Laboratoire de Mathématiques et Informatique pour la Complexité et les Systèmes

MICS

Présente

L'AVIS DE SOUTENANCE

de **Madame Charlotte Claye**

à l'école doctorale INTERFACES

CentraleSupélec, Université Paris Saclay, qui soutiendra publiquement ses travaux de thèse de doctorat intitulés :

«Concept-based interpretability of foundation models for medical research in immuno-inflammation»

Sous la Direction de Madame Wassila Ouerdane, Madame Céline Hudelot et l'encadrement de Monsieur Julien Duquesne.

Le mardi 2 juin à 14h

À l'école CentraleSupélec, en **amphi V** - Bâtiment Eiffel.

Membres du jury :

Nataliya SOKOLOVSKA, Professeure, Sorbonne Université, Rapporteuse,

Grégoire MONTAVON, Professeur, Charité Berlin, Rapporteur,

Flora JAY, Chargée de recherche CNRS, UP-Saclay, Examinatrice

Romain GIOT, Maître de conférences, Université de Bordeaux, Examinateur

Marie-Jeanne LESOT, Professeure des universités, UPMC, Examinatrice

Résumé :

Les maladies auto-immunes touchent une proportion croissante de la population mondiale et représentent un véritable défi de société. La collecte à grande échelle de données biologiques de haute résolution ouvre de nouvelles perspectives pour mieux comprendre ces maladies, et les modèles de deep learning entrainés de manière auto-supervisée sont une voie pour tirer pari de ces données

complexes. Dans ce travail, nous explorons l'interprétabilité basée concepts comme levier pour extraire des hypothèses scientifiques à partir de ces modèles de deep learning, avec des défis spécifiques aux données biologiques de grande dimension et complexes:

(i) Dans une première partie, nous proposons un cadre modulaire pour l'interprétabilité post-hoc non supervisée par concepts, regroupant de nombreuses méthodes de la littérature et mettant en avant les choix critiques. Ce cadre est ensuite appliqué à un cas d'étude en NLP avec un focus sur l'extraction et l'interprétation de concepts au sein de représentations de textes longs.

(ii) Dans une deuxième partie, nous introduisons une méthode en deux étapes pour l'interprétabilité basée concepts des modèles de type MIL en histologie. Appliquée au diagnostic du Syndrome de Sjögren à partir de biopsies de glandes salivaires, elle permet à des pathologistes d'interpréter des concepts histologiques, de valider le fonctionnement du modèle et de générer une hypothèse de recherche.

(iii) Dans une troisième partie, nous nous intéressons à l'interprétation des modèles de fondation pour la transcriptomique unicellulaire, une modalité particulièrement opaque et de grande dimension. Nous introduisons des méthodes dédiées pour interpréter les concepts et montrons que les autoencodeurs épars (Top-K SAE) permettent d'identifier des concepts plus interprétables que les neurones individuels du modèle.

Abstract:

Autoimmune diseases affect a growing proportion of the world's population and represent a real societal challenge. Large-scale collection of high-resolution biological data opens new perspectives for better understanding these diseases, and deep learning models trained with self-supervision are one avenue for leveraging this complex data. In this work, we explore concept-based interpretability as a means of extracting scientific hypotheses from these deep learning models, addressing challenges specific to high-dimensional and complex biological data:

(i) In a first part, we propose a modular framework for unsupervised post-hoc concept-based interpretability, encompassing numerous methods from the literature and highlighting critical design choices. This framework is then applied to a case study in NLP, with a focus on the extraction and interpretation of concepts within representations of long texts.

(ii) In a second part, we introduce a two-step method for concept-based interpretability of MIL-type models in histology. Applied to the diagnosis of Sjögren's Syndrome from salivary gland biopsies, it enables pathologists to interpret histological concepts, validate the model's behavior, and generate a research hypothesis.

(iii) In a third part, we focus on the interpretation of foundation models for single-cell transcriptomics, a particularly opaque and high-dimensional modality. We introduce dedicated methods for interpreting concepts and show that sparse autoencoders (Top-K SAE) make it possible to identify more interpretable concepts than individual neurons within the model.