Master i informatique
Travail Encadré de Recherche
Juin 2018

Rapport QuChemPedia

Sous titre

Auteur
Jules Leguy

Encadrants
Benoit Da Mota
Thomas Cauchy

Table des matières

Chapitre 1

Introduction

Chapitre 2

Description des problèmes

Chapitre 3

Représentations géométriques moléculaires

3.1 Matrice des coordonnées atomiques

La matrice des coordonnées atomiques est la façon la plus simple de représenter la géométrie d'une molécule. L'intérêt de cette représentation est qu'elle est utilisée par les chimistes (fichiers .mol, .xyz + utilisation dans les logiciels de calcul ?). Il s'agit donc pour nous d'une représentation d'entrée et de sortie. Nos données d'apprentissage contiennent pour chaque molécule une matrice des positions, en plus des numéros et masses atomiques, et nous devons être capables de fournir cette représentation en sortie de nos prédictions, pour que nos résultats soient utilisables par les chimistes.

Formellement, la matrice des coordonnées atomiques d'une molécule contient les coordonnées de chaque atome dans un repère cartésien orthonormé à trois dimensions.

x_1	y_1	z_1
x_2	y_2	z_2
$\vdots$	$\vdots$	$\vdots$
x_n	y_n	z_n

FIGURE 3.1 – Matrice des coordonnées atomiques (molécule de taille n)

Si cette représentation de la géométrie des molécules est très commode pour les chimistes, elle n'est pas utilisable telle quelle dans nos modèles prédictifs. Nous cherchons en effet à prédire des distances (ou des différences de distances, voir REF DELTA_DIST...) entre des points. Donner les coordonnées brutes aux modèles implique qu'ils devraient *apprendre* les outils mathématiques permettant de calculer des distances entre des points, ce qui constitue en soi une tâche complexe. C'est pourquoi nous allons définir un ensemble de représentations géométriques, toutes basées sur les distances plutôt que les positions, et adaptées aux différentes prédictions que nous souhaitons effectuer.

3.2 Matrice réduite des distances inter-atomiques

3.2.1 Motivation

Cette représentation est issue du travail qui a été fait précédemment sur ce projet, et consiste à représenter une molécule par ses distances inter-atomiques. L'intérêt de cette représentation est que les réseaux de neurones qui l'utilisent travaillent dans des repères relatifs. Lorsqu'ils effectuent des prédictions, ils n'ont pas besoin de *comprendre* les notions mathématiques de géométrie permettant de déduire la position d'un point dans un repère à partir de ses distances à d'autres points, contrairement à la représentation décrite en (REF REPR ABS). Cette représentation est donc très commode pour les modèles prédictifs dont l'objectif est de corriger les distances entre deux atomes, puisqu'elle est basée sur les distances entre les paires d'atomes.

De plus, l'utilisation d'une représentation basée sur les distances relatives permet d'offrir une représentation unique pour les molécules ayant des ensembles d'atomes pouvant effectuer des rotations, contrairement aux

représentations basées sur les coordonnées (REF REPR COORDS) ou sur des distances à des points fixes (REF REPR DIST ABS).

Lorsque les modèles utilisent cette représentation en sortie, ou plus précisément que l'on déduit la matrice réduite des distances inter-atomiques de la sortie du modèle (voir REF SORTIE DELTA_DIST+H), nous devons toutefois trouver une méthode (voir REF PRINC RECONSTRUCT) pour reconstruire les molécules sous la forme d'une matrice de coordonnées (REF REPR COORDS).

3.2.2 Formalisation

Pour ne pas surcharger les modèles d'information, nous ne travaillons pas sur la matrice de distances inter-atomiques complète, mais sur un sous-ensemble de cardinalité minimale de cette matrice telle que nous pouvons reconstruire sans ambiguïté un ensemble de coordonnées représentant les positions des atomes de la molécule. La matrice des distances étant symétrique et la diagonale étant nulle, toute l'information est contenue dans chaque demi-matrice triangulaire privée de la diagonale.

	a_0	a_1	a_2	a_3	a_4	...	a_{n-4}	a_{n-3}	a_{n-2}	a_{n-1}	a_n
a_0	$\mathbf{d_{0,0}}$	$\mathbf{d_{0,1}}$	$\mathbf{d_{0,2}}$	$\mathbf{d_{0,3}}$	$\mathbf{d_{0,4}}$	**...**	$\mathbf{d_{0,n-4}}$	$\mathbf{d_{0,n-3}}$	$\mathbf{d_{0,n-2}}$	$\mathbf{d_{0,n-1}}$	$\mathbf{d_{0,n}}$
a_1	$\mathbf{d_{1,0}}$	$\mathbf{d_{1,1}}$	$\mathbf{d_{1,2}}$	$\mathbf{d_{1,3}}$	$\mathbf{d_{1,4}}$	**...**	$\mathbf{d_{1,n-4}}$	$\mathbf{d_{1,n-3}}$	$\mathbf{d_{1,n-2}}$	$\mathbf{d_{1,n-1}}$	$\mathbf{d_{1,n}}$
a_2	$\mathbf{d_{2,0}}$	$\mathbf{d_{2,1}}$	$\mathbf{d_{2,2}}$	$\mathbf{d_{2,3}}$	$\mathbf{d_{2,4}}$	**...**	$\mathbf{d_{2,n-4}}$	$\mathbf{d_{2,n-3}}$	$\mathbf{d_{2,n-2}}$	$\mathbf{d_{2,n-1}}$	$\mathbf{d_{2,n}}$
a_3	$\mathbf{d_{3,0}}$	$\mathbf{d_{3,1}}$	$\mathbf{d_{3,2}}$	$\mathbf{d_{3,3}}$	$\mathbf{d_{3,4}}$	**...**	$\mathbf{d_{3,n-4}}$	$\mathbf{d_{3,n-3}}$	$\mathbf{d_{3,n-2}}$	$\mathbf{d_{3,n-1}}$	$\mathbf{d_{3,n}}$
a_4	$\mathbf{d_{4,0}}$	$\mathbf{d_{4,1}}$	$\mathbf{d_{4,2}}$	$\mathbf{d_{4,3}}$	$\mathbf{d_{4,4}}$	**...**	$\mathbf{d_{4,n-4}}$	$\mathbf{d_{4,n-3}}$	$\mathbf{d_{4,n-2}}$	$\mathbf{d_{4,n-1}}$	$\mathbf{d_{4,n}}$
⋮	⋮	⋮	⋮	⋮	⋮	⋱	⋮	⋮	⋮	⋮	⋮
a_{n-4}	$\mathbf{d_{n-4,0}}$	$\mathbf{d_{n-4,1}}$	$\mathbf{d_{n-4,2}}$	$\mathbf{d_{n-4,3}}$	$\mathbf{d_{n-4,4}}$	**...**	$\mathbf{d_{n-4,n-4}}$	$\mathbf{d_{n-4,n-3}}$	$\mathbf{d_{n-4,n-2}}$	$\mathbf{d_{n-4,n-1}}$	$\mathbf{d_{n-3,n}}$
a_{n-3}	$\mathbf{d_{n-3,0}}$	$\mathbf{d_{n-3,1}}$	$\mathbf{d_{n-3,2}}$	$\mathbf{d_{n-3,3}}$	$\mathbf{d_{n-3,4}}$	**...**	$\mathbf{d_{n-3,n-4}}$	$\mathbf{d_{n-3,n-3}}$	$\mathbf{d_{n-3,n-2}}$	$\mathbf{d_{n-3,n-1}}$	$\mathbf{d_{n-3,n}}$
a_{n-2}	$\mathbf{d_{n-2,0}}$	$\mathbf{d_{n-2,1}}$	$\mathbf{d_{n-2,2}}$	$\mathbf{d_{n-2,3}}$	$\mathbf{d_{n-2,4}}$	**...**	$\mathbf{d_{n-2,n-4}}$	$\mathbf{d_{n-2,n-3}}$	$\mathbf{d_{n-2,n-2}}$	$\mathbf{d_{n-2,n-1}}$	$\mathbf{d_{n-2,n}}$
a_{n-1}	$\mathbf{d_{n-1,0}}$	$\mathbf{d_{n-1,1}}$	$\mathbf{d_{n-1,2}}$	$\mathbf{d_{n-1,3}}$	$\mathbf{d_{n-1,4}}$	**...**	$\mathbf{d_{n-1,n-4}}$	$\mathbf{d_{n-1,n-3}}$	$\mathbf{d_{n-1,n-2}}$	$\mathbf{d_{n-1,n-1}}$	$\mathbf{d_{n-1,n}}$
a_n	$\mathbf{d_{n,0}}$	$\mathbf{d_{n,1}}$	$\mathbf{d_{n,2}}$	$\mathbf{d_{n,3}}$	$\mathbf{d_{n,4}}$	**...**	$\mathbf{d_{n,n-4}}$	$\mathbf{d_{n,n-3}}$	$\mathbf{d_{n,n-2}}$	$\mathbf{d_{n,n-1}}$	$\mathbf{d_{n,n}}$

FIGURE 3.2 – Matrice complète des distances inter-atomiques d'une molécule

De plus, nous n'avons besoin que des distances à quatre points pour retrouver la position de chaque atome (voir REF RECONSTRUCT), nous nous contentons donc de garder les quatre premières distances de chaque ligne de la matrice triangulaire supérieure privée de la diagonale.

	a_0	a_1	a_2	a_3	a_4	...	a_{n-4}	a_{n-3}	a_{n-2}	a_{n-1}	a_n
a_0	$d_{0,0}$	$\mathbf{d_{0,1}}$	$\mathbf{d_{0,2}}$	$\mathbf{d_{0,3}}$	$\mathbf{d_{0,4}}$	...	$d_{0,n-4}$	$d_{0,n-3}$	$d_{0,n-2}$	$d_{0,n-1}$	$d_{0,n}$
a_1	$d_{1,0}$	$d_{1,1}$	$\mathbf{d_{1,2}}$	$\mathbf{d_{1,3}}$	$\mathbf{d_{1,4}}$	...	$d_{1,n-4}$	$d_{1,n-3}$	$d_{1,n-2}$	$d_{1,n-1}$	$d_{1,n}$
a_2	$d_{2,0}$	$d_{2,1}$	$d_{2,2}$	$\mathbf{d_{2,3}}$	$\mathbf{d_{2,4}}$	...	$d_{2,n-4}$	$d_{2,n-3}$	$d_{2,n-2}$	$d_{2,n-1}$	$d_{2,n}$
a_3	$d_{3,0}$	$d_{3,1}$	$d_{3,2}$	$d_{3,3}$	$\mathbf{d_{3,4}}$	...	$d_{3,n-4}$	$d_{3,n-3}$	$d_{3,n-2}$	$d_{3,n-1}$	$d_{3,n}$
a_4	$d_{4,0}$	$d_{4,1}$	$d_{4,2}$	$d_{4,3}$	$\mathbf{d_{4,4}}$	...	$d_{4,n-4}$	$d_{4,n-3}$	$d_{4,n-2}$	$d_{4,n-1}$	$d_{4,n}$
⋮	⋮	⋮	⋮	⋮	⋮	⋱	⋮	⋮	⋮	⋮	⋮
a_{n-4}	$d_{n-4,0}$	$d_{n-4,1}$	$d_{n-4,2}$	$d_{n-4,3}$	$d_{n-4,4}$	...	$d_{n-4,n-4}$	$\mathbf{d_{n-4,n-3}}$	$\mathbf{d_{n-4,n-2}}$	$\mathbf{d_{n-4,n-1}}$	$\mathbf{d_{n-4,n}}$
a_{n-3}	$d_{n-3,0}$	$d_{n-3,1}$	$d_{n-3,2}$	$d_{n-3,3}$	$d_{n-3,4}$	...	$d_{n-3,n-4}$	$d_{n-3,n-3}$	$\mathbf{d_{n-3,n-2}}$	$\mathbf{d_{n-3,n-1}}$	$\mathbf{d_{n-3,n}}$
a_{n-2}	$d_{n-2,0}$	$d_{n-2,1}$	$d_{n-2,2}$	$d_{n-2,3}$	$d_{n-2,4}$	...	$d_{n-2,n-4}$	$d_{n-2,n-3}$	$d_{n-2,n-2}$	$\mathbf{d_{n-2,n-1}}$	$\mathbf{d_{n-2,n}}$
a_{n-1}	$d_{n-1,0}$	$d_{n-1,1}$	$d_{n-1,2}$	$d_{n-1,3}$	$d_{n-1,4}$	...	$d_{n-1,n-4}$	$d_{n-1,n-3}$	$d_{n-1,n-2}$	$d_{n-1,n-1}$	$\mathbf{d_{n-1,n}}$
a_n	$d_{n,0}$	$d_{n,1}$	$d_{n,2}$	$d_{n,3}$	$d_{n,4}$	...	$d_{n,n-4}$	$d_{n,n-3}$	$d_{n,n-2}$	$d_{n,n-1}$	$d_{n,n}$

FIGURE 3.3 – Matrice réduite des distances inter-atomiques d'une molécule (en gras)

$\mathbf{d_{0,1}}$	$\mathbf{d_{0,2}}$	$\mathbf{d_{0,3}}$	$\mathbf{d_{0,4}}$
$\mathbf{d_{1,2}}$	$\mathbf{d_{1,3}}$	$\mathbf{d_{1,4}}$	$\mathbf{d_{1,5}}$
$\mathbf{d_{2,3}}$	$\mathbf{d_{2,4}}$	$\mathbf{d_{2,5}}$	$\mathbf{d_{2,6}}$
$\mathbf{d_{3,4}}$	$\mathbf{d_{3,5}}$	$\mathbf{d_{3,6}}$	$\mathbf{d_{3,7}}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\mathbf{d_{n-4,n-3}}$	$\mathbf{d_{n-4,n-2}}$	$\mathbf{d_{n-4,n-1}}$	$\mathbf{d_{n-4,n}}$
$\mathbf{d_{n-3,n-2}}$	$\mathbf{d_{n-3,n-1}}$	$\mathbf{d_{n-3,n}}$	$\mathbf{0}$
$\mathbf{d_{n-2,n-1}}$	$\mathbf{d_{n-2,n}}$	$\mathbf{0}$	$\mathbf{0}$
$\mathbf{d_{n-1,n}}$	$\mathbf{0}$	$\mathbf{0}$	$\mathbf{0}$

FIGURE 3.4 – Matrice réduite des distances inter-atomiques d'une molécule

3.2.3 Reconstruction des molécules

Lorsqu'un modèle a pour sortie une matrice réduite des distances inter-atomiques lorsqu'il effectue des prédictions, il faut définir une méthode pour reconstruire une matrice des coordonnées (ref REPR MAT COORDS) de façon automatique à partir de cette sortie, la seule contrainte étant que la distance relative entre chaque paire d'atomes soit respectée. Il ne s'agit pour autant pas d'une tâche triviale, elle s'est en effet avérée impossible en pratique pour les grosses molécules à cause de la propagation des erreurs qu'elle induit (voir REF PROPAG ERREURS).

3.2.3.1 Formalisation de la méthode de reconstruction

Nécessité et limite de l'introduction d'un atome fictif Notre méthode de reconstruction des atomes doit permettre de respecter la chiralité [1] des molécules. Or, en déduisant uniquement la position d'un atome de ses distances aux quatre atomes précédents, il existe des cas pour lesquels il existe plusieurs solutions pour la position de l'atome (deux si les quatre atomes précédents sur un même plan ou une infinité si les quatre atomes précédents appartiennent à une droite). Pour pallier ce problème, la méthode retenue précédemment a été d'introduire un nouveau point (que l'on nomme atome fictif) et que l'on place arbitrairement dans la molécule, à une position telle qu'il n'appartient pas au plan formé par les trois premiers atomes, ou à la droite formée par les trois premiers atomes s'ils sont alignés. De cette façon, les atomes suivants seront placés sans ambiguïté.
Cependant, on peut imaginer des cas pour lesquels la technique de l'introduction d'un atome fictif ne permet pas de lever l'ambiguïté, notamment pour les molécules possédant une chaîne de carbones liés par des doubles liaisons (et formant donc une droite). La méthode ne permettra pas dans ce cas de déterminer les positions des atomes en bout de chaîne tel que leurs distances relatives soient respectées, cette information étant perdue lors de la création de la matrice réduite des distances inter-atomiques.
Cette représentation n'est donc pas viable en pratique. Cela fait partie des raisons (voir également REF PB SQRT) pour lesquelles nous sommes passés à la représentation par matrice réduite des distances à des points fixes (REF MATR RED FIXES).

Placement de l'atome fictif Puisque l'on définit la position de chaque atome en fonction de ses distances aux quatre atomes précédents, on doit d'abord placer les quatre premiers atomes de façon partiellement arbitraire. Le premier atome de la molécule dans notre représentation étant l'atome fictif a_0, nous commençons par le placer à la position qui lui a été attribuée.

Placement de l'atome $\mathbf{a_1}$ Une fois l'atome a_0 placé, il existe une infinité de solutions pour la position de l'atome a_1. On peut en effet le placer à tout point appartenant à la surface de la sphère de centre a_0 et de rayon $d_{0,1}$.

Placement de l'atome $\mathbf{a_2}$ L'atome a_2 appartient au cercle solution de l'intersection entre les sphères de centres a_0 et a_1 et de rayons $d_{0,2}$ et $d_{1,2}$. On choisit donc arbitrairement une position appartenant à ce cercle.

1. Un composé chimique est dit chiral s'il n'est pas superposable à son image dans un miroir. (https://fr.wikipedia.org/wiki/Chiralité_(chimie))

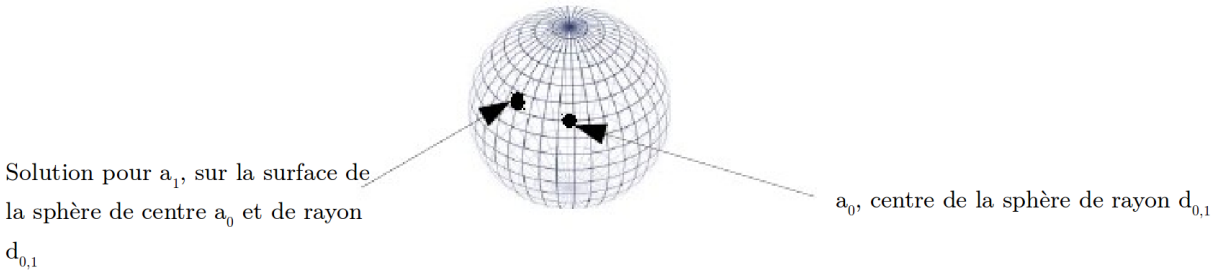

FIGURE 3.5 – Placement de l'atome a_1 (image extraite du rapport de N.Roux)

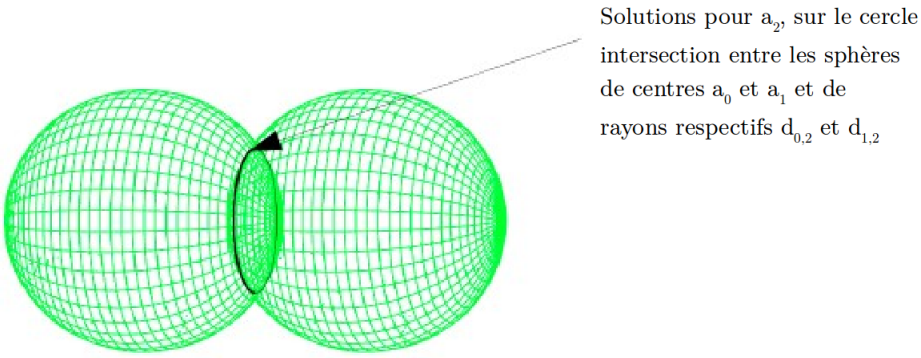

FIGURE 3.6 – Placement de l'atome a_2 (image extraite du rapport de N.Roux)

Placement de l'atome a_3 Dans le cas général, il existe deux solutions pour le placement de l'atome a_3, l'intersection non nulle de trois sphères étant deux points si tous les points ne sont pas sur un même plan ou une même droite. On choisit arbitrairement un point parmi ces deux solutions, car il n'y a pas à ce stade d'ambiguïté de chiralité de la molécule, une molécule composée de trois atomes ne possédant pas de chiralité (l'atome fictif a_0 ne fait pas partie de la molécule).

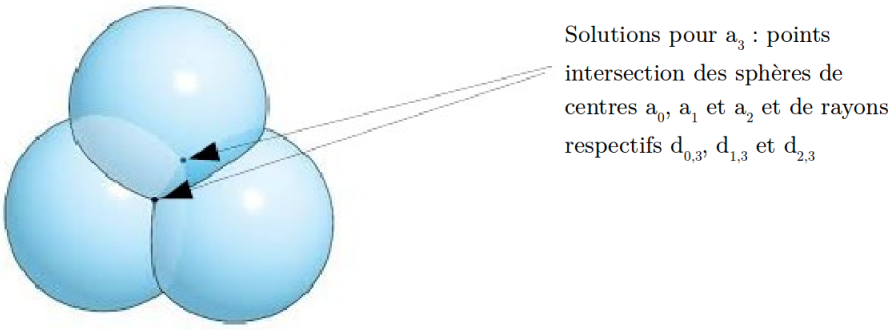

FIGURE 3.7 – Placement de l'atome a_3 (image extraite du rapport de N.Roux)

Placement de l'atome a_n Pour placer l'atome a_n (n étant strictement inférieur à la taille de la molécule), nous généralisons la méthode de placement de l'atome a_3. Plutôt que de travailler sur l'intersection de quatre sphères, nous travaillons toujours sur l'intersection de trois sphères et nous utilisons la dernière distance pour discriminer les deux solutions obtenues. Cela facilite grandement la résolution des équations mathématiques associées et permet d'obtenir des solutions sensiblement équivalentes.
Formellement, nous calculons les positions des deux points solutions de l'intersection des trois sphères de centres a_{n-4}, a_{n-3}, et a_{n-2} et de rayons $d_{n-4,n}$, $d_{n-3,n}$ et $d_{n-2,n}$, et nous discriminons les deux solutions selon la distance

$d_{n-1,n}$.

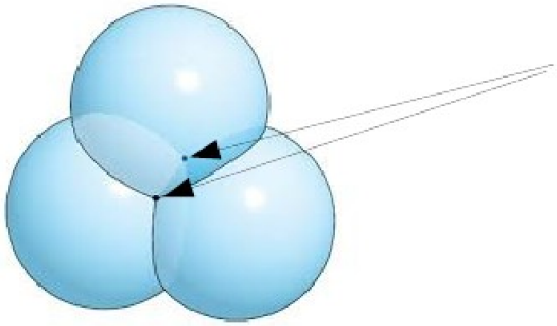

FIGURE 3.8 – Placement de l'atome a_n (image extraite du rapport de N.Roux)

3.2.3.2 Reconstruction automatique des positions en utilisant un solveur

Nous développons ici une méthode permettant de déterminer les coordonnées d'un atome quelconque en utilisant un solveur d'équations non linéaires [2]. Nous utilisons la bibliothèque Sympy [3].

Tout d'abord, l'atome fictif a_0 doit être placé à la position qui lui a été attribuée (REF PLACEMENT AT FICTIF). Nous plaçons ensuite arbitrairement les trois atomes suivants, de sorte que leurs distances relatives soient respectées. Pour simplifier le problème, nous effectuons une translation temporaire telle que $a_0{}'$ est à l'origine du repère. Nous plaçons alors $a_1{}'$ sur l'axe x, à une distance $d_{0,1}$ de l'origine, et $a_2{}'$ sur le plan tel que $z = 0$, à une position telle que les distances $d_{0,2}$ et $d_{1,2}$ sont respectées. Pour finir, nous plaçons $a_3{}'$ à l'une des deux solutions de l'intersection des sphères associées au problème (REF PLACEMENT A3). Le choix de la solution est arbitraire car la reconstruction de la bonne chiralité de la molécule ne dépend pas du placement des trois premiers atomes non fictifs.

$$a'_0 \begin{cases} x'_0 = 0 \\ y'_0 = 0 \\ z'_0 = 0 \end{cases} \quad a'_1 \begin{cases} x'_1 = d_{0,1} \\ y'_1 = 0 \\ z'_1 = 0 \end{cases} \quad a'_2 \begin{cases} x'_2 = \frac{d^2_{0,2} - d^2_{1,2} + x^2_1}{2x'_1} \\ y'_2 = \sqrt{d^2_{2,0} - x'^2_2} \\ z'_2 = 0 \end{cases} \quad a'_3 \begin{cases} x'_3 = \frac{d^2_{0,3} + x'^2_1 - d^2_{1,3}}{2x'_1} \\ y'_3 = \frac{-2x'_3x'_2 + d^2_{0,2} - d^2_{2,3} + d^2_{0,3}}{2y'_2} \\ z'_3 = \sqrt{-x'^2_3 - y'^2_3 + d^2_{0,3}} \end{cases}$$

FIGURE 3.9 – Placement des atomes $a_0{}'$, $a_1{}'$, $a_2{}'$ et $a_3{}'$

Une fois que les quatre premiers atomes sont placés, nous leur appliquons une translation selon le vecteur $\vec{a_0}$, de sorte que l'atome fictif soit à sa position originale, et que les distances relatives des atomes a_0, a_1, a_2 et a_3 soient toujours consistantes. Nous faisons alors appel au solveur pour résoudre les équations associées au placement des autres atomes de de la molécule. Pour chaque atome, nous sélectionnons la solution respectant au mieux la distance $d_{n-1,n}$ (REF PLACEMENT AN).

Limites de l'approche par solveur L'utilisation d'un solveur calculant les solutions au cas par cas pose deux problèmes importants. Le premier concerne les performances de la solution. En effet, la résolution des systèmes d'équations consomme beaucoup de ressources et prend donc un temps non négligeable si on souhaite appliquer la méthode à un grand nombre de molécules.
Le second problème est lié à la propagation des erreurs lors de la reconstruction (REF RECONSTRUCT TRILAT). À cause du manque de précision de certaines valeurs, certaines intersections de sphères sont vides. Le

2. https://en.wikipedia.org/wiki/Nonlinear_system
3. http://www.sympy.org/fr/

$$\begin{cases} d^2_{n-4,n} = (x_n - x_{n-4})^2 + (y_n - y_{n-4})^2 + (z_n - z_{n-4})^2 \\ d^2_{n-3,n} = (x_n - x_{n-3})^2 + (y_n - y_{n-3})^2 + (z_n - z_{n-3})^2 \\ d^2_{n-2,n} = (x_n - x_{n-2})^2 + (y_n - y_{n-2})^2 + (z_n - z_{n-2})^2 \end{cases}$$

FIGURE 3.10 – Équations de sphères permettant d'obtenir la position d'un atome quelconque de la molécule

solveur renvoie alors des solutions imaginaires que nous ne pouvons pas interpréter. Ce problème se manifeste avant tout sur les molécules de taille importante, mais il est impossible de déterminer une taille limite au delà de laquelle nous ne pouvons pas reconstruire les molécules. Cela implique qu'il existe des molécules que nous ne pouvons pas reconstruire, et que nous ne pouvons pas déterminer à l'avance si une molécule donnée peut être reconstruite.

3.2.3.3 Reconstruction automatique des positions en utilisant des équations de trilatération

Afin de pallier les problèmes liés à l'utilisation d'un solveur pour construire l'ensemble des positions des atomes d'une molécule à partir de la matrice réduite des distances inter-atomiques, nous utilisons une méthode permettant de calculer les positions de chaque point à partir d'un ensemble d'équations. Cette méthode est décrite sur Wikipédia [4]. Il s'agit d'une méthode de trilatération de points, c'est à dire que l'on cherche à déterminer la position d'un point en fonction de ses distances à trois points dont les positions sont connues, par opposition à la triangulation [5] pour laquelle on détermine la position d'un point en fonction de ses angles à des points dont les positions sont connues.

De même que pour la méthode utilisant un solveur (REF SOLV), nous commençons par placer l'atome fictif a_0 à la position qui lui a été attribuée, puis les atomes a_1, a_2 et a_3 de façon arbitraire telle que les distances relatives des atomes a_i, $i \in \{0, ..., 3\}$ soient respectées. Nous utilisons pour cela les équations décrites en (REF FIG PLACEMENT).

Une fois les quatre premiers atomes placés, nous cherchons à placer l'atome a_n de la molécule en fonction de ses distances aux quatre atomes précédents. Nous calculons les solutions en considérant que $a_{n\text{-}4}{}'$ est à l'origine du repère, que $a_{n\text{-}3}{}'$ est sur l'axe x, et que $a_{n\text{-}2}{}'$ est sur le plan tel que $z = 0$, puis nous effectuons une translation des solutions dans le système de coordonnées original. Pour cela, nous définissons les quantités et vecteurs suivants.

La notation $\hat{u}$ indique un vecteur u de norme 1, et nous considérons que $\overline{a_i}$ représente le vecteur allant de l'origine au point a_i, dans le but de simplifier l'écriture des équations.

Vecteur unitaire dans la direction de a_{n-4} à a_{n-3} :

$$\hat{e_x} = \frac{\overline{a_{n-3}} - \overline{a_{n-4}}}{d_{n-4,n-3}}$$

Ordre de grandeur signé de la composante x dans le nouveau système de coordonnées du vecteur $\overline{a_{n-4}a_{n-2}}$:

$$i = \hat{e_x} \cdot (\overline{a_{n-4}} - \overline{a_{n-2}})$$

Vecteur unitaire dans la direction y par rapport à $\hat{e_x}$:

$$\hat{e_y} = \frac{\overline{a_{n-2}} - \overline{a_{n-4}} - i\hat{e_x}}{\left\|\overline{a_{n-2}} - \overline{a_{n-4}} - i\hat{e_x}\right\|}$$

Vecteur unitaire dans la direction z par rapport à $\hat{e_x}$ et $\hat{e_y}$:

4. https://en.wikipedia.org/wiki/Trilateration
5. https://fr.wikipedia.org/wiki/Triangulation

$$\hat{e_z} = \hat{e_x} \times \hat{e_y}$$

Ordre de grandeur signé de la composante y dans le nouveau système de coordonnées du vecteur $\overline{a_{n-4}a_{n-2}}$:

$$j = \hat{e_y} \cdot (\overline{a_{n-4}} - \overline{a_{n-2}})$$

On calcule alors les deux solutions pour a'_n selon les équations suivantes.

$$a'_n \begin{cases} x'_n = \frac{d^2_{n-4,n} - d^2_{n-3,n} + d^2_{n-4,n-3}}{2d_{n-4,n-2}} \\ y'_n = \frac{d^2_{n-4,n} - d^2_{n-2,n} + i^2 + j^2}{2j} - \frac{i}{j} x'_n \\ z'_n = \pm\sqrt{d^2_{n-4,n} - x'^2_n - y'^2_n} \end{cases}$$

Enfin, nous translatons les deux solutions a'_n dans le système de coordonnées original selon le vecteur suivant.

$$\overline{p} = \overline{a_{n-4}} + x'_n\hat{e_x} + y'_n\hat{e_y} + z'_n\hat{e_z}.$$

Nous obtenons alors deux solutions a_n, et nous sélectionnons celle telle que la distance $d_{n-1,n}$ est la plus cohérente.

Performances et limites (propagation des erreurs) Les équations de trilatération permettent de calculer la matrice des coordonnées de façon très rapide. Néanmoins, de même que la méthode utilisant un solveur d'équations, cette méthode souffre d'un problème de propagation des erreurs intrinsèque à la représentation par matrice réduite des distances inter-atomiques. En effet, lorsque l'on calcule les coordonnées d'un atome à partir de ses distances aux quatre atomes précédents, et que l'on compare ces distances aux distances aux mêmes points de la position nouvellement calculée, on s'aperçoit qu'elles ne sont pas parfaitement identiques. L'erreur est très faible (de l'ordre de 10^{-25} m) et est individuellement très au-delà de la précision requise en chimie quantique (environ 10^{-15} m), mais elle finit par devenir trop importante du fait de sa propagation au fil des calculs, la position de chaque atome étant calculée à partir de ses distances aux quatre atomes précédents.
La présence d'une racine carrée dans les équations (calcul de z'_n) accélère la propagation des erreurs. En effet, après quelques itérations et quelques faibles erreurs, les intersections de sphères deviennent vides, ce qui se traduit dans nos équations par le calcul de la racine d'un nombre négatif. Pour parer cela, nous considérons que le contenu de la racine vaut zéro lorsqu'il est négatif, mais cela introduit une erreur importante et augmente donc la fréquence des intersections vides dans le calcul de la position des atomes suivants.
Afin de retarder l'apparition des erreurs dépassant le seuil toléré, nous aurions pu ajuster les valeurs de x'_n et y'_n lorsque l'on considère que le contenu de la racine est nul selon l'équation ci-dessous. Toutefois, cela n'aurait pas constitué une solution viable car le problème aurait été simplement déplacé dans le temps, l'erreur se propageant tout de même.

$$x'^2_n = d^2_{n-4,n} - y'^2_n$$

FIGURE 3.11 – Optimisation de x'_n et y'_n lorsque l'on considère que z'_n est nul

Test de la reconstruction Les tests ont montré que l'on pouvait reconstruire les positions des atomes des molécules avec cette méthode de façon fiable pour les molécules de taille inférieure ou égale à 20 atomes. La méthode de test est la suivante. On génère des coordonnées aléatoirement pour 100000 molécules de taille (nombre d'atomes) variable. On utilise la méthode de reconstruction puis on calcule la nouvelle matrice réduite des

distances inter-atomiques sur les nouvelles positions. On fait la différence des deux matrices de distances pour obtenir les erreurs et on considère que la reconstruction a été un succès si aucune composante de la matrice des erreurs n'est supérieure à un seuil. Ce seuil est choisi à 10^{-15} m, car il correspond à la précision au delà de laquelle les chimistes considèrent qu'il ne s'agit plus d'information mais de bruit. Les résultats sont données dans le tableau suivant.

Taille des molécules	10	15	20	25	30	35	40
Molécules mal reconstruites	0	0	0	18	132	468	1752

FIGURE 3.12 – Test de la méthode de reconstruction pour la matrice des distances inter-atomiques

3.3 Matrice des distances à des points fixes

3.3.1 Motivation

La matrice des distances à des points fixes a pour objectif de corriger les défauts de la représentation géométrique moléculaire par matrice réduite des distances inter-atomiques (REF MATR RED DIST REL). Cette dernière possédait en effet le défaut majeur de ne pas être systématiquement réversible en matrice des coordonnées atomiques (REF REPR MAT COORDS). Ce défaut était dû à la propagation des erreurs induite par le fait que les positions des atomes étaient calculées à partir du calcul de la position des atomes précédents (REF REPR DIST REL RECONSTRUCT). Pour parer cela, nous définissons une représentation telle que la position de chaque atome est définie à partir de distances à quatre points fixes du repère. Les erreurs, même si elles existent toujours à des valeurs minimes (autour de 10^{-25} m), ne se propagent donc plus lors de la reconstruction des positions des atomes.
Un autre problème résolu par cette nouvelle représentation est qu'il n'existe plus de molécule dont on ne peut pas reconstruire les positions à cause d'une géométrie plane ou linéaire (REF AT FICTIF), le calcul de la position de chaque atome dépendant désormais de la distance à quatre points de l'espace que l'on choisit tels qu'ils n'appartiennent pas à un même plan.

3.3.2 Formalisation

Formellement, la matrice contient donc les distances de chaque atome d'une molécule à quatre points fixes du repère. Nous choisissons arbitrairement comme points l'origine du repère, et le point sur chaque axe de distance 1 à l'origine. Ce choix est justifié par le fait que les points ont une distance à l'origine du même ordre de grandeur que les coordonnées des atomes dans les données (10^0 à 10^1). Cela permet donc d'avoir suffisamment d'information pour calculer la position des atomes avec une précision suffisante lors de la reconstruction de la matrice des coordonnées atomiques.

$$p_0(0,0,0) \quad p_1(1,0,0) \quad p_2(0,1,0) \quad p_3(0,0,1)$$

FIGURE 3.13 – Points fixes

$\mathbf{d_{a_0,p_0}}$	$\mathbf{d_{a_0,p_1}}$	$\mathbf{d_{a_0,p_2}}$	$\mathbf{d_{a_0,p_3}}$
$\mathbf{d_{a_1,p_0}}$	$\mathbf{d_{a_1,p_1}}$	$\mathbf{d_{a_1,p_2}}$	$\mathbf{d_{a_1,p_3}}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\mathbf{d_{a_n,p_0}}$	$\mathbf{d_{a_n,p_1}}$	$\mathbf{d_{a_n,p_2}}$	$\mathbf{d_{a_n,p_3}}$

FIGURE 3.14 – Matrice des distances à des points fixes (molécule de taille n)

3.3.3 Reconstruction des molécules

De même que pour la représentation par matrice réduite des distances inter-atomiques (REF MAT DIST REL), nous devons être capables de passer d'une matrice des distances à des points fixes à une matrice des coordonnées atomiques, afin que les résultats des modèles prédictifs puissent être utilisés par des chimistes.

La méthode de reconstruction des positions atomiques est très similaire pour les deux représentations. Nous utilisons également les équations de trilatération d'un point à partir des distances à trois points dont les positions sont connues, en utilisant la dernière distance comme un moyen de choisir la bonne solution (voir REF RECONSTRUCT MAT DIST REL). Du fait que la position des quatre points de référence soit fixe et qu'ils suivent les contraintes que nous imposions lors de la translation dans un système de coordonnées plus simple, les équations se trouvent néanmoins simplifiées. En effet, p_0 est à l'origine du repère, p_1 est sur l'axe x et p_2 est sur le plan tel que $z = 0$. Pour rappel, nous résolvons le problème de placement de point dans le système de coordonnées simplifié, puis nous effectuons une translation des solutions dans le système de coordonnées original. Or, nos points de référence se trouvent être les vecteurs unitaires dans chaque direction des deux systèmes de coordonnées. Nous obtenons donc directement les solutions dans le système de coordonnées original.
La méthode complète est décrite sur Wikipedia[6]. Nous en extrayons les équations suivantes pour le placement général d'un atome d'une molécule.

$$a_n \begin{cases} x_n = \frac{d^2_{a_n,p_0} - d^2_{a_n,p_1} + 1}{2} \\ y_n = \frac{d^2_{a_n,p_0} - d^2_{a_n,p_2} + 1}{2} \\ z_n = \pm\sqrt{d^2_{a_n,p_0} - x'^2_n - y'^2_n} \end{cases}$$

Nous obtenons alors deux solutions a_n, et nous sélectionnons celle telle que la distance d_{a_n,p_3} est la plus cohérente.

3.4 Représentation locale des liaisons covalentes

3.4.1 Motivation

Cette représentation géométrique s'éloigne des représentations précédentes pour plusieurs raisons. Premièrement, elle s'inscrit dans l'idée de formuler des problèmes plus simples (REF MOD DIST REL) suite à l'échec des modèles utilisant les représentations précédentes (REF MOD DELTA DIST). Pour cette raison, nous n'allons plus chercher à représenter des molécules complètes mais uniquement des liaisons covalentes[7] entre des paires d'atomes au sein des molécules. Cette représentation doit contenir des informations permettant aux modèles l'utilisant de prédire la longueur de la liaison représentée, sans bien-sûr l'enregistrer directement.
En second lieu, la contrainte majeure de la nécessité d'être capable de reconstruire la matrice des coordonnées atomiques à l'issue des prédictions des modèles utilisant cette représentation disparaît. En effet, si l'on peut imaginer des représentations similaires (REF REPR ANGLES) et un assemblage de modèles (REF MODULES) qui permettraient de reconstruire la matrice de coordonnées atomiques d'une molécule convergée (REF DEF CONVERG), il s'agit d'objectifs hors de notre portée pour le moment, notre objectif étant dans un premier temps de valider notre capacité à prédire des géométries moléculaires.

3.4.2 Classes positionnelles

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, qui dépend de leur position relativement aux atomes de la liaison. C'est pour cette raison que nous formalisons 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.

6. https://en.wikipedia.org/wiki/Trilateration
7. Une liaison covalente est une liaison chimique dans laquelle deux atomes se partagent deux électrons (un électron chacun ou deux électrons venant du même atome) d'une de leurs couches externes afin de former un doublet d'électrons liant les deux atomes. (Wikipédia)

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.

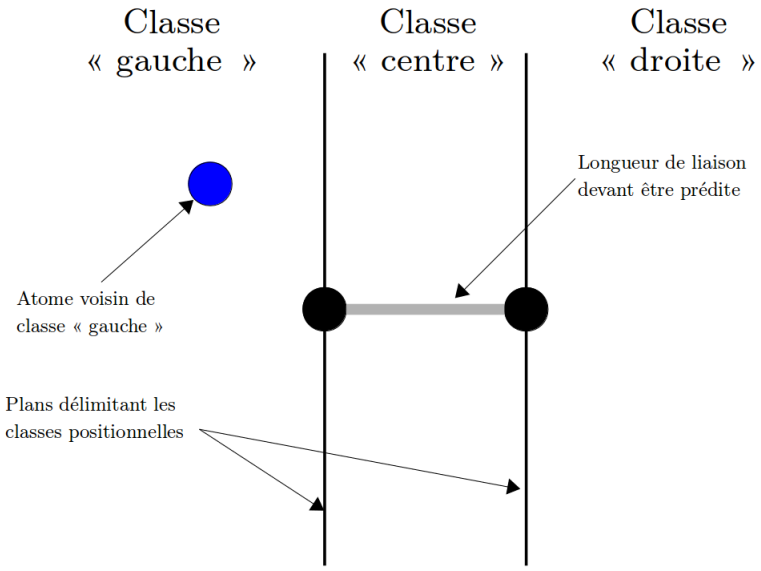

FIGURE 3.15 – Classes positionnelles au voisinage d'une liaison covalente

3.4.3 Distances aux atomes de la liaison

L'influence des atomes au voisinage de la liaison dépend également de leur distance à chacun des deux atomes de la liaison. Plus l'atome voisin est près, plus son influence est forte. C'est pourquoi notre représentation contient également cette information.
En fonction des modèles qui l'utilisent, on va éventuellement appliquer une fonction à ces distances, afin de mieux rendre compte de l'influence réelle des atomes au voisinage. Si les réseaux de neurones sont capables d'approximer ces fonctions lors de l'apprentissage, d'autres modèles comme les SVM (REF SVM) ne le sont pas et l'application de ces fonctions est donc nécessaire pour espérer obtenir de bons résultats. Ces fonctions sont les suivantes.

- Fonction identité : distance brute
- Fonction inverse : influence inversement proportionnelle à la distance, relation d'ordre identique à la réalité chimique.
- Fonction inverse du carré : influence inversement proportionnelle au carré de la distance, relation d'ordre identique à la réalité chimique et rend mieux compte de l'influence réelle des atomes qui suit la loi de Coulomb [8] en $\frac{1}{d^2}$.

3.4.4 Sélection de l'information la plus pertinente

L'influence des atomes au voisinage de la liaison étant proportionnelle à l'inverse du carré de leurs distances aux atomes de la liaison, elle décroît rapidement lorsque l'on s'éloigne de la liaison. L'influence des atomes n'étant pas au voisinage direct est ainsi négligeable. Dans le but de ne pas saturer l'entrée des modèles d'information inutile, nous n'enregistrons alors que les informations (classes positionnelles, distances et autres informations non géométriques spécifiques aux différents modèles) concernant les atomes au voisinage proche de la liaison.

8. https://fr.wikipedia.org/wiki/Loi_de_Coulomb_(électrostatique)

Formellement, nous enregistrons ces informations pour les atomes dont la distance à au moins un des atomes de la liaison est inférieure à un seuil ϵ donné.

Un autre avantage de cette sélection est qu'il existe des molécules aux géométries particulières (repliées) telles que des atomes au voisinage d'une liaison ont très peu d'influence sur sa longueur (ne forment aucune liaison covalente avec les deux atomes de la liaison), et dont la proximité va induire les modèles en erreur. La sélection des atomes au voisinage le plus proche de la liaison avec un seuil ϵ bien choisi va permettre de résoudre ces problèmes.

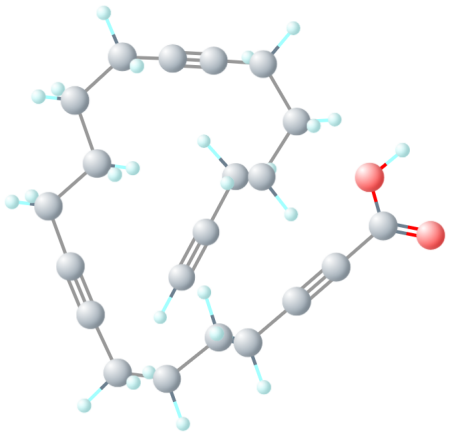

FIGURE 3.16 – Exemple de molécule repliée (CID Pubchem 328310)

Chapitre 4

Donnees

Chapitre 5

Prédiction de longueurs de liaisons convergées

Chapitre 6

Prédiction de géométries moléculaires convergées

6.1 Motivation et méthodologie

L'objectif des modèles prédictifs que l'on décrit dans ce chapitre est de prédire la géométrie convergée (REF GEOM CONVERG) d'une molécule complète, à partir d'une géométrie non convergée. Ils sont issus d'une tentative de reproduction de résultats antérieurs, afin de confirmer la méthode élaborée lors des stages précédents sur le projet QuChemPedIA.
Chronologiquement, ces modèles ont constitué la première partie de mon travail, avant de passer aux modèles tentant de prédire les longueurs de liaisons (REF DIST_REL), à cause de l'impossibilité de produire des prédictions de qualité suffisante (REF RESULTATS).

L'objectif à terme de ces modèles est de pouvoir constituer une alternative au DFT (REF DFT) pour calculer rapidement la géométrie convergée d'une molécule. Cela nécessite de produire des prédictions d'une très grande précision. Cependant, le but ici est avant tout de valider une méthode et notre capacité à produire des prédictions d'ordre géométrique. Nous ne cherchons donc pas à créer un modèle effectuant de très bonnes prédictions, mais plutôt à définir une représentation des données et un ensemble de paramètres permettant d'obtenir de bons résultats.

Introduction de bruit Afin de prédire des géométries moléculaires convergées 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 d'optimisation par mécanique moléculaire (REF MÉCA MOL), et l'ensemble de géométries convergées par le DFT (REF DFT) associé. Cela constituerait en effet un ensemble de données homogène qui aurait l'avantage d'être comparable aux données que l'on utiliserait dans un cas d'utilisation réel. Malheureusement, nous ne possédons pas de telles données. Nous possédons les géométries convergées issues de la base PubChemQC (REF PUBCHEMQC) mais pas les géométries à partir desquelles elles ont été calculées. S'il est théoriquement possible de calculer la géométrie optimisée en mécanique moléculaire de toutes les molécules de la base PubChemQC en utilisant le programme Open Babel [1], la perte de l'ordre des atomes lors de l'optimisation rend la procédure impossible en pratique.
L'alternative retenue lors des stages précédents est d'introduire du bruit (REF PREP DONNEES BRUIT) dans les coordonnées des géométries optimisées, et d'entraîner les modèles à prédire ce bruit. La différence entre la géométrie bruitée et le bruit prédit permet alors d'obtenir la géométrie optimisée par le modèle. L'introduction de bruit ne garantit donc pas que les modèles se généraliseront aux données réelles, mais semble tout de même raisonnable pour tenter de valider la méthode, puisque nous entraînons des modèles dont l'objectif est de déplacer les atomes d'une molécule de sorte à obtenir une géométrie convergée.

Modèles Cinq modèles différents ont été entraînés. Ils diffèrent par les représentations utilisées en entrée et en sortie, les caractéristiques des molécules dont on tente de prédire la géométrie convergée, et les paramètres

1. http://openbabel.org/wiki/Main_Page

propres aux réseaux de neurones comme les fonctions de coût ou la topologie. Nous allons répertorier les différentes caractéristiques utilisées mais pas modèle par modèle, puisque aucun ensemble de caractéristiques n'a produit de résultats significativement meilleurs que les autres (REF RESULTATS). Cependant, une table des caractéristiques utilisées modèle par modèle est disponible en annexe (REF CARAC ANNEXES).

6.2 Données

6.3 Évaluation des performances

6.4 Architectures

6.5 Optimisation des modèles

6.6 Performances

Chapitre 7

Conclusion

Annexes

Modèle	Tailles molécules	Repr. géom. entrée	Repr. géom. sortie	Numéros atomiques	Masses atomiques	Distances inter at. fictifs	Fonction de coût	Profondeur	Largeur	Taille entrée	Bruit
DELTA_DIST+H_01	0 - ∞	Matr. dist. pts. fixes	Matr. dist. pts. fixes	Non	Oui	Non	RMSE partiel/total	4	8650	1000	+/++
DELTA_DIST+H_02	0 - ∞	Matr. dist. pts. fixes	Matr. dist. pts. fixes	Non	Oui	Oui	RMSE partiel	4	8650	1020	+
DELTA_DIST+H_03	0 - ∞	Matr. dist. pts. fixes + Matr red. dist. inter-at.	Matr. dist. pts. fixes	Non	Oui	Non	RMSE partiel	3	9000	1800	++
DELTA_DIST+H_04	0 - ∞	Matr. dist. pts. fixes + Matr red. dist. inter-at.	Matr red. dist. inter-at.	Non	Oui	Non	RMSE partiel	3	9000	1800	++
DELTA_DIST+H_05	2 - 60	Matr. dist. pts. fixes	Matr. dist. pts. fixes	Oui	Oui	Non	RMSE partiel	3	360	360	++