Equipe DSIMB, UMR-1134, Paris, (www.dsimb.inserm.fr)
Contact: catherine.etchebest@inserm.fr, julien.diharce@inserm.fr

Keywords: Biomedicines, bioinformatics, simulations, protein, modeling, prediction, artificial intelligence, machine learning, optimization, data analysis, programming, computer science, structural biology, molecular visualization, antibodies, aggregation.

Candidate profile and skills: The candidate should preferably have a Master's degree in bioinformatics or data science in biology, or theoretical biochemistry/molecular modeling.

Pre-requisites:

Advanced knowledge of a programming language
Knowledge of learning methods & biostatistics
Knowledge of the concepts of physical chemistry of biomolecules
Knowledge in molecular simulations

Qualities: Dynamism, curiosity, autonomy, imagination and strong adaptation capacities. The candidate will have to evolve in an academic and industrial environment and respond efficiently to different constraints. He/she must be able to collaborate and to communicate both orally and in writing in French and English.

Summary of the PhD project

During the production of a biomedical product, industrialists notice significant losses of active product. These losses of activity result from the different steps implemented, such as purification or storage. Environmental conditions, such as temperature and pH variations, play a key role in maintaining an active product.
Audensiel Healthcare has proposed to set up a decision support tool named AIDRUG for the design of biomedicines, from the sequence design until the implementation of optimal conditions for its industrial production. The PhD project assumes that the loss of product activity is due to structural and/or dynamic changes that can lead to denaturation or aggregation phenomena.
Thus, it becomes crucial to study in detail the risks associated with the very low temperatures encountered during the cryopreservation of the product, to examine the impact of certain excipients as well as of the pH. A first step will be to generate data using advanced conformational sampling methods and those dedicated to simulate the effect of pH. The impact of these factors on the functional areas will be explored. The elucidation of the mechanisms responsible for changes in these regions will allow the identification of relevant descriptors that can be exploited in the second step. This second step will consist in training

methods to predict the observed effect and thus in proposing the optimal parameters limiting this risk.

Objective and context

The implementation of the AIDRUG tool is based on the exploitation of data via learning algorithms. However, few experimental data are available due to their confidentiality at the industrial level. In order to overcome this deficiency, it is possible to simulate data using powerful molecular modeling tools. During a previous thesis, data were generated for a therapeutic antibody, pembrolizumab. The interest was focused on the impact of temperature on the conformation of the antibody and the risk of aggregation. Thanks to molecular dynamics simulations and the implementation of dedicated analysis tools, it was possible to i) show the great flexibility of the antibody considered and ii) identify states accessible at certain temperatures presenting an increased risk of aggregation and thus a risk of loss of activity. The originality of the approach consisted in estimating the risk thanks to a function combining structural and dynamic properties obtained by simulation. It was thus possible to show that the current tools available are limited because they only consider local or at least very restricted conformational changes. They must therefore be revisited in order to take into account the major changes observed by the simulations.
Based on these promising but still preliminary results, Audensiel Healthcare, in partnership with the DSIMB team of UMR-S1134, proposes to further develop this approach by refining in particular the risk score function and by extending the study to other environmental factors influencing the activity of a biomedicine.
The thesis project is thus composed of two major parts: i) a fundamental research part aiming at better understanding the molecular mechanisms resulting in the loss of function of the biomedicine under different conditions ii) an industrial part consisting in the implementation of the prediction tool. The intertwining of these two axes is crucial because it is through the fine analysis of the results of part i) that part ii) can be optimized by selecting the most suitable descriptors.
The results obtained during the previous thesis showed that important risks of loss of function were linked to low temperatures, such as those encountered during the cryopreservation of the product. In the first step, we will further study these low temperatures, by examining in particular the role of the solvent, to finally determine the best cryoprotectant. The protein of interest will still be pembrolizumab. The impact will be evaluated first on the preservation of the tertiary (or quaternary if needed) structure of the protein using advanced conformational sampling methods. Native contacts, tertiary and secondary structure stability will be precisely assessed as they are prerequisites to maintain function. Putative changes in surface properties will be considered in particular at the level of binding sites. Both the interaction properties of the biomedicine with its target and its capacity for irreversible self-assembly leading to aggregation will be examined. The quantification of this interaction will be performed using free energy calculations. Statistical potentials or approaches based on

molecular mechanics calculations will be used. To date, no study of this type has been conducted.
In a second phase, the effect of pH will be examined. Indeed, the pH range visited during the production of a biomedicine ranges from 3 to 7 to 8 pH units. Such a variation has a major consequence on the titratable residues. The previous thesis showed that the risks of aggregation involved precisely such residues. It will be necessary to evaluate for which pH the risks are maximal and to propose modifications of the most sensitive residues. In a third phase, the impact of chemical agents such as metals or fatty acids, added during the different production phases, on the function will also be studied. The approach will be the same as for the two previous phases.
In parallel, a method based on learning approaches based on artificial intelligence will be implemented to i) predict the conformational states from the 3D structure and ii) predict the self-assembly capabilities. The first axis can be achieved using an autoencoder approach as it has already been done for other protein systems. The challenge here lies in the size of the system and its great flexibility. For the second axis, the learning will be done on the self-assembly data produced by the simulations and will aim at predicting the risk of self-assembly thanks to the data characterizing the involvement of each residue in a protein-protein interface. The major difficulty lies in identifying the relevant factors involved in this risk. The solvent surface accessible area is a frequently used measure as well as the nature of the residue considered. However, these two properties are often not sufficient to predict the risk of aggregation with a high confidence. Thanks to the nature of the data produced, it will be possible to broaden the nature of the factors considered and thus to refine the prediction.
This procedure will be carried out for each of the factors examined. Finally, an approach combining these different factors will be developed.

Financial support and supervision:

The thesis will be financed by a CIFRE contract and will take place on two sites, the academic laboratory (DSIMB, UMR-1134, Paris) and the Audensiel Healthcare company (located very close to Paris). The PhD student will be affiliated to the doctoral school BioSPC of the Université Paris Cité and will be supervised in the laboratory by Prof. Catherine Etchebest and Dr Julien Diharce.

Material and financial conditions of the research project

For the realization of the research project, the student will benefit from a fixed and portable computer station allowing him/her to access his/her data remotely. He/she will be able to access the network computer storage space and the computing cluster of the team and the national centers. Requests for computing hours will be made every year. It should be noted that the team has so far obtained 6 million hours to complete the first thesis.

Valorization of the PhD work:

This research project will be carried out within the framework of a collaboration agreement between Audensiel Healthcare and the University in which the conditions of valorization and intellectual property will be specified.
The analysis of the data resulting from the simulations will be valorized through publications.

Start of the thesis on October 1, 2023