

Des réseaux de neurones pour prédire des distances interatomiques extraites d'une base de données ouverte de calculs en chimie quantique

Jules Leguy *, Thomas Cauchy**, Béatrice Duval *, Benoit Da Mota*

*Laboratoire LERIA, Université d'Angers, 2 bd Lavoisier, 49045 Angers, France
{beatrice.duval, benoit.damota}@univ-angers.fr

**Laboratoire MOLTECH-Anjou, Université d'Angers, CNRS UMR 6200,
2 bd Lavoisier, 49045 Angers, France
thomas.cauchy@univ-angers.fr

Résumé. Le calcul de la géométrie de l'état fondamental d'une molécule est le point de départ de l'immense majorité des travaux en chimie quantique moléculaire. La base de données ouverte PubChemQC met à disposition les résultats de calculs des états fondamentaux pour plus de trois millions de molécules. Nous avons extrait les géométries convergées afin d'entraîner des modèles d'apprentissage automatique. Prédire la géométrie complète serait une avancée remarquable. Nos premiers résultats suggèrent qu'il est difficile d'entraîner un réseau de neurones sur cette tâche complexe. Par contre, nous démontrons qu'un réseau de neurones est capable de prédire précisément une distance entre deux atomes. L'objet d'étude de ce travail est la distance la plus complexe en chimie organique, la distance carbone-carbone. Les meilleurs résultats sont obtenus en limitant la quantité d'information grâce à une distance seuil autour de chaque carbone.

1 Introduction

La chimie moléculaire se définit comme l'étude d'entités discrètes (appelées molécules) et correspond à la communauté la plus large de chimistes. Des centaines de millions de molécules sont connues, contenant généralement moins d'une centaine d'atomes et moins d'un millier d'électrons. Les propriétés chimiques de ces molécules dépendent des positions des noyaux et des électrons qui peuvent être calculées de manière approchée par des méthodes issues de la mécanique quantique. Avec la démocratisation de la puissance de calcul, la chimie informatique est devenue une partie essentielle de la recherche en chimie moléculaire. Mais, selon les différents paramètres utilisés, ces calculs peuvent durer de quelques heures à quelques milliers d'heures par molécule. L'apprentissage automatique et plus généralement l'intelligence artificielle appliquée à des données de chimie moléculaire promet de révolutionner la chimie dans un futur proche (Schneider, 2018; Tabor et al., 2018). Avec la récente abondance de données en chimie quantique moléculaire, de nombreux travaux ont vu le jour à un rythme accru depuis 2017. Les modèles employés sont majoritairement de deux types : les réseaux de neurones