# 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

(Schütt et al., 2017, 2018; Gubaev et al., 2018; Hy et al., 2018; Sinitskiy et Pande, 2018) et les méthodes à noyaux de type Support Vector Machine (SVM) ou Gaussian Process Regressions (GPR) (Nakata et Shimazaki, 2017; Bartók et al., 2017; Musil et al., 2018). Aujourd'hui, les travaux se concentrent sur la prédiction de valeurs finales, au sens où si l'énergie totale de la molécule est l'objet d'étude, alors un modèle prédisant cette énergie est entraîné. La plupart des travaux présentent des résultats prometteurs, mais travaillent sur des jeux de données très restrictifs en termes de taille et de variété de molécules ; principalement le jeu de données QM9 avec 1 million de couples géométrie/énergie sur seulement 7165 molécules contenant au maximum 23 atomes.

Les propriétés moléculaires les plus étudiées en chimie quantique concernent la réactivité d'une molécule (localisation des électrons les plus énergétiques, *etc.*) ou ses propriétés d'absorption et d'émission de lumière visible qui dépendent des états excités de la molécule. Dans tous ces cas, une description précise de l'état fondamental est nécessaire. Cela signifie connaître la position d'équilibre des noyaux, ce que l'on appelle la géométrie convergée de l'état fondamental, et connaître les fonctions d'onde des électrons. Ainsi prédire la géométrie complète à partir d'une méthode d'apprentissage automatique serait une importante avancée, permettant notamment d'économiser beaucoup de temps de calculs et permettant à terme d'accélérer et guider le criblage de nouvelles molécules. Un point crucial pour l'apprentissage automatique est la disponibilité de données homogènes ou tout du moins comparables. Or, les calculs en chimie quantique sont toujours des méthodes approchées car la résolution analytique de l'équation de Schrödinger n'est pas possible pour des systèmes contenant plusieurs électrons. Ne sont donc comparables que des calculs effectués avec les mêmes approximations de calculs (sur l'opérateur mathématique ou sur les fonctions d'onde électronique). Des bases de données de calculs homogènes sont très rares en chimie moléculaire. Il existe des bases de données expérimentales de tailles importantes dont la plus conséquente est le projet PubChem contenant plus de 96 millions de molécules (Wang et al., 2009). Il existe aussi au moins cinq bases de données théoriques pour des systèmes de la chimie des solides (comme NoMaD par exemple), mais leurs méthodes de calcul sont malheureusement radicalement différentes et assez incompatibles avec la chimie moléculaire (fonctions mathématiques localisées contre fonctions mathématiques périodiques). À l'échelle moléculaire, depuis 2013 le projet "Clean Energy" d'Harvard contient plus de 2 millions de molécules calculées afin d'en estimer leurs potentiels comme matériau photovoltaïque (`https://cepdb.molecularspace.org/`). Malheureusement, les données des calculs ne sont pas disponibles et ces calculs auraient aussi pu servir à bien d'autres applications. Finalement, une base de données de calculs en chimie moléculaire, PubChemQC (Nakata et Shimazaki, 2017), a été construite par un laboratoire japonais. Elle avait pour objectif ambitieux de calculer avec des paramètres constants tous les composés de la base PubChem. Le projet est au point mort après 3,5 millions de composés calculés, mais il s'agit de la source de données primaires, libre d'accès, la plus homogène et la plus large en chimie moléculaire. Elle est beaucoup plus représentative de l'espace moléculaire que le jeu de données QM9. Nous avons donc utilisé cette source pour l'apprentissage de nos modèles.

# 2 Préliminaires

Notre objectif à terme est de pouvoir se passer du calcul de mécanique quantique ou tout du moins de prédire un bon point de départ pour l'accélérer de façon substantielle. Le premier

problème qu'il faut résoudre est donc de prédire précisément la position des atomes (section 3), problème qui peut être décomposé en la prédiction de la longueur d'une liaison covalente (section 4) et d'angles. Cette longueur de liaison covalente entre deux atomes est un équilibre entre la répulsion des noyaux de charge positive, la répulsion entre les électrons de charge négative et l'attraction entre les électrons et les noyaux. Ainsi la distance d'équilibre dépend de la nature des atomes (carbone, hydrogène, oxygène...) participant à la liaison, mais est également influencée par les atomes au voisinage de la liaison car ils peuvent par exemple attirer à eux une partie des électrons et donc modifier l'équilibre de la liaison. L'influence des atomes du voisinage peut être plus ou moins forte selon leurs positions relatives à la liaison.

Les calculs dont les résultats sont disponibles sur la base PubChemQC (Nakata et Shimazaki, 2017) ont été réalisés à l'aide du logiciel de chimie quantique GAMESS avec comme paramètres la fonctionnelle B3LYP (approximation sur l'opérateur hamiltonien), l'ensemble de fonctions de base 6-31G* (approximation sur les fonctions monoélectroniques), le tout en *closed shell* et phase gazeuse. Nous avons récupéré pour cette étude la géométrie issue de l'optimisation de l'état fondamental. Ce sont ces données qui serviront de cibles à nos modèles prédictifs. Nous avons effectué un premier filtre grossier afin d'enlever les molécules vides ou contenant un unique atome. Afin de limiter la taille des entrées de nos modèles, nous avons fixé une taille maximale de 60 atomes (bien supérieure aux 23 atomes du jeu de données QM9), ce qui permet de garder la quasi-totalité des molécules de cette base. Nos travaux préliminaires de curation manuelle des données nous permettent d'affirmer qu'une partie de ces calculs sont faux, au sens où il n'arrivent pas à optimiser l'état fondamental de la molécule initialement demandée. Il s'agit de calculs qui ont convergé vers une autre molécule par une modification de certaines fonctions chimiques ou en plusieurs autres molécules par une dissociation. Nous considérons dans un premier temps que ces données sont valorisables en terme d'apprentissage. Cette hypothèse ne peut pas être vérifiée actuellement faute de procédure automatique de nettoyage de la base de données, qui aurait permis de comparer les performances de nos modèles avec ou sans ces calculs.

Afin d'évaluer la qualité des prédictions lors de l'entraînement et pour guider les modèles lors de la procédure d'optimisation des poids, nous utilisons l'erreur quadratique moyenne (*Root-Mean-Square Error* ou *RMSE*). Pour $\hat{y}_i$ la valeur prédite pour la variable $y_i$ pour un exemple $i$, le *RMSE* de $N$ prédictions se définit comme suit :

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}{N}}$$

Lors de la prédiction d'une géométrie complète, nous adaptons cette fonction afin de prendre en compte la prédiction d'un vecteur de distances restreint aux sorties correspondant à des atomes en entrée. En effet, le nombre d'atomes variant d'une molécule à une autre, il faut masquer le vecteur de sortie. Pour $\hat{y}_{i,j}$ la valeur prédite pour la variable $y_{i,j}$ pour l'atome $j$ d'une molécule $i$ possédant $A_i$ atomes, le *PRMSE* de $N$ prédictions se définit comme suit :

$$\mathrm{PRMSE} = \sqrt{\sum_{i=1}^{N} \frac{\sum_{j=1}^{A_i}(\hat{y}_{i,j} - y_{i,j})^2}{A_i}}$$

Sans le masquage du *PRMSE*, le modèle apprendrait surtout à prédire des valeurs nulles pour les sorties ne correspondant pas à des atomes en entrée, ce qui constitue une tâche très simple et éloignée de nos objectifs.

L'ensemble de nos traitements ont été réalisés en Python à l'aide des bibliothèques TensorFlow et Scikit-Learn.

# 3 Prédiction de la géométrie complète

## 3.1 Données et modèles

**Représentation géométrique.** Un modèle naïf consisterait à utiliser en entrée une matrice des distances interatomiques, ce qui a été utilisé avec succès par (Schütt et al., 2017) pour prédire l'énergie totale d'une molécule. Les distances relatives ont comme bonne propriété d'être indépendantes d'un repère absolu. Au-delà de quelques atomes cette représentation ne peut pas passer à l'échelle. Il est alors possible de penser à utiliser la trilatération afin de reconstruire des coordonnées avec les distances prédites à partir de 4 distances relatives. En pratique, l'accumulation d'imprécisions rend la reconstruction impossible. Nous avons finalement choisi de représenter nos positions atomiques par des distances à 4 points fixes d'un repère orthonormé. La promesse de l'apprentissage profond étant de pouvoir se passer d'ingénierie des descripteurs, nous fournissons des descripteurs géométriques simples et laissons à la charge du réseau de neurones la projection dans un espace adapté de variables latentes.

**Introduction de bruit.** Afin de prédire des géométries moléculaires optimisées à partir de géométries moléculaires non convergées, la situation idéale serait que les modèles apprennent à partir d'un ensemble de géométries non convergées issues de mesures ou de résultats de mécanique moléculaire, c'est à dire un modèle théorique moins sophistiqué. Malheureusement, la base PubChemQC ne fournit que les géométries optimisées en mécanique quantique et lors de nos essais de génération nous avons été confrontés aux problèmes de l'ordre des atomes, compliquant sérieusement le calcul du RMSE. Nous avons choisi dans un premier temps d'insérer un bruit contrôlé afin de valider notre méthodologie. Le modèle devra alors prédire le bruit afin de le soustraire à la géométrie bruitée. L'introduction de bruit ne garantit donc pas que le modèle se généralisera aux données réelles, mais nous permet de valider la méthode. Le bruit que l'on introduit est un bruit gaussien (normal et identiquement distribué) de moyenne nulle. Le paramètre d'écart-type $\sigma$ permet de contrôler l'amplitude avec précision, tout en générant quelques cas extrêmes. Le déplacement des atomes doit être suffisamment important pour que la tâche d'optimisation de la géométrie moléculaire soit difficile et comparable à des cas d'utilisation réels, mais suffisamment modérée pour que l'on n'inverse pas la position de couples d'atomes. Le déplacement des atomes est réalisé sur les coordonnées, *ie.* avant le calcul des distances. Nous avons estimé qu'il était réaliste chimiquement bruiter les coordonnées des atomes avec un bruit gaussien de paramètre $\sigma = 17,32$ pm. Dans approximativement 95% des cas (i.e. $2\sigma$), l'atome se retrouve ainsi à une distance comprise entre 0 et 60 pm de sa position initiale, ce qui est raisonnable. Pour cette tâche nous disposons de 2,5 millions de molécules. Afin d'évaluer la performance de notre modèle, nous calculons le *PRMSE* après introduction du bruit, ce qui correspond à environ $17,31$ pm.

| **Paramètres** | **Valeurs** |
|---|---|
| Taux d'apprentissage (*learning rate*) | 0,1 ; 0,0001 ; 0,00001 |
| Dégradation des coefficients (*weight decay*) | 0,001 ; 0,01 ; 0,1 |
| Epsilon (Adam optimizer) | 0,0001 ; 1000 |
| Initialisation des poids | 0,002 ; 0,2 |
| Fonction d'activation de la couche de sortie | linéaire |
| Taille de lot (*batch size*) | 500 ; 2000 |
| Époques d'entraînement | 3 |
| Fonction d'activation des couches cachées | elu, crelu |
| Largeur des couches cachées | 360 |
| Nombre de couches cachées | 3 ; 7 |
| Taux de désactivation (*dropout*) | 0,03 ; 0,07 |

TAB. 1 – *Grille des paramètres pour la recherche par quadrillage pour le modèle tentant de prédire la géométrie complète d'une molécule.*

**Modèles.** En plus des données géométriques, nous fournissons aux modèles des informations concernant la nature de chaque atome, *ie.* la masse et le numéro atomique, soit six descripteurs par atome. Les modèles prédictifs possédant une entrée de taille fixe et les molécules une taille variable (nombre d'atomes), nous adaptons la représentation des molécules en prévoyant une couche d'entrée capable de supporter des molécules jusqu'à 60 atomes. Lorsqu'une molécule est de taille inférieure à la taille maximale, les caractéristiques des atomes non définis sont fixées à zéro (*padding*). De même, l'évaluation du modèle est réalisée à l'aide du *PRMSE*. Les modèles testés sont tous des réseaux de neurones possédant des architectures simples. Ils sont composés d'une couche d'entrée (360 neurones), d'une couche de sortie (240 neurones) et d'un certain nombre de couches internes de taille fixe (360 neurones) et entièrement connectées, c'est à dire que chaque neurone d'une couche est connecté à tous les neurones de la couche suivante. Le nombre de couches varie en fonction des modèles (*cf.* table 1). Nous avons pris quelques précautions afin d'éviter le sur-apprentissage de nos modèles, notamment avec le taux de désactivation aléatoire des neurones (*dropout*) et la dégradation des coefficients (*weight decay*). Le temps d'exécution de l'entraînement d'un modèle limite grandement la possibilité d'entraîner des modèles avec des jeux de paramètres variés et un nombre élevé de validations croisées. Il faut donc effectuer un compromis entre la quantité de modèles différents à entraîner, le nombre d'entraînements de chacun de ces modèles et le nombre d'époques. Nous avons effectué une recherche par quadrillage (*cf.* table 1) décrivant les paramètres de 576 modèles différents avec une validation croisée à deux échantillons (*2-fold CV*), soit un total de 1152 entraînements. Puis le même jeu de paramètres a été utilisé afin d'entraîner le modèle sur l'ensemble des données d'entraînement (90 % du jeu de données original) en augmentant le nombre d'époques à 5. Les résultats que nous présentons sont les performances réalisées sur des données mises de côté avant l'entraînement, soit 10 % du jeu de données.

## 3.2 Résultats

À l'issue de la recherche, en dehors de quelques modèles encore moins performants, les performances sont très similaires. Les meilleurs modèles travaillant sur des données ayant un

| | **cibles** | **prédictions** | **\|erreurs\|** |
|---|---|---|---|
| Moyenne | -0,82 | -0,23 | 13,83 |
| Médiane | -0,82 | -0,13 | 11,69 |
| Écart-type | 17,31 | 10,45 | 10,45 |
| Minimum | -94,80 | -9,57 | 0,00 |
| Maximum | 97,24 | 1,23 | 97,80 |

TAB. 2 – *Analyse statistique des valeurs cibles (Δ de distance engendré par le bruit), des prédictions (Δ de distance prédit) et des erreurs absolues en prédiction (en pm).*

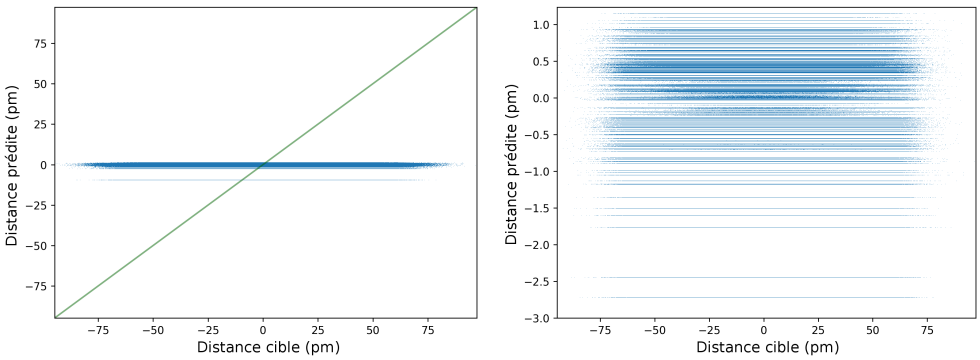


FIG. 1 – *Prédictions en fonction des cibles pour le modèle prédisant une géométrie complète. À droite, le zoom permet d'observer des prédictions discrètes avec un nombre fini de valeurs.*

bruit de PRMSE de 17,31 pm effectuent des prédictions de PRMSE à 10,45 pm (*cf.* table 2). Cela revient à réduire l'erreur à environ 60 % de sa valeur initiale, et donc à prédire 40 % du bruit. Il s'agit d'un gain qui pourrait être non négligeable, même si ce n'est pas réellement utilisable pour optimiser la géométrie des molécules. Toutefois, l'analyse détaillée révèle un comportement inattendu du modèle et remet en cause la nature du bruit introduit.

En effet, l'analyse statistique des données bruitées révèle qu'ajouter le bruit sur les coordonnées plutôt que sur les distances a plus éloigné les atomes de l'origine du repère en moyenne (0.82 pm, *cf.* table 2). Les prédictions de notre modèle s'étendent entre -9,6 pm et 1,2 pm, alors qu'elles devraient s'étendre entre -94,8 pm et 97,2 pm. Le modèle n'arrive donc pas à suffisamment déplacer les atomes pour obtenir les géométries convergées. Pire, il semble tout juste capable de prédire une partie du biais de déplacement en prédisant en moyenne -0.23 pm avec très peu de dispersion. Cet effet est d'autant plus flagrant sur la figure 1. Il est possible de remarquer aussi que le modèle, malgré un très grand nombre de paramètres, prédit un faible nombre de valeurs discrètes. Le modèle apprend très peu, voire n'apprend rien en terme de chimie. Nous avons essayé d'introduire un bruit plus faible ou de l'introduire directement sur les distances, mais nous avons obtenu des résultats similaires. Cette expérience, montre la complexité du problème à résoudre, cependant la tâche ne nous semble pas impossible et nous donnerons quelques pistes à la fin de cet article.

| Classe pos. | | | Distances | | Masse atomique | Numéro atomique | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **g** | **c** | **d** | | | | **H** | **He** | **Li** | **Be** | **B** | **C** | **N** | **O** | **F** |
| 1 | 0 | 0 | $d_{C_1,1}$ | $d_{C_2,1}$ | 14,007 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| 0 | 0 | 1 | $d_{C_1,2}$ | $d_{C_2,2}$ | 15,999 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

TAB. 3 – *Représentation des données d'une liaison en entrée des modèles tentant de prédire des distances carbone-carbone. Pour un atome $k$ du voisinage de la liaison, la distance au premier (resp. second) atome de carbone est notée $d_{C_1,k}$ (resp. $d_{C_2,k}$).*

# 4 Prédiction d'une distance particulière

Les modèles décrits dans cette section ont pour objectif de prédire la distance entre des atomes partageant une liaison covalente au sein d'une molécule. L'objectif n'est donc plus de résoudre le problème de prédiction d'une géométrie moléculaire convergée complète, mais plutôt d'en résoudre une version locale simplifiée.

## 4.1 Données et modèles

**Problème et données.** La liaison carbone-carbone est la liaison chimique la plus complexe de la chimie organique. Nous en avons extrait 6,5 millions de la base PubChemQC, dont 80 % servent à l'entraînement de nos modèles et 20 % à la validation. La représentation de la distribution de cette distance dans notre jeu de données montre une dispersion importante, entre 115 et 160 pm, avec une forte prédominance de liaisons entre 150 et 155 pm (dite simple liaison) et autour de 140 pm (dite double liaison). On retrouve toutefois un certain nombre de triple liaisons vers 120 pm et des liaisons intermédiaires entre ces trois représentations limites (voir graphique en bas à droite de la figure 2). Une précision en dessous du picomètre permettrait de considérer une géométrie prédite comme fiable.

**Représentation géométrique.** La longueur d'une liaison covalente entre deux atomes dépend du type des atomes formant la liaison, mais également de l'influence des atomes au voisinage de la liaison. L'influence des atomes du voisinage dépend de leur position relative à la liaison. C'est pour cette raison qu'en plus des distances, nous introduisons la notion de classe positionnelle qui va représenter de quel côté de la liaison chaque atome se trouve. Les atomes peuvent donc être « à gauche », « au centre » ou « à droite » de la liaison. Formellement, on compare la position des atomes aux deux plans normaux à la liaison et passant par les atomes de la liaison. Si un atome est entre les deux plans, il est de classe « centre », sinon il est de classe « gauche » ou « droite » en fonction du plan dont il est le plus proche. Puisque l'on se place dans le repère relatif de la liaison et qu'il n'y existe pas de notion absolue de gauche ou de droite, ces deux classes sont interchangeables à condition que les atomes appartenant à une classe soient tous à distance minimale du même plan.

**Horizon.** L'influence des atomes au voisinage étant inversement proportionnelle à leur distance aux atomes de la liaison, elle décroît rapidement lorsque ils s'en s'éloignent. Donc,

| **Paramètres** | **Valeurs** |
|---|---|
| Taux d'apprentissage (*learning rate*) | 0,01 |
| Dégradation des coefficients (*weight decay*) | 0,001 |
| Epsilon (Adam optimizer) | 0,001 |
| Initialisation des poids | 0,001 |
| Fonction d'activation de la couche de sortie | linéaire |
| Taille de lot (*batch size*) | 10000 |
| Époques d'entraînement | 300 |
| Fonction d'activation des couches cachées | elu |
| Largeur des couches cachées | 870 |
| Nombre de couches cachées | 3 |
| Taux de désactivation (*dropout*) | 0,02 |

TAB. 4 – *Paramètres des modèles tentant de prédire des distances carbone-carbone.*

l'influence des atomes qui ne sont pas au voisinage direct peut être considérée comme négligeable. Dans le but de tester cette hypothèse et de simplifier la tâche à notre modèle, dit « avec horizon » , nous avons choisi d'implémenter un seuil au-delà duquel les voisins ne sont plus considérés. En pratique, ce seuil a été choisi pour correspondre à une réalité chimique : garder uniquement les distances pouvant correspondre à des liaisons covalentes proches de la liaison carbone-carbone étudiée, soit 200 pm.

**Modèles.** En plus des informations géométriques, nous ajoutons la masse et le numéro atomique de chaque atome au voisinage de la liaison. Le numéro atomique est encodé de façon booléenne (*one-hot encoding*). Cela a pour but de ne pas instaurer de relation d'ordre entre les différents atomes et donc a priori de mieux guider les modèles lors de l'apprentissage. Cela implique toutefois de déterminer une limite aux numéros atomiques des atomes acceptés par un modèle. En effet, cet encodage coûte un attribut pour chaque numéro atomique accepté et cela pour chaque atome au voisinage de la liaison. Afin de travailler sur des modèles de taille raisonnable, nous acceptons les atomes de numéro atomique inférieur ou égal à celui du fluor, ce qui correspond à 9 attributs encodant le numéro atomique pour chaque atome du voisinage. La classe positionnelle de chaque atome par rapport à la liaison est également représentée en *one-hot encoding*. Ainsi, il faut 15 attributs par atome dans le voisinage. La grande majorité des molécules de notre jeu de données étant de taille inférieure à 60 et les deux atomes composant la liaison n'apparaissant pas dans les entrées, nous choisissons de limiter le voisinage de la liaison à 58 atomes, soit une couche d'entrée de taille 870. Les molécules possédant un nombre variable d'atomes et l'entrée des modèles étant de taille fixe, nous effectuons une procédure de *padding* des données : lorsqu'une liaison possède moins de 58 voisins, les blocs correspondant aux atomes non définis valent zéro. La table 3 illustre les entrées de nos modèles. Ceux-ci possèdent 3 couches cachées entièrement connectées de largeur 870 et un unique neurone de sortie dont l'objectif est de prédire la distance entre les deux atomes de carbone. Nous avons pris quelques précautions afin d'éviter le sur-apprentissage de nos modèles, notamment avec le taux de désactivation aléatoire des neurones (*dropout*) et la dégradation des coefficients (*weight decay*) (*cf.* table 4). Les résultats que nous présentons sont les performances réalisées sur des données mises de côté avant l'entraînement, soit 20 % du jeu de données.

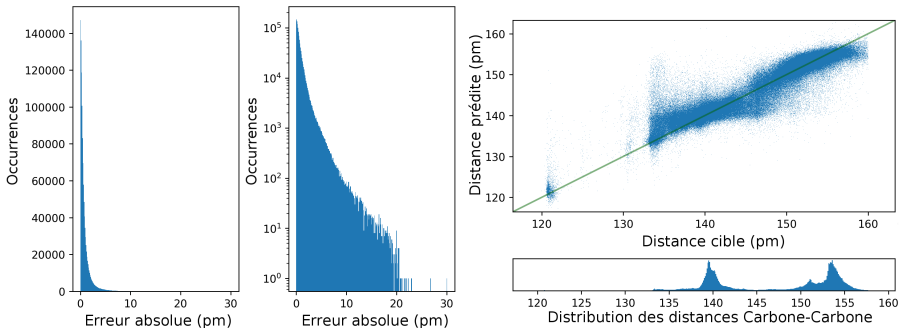


FIG. 2 – *Analyse graphique du modèle tentant de prédire des distances carbone-carbone sans horizon. À gauche, l'histogramme de distribution des erreurs. Au centre, l'histogramme de distribution des erreurs en échelle logarithmique. En haut à droite, le tracé des distances prédites (en ordonnée) en fonction des distances cibles (en abscisse) à mettre en relation avec l'histogramme de distribution des distances cibles en bas à droite.*

## 4.2 Résultats

Le tableau 5 fournit les résultats de l'analyse statistique des erreurs de prédiction des modèles. Les deux modèles obtiennent des performances très satisfaisantes qui permettent d'envisager leur utilisation en pratique. La restriction au plus proche voisinage améliore significativement les performances sur notre jeu de données. Les analyses graphiques des erreurs et des prédictions (figure 2 et 3) des modèles prédisant les longueurs de liaisons entre des atomes de carbone font nettement apparaître la diminution des erreurs importantes. Malgré la quantité de données disponibles, l'espace réel présente une concentration importante sur deux types de distances. Le modèle sans horizon a tendance à ramener, entre autres, les liaisons très courtes (< 130 pm) vers 140 pm. Avec le seuil de 200 pm, une meilleure continuité des prédictions entre les différents types de liaisons apparaît. Soit le modèle sans horizon, plus complexe, ne dispose pas d'assez d'exemples pour bien prédire les distances ayant un faible effectif, soit il n'a pas encore convergé. En ajoutant l'horizon, le modèle est plus simple et possède suffisamment d'exemples pour converger rapidement vers une meilleure solution.

| **Métrique** | **Sans horizon** | **Avec horizon** |
|---|---:|---:|
| Moyenne | 0,833 | 0,342 |
| Médiane | 0,460 | 0.267 |
| Écart-type | 1,207 | 0,337 |
| Minimum | 0,000 | 0,000 |
| Maximum | 30,114 | 26,217 |
| Erreur relative moyenne | 0,006 | 0,002 |

TAB. 5 – *Analyse statistique des erreurs des modèles tentant de prédire des distances carbone-carbone (en pm).*

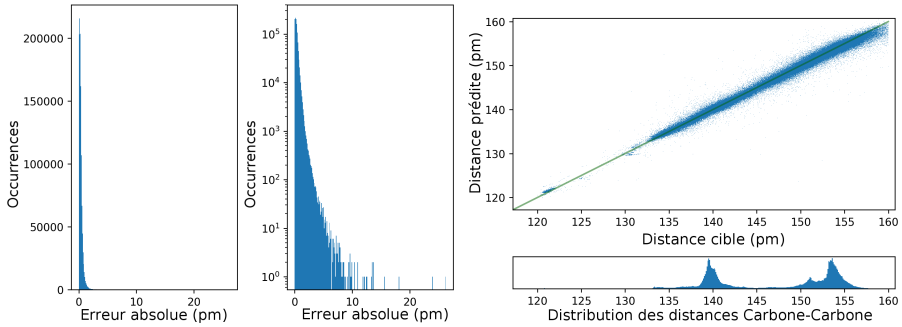


FIG. 3 – *Analyse graphique du modèle tentant de prédire des distances carbone-carbone avec horizon. À gauche, l'histogramme de distribution des erreurs. Au centre, l'histogramme de distribution des erreurs en échelle logarithmique. En haut à droite, le tracé des distances prédites (en ordonnée) en fonction des distances cibles (en abscisse) à mettre en relation avec l'histogramme de distribution des distances cibles en bas à droite.*

# 5 Conclusion et perspectives

Nous avons réalisée une tentative ambitieuse en essayant de prédire la géométrie complète de molécules à partir d'une base de données (PubChemQC) large, diversifiée et imparfaite. La tâche que nous avons tentée d'accomplir avec ces modèles est théoriquement possible, cependant l'approche directe, la plus simple, est particulièrement inefficace. Le fait que le modèle effectue des prédictions constantes et l'impossibilité de produire de meilleurs résultats à l'issue de la recherche par quadrillage ont mené à l'abandon de la méthode pour prédire des géométries moléculaires convergées, au profit d'une méthode plus locale. Toutefois, nous pouvons essayer d'en tirer quelques explications et de nouvelles pistes. Premièrement, les modèles que nous avons entraînés sont des modèles aux architectures relativement simples, avec un nombre de neurones et de connexions limité par les capacités matérielles actuelles. Des architectures plus complexes auraient pu mener à de meilleures performances pour les mêmes données. Un autre écueil pourrait être le manque de données. Même si nous travaillons sur un jeu de données conséquent, il s'agit peut-être d'une quantité insuffisante pour une tâche aussi complexe. Il est également possible que le problème soit lié à notre méthodologie et notamment à l'ajout du bruit sur les données à prédire. Enfin, il est probable, et c'est cette piste de travail que nous souhaitons privilégier pour la suite, qu'il nous manque les bons descripteurs des molécules en entrée des modèles. En effet, les travaux récents mêlant chimie moléculaire et apprentissage obtiennent des résultats très convaincants en utilisant des filtres de convolution reflétant les lois fondamentales de la physique et ayant les propriétés recherchées pour ce type d'application : invariance à l'indexation et à la translation/rotation des atomes (Schütt et al., 2018). La même logique a été déclinée pour l'utilisation de méthodes à noyaux (Bartók et al., 2017; Musil et al., 2018). Les travaux de Sinitskiy et Pande (2018) utilisent une représentation discrétisée dans l'espace (volume 3D) et entraînent des réseaux de neurones convolutifs. Il faut tout de même noter que des distances interatomiques ont été utilisées avec succès par Schütt et al. (2017) afin de prédire l'énergie totale d'une molécule en fonction de sa géométrie. Nous avons donc choisi dans un premier temps d'étudier un sous-problème plus simple.

Les modèles tentant de prédire la distance carbone-carbone travaillent sur des données parfaites, c'est à dire qu'il prédisent des longueurs de liaisons dans des molécules dont la géométrie a déjà été optimisée. Cela nous permet de confirmer notre capacité à effectuer des prédictions d'ordre géométrique en utilisant des distances interatomiques. Afin de prédire avec une haute précision l'immense majorité des distances de la base de données, de la connaissance métier a été introduite dans le modèle d'apprentissage par le biais d'un seuil. Ce seuil permet de mieux discriminer l'environnement proche ayant un fort impact sur la distance calculée. Cette information, relativement simple, limite aussi la taille des données à fournir au modèle. Nous avons également entraîné des modèles sur des liaisons plus simples comme la liaison carbone-hydrogène et la liaison oxygène-hydrogène et les performances sont du même ordre de grandeur. En complément, nous avons testé des modèles de type *support vector machine* (SVM) et *Kernel Ridge Regression* (KRR) sans obtenir de résultats aussi convaincants. Au final, seule une dizaine de cas sur plusieurs millions d'exemples semble poser des problèmes. Une application inattendue de notre modèle est la mise en évidence d'un défaut de curage de la PubChemQC avec des résultats ayant mal été calculés Ainsi notre modèle a été capable de s'entraîner sur des données imparfaites sans sur-apprendre et sa capacité en généralisation permet de mettre en exergue une partie des données de mauvaise qualité dans les données d'origine. Notre modèle peut donc être utilisé afin de vérifier qu'une molécule ne possède pas une longueur de liaison carbone-carbone aberrante ou au contraire, mettre en avant les situations exceptionnelles, importantes en réactivité chimique. Cette piste nous intéresse particulièrement dans le cadre du projet QuChemPedIA, dont un des volets vise à fournir une base de données libre, collaborative et nettoyée pour la chimie quantique moléculaire. La suite de ce travail sur les modèles localisés serait de constituer une procédure itérative combinant différents modèles (réseaux de neurones et modèles à noyaux) et d'ajouter la notion d'angles.

## Remerciements

Ce travail a été financé par un projet d'amorçage de la commission de la recherche de l'Université d'Angers (QuChemPedIA). Les moyens de calcul ont été mis à disposition par le laboratoire LERIA, merci à Jean-Mathieu Chantrein pour son aide.

## Références

Bartók, A. P., S. De, C. Poelking, N. Bernstein, J. R. Kermode, G. Csányi, et M. Ceriotti (2017). Machine learning unifies the modeling of materials and molecules. *Science Advances 3*(12), e1701816.

Gubaev, K., E. V. Podryabinkin, et A. V. Shapeev (2018). Machine learning of molecular properties : Locality and active learning. *The Journal of Chemical Physics 148*(24), 241727.

Hy, T. S., S. Trivedi, H. Pan, B. M. Anderson, et R. Kondor (2018). Predicting molecular properties with covariant compositional networks. *The Journal of Chemical Physics 148*(24), 241745.

Musil, F., S. De, J. Yang, J. E. Campbell, G. M. Day, et M. Ceriotti (2018). Machine learning for the structure–energy–property landscapes of molecular crystals. *Chemical Science 9*(5), 1289–1300.

Nakata, M. et T. Shimazaki (2017). PubChemQC Project : A Large-Scale First-Principles Electronic Structure Database for Data-Driven Chemistry. *Journal of Chemical Information and Modeling 57*(6), 1300–1308.

Schneider, G. (2018). Generative Models for Artificially-intelligent Molecular Design. *Molecular Informatics 37*(1-2), 1880131.

Schütt, K. T., F. Arbabzadah, S. Chmiela, K. R. Müller, et A. Tkatchenko (2017). Quantum-chemical insights from deep tensor neural networks. *Nature Communications 8*, 13890.

Schütt, K. T., H. E. Sauceda, P.-J. Kindermans, A. Tkatchenko, et K.-R. Müller (2018). SchNet – A deep learning architecture for molecules and materials. *The Journal of Chemical Physics 148*(24), 241722.

Sinitskiy, A. V. et V. S. Pande (2018). Deep Neural Network Computes Electron Densities and Energies of a Large Set of Organic Molecules Faster than Density Functional Theory (DFT). *arXiv :1809.02723 [physics]*. arXiv : 1809.02723.

Tabor, D. P., L. M. Roch, S. K. Saikin, C. Kreisbeck, D. Sheberla, J. H. Montoya, S. Dwaraknath, M. Aykol, C. Ortiz, H. Tribukait, C. Amador-Bedolla, C. J. Brabec, B. Maruyama, K. A. Persson, et A. Aspuru-Guzik (2018). Accelerating the discovery of materials for clean energy in the era of smart automation. *Nature Reviews Materials 3*(5), 5–20.

Wang, Y., J. Xiao, T. O. Suzek, J. Zhang, J. Wang, et S. H. Bryant (2009). PubChem : a public information system for analyzing bioactivities of small molecules. *Nucleic Acids Research 37*, W623–W633.

# Summary

The calculation of the geometry of a molecule's fundamental state is the starting point for the vast majority of molecular quantum chemistry research. PubChemQC, an open database, provides the results of fundamental state calculations for more than three million molecules. We have extracted the converged geometries to train machine learning models. Predicting the complete geometry would be a remarkable step forward. Our initial results suggest that it is difficult to train a neural network on this complex task. On the other hand, we demonstrate that a neural network is capable of accurately predicting a distance between two atoms. The subject of this work is the most complex distance in organic chemistry, the carbon-carbon distance. The best results are obtained by limiting the amount of information through a cut-off distance around each carbon.