Rapport de stage de Master 1 Physique fondamentale

Présentation des travaux de Yanis Datti et Alban Degezelle

Construction et benchmarking d'une base de données de polarisabilités atomiques

Sous la direction de Pr Benoit Guillot, Professeur des Universités à l'Université de Lorraine et de Monsieur Théo Leduc, post-doctorant au laboratoire de Cristallographie, Résonance Magnétique et Modélisations (CRM2).

TABLE DES MATIÈRES

Avant-propos

Abstract

L'études des biomolécules et de leurs propriétés est un domaine phare des sciences physiques et biologiques. Les fonctions qu'assurent ces dernières dans les cellules sont d'une importance capitale sur le plan de la santé, de nouveaux défis en pharmacologie ou en biologie structurale reposent sur la connaissance exacte du fonctionnement de ces « ouvriers du vivant ». Une de ces propriétés importantes est la polarisabilité, qui résulte de l'interaction entre un champ électrique et les porteurs de charge dans un milieu. Une méthode développée par le CRM^2 alliant calculs théoriques et données expérimentales transférées permet d'analyser ces polarisations de manière inédite. Notre objectif ici est de valider cette modélisation en la confrontant à des résultats purement théoriques.

Abstract (EN)

The study of biomolecules and their properties is a key domain of biologic and physic sciences. The functions that they fulfil in cells are of crucial importance in the health front, new challenges in pharmacology or structural biology are based on precise knowledges of the functioning of these molecules. One of these important properties is the polarization, which describes the result of the interaction between an electric field and charge carriers in an environment. A method developed in the CRM^2 laboratory combining theorical calculations and transferred experimental data allows to analyze these polarizations in a unique way. Our goal here is to validate this modelling by confronting it with purely theorical results.

Etat de l'art

L'équipe BioMod du laboratoire CRM2 est spécialisée dans les méthodes d'études de bio-structures (RMN, Diffraction Rayon-X, Dynamique moléculaire). Avec ces outils, ils ont commencé le développement d'une nouvelle méthode originale, basée sur les densités de charge multipolaires et la cristallographie haute résolution, pour avoir accès aux interactions électrostatiques précises des molécules.

La méthode a fait l'objet de nombreuses publications et de thèses par le laboratoire au cours de ces dernières années. Notre participation intervient dans la continuité de ces travaux, dont la thèse de Theo Leduc soutenue en 2019[1]. L'étude des propriétés moléculaires est un atout considérable pour aider à les comprendre et les utiliser. Dans cette démarche, l'utilisation de la diffraction X est particulièrement adaptée pour aller étudier les propriétés de la matière cristalline. Par définition un cristal est un assemblage périodique de ses constituants (atomes, molécules, ions). Cependant l'application de ces méthodes aux biomolécules pose un problème, : les cristaux de protéines sont en général beaucoup moins réguliers et organisés que les cristaux de « petites molécules ». Les données obtenues par diffraction par rayons X par exemple, sont de nature beaucoup moins précises que dans le cas de petites molécules. Cette précision sur les données de diffraction X va se répercuter sur la qualité des modélisations des densités de charges, donnée centrale de la méthode.

L'approche développée par nos encadrants Théo Leduc et Benoit Guillot ainsi que leurs collègues est la suivante. Des densités électroniques sur des petites molécules sont obtenues par des calculs théoriques, donnant aussi accès à des polarisabilités. De ça, il est possible d'extraire des données de polarisabilités atomiques et moléculaires brutes. C'est là que la méthode devient hybride, car en effet, nous allons créer avec ces polarisabilités atomiques une base de données de polarisabilités moyennes et transférables, qui vont être ajoutées aux densités électroniques

décrites dans une librairie de paramètres atomiques également transférables déjà développée au CRM2. En appliquant le principe de transférabilité (cf. : partie Modélisation), on obtiendra donc des densités électroniques polarisées et transférées. On peut donc en déduire les énergies de polarisation correspondantes et les confronter aux valeurs obtenues indépendamment par des méthodes théoriques.

L'étape traitée durant ce stage concerne principalement la construction et la validation, à partir de données théoriques, d'un ensemble de polarisabilités atomiques compatibles avec l'hypothèse de transférabilité avec la librairie ELMAM2.

La polarisabilité traduit la propension à voir les charges se réorganiser sous l'action d'un champ électrique externe, ce qui va entrainer un moment dipolaire induit. Ce moment dipolaire est lié au champ électrique externe du milieu et à la polarisabilité dans une relation de nature tensorielle étant donné que la polarisabilité d'une distribution de charge est en général anisotrope. Ce champ électrique peut être calculé en utilisant le Modèle Multipolaire de Hansen et Coppens[2] (abrégé HCMM en anglais). Le modèle HCM décrit la densité électronique par un ensemble de paramètres de populations électroniques associées à des fonctions harmoniques sphériques.

Les énergies de polarisation obtenues devront reproduire ou du moins grandement approcher celles qui sont calculées théoriquement, par d'autres méthodes. Les sets de polarisabilités que l'on aura mis au point, associés à la densité électronique multipolaire transférée, nous permettrons de calculer avec précision des énergies de polarisation dans des dimères de petites molécules organiques. Pour confronter nos résultats on prendra comme référence la méthode SAPT[3] (*Symmetry Adapted Perturbation Theory*). Celle-ci consiste à différencier l'énergie d'interaction intermoléculaire totale en somme de plusieurs contributions dont une de polarisation. On réalisera les tests sur le Dataset S66[4,5]

Modélisation

La modélisation des densités électroniques via le principe de transférabilité se fait à l'aide du logiciel MoProViewer et la base de données ELMAM2. ELMAM2[6] est une banque de données de densités électroniques de nombreux groupes fonctionnels pouvant être trouvés dans les molécules biologiques. Ces données ont été obtenues via des méthodes expérimentales rigoureuses telles que la diffraction des rayons X à haute résolution. Ces informations ont été affinées (contre les données de diffraction) dans le modèle multipolaire d'Hansen & Coppens. Ce modèle consiste à décrire la densité électronique d'un atome comme la somme d'une contribution de ses électrons de valence, des électrons de cœur et de fonctions de types harmoniques sphériques représentant le caractère anisotrope de la densité de valence, lorsque l'atome forme des liaisons dans une molécule (annexe n°3). Ces fonctions harmoniques sphériques sont orientées en 3D grâce à un repère orthonormé local, centré sur le noyau de l'atome considéré.

L'objectif est ensuite de trier ces données en fonction des similitudes trouvées dans les diverses molécules. En effet, pour un atome dit « central » donné, on peut déterminer différents critères permettant de le caractériser plus précisément et le mettre en évidence dans différentes molécules. Ces critères sont notamment l'élément chimique, son hybridation et le nombre et le type de premiers voisins covalents. Ainsi, pour toutes les données de la bibliothèque ELMAM2[6], les densités électroniques de chaque atome partageant ces mêmes caractéristiques ont été regroupées et moyennées. Au total 68 types d'atomes associé à des paramètres de densité

électronique ont été créés grâce à cette approche. Il est important de noter que cette approche se base donc uniquement sur l'environnement covalent de l'atome étudié, la densité électronique qui en ressort n'est donc pas polarisée par les interactions intermoléculaires. Ces densités électroniques obtenues, l'objectif sera de les associer aux polarisabilités théoriques correspondantes calculées en amont afin d'avoir cette fois ci les densités électroniques polarisées. La méthode lie donc à la fois les données théoriques obtenues via calculs quantiques de polarisabilités et les résultats expérimentaux de densités électroniques présents dans la base ELMAM2.

Ces données sont ensuite utilisées par MoProViewer afin de représenter les densités électroniques d'une molécule choisie grâce au principe de transférabilité. Les atomes de la molécule sont comparés aux 68 groupes présents dans la banque de données, en utilisant les mêmes critères que leur création, et s'il y'a correspondance, alors on lui associe la densité électronique obtenue au préalable. Ainsi, on peut représenter un modèle précis de densité électrique d'une molécule sans passer par de longs calculs.

Validation

Une fois obtenues, les énergies de polarisations transférées vont être comparées à des énergies de polarisation via la méthode théoriques SAPT. De ce fait, si les énergies de polarisation obtenues entre le principe de transférabilités et celles théorique correspondent, cela permettrait de valider l'utilisation d'une telle méthode, en incluant des effets de polarisation. Ceci permettra donc de pouvoir utiliser une banque de données pour décrire les densités électroniques des biomolécules, plutôt que des valeurs théoriques, ou alors quand des techniques telles que la diffraction à rayon X à haute résolution ne sont pas adaptées.

Contexte bibliographique

Le contexte du stage s'appuie essentiellement sur les travaux effectués par Théo Leduc durant sa thèse[1] rapportée en décembre 2019. En amont du stage nous avons donc effectué des recherches bibliographiques sur le sujet avec l'aide de nos encadrants. En plus de la thèse de Mr Leduc, la source d'information principale sur laquelle nous nous sommes appuyés est l'article de 2019[7] publié dans Journal of Physical Chemistry A par l'équipe BioMod, dans lequel on présente la méthode d'obtention des polarisabilités atomiques associés au modèle multipolaire. Les résultats présentés dans ce dernier sont ceux de la création et de l'application de la méthode décrite précédemment au dataset S66. Ce set de données de dimères de molécules organiques (avec différentes distances d'interaction) a été créé par Jan Rez, Kevin E. Riley et Pavel Hobza. Ce dernier a fait l'objet de deux publications[4,5].

Mode de lecture du rapport

Tout au long du stage, deux fronts de travail avançaient en parallèle : d'un côté la partie physique et interprétation et de l'autre la partie programmation. A chaque étape de ce rapport correspond donc une partie du code *MoProFileWorker*. Nous avons rédigé le présent travail de manière à rendre optionnelle la lecture du code pour la compréhension, néanmoins à chaque étape importante du rapport, nous ferons référence à sa partie correspondante dans le code entre parenthèses.

Exemple : la méthode est automatisée pour les 12 configurations possibles (**I.**). La parenthèse fait référence aux parties du code décrit dans son petit manuel de lecture en annexe.

1. Premier traitement : Créations des repères locaux et des tenseurs de polarisabilité.

On l'a vu précédemment, la transférabilité d'une propriété donnée est définie comme la conservation de cette propriété entre deux objets similaires dans des systèmes différents (ici, la densité électronique). La construction de polarisabilités atomiques moyennes va être le premier pas vers la vérification de cette transférabilité des densités électroniques sur les polarisabilités moléculaires. La description des tenseurs de polarisabilité atomiques dans un repère locale définis de manière à faciliter sa représentation fait sens ici. En effet l'objectif est ici de créer, d'après les données fournies qui sont un ensemble de fichiers moléculaires au format *MoProViewer* appelé le « *BigSet* » une base de données de polarisabilités atomiques moyennes.

Le *BigSet* est constitué de 168 fichiers moléculaires, parmi eux, sont présents 161 molécules 104 structures de mono-, di- et tri-peptides, 57 dimères non-covalents et 7 monomères. Cependant, l'obtention des polarisabilités de chaque atome de ces systèmes moléculaires s'est faite de deux façons différentes. La première est le « S66 », il contient 57 dimères et les 7 monomères, les polarisabilités ont été obtenu directement au *CRM²*, plus particulièrement par Emmanuel Aubert avec des calculs DFT *(Density functional theory*). La seconde source concerne le reste des 104 dimères. Leurs polarisabilités atomiques ont été calculées par des collaborateurs Polonais du laboratoire, à savoir la professeure Anna Krawczuk et la docteure Katarzyna Rzesikowska, à l'aide du logiciel Polaber. Par la suite, on appellera le premier « S66 » et le second « SAK ».

Les méthodes de calculs entre ces deux sources diffèrent principalement des fonctions de base utilisées pour décrire les fonctions d'onde lors des calculs DFT. Le S66 utilise des fonctions gaussiennes tandis que le set SAK utilise des fonctions de type Slater, impliquant un temps de calcul supérieur. Une seconde différence réside dans le calcul des polarisabilités pour le S66. En effet, le terme de transfert de charge, décrivant les polarisabilités de liaison après application d'un champ électrique, a été négligé quand le champ appliqué était très faible.

Cette première partie se consacre au traitement primaire du jeu de données avec notamment la création des repères locaux et des tenseurs de polarisabilité et ainsi créer un fichier de sortie regroupant toutes les informations utiles dans l'interprétation pour chaque atome du set. Les données contenues dans ce fichier nous serviront pour tout le reste du présent travail.

Pour bien comprendre le vocabulaire utilisé au long de ce rapport il est nécessaire de comprendre le format des fichiers moléculaires ainsi que les informations qu'ils contiennent. Une description de ces derniers se trouve en annexe.

1.1 Création des repères locaux pour chaque atome.

On construit les repères locaux en se basant sur les repères de définition. On à 12 configurations possibles pour ces derniers : XY XZ YZ YX ZX ZY et leur homologue dans le cas avec bissectrice nommés bXY bXZ bYZ bYX bZX bZY. Le choix de ces configurations a été fait par l'équipe du CRM2 afin de symétriser au maximum les représentations de densités électroniques de la librairie ELMAM2 au sein des molécules.

Illustration des deux cas avec (figure 2) et sans (figure 1) bissectrice sur une molécule d'eau :

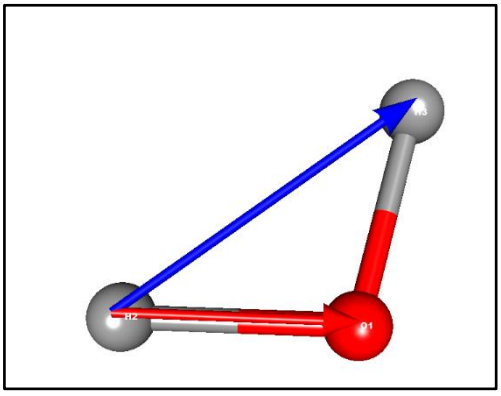


Figure 1 : Molécule d'eau et repère de définition de l'atome H2.

Repère de définition de l'atome H2, son premier voisin est celui de la liaison covalente directe avec l'oxygène (flèche rouge). Son deuxième voisin H3 définit l'orientation du deuxième axe du repère locale. Ici l'atome H2 a une configuration ZX, son premier voisin définit donc u_z comme premier axe du repère cartésien local le long de la liaison. A partir de u_z et de la direction du deuxième voisin (flèche bleu) on construit le deuxième axe perpendiculaire à u_z dans le plan de la molécule.

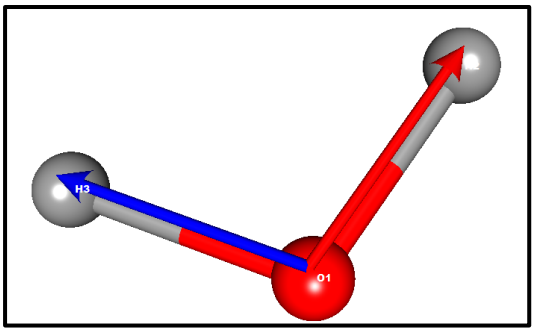


Figure 2 : Molécule d'eau et repère de définition de l'atome O1.

Repère de définition de l'atome O1, ses deux voisins sont les atomes d'hydrogènes. Sa configuration bXY définit le premier axe du repère local u_x comme la bissectrice de l'angle entre les deux vecteurs dir1 et dir2 ci-contre. Ainsi on construit de la même manière l'axe u_y orthogonal à u_x dans la direction du deuxième voisin et dans le plan de la molécule. Enfin, u_z sera obtenu par produit vectoriel entre u_x et u_y.

Le repère orthonormé est représenté figure 3.

Création des repères locaux, utilisation de l'algorithme de Gram-Schmidt

Pour chaque atome on accède aux positions de l'atome considéré ainsi que celles de ses deux voisins. On va ainsi utiliser le procédé de Gram-Schmidt pour construire une base orthonormale selon la configuration indiquée pour l'atome dans les fichiers moléculaires.

$$proj_u(v) = \frac{\vec{u}.\vec{v}}{\vec{u}.\vec{u}}\vec{u}$$

$$\vec{u}_1 = \vec{v}_1$$

$$\vec{u}_2 = \vec{v}_2 - proj_{u_1}(\vec{v_2})$$

$$\vec{u_3} = \vec{u_1} \times \vec{u_2}$$

Puis normalisation à l'unité : $\vec{e_i} = \frac{\vec{u_i}}{\|u_i\|}$, i =1,2,3

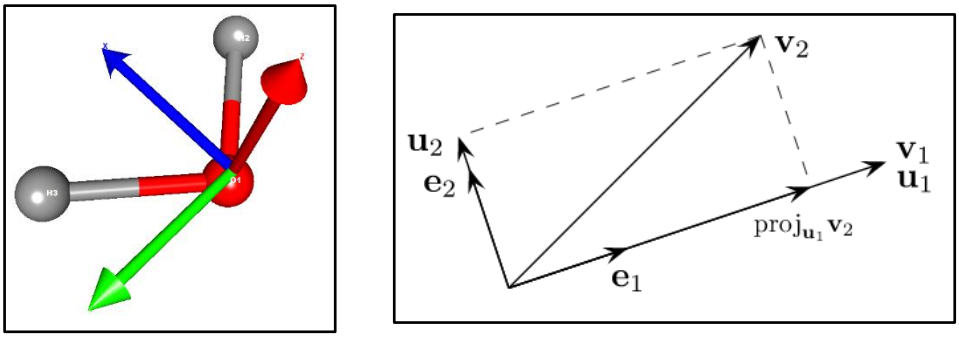


Figure 3 : A gauche, le repère orthonormé construit de l'atome O1 discuté précédemment. A droite, la visualisation des deux premières étapes de l'algorithme de Gram-Schmidt

1.2 Transformation des tenseurs de polarisabilité.

Les tenseurs de polarisabilité de chaque atome présent dans les fichiers moléculaires du jeu de données sont au départ exprimé dans le repère cartésien de la molécule. On applique un changement de base pour exprimer ces derniers dans leur repère local.

$$T_{local} = P^{-1} T_{molécule} P \qquad (1)$$

Ou P est la matrice de passage de la base de la molécule vers la base locale et T le tenseur de polarisabilité de l'atome. La représentation de ces tenseurs de rang 2 dans une base orthonormée est un ellipsoïde, *MoProViewer* permet de les représenter (figure 4). Nous avons donc créé un code pour automatiser la méthode de création de repère locaux (**I. 1,2**) et de changement de base (**I. 3,4,5**) des 4366 atomes du *BigSet*. Un fichier de sortie du code regroupe toutes les informations pour chaque atome du set (**I. 6**). Les informations contenues dans ce fichier sont la base de travail de tout ce qui suit.

A partir de ce fichier on construit une bibliothèque de type atomique en nous appuyant sur celles déjà crées auparavant par le CRM2.

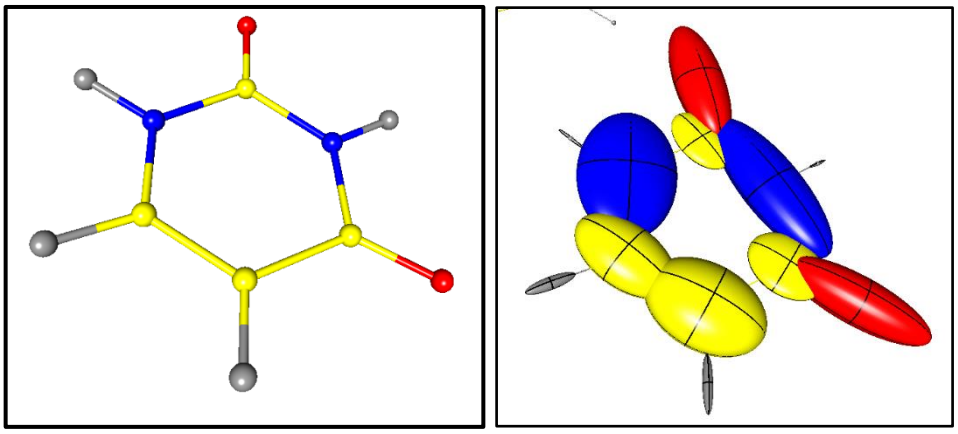


Figure 4 : Molécule d'uracile avec et sans ellipsoïde de polarisabilité

2. Détermination des types atomiques

Le principe de transférabilité repose sur le fait que l'on peut considérer, dans la limite d'une tolérance, des atomes comme semblables du point de vue d'une propriété conservée : (ici la densité électronique). On regroupe les atomes satisfaisant cette condition sous un même type atomique. Pour les atomes, la librairie ELMAM2[6] construite par le CRM2 contient l'ensemble des types atomiques présents dans cristaux qui ont permis sa construction, c'est-à-dire les types d'atomes rencontrés dans les petites biomolécules et les molécules organiques simples. Notre démarche de travail est similaire : elle se base sur les fichiers moléculaires du *BigSet*, la création des types dans le présent travail repose donc sur des informations contenues dans ces derniers. Ceci forcera quelques approximations par la suite lors de la définition des types atomiques, avec de possibles conséquences sur les polarisabilités moléculaires calculées.

2.2 Etude statistique des polarisabilités des types atomiques

La méthode choisie porte sur l'étude statistique des trois valeurs propres des tenseurs de polarisabilité. Pour chaque type atomique on va représenter la dispersion des ses trois valeurs propres. On s'attend à une distribution proche d'une distribution normale pour les types atomiques avec un nombre d'occurrence dans le *BigSet* assez élevé. Pour le vérifier, les dispersions du spectre des valeurs propres sont soumises à un test d'Agostino K^2(annexe n°4).

Les résultats montrent que très peu de types atomiques obéissent à une loi de distribution normale et ce malgré de très fortes occurrences dans certains cas. Après représentation (**II. 4**) un problème est clairement apparu. La distribution des valeurs propres semble présenter un **caractère bimodal**, ce qui correspond à deux populations de tenseurs de polarisabilités pour le même type atomique.

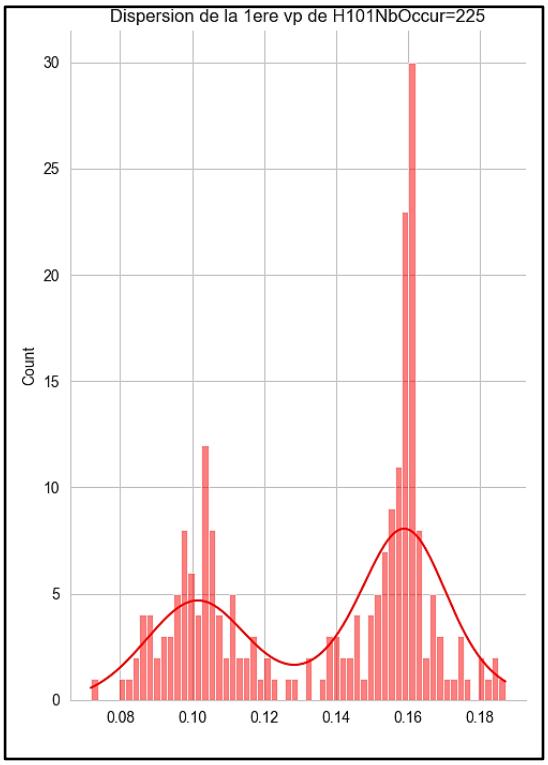


Figure 5 : Dispersion de la première valeur propre du tenseur du type H101 sur l'ensemble du BigSet (en Å³)

On observe figure 5, la dispersion de la première valeur propre du type **H101**, (hydrogène lié à un carbone sp^2, lui-même lié à deux autres atomes de carbone) contenu 225 fois dans le *BigSet*. Au premier abord, nous pensions avoir rassemblé deux types atomiques hydrogènes différents en un seul. Nous sommes donc revenus au fichier original contenant notre tri des différentes populations possibles dans le type H101 et avons observés ces atomes avec *MoProViewer*. Aucune différence notoire n'est ressortie. En observant, pour les atomes dont la première valeur propre de son tenseur de polarisabilités se situe dans chacun des pics d'occurrence, on observe que les atomes constituant un pic d'occurrence sont contenus dans le S66 et les autres correspondant au deuxième pic contenues dans les molécules du SAK.

Pour vérifier si l'origine de ce caractère bimodal est bel et bien dû aux origines différentes des fichiers présents dans le *BigSet*, on va de nouveau effectuer le procédé de création de type atomique comme vu précédemment ainsi que de nouveau analyser la dispersion des valeurs propres de chaque type mais cette fois en **traitant séparément** les fichiers obtenus par Anna Krawczuk et les fichiers du S66.

Les résultats dans le cas du type H101, sont visibles Figure 6.

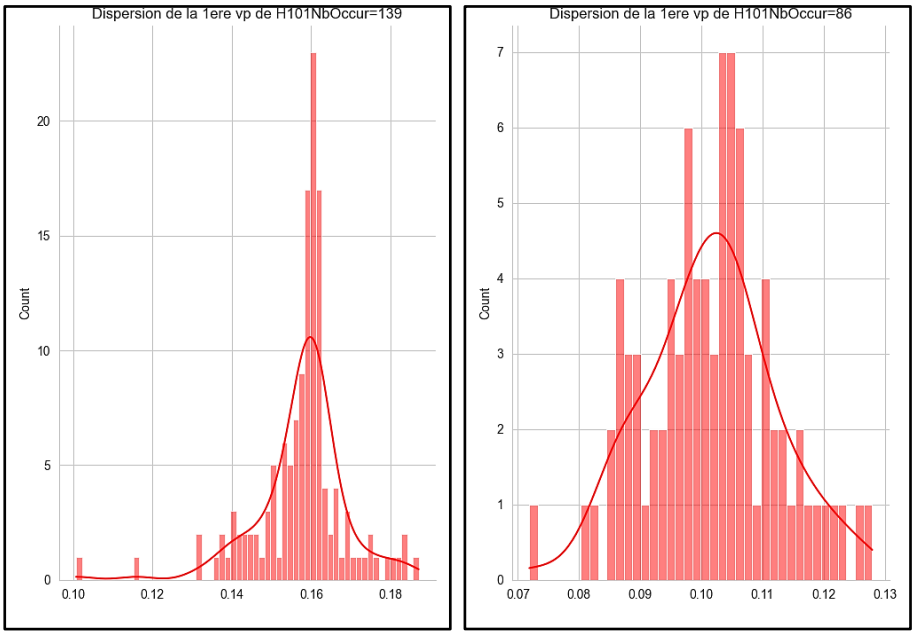


Figure 6 : Dispersion de la première valeur propre du type H101 pour le jeu de données S66 (à gauche) et pour SAK (à droite)

On observe bel et bien que les deux histogrammes de la figure 6 constituent les pics d'occurrence de la figure 5. Le test d'Agostino effectué sur chaque dispersion de valeur propre pour les deux cas différents se révèle cette fois vrai pour la majeure partie des types atomiques dont les occurrences sont assez élevées.

Nous sommes donc assez confiants sur le fait que le caractère bimodal de ces distributions ne provient pas d'un tri des types atomiques erroné mais du fait que pour un même type atomique, deux méthodes de calculs des tenseurs de polarisabilités ont été utilisées sur deux parties des données initiales.

Si l'on choisit de calculer les tenseurs moyens sur l'ensemble du *BigSet* en ne prenant pas en compte le caractère bimodal des polarisabilités de certains types d'atomes, nous risquons de faire une erreur qui va se cumuler de manière additive sur la polarisabilité moléculaire, du

fait de la propriété d'additivité des tenseurs de polarisabilité. Reprenons le cas du type H101, on remarque que l'écart entre les deux pics du spectre de la première valeur propre est presque de 50% (Figure 5) . Ce qui peut être significatif si une molécule contient un grand nombre d'hydrogène de ce type.

Selon nous trois approches sont à effectuer et à comparer :

- Méthode n°1 : La première serait de calculer les polarisabilités moléculaires sur les deux parties des données séparément, ainsi on aurait aucun souci de distributions de valeurs propres bimodales et les tenseurs moyens d'un même type seraient pertinents. Cependant cela ne nous permettrait pas de comparer les différentes méthodes conçues pour obtenir les fichiers d'origines à moins que dans le S66 et le SAK on retrouve la même molécule (On pourrait tenter de se pencher sur la polarisabilité de certains groupement chimiques communs aux deux groupes pour pouvoir comparer mais là encore il faudrait un échantillon statistique assez important pour tirer de possibles conclusions sur la comparaison).
- Méthode n°2 : La deuxième consisterait dans un premier temps à calculer les tenseurs moyens en utilisant uniquement les données du S66 puis du celles du SAK. Ensuite, on répertorie toutes les molécules, parmi celles fournies, où les types atomiques sont communs aux deux sets.
  On calculera ensuite la polarisabilité moléculaire de toutes les molécules dont les types atomiques sont présents dans les deux séries de fichiers, **en utilisant les deux tenseurs moyens différents pour chaque type**. De cette manière on pourrait comparer si l'origine des fichiers influe sur l'écart entre les polarisabilités transférées et les polarisabilités théoriques et donc si les deux bases de données de polarisabilités atomiques sont transférables.

- Méthode n°3 : La troisième et la plus naïve, énoncée auparavant, serait de calculer la polarisabilité moléculaire sur tout le *BigSet* en faisant abstraction du caractère bimodal des valeurs propres.

Pour mieux visualiser ce problème de bimodalité sur les types communs entres les deux jeux de données, on compare les moyennes des traces des tenseurs de polarisabilité pour ces types. (Figure 7 et 8).

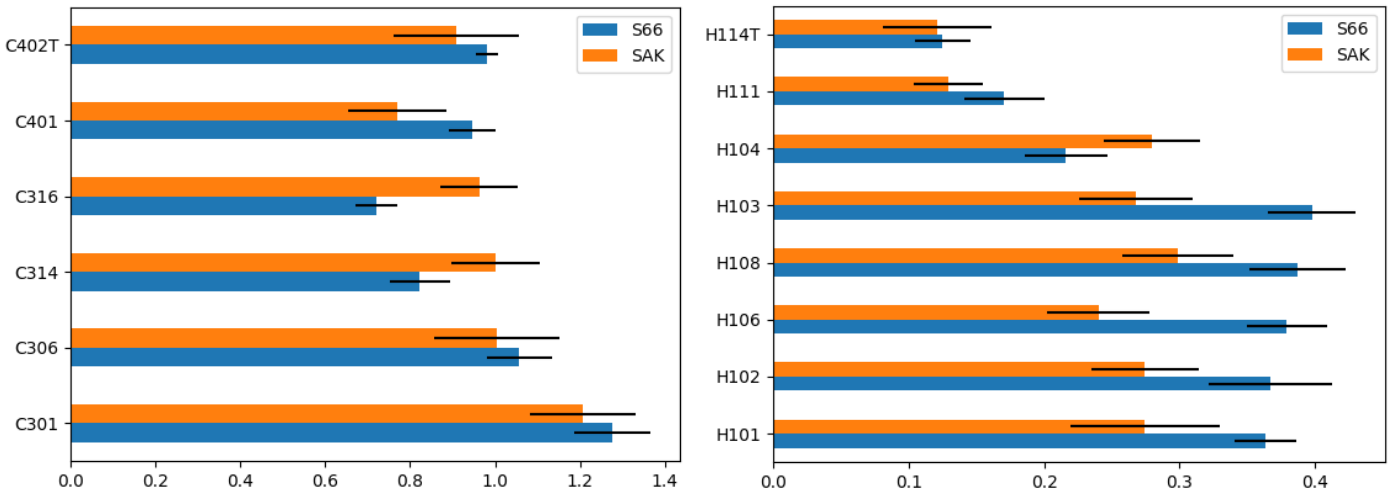


Figure 7 : Moyenne des traces des tenseurs de polarisabilité pour chaque type atomiques hydrogènes (à droite) et carbones (à gauche) communs aux séries de fichiers S66 et SAK, le STD est représenté par le trait noir.

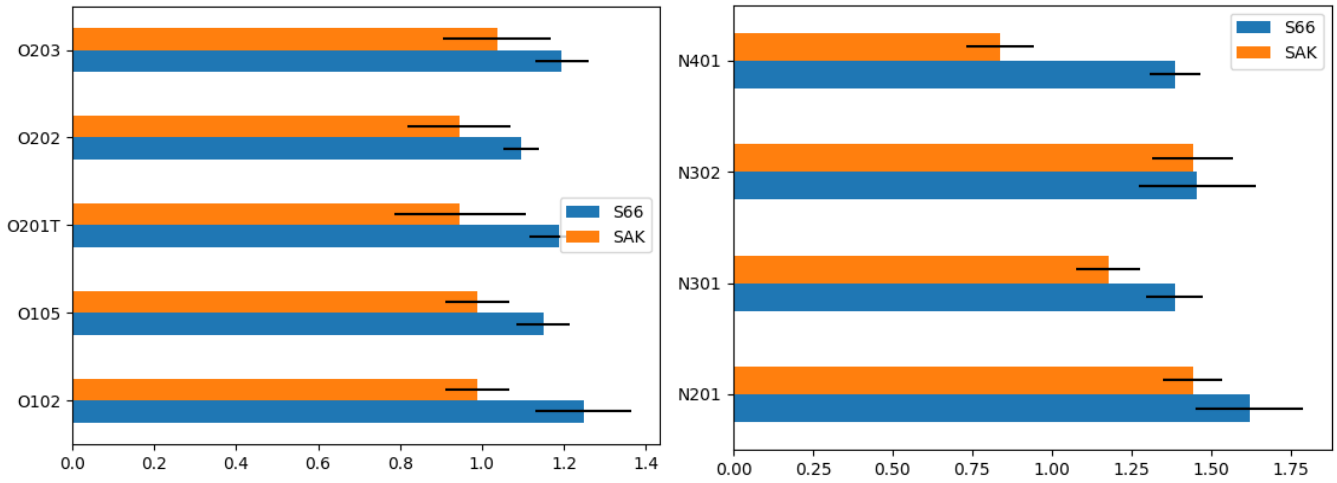


Figure 8 : Moyenne de la trace des tenseurs de polarisabilité diagonalisés pour les types atomiques oxygènes et azotes communs aux séries de fichiers S66 et SAK, le STD est représenté par le trait noir

Sur certains types, on observe quasiment aucune différence (ex : figure 8, N302) par opposition à d'autres où la distinction est remarquable (ex : figure 7 : H108, H102, H101, H106). Le traitement séparé du *BigSet* en deux sous-ensembles permet donc de régler le problème de bimodalité des populations de tenseurs de polarisabilité.

Cependant pour s'en assurer on vérifie chaque dispersion pour chaque type et cela sur les deux sous-ensembles de données, et à raison. Effectivement on peut émettre quelques réserves sur le fait que la présence de deux modes de dispersion soit dans certains cas, uniquement liée à l'origine des deux sets de données différents. Exemple ici figure 9, avec le type atomique **C316**.

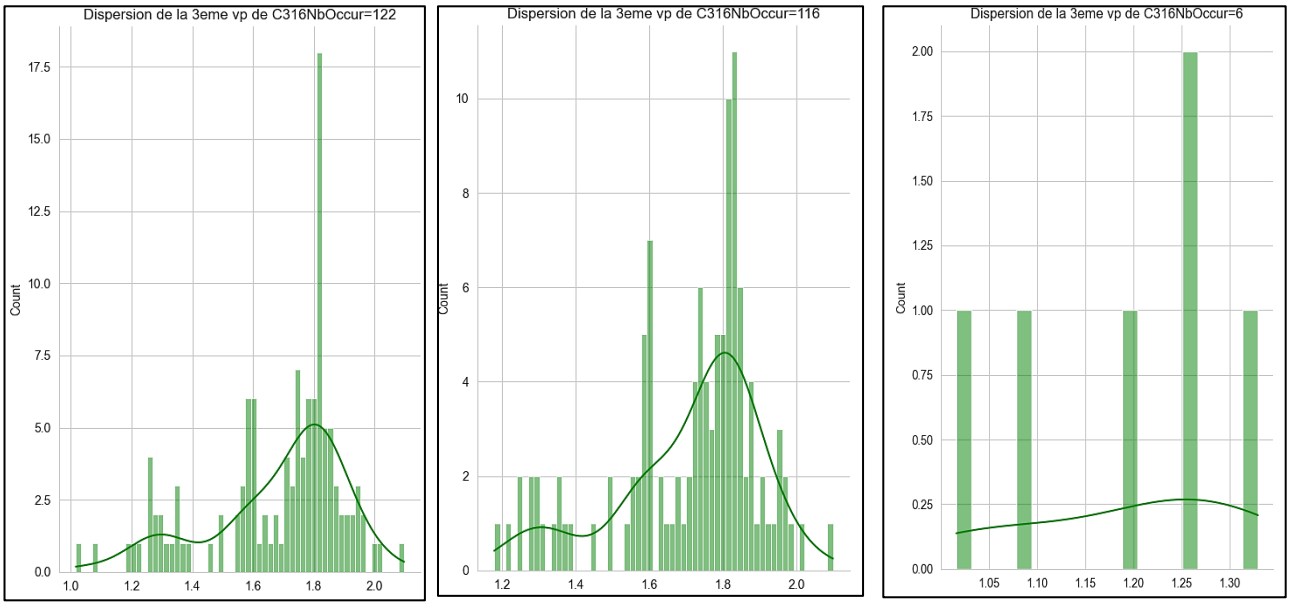


Figure 9 : Dispersion de la troisième valeur propre du tenseur de polarisation du type C316 pour un traitement sur : tout le BigSet, le SAK, le S66 (de gauche à droite)

Malgré le traitement séparé, la courbe de densité de la deuxième image de la figure 9 montre que dans le jeu de fichiers SAK, deux populations de tenseurs de polarisabilités semblent encore être rassemblées sous la même étiquette C316, une dont la 3ème valeur propre est autour de 1.8 $Å^3$ et une autour de 1.3 $Å^3$. Cette deuxième population se retrouve également dans les fichiers type S66. L'observation des densités électroniques de déformation qui suit sur des atomes de type C316 met en évidence l'origine de ce phénomène.

Mode de lecture des cartes 2D de densités électroniques de déformation (Figures 10, 11 et 12)

Le carbone étudié est ici le C29 (figure 12), *MoProViewer* permet de tracé une carte de densité de déformation électronique 2D dans un plan défini par trois atomes sélectionnés. Les lignes de niveaux correspondent à la variation $\Delta\rho = \rho_{\text{réel}} - \rho_{\text{sphérique}}$ entre la densité électronique multipolaire et la densité électronique sphérique. Les zones bleues correspondent donc à un $\Delta\rho > 0$, plus d'électron sont localisés dans ses régions et les zones rouges à un $\Delta\rho < 0$, la ou des électrons ont été délocalisés. Le trait jaune représente un $\Delta\rho = 0$.

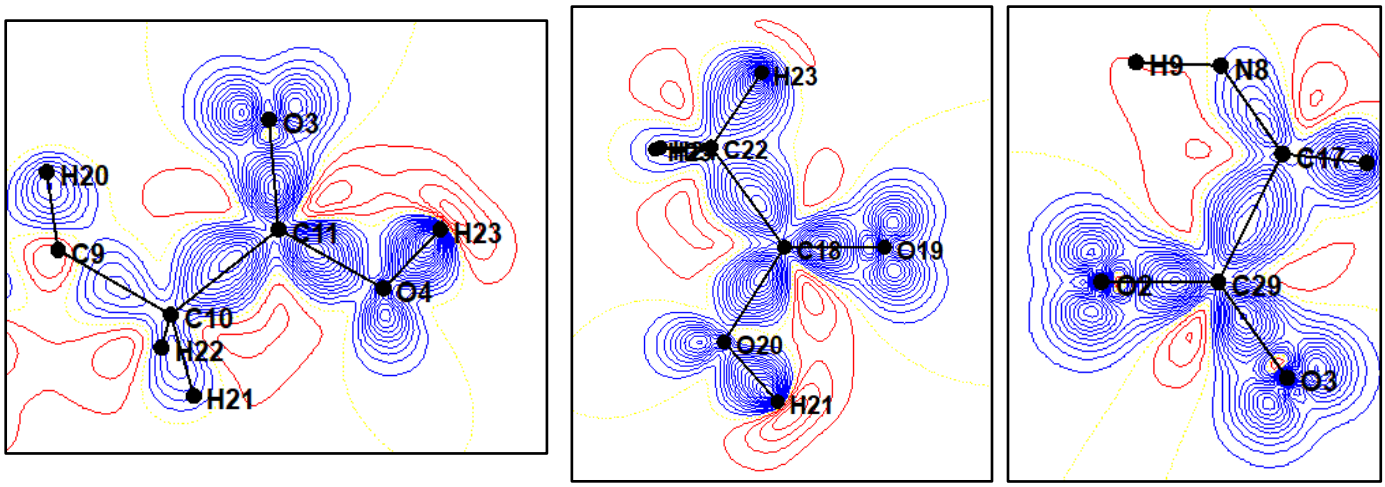


Figure 10 : (à gauche) Densité de déformation électronique du C11 dans le Fichier GLUGLY (Type SAK) correspondant à une valeur propre n°3 autour de 1.30 $Å^3$

Figure 11 : (au centre) Densité de déformation électronique du C18 dans le Fichier PentaneAcOH (Type S66) correspondant à une valeur propre n°3 autour de 1.30 $Å^3$

Figure 12 :(à droite) Densité de déformation électronique autour du C29 dans le Fichier XUDOVH (Type SAK) correspondant à une valeur propre n°3 autour de 1.85 $Å^3$

On constate en comparant les densités électroniques autour des carbones C11 et C29 des figures 10 et 12 que pour systèmes présent dans le set SAK, un même atome de carbone de type C316 a effectivement le même environnement proche (ses voisins/liaisons covalentes sont identiques). Cependant, si l'on regarde au-delà, la présence d'un hydrogène lié à un des voisins oxygène du carbone considéré change complètement le groupe caractéristique en présence et de ce fait modifie la densité électronique dans cette zone. Le carbone C316 est impliqué dans deux fonctions chimiques : celle d'un groupement acide carboxylique COOH et celle de sa base conjugué COO-.

Tous deux auront donc un tenseur de polarisabilité sensiblement différent et c'est pour cela qu'on observe une différence sur la distribution des valeurs propres. Les tenseurs de polarisabilité des carbones impliqués dans un groupe carboxylique auront une troisième valeur

propre autour de 1.3 $Å^3$(figure 10 et 11) et ceux impliqués dans un groupement carboxylate (figure 11) autour de 1.85 $Å^3$.

Pour conclure sur le problème de bimodalité des polarisabilités atomiques, il est avéré que l'origine de ce problème vient majoritairement du fait que notre base de données est composée de données d'origines différentes ou les polarisabilités atomiques n'ont pas été obtenues de la même manière. Cependant, le cas du carbone C316 impliqué dans les configurations COO- et COOH montre que : bien que considéré comme transférable du point de vue de la densité électronique, l'approximation n'est plus valable pour les polarisabilités. Selon nous, cela montre la nécessité de **séparé le type C316 en deux types différents** étant donné leur environnement chimique éloigné.

3. Création des tenseurs moyens pour chaque type atomique

Les trois approches étant désormais définies, on peut en se basant sur ces dernières créer des tenseurs moyens pour chaque type atomique afin de créer une base de données répertoriant ces tenseurs avec pour but de transférer les polarisabilités atomiques, en même temps que la densité électronique ELMAM2[6], sur n'importe quelle molécule pour peu qu'elle possède les atomes parmi les 56 types atomiques que nous avons à disposition.

Pour la création de ces tenseurs moyens, deux méthodes sont possibles. La première, la plus intuitive, consiste à tout simplement moyenner élément par élément chaque tenseur d'un type donné afin de n'avoir qu'un seul tenseur moyen en sortie. Cette approche est rendue possible par le fait que ces tenseurs sont maintenant décrits dans un repère local, propre à chaque type d'atome. La deuxième méthode, plus complexe, sollicite l'utilisation des valeurs et vecteurs propres de chaque tenseur et de construire le tenseur moyen à partir de ceux-ci.

La principale différence est que la première méthode dans certains cas, ne restitue pas correctement l'anisotropie des tenseurs inclus dans la procédure de moyenne, ce qui n'est pas le cas en moyennant à partir des valeurs et vecteurs propres étant donné que ces derniers sont intrinsèquement liés à l'orientation des ellipsoïdes qui les représentent. Il serait donc préférable en toute rigueur de travailler avec la seconde méthode afin d'avoir des polarisabilités moyennes aussi précises que possible.

Cependant, la méthode par valeurs propres implique un travail sur les contraintes de symétrie des tenseurs atomiques ne pouvant pas être traitées dans le temps imparti (cette méthode est celle utilisée par le module *Beluga* dans *MoProViewer*, elle est détaillée dans la thèse de Théo Leduc[1]). Compte tenu de ces considérations, nous avons décidé d'utiliser la technique des tenseurs moyens obtenus via les moyennes éléments par éléments (**II. 3**) pour effectuer les trois approches énoncées précédemment, on a donc de manière générale, et pour chaque type atomique :

$$\alpha_{moyen} = \begin{bmatrix} \langle \alpha_{xx} \rangle & \langle \alpha_{xy} \rangle & \langle \alpha_{xz} \rangle \\ \langle \alpha_{yx} \rangle & \langle \alpha_{yy} \rangle & \langle \alpha_{yz} \rangle \\ \langle \alpha_{zx} \rangle & \langle \alpha_{zy} \rangle & \langle \alpha_{zz} \rangle \end{bmatrix}$$

3.1 Création des fichiers ELMAM à partir de chaque base de données.

Pour ces trois cas, il nous a maintenant fallu créer un fichier avec une syntaxe de type « ELMAM » afin qu'il soit lisible par *MoProViewer* par la suite. Ces trois fichiers représentent donc trois bases de données de polarisabilités atomiques d'origines différentes, associées aux paramètres de densité électroniques transférables. Là aussi, un script python nous a permis d'écrire ces tenseurs moyens pour chaque type dans le fichier texte formaté en adéquation avec les 3 cas possibles.

```
ATOM      H101  H       ZX    C   C   -   C        Hc[cc]
NBOND     1   CYCLE   0       CHIV    -    CHIR    0
SYMPLM    cy   CONVAL  -   CONPLM  -    CONKAP  -
DIST      H-C  1.083
DIST_ESD       0.003
KMD       1.15575 1.18064  0.918  0.000  0.000  0.000  0.145
KMD_ESD   0.01965 0.01174  0.051  0.000  0.000  0.000  0.018
QUA                        0.068  0.000  0.000  0.000  0.000
QUA_ESD                    0.017  0.000  0.000  0.000  0.000
ALPHA        0.16330  0.20356  0.76515  -0.00037  0.01382  -0.00207
ALPHA_ESD    0.00000  0.00000  0.00000  0.00000  0.00000  0.00000
TEXT      H-c aromatic atom
```

Figure 13 : Exemple de disposition des informations pour le type « H101 » respectant la syntaxe ELMAM pour les fichiers issus du S66 isolés. Les six éléments du tenseur moyen symétrique y sont encadrés en rouge.

Par la suite, nous pourrons transférer les densités multipolaires et les polarisabilités atomiques moyennes sur les atomes de n'importe quelle molécule compatible. La nécessité de créer un fichier respectant la syntaxe de type « ELMAM » (Figure 13) vient du fait que le module *Beluga* de *MoProViewer* permet de transférer facilement sur une molécule en utilisant ces fichiers et un fichier moléculaire en .par. En sortie, il ajoute automatiquement une ligne légendée « TPOL » pour chaque atome avec les composantes du tenseur moyen associé au type de ce dernier, écrit dans le repère local propre au type d'atome considéré (Figure 14).

```
ATOM 1 N1 xyz 2    1.572481 0.254549 -0.256481    1.0000    1    N
bXY     C2     C10     OCT    K1    V0    M0    Q0
UANI 1.641874 2.157681 0.953464 -0.063292 -0.103934 -0.222129
5.12364 0.000 -0.129 -0.008 0.000 -0.101 0.000 0.000 0.041 -0.008
0.000 -0.031 0.007 0.000 0.000 -0.121 0.006
TPOL 1.78279 2.12593 0.95436 -0.06660 -0.03722 0.02143
```

Figure 14 : Exemple d'ajout de la ligne « TPOL » pour un atome d'azote de type « N201 », le transfert est effectif.

Les trois bases de données ELMAM créées sont respectivement nommées *ELMAM SAK*, *ELMAM BigSet* et *ELMAM S66* dans la suite de ce travail en référence au jeu de données correspondant (**II. 8**).

4. Calcul des polarisabilités moléculaires : Méthode n°1 : Travail sur les deux sets séparés.

4.1 La polarisabilité moléculaire

La polarisabilité moléculaire est définie comme la somme des polarisabilité atomiques. C'est une somme tensorielle donc il est primordial d'exprimer tous les éléments sommés dans le même système de coordonnées. Pour cela on exprime chaque polarisabilité atomique dans le repère cartésien de la molécule. Pour une molécule de n atomes on a donc :

$$\sum_{i=1}^{n} \alpha_i^* = \alpha_{mol} \quad (2)$$

Avec α_i^*, le tenseur de polarisabilité de l'atome i exprimé dans le repère de la molécule.

On a deux polarisabilités moléculaires pour chaque molécule : une théorique, directement calculée en sommant les tenseurs inscrits dans les données initiales dont nous disposions, et une dite transférée, calculée en associant pour chaque atome la polarisabilité atomique moyenne associer à son type atomique définit par la bibliothèque ELMAM choisie. On calcule ensuite la polarisabilité moléculaire en diagonalisant le tenseur de polarisabilité moléculaire et en calculant le tier de sa trace. C'est cette grandeur que l'on va comparer dans le cas théorique et transféré.

Etant donné que la polarisabilité moléculaire est une somme de polarisabilités atomiques anisotropes, étudier l'anisotropie pour évaluer sa transférabilité fait sens. Dos Santos,

Krawczuk et Macchi[8] utilisent l'indicateur $\Delta\alpha$ pour évaluer l'anisotropie d'un tenseur de polarisabilité $\boldsymbol{\alpha}$.

$$\Delta\alpha = \left[\frac{3.trace(\boldsymbol{\alpha}^2) - trace(\boldsymbol{\alpha})^2}{2}\right]^{\frac{1}{2}} \quad (3)$$

Les polarisabilités et l'indicateur $\Delta\alpha$ sont homogènes à un volume et exprimées en Å^3.

4.2 Transfert des deux jeux de données S66 et SAK avec leur propre base de données de polarisabilité atomique (II. 6).

Les polarisabilités moléculaires issues des données initiales et reconstitués à l'aide de nos polarisabilités atomiques moyennes par types d'atomes sont comparées figures 15.

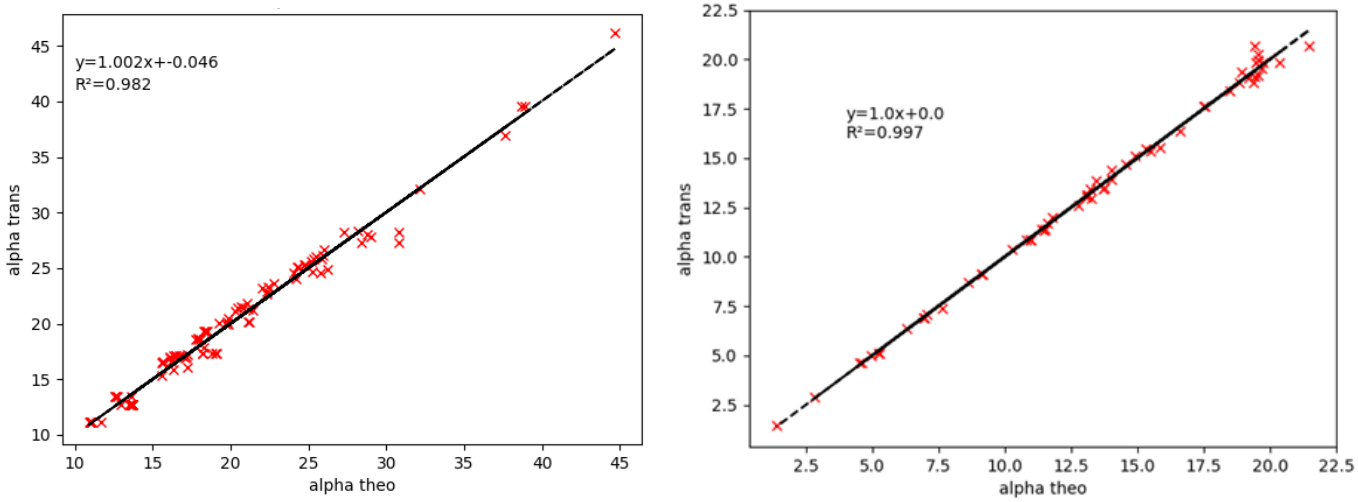


Figure 15 : Comparaison entre polarisabilités moléculaires isotropes théoriques et transférées sur le set S66 (à droite) et sur le SAK (à gauche) en Å^3.

Le coefficient directeur de la régression linéaire est dans les deux cas est unitaire, et les coefficients de détermination eux aussi sont excellents**. <u>Les polarisabilités moléculaires sont donc transférables avec ces bases de données de polarisabilités atomiques</u>**. L'anisotropie des polarisabilités moléculaires transférées est moins bien reproduite (coefficient de détermination et coefficient directeur de régression éloignés de l'unité pour les deux cas, cf : annexe n°5) dû à la méthode utilisée pour moyenner les tenseurs décrite partie 3.

5. Calcul des polarisabilités moléculaires : Méthode n°2 : Comparaison du transfert du S66 avec les paramètres des types atomiques obtenues sur le set de données SAK et S66.

5.1 Détermination des types ou le transfert est possible (II. 9)

Le S66 est constitué de dimères et de monomères. Le SAK est complètement différent et par conséquent on ne pourra transférer que les molécules constituées de types atomiques présents dans les deux sets. Nous sommes cependant obligés de nous limiter à des dimères extraits du S66. En effet, par la suite il faudra utiliser des références d'énergie de polarisation théoriques à comparer aux énergies de polarisation transférées que nous calculerons partie 8. Les références théoriques sont pour le moment seulement disponibles pour les dimères du S66, et c'est pourquoi nous travaillerons uniquement sur ce jeu de données. Mais nous pouvons tout à fait imaginer les mêmes analyses en incluant les molécules du SAK si les énergies de polarisation théoriques étaient disponibles.

Ce qui nous permet de traiter entièrement seulement 22 systèmes moléculaires (Annexe n°1) du S66 incluant 15 dimères construits à partir de 7 monomères.

5.2 Comparaison des polarisabilités moléculaires et de leur anisotropie en Å^3 sur les 15 dimères sélectionnés transférés par ELMAM S66 et ELMAM SAK.

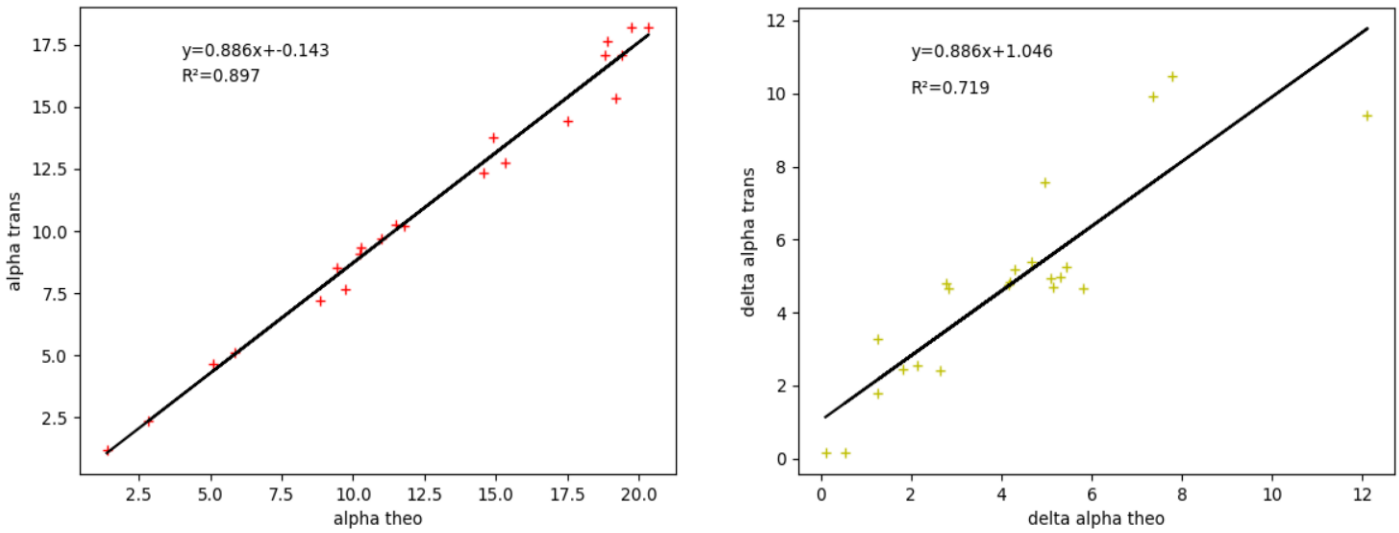


Figure 16 : A gauche, la comparaison entre la polarisabilité moléculaire isotrope transférée par ELMAM SAK et la polarisabilité moléculaire isotrope théorique sur les 15 dimères du S66 choisis. A droite, la comparaison de leur anisotropie.

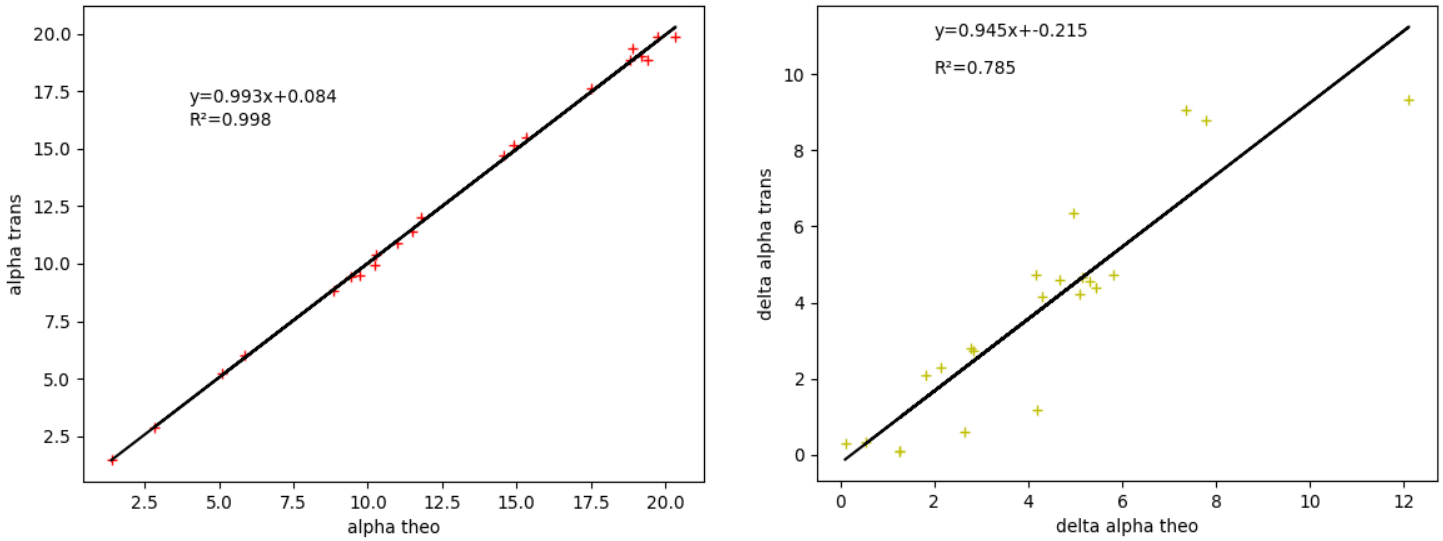


Figure 17 : A gauche, la comparaison entre la polarisabilité moléculaire isotrope transférée par ELMAM S66 et la polarisabilité moléculaire isotrope théorique sur les 15 dimères du S66 choisis. A droite, la comparaison de leur anisotropie.

La régression sur les polarisabilités moléculaires isotropes donne un coefficient directeur plus éloigné de l'unité que lorsque l'on a transféré chaque set avec la base de données ELMAM lui correspondant (partie 4.2, figures 15). Cependant on remarque que pour le transfert du S66 par la base donnée ELMAM SAK (Figure 16), le résultat n'est pas aberrant, les polarisabilités semblent transférables d'un set à l'autre.

Les polarisabilités moléculaires des 15 dimères du S66 sont plus proches des valeurs théoriques lorsque l'on travaille avec la bibliothèque *ELMAM S66* (coefficient directeur de 0.993 contre 0.886 avec *ELMAM SAK*, figure 16 et 17). On observe également que ces régressions restent correctes même avec les quelques approximations sur les types atomiques que nous avons été forcés de faire, une partie est consacrée à la description et justification de ces dernières annexe n°7.

En ce qui concerne l'anisotropie (figures de droite 16 et 17), on a encore une fois une régression bien moins pertinente comme lors du traitement sur les deux sets séparés. Cela s'explique par le fait que l'on travaille toujours sur des tenseurs moyennés élément par élément et aussi dû au fait que le SAK comporte majoritairement des molécules au polarisabilités moléculaires plus anisotropes que celles du S66.

On va donc par la suite, se focaliser sur l'études des 15 dimères du S66 que l'on a transférés avec ces bases de données d'origines différentes. Une fois polarisées avec l'outil *Beluga MP Polarizer* de *MoProViewer*, on va calculer leurs énergies de polarisations pour pouvoir confronter ces dernières à d'autres résultats obtenus par des méthodes purement théoriques (méthode SAPT[3]).

6. Calcul des polarisabilités moléculaires : Transfert avec ELMAM BigSet sur tout le BigSet. (**II. 6**)

6.1 Comparaison entre les polarisabilités moléculaires isotropes transférées par ELMAM BigSet et les polarisabilités moléculaires isotropes théoriques ainsi que de leur anisotropie en $Å^3$.

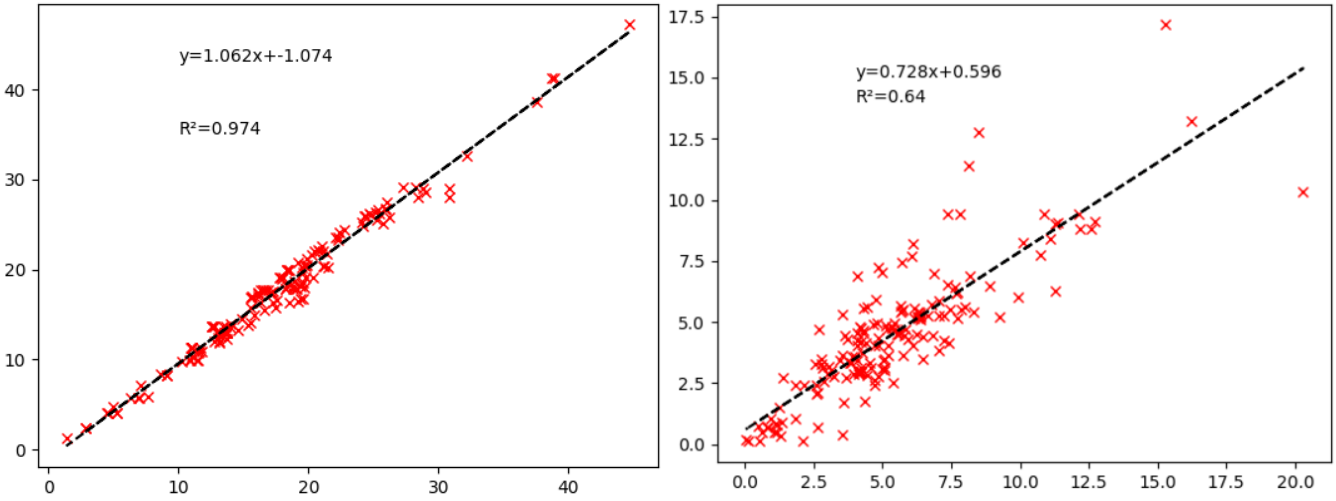


Figure 18 : Comparaison entre les polarisabilités moléculaires isotropes transférées par ELMAM BigSet et les polarisabilités moléculaires isotropes théorique (à gauche). A droite on compare l'anisotropie sur les mêmes molécules.

En travaillant naïvement sur tout le set, on remarque que l'on garde une régression sur les polarisabilités moléculaire avec un coefficient directeur de régression quasiment unitaire (figure 18) alors que le caractère bimodal du set de donnée est ici passé sous silence. On pourrait considérer **les polarisabilités atomiques de la base de données ELMAM BigSet comme transférable bien qu'elles soient hétérogènes.** Une mauvaise estimation de l'anisotropie (figure 18) est encore une fois visible pour les raisons énoncées au début de la partie 3.

Cette méthode n'est donc en soit pas pertinente mais montre que l'hétérogénéité de la base de données d'origine est indétectable sans étude approfondie des polarisabilités.

7. Polarisations des 15 dimères du S66 sélectionnés avec les tenseurs moyens définis par les deux sets

On a donc accès à 15 dimères transférables par deux bibliothèques ELMAM différentes que l'on a mises au point pour le transfert précédemment. Le module *Beluga* de *MoProViewer* fournit un outil de polarisation de la densité électronique d'un dimère, le fonctionnement est détaillé chapitre 2 partie 2.1 de la thèse de Théo Leduc[1]. On va rappeler rapidement son fonctionnement à l'aide d'un workflow (Figure 19).

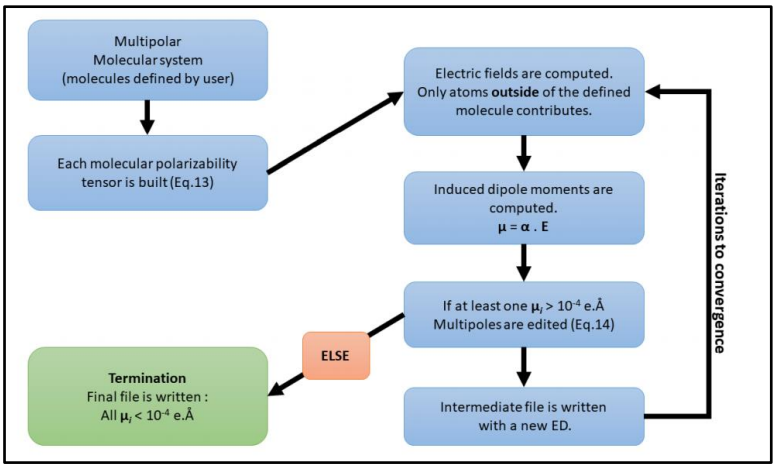


Figure 19 : Source : « Vers un potentiel multipolaire quantitatif et transférable aux macromolécules biologiques : une étude méthodologique des effets de la polarisabilité[1] », Theo Leduc

Une sélection d'atomes au sein de la molécule est polarisée par le champ électrique de tous les atomes non sélectionnés et donc des moments dipolaires atomiques induit sont calculés itérativement sur les deux groupes d'atomes, jusqu'à un seuil de convergence placé à $10^{-4}\ e.Å$, via la formule :

$$\vec{\mu} = \alpha.\vec{E} \qquad (4)$$

Si ce seuil est dépassé, les populations dipolaires du modèle multipolaire (cf : annexe n°3) vont être modifiées, cet algorithme va donc effectuer plusieurs cycles jusqu'à atteindre la condition d'arrêt sur le moment induit. Le nombre de cycle maximal est fixé par l'utilisateur.

Une fois les deux monomères polarisés mutuellement, on va pouvoir représenter les densités électroniques polarisées souhaitées. *MoProViewer* permet de visualiser ces densités électroniques de déformation.

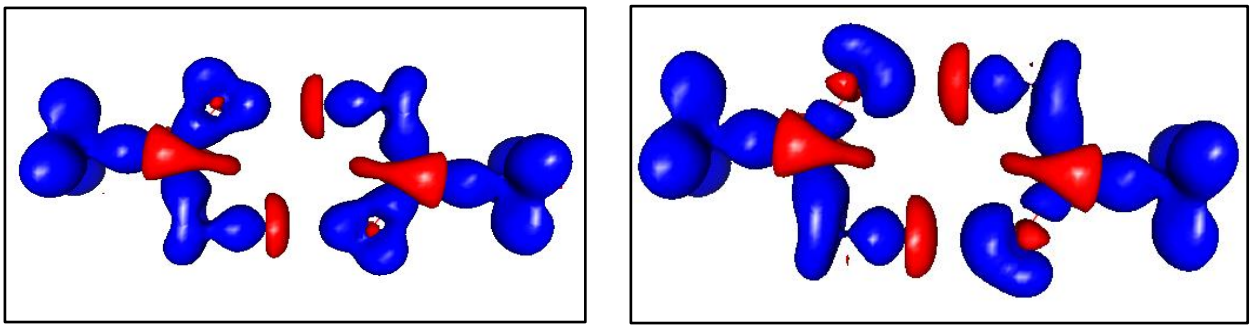


Figure 20 : Densité électroniques de déformation avant et après polarisation du dimère AcOH-AcOH. Les zones rouges correspondent à des électrons délocalisés, vers les zones bleues. Les iso-surfaces sont de ±0.2 e/Å³

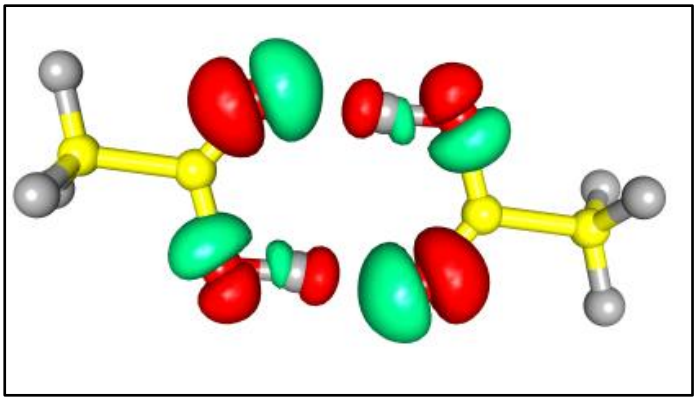


Figure 21 : Différence entre les densités électroniques après et avant polarisation du dimère AcOH-AcOH. La différence met en évidence les dipôles induits, en rouge les électrons délocalisés vert les régions en vert. L'échelle utilisée pour cette figure est différente de celle de la figure précédente pour un soucis de représentation.

8. Validation du principe de transférabilité pour des dimères polarisés : Comparaison des énergies de polarisations.

Avec les 15 dimères du S66 étudiés, nous allons maintenant calculer la contribution de polarisation à leur énergie d'interaction électrostatique afin de les comparer aux valeurs théoriques et valider ou non l'hypothèse de transférabilité des tenseurs de polarisabilités moyens.

8.1 Description des énergies comparées

Dans ce présent travail, on utilise comme référence théorique les énergies obtenues via la méthode « SAPT » (*Symmetry Adapted Perturbation Theory* [3]).

$$E_{SAPT} = E_{ELEC} + E_{IND} + E_{DISP} + E_{EX-CORR} \quad (5)$$

C'est une méthode qui calcule les diverses contributions énergétiques des dimères à partir de calculs de perturbations théoriques. Le terme qui nous intéresse dans la précédente formule est donc E_{IND}, c'est ce dernier qui contient les informations sur l'énergie de polarisation théorique, que l'on va comparer avec l'énergie de polarisation obtenue via nos polarisabilités transférées.
Cette énergie de polarisation transférée est calculable comme suit :

$$E_{POL,DB} = E_{ELEC,POLARIZED} - E_{ELEC,DB} \quad (6)$$

Avec $E_{ELEC,DB}$ qui représente l'énergie d'interaction électrostatique entre les deux dimères avant application de la polarisation, et $E_{ELEC,POLARIZED}$, énergie d'interaction électrostatique en tenant compte de la polarisation mutuelle des monomères impliqués. La différence entre ces deux termes donne donc l'énergie de polarisation, qui sera naturellement négative. Lors de cette comparaison, il est à noter qu'une certaine marge d'erreur est attendue. En effet, si le terme d'énergie théorique E_{IND} prend bien en compte les dipôles induits lors de la polarisation des molécules formant le dimère, il contient aussi d'autres contributions telles que les transferts de charge ou d'autres contributions d'induction de moments d'ordre supérieurs.

8.2 Calcul des énergies transférées

On a donc, pour chacun des 15 dimères du S66 à notre disposition, des modèles de densité électronique non polarisés (simplement transférées) et d'autres polarisées. La méthode sera la même peu importe la base de données utilisée, de manière à comparer uniquement les polarisabilités utilisées. On prend pour un dimère donné le fichier non polarisé et on calcule grâce à *MoProViewer* l'énergie électrostatique totale du dimère. On répète ensuite la même méthode pour le dimère polarisé, ce qui nous donne donc pour un dimère l'énergie électrostatique quand il est polarisé et quand il ne l'est pas.

On calcule l'énergie de polarisation transférée $E_{POL,DB}$ comme dans (2) et on la compare avec E_{IND} issue de calculs SAPT[3], dans le cas des fichiers transférés avec l'ELMAM SAK et celui du S66.

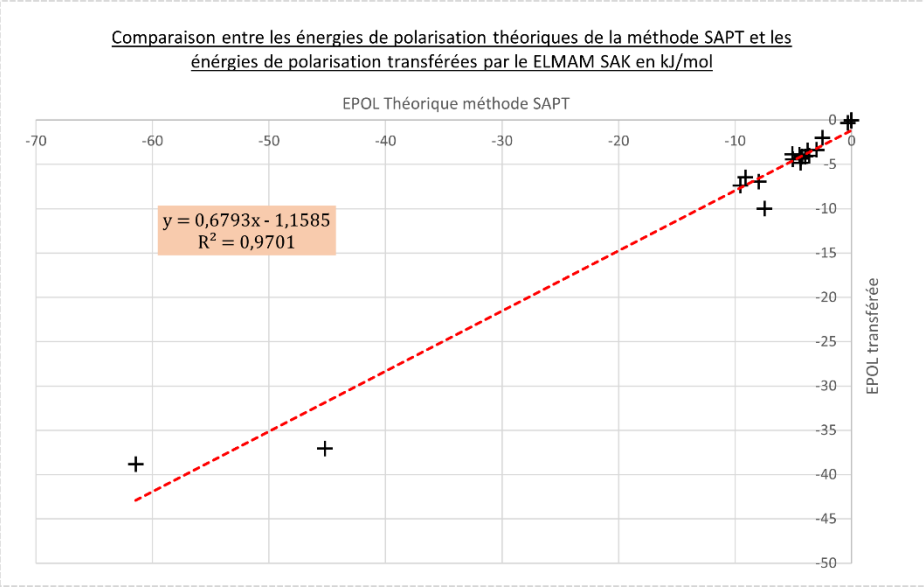


Figure 22 : Utilisation des polarisabilités moyennes issues du jeu de données SAK.

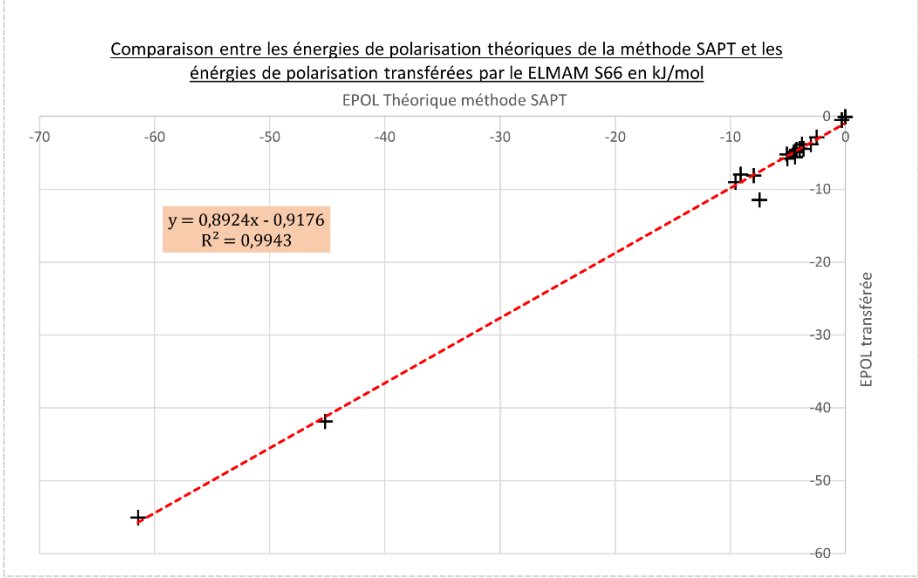


Figure 23 : Utilisation des polarisabilités moyennes issues du jeu de données S66

Dans le cas des énergie obtenues via le transfert à partir du set S66 (Figure 23), on constate une régression plutôt satisfaisante avec un coefficient directeur égal à 0.8928 et un R^2 de 0. 9948. L'hypothèse de transférabilité semble ici validée pour cette série de données de polarisabilité et ce jeu de dimères.

En revanche, dans le cas du SAK (Figure 22), on obtient un coefficient directeur de 0.6789, beaucoup moins satisfaisant. On remarque sur le graphique que c'est majoritairement dû à 3 valeurs aberrantes, elles correspondent aux dimères :

18_WaterPyridine100_05_TheorPolDim_01 (a)

20_AcOHAcOH100_02_TheorPolDim_ (b)

21_AcNH2AcNH2100_02_TheorPolDim_01 (c)

On peut donc imaginer que **l'hypothèse de transférabilité des polarisabilités moyennes moins valable** sur ces trois dimères. On pourrait se pencher sur ces dimères en particulier et essayer de trouver sur quels types atomiques précisément l'hypothèse de transférabilité est moins vérifiée.

La limite de l'hypothèse de transférabilité sur ces 3 dimères pourrait directement provenir de la différence des méthodes utilisées pour calculer les tenseurs théoriques. En effet, comme dit dans la première partie, les tenseurs calculés par l'équipe de Anna Krawczuk ont comme principale différence de toujours tenir compte des polarisation de liaison, ce qui n'est pas le cas des tenseurs calculés par le *CRM²*, ce qui pourrait expliquer cet écart de valeur sur des molécules très polarisables.

Les dimères (b) et (c) sont en effet fortement polarisables due notamment à la présence d'un groupe COOH ou amide. De plus leur disposition (Figure 24) accentue le phénomène.

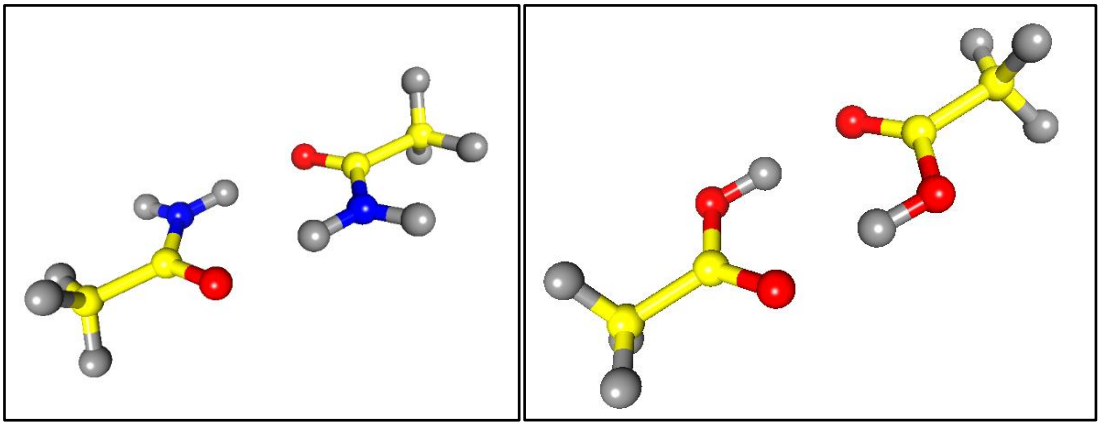


Figure 24 : Dimère d'acide éthanoïque (b) (à droite) et acétamide (c) (à gauche), les groupes amide et acide carboxylique sont dans le même plan. Les dimères sont donc très polarisables.

Leur énergie de polarisation est donc bien supérieure aux autres, une nouvelle régression linéaire limitée aux énergies de polarisation les plus faibles semble indiquée (figure 25 et 26).

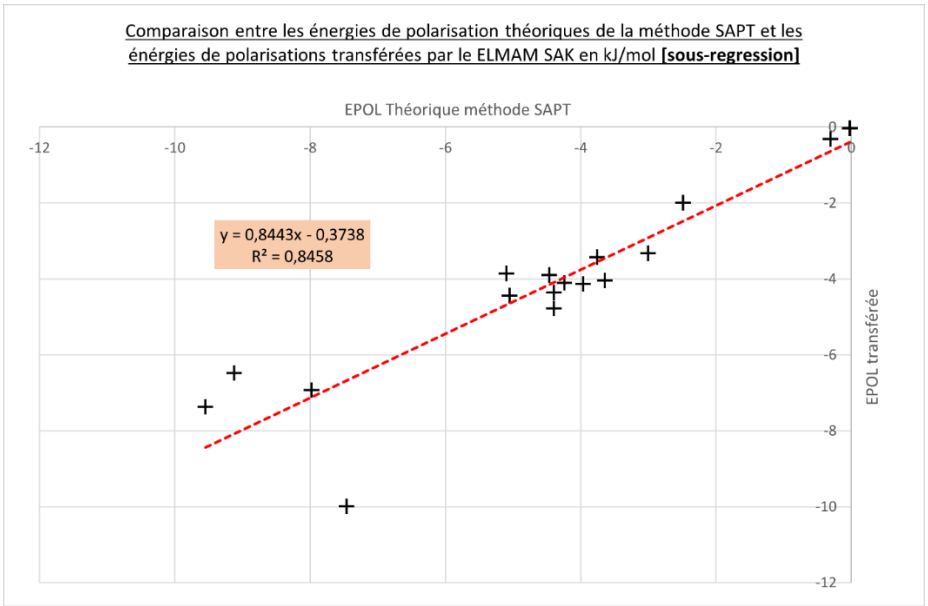


Figure 25 : Utilisation des polarisabilités moyennes issues du jeu de données SAK.

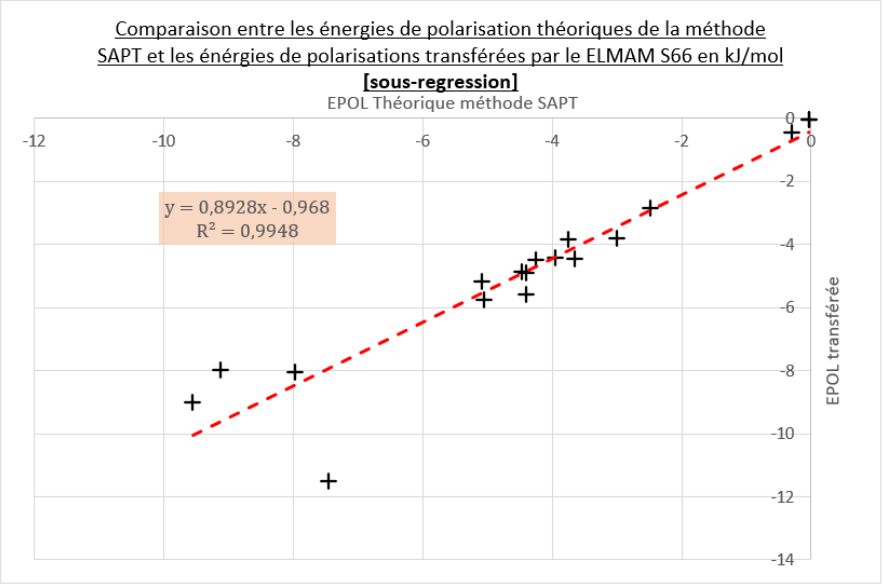


Figure 26 : Utilisation des polarisabilités moyennes issues du jeu de données S66.

Dans le cas de l'utilisation de l'ELMAM S66 (Figure 26), la pente de la courbe a légèrement baissé, passant à 0.8733, mais reste clairement acceptable. Quant au cas où nous avons utilisé l'ELMAM SAK (Figure 25), le coefficient directeur est maintenant de 0.8442, bien plus proche de l'unité qu'il ne l'était auparavant. Ce qui montre que pour ces dimères, l'hypothèse de transférabilité s'avère vérifiée.

Pour conclure sur la comparaison des énergies de polarisations. Le bon accord sur la régression entre la méthode de calcul SAPT et les énergies obtenues par transfert des polarisabilités moyennes du S66 sur les 20 dimères sélectionnés montre que pour **tous ces dimères**, l'hypothèse de transférabilité reste valide bien que menant à des énergies de polarisations sous-estimées. Cette sous-estimation peut être du aux méthodes utilisées en amont pour l'obtention du jeu de données S66 mais aussi à l'absence de transfert de charge ou d'induction de moments d'ordre supérieur (cf. Partie 1). Le cas du transfert avec le jeu de

données issues du SAK quant à lui ne permet pas de valider l'hypothèse de transférabilité sur tous les dimères notamment sur certains dimères très polarisable comme celui d'acétamide ou d'acide acétique.

9. Conclusions

Ce travail s'inscrit dans la continuité des recherches du laboratoire *CRM²* afin de valider le principe de transférabilité dans le cas des tenseurs de polarisabilité d'atomes répertoriés par types suivant leur environnement chimique. Cette validation s'est faite premièrement par la détermination des types atomiques présents dans le *BigSet* via diverses analyses statistiques, puis par construction de tenseurs de polarisabilité moyens pour les types en question.

Ce problème de bimodalité à mis en évidence que la différence de méthodes théoriques de calculs de tenseurs de polarisabilité fait pour une partie du *BigSet* directement par le *CRM²* (S66) et une autre par l'équipe de Anna Krawczuk (SAK) aurait des conséquences plus importantes sur la transférabilité que prévues à l'origine. Le calcul des polarisabilité moléculaires sur les deux jeux séparés valide l'hypothèse de transférabilité (de même que le calcul sur tout le *BigSet* avec un jeu de polarisabilité hétérogène) mais la comparaison est moins exploitable car effectuée sur des statistiques différentes. En limitant les calculs aux 15 dimères du S66 transférables par les deux bases de données de polarisabilités atomiques, on constate que les polarisabilités moléculaires sont bien reproduites.

La comparaison des énergies de polarisations sur les 15 dimères évoqués précédemment, avec celles calculées par la méthode SAPT, met une nouvelle fois en lumière la distinction entre les deux jeux de données. Si les dimères du S66 transférés à partir de cette même source présente encore une fois une bonne compatibilité avec les données théoriques, validant cette fois ci le principe de transférabilité même après polarisation des dimères, c'est nettement moins le cas quand ils sont transférés à partir des données du set SAK, plus précisément, les dimères de nature très polarisables montrent une énergie de polarisation relativement plus éloignée de de la valeur théorique.

Une analyse comparative des deux méthodes d'obtention des jeux de données plus complètes pourrait améliorer la création de futures bases de données de polarisabilités atomiques transférables.

10. Remerciements

Nous souhaitons remercier notre encadrant le professeur Benoît Guillot ainsi que son post-doctorant Théo Leduc de nous avoir permis de faire ce stage durant ces deux mois. Leurs très nombreux conseils et leur sympathie à notre égard nous ont permis d'effectuer ce présent travail dans d'excellentes conditions et d'acquérir de nouvelles connaissances autant en physique qu'en programmation. Un grand merci aussi à toute l'équipe du CRM² que nous avons pu côtoyer brièvement et qui ont su être à l'écoute lors de la présentation de notre sujet.

Références

(1)
Leduc, T. Vers un potentiel multipolaire quantitatif et transférable aux macromolécules biologiques : une étude méthodologique des effets de la polarisabilité. **2019**, 326.

(2)
Hansen, N. K.; Coppens, P. Testing Aspherical Atom Refinements on Small-Molecule Data Sets. *Acta Crystallographica Section A* **1978**, *34* (6), 909–921. https://doi.org/10.1107/S0567739478001886.

(3)
Li, A.; Muddana, H. S.; Gilson, M. K. Quantum Mechanical Calculation of Noncovalent Interactions: A Large-Scale Evaluation of PMx, DFT, and SAPT Approaches. *J. Chem. Theory Comput.* **2014**, *10* (4), 1563–1575. https://doi.org/10.1021/ct401111c.

(4)
Řezáč, J.; Riley, K. E.; Hobza, P. S66: A Well-Balanced Database of Benchmark Interaction Energies Relevant to Biomolecular Structures. *J. Chem. Theory Comput.* **2011**, *7* (8), 2427–2438. https://doi.org/10.1021/ct2002946.

(5)
Řezáč, J.; Riley, K. E.; Hobza, P. Extensions of the S66 Data Set: More Accurate Interaction Energies and Angular-Displaced Nonequilibrium Geometries. *J. Chem. Theory Comput.* **2011**, *7* (11), 3466–3470. https://doi.org/10.1021/ct200523a.

(6)
Domagała, S.; Fournier, B.; Liebschner, D.; Guillot, B.; Jelsch, C. An Improved Experimental Databank of Transferable Multipolar Atom Models – ELMAM2. Construction Details and Applications. *Acta Crystallogr A Found Crystallogr* **2012**, *68* (3), 337–351. https://doi.org/10.1107/S0108767312008197.

(7)
Leduc, T.; Aubert, E.; Espinosa, E.; Jelsch, C.; Iordache, C.; Guillot, B. Polarization of Electron Density Databases of Transferable Multipolar Atoms. *J. Phys. Chem. A* **2019**, *123* (32), 7156–7170. https://doi.org/10.1021/acs.jpca.9b05051.

(8)
Dos Santos, L. H. R.; Krawczuk, A.; Macchi, P. Distributed Atomic Polarizabilities of Amino Acids and Their Hydrogen-Bonded Aggregates. *The Journal of Physical Chemistry A* **2015**, *119* (13), 3285–3298. https://doi.org/10.1021/acs.jpca.5b00069.

(9)
Jelsch, C.; Pichon-Pesme, V.; Lecomte, C.; Aubry, A. Transferability of Multipole Charge-Density Parameters: Application to Very High Resolution Oligopeptide and Protein Structures. *Acta crystallographica. Section D, Structural biology* **1998**, *D54, Part 6* (2), 1306–1318. https://doi.org/10.1107/S0907444998004466.

Annexe

1. Liste des 20 fichiers du data66 transférés avec les tenseurs moyens créer depuis les fichiers SAK.

'01_2701_01WaterWater100_02rot_The.txt',
Dimère Eau-Eau
'18_WaterPyridine100_05_The.txt',
Dimère Eau-Pyridine, l'hydrogène de l'eau face à l'azote de la pyridine
'20_AcOHAcOH100_02_The.txt',
Dimère Acide acétique-Acide acétique dans le plan, l'hydrogène du groupe carboxyle face à l'oxygène de l'autre
'21_AcNH2AcNH2100_02_The.txt',
Dimère Acétamide-Acétamide dans le plan, un de leur hydrogène respectif du groupe NH2 face à l'oxygène de l'autre
'24_BenzeneBenzenepipi100_02_The.txt',
Dimère Benzène-Benzène, en interaction parallèle entre eux
'25_PyridinePyridinepipi100_02_theorPol_01'
Dimère Pyridine-Pyridine, en interaction parallèle entre eux
'27_BenzenePyridinepipi100_02_The.txt',
Dimère Benzène-Pyridine, en interaction parallèle entre eux
'34PentanePentanepolar_trfcellrotpol_01.txt',
Dimère Pentane-Pentane, en interaction parallèle entre eux
'38_CyclopentaneCyclopentane100_02_The.txt',
Dimère Cyclopentane-Cyclopentane, en interaction parallèle entre eux
'39_BenzeneCyclopentane100_02_theorPol_01'
Dimère Benzène-Cyclopentane, en interaction parallèle entre eux
'47BenzeneBenzeneTSpolar_trfce.txt',
Dimère Benzène-Benzène, en interaction perpendiculaire entre eux
'48_PyridinePyridineTS100_02_The.txt',
Dimère Pyridine-Pyridine, en interaction perpendiculaire entre eux
'49_BenzenePyridineTS100_02_theorPol_01'
Dimère Benzène-Pyridine, en interaction perpendiculaire entre eux
'52_BenzeneAcOHOHpi100_02_TheorPol_01'
Dimère Benzène-Acide acétique, en interaction perpendiculaire entre eux
'53_BenzeneAcNH2NHpi100_02_TheorPol_01'
Dimère Benzène-Acétamine, en interaction perpendiculaire entre eux
'54_BenzeneWaterOHpi100_05_The.txt',
Dimère Benzene-Eau, en interaction perpendiculaire entre eux
'58_PyridinePyridineCHN100_05_The.txt',
Dimère Pyridine-Pyridine, dans le même plan
'61_PentaneAcOH100_02_The.txt',
Dimère Acide acétique-Pentane, parallèle entre eux
'62_PentaneAcNH2100_02_The.txt',
Dimère Acétamide-Pentane, parallèle entre eux
'63BenzeneAcOHpolar_cellp.txt',
Dimère Acide acétique-Benzène, parallèle entre eux

2. Liste des 56 types atomiques présents dans le *BigSet* après approximation :

[C301, C304, C305, C306, C307, C308T, C312, C314, C316, C4, C401, C402T, C403, C407, C410, C411, C413T, C414, Cohhh, Cl, H1 , H101, H102, H105, H106, H107, H108, H103, H104, H109T, H110T, H111, H113, H111T, H112, H114T, H115, N201, N301, N302, N303, N304, N401, N402T, O102, O104, O105, O106, O1B, O201T, O202, O203, O204T, Oh1 , P1, S201T]
Leurs caractéristiques sont explicitées dans la base de données ELMAM2.

3. Description des fichiers format MoProViewer du BigSet

ATOM 1 O1 xyz 2 -0.392018 -0.384719 0.076071 1.0000 1 O 1

bXY H2 H3 OCT K1 V0 M0 Q0 2

UANI 1.481586 0.995672 1.181808 -0.064541 -0.018233 -0.020071 3

6.34000 0.000 -0.082 0.000 0.000 0.049 0.000 0.000 0.021 0.000 4

0.000 -0.039 0.000 0.000 0.000 -0.079 0.000

Ligne 1 : numéro de l'atome dans le fichier, nom de l'atome, coordonnées de l'atome dans une base cristallographique.

Ligne 2 : Orientation du repère de définition, première direction, deuxième direction

Ligne 3 : Type de tenseurs de polarisabilité (isotrope ou anisotrope), coefficients du tenseur de polarisation (6 éléments supérieur droit d'une matrice 3x3 symétrique, respectivement xx yy zz xy xz yz)

Ligne 4 : coefficients de population dipolaire, coefficients de populations quadrûpolaire

Les coefficients de population correspondent aux coefficients de combinaison linéaire $P_{l,m}$ dans l'équation ci-dessous extraite de la description multipolaire de la densité électronique de Hansen et Coppens[2].

$$\rho_{atom}(r,\theta,\phi) = P_{core}\rho_{core}(r) + P_{val}\kappa^3\rho_{val}(\kappa r) + \sum_{l=0}^{l_{max}} \kappa'^3 R_l(\kappa' r) \sum_{m=-l}^{l} P_{lm\pm} Y_{lm\pm}(\theta,\phi)$$

4. Test d'Agostino

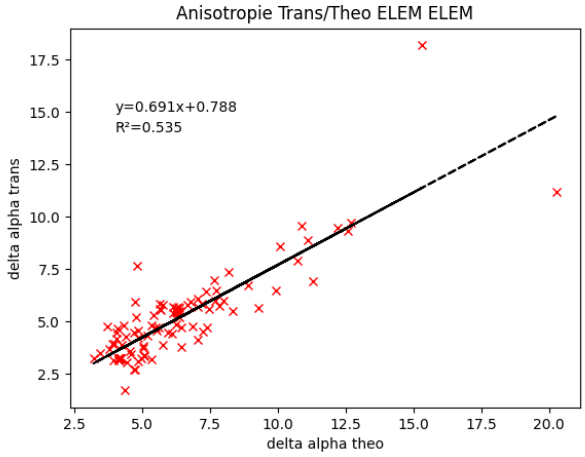
Test d'Agostino utilisé est celui de la librairie *scipy.stats.normaltest.*

Références : https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.normaltest.html

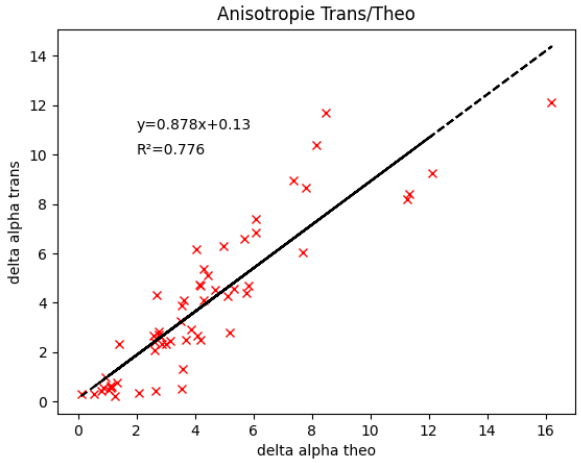
On n'a ici évalué que la validité ou non de l'hypothèse de normalité selon une précision fixé 5% (*pvalue* dans la fonction python). Pour pouvoir remarquer les *outliers* et évaluer la qualité des séries statistiques rapidement.

La description complète du critère de normalité du test d'Agostino : https://en.wikipedia.org/wiki/D%27Agostino%27s_K-squared_test#Omnibus_K2_statistic

5. Anisotropie pour les deux cas présentés en méthode n°1.



Comparaison entre l'indicateur de l'anisotropie des polarisabilités moléculaires théoriques et transférés sur le set SAK *en* $Å^3$



Comparaison entre l'indicateur de l'anisotropie des polarisabilités moléculaires théoriques et transférés sur le set $S66$ *en* $Å^3$, *en utilisant la méthode de moyenne éléments par éléments.*

6.

MANUEL D'UTILISATION/LECTURE DU CODE POUR TRAITEMENT D'UNE BASE DE FICHIERS MORPRO.

Programme créé par Datti Yanis et Degezelle Alban au CRM2

Language: Python 3.9.0 (tags/v3.9.0:9cf6752) sur Windows (64 bits).

IDE : Pyzo

Description général du code

Ce code python a été utilisé et développé tout au long du stage au CRM2 en autodidacte sur le langage Python. Il contient les modules de traitement des fichiers moléculaires ainsi que ceux de l'analyse statistique des types atomiques et la représentation des données pour le rapport de stage.

Sommaire

Le code est construit en plusieurs parties, séparés par des lignes de commentaires en double ## pour les titres. Les titres du sommaire suivant sont identiques a ceux dans le code en .py.

Les parties en # sont des commentaires ou des parties intermédiaires qui sont désactivées car déjà utilisées.

II. Détermination des types atomiques
1. Premier tri (grande précision)
2. Affinage du tri automatique après tri manuel
3. Créations des bibliothèques numériques pour chaque type atomique
4. Procédés d'analyse statistique des polarisabilités.

- Toggle manuel d'un histogramme de distribution de valeur propre
- Différentes bibliothèques de paramètres statistiques
- Test d'Agostino K2 sur les distributions de valeur propre

5. Sortie d'un fichier txt avec le type atomique pour chaque atome du set
6. Calcul des polarisabilités moléculaire avec tenseurs moyens élément par élément
   - Transférés
   - Théoriques
   - Sortie de fichiers .txt avec les infos par type atomique
7. Calcul des polarisabilités moléculaires avec méthodes des vecteurs propres (Incomplet)
8. Sorties des fichiers ELMAM complétés par nos données
9. Transfert le set S66 par la base de données ELMAM SAK

III. Partie représentation graphique des comparaisons
1. Polarisabilités moléculaires isotropes et anisotropie
2. Polarisabilité moléculaire et anisotropie des 15 dimères du S66 transféré par ELMAMA SAK

3. Mode d'utilisation

Le code est construit pour traiter simplement une liste de fichiers moléculaire au format MoPro(.par).

La condition d'utilisation de la série de fichiers à traiter partie c. porte sur l'unicité des noms atomiques au seins d'un même fichier.

Le fichier de sortie généré en I. 6.(« *POL_OUTPUT_TRANSFORM.txt* ») regroupe toutes les infos nécessaire au tri des types atomiques. Le code ne requière aucune intervention manuelle jusqu'à ce stade. Il permet de créer tenseur et base pour chaque atome du set ainsi que de sortir un fichier txt regroupant tout le set donné et les toutes les informations par atome.

La partie II. Contient des éléments de code nous ayants permis de créer les types atomiques, cette dernière a nécessité un **tri manuel** pour plus de précision.

Néanmoins à partir de la partie II.3. le code fonctionne en autonomie en se basant sur le fichier généré en I.6 et avec un fichier de description des types atomiques ayant la syntaxe du fichier « *listee_type_atom_done.txt* » (cf : fichiers fournis) correspondant à votre base de données. Cette partie manuelle de tri des types atomiques serait automatisable avec plus de temps de travail.

Au vu des discussion dans le rapport de stage, nous avons créés plusieurs fichiers de code pour traiter différemment la série de donnée en présence. Certaines parties sont ajoutées (en bleu) à l'originale.

⚠ Même si cela est précisé dans le fonctionnement du code : Les fichiers *POL_OUTPUT_TRANSFORM.txt* et *listee_type_atom_done.txt* doivent être **dans le même dossier**

7. Approximation sur certains types atomiques.

La création des types atomiques dans notre méthode de travail s'est basée sur les fichiers format *MoProViewer* du *BigSet*. Travailler de cette manière ne nous renseigne pas sur la forme de la molécule (groupe cyclique etc.), on a rencontré plusieurs problèmes sur les groupes caractéristiques cycliques type benzène/uracile. En effet pour plusieurs atomes on trouve dans ces molécules cycliques des atomes avec des types atomiques identiques à ceux trouvés dans des molécules linéaires, auxquels correspondent des tenseurs différents. Notre algorithme n'est pas assez élaboré pour ce genre de subtilité, nous avons donc fait le choix d'affecter aux valeurs de configuration cyclique, les polarisabilités de configuration non cyclique (ex : C403/C403T). D'autres cas pose un problème avec la construction de notre algorithme, nous avons donc fait les choix suivants :

Cas du C403T/C403

On a affecté au type C403T les valeurs de polarisations du type C403, le type C403T.

Cas du N401/N306

Ici la présence d'une liberté sur les voisins éloignés possibles (wildcard « x » , dans les fichiers ELMAM) pose un problème, n'ayant pas accès à ce type d'information dans les fichiers originaux, on a été forcé d'affecter les polarisabilités moyenne correspondant à N401 dans le type N306.

Extrait de ELMAM_BASE_SAK

ATOM N306 N ZX C H - CHH Nc[hxx]hh WILDCARD x=c|h

ATOM N401 N ZX C H - CHHH Nchhh

Pour ces deux types, les paramètres multipolaires sont identiques.

Cas du O102/O102T

O102/O102T est l'oxygène impliqué d'un un groupe amide, O102 est exclusivement présent dans les peptides, on considère les deux types comme identiques étant donné que leurs populations multipolaires et leur repère de définition sont identiques et que nous ne parvenons pas à construire un code de détection de forme peptidique. L'utilisation de wildcard dans ELMAM pose ici un souci.

Cas du H101/H117

La configuration géométrique du type H101 (H ZX C C, hydrogène lié au carbone dans des molécules aromatiques) est identique à celle de H117 de même que leurs populations multipolaires respectives. La différence se situe au niveau des voisins plus éloignés, un atome de type H117 ne se rencontre que dans l'éthène. On a donc considéré leur tenseur moyen comme identique.

Cas du N302/N302T

De même ici le cas cyclique N302T (N bXY C C) présent dans les uraciles est considéré équivalent au type N302 qui lui représente les azotes présents dans les peptides.