

Federated Generative Modeling for Predictive Analysis of Microbiome Data

La communauté microbienne complexe qui vit dans l'intestin humain, plus connue sous le nom de microbiote intestinal, présente de nombreuses fonctions importantes pour son hôte et est désormais reconnue comme un facteur crucial du maintien de la santé. Ainsi, lorsque ce microbiote est altéré, il peut jouer un rôle dans l'apparition, la chronicité et la sévérité de nombreuses maladies intestinales et extra-intestinales. L'amélioration et la baisse du coût des techniques de séquençage ont donné accès à une quantité massive de données qui pourraient être utilisées pour le diagnostic, le pronostic, et la prédiction de la réponse au traitement. Cependant, la structure complexe du microbiote intestinal ainsi que les caractéristiques spécifiques des données séquencées soulèvent des défis statistiques majeurs.

Dans ce projet, nous proposons d'utiliser les données de microbiote intestinal pour prédire des variables médicales d'intérêt, notamment la réponse au traitement. Certaines tâches de prédictions peuvent être particulièrement complexes, soit à cause du type de variable à prédire (par exemple la prédiction d'un temps de rechute, qui fait éventuellement intervenir des données censurées), soit à cause de la nature des données disponibles (données hétérogènes provenant de cohortes différentes). Enfin, nous souhaitons fournir des résultats interprétables avec des marges d'erreurs quantifiables qui peuvent être utilisés en pratique à des fins médicales.

D'un point de vue méthodologique, nous proposons d'explorer spécifiquement des modèles génératifs conditionnels, avec différentes implémentations pratiques tirant profit des modèles d'attention utiles pour prendre en compte la multimodalité des données et un conditionnement partiel permettant de masquer un sous-ensemble de covariables afin de rendre le modèle flexible à l'hétérogénéité du jeu de données. Dans un deuxième axe, nous proposons d'introduire de nouvelles méthodologies fédérées pour entraîner des modèles génératifs décrivant la distribution du microbiote intestinal, cadre particulièrement adapté à l'hétérogénéité des données (populations, protocoles de prélèvement, séquençage).