

données et algo

pour une ville intelligen

SÉMINAIRE DE LAURENT VIDAL

Détermination de structures moléculaires avec estimation de confiance par diffusion few-step optimisée

Remonter à la structure d'une molécule à partir de données spectroscopiques est un problème central notamment pour la conception de matériaux et de médicaments. À l'heure actuelle, cette tâche repose encore largement sur l'expertise humaine et peut nécessiter des simulations in silico coûteuses en temps et en ressources.

Une fois adaptées aux contraintes chimiques et physiques du problème, les méthodes génératives en apprentissage profond — et en particulier les modèles de type diffusion — offrent une alternative prometteuse : elles peuvent générer directement des structures candidates à partir des spectres et aider à les hiérarchiser. Deux difficultés subsistent cependant : leur échantillonnage reste relativement coûteux, et la confiance associée aux structures prédites est rarement étudiée ou évaluée dans la littérature, limitant leur applicabilité en pratique.

Je présenterai une preuve de concept visant à lever ces deux verrous. Ce travail a été mené avec Jérémie Cabessa (UVSQ), Marie-Pierre Gaigeot (UEVE), Cosmin Marinica (CEA) et Thibault Charpentier (CEA).

Nous introduisons d'abord **Optimized Time Sub-Sampling (O-TSS)**, une méthode permettant de construire systématiquement, et à très faible coût, des grilles d' échantillonnage optimisées pour les modèles de diffusion discrète few-step. Cette accélération rend possible la génération répétée de structures candidates, que nous exploitons pour calculer la masse empirique des candidats générés, et construire un score de confiance. Nous testons ces scores sur la génération de graphes moléculaires à partir de spectres RMN carbone et proton. En réordonnant les candidats selon notre score de confiance, notre démonstrateur double la précision de ses prédictions et atteint, dans certains régimes, des performances comparables à l'état de l'art, avec un budget de calcul 11 fois inférieur et un modèle comportant 4 fois moins de paramètres.
Je profiterai de ce séminaire pour réintroduire de manière pédagogique les principes de la diffusion sur graphes et les architectures utilisées dans ce travail.