Titre : Classification et caractérisation des particules biopharmaceutiques par apprentissage profond à l'intérieur d'une seringue

Title: Classification and characterization of biopharmaceutical particles by deep learning inside a syringe

Laboratoires partenaires impliqués : IBISC (UEVE) **durée totale du stage** 6 mois (M2) ou 3 mois (M1) **date de début et de fin du stage** 15/02/2022 au 1/09/2022

MOTS CLÉS machine learning, deep tech, VAE, méthode de caractérisation, Micro-Flow Imaging, caractérisation de couches minces

Contexte et objectifs

L'imagerie par micro-flux (MFI) est une méthode sensible, simple et automatisée pour l'analyse des particules sous-visibles et des agrégats de protéines translucides qui fournit la taille, le nombre et la morphologie des particules. Les agences qualité demandent plus d'informations sur la nature des particules au sein d'un produit pharmaceutique.

La caractérisation mécanique, chimique et topographique de l'huile de silicone régulière ou réticulée (PDMS) à l'intérieur d'un cylindre en verre - la seringue - est très difficile [4]. Les particules sont potentiellement présentes dans tous les échantillons biopharmaceutiques et peuvent provenir de diverses sources : processus de production, excipients, produits de dégradation, impuretés particulaires, etc. et se présenter sous différentes tailles et forme.

La présence de particules peut influencer négativement la sécurité et l'efficacité de la thérapie. C'est pourquoi la caractérisation des particules est si cruciale.

La plupart des méthodes de caractérisation sont chronophages, destructives ou non évolutives [1]. Par exemple, certaines pratiques actuelles (par exemple l'AFM standard) limitent la caractérisation à quelques échantillons en une semaine pour la préparation, l'étalonnage et la mesure.

La modélisation par apprentissage statistique est une des solutions possibles, avec l'avantage indéniable de pouvoir modéliser des quantités massives de données et de découvrir des représentations de haut niveau [2]. L'apprentissage en profondeur s'est rapidement imposé comme un standard dans plusieurs domaines, battant les records de diverses méthodes de pointe [2]. Dans ce projet nous nous intéresserons plus spécifiquement à la classification des particules avec l'objectif de traiter plus de 35 000 images par minutes. Ces méthodes nécessitent également un grand nombre d'exemples annotés pour former un modèle. Ce projet aborde le problème de l'apprentissage par transfert et de l'augmentation des données dans un contexte de données insuffisantes.

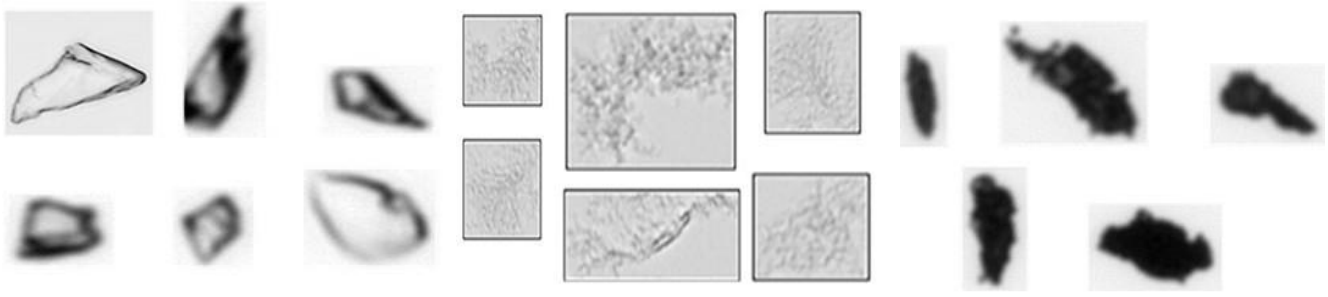

Figure 1 : MFI donne des comptages et des tailles plus précis avec des détails morphologiques complets pour toutes les particules subvisibles dans votre échantillon d'agrégats de protéines.

Méthodes

L'analyse par apprentissage profond des données MFI (*Micro-Flow Imaging*) est presque inexistante. La raison est le manque de grands ensembles de données de particules MFI disponibles en raison de leur coût d'acquisition et de la variabilité des compositions, etc. La difficulté d'obtenir une quantité suffisante de données d'entraînement fiables spécifiques à une classe pour une approche automatique supervisée nécessite l'étude de nouvelles stratégies. Une solution suggérée par des études très récentes [3], propose de développer de nouvelles fonctions génériques de saillance ou d'utiliser la méthode d'augmentation de données pour construire une classification robuste ainsi que d'autres paramètres tels que la texture ou la forme.
L'apprentissage par réseaux de neurones inspirés des réseaux en échelle ou des réseaux de régime ou d'auto-encodage adversaire, modèle curriculaire, etc. sera privilégié dans ce stage.
La solution au problème est de segmenter les particules en suspension dans le fluide par apprentissage automatique (pour surmonter les techniques traditionnelles de segmentation de traitement d'image) et pour cela (a) l'algorithme serait basé en grande partie sur les données d'imagerie MFI déjà disponibles (b) prendre en compte la multimodalités des données (granulométrie, NdG, texture, forme).
L'algorithme de classification devra être plus discriminant que ceux utilisés aujourd'hui sur les images MFI, tout en permettant une analyse rapide des particules présentes sur les images, le débit des machines MFI étant aujourd'hui très élevé, de l'ordre de 200 µl/min avec une concentration de 175 000 particules /ml).
Ainsi, sur la base des informations partagées à ce stade, deux approches algorithmiques nous semblent intéressantes :

- utiliser le deep learning sur un système USR (Ultrafast Shape Recognition),
- adapter les réseaux de neurones modernes de reconnaissance d'images dédiés au edgecomputing connus pour être légers et peu gourmands en CPU.

De nombreuses publications dans divers domaines disciplinaires (physique, géologie, imagerie médicale, etc.) ont déjà démontré l'efficacité de la multimodalité pour l'apprentissage automatique. A noter que les caractéristiques physiques mentionnées dans le sujet (taille, diamètre, rapport hauteur/largeur, circularité, aire, périmètre, intensité) ne sont pas prises en compte explicitement car ces informations sont de facto présentes dans l'image.
En l'absence de vérité terrain, la caractérisation granulométrique est indispensable pour vérifier les performances du programme et la qualité de la segmentation.

Références

[1] Nayyer Aafaq, Ajmal Mian, Wei Liu, Syed Zulqarnain Gilani, and Mubarak Shah. 2019. Video description: A survey of methods, datasets, and evaluation metrics. *ACM Computing Surveys* (CSUR) 52, 6 (2019), 1–37.

[2] LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning, *Nature*, 521(7553) :436–444.

[3] Obeso, A. M., Benois-Pineau, J., Guissous, K., Gouet-Brunet, V., García Vázquez, M. S., and Ramírez Acosta, A. A. (2018). Comparative study of visual saliency maps in the problem of classification of architectural images with deep CNNs. In 2018 Eighth International Conference on Image Processing Theory, Tools and Applications (IPTA), pages 1–6.

[4] D.J. Houde, A.S. Berkowitz, eds., Biophysical Characterization of Proteins in Developing Biopharmaceuticals, 1st ed., Newnes, 2014

Programme de travail

étape 1 : collecter les données existantes étape

étape 2 : Faisabilité. Prototype d'un modèle de réseau de neurones. 2. Formation en réseau sur la base créée en « étape 1 ».

Résultats attendus

La création d'un classifieur capable de prédire et discriminer des formes similaires dans des images basse résolution pour des objets de taille >5µm. Les classes seront le verre, la cellulose, les imitations de protéines et les agrégats de lysozyme.
Les indicateurs de performance attendus sont : (a) la précision des prédictions (b) la répétabilité du processus de prédiction (c) la robustesse en situation dégradée (d) la satisfaction du responsable qualité Les critères de décision/performance potentiels sont : le temps d'identification, les indices de performance de classification (sensibilité, faux positifs, etc.), la reproductibilité des résultats

Profil et compétences recherchées

Le candidat devra fournir un CV complet mentionnant son age, ses notes depuis la 12e année de bachelor ainsi qu'une lettre de motivation.
Capacité à comprendre et à développer des algorithmes d'apprentissage adaptatif et à traiter données médicales, les indexer et les exploiter dans un système opérationnel pour réaliser la mission décrite ci-dessus. Compétences en programmation : Python ou C/C++. Une pratique de Tensorflow et Pytorch serait un plus. La pratique du français n'est pas obligatoire.

Qualités professionnelles recherchées

autonomie, sens du relationnel pour interagir avec les équipes de recherche et d'entreprise, motivation pour les nouvelles technologies, créativité pour mettre en place une solution innovante.

Encadrement et conditions scientifiques Le projet est pluridisciplinaire, à l'interface de l'apprentissage automatique, de l'informatique et de la physique. L'étudiant sera encadré par Vincent Vigneron, Jean-Philippe Congé et Hichem Maaref du laboratoire IBISC (Univ d'Évry, Université Paris-Saclay). Tous maîtrisent le machine learning, le traitement du signal et des images.

Contact:

jean-Philippe Congé **congej@yahoo.fr**
Vincent Vigneron **vincent.vigneron@univ-evry.fr** Hichem Maaref
hichem.maaref @univ-evry.fr
Téléphone : +33 1 69 47 75 45