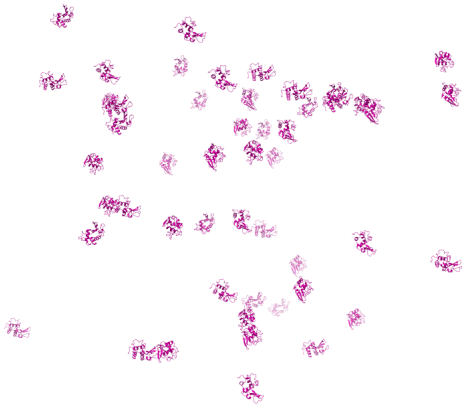

Figure 13: Random distribution of 50 lysozyme proteins from the Protein Data Bank built with ChimeraX

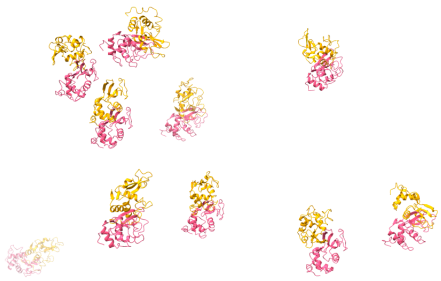

Figure 14: Random distribution of 10 lysozyme dimers built with ChimeraX

- **Fibres**

  Fibres hold a main role in our procedure since it is this kind of aggregates that we are the more interested in. Therefore, to well represent them, we decided to create four different shapes of fibres : straight fibres, twisted fibres, bent fibres and helical fibres.

  We define in each case a function that write in a command file the commands to execute using a loop on the number of proteins that are going to constitute the fibre. We give in input the distance we want to have between monomers and we write the move command adequately to the position of the model in the fibre. For instance, if we move the second protein to a distance of 20 Åin the z-direction to the first one, then we will need to move the third one of 40 Åin the same direction from the first one - using the protein number 1 as reference for the coordinate system. This is due to the fact that each time we open a new PDB file, it places the model at the same position we opened the first one. Another solution requires the use of the command "view position sameAs" which enables us to place the model we are considering to the same position than another model of our choice. Once the translation has been performed on the model considered and depending on the kind of fibre we want to form, we turn our model using an angle and axis given in input.

  We detail in the following the specificities of each shape.

  - Straight fibres : This is the simplest type of fibre where each monomer is simply translated from a given distance - which can be constant or not - relatively to the previous monomer.
  - Twisted fibres : There are basically straight fibres from which we just added a rotation of a given angle around a fixed axis of rotation for each monomer. However,

for that to work, it is necessary that the axis of translation be the same as the axis of rotation - these axis being the same for all the proteins.

- Bent fibres : Conversely to the twisted fibres, here the axis of rotation must be orthogonal to the axis of translation. Thus, it can be described as a straight fiber that has been bent since we do not modify the rotation on the axis of translation.
- Helical fibres : To create helical fibres, the axis of rotation must not be orthogonal to the axis of translation. Apart from this difference, the code is exactly the same as a bent fibre.

We show below an example of these four types of fibres using the lysozyme protein as building block.

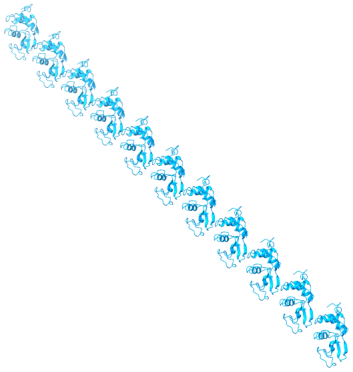

Figure 15: Straight fibre built with ChimeraX and composed of 10 lysozyme proteins from the Protein Data Bank

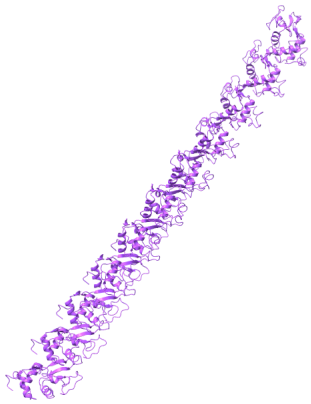

Figure 16: Twisted fibre built with ChimeraX and composed of 10 lysozyme proteins from the Protein Data Bank

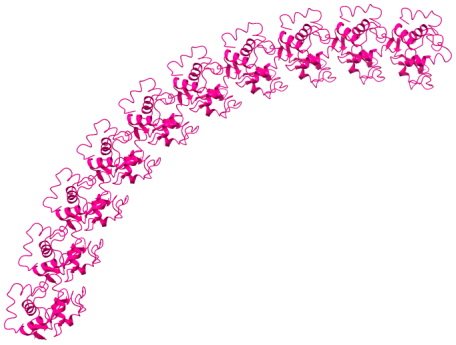

Figure 17: Bent fibre built with ChimeraX and composed of 10 lysozyme proteins from the Protein Data Bank

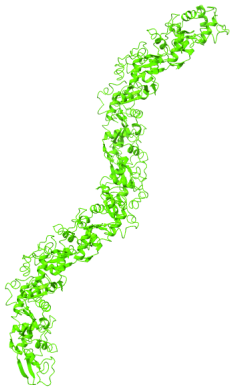

Figure 18: Helical fibre built with ChimeraX and composed of 10 lysozyme proteins from the Protein Data Bank

- **Sheets**

  We have simply defined sheets as the assembly of several fibres. Therefore, once we have created the fibres and save them under PDB files, we can then open them in ChimeraX and move them from one another exactly in the same way as with monomers. Thus, we have created both regular sheets by the association of straight fibres as well as irregular sheets by the association of fibres with random configurations (see section 3.4).

  We show below an example of a regular sheet constructed by this process with the lysozyme protein.

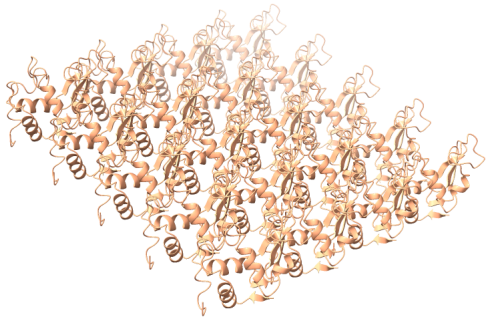

Figure 19: Sheet built with ChimeraX and composed of 20 lysozyme proteins from the Protein Data Bank

- **Crystals**

  We defined crystals as the association of regular sheets saved under PDB files that we opened in ChimeraX and translated of a given distance such that it forms a crystal. We similarly built "irregular crystals" by the association of irregular sheets. However, we only created crystals with parallelepiped shapes.

  We show below an example with the lysozyme protein.

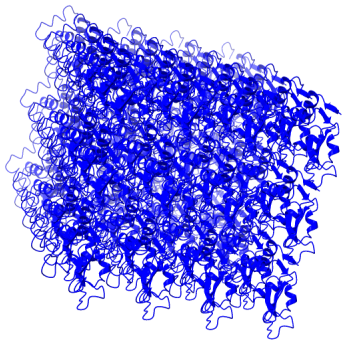

Figure 20: Crystal built with ChimeraX and composed of 48 lysozyme proteins from the Protein Data Bank

Thus, we are able to create various shapes of aggregates using known 3D-shapes of proteins. These aggregates we created are fictitious and do not represent real aggregates. Nevertheless, as long as we are only trying to characterize them by their dimension, having the shape desired is all we need. However, we have neglected so far the way our proteins interact with each other.

3.3 We choose not to take into account the short-range interactions between our monomers as building our aggregates

So far, we simply described how we were able to build our aggregates. However, we never justified the fact that we can just move and rotate arbitrarily one protein relatively to another without taking into account the actual interactions between these proteins to finally create an aggregate. Therefore, it is important to test whether or not we should create our aggregates according to the match between the electron density maps of the monomers as we aggregate them. We will show that, even if our building process is an artificial stacking of proteins, to neglect their interactions is a reasonable hypothesis notably thanks to the broad set of parameters encompassed in our generation of aggregates and to the range of q we are considering.

There actually exists a command in ChimeraX - named fitmap - that allows us to perform rigid-body local optimization to fit an atomic model into the electron density map of another model. The electron density map of the protein to which we want to dock a second protein is obtained thanks to the molmap command. Then it suffices to apply the fitmap command to the second model and the software adjust its position relatively to its distance to the first monomer using the map defined previously by molmap. We show below an example of this operation with actins. On the left we can see how we moved and turned the second protein from the first one. Then we applied the fitmap command as described and we show the result on the right.

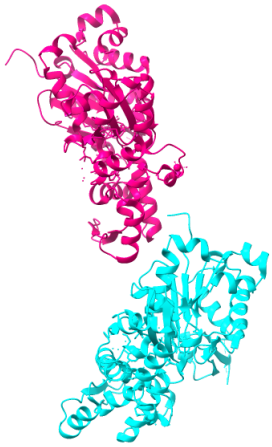

Figure 21: Picture taken before the use of fitmap between two monomers of actin arbitrarily arranged. The pink protein is the protein number 1.

Figure 22: Picture taken after the use of fitmap between the same two monomers.

As we can see the difference between the two pictures is not obvious. Indeed, the blue protein which corresponds to the second protein has only been slightly rotated. In many cases the change induced by fitmap is unnoticeable which already suggest that it will not yield significantly different scattering signals.

The fit obtained between the two monomers can then be retrieved in a position file containing the translation vector and the rotation matrix of the model considered. From the rotation matrix we can get back the axis and angle of rotation using the mathematical properties of 3x3 rotation matrix. Thus, we now know which position must have our protein relatively to the previous one in the fibre to optimize the fitting of their two electron density maps and we just have to use these new data in input of our usual building process. For instance, if we consider a straight fiber, since we want to propagate exactly the same relative positions between monomers we only have to retrieve the translation vector and rotation matrix obtained between two monomers and given by fitmap. Once this is done, we use the function that enables us to create a straight fiber at the difference that this time we are not using an arbitrary translation vector neither arbitrary axis and angle of rotation.

When such an aggregate has been built, we can compare its scattering intensity to the scattering intensity of the same aggregate built without using fitmap on its monomers. The simulation that gives us the scattering intensity is detailed in the next section, at the subsection 4.1. We show below an example of the scattering intensities we obtained for straight fibres in both cases with the same parameters - i.e. the same initial distance between the monomers as well as the number of proteins.

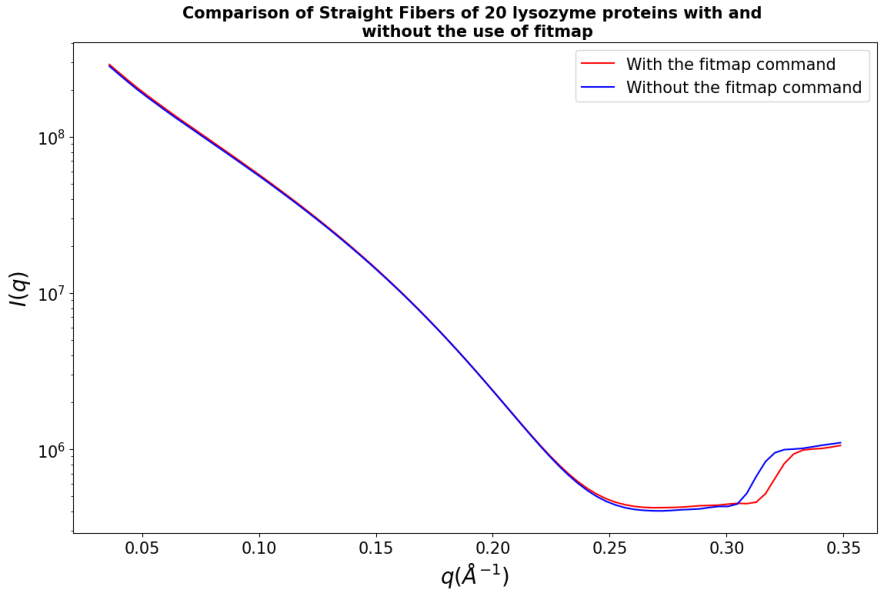

Figure 23: Example of the influence of fitmap on the scattering intensity of a straight fiber built from lysozyme proteins

As we can see, differences appears only at values of q higher than 0.30 $Å^{-1}$ - which means we are already considering differences only distinguishable under 2 nm of distance. This is consistent with our precedent remark on the two pictures showed. Moreover, these differences do not show a strictly different behaviour between the two profiles. In the end, the notable differences between the fibres built both with and without fitmap seem to only result of a change in the angle of rotation between the consecutive monomers which does induce changes in the scattering profiles. Indeed, the fitmap command mainly rotate the protein added given its chosen distance from the previous protein. To confirm this point, we performed several tests to study the influence of the angle of rotation between monomers on the scattering signal without the use of fitmap. We show below one example yielded in the case of twisted fibres composed of 20 lysozymes. All the parameters given in input are the same except for the angle of rotation that we chose equal to 0,30,60 or 90 degrees.

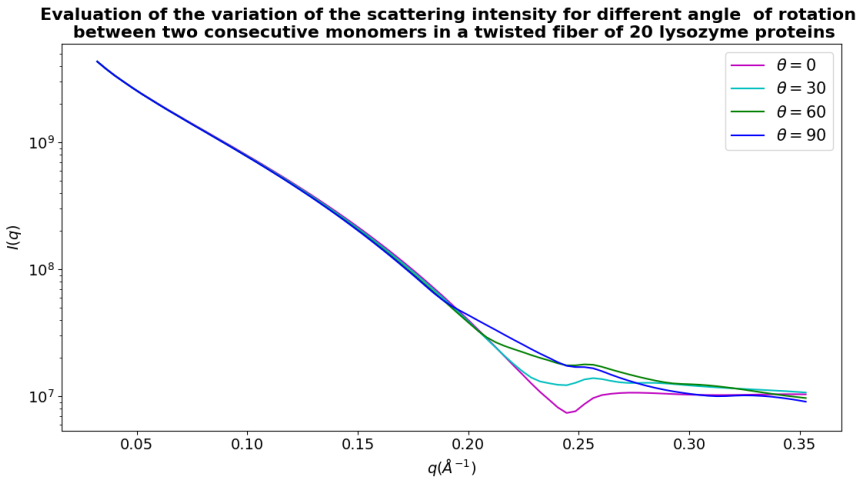

Figure 24: Variation of the angle of rotation between consecutive monomers for a same fibre

As with the previous diagram where we had used fitmap, a gap appears between the curves at $q \geq 20\text{Å}^{-1}$ and then disappear at $q \simeq 0.35\text{Å}^{-1}$ - the curves coinciding again. Our point was to show here that the differences induced by fitmap and showed previously will simply be hidden by the differences induced by a change of parameters. Hence, we chose not to optimize the local contact between monomers since it will not change the exploitable and interesting data contained in our scattering profiles.

Thus, we justified that we could arbitrarily arrange our monomers within our aggregates without any lost of important information notably because of the low range of q we are studying and of the differences already yielded by variations in our set of parameters when building our aggregates. We even decided to go further to stress out the general features associated to the scattering profiles of each category of aggregate characteristic of a different dimension. Indeed, the aim being to distinguish different shapes, we built aggregates where we shuffled the orientation of the monomers within the aggregates.

3.4 To find the general characteristics of our future simulated scattering profiles, we shuffle the orientation of the monomers within our aggregates

Since the aggregates we have created are artificial, we want to make sure that the information we retrieve is indeed dimension-dependent and not a consequence of our building process. Therefore, we decided to also include structures where we would have shuffled the orientation of the monomers within our aggregates. To do so, we had to proceed in two steps.

First, we need to create the regular aggregate as presented before and save the files of the position of each model. Then, we retrieve from these the translation vector for each monomer in the final fibre.

Once this is done, we define another function that is going to create a new fibre - our final aggregate with monomers whose orientations are shuffled . For that, for each new opened model, we first turn it using a randomly picked angle of rotation and axis of rotation. The angle of rotation is uniformly picked between 0 and 360°. As for the axis of rotation that is defined by three coordinates x,y,z, we uniformly pick each one of them between 0 and 1. It is only afterwards this operation that we move our model using the saved translation vector associated from the first fibre. Thus, we conserve the general shape of the aggregate but the monomers have all been randomly rotated between one another. We show an example of such an aggregate below.

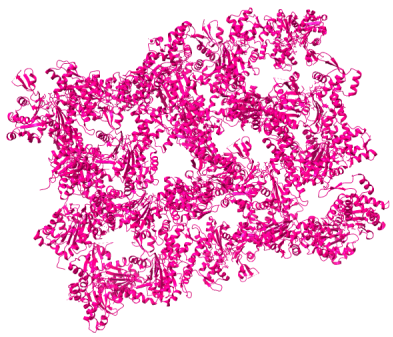

Figure 25: Irregular sheet built from straight fibres of actin monomers whose orientations have been shuffled

We have created such aggregates for each category - random distributions, fibres, sheets and crystals. We chose not to differentiate them from the other aggregates of their categories in the following, the point being to be as generic as possible. We summarize in the next subsection the contents of our database as well as the parameters that have been varied through the different shapes of aggregates.

3.5 We have a database of 282 aggregates of very various shapes and sizes

To exploit the PDB files yielded by our building process, we created a directory under which we saved all our aggregates using a unified nomenclature :

(name_structure+precision)_nb_prot_(number of proteins)_prot_(name of the protein)_.pdb

The precision indicates if we have a regular aggregate or not. Thus, if we have a regular aggregate, the precision is a certain number of symbol "+" added, this number depending on the number of previous files having the same name. On the other hand, if we have an aggregate where the orientations of the monomers have been shuffled, the precision is simply a number corresponding to the order of creation of the file still depending on the number of previous files with the same name. The name of the structure is the initials of the structure created -

for instance "RD" for random distribution, "HF" for helical fiber etc. This organization will help us in our analysis as we will see in the section 4 but also to have a traceability of our aggregates when performing our analysis in case we would obtain unreasonable results.

We now have 282 files : 55 random distributions, 177 fibres, 28 sheets and 22 crystals. This inequitable distribution comes from the fact that we first mainly focused on fibres but also from the fact that crystals and sheets takes more time to create and more computational time for their scattering simulation. Ideally, we should continue to create aggregates until having an equal distribution of files in these 4 categories.

However, if the inequitable distribution is a problem, we still have created in each category very various aggregates on which we control a lot of different parameters. We recapitulate in the following which parameters have been varied and on which range it has been varied:

- the number of proteins : this parameter largely vary from the category of aggregate considered, notably because sheets and crystals require a lot more proteins than for a fibre if we want to have an aggregate extended in several dimensions. Thus, for random distributions and fibres, this number was kept under 50 proteins - mainly between 10 and 20 - whereas for sheets and crystals we can have aggregates up to 360 proteins.

- the distance between monomers : Usually, the gap between monomers has been varied between 0 - the proteins are interlocked - and 0.2 times the size of the protein considered. Indeed, we roughly estimated that above such a value the gap between the monomers was too important to consider the aggregate as an effective aggregate.

- the angle and axis of rotation : this parameter is particularly relevant for helical and bent fibres for which, even in the case where the orientations of the monomers have been shuffled, these two parameters still are at the foundation of the general shape of the final aggregate. The angle of rotation used for bent fibres was restrained to rather small angles i.e. under 90° and mainly between 5 and 20°, otherwise the proteins formed a circle which was not our purpose. As to helical fibres, by varying the angle and axis of rotation, we yielded helical fibres with a pitch roughly varying between 1 and 7 times the size of the protein considered. As for the radius, it roughly varies between 0.2 and 2.5 times its size.

Hence, we were able to built a database of very diversified aggregates and if it does need to be complemented with more files, notably for the sheets and the crystals, it can easily be done from our building process - the limiting operation being the simulation of these PDB files. Regarding the number of aggregates from each different protein, we have 38% of aggregates of lysozymes, 14% of aggregates of actin, 23.5% of aggregates of alpha-synuclein and finally 24.5% of aggregates of insulin. The low number of aggregates of actin comes from the fact that the actin was the most massive protein - hence its computational time was much longer and we chose to privilege the smallest proteins. The lysozyme is also more represented because it is a protein model.

Therefore, we have written functions in Python scripts that enable us to create different types of aggregate from any PDB file by writing a command file readable by ChimeraX. Each aggregate created is then saved as a new PDB File. Especially, for each created aggregate we control the number of proteins, the distance between the monomers, their disposition and rotation and therefore the shape of the aggregate. Now, in order to test if we are able to retrieve

the dimension of these numerically built aggregates from X-ray scattering experiments, we need to simulate their scattering intensity and analyse them through appropriate analytical methods.

4 We use data mining on the simulations of the scattering profiles of our numerically built aggregates to retrieve their dimension

In the following, we will present in the subsection 4.1 the software used to yield the scattering profiles of our artificial aggregates and the parameters we took in input. Then, in order to differentiate the different scattering profiles according to the dimension of the aggregate they are associated with, we chose to use data mining analytical methods. Thus we will show in the subsection 4.2 that the results yielded by a two principal components analysis do not enable us to differentiate fibres from monomers - although this technique had given interesting results from the experimental data formerly presented. Finally in the subsection 4.3 we will present the use of machine learning to predict the dimension of our aggregates which yielded promising results although further work is needed in order to confirm these results.

4.1 We simulate the scattering intensities of our numerically built aggregates with the software CRYSOL

Now that we have PDB files containing information on the atomic coordinates of the proteins constituting our aggregates, we need to use them to infer their scattering profiles. There exists plenty of adequate software that enable us to compute such scattering profiles from PDB files and which have long been proved to yield simulations in agreement with the experiments - hence our choice to use an already existing one. We will present here the software chosen. More particularly we will explain the parameters we chose for the simulations.

CRYSOL is a program evaluating the scattering of a solution composed of macromolecules with known atomic structure that is widely used to simulate SAXS data [20,21,8]. It especially takes a PDB file as an input and reads the atomic coordinates of the structure.

To perform the simulations, several parameters can be chosen :

- the number of harmonics : this number defines the resolution of the calculated curve. Indeed, to compute the scattering intensity of a given PDB file, CRYSOL uses a mathematical representation of the protein where the protein is expanded in terms of an infinite basis of spherical harmonics. The orthogonality properties of the basis functions simplify the averaging of the harmonic series from which an overall scattering intensity can be computed[22]. These basis functions are built from spherical Bessel functions, and normalized spherical harmonics of degree m and order L.(See [20] for more details). In our case, since we consider rather large protein aggregates, the software warned us to use the maximum of harmonics available - that is 99 harmonics.

- the range of q : we estimated that limiting our study to a range of scattering vectors in [0.03,0.35] $Å^1$ was enough to make out the shapes of the aggregates. Indeed, using the inverse proportionality relation introduced in the first section, $q = 0.03Å^1$ corresponds

to an object of characteristic size L around 21 nm and $q = 0.35\text{Å}^1$ corresponds to $L \simeq$ 2 nm. We purposely chose to not go below q=0.03 Å^1 since the X-ray crystallography data we aim to use as an input to our procedure do not have the scattering intensity below this value of q.

- the number of points : this corresponds to the number of intensities $I(q)$ that the software will compute on the range of q chosen. The maximum number of points available actually depends on the other chosen parameters although this maximum can never go above 5000 points. In order to minimize the computational time, we chose to arbitrarily fix this number to 2500 points. Indeed, regarding the previous parameters chosen, the maximum of points available was 4000 points. However, we visually compared scattering profiles of a same PDB file for different values of the number of points and estimated that we would not gain much to go above 2500 points, notably concerning the computational time. Moreover, to gain some additional computational time, we chose to reduce this number to 800 points and then use a linear extrapolation function to retrieve the same total number of calculated intensities on the same range of q. This choice was driven by a test on the output curves given with different number of points which stressed the fact that even with fewer points, the curves were all indistinguishable from one another.

CRYSOL then returns in output two files containing the scattering intensities of our PDB file. We show below an example of the different curves contained in these files on the range of q chosen.

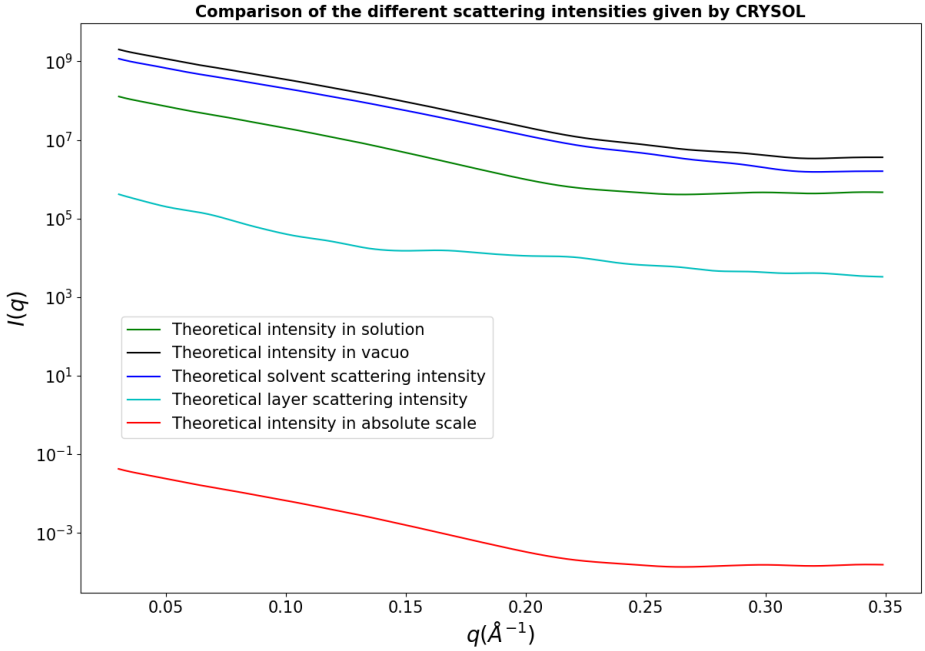

Figure 26: Comparison of the different scattering intensity given by CRYSOL on the range $q \in [0.03, 0.35]\text{Å}^1$ for a PDB File coding a bent fiber of 10 lysozyme proteins (see Figure 17)

Thus, we first have the theoretical intensity in solution which is the theoretical intensity that the proteins in solution should yield considering the electron density map given in input. This theoretical intensity in solution actually takes into account several parameters, notably the theoretical scattering of the solvent surrounding the proteins as well as the theoretical border layer scattering - that is the scattering intensity of the thin layer of solvent that directly surrounds the proteins in solution. These scatterings must be subtracted to the theoretical scattering in vacuo to yield the theoretical scattering intensity in solution - adding also the factor of absorption estimated by the software on the scattering intensity - the aim being that this is taken in consideration in experimental scattering intensities.

Therefore, the main difference observed on the vertical axis of the graph is explained by the definitions on the various scattering intensities presented : in vacuo there is no absorption considered hence the high value, which is practically the same for the solvent scattering. As for the border layer scattering, it is a thin layer so it does not contain a lot of molecules and the scattering intensity is proportional to the number of molecules. Finally, the scattering in solution has a middle value given by its construction detailed above. As for the theoretical scattering intensity, this one takes into account the fact that a proportion of photons are not scattered by the sample due to the absorption of the material[24]. Hence, it corresponds to the actual intensity scattered by the sample compared to the incident intensity. Then the number of photons scattered in a given direction have to be normalised with respect to the number of photons transmitted through the sample to get the absolute intensity from the 2D-pattern. This intensity in absolute units (cm^{-1}) is actually required to deduce several interesting information on the proteins studied from SAXS experiments [16][17]. Moreover, the '.abs' file returned by crysol also takes into account the concentration of proteins used in the sample. Hence the very low value on the axis compared to the other scattering intensities.

Taking into account all these informations, the theoretical intensity in solution and the theoretical intensity in absolute scale are the ones that interest us since they are the ones corresponding to the scattering intensities of our aggregates - and measured experimentally-, the intensity in absolute scale being just a specific normalization of the intensity in solution depending on the number of transmitted photons. Thus, we arbitrarily chose to focus on the theoretical intensities in absolute scale in the following. However, to be able to compare the scattering intensities of different proteins which might have different order of magnitudes since it depends on the electron density, we chose to always normalize the intensity vectors given by CRYSOL such as their norm is equal to 1 in our analyses.

Now that we have obtained the simulated scattering intensity of our different aggregates, we need to find criteria to differentiate them according to the dimension of the aggregates considered. None of the analysis methods presented in section 2.2 allow us to characterize the dimension of aggregates, notably because of the lack of data on the range where these techniques are efficient but also because of the complexity if the objects we consider here. Thus, if we first tried to visually analyze the data obtained by performing several experiments - notably by studying how our scattering profiles changed regarding one parameter that we varied - nothing obvious came out. Hence, we decided to move towards methods allowing us to find common characteristics between the different scattering profiles of our database according to the size of the aggregates associated with these profiles by comparing our data on a deeper level. Therefore, we chose to use data mining - a method transforming data into useful information by establishing relationships between the data or by spotting patterns.

4.2 A principal components analysis is not enough to differentiate aggregates with different dimensions

The first analytical method we employed was a principal component analysis (PCA). This choice was mainly made because of its successful application to the experimental data introduced in the section 2.4. Thus, we will show here the results obtained by principal component analysis with two and four principal components from which we will conclude that either this analytical method is not discriminating enough or we have an issue regarding our reasoning or our data.

A principal component analysis is very useful in cases of large dataset. Indeed, extracting the important information from large datasets can be made difficult by the mass of information. Hence, a PCA aims to reduce the dimensionality of large datasets by transforming the previous large set of variables into a smaller one still containing the important information. Thus, a two components PCA means that we have reduced our set of 2269 variables - each value of $I(q)$ for $q \in [0.03; 0.35]\mathring{A}^{-1}$ - to a set of only two variables which correspond to the directions in which there is the most variance of the data.

As the experimental data presented in the section 2 were analysed by a two component analysis, we tried to do the same with our simulated scattering data. We show the results below. We added to these projections the plots of the two principal components in function of q. These components are the linear combination of q for which we observe the most variance. We also indicate on these plots the percentage of the total variance explained by each principal component.

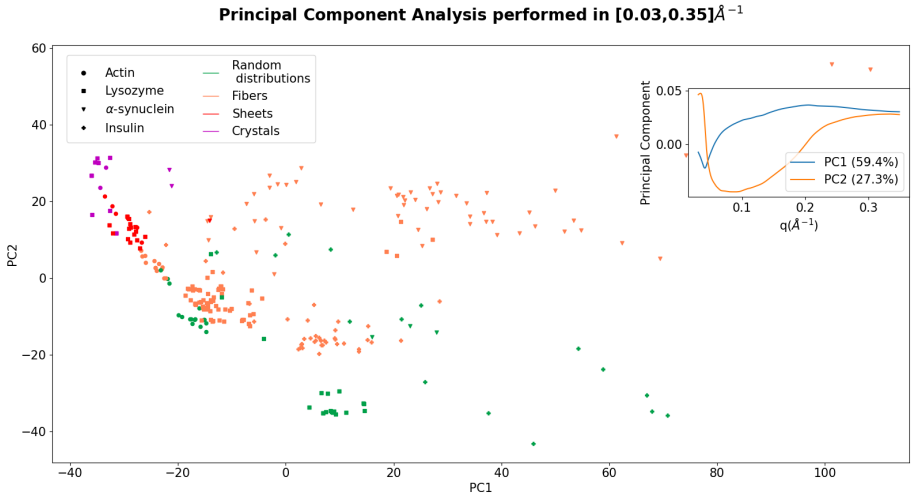

Figure 27: 2 principal component analysis performed in q$\in [0.03; 0.35]\text{Å}^{-1}$ on the theoretical intensities of our database of aggregates

We can already notice that the points corresponding to different proteins are well mixed which indicates that we have scattering profiles that are close although coming from different

proteins with different shapes and sizes which is encouraging. Nevertheless, we can see that if clusters seem to appear for each aggregate characteristic of a different dimension there still are overlaps between them. However, since we actually consider four different aggregate dimensionalities here - whereas the experimental data was only comparing fibres and monomers - we might need to use 4 principal components rather than 2 to distinguish the four characteristic types of aggregates. We show below the results given by this new test.

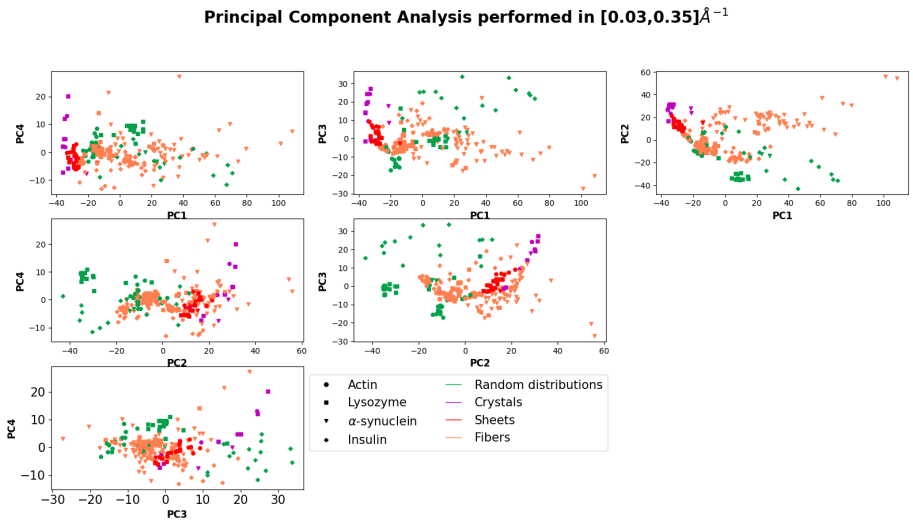

Figure 28: 4 principal components analysis performed in $q \in [0.03; 0.35]\text{Å}^{-1}$ on the theoretical intensities of our database of aggregates

It seems even more difficult from these new results to distinguish specific clusters associated with each characteristic type of aggregate. Thus, despite the apparition of a clustering for the different categories, the clustering is not sharp enough to systematically determine the dimension of an aggregate. The issue might come from the fact that we cannot well visualize the 4D space. One idea would be to perform a three components PCA and visualize the results on a 3D graph. We should also try to perform our two components PCA on only our monomers and fibres to really compare the results to the experimental ones. However, the internship coming to an end, we chose to primarily focus on other methods of analysis that we expected to yield more direct results on whether it is possible to retrieve the dimension from our scattering data. Hence, we chose to try machine learning.

4.3 Training a machine learning algorithm on the scattering files from our numerical database yields promising results

Supervised learning is a subcategory of machine learning where we train an algorithm on a labeled dataset - that is a dataset containing objects of different nature, these objects being all labeled regarding their nature. The ultimate goal is for the algorithm to correctly predict unlabeled data from its training. This technique is able to detect any general feature of each

category of objects we give in input to learn to recognize such objects and classify them accordingly. More particularly, it finds a mapping function to map the input variable with the output variable through a process that is not well understood yet using a neural network with several layers. We apply this technique to the scattering profiles of our database. Our aggregates are labeled by a 4-dimensional vector whose coordinates are equal to zero apart the coordinate corresponding to the dimension of the aggregate which is then equal to 1. For instance, a fibre is labelled with the following vector : [0,1,0,0] whereas a random distribution is labelled as [1,0,0,0]. We show that the results are quite promising even if improvements could be made.

For the implementation of the algorithm, we mainly referred to the lectures given by Andrew Ng on Coursera [23] notably for the different parameters taken in input as the learning rate which determines the step size at each iteration while moving toward a minimum of a loss function. We took three layers, the two hidden layers being composed of respectively 25 and 12 neurons. We then distributed our scattering profiles between the training set and the test set. According to our reference [23], a ratio of 9 to 1 for the number of files in respectively the training and the test set is well suited. However, the main point is above all to have the same distribution of the different categories of aggregates for both sets. We sum-up the composition of both these sets in the array below :

Training set 251 aggregates	Test set 31 aggregates
19.5% random distributions 62.5% fibres 9.96% sheets 7.96% crystals	19.35% random distributions 64.5% fibres 9.67% sheets 6.45% crystals

Hence we arrived at a ratio of $\frac{31}{251}$ that is 12.3% which is a little above the goal ratio. However, the distributions of both sets are quite close which is more important. To evaluate the efficiency of the algorithm of machine learning, we use the accuracy which is defined as the proportion of right predictions. Accuracies are given in output of the algorithm for both sets. More particularly, the accuracy of the training set corresponds to the proportion of correct predictions of the algorithm after its training. We obtained an accuracy of 92.8% for the training set and an accuracy of 90.3% for the test set. We checked that these results were indeed due to the labeling of our files. For that, we shuffled the labels in the train set and launched our algorithm. This gave us a very low accuracy for the test set - that is below 20%. We reiterated the experiment by shuffling this time the labels in the test set which also gave us a very low accuracy. Thus, we can already be quite satisfied by these results.

To better understand these results, and especially the wrong predictions, we evaluated which proportion of bad predictions we got for each type of aggregate. These results are summarized in figures 29 and 30 for both the training and the test sets. Each box of indices

i,j represents how much aggregates of the dimension i is predicted to have the dimension j. Hence,the right predictions are localized on the diagonal. The first line corresponds to random distributions, the second to fibres, the third to sheets and the fourth to crystals.

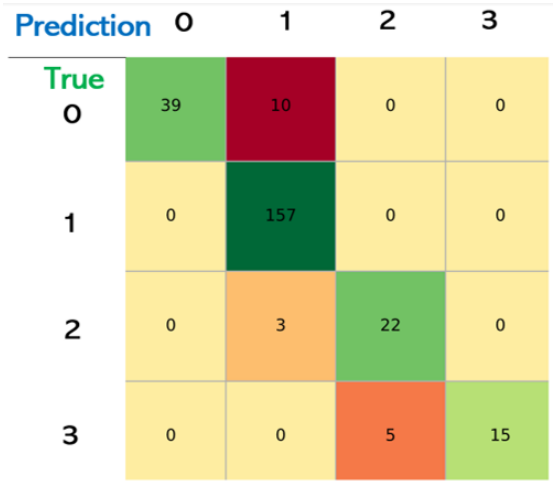

Figure 29: Diagram representing the wrong and right predictions for each category of aggregate for the training set

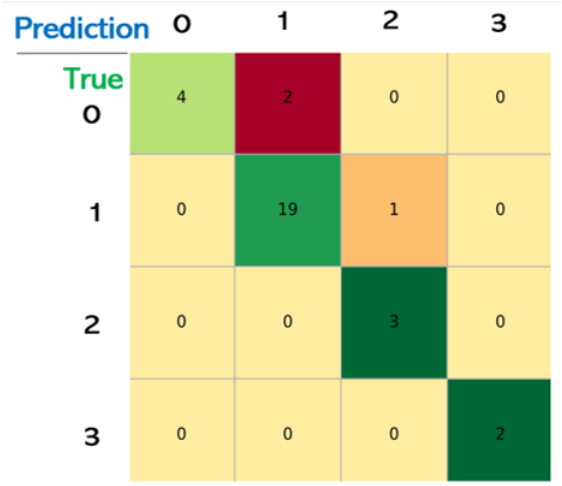

Figure 30: Diagram representing the wrong and right predictions for each category of aggregate for the test set

We can see in the case of the training set that we notably have wrong predictions for random distributions, sheets and crystals which is consistent with the fact that we have much more fibres in this set. Hence, the algorithm has been well trained to detect the common features between fibres thanks to the broad set of fibres contained in our training set whereas it learned less on the other types of aggregate. Thus, we expect that adding crystals, sheets and random distributions to our sets would improve the results. What is more concerning is the fact that the wrong predictions are mainly random distributions which have been wrongly predicted as fibres whereas this in the most important distinction we wish to make when studying the experimental data. However, we can find that these wrong predictions mostly occurred for random distributions of dimers which are very rare in our database - only a fifth of the random distributions which are already underrepresented.

Nevertheless, a reassuring fact is that in each case and in each set, the distribution of proteins with wrong predictions is quite close from the proportions of the database for each protein, confirming the possibility to predict the dimension of any kind of protein aggregates with its normalized scattering intensity regardless of the protein. To make sure of that result, we have relaunched our program with exactly the same parameters but at the difference that we added three files in our test set - a protein not taken in account in our database, tubulin (see section 2.4) - these three files corresponding to a twisted fibre, an helical fibre and a random distribution. Our algorithm then correctly predicted the labels of these three files, even if it did not trained on aggregates of this protein.

Thus, supervised learning appears to be well suited to answer our problem. Notably, we have an algorithm already ready to predict from given experimental X-ray crystallography data restrained on the range $[0.03;0.35]\text{Å}^{-1}$ the dimension of the aggregate contained in the solution. However, if the accuracy obtained are satisfying as to whether it is possible to retrieve the dimension, there still is place for improvement, notably by widening our database.

5 Conclusion

Obtaining experimental data on protein fibres is essential to test the hypothesis that fibres in general arise from generic physics principles. The large amount of crystallography data would have been perfect to test the hypothesis if it was not for the scales studied in such experiments which are smaller than the scale of protein aggregates. Yet, there exists a small overlap of these scales. Hence, we formulated the hypothesis that these crystallography data could contain some hidden information, residues of the data yielded on larger scales. Therefore, to test this hypothesis, we established a strategy consisting into building our own numerical database of protein aggregates, simulating their scattering intensities and finally use elaborated data analysis methods. This led us to machine learning which produced satisfying results. Indeed, we can now determine the dimension of our numerical aggregates from their scattering intensities with a high accuracy.

However, a few points could be discussed. First, the fact that we only studied four proteins. Indeed, the number four has been arbitrarily chosen. Hence, it is possible that our algorithm will not be able to determine the dimension of aggregates formed with very different proteins from the four we chose. Nevertheless, this problem could be tackled by enlarging the database. Another issue is that we did not tried our algorithm on more complex aggregates that would not have a well-defined shape and therefore a well-defined dimension. We could nonetheless try to resolve this problem by building such aggregates and adding them in our database. Finally, as we explained at the end of section 4.2, the use of a principal component analysis could have been deepened. The advantage of this second method would notably be - if we could make it work - to confirm or not the predictions given by our algorithm on experimental data.

Therefore, the next stage now is to test our algorithm on experimental data. The first step would be to test the algorithm on protein fibres whose experimental data is already available on the online databases. Then, we would need to realize our own experiments to further check the accuracy on the predictions of the algorithm.

6 Bibliography

[1] URL : https://www.britannica.com/science/protein

[2] Official website of LPTMS, URL : http://lptms.u-psud.fr/en

[3] Michael F. Hagan and Gregory M. Grason (2021). Equilibrium mechanisms of self-limiting assembly.*Rev. Mod. Phys. 93, 025008.*

[4] Martin Lenz and Thomas A. Witten (2017). Geometrical frustration yields fibre formation in self-assembly. *Nature Physics*, 13, pages 1100–1104.

[5] URL : https://www.khanacademy.org/science/biology/macromolecules/proteins-and-amino-acids/a/introduction-to-proteins-and-amino-acids

[6] Website of the Protein Data Bank, URL : https://www.rcsb.org/

[7] Gregory M. Grasona (2016). Perspective: Geometrically frustrated assemblies. *The Journal of Chemical Physics*,volume 145,Issue 11.

[8] Antonio Cupane and Matteo Levantino (2016). Investigating protein structure and dynamics through wide-angle X-ray solution scattering. *IL NUOVO CIMENTO C*, Vol. 39, Article 303

[9] Karol Vegso, Matej Jergel, Peter Sifflovic, Mario Kotlar, Yuriy Halahovets, Martin Hodas, Marco Pelletta and Eva Majkova (2016). Real-time SAXS study of a strain gauge based on a self-assembled gold nanoparticle monolayer. *Sensors and Actuators A: Physical*, volume 241, pages 87-95.

[10] Lecture of Brigitte PANSU, Paris-Saclay university, Physics department. Small Angle scattering: from individual objects to collective behavior investigation

[11] Anfred Roessle, EMBO Course 2012, Fach Hochschule Lübeck, University of applied Sciences

[12] URL : https://en.wikipedia.org/wiki/Cytoskeleton

[13] Damien JEANGERARD, Martin SAVKO, Lidia CICCONE, Kewin DESJARDINS, Antoine LE JOLLEC, Ahmed HAOUZ, William SHEPARD. Poster of synchrotron SOLEIL : From Plate Screening to Artificial Intelligence: Innovative developments on PROXIMA 2A at Synchrotron SOLEIL

[14] Lilian Jacquamet, Jeremy Ohana, Jacques Joly, Franck Borel, Michel Pirocchi, Philippe Charrault, Alain Bertoni, Pascale Israel-Gouy, Philippe Carpentier, Frank Kozielski, Delphine Blot, and Jean-Luc Ferrer (2004). Automated Analysis of Vapor Diffusion Crystallization Drops with an X-Ray Beam. *Structure*,Vol. 12, 1219–1225.

[15] Roberto Dominguez and Kenneth C. Holmes (2011). Actin Structure and Function? *Annu Rev Biophys.*

[16] Leonidas Stefanis (2012). α-Synuclein in Parkinson's Disease. *Cold Spring Harb Perspect Med.*

[17] Carsten Janke and Maria M. Magiera(2020). The tubulin code and its role in controlling microtubule properties and functions. *Nature Reviews Molecular Cell Biology*,volume 21, pages 307–326.

[18] Eva-Maria Mandelkow and Eckhard Mandelkow, (2012). Biochemistry and Cell Biology of Tau Protein in Neurofibrillary Degeneration. *Cold Spring Harb Perspect Med.*

[19] URL : https://www.rbvi.ucsf.edu/chimerax/

[20] D. Svergun, C. Barberato and M. H. J. Koch (1995). CRYSOL – a Program to Evaluate X-ray Solution Scattering of Biological Macromolecules from Atomic Coordinates. *Journal of Applied Crystallography*

[21] Jaydeep Bardhan, Sanghyun Park and Lee Makowski (2009). SoftWAXS: a computational tool for modeling wide-angle X-ray solution scattering from biomolecules. *Journal of Applied Crystallography*

[22] Daniel K.Putnam, Edward W.Lowe Jr. and Jens Meiler (2013). RECONSTRUCTION OF SAXS PROFILES FROM PROTEIN STRUCTURES. *Computational and Structural Biotechnology Journal* Volume 8, Issue 11

[23] Courses on Machine learning of Andrew Ng on Cousera, URL: https://fr.coursera.org/courses?query=machine

[24] URL : https://iramis.cea.fr/nimbe/Phocea/Vie_des_labos/Ast/ast_sstechnique.php?id_ast=1065