Applied Structural Biology Report

Study of *Staphylococcus aureus* leukotoxins LukD through X-ray crystallography and LukE through native mass spectrometry

This manuscript was compiled on May 24, 2022

Nessim Louafi, Ali Kanso
M1 Quantitative Biology, University of Montpellier
nessim.louafi@etu.umontpellier.fr
ali.kanso@etu.umontpellier.fr

***Staphylococcus aureus* is a gram-negative bacterium commonly found in human microbiota. Its relation with our body is mainly commensal but, in some cases, pathogenic and can lead to several harmful diseases. The study of the molecular details of transition between commensal and pathogenic is crucial to develop new medications and prevent similar pathogenic infections. Precisely, studying the structure of *S.aureus* leukotoxins can offer a novel inside of the infection mechanism. Using the central dogma of structural biology that the structure of a protein is closely related to its function one can hope to understand the role of *S.aureus* secreted proteins LukE and LukD by studying their molecular structures. In this report we show that the combination of two different techniques namely X-ray crystallography and native mass spectrometry can enable us to gain more insights on *S.aureus* secreted toxins LukE and LukD.**

| X-ray Crystallography | Mass Spectrometry | Data Processing| LukE | LukD |

INTRODUCTION

Staphylococcus aureus produces many virulent factors amongst which leukotoxins, a class of pore forming proteins able to disrupt the functioning of leukocytes [1]. Commonly , leukotoxins are classified by their rate of migration on carboxy-methylcellulose creating F(fast eluted) proteins and S(slow) eluted proteins [2]. Originally these proteins were studied using a variety of biochemical and immunological assays. Later with the democratization of techniques such as X-ray crystallography, electronic microscopy and mass spectrometry, structural analysis of leukotoxins became a greater part of the literature. Structural analysis allows the study of the function of a given protein with a precision at the amino acid level.

X-ray crystallography:

X-ray crystallography is a technique that aims at solving a three-dimensional molecular structure from a crystal. To do so, A purified sample at high concentration is crystallized and the crystals are exposed to an x-ray beam [3]. This results in diffraction patterns that can then be processed.

In this report we will go through the different steps of data processing and analysis of 250 diffraction images of LukD, arriving at a 3D structure of our protein. We will also discuss further on the refinement of the model to obtain a complete atomic structure.

Mass spectrometry:

Mass spectrometry (MS) is a chemical analysis technique that permits studying phenomena at the molecular level. It enables direct identification of molecules based on the mass-to-charge ratios as well as fragmentation patterns (dissociation of energetically unstable molecular ions)[4]. Later in this report, we will discuss how we prepared LukE for ionization and used MS to obtain quantitative data and study the stoichiometry of LukE and the binding of a peptide.

MATERIAL & METHODS:

X-ray crystallography:

We first investigated the structure of LukD. 250 crystallographic diffraction images, with different orientations of LukD crystal, were collected with rotating crystal method [5].

Data processing:

The data were processed with xia2 using DIALS (Diffraction Integration for Advanced Light Sources) on CCP4 [6]. We went through 4 major stages: Spot finding, autoindexing, parameter refinement, and Integration. We first started by performing a spot search on the entire dataset. The spots position is determined by pixels that are above local threshold values. Then the reflections were indexed automatically, and a miller index (h, k, l) was assigned for each reflection. In this step, the spots were mapped to reciprocal space and a 3D Fast Fourier Transform (FFT) was used to estimate basis vectors of the unit cell and complete the model for the diffraction geometry. This model was then refined to improve the agreement between spot locations as predicted from the model and their locations on the image data [7]. This step is crucial in data processing to obtain accurate unit cell parameters and thus more accurate integrated intensities. Spot’s intensities were then integrated to be able to phase and obtain an electron density map. The phasing of the dataset was done through molecular replacement using the crystalized structure of homologous protein (LukE, pdb :3ROH) with 32% sequence identity. By placing the structure of LukE in the experimental crystal lattice and using Fourier synthesis we can retrieve the phase of this known structure and use it on the data. The model built was then further refined by checking basic geometry, going to detailed chemical and folding plausibility.

Native MS:

Sample preparation:

We have been given two tubes, one contains a standard well analyzed protein (100uM), and the other one contains LukE. Salts such as sodium or phosphates and detergents used during the preparation of the sample, must be removed to avoid interference during ionization for mass spectrometry analysis. To do so we used Bio-Spin 6 columns (MW exclusion limit of 6 kDa) and replaced the buffer with Ammonium Acetate (200mM, pH = 7,4).

The column was washed and then equilibrated with four times 500µL of ammonium acetate buffer and centrifuged for 1 minute at 900g. 20µL of sample was added to the column and centrifuged at 900g for 4 minutes. The concentration of the protein is then estimated according to the output volume of the column. The concentration was further adjusted by obtaining a rapid spectrum and adjusted to have the best resolution on the mass over charge spectra.

Data analysis:

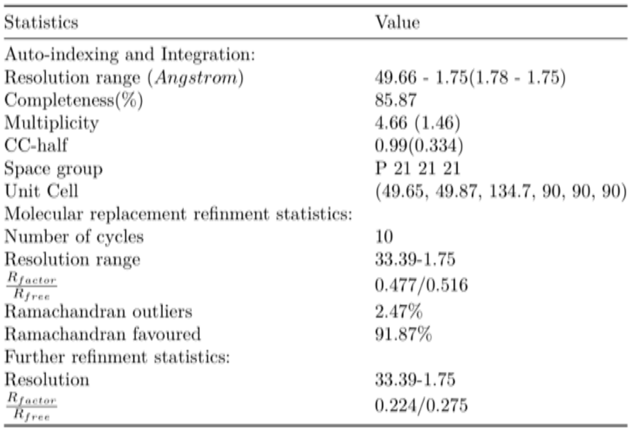
The images were acquired using Masslynx software. The data was preprocessed using the acquisition software analysis tool and further analyzed using a custom-made python pipeline that can be found on GitHub. The data was fitted using a multiple gaussian fitting then displayed as n separate gaussians. The number of gaussians was estimated by observing the spectra. The fitting was done by minimization of a standard least error function between the data and the theoretical gaussian distribution. The mean of the fitted gaussians was used to calculate the mass of each species. The charge state of the average peak for each distribution is calculated from:

$$\frac{n}{(\mathrm{n}_{+1}) - n}$$

With n being the *m/z* of a peak in the isotopic distribution and n+1 the *m/z* of the adjacent peak in the distribution.

RESULT AND DISCUSSION:

X-ray crystallography:



Statistics	Value
Auto-indexing and Integration:	
Resolution range (*Angstrom*)	49.66 - 1.75(1.78 - 1.75)
Completeness(%)	85.87
Multiplicity	4.66 (1.46)
CC-half	0.99(0.334)
Space group	P 21 21 21
Unit Cell	(49.65, 49.87, 134.7, 90, 90, 90)
Molecular replacement refinment statistics:	
Number of cycles	10
Resolution range	33.39-1.75
$\frac{R_{factor}}{R_{free}}$	0.477/0.516
Ramachandran outliers	2.47%
Ramachandran favoured	91.87%
Further refinment statistics:	
Resolution	33.39-1.75
$\frac{R_{factor}}{R_{free}}$	0.224/0.275

Table 1: Summary of the different statistics for the crystallography data analysis

250 Diffraction images of LukD crystal were processed with xia2 using DIALS for indexing, refinement, and integration[7]. This revealed geometry parameters of the predicted model (space group, unit cell parameters, resolution range), and reported statistics to assess the quality of the experimental data (multiplicity, completeness, half-dataset correlation coefficient (cc-half)). The overall crystallography statistics are shown in table1. There are two general indicators for the quality of diffraction data: resolution and data completeness. The resolution limit reported automatically by xia2 is estimated to be 49 Å for the lowest resolution and 1.75 Å for the highest. This resolution can be only nominal and need to be confirmed by looking at the average ratio of reflection intensity to its estimated error ⟨I/σ(I)⟩. Our data showed an overall ⟨I/σ(I)⟩ of 19.2, whereas in the high-resolution shell it drops to near 0.2 < 2.0 suggesting that the true resolution is not as good [8]. Taking this in account we assume that we will still have an electron-density map of low error probability but needs to be manually inspected. These errors can be caused, for example, by insufficient completeness of data. Data completeness is the comparison of collected crystallographic reflections with all theoretically possible unique reflections within the measured data set. With a reported completeness of 86% we considered the data as exploitable and usable for phasing. Phases were determined by molecular replacement using PDB entry 3ROH as the search model. As shown in table1 the process lasted 10 cycles. The final statistic for the phasing is the ratio R_{factor}/R_{free}. Crystallographic R value is a standard indicator for assessing the agreement of a refined model with the data [9]. R_{factor} is a measure of how well the model predicts the data by comparing reflection amplitudes observed in experiment values and those calculated from the model. R_{free} is calculated analogously to R_{factor}, but only on a smaller set of reflections randomly chosen from the dataset and never included in the refinement (A low value of R_{free} is an indicator of successful refinement). By comparing these two measurements we can assess potential overfitting. Overfitting describes the situation in which the noise is affecting the prediction of a model and thus the model becomes unable to generalize the data. A prediction should be independent from the sample size (i.e.: be influenced by noise) resulting in equal values of the two R. The final step of the analysis is thus the model refinement in order to decrease the R_{factor}. This may be done interactively when done manually, or non-interactively using automated fitting tools. To fix the errors in the model, we inspected the electron density around the questionable residues and made necessary corrections using the coot software [10]. We could not complete this step, but if we did, we would have gone through a loop of refinement and model fixing until the electron density difference shows no more significant errors. Later on, we have been able to try an automated way for model refinement by the program REFMAC5 [11]. We see that the final refinement allowed to decrease the R-factor to 0.2. The figure 1 shows the final 3D model of LukD after refinement and the full data processing steps are illustrated in annex 2. These final statistics that we gathered are for validation. Ramachandran plots describe dihedral angles (ϕ, ψ) of the two-dimensional distribution of the protein backbone[12]. We see that for our model we have a Ramachandran outlier of 2.5%, meaning that only 2.5% of the amino acids modeled (7

residues out of 288) were not geometrically possible and have non-favorable dihedral angles.

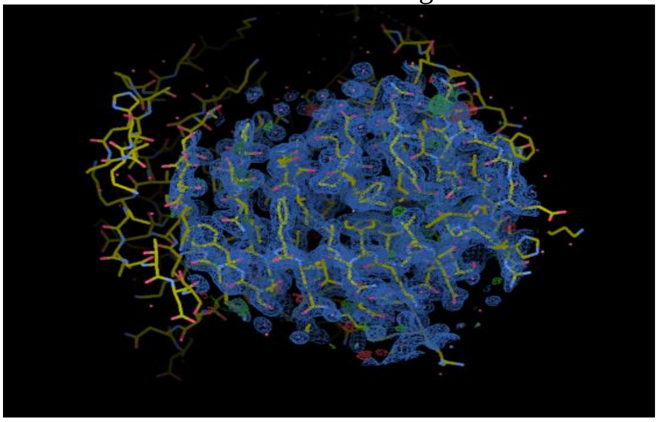


Figure 1 : 3D structure of the protein LukD

Now that we studied the structure of the protein LukD, we changed techniques and tried to study the homologous protein LukE by mass spectrometry.

Mass spectrometry:

The first experiment was aimed at understanding the impact of concentration and trap voltage on the mass over charge spectra on a standard protein: Concanavalin A. The standard protein was tested at two different concentrations namely 12.5 and 50 µM. When increasing the concentration, we saw that the resolution of the spectra decreased, furthermore we observed that experimentally when the concentration is high the electro-spray was less stable and didn't allow for long acquisition time. The difference in spectral resolution can be found in annex 1. The concentration was thus set at 12.5µM for further analysis. As shown in figure 3, the data was fitted with three gaussians suggesting three different stoichiometric states for the standard protein. From the sequence of the protein, the monomeric weight is estimated at 25.5kDa (UniProtKB - P81461). The three populations observed had a molecular weight of respectively: 23.4, 46.8 and 96.9 kilo Daltons.

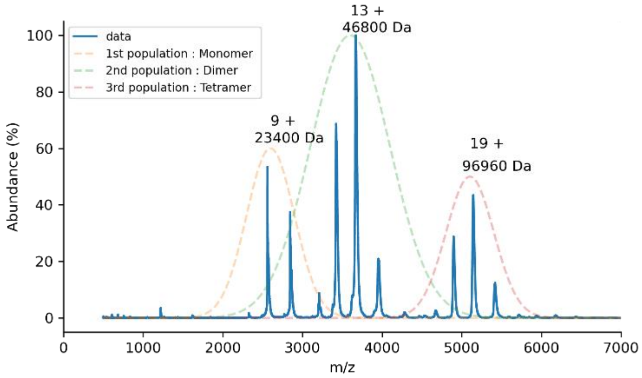


Figure 3: Assignment of the standard protein mass over charge spectra

We can thus conclude that the population observe correspond respectively to the monomeric state, a dimeric state and a tetrameric state. The difference between the expected weight and the calculated weight is due to the method of computation based on the fitting. Once the spectrum was assigned, we aimed at studying the effect of ion activation on the protein. In this experimental setup the sample is colliding with a gas, in our case argon. The energy of the collision of the protein with each gas particle will be accumulated and will eventually lead to the progressive dissociation of the protein. With this method we will be able to study the folding of the protein. From figure 4 we can see that when increasing the voltage applied to the system, i.e the acceleration of the sample towards the gas, we see an increase in the number of peaks in the low m/z region. This is the illustration of the phenomenon described above. Moreover, we see that the dissociation starts at a certain threshold in voltage and by further increasing the voltage we only increase the intensity of the newly created peaks and not the general number. This suggests that there is a limited number of possible dissociation states for this protein which could be related to the folding history of such protein if further experiments were performed.

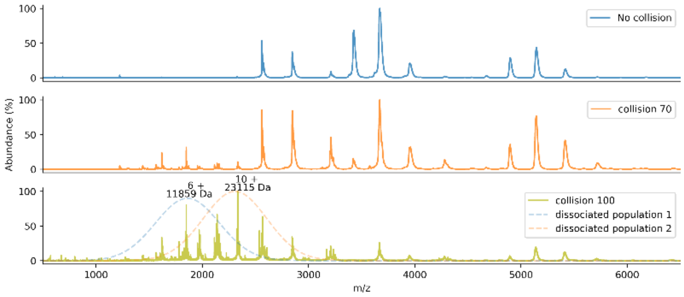


Figure 4: Effect ion activation on the m/z spectra
The spectra were acquired on the software: Masslynx and processed using a python pipeline. The CE voltage applied was increased from 0 to 70 to 100 and spectra were recorded for each voltage.

Quantitatively, we observe two new populations corresponding to the dissociated protein with weights of respectively: 11.8kDa and 23.1 kDa. This suggest that the monomeric state dissociate in two hence the 11.8kDa. The other population probably corresponds to the monomeric state but with a higher charge state due to the partial unfolding of the protein. These conclusions are to be nuanced by the charge state computation that might be incorrect at low *m/z* due to the difficulty to assign a peak to a distribution.

Having the structural analysis of x-ray crystallography for the protein LukD and having the optimal experimental conditions. We set out to study the protein LukE and its association to an unknown peptide ligand. As for the standard protein the assignment was performed using multiple gaussians. The first observation was that 3 apparent distributions showed, however when looking at the mass only two populations differed. Indeed, as shown in figure 4, at low m/z two distributions were fitted. Our hypothesis is that the lowest population (2200 m/z) corresponds to an unfolded state of LukE which would allow more protonation due to exposed amino acids thus increasing the charge without increasing the weight significantly. When the protein is folded there is less protonation possible but however we see sodium binding in the form of adducts figure 5.B. that we don't see at lower m/z as the protein is full of protons and thus does not bind sodium. We only assume that the adduct is in fact sodium from the added weight but complementary experiments could be performed to verify this. Figure 5.C shows the dimer state of the protein, in this state we also see the presence of adducts that add ± 100 Da.

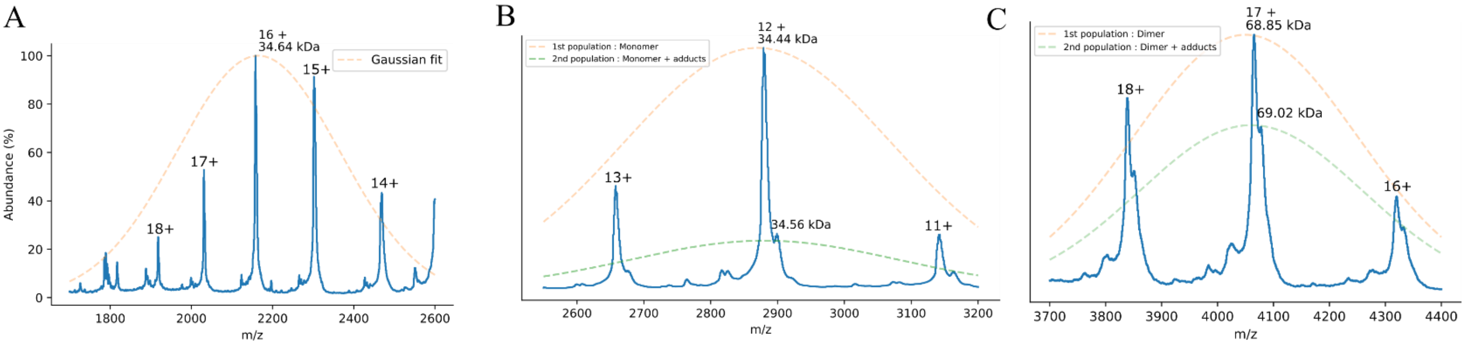


Figure 5: LukE is a protein present in two different stoichiometry in solution.
A.) m/z spectra for the monomeric state of LukE with a gaussian fit used for the mass calculation. B.) m/z spectra for another monomeric state of LukE with multiple gaussian fit used for the mass calculations. C) m/z spectra for the dimeric state of the protein LukE with multiple gaussian fitting.

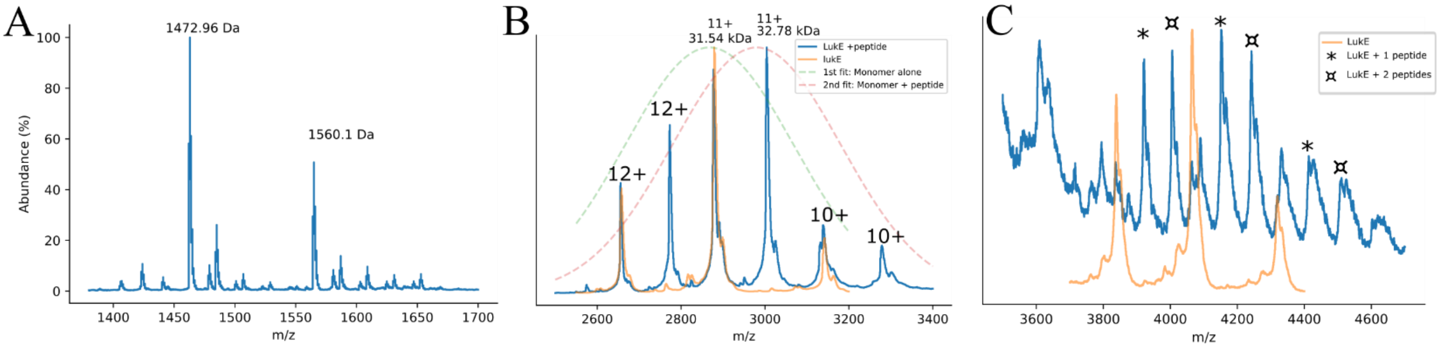


Figure 6: The peptide has a mass of 1.5 kDa and a single binding site on LukE
A.) m/z of the isolated peptide region. The 2 fitting respectively correspond to hydrolyzed and unhydrolyzed peptide. B.) m/z spectra of LukE monomer region with in blue LukE+ peptide and its fit in red, in orange LukE without peptide and its fit in green. C.) m/z spectra for the dimeric region on LukE. LukE alone is represented in orange, LukE plus one peptide is the asterisk and LukE with 2 peptides is represented with ¤

We chose to call the first state the monomeric state as we had access to the sequence of LukE and thus the theoretical weight of the full protein. By comparing the results obtained from the fitting we concluded that the first population at 2200 and 2900 m/z corresponded to a monomeric stoichiometry. Subsequently, the population at around 4000 m/z was labeled as dimeric.

We then investigated the binding of LukE to a peptide. After the assignment of the LukE spectra we expect three regions when adding the peptide: a peptide only region, a monomer with one or more binding sites and then a dimer with two times the binding sites of the monomer. Figure 6 shows successively the three regions of interest in the complete spectrum namely: the peptide alone, LukE monomer and monomer bound to peptide and finally dimer and dimer bound to one or two peptides. In the peptide-only region we observe two populations corresponding to a hydrolyzed or unhydrolyzed peptide. In the monomeric region we see two distinct populations which implies that LukE possesses only one binding site. The difference in weight equals thus to the peptide weight of around 1500 Da which was shown in figure 6A. Finally, in the dimeric region, as expected we see 3 populations that correspond to the dimer bound to zero up to two peptides and thus the corresponding weights of : 69,70 and 72 kDa .

CONCLUSION:

In summary, we have been able to utilize X-ray crystallography and Mass Spectrometry to study the structure of two different toxins (LukE & LukD) on the atomic level and obtain quantitative biological data.
Processing the data of LukD X-ray crystallography, we constructed a 3D model of the protein by molecular replacement using LukE. We also refined the model to arrive at one where only 2.5 of the amino acids modeled were not geometrically possible. With the MS data acquired from LukE, we have been able to identify the overall stoichiometry of the protein. This showed that the protein exists in a monomeric state but can also assemble as a dimer in solution. We have also seen a single binding domain that is found in both the monomeric and the dimeric form of the protein
For future work and more thorough structural interpretation, we can try and use both techniques on the same protein. Studies showed that mass spectrometry can assist and expedite high resolution X-ray structure determination through each stage of the process of protein crystallography [13]. Furthermore, a better computation method to compute weights could be implemented. This method would require an automatic peak and charge state assignment which could be challenging to program. From this assignment the weight could then be an average of all the peaks in one distribution and not only the most abundant one and thus the measure would be more precise. Finally, this method would allow to evaluate a standard deviation on the mass measurement which could lead to more precise comparisons between stoichiometric states.

REFERENCES:

[1] B. A. Diep *et al.*, « Polymorphonuclear leukocytes mediate Staphylococcus aureus Panton-Valentine leukocidin-induced lung inflammation and injury », *Proc Natl Acad Sci U S A*, vol. 107, nº 12, p. 5587-5592, mars 2010, doi: 10.1073/pnas.0912403107.

[2] A. M. Woodin, « Purification of the two components of leucocidin from Staphylococcus aureus », *Biochem J*, vol. 75, p. 158-165, avr. 1960, doi: 10.1042/bj0750158.

[3] M. S. Smyth et J. H. J. Martin, « x Ray crystallography », *Molecular Pathology*, vol. 53, nº 1, p. 8-14, févr. 2000, doi: 10.1136/mp.53.1.8.

[4] P. L. Urban, « Quantitative mass spectrometry: an overview », *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 374, nº 2079, p. 20150382, oct. 2016, doi: 10.1098/rsta.2015.0382.

[5] B. E. Warren, « X-Ray Diffraction Methods », *Journal of Applied Physics*, vol. 12, nº 5, p. 375-384, mai 1941, doi: 10.1063/1.1712915.

[6] M. D. Winn *et al.*, « Overview of the CCP4 suite and current developments », *Acta Cryst D*, vol. 67, nº 4, Art. nº 4, avr. 2011, doi: 10.1107/S0907444910045749.

[7] D. G. Waterman *et al.*, « Diffraction-geometry refinement in the DIALS framework », *Acta Cryst D*, vol. 72, nº 4, Art. nº 4, avr. 2016, doi: 10.1107/S2059798316002187.

[8] A. Wlodawer, W. Minor, Z. Dauter, et M. Jaskolski, « Protein crystallography for non-crystallographers, or how to get the best (but not more) from published macromolecular structures: Protein crystallography for non-crystallographers », *FEBS Journal*, vol. 275, nº 1, p. 1-21, janv. 2008, doi: 10.1111/j.1742-4658.2007.06178.x.

[9] P. A. Karplus et K. Diederichs, « Linking Crystallographic Model and Data Quality », *Science*, vol. 336, nº 6084, p. 1030-1033, mai 2012, doi: 10.1126/science.1218231.

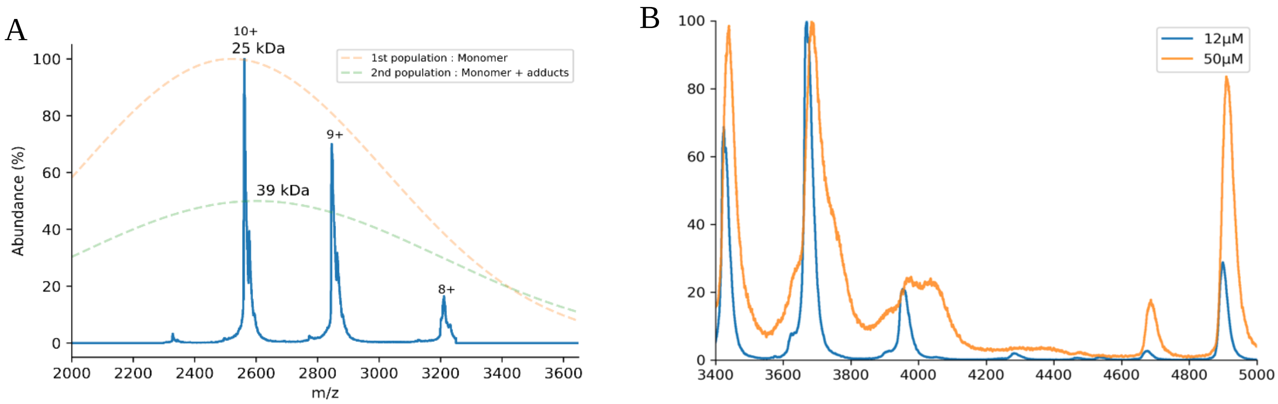
[10] P. Emsley, B. Lohkamp, W. G. Scott, et K. Cowtan, « Features and development of *Coot* », *Acta Crystallogr D Biol Crystallogr*, vol. 66, nº 4, p. 486-501, avr. 2010, doi: 10.1107/S0907444910007493.

[11] G. N. Murshudov *et al.*, « REFMAC5 for the refinement of macromolecular crystal structures », *Acta Cryst D*, vol. 67, nº 4, Art. nº 4, avr. 2011, doi: 10.1107/S0907444911001314.

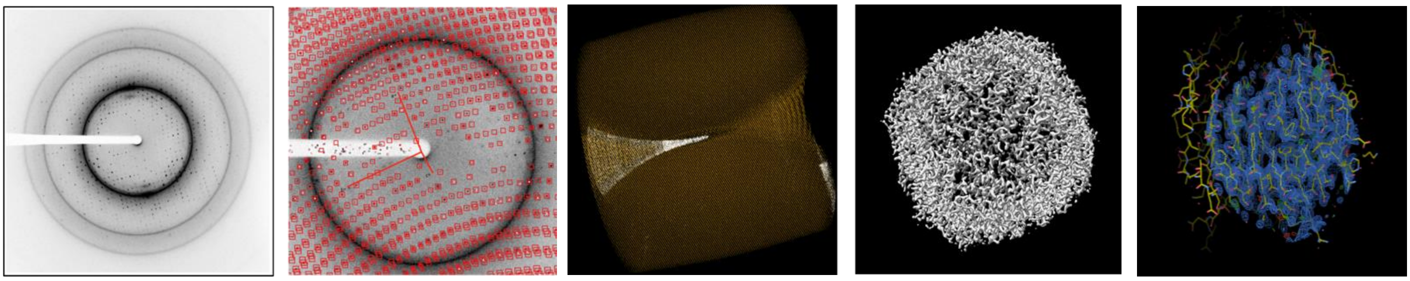
[12] O. V. Sobolev *et al.*, « A Global Ramachandran Score Identifies Protein Structures with Unlikely Stereochemistry », *Structure*, vol. 28, nº 11, p. 1249-1258.e2, nov. 2020, doi: 10.1016/j.str.2020.08.005.

[13] S. L. Cohen et B. T. Chait, « Mass Spectrometry as a Tool for Protein Crystallography », *Annual Review of Biophysics and Biomolecular Structure*, vol. 30, nº 1, p. 67-85, 2001, doi: 10.1146/annurev.biophys.30.1.67.

SUPPLEMENTERY MATERIALS :



Annex 1: Spectral analysis of the standard protein.
The Data were acquired on the software: Masslynx and processed using a python pipeline. The data was gaussian fitted, the mean and sd of the fit was extracted to calculate the mass of the species. A.) Monomeric state of the protein ConcA with the presence of an adduct. B.) The effect of the concentration of the sample on the resolution of the spectra.



Annex 2: LukD crystallography data processing and model refinement steps
A.) representation of the diffraction image. B.) Indexation and integration of the diffraction spots. C.) Representation of the reciprocal crystal lattice. D.) Real-space mapping resulting from the Fourier Transform of the reciprocal lattice E.) Final 3D electron density mapping after refinement.