AVERTISSEMENT

Ce document est le fruit d'un long travail approuvé par le jury de soutenance et mis à disposition de l'ensemble de la communauté universitaire élargie.

Il est soumis à la propriété intellectuelle de l'auteur. Ceci implique une obligation de citation et de référencement lors de l'utilisation de ce document.

D'autre part, toute contrefaçon, plagiat, reproduction illicite encourt une poursuite pénale.

Contact : ddoc-theses-contact@univ-lorraine.fr

LIENS

Code de la Propriété Intellectuelle. articles L 122. 4
Code de la Propriété Intellectuelle. articles L 335.2- L 335.10
http://www.cfcopies.com/V2/leg/leg_droi.php
http://www.culture.gouv.fr/culture/infos-pratiques/droits/protection.htm

UNIVERSITÉ DE LORRAINE

Ecole Doctorale C2MP (Chimie-Mécanique-Matériaux-Physique)

Thèse

Présentée et soutenue publiquement pour l'obtention du titre de

DOCTEUR DE l'UNIVERSITE DE LORRAINE

Mention : « Physique »

par **Théo LEDUC**

Vers un Potentiel Multipolaire Quantitatif et Transférable aux Macromolécules Biologiques : Une Étude Méthodologique des Effets de la Polarisabilité

Le 19 Décembre 2019

Membres du jury :

Rapporteurs :	**Mme Maria MITEVA**	**Directrice de Recherche INSERM, MTi, Université Paris-Diderot, Paris**
	Mr Nour-Eddine GHERMANI	**Professeur, Institut Galien Paris-Sud, Université Paris-Sud, Paris**
Examinateurs :	**Mme Francesca INGROSSO**	**MCF HDR, LPCT, Université de Lorraine, Nancy**
	Mr Éric CHABRIERE	**Professeur, MEPHI, Aix-Marseille Université, Marseille**
	Mr Benoît GUILLOT	**Professeur, CRM², Université de Lorraine, Nancy, directeur de thèse**
Membres invités:	**Mr Enrique ESPINOSA**	**Professeur, CRM², Université de Lorraine, Nancy**

UMR CNRS 7036, Laboratoire de Cristallographie, Résonnance Magnétique et Modélisations, Faculté des Sciences et Technologies de Nancy – Université de Lorraine BP 70239 – 54506 Vandœuvre-lès-Nancy Cedex

Sommaire

Résumé vulgarisé / Intelligible Abstract **7**

Remerciements / Acknowledgments **9**

Abréviations **10**

Introduction **11**

État de l'Art **17**

1 - Cristallographie par diffraction des rayons X **19**

2 - Le facteur de structure **23**

3 - Modèle multipolaire de la densité électronique **27**

4 - Potentiel électrostatique **29**

5 - Énergie d'interaction électrostatique **30**

6 - Topologie de la densité **31**

7 - Topologie du potentiel **33**

8 - Induction dipolaire & polarisabilité **35**

9 - Modèle d'Applequist **37**

10 - Modèles atténués (Thole, Birge) **41**

11 - Modèle topologique de la polarisabilité **44**

12 - Calculs d'énergie de polarisation intermoléculaire E_{POL} **46**

13 - Bases de données **49**

14 - Estimation de la « distance » entre tenseurs de polarisabilité **51**

15 - Cinétique enzymatique et titration isotherme par calorimétrie **52**

Chapitre 1 : Polarisabilités Atomiques Topologiques 59

1.1 - *MoProSuite* et son module de polarisation **62**

1.2 - Transfert et comparaison de polarisabilités atomiques anisotropes **63**

1.3 - Méthode d'Umeyama **65**

1.4 - Permutations : génération et composition **67**

1.5 - Molécules Fantômes **74**

1.6 - Rapport de comparaison entre tenseurs de polarisabilité atomique **77**

1.7 - « Polarization of electron density databases of transferable multipolar atoms » *Partie 1 : Polarisabilités* **79**

1.7.a - Base de données de référence **79**

1.7.b - Calcul des polarisabilités moléculaires et atomiques **80**

1.7.c - Polarisabilités moléculaires théoriques **81**

1.7.d - Polarisabilités atomiques isotropes **83**

1.7.e - Polarisabilités anisotropes par éléments chimiques **87**

1.7.f - Polarisabilités anisotropes par types atomiques ELMAM2 ..... **92**

1.7.g - Polarisabilités et incertitudes sur les populations dipolaires ELMAM2 .... **95**

1.7.h - Variations des valeurs propres de la polarisabilité **97**

1.7.i - Variations d'orientation des polarisabilités **99**

1.8 - Amélioration du set de polarisabilités atomiques théoriques **101**

Chapitre 2 – Polarisation *in Situ* d'Atomes Multipolaires 107

2.1 - Procédure de polarisation **109**

2.2 - Contrôle par l'utilisateur **112**

2.3 - *Output* et visualisations **117**

2.4 - « Polarization of electron density databases of transferable multipolar atoms » *Partie 2 : Polarisation et Énergies* **122**

2.4.a - Interaction électrostatique permanente **122**

2.4.b - Transfert de charge et autres contributions à l'induction de moments dipolaires **128**

2.4.c - Énergies d'induction : détermination de la meilleure méthode de calcul des polarisabilités **132**

2.4.d - Dipôles induits moléculaires moyens **134**

2.4.e - Calculs d'E_{POL} et E_{IND} sur les ensembles de données S66 & S66x8 ..... **136**

2.4.f - Somme des contributions électrostatiques permanentes et induites **141**

2.4.g - Répartition des erreurs pour les calculs d'énergie **144**

2.5 - Conclusion de l'étude sur les jeux de données S66 et S66x8 ..... **145**

Chapitre 3 - Construction et Application d'une Base de Données de Polarisabilités Atomiques Transférables **149**

3.1 - L'approche des bases de données : considérations générales liées à la librairie ELMAM2 ..... **152**

3.2 - Création d'une base pELMAM3 (polarizable Experimental Library of Multipolar Atom Models) **155**

3.3 - Détection de boucles **158**

3.4 - Moyennes de tenseurs définis positifs **163**

3.5 - Ajout de contraintes de symétrie à la méthode précédente **170**

3.6 - Construction de librairie de polarisabilités indépendantes **171**

3.7 - Applications biochimiques **179**

3.7.a - Interactions protéine-ligand : généralités et modélisations **179**

3.7.b - L'interaction streptavidine-biotine **182**

3.7.c - Comparaison des énergies d'interactions des streptavidines mutantes ... **188**

3.7.d **-** Conclusions partielles **205**

3.7.e - L'interaction streptavidine-biotine du point de vue topologique **212**

3.7.f - Conclusions sur l'approche topologique **224**

Chapitre 3-bis : Améliorations Contextuelles de la Base de Données ELMAM2 .... **231**

Chapitre 4 : Au-Delà du Modèle des Polarisabilités Indépendantes, et Autres Perspectives **253**

4.1 - Trypsine et arginase **255**

4.2 - Application à un cluster de complexe urée-éther-couronne **260**

4.3 - Énergie de pénétration liée au phénomène de polarisation **264**

4.4 - Cartes de recouvrement de densité électronique, méthodes de calcul des volumes des isosurfaces **268**

4.5 - Atténuation anisotrope et adaptation du tenseur de Birge au modèle HCM **271**

4.6 - Molécules velues **276**

Conclusions Générales **283**

Annexes **291**

Références **312**

Résumé / Abstract **324**

Résumé Vulgarisé.

Le terme ligand désigne une molécule capable de se fixer sur une protéine. L'étude de cette fixation est utile au design de nouveaux médicaments mais est aussi d'intérêt fondamental pour comprendre le fonctionnement de ces protéines qui assurent de nombreuses fonctions essentielles de la cellule. Ce manuscrit présente différents outils pour simuler une contribution particulière de cette interaction : celle entre dipôles induits qui émane des forts champs électriques que les molécules génèrent dans leur voisinage. C'est le phénomène de polarisation. La modélisation de cet effet utilise et confronte une combinaison inédite de données expérimentales (cristallographie à rayons X) et théoriques (simulations quantiques). Le modèle développé ouvre des possibilités novatrices d'analyse des interactions protéines-ligand.

Intelligible Abstract.

The term ligand refers to a molecule capable of binding to a protein. The study of this fixation is useful for drug design but is also of fundamental interest to understand the inner workings of proteins, which provide many essential functions of the cell. This manuscript presents various tools to simulate a specific contribution to this interaction: the interaction between induced dipoles. This interaction arises from the strong electric fields that the molecules generate in their neighborhood. This is the polarization phenomenon. The modeling of this effect uses and confronts an unprecedented combination of experimental data (X-ray crystallography) and theoretical data (quantum simulations). The developed model opens up new possibilities for analyzing protein-ligand interactions.

Remerciements / Acknowledgments

First and foremost, I would like to warmly thanks Anna Krawczuk, Emmanuel Aubert *and* Katarzyna Rzesikowska*. The precious data they provided had an tremendous impact on this project: without their calculations, the benchmarkings and the proof-of-concept studies presented in these pages would have been limited to rough outlines. Instead, their work allowed to showcase the full potential of the tools I developed during this PhD. For this, I can only express gratitude and a deep respect for their skills. For the fascinating scientific discussions, I thank them all.* Anna*,* Katarzyna*, I hope we will collaborate again in the future. Because of your sympathy and fellowship, your stays in France left cherished memories.*

Je souhaite remercier Maria Miteva et Nour-Eddine Ghermani pour avoir accepté d'être rapporteurs de ce manuscrit. Également, merci à Fransceca Ingrosso et Éric Chabrière d'avoir accepté de faire partie du jury.

Ensuite, j'aimerais remercier Benoît Guillot, mon directeur de thèse, pour m'avoir laissé une chance de faire mes preuves, il y a bien longtemps maintenant. Benoît m'a toujours laissé poursuivre mes propres intérêts scientifiques, sans négliger la transmission de son bagage exceptionnel de cristallographe et de physicien. Grâce à sa rigueur, sa grande culture scientifique, sa clarté d'esprit, les passionnantes lectures qu'il m'a conseillées, et également son amitié, je considère Benoît comme un véritable mentor. Merci pour tout, mais surtout, merci de m'avoir montré qu'il y a peut-être une place, quelque part, pour moi et pour mon savoir-faire.

À mes collègues du bureau 320, Vedran Vuković, Julien Cappele et Mathieu Schwartz, merci pour votre sympathie, pour nos passionnants échanges scientifiques mais aussi pour les éclats de rire. Je tiens à remercier Hassan Koder, avec qui j'ai eu la chance de découvrir l'ESRF. Eleonora Carocci mérite une mention spéciale : merci pour le soutien, que j'espère mutuel, tu es une véritable amie, *Norita*. Merci également à mes autres ami.e.s docteur.e.s, en particulier Heidi Schwartz et Dafné Wilhelm, mais aussi Fabien Deligey. Pascale Tsan et Axel Gansmüller méritent amplement ma gratitude, pour leur confiance avec les étudiant.e.s de l'IUT Nancy-Brabois. Enseigner a été une expérience forte et formatrice de ces trois ans passés au laboratoire, j'espère avoir été à la hauteur. Pour leurs inestimables conseils scientifiques et leur grande expérience, je souhaite remercier chaleureusement Enrique Espinosa et Christian Jelsch. Merci également à El-Eulmi Bendeif et Sébastien Pillet, grâce à vous j'ai eu l'opportunité de découvrir avec émerveillement le travail au synchrotron de Grenoble. Pour nos interventions communes de vulgarisation scientifique, je souhaite remercier Gwendall Kervern, Nicolas Claiser, Emmanuel Wenger (ainsi que Benoît, Axel et El-Eulmi). Une mention spéciale revient à Étienne Haouy, merci pour m'avoir impliqué dans ces multiples évènements, c'était un plaisir et une chance. À tous les membres du CRM² que je n'ai plus la place de citer (Dominik Schaniel, Akihiro Umayahara, Samuel Ahoulou, Maxime Deutsch, Pierrick Durand, Claude Lecomte…), merci à tous pour votre accueil chaleureux, les repas, cafés, opinions et rires partagés. Ce laboratoire est une gemme.

À mes parents, Christine, Christophe, merci pour votre soutien inconditionnel. À mes ami.e.s, Mathilde M., Kostia F., Sacha D., Antoine V., Caroline B., Adrien P., Nicolas O., Christine S., Mehdi T., Zacharia D., Jonathan G., Lou F., Thibaut J., Maxime G., Anne-Sophie S., Manon P., et je ne peux tous les nommer, ce travail vous est dédié. Vous m'avez aidé à garder la force de me battre chaque jour. À l'avenir, que je puisse partager les fruits de cette détermination avec vous.

Abréviations

ADP, *atomic displacement parameter*, paramètre de déplacement atomique ;

BP, *base pair*, paire de bases ;

BPL, *bond path length*, longueur du chemin d'interaction ;

BCP, *bond critical point*, point critique de la liaison ;

DB, *database*, base de données ;

DFT, *density-functional theory*, théorie fonctionnelle de la densité ;

ED, *electron density*, densité électronique ;

HCMM, modèle multipolaire d'Hansen et Coppens ;

ELMAM, *experimental library of multipolar atom model* ;

EP/MM, potentiel exact / modèle multipolaire ;

ESP, *electrostatic potential*, potentiel électrostatique ;

F2_25, *N*-(2,5-fluorophényl)-benzamide ;

Iso., isotrope ;

ITC, *isothermal titration calorimetry*, titration isotherme par calorimétrie ;

MARE, *mean absolute relative error*, erreur moyenne relative absolue ;

QTAIM, *Quantum Theory of Atoms In Molecules* ;

SSD, *sample standard deviation*, écart-type standard de l'échantillon ;

RMSD, *root mean square deviation*, racine de la déviation quadratique moyenne ;

SAPT, *symmetry-adapted perturbation theory* ;

TS, T-stack ;

UBDB, *University of Buffalo Database* ;

WT, *wild-type*, variant sauvage.

Introduction

Les relations entre la structure d'un objet macroscopique, de la vie courante et ses propriétés sont souvent triviales. L'esprit humain associe instinctivement ces propriétés à des fonctions. Qui n'a jamais détourné un outil de son usage initial ? Il est possible d'ouvrir une boite de conserve avec un clou et un pot de confiture, simplement en donnant une nouvelle fonction à la structure de ces objets. Lorsque l'on parle d'objets biologiques par contre, des tissus aux biomolécules en passant par les cellules, les moyens à mettre en œuvre pour découvrir les relations structure-fonction explosent en complexité. Les enjeux liés à la découverte de ces relations explosent également, touchant des domaines cruciaux tels que par exemple la pharmacologie et la médecine[1] ou l'écologie[2].

La biologie structurale est la branche de la science qui s'intéresse à produire des structures 3D – plus ou moins précises – des molécules complexes qui constituent les êtres vivants. Ces images sont souvent complétées par des modèles afin de rendre tangibles ces objets dont la taille est de l'ordre du nanomètre. À cette fin, la biologie structurale utilise des techniques telles que la cristallographie par diffraction des rayons X, la résonnance magnétique nucléaire, la microscopie électronique ou la spectrométrie de masse, qui sont bien évidemment complémentées par les informations obtenues de la biochimie, de la biologie moléculaire et cellulaire. La biologie structurale est une inestimable source d'informations pour la découverte des mécanismes qui sous-tendent le fonctionnement de chaque être vivant : respiration, photosynthèse, immunisation… Ces processus biologiques, et tant d'autres, sont de nos jours connus à l'échelle moléculaire, même s'il reste de nombreuses découvertes à réaliser.

La production de cette information structurale, disponible pour tous grâce à des initiatives telles que la *Protein DataBank* (PDB[3]), est une des prérogatives de l'équipe BioMod au laboratoire CRM^2, où cette thèse a été réalisée. Le laboratoire CRM^2 est spécialisé dans la cristallographie par diffraction des rayons X et la résonnance magnétique nucléaire. L'utilisation de données cristallographiques est centrale à cette thèse. Une autre mission de l'équipe BioMod est de développer et d'améliorer la modélisation de données cristallographiques, en particulier pour les études de cristallographie quantique (un domaine anciennement appelé « densité de charge »), qui concerne donc la détermination de la structure et des propriétés des molécules à l'échelle subatomique, et l'adaptation de ces outils aux macromolécules biologiques.

En effet, le CRM^2 est loin de se limiter à l'étude des molécules biologiques. La plupart des études cristallographiques à très haute résolution concernent des composés organiques ou organométalliques[4]. C'est dans ce contexte que l'application du paradigme QTAIM (*Quantum Theory of Atoms In Molecules*)[5,6,7,8,9] à des densités électroniques expérimentales de très haute qualité a été développé – entre autres laboratoires – au CRM^2. Ce paradigme s'intéresse principalement à la forme (topologie) du champ de densité électronique, qui est observé lors de l'expérience de cristallographie.

Le présent travail s'inscrit dans un effort de recherche à long terme qui vise à appliquer ce paradigme topologique aux macromolécules biologiques. Des publications récentes[10] montrent que le sujet est un domaine de recherche actif. L'objectif général de cette thèse est de permettre l'étude topologique du champ de potentiel électrostatique (calculé entre autres à partir du champ de densité électronique), au sein de macromolécules biologiques. La topologie du potentiel électrostatique est moins développée que le paradigme QTAIM qui s'intéresse à la densité électronique. Cette nouvelle méthode de description des structures observées n'a pas encore été évaluée pour sa capacité à expliquer à des relations structure-fonction dans un contexte biochimique.

L'objectif pratique de ce travail de recherche est donc – dans un premier temps – de rendre applicable la description topologique au champ de potentiel électrostatique, en contribuant au développement d'outils spécialisés. Dans un premier temps, cet objectif sera restreint au site de fixation de ligands (petites molécules) sur une protéine d'intérêt. Les études de cristallographie quantique sur des complexes protéiques sont encore exceptionnelles[11,12,13]. Les cristaux de protéine sont en général beaucoup moins réguliers, organisés et volumineux que les cristaux de petites molécules. La conséquence est une qualité inférieure des données de diffraction X et donc de la densité électronique, rendant impossible l'application du paradigme QTAIM ou le calcul d'un potentiel électrostatique précis. Cet état de fait impose de reconstruire la densité électronique au-delà des limites imposées par la résolution cristallographique pour les macromolécules.

L'obtention de cette densité de haute qualité passe par une approche fondée sur une base de données de densité électronique, expérimentale et *transférable*. Ce concept de transférabilité, détaillé par la suite, peut se résumer comme une approximation raisonnable de la densité électronique, dont les points communs ont été observées dans

différents cristaux. La densité électronique moyenne est représentée par un ensemble de paramètres atomiques ; ces paramètres sont stockés dans la base de données. Cette densité moyenne est ensuite utilisée telle qu'elle, sur la structure à reconstruire, à la manière de blocs de construction.

C'est à ce moment que la méthode présentée ici devient hybride – c'est-à-dire comportant des éléments expérimentaux et théoriques – car pour obtenir un potentiel électrostatique correct, il convient d'assumer les limites du principe de transférabilité et de tenter de *contextualiser* la densité électronique simulée, la rendre sensible à son environnement chimique après qu'elle y ait été transférée. Cette interaction d'une molécule avec son environnement passe par l'évaluation du champ électrique, et le calcul de déformations de la densité électronique en réaction à ce champ extérieur. Cette réorganisation des électrons est appelée ici polarisation.

Déjà en 2016, mon projet de master 2 au CRM^2 portait déjà sur une première tentative de l'utilisation d'un modèle simple pour prendre en compte ces effets de polarisation. L'objectif général d'obtenir le meilleur potentiel électrostatique possible à partir de densités électroniques transférables s'est donc très rapidement orienté vers ces effets de polarisation *in situ*. Obtenir une densité électronique la plus fiable possible est le principal objectif, en tenant compte de ces effets de polarisation, et donc obtenir une topologie exploitable du champ de potentiel électrostatique au sein du site de fixation des protéines.

Le potentiel électrostatique est une propriété moléculaire d'intérêt : il est en effet connu que les forces de nature électrostatiques jouent un rôle prépondérant dans de nombreuses de leurs fonctions (reconnaissance moléculaire, formation de complexes…). Cependant la motivation à poursuivre cet objectif s'étends à la prédiction d'autres propriétés moléculaires. Certaines propriétés optiques (par exemple, l'indice de réfraction) sont intrinsèquement liées au modèle développé ici, du fait de la nature électromagnétique de la lumière.

Un intérêt tout particulier sera donné aux prédictions d'énergies d'interactions électrostatiques. La polarisation correspond, dans notre modèle attaché à la densité électronique, à une reconfiguration dans l'espace des électrons, chargés négativement, autour des noyaux chargés positivement. Appliquer le champ électrique aux densités

électroniques transférables rapproche les électrons des zones où le potentiel est plus positif. Cette réorganisation aboutit donc à un état plus stable, de moindre énergie, du système simulé.

En effet, toute évolution d'un système chimique dépend fondamentalement de différences d'énergie entre l'état initial et l'état final. Plus encore, la vitesse de ces changements est dictée par ces balances énergétiques. D'un point de vue très pragmatique, ce phénomène explique la capacité de certains médicaments que nous absorbons à se coller à une seule cible et altérer son activité : les inhibiteurs de la protéase du VIH (utilisés dans la trithérapie) se fixent à cette enzyme-clé du cycle viral[14] ; les inhibiteurs sélectifs de la recapture de la sérotonine atténuent les symptômes de la dépression en bouchant physiquement une pompe moléculaire, la SERT[15]. Pour assurer cette présence au cours du temps, la forme de ces molécules a été sélectionnée pour abaisser leur énergie au maximum lorsqu'elles sont au contact de leur cible. C'est pour cette raison que la détermination d'énergies d'interactions électrostatiques précises est un objectif capital de cette thèse et l'importance de la polarisation sera abordée pour décrire ces interactions à courte portée, presque de contact.

Le lien entre énergie et densité électronique mesurée par cristallographie est une question de recherche contemporaine. La *Cambridge Structural Database* (CSD[16]) recense et stocke les structures cristallographiques à l'échelle mondiale. Cette organisation promeut chaque année un concours de prédiction de structure cristalline[17]. L'intérêt fondamental est de comprendre exactement comment une même molécule peut générer des cristaux différents, appelés polymorphes. Le domaine de l'ingénierie cristalline s'attache à cette question, afin de produire de nouveaux matériaux cristallins, pour leurs propriétés optiques[5] ou pharmacocinétiques[18] par exemple. Les méthodes présentées ici sont satellites – et donc complémentaires – de ces efforts, vu que cette thèse offre une utilisation différente des densités électroniques « cristallographiques » pour prédire des énergies d'interaction.

Les différentes méthodes utilisées pour construire le modèle de densité électronique, autant expérimentales (cristallographie, transférabilité) que théoriques (polarisation, QTAIM, etc.) sont détaillées dans leurs sections dédiées. Comme pour de nombreux autres projets de thèse, ce qui devait relever de l'adaptation d'un modèle bien connu à un contexte nouveau s'est transformé en une enquête méthodologique poussée.

Du fait de contraintes temporelles, le modèle reste inachevé à ce jour mais cette enquête a d'ores et déjà été fructueuse en termes de publication et de développement logiciel. La publication de résultats présentés – ou ébauchés – ici continuera donc dans un futur proche. L'évaluation des moyens pour obtenir ce potentiel électrostatique tenant compte des effets de polarisation, et les conséquences de cette modélisation font bien entendu partie intégrante de cet objectif. Pour rendre compte de la nature hybride de cette thèse, la reproduction de bancs de test théoriques et de propriétés expérimentales est mise à profit.

État de l'art

Comme il a été vu dans l'introduction, le travail présenté touche à des domaines hétéroclites. Des questions liées aux liaisons chimiques, à la thermodynamique ou à l'électrostatique se mélangent à des questions d'implémentation logicielle ou de pertinence biochimique. Il est donc essentiel de faire un état des connaissances ordonné. La base du projet repose sur la détermination de structures cristallines par diffraction des rayons X, qui sera introduite de prime abord. Ensuite, il est nécessaire d'accorder une attention particulière à la description des structures à résolution subatomique à l'aide du modèle multipolaire d'Hansen et Coppens (HCM). Une introduction au paradigme QTAIM, et son extension à la topologie du potentiel électrostatique permettra la transition vers le centre du projet, qui concerne les polarisabilités et le phénomène de polarisation de la densité électronique multipolair. Les autres éléments méthodologiques nécessaires pour obtenir les énergies d'interactions électrostatique et d'induction seront passés en revue. Les méthodes de référence utilisées pour évaluer les calculs présentés, d'origine théorique ou et expérimentale, seront également couvertes.

1 - Cristallographie par diffraction des rayons X

La théorie de la diffraction s'intéresse à l'interaction entre les ondes et, notamment, la matière périodique. Elle est particulièrement adaptée à l'étude de la matière cristalline, c'est-à-dire la cristallographie. De manière routinière, la cristallographie utilise des rayons X (rayonnement de longueur d'onde λ entre 10 nm et 0.01 nm), dont la production est aisée sur des machines de laboratoire. L'interaction entre ce type de rayonnement et la matière cristalline produit un phénomène de diffraction permettant de sonder le contenu atomique, ou moléculaire, des cristaux. Il est également possible de réaliser les mêmes expériences avec des électrons (comme par exemple en microscopie électronique *dark field*), ou des neutrons (ce qui nécessite un réacteur nucléaire pour leur production). La découverte de la diffraction par des cristaux est date de l'année 1912, quand Max von Laue découvre le phénomène de diffraction des rayons X qui lui vaudra le prix Nobel de physique 1914. Ce phénomène – montré Figure 1 – permit de confirmer le caractère ondulatoire des rayons X et amènera William et Lawrence Bragg à utiliser le caractère ordonné et périodique des cristaux pour trouver les premières structures atomiques.

Le cristal est un état de la matière, une phase condensée dont la particularité est de présenter un ordre à longue portée, et le plus souvent, une périodicité de l'agencement de ses constituants. Un cristal est formellement défini comme la convolution d'un motif et d'un réseau. Le motif est ici un ou plusieurs atomes ou molécules, qui se répètent régulièrement dans les trois directions de l'espace. Le réseau représente l'agencement du motif par des opérations de translation (comme un mur de briques). Dans l'espace tridimensionnel, on dénombre 14 empilements différents (réseaux de Bravais), qui décrivent les 14 façons de paver l'espace 3D. Le motif répété est la maille élémentaire, qui peut elle-même avoir une organisation interne. Une maille peut contenir plusieurs fois le motif répété si représenter cette répétition demande d'autres opérations de symétrie (rotations, miroirs, centres d'inversions…). Cette description de la symétrie locale utilise les 230 groupes d'espace qui recensent toutes ces combinaisons de symétries compatibles avec la symétrie de translation des réseaux. Dans la maille, le motif répété par les opérations du groupe d'espace est appelé unité asymétrique. La convolution de l'unité asymétrique et du groupe d'espace donne ainsi la maille, alors que la convolution de la maille et du réseau donne le cristal.

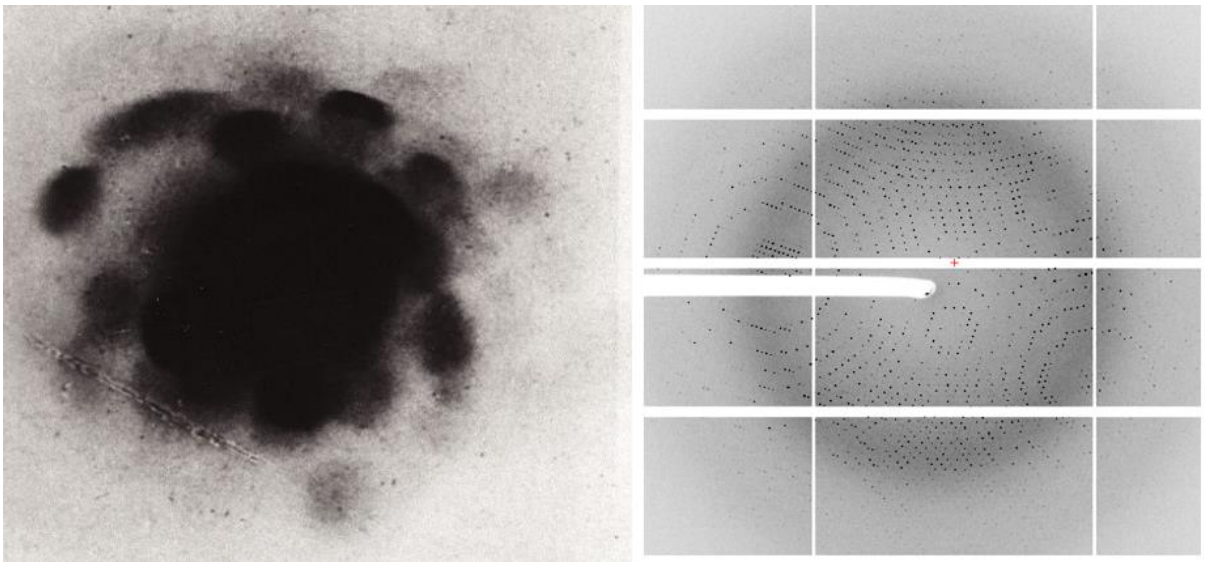


Figure 1. Le premier cliché de diffraction de l'histoire, pris par Max von Laue, à partir d'un cristal de $ZnSO_4$ (à gauche). Un cliché de diffraction d'un cristal de protéine OrfG, enregistré au synchrotron SOLEIL, ligne Proxima-2 (à droite, image gracieusement fournie par Julien Cappele). Plus de 106 ans de raffinement des techniques expérimentales en cristallographie séparent les deux clichés.

Pour comprendre la position des taches de diffraction (réflexions), il est nécessaire de définir le concept de plan diffractant ou réticulaire. La nature ondulatoire de la lumière crée des interférences lorsqu'elle passe dans un réseau dont le pas est de l'ordre de grandeur de sa longueur d'onde. C'est l'expérience des fentes d'Young. Au sein du cristal, la longueur d'onde des rayons X leur permet d'interagir avec les atomes qui vont diffuser les rayons. Les espaces périodiques entre les atomes du cristal vont alors se comporter comme un réseau dans une expérience classique de diffraction. On peut mesurer ainsi le réseau qui constitue le cristal. Dans un réseau, on définit les plans réticulaires comme tout plan passant par 3 nœuds du réseau. Les vecteurs normaux à ces plans définissent des familles de plans réticulaires, qui partagent le même vecteur normal, et sont séparés par la distance interréticulaire d, qui apparaît dans la loi de Bragg :

$$2d \cdot \sin(\theta) = n \cdot \lambda \qquad (\mathbf{1})$$

Où θ est le demi-angle entre le faisceau transmis et le faisceau diffracté considéré et n est un entier, l'ordre de diffraction. Chaque famille de plan diffracte l'onde incidente comme un réseau, créant des franges d'interférence espacées périodiquement. Le minimum mesurable de d correspond à la résolution spatiale maximale. Les taches de diffraction sur le panneau droit de la Figure 1 sont régulièrement espacées. Pour décrire cet agencement régulier, on utilise un espace virtuel, l'espace réciproque.

L'espace direct en cristallographie est décrit par les 3 vecteurs de base de la maille, **a**, **b** et **c**. Le volume de la maille est $V = |\det(\mathbf{a}, \mathbf{b}, \mathbf{c})|$. L'espace réciproque est pavé par une maille réciproque décrite par les vecteurs **a***, **b*** et **c*** (où '⨯' représente le produit vectoriel) :

$$\mathbf{a}^* = \frac{(\mathbf{b} \times \mathbf{c})}{V} \;; \qquad \mathbf{b}^* = \frac{(\mathbf{c} \times \mathbf{a})}{V} \;; \qquad \mathbf{c}^* = \frac{(\mathbf{a} \times \mathbf{b})}{V} \qquad (\mathbf{2})$$

Les distances dans l'espace réciproque se mesurent donc en $Å^{-1}$. On peut interpréter cette unité comme une fréquence spatiale et, en effet, le réseau réciproque est la transformée de Fourier du réseau direct. Les réflexions sont situées au nœuds de l'espace réciproque et portent chacune un ensemble de coordonnées (h, k, l) relatives à **a***, **b*** et **c*** respectivement. (h, k, l) sont les indices de Miller. Pour cartographier l'espace réciproque, on utilise la sphère d'Ewald. La sphère d'Ewald est une construction géométrique dans l'espace réciproque de rayon $1/\lambda$ qui encapsule les conditions de

réflexion (existence d'une tache de diffraction). Si un nœud de l'espace réciproque est sur la sphère, une tache de diffraction apparaît. La sphère contient l'origine du réseau réciproque, puisque la réflexion (*0,0,0*) est toujours présente. Le centre de la sphère est dans la direction du faisceau incident (voir Figure 2). Pour cartographier le réseau réciproque, on tourne la sphère autour de l'origine en tournant le cristal dans l'espace direct de façon à maximiser le nombre de nœuds du réseau réciproque se trouvant en condition de diffraction. On recensera de cette manière le plus grand nombre possible de nœuds de l'espace réciproque.

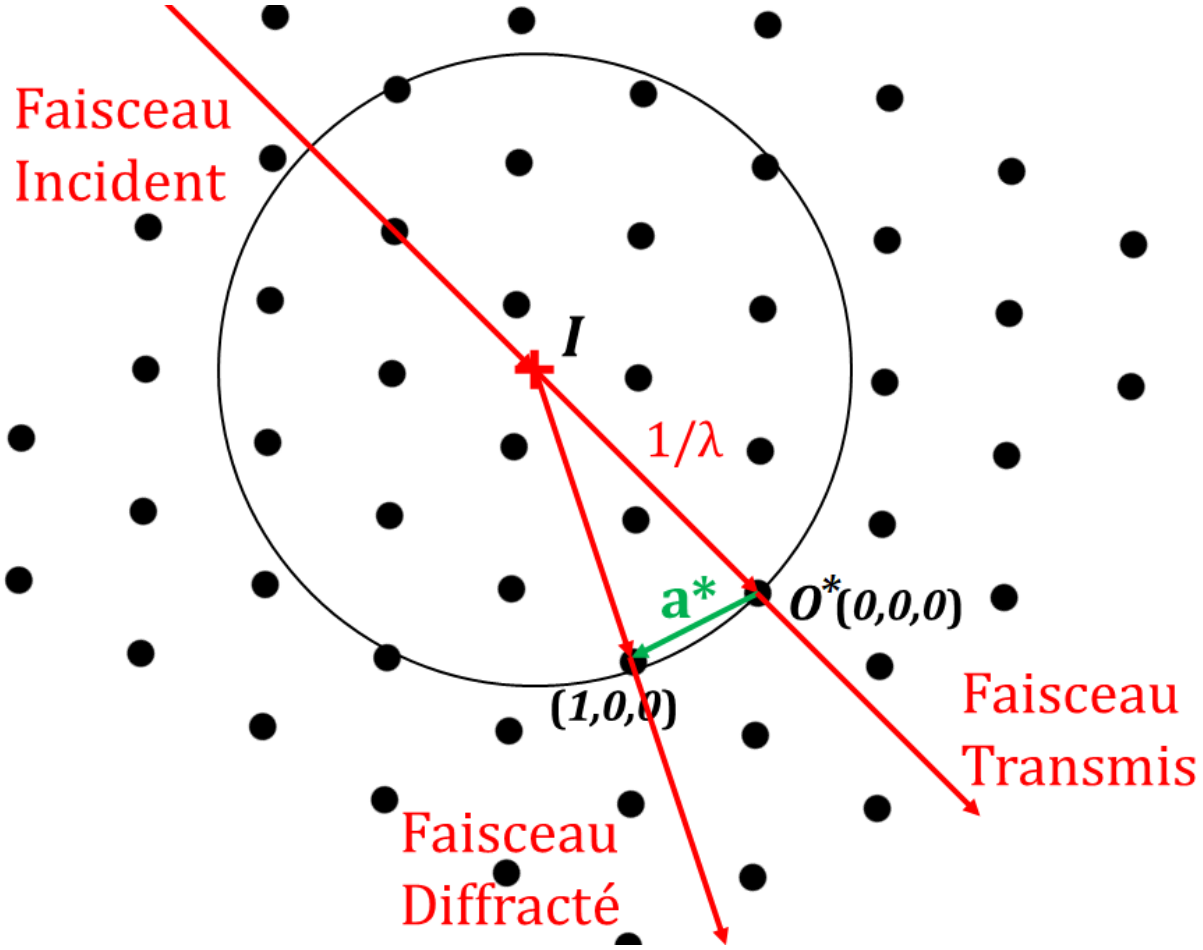


Figure 2. La sphère d'Ewald (de rayon $1/\lambda$) décrit les conditions de diffraction d'un cristal dans l'espace réciproque. Le cristal est formellement positionné en *I*. Tourner le cristal dans le faisceau incident revient à faire tourner le réseau réciproque autour de son origine, *O**. Les faisceaux diffractés apparaissent dès qu'un nœud du réseau réciproque intersecte la sphère d'Ewald.

2 - Le facteur de structure

Chaque réflexion diffractée par un cristal est associée à un nombre complexe $\boldsymbol{F}_{hkl}$, le facteur de structure. Seule l'intensité $\boldsymbol{I}_{hkl}$ de la réflexion est mesurable expérimentalement. Cette intensité $\boldsymbol{I}_{hkl}$ est proportionnelle au carré de l'amplitude $|\boldsymbol{F}_{hkl,obs}|^2 = \boldsymbol{F}_{hkl} \cdot \boldsymbol{F}_{hkl}^*$, où '*' désigne le conjugué complexe de $\boldsymbol{F}_{hkl}$.

$$\boldsymbol{F}_{hkl} = \sum_j f_j e^{2\pi i(hx_j + ky_j + lz_j)} \qquad (3)$$

Où la sommation court sur tous les atomes *j* de la maille, de facteur de diffusion f_j. Ce facteur de diffusion dépend de l'angle θ (on aura de faibles intensités aux grands angles et de fortes au petits angles) et décrit les propriétés de diffusion d'un atome isolé. $\boldsymbol{F}_{hkl}$ dépend également de la position de chaque atome dans la maille. Réécrire $\boldsymbol{F}_{hkl}$ comme une équation d'onde fait apparaître son amplitude (mesurable) et sa phase :

$$\boldsymbol{F}_{hkl} = |\boldsymbol{F}_{hkl}| \cdot e^{i\varphi(hkl)} \qquad (4)$$

où $\varphi(hkl)$ est la phase associée au rayon diffracté (h,k,l). Individuellement, chaque $\boldsymbol{F}_{hkl}$ contient une information sur l'onde diffusée par chacun des atomes, et donc, indirectement, sur leurs positions dans la maille. La Figure 3 matérialise comment le déphasage des ondes individuelles affecte l'amplitude de $\boldsymbol{F}_{hkl}$ (ainsi que sa phase).

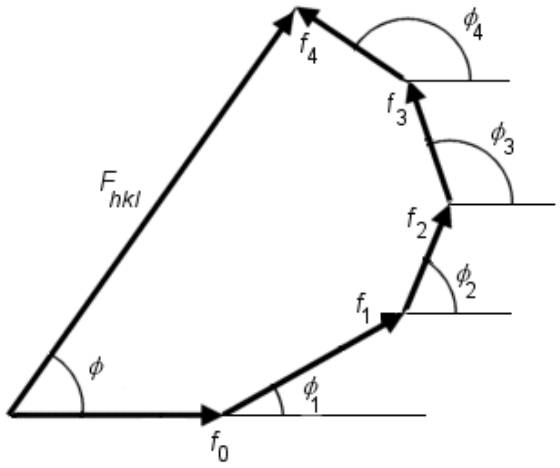


Figure 3. Les ondes diffusées par chaque atome de la maille, puis diffractées par le réseau, s'additionnent (de manière constructive) sur chaque réflexion. De même, le facteur de structure $\boldsymbol{F}_{hkl}$ de la réflexion est la somme des facteurs de diffusion atomique $f_{0,\ldots,4}$ des atomes de la maille (ici, 5), affectés de leurs phases individuelles. Reproduit de la page « *Structure Factor* » sur : http://reference.iucr.org.

Pour retrouver la phase associée à chaque onde diffractée, on cherche un ensemble de phases par un processus itératif. On l'a vu, la phase de chaque onde dépend des coordonnées de chaque atome de la maille. On crée donc un modèle virtuel du cristal et on calcule l'ensemble de ses $|\boldsymbol{F}_{hkl,calc}|$. On raffine le modèle à chaque étape pour minimiser l'écart quadratique entre les intensités $\boldsymbol{I}_{hkl,obs}$ observées et les intensités $\boldsymbol{I}_{hkl,calc}$ calculées. Le résiduel *R* suivant est souvent utilisé pour apprécier la qualité du modèle :

$$R = \frac{\sum_{h,k,l} abs(|\boldsymbol{F}hkl,obs| - |\boldsymbol{F}hkl,calc|)}{\sum_{h,k,l} |\boldsymbol{F}hkl,obs|} \qquad (\mathbf{5})$$

Où *abs*() désigne ici la fonction valeur absolue pour éviter la confusion avec l'amplitude de $\boldsymbol{F}_{hkl}$. D'autres méthodes d'affinement existent, qui ne fonctionnent pas par minimisation de l'écart quadratique (par exemple celles utilisant la maximisation de l'entropie[19]). $\boldsymbol{F}_{hkl,calc}$ est calculé directement à partir des paramètres du modèle. On utilise une opération de transformée de Fourier inverse pour reconstituer la densité électronique dans la maille à partir des $\boldsymbol{F}_{hkl,calc}$.

Pour calculer $\boldsymbol{F}_{hkl,calc}$, le facteur de diffusion f_j, qui représente la diffusion de l'onde par un atome isolé, est nécessaire à cette opération. f_j est dépendant de l'angle 2θ entre le faisceau diffracté et incident, et de la longueur d'onde λ, comme montré sur la Figure 4. Pour les rayons X, le facteur de diffusion f_j dépend du nuage électronique qui interagit avec le faisceau incident (en particulier les électrons de cœur). f_j est la transformée de Fourier de la densité électronique de l'atome isolé.

Dans tous les cas pratiques on commencera par utiliser un facteur de diffusion isotrope. C'est le modèle *IAM*, pour *Independant Atom Model*, qui assume une répartition sphériquement symétrique des électrons autour du noyau (à l'instar d'un atome isolé dans le vide). Le facteur de diffusion f_j, transformée de Fourier de cette densité électronique atomique sphérique, permet de lier la densité à la diffraction par l'édifice cristallin. Par extension, la densité électronique au sein de la maille est donc la transformée de Fourier inverse des facteurs de structure complexes.

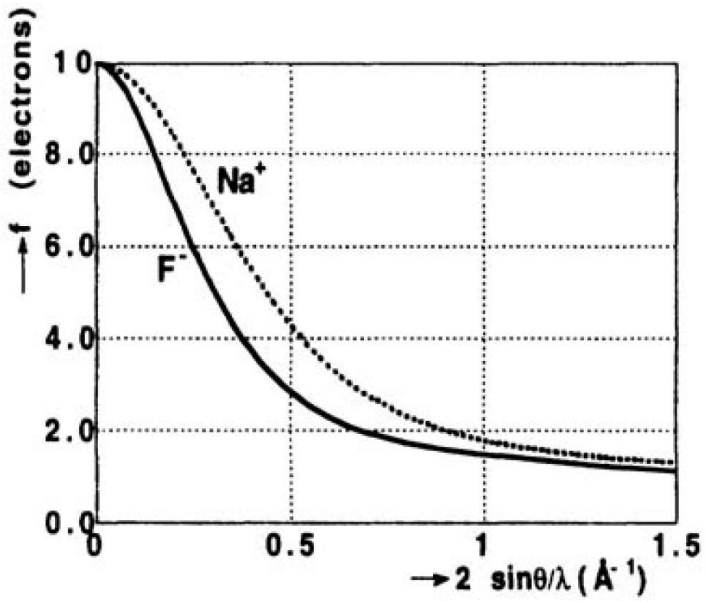


Figure 4. Facteur de diffusion atomique pour deux atomes iso-électroniques, F- et Na+. Dans les deux cas, la réflexion (*000*), c'est-à-dire le faisceau transmis, a une intensité proportionnelle au nombre d'électrons diffractant. En termes d'unités, le facteur de diffusion est exprimé proportionnellement au facteur de diffusion d'un électron isolé. Reproduit de « *X-Ray Charge Density and Chemical Bonding* »[20], chapitre I, page 11.

Le modèle *IAM* est une bonne première approximation car d'autres variables vont être prises en compte dans le modèle. Les deux principales sont liées au désordre du cristal réel : l'expérience de cristallographie permet d'observer la densité électronique, moyennée sur le temps (la durée de collecte) et l'espace (dans le cristal réel, les atomes sont agités de vibrations thermiques et toute les mailles ne sont pas forcément parfaitement identiques). Ce désordre peut être représenté de manière statique (conformations alternatives, représentées par les facteurs d'occupation) ou par le moyen de paramètres de déplacement atomiques (*ADP*, notés B_{eq} ou B_{ij} en fonction de leur caractère iso- ou anisotrope). L'agitation thermique des atomes est – entre autres – modélisé par les *ADP* (*Atomic Displacement Parameters*), dont l'amplitude moyenne diminue avec la température de collecte.

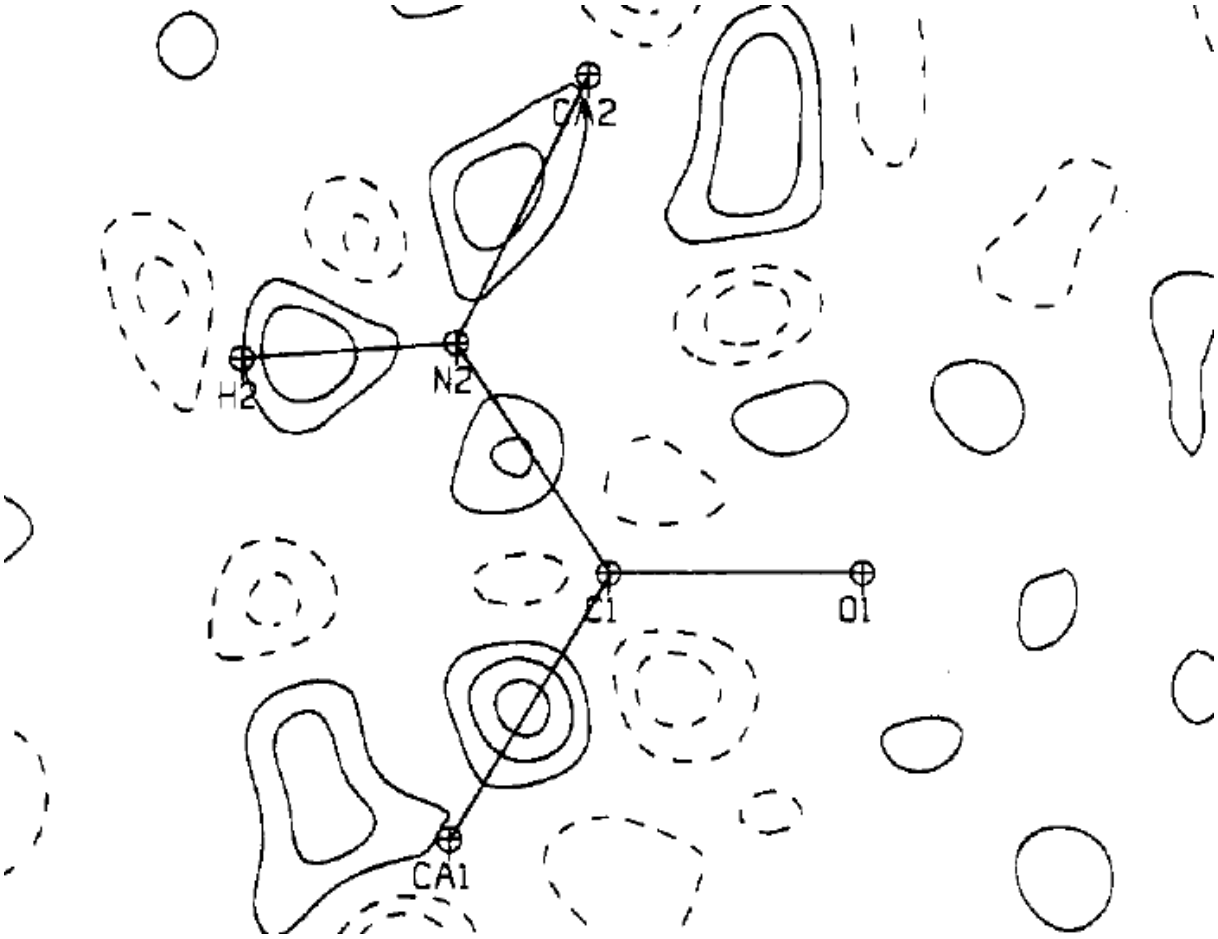


Figure 5. Densité résiduelle d'une liaison peptidique après un affinement *IAM*. Contours positifs en traits continus, contours négatifs en pointillés (±0.05 e·Å^{-3}). Reproduit de Pichon-Pesme et al., 1995[21].

Une fois le désordre pris en compte dans le modèle, il devient raisonnable d'envisager une description plus fine de la densité électronique. La densité résiduelle, présentée Figure 5, est caractéristique d'un cristal de haute qualité pour laquelle l'affinement *IAM* n'offre pas une description adéquate de la densité électronique. Pour obtenir ce résiduel, on utilise les phases $\varphi_{calc}(hkl)$, calculées à partir du modèle atomique, et la différence entre les amplitudes des facteurs de stucture observés lors de l'expérience et les amplitudes calculées à partir du modèle : $|\boldsymbol{F}_{hkl,obs}|-|\boldsymbol{F}_{hkl,calc}|$. Les contours positifs montrent notamment l'accumulation d'électrons au sein des liaisons covalentes, qu'un modèle d'atome sphérique ne peut représenter.

3 - Modèle multipolaire de la densité électronique

Le modèle multipolaire d'Hansen & Coppens (HCMM ou modèle HCM)[22] a été développé pour modéliser la densité électronique observable par la diffraction à haute résolution des rayons X sur monocristal. Ce modèle est au centre du présent travail. Le principal attrait du HCMM est de permettre la modélisation de densités électroniques lorsque la résolution des données de diffraction dépasse les capacités du modèle *IAM*. Le modèle HCM permet d'accéder aux facteurs de structure adaptés pour prendre en compte le type de densité électronique résiduelle présenté Figure 5. Pour la biocristallographie, la nécessité pratique d'aller au-delà de l'*IAM* est une exception rare, étant donné la qualité en général limitée des cristaux formés de macromolécules biologiques.

Le modèle HCM divise la densité électronique moléculaire en une somme de contributions pseudo-atomiques. La densité électronique de chaque pseudoatome est à son tour composée de 3 parties. Les deux premiers termes de l'équation (6) représentent des densités électroniques sphériques, associées respectivement aux électrons de cœur et aux électrons de valence du pseudoatome, et centrées sur son noyau. Le troisième terme, correspondant à la densité de déformation, modélise la déviation de la densité électronique par rapport à une répartition sphérique des électrons autour du noyau. Cette déviation est modélisée à partir de fonctions harmoniques sphériques, centrées sur le noyau et modulées par une fonction radiale R_{nl}.

$$\rho(\mathbf{r}) = \rho_{core}(r) + P_v\kappa^3\rho_{val}(\kappa r) + \sum_{l=0}^{l\,max}\kappa'^3 R_{nl}(\kappa' r)\sum_{m=0}^{+l} P_{l\pm m}y_{l\pm m}(\theta,\varphi) \quad (6)$$

Dans cette expression, ρ_{core} et ρ_{val} sont des densités électroniques de symétrie sphérique. Les deux utilisent les fonctions orbitales paramétrés par Roetti & Clementi[23]. ρ_{core} est considéré fixe durant l'affinement multipolaire. ρ_{val} est normalisée à un électron, le coefficient P_v permet l'ajustement de la population de valence à la densité électronique observée ou calculée. $y_{l\pm m}$ sont les fonctions harmoniques sphériques réelles, pondérées par les populations électroniques $P_{l\pm m}$. κ et κ' sont respectivement des coefficients d'expansion et de contraction pour la partie radiale des densités de valence et de déformation (R_{nl}). P_v, $P_{l\pm m}$, κ et κ' sont des paramètres ajustables lors de l'affinement cristallographique.

La fonction R_{nl} est une fonction de densité de type Slater, et répond à l'équation suivante :

$$R_{n_l}(r) = \frac{\zeta^{n_l+3}}{(n_l+2)!}\, r^{n_l} e^{-\zeta r} \qquad (7)$$

Où ζ est l'exposant de Slater, optimisé par Roetti & Clementi, 1974[23] et n_l est un entier naturel faisant office de nombre quantique principal. Les valeurs de n_l dépendent de l'ordre l de l'harmonique sphérique considérée et de l'espèce chimique du pseudoatome.

Les facteurs de diffusion $f_{j,core}$, $f_{j,val}$ et $f_{j,def}$ – respectivement – des densités ρ_{core}, ρ_{val} et de la densité de déformation sont additifs et peuvent être trouvés dans Coppens, 1997[20]. Bien entendu, le facteur de diffusion de la densité de déformation (orientée dans l'espace réel) sera anisotrope et dépendra des directions de l'espace réciproque considérées.

L'estimation de la charge atomique est possible au travers du paramètre P_v (charge multipolaire, à différencier d'autres méthodes décrites plus bas), ainsi qu'au travers de la population de déformation $P_{0,0}$ (en pratique $P_{0,0}$ n'est utilisé que pour la description de métaux de transition possédant une orbitale *s* très diffuse). Cette charge atomique est effectivement le premier moment électrostatique (ordre zéro) porté par le pseudoatome. Les populations de déformation d'ordre $l > 0$ modifient uniquement la répartition des électrons autour du noyau. De ce fait, seule cette densité de déformation décrit les moments électrostatiques d'ordre supérieur : dipôles ($l = 1$), quadrupôles ($l = 2$), octupôles ($l = 3$), etc. Cette description des moments électrostatiques comme de simples fonctions des paramètres multipolaires est d'un intérêt particulier pour le calcul de propriétés électrostatiques (potentiels, énergies et champ électrique)[20,24,25,26]. Le modèle HCM utilise uniquement des moments sans trace (*traceless convention*) de sorte que seules les harmoniques sphériques d'ordre l contribuent au moment électrostatique d'ordre l^*.

---

[*] *: On peut montrer que l'expression suivante :*

$$\boldsymbol{\mu} = \iiint \mathbf{r} \cdot \rho(\mathbf{r}) r^2 \sin\theta \, dr d\theta d\varphi$$

Cette représentation n'exclut pas l'émergence de moments moléculaires d'ordre *l* à partir de contributions atomiques d'ordre inférieur à *l* (exemple : l'agencement de charges atomiques dans une molécule peut contribuer au moment dipolaire moléculaire).

4 - Potentiel électrostatique

Le schéma EP/MM (*Exact Potential/Multipolar Model*)[27] est une méthode permettant de calculer le potentiel électrostatique et l'énergie d'interaction électrostatique en tant que somme d'une contribution à courte portée et d'une contribution à longue portée. Pour les atomes plus proches qu'un seuil fixé à 5 Å, une intégration explicite de la densité électronique est réalisée. On considère donc le potentiel au point **r** en tenant compte de la charge contenue dans chaque volume infinitésimal autour de **r**.

L'équation suivante est utilisée pour calculer le potentiel exact (EP) généré par une densité de charge moléculaire constituée d'une densité électronique $\rho(\mathbf{r})$ et d'un ensemble de *i* noyaux :

$$\varphi(\mathbf{r}) = \sum_i \frac{Z_i}{|\mathbf{r}-\mathbf{r}_i|} - \int \frac{\rho(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|} d\mathbf{r}' \qquad (8)$$

Où $\mathbf{r}_i$ est la position du i^{eme} noyau de charge Z_i. À des distances plus longues, le recouvrement des densités électroniques devient négligeable, le potentiel des moments électrostatiques ponctuels (MM) est utilisé, comme décrit par Buckingham[24,25,26]. Le calcul EP/MM est implémenté dans le module *VMoPro* de *MoProSuite*[28].

Bojarowski *et al.* ont utilisé le schéma EP/MM pour évaluer leur méthode rapide de calcul d'énergie (aug-PROmol)[29,30]. Un autre exemple d'utilisation d'un modèle explicite de densité électronique peut être trouvé dans Vandenbrande *et al.*, où un champ de force est construit à l'aide d'une fonction décrivant le terme de pénétration (pour tenir compte du chevauchement entre la densité de différentes molécules dans les régions intermoléculaires) et paramétré utilisant la méthode EP/MM[31]. Il est à noter que leur

---

S'annule pour les densités multipolaires sphériques de cœur et de valence, et pour toutes les densités de déformations de l ≠ 1.[20]

champ de force ne contient pas de terme de polarisation explicite. Dans leur travail, l'énergie de polarisation est calculée par une fonction empirique des distances interatomiques.

Dans la présente approche, des multipôles distribués sont utilisés pour calculer l'énergie d'interaction électrostatique entre deux entités (atomes ou molécules). On souhaite, dans ce travail, prendre en compte le contexte non-covalent des atomes dans ces interactions, en rapport avec le modèle de densité électronique décrit page 27 et page 49. Ces interactions passent par le calcul d'un champ électrique externe. Le calcul du potentiel électrostatique à partir de la densité multipolaire donne accès, par différenciation numérique (série de Taylor), au vecteur de champ électrique:

$$\mathbf{E}(\mathbf{r}) = -\nabla\varphi(\mathbf{r}) \quad (9)$$

5 - Énergie d'interaction électrostatique

E_{ELEC}, l'énergie d'interaction électrostatique EP/MM[27] entre une densité électronique ρ'(r) avec un ensemble de i' noyaux de charge $Z_{i'}$ et le potentiel précédemment défini $\varphi(\boldsymbol{r})$ – calculé à partir de l'autre densité de charge $\rho(\mathbf{r})$ – est décrite par l'équation (10).

$$E_{ELEC} = \sum_{i'} Z_{i'}\varphi(\mathbf{r}_{i'}) - \int \rho'(\mathbf{r})\varphi(\mathbf{r})\, d\mathbf{r} \quad (10)$$

Cette seconde intégration explicite du produit entre densité électronique et potentiel électrostatique est réalisée pour les contributions dépendant d'atomes plus proches que 5 Å. Tout comme dans la section précédente, dans VMoPro, ce calcul exact est réalisé par l'intégration à courte distance (c'est le cas EP), alors que les interactions à longue distance sont calculées par interaction de moment ponctuels (c'est le cas MM) telle que reporté par Buckingham[24,25,26] ou Stone[32].

6 - Topologie de la densité

L'analyse topologique de la densité a été introduite par R.F.W. Bader en 1987[9] comme une méthode rigoureuse de partage de l'espace en « territoires » atomiques, les bassins, en se basant sur la topologie de la densité électronique, au sens analytique du terme. Cette méthode, souvent appelée QTAIM (*Quantum Theory of Atom In Molecules*), s'attache uniquement à la forme du champ de densité électronique et est indépendante du modèle utilisé pour la représenter[33]. On s'intéresse ici à décrire la forme du champ scalaire $\rho(\mathbf{r})$ par son gradient, $\nabla\rho(\mathbf{r})$, et son Laplacien, $\nabla^2\rho(\mathbf{r})$. Quatre structures de dimension(alité) croissante émergent naturellement de l'étude de la densité électronique moléculaire : les points critiques, les chemins d'interaction (*bond paths*), les surfaces de flux-nul et enfin les bassins topologiques.

Les points critiques sont les points de l'espace où le gradient de $\rho(\mathbf{r})$ s'annule :

$$\nabla\rho(\mathbf{r_{CP}}) = 0 \qquad (11)$$

Les signes des valeurs propres de la matrice hessienne au point critique, $\nabla^2\rho(\mathbf{r}_{CP})$, permettent de définir 4 sortes de points critiques. La signature d'un point critique se note (a, ±b), où a est le nombre de courbures non-nulles en ce point et b la somme des signes de ces courbures. Les points critiques (3, −3) correspondent à des maxima locaux du champ. Toutes les positions nucléaires d'une molécule sont donc associées à un point de ce type. Ces points critiques peuvent aussi exister – dans de très rares cas – hors des positions nucléaires. La mise en évidence expérimentale de ces attracteurs non-nucléaires (*Non-Nuclear Attractors*) est très récente[34].

Les minimums locaux ont une signature (3, +3), et sont souvent appelés points cages. On les rencontre typiquement au sein de molécules présentant une cavité fermée (adamantanes, fullerènes…).

La présence de deux maxima locaux voisins entraîne l'existence d'un point critique d'un autre type, un point selle de signature (3, −1), aussi appelé *BCP* (*bond critical points*). Dans une direction (celle qui rejoint les deux maxima locaux), ce point correspond à un minimum de densité électronique, et à un maximum dans les deux autres directions. Leur position, la valeur de la densité électronique, celle du Laplacien et les valeurs de la matrice hessienne en ce point fondent la vision topologique des interactions interatomiques et en permettent une caractérisation et une description rigoureuse.

Enfin, les points critiques de cycles (*ring critical points*) ont une signature (3, +1). On les trouvera habituellement au centre de cycles aromatiques.

La présence de points selle entre deux maxima locaux permet de définir la structure topologique de dimension 1, le chemin d'interaction (*bond path*). Ces chemins sont définis comme le chemin de gradient maximal reliant un point (3, –1) aux deux plus proches points (3, -3)[35]. Ces chemins d'interaction existent pour toutes les interactions atomiques : covalentes, ioniques, van der Waals, etc. et ont été explorée par différents auteurs[36,37,38,39].

Dans la matière condensée, deux atomes reliés par un chemin de gradient interagissent à courte ou moyenne portée. Ce voisinage se matérialise par une surface, qui contient le point selle : la surface de flux nul (ZFS). Ces surfaces (structures topologiques de dimension 2) sont définies par un vecteur normal $\mathbf{n}$ qui répond, pour tout point $\mathbf{r}_{ZFS}$ de la surface, à :

$$\boldsymbol{\nabla}\rho(\mathbf{r_{ZFS}}) \cdot \mathbf{n}(\mathbf{r_{ZFS}}) = 0 \qquad \mathbf{(12)}$$

où '·' symbolise le produit scalaire. Le gradient du champ scalaire considéré (ici par exemple la densité électronique) est perpendiculaire à la normale, en tout point de la surface de flux nul. La surface n'est traversée par aucune ligne de gradient (Le gradient du chemin d'interaction devient nul exactement au point critique de la surface). Cette surface permet de séparer l'espace en régions où toutes les lignes de champ mènent au même maximum local.

Ce partage de l'espace par des surfaces de flux nul permet, dans le cas d'un système moléculaire, de définir la structure topologique de dimension 3, nommée bassin. Dans le champ de densité électronique, tout point du bassin est relié par une ligne de gradient à un même maximum correspondant à la position d'un noyau. Pour l'immense majorité des systèmes, la correspondance noyau – point critique (3, –3) est 1 : 1, ce qui fonde le terme bassin atomique utilisé dans l'analyse QTAIM (la seule exception étant la rare présence d'un NNA). La définition de bassins atomiques attribue, de manière naturelle et univoque, la densité électronique d'une région de l'espace à un atome. Cette manière de partager la densité est bien entendu indépendante de la partition en atomes multipolaires, vue plus haut. Il est donc possible de partager d'autres propriétés moléculaires, dérivées de la densité électronique, en contributions atomiques topologiques, comme par exemple les

moments électrostatiques (charges, dipôles, quadrupôles, etc.). La partition de ces propriétés dépend du volume attribué à chaque atome, comme ici pour la charge q_Ω d'un bassin atomique :

$$q_\Omega = Z_\Omega - \int_\Omega \rho_{total}(\mathbf{r})\, d\mathbf{r} \qquad (\mathbf{13})$$

où Z_Ω est la charge du noyau centrant le bassin. Dans la matière condensée, les bassins atomiques ont tous un volume fini. Pour des molécules isolées, la densité électronique s'étend formellement jusqu'à l'infini. L'intégration de propriétés sur une densité électronique infinie suppose habituellement de fermer les bassins ouverts par une surface définie de manière à ignorer une partie négligeable de $\rho(\mathbf{r})$.

7 - Topologie du potentiel

Le potentiel électrostatique $\varphi(\mathbf{r})$ est un champ scalaire, tout comme la densité électronique. Il peut donc être analysé en utilisant le formalisme QTAIM. Plusieurs différences transforment cependant l'interprétation des entités topologiques de $\varphi(\mathbf{r})$ comparées aux entités trouvées dans le champ de densité électronique $\rho(\mathbf{r})$.

Si la densité électronique est positive en tout point de l'espace, le potentiel peut être indifféremment positif ou négatif. Les positions nucléaires sont encore une fois associées à des points critiques (3, –3), un potentiel fortement positif étant attendu à proximité des noyaux, voir équation (8). Il n'y a cependant plus de garantie qu'un maximum local corresponde à une position nucléaire. Les minima locaux, points critiques (3, +3), ne sont pas associés à des déplétions locales en électrons. Des minima de $\varphi(\mathbf{r})$ apparaissent dans les zones de concentration locale en électrons : ils peuvent par exemple signer la présence de doublets non-liants. La position de ces minima peut par exemple être relié à la force de liaisons hydrogène[40]. Si des points selle apparaissent au niveau des liaisons covalentes, ils ne sont pas superposables aux points critiques correspondants de la densité électronique. Il est à noter que le Laplacien du potentiel électrostatique, $\nabla^2\varphi(\mathbf{r})$, est proportionnel à la densité électronique [équation (14)]. On s'attend donc à une forte courbure de $\varphi(\mathbf{r})$ au point critique de liaisons covalentes riches en électrons.

$$-\nabla^2\varphi(\mathbf{r}) = \frac{\rho_{total}(\mathbf{r})}{\varepsilon_0} \qquad (\mathbf{14})$$

Où $\rho_{total}(\mathbf{r})$ symbolise la densité de charge totale (électrons et noyaux). Les lignes de gradient de $\varphi(\mathbf{r})$ ont une signification physique bien définie, $-\nabla\varphi(\mathbf{r})$ est en effet le champ électrique. Plusieurs conséquences en découlent pour les structures topologiques de dimension supérieure, c'est-à-dire les surfaces de flux nul et les bassins.

Le champ électrique ressenti par une charge électrique ponctuelle définit la direction et la magnitude de la force qui s'exerce sur cette charge. Les chemins d'interaction entre extrema correspondent toujours au maximum du gradient entre ces deux points. Ces chemins ne sont pas seulement l'avatar d'interactions covalentes (ou de courte portée) dans $\varphi(\mathbf{r})$. Ils représentent aussi les chemins de moindre résistance où une attaque nucléophile peut se produire (en suivant $\mathbf{E}(\mathbf{r})$, c'est-à-dire en remontant le sens du gradient). Inversement, les zones nucléophiles d'une molécule (réactives aux espèces électrophiles) sont localisées en suivant le gradient $\nabla\varphi(\mathbf{r})$ vers un minimum local. Ces considérations sont bien entendu pertinentes lorsqu'on considère la périphérie d'un système moléculaire pour déterminer ses interactions et/ou sa réactivité.

Les surfaces de flux nul de $\varphi(\mathbf{r})$ sont déterminées de la même manière que celles de $\rho(\mathbf{r})$. Leurs formes et leurs positions diffèrent cependant et d'autres points critiques que les (3, –1) peuvent être contenus dans la surface, notablement des minima (3, +3). Le théorème de Gauss stipule que le flux net du champ électrique traversant une surface fermée est proportionnel à la charge contenue dans le volume défini par cette surface. Aucune ligne de gradient de $\varphi(\mathbf{r})$ ne traverse la surface de flux nul, par conséquent, tout bassin topologique de $\varphi(\mathbf{r})$ est de charge nulle s'il est fermé. Une interprétation naturelle de cette partition de l'espace à l'aide de φ-bassins est que seule la densité électronique située hors de cet ensemble de bassins fermés et neutres est disponible pour les interactions électrostatiques à longue portée[41], car l'influence électrostatique majoritaire au sein d'un φ-bassin atomique est le noyau qui compense la charge des électrons. À cet endroit, toutes les lignes de champ ont le noyau pour origine. À l'inverse, pour un cation isolé, on trouvera ainsi des lignes de champ s'étendant du noyau à l'infini. Pour un anion, les lignes de champ en provenance de l'infini se finissent sur la surface de flux nul qui encapsule exactement la partie neutre de la densité de charge.

8 - Induction dipolaire & polarisabilité.

La polarisabilité d'un objet (ici une molécule) représente sa capacité à développer un moment électrostatique en réponse à un champ électrique. Ces moments électrostatiques induits dépendent de la réorganisation des charges dans une molécule. Cette capacité à se réorganiser est quantifiée par la polarisabilité moléculaire. La polarisabilité moléculaire dépend de deux facteurs car une molécule a deux possibilités pour développer un moment électrostatique. Dans une molécule flexible, les atomes peuvent bouger les uns par rapport aux autres ce qui peut induire des moments électriques (en éloignant des atomes de charges différentes). Cette composante – dite de relaxation – de la polarisabilité n'est pas traitée ici. On s'intéresse plutôt à la polarisabilité dite électronique. La polarisabilité électronique dépend uniquement de la réorganisation des électrons autour des noyaux de la molécule[42].

Le présent travail s'intéresse à des géométries moléculaires optimisées ou observées par cristallographie, donc seule la polarisabilité électronique sera traitée ici (géométrie fixée). L'importance relative de la composante de relaxation varie en fonction des molécules et est mesurable à partir d'un spectre infrarouge complet. La polarisabilité électronique est la seule contribuant à l'indice de réfraction d'un milieu (équation de Lorentz-Lorenz). Les polarisabilités électroniques et de relaxation contribuent toutes deux à la constante diélectrique d'un milieu (équation de Clausius-Mossotti).

En plus de ces importantes propriétés macroscopiques, des propriétés microscopiques d'une molécule sont aussi dépendantes de la polarisabilité. Par exemple, les forces de dispersion entre molécules dépendent de la polarisabilité, ces forces sont très importantes pour modéliser l'interaction de molécules aromatiques. Ici, on s'intéresse à la description explicite de l'induction statique, dans le contexte de la modélisation cristallographique via le modèle HCM. Cette induction est relative aux forces de Debye entre moments électrostatiques induits et permanents. Ici, on ne s'attachera pas à décrire ces forces de manière statistique, mais explicite, à l'échelle de la molécule et de l'atome. Le moment induit $\boldsymbol{\mu}$ (en e·Å) est décrit, au niveau dipolaire pour un champ électrique $\mathbf{E}(\mathbf{r})$ homogène et permanent (de fréquence nulle, en e·Å^{-2}) :

$$\boldsymbol{\mu} = \boldsymbol{\alpha} \,.\, \mathbf{E}(\mathbf{r}) \qquad (15)$$

La polarisabilité atomique $\boldsymbol{\alpha}$ a donc la dimension d'un volume. $\boldsymbol{\alpha}$ peut être un scalaire positif ou un tenseur symétrique et défini-positif (3^2 éléments). Si la polarisabilité scalaire (isotrope) est nécessaire, la relation $\boldsymbol{\alpha}_{\text{iso}} = (1/3)\cdot\text{Trace}(\boldsymbol{\alpha})$ est utilisée. Le modèle d'Hansen et Coppens utilise des harmoniques sphériques centrées sur le noyau pour modéliser la densité électronique. L'évaluation de la polarisation dipolaire de pseudoatomes utilisera donc des champs électriques $\mathbf{E}(\mathbf{r})$ calculés aux coordonnées des noyaux atomiques. Le présent travail s'attache à deux descriptions complémentaires de la polarisabilité dipolaire, électronique et statique, d'une molécule. L'une s'attache à la densité électronique et à sa partition topologique en contributions atomiques (QTAIM). Les propriétés atomiques liées à la densité électronique peuvent être intégrées sur ces bassins, ici les moments dipolaires et la réponse de la densité à des champs électriques externes (via la polarisabilité des bassins). L'autre formalisme décrit la même propriété par des moments électrostatiques ponctuels. L'utilisation de moments ponctuels pour décrire la polarisation demande une adaptation pour représenter cet effet dans la densité électronique du modèle HCM, ce qui fait l'objet d'une partie du travail présenté dans ce manuscrit.

Le modèle HCM est un pont entre ces deux représentations du moment dipolaire et de la polarisabilité. D'un côté le modèle HCM permet de reconstituer le champ de densité permettant une analyse QTAIM. De l'autre côté, le moèle HCM représente la densité électronique par des harmoniques sphériques, lesquelles ont des contributions directes aux différents moments électrostatiques atomiques. En effet, pour un atome multipolaire isolé, son moment dipolaire est relatif à sa densité de déformation, et est uniquement décrit par la partie dipolaire ($l = 1$), de populations $P_{1,\pm m}$, correspondant aux harmoniques sphériques $y_{1,\pm m}$. Encore une fois, la partition de ces propriétés atomiques n'est pas équivalente entre la partition HCM et celle de Bader (voir page 44). Le calcul des composantes du moment dipolaire atomique, à partir de l'expression du modèle HCM [équation (6)], permet de montrer que celles-ci ne dépendent que des populations dipolaires $P_{1,\pm m}$ (selon Coppens[20], chapitre 7, page 148) :

$$P_{1,+1} = \text{u}.\,\boldsymbol{\mu_x} \qquad P_{1,-1} = \text{u}.\,\boldsymbol{\mu_y} \qquad P_{1,0} = \text{u}.\,\boldsymbol{\mu_z} \qquad \text{avec u} = -\frac{3\kappa'\zeta}{20} \qquad \mathbf{(16)}$$

9 - Modèle d'Applequist

La description des moments dipolaires en moments ponctuels centrés sur les positions nucléaires est utilisée de longue date (voir l'article de Buckingham & Utting, 1970[25]). L'approche trouve ses origines dans l'électrostatique classique. Dans ce formalisme, les polarisabilités représentent la réponse (en termes de moments électriques) d'une de ces particules ponctuelles au champ électrique appliqué au système. Le champ électrique est ici exprimé comme une série de Taylor. Dans le présent manuscrit, le terme polarisabilité est utilisé spécifiquement pour décrire $\boldsymbol{\alpha}$, la réponse dipolaire à un champ électrique permanent. La définition de Buckingham et Utting est applicable à tous les moments, en réponse à chaque niveau de l'expansion du champ, elle est donc plus générale que celle employée ici.

La polarisabilité moléculaire dipolaire est usuellement considérée comme additive, en vertu de l'additivité des moments dipolaires dans une molécule neutre, et en vertu de l'additivité des réfractions molaires dans l'équation de Lorentz-Lorenz. Cette vision exclut l'interaction entre les dipôles induits et est une approximation valide pour les gaz (molécules éloignées) et la matière condensée isotrope (où toutes les orientations possibles d'interaction sont réalisées, liquide où solide amorphe)[42]. Malgré une certaine transférabilité (rapportée par Applequist[43]) des polarisabilités atomiques isotropes additives, ce modèle pèche par sa simplicité : les atomes dans une molécule interagissent à courte distance, les interactions deviennent directionnelles, donc anisotropes. Ces questions, liées aux distances et à l'anisotropie des interactions, seront abordées tout au long du manuscrit.

L'utilisation de centres polarisables ponctuels en interaction pour décrire l'indice de réfraction de gaz diatomiques remonte à Silberstein, 1917[44,45]. Applequist publie en 1972 une méthode pour généraliser la théorie de Silberstein à des molécules d'un quelconque nombre d'atomes. Cette méthode est cohérente avec le formalisme de Buckingham, où l'énergie d'interaction électrostatique est décomposée à partir de l'expansion multipolaire (Buckingham, 1988[26]) en une somme de contributions dépendant de tenseurs d'interaction. Le tenseur de second ordre $\boldsymbol{T}_{\alpha\beta}$ de Buckingham gouverne l'interaction entre dipôles et correspond au tenseur $\boldsymbol{T}_{ij}$ d'Applequist. L'apport d'Applequist apparaît dans le schéma de partition des polarisabilités entre les centres atomiques ponctuels. Son formalisme représente l'ensemble des dipôles induits sur une

molécule par un unique vecteur (3·*n* composants pour *n* atomes). Les champs électriques aux coordonnées nucléaires sont exprimés de la même manière. Il est donc possible d'écrire un tenseur de polarisabilité moléculaire, comme suit.

$$\begin{bmatrix} \boldsymbol{\mu_1} \\ \boldsymbol{\mu_2} \\ \vdots \\ \boldsymbol{\mu_n} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\alpha}_1 & 0 & \cdots & 0 \\ 0 & \boldsymbol{\alpha}_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \boldsymbol{\alpha}_n \end{bmatrix} \cdot \begin{bmatrix} \boldsymbol{E_1} \\ \boldsymbol{E_2} \\ \vdots \\ \boldsymbol{E_n} \end{bmatrix} \qquad (\mathbf{17})$$

Où $\boldsymbol{\alpha}_n$ est la polarisabilité de l'atome *n*, représenté par un tenseur de 3×3 éléments. L'équation ci-dessus utilise des polarisabilités additives, c'est-à-dire que la polarisabilité de la molécule est égale à la somme (élément par élément) des polarisabilités atomiques. Dans ce modèle, il n'y a pas d'interaction entre les moments induits : le champ électrique généré par le dipôle induit sur l'atome *i* n'a aucun effet sur les moments induits aux autres atomes de la même molécule. Applequist utilise des polarisabilités isotropes, mais permettre l'interaction entre dipôles induits demande d'écrire ces polarisabilités sous forme matricielle.

Il est possible d'utiliser le tenseur $\boldsymbol{T}_{ij}$ pour connaître le champ électrique $\mathbf{E}$ (à l'origine du référentiel) généré par un dipôle ponctuel $\boldsymbol{\mu}_i$ situé au point $\mathbf{r}_i$:

$$\mathbf{E} = -\boldsymbol{T}_{ij}(\mathbf{r}_i) \cdot \boldsymbol{\mu}_i \qquad (\mathbf{18})$$

On place la molécule dans un champ électrique externe. La molécule est polarisable. Elle est alors considérée dans ce modèle comme un ensemble de dipôles induits ponctuels. C'est à l'aide de l'équation (19) que Applequist permet aux dipôles induits de se « voir » au sein de la molécule.

Le champ électrique généré par le moment $\boldsymbol{\mu}_j$, associé à un atome situé aux coordonnées $\mathbf{r}_j$, sur un atome *i* situé en $\mathbf{r}_i$, s'écrit $-\boldsymbol{T}_{ij}(\mathbf{r}_{ij})\cdot\boldsymbol{\mu}_j$, où $\mathbf{r}_{ij}$ est le vecteur interatomique. Le dipôle induit (total) $\boldsymbol{\mu}_i$ sur l'atome *i* dépend donc du champ externe $\mathbf{E}_i$ et de tous les autres dipôles induits de la molécule :

$$\boldsymbol{\mu}_i = \boldsymbol{\alpha}_i\left[\mathbf{E}_i - \sum_{i \neq j} \boldsymbol{T}_{ij}\left(\mathbf{r}_{ij}\right) \cdot \boldsymbol{\mu}_j\right] \qquad (\mathbf{19})$$

$$\text{où } \mathbf{r}_{ij} = \mathbf{r}_j - \mathbf{r}_i$$

La formulation matricielle de l'équation (17) permit à Applequist d'introduire le tenseur $\boldsymbol{T}_{ij}$ – en tant que couplage entre les dipôles induits $\boldsymbol{\mu}_i$ et $\boldsymbol{\mu}_j$ – au sein du tenseur de polarisabilité moléculaire. Pour un dipôle induit ponctuel :

$$\boldsymbol{T}_{ij}(\mathbf{r}_{ij}) = \frac{1}{|\mathbf{r}_{ij}|^3}\delta_{ij} - \frac{3}{|\mathbf{r}_{ij}|^5}\mathbf{r}_{ij} \otimes \mathbf{r}_{ij}^t \qquad (\mathbf{20})$$

Où '⊗' et δ_{ij} représentent respectivement le produit et le delta de Kronecker (matrice identité). Le tenseur $\boldsymbol{T}_{ij}$ remplace alors les zéros hors diagonale de l'équation (17), aboutissant à l'expression suivante :

$$\begin{bmatrix}\boldsymbol{\mu}_1\\ \boldsymbol{\mu}_2\\ \vdots\\ \boldsymbol{\mu}_n\end{bmatrix} = \begin{bmatrix}\boldsymbol{\alpha}_1^{-1} & \boldsymbol{T}_{12} & \cdots & \boldsymbol{T}_{1n}\\ \boldsymbol{T}_{12} & \boldsymbol{\alpha}_2^{-1} & \cdots & \boldsymbol{T}_{2n}\\ \vdots & \vdots & \ddots & \vdots\\ \boldsymbol{T}_{1n} & \boldsymbol{T}_{2n} & \cdots & \boldsymbol{\alpha}_n^{-1}\end{bmatrix}^{-1} \cdot \begin{bmatrix}\mathbf{E}_1\\ \mathbf{E}_2\\ \vdots\\ \mathbf{E}_n\end{bmatrix} \qquad (\mathbf{21})$$

$\boldsymbol{T}_{ij}$ est exprimé en Å^{-3}, pour des champs électriques en e·Å^{-2}, des moments dipolaires en e·Å et des polarisabilités en Å^3. Il est à noter que les polarisabilités atomiques utilisées ici ne peuvent pas être équivalentes à celles utilisées précédemment. Par opposition au modèle de polarisabilités additives de l'équation (17), on parlera ici de polarisabilités interactives.

Pour obtenir la polarisabilité moléculaire, on additionne élément par élément chaque sous-matrice 3×3 du tenseur moléculaire inversé de l'équation (21). Cette matrice 3×3 peut ensuite être diagonalisée pour obtenir les valeurs propres de la polarisabilité moléculaire. Cette méthode correspond à une partition alternative de cette propriété entre les atomes. Applequist révèle cette différence sur la molécule de méthane où plusieurs choix de polarisabilités isotropes pour H et C permettent d'obtenir la même polarisabilité moléculaire. Pire encore, d'autres combinaisons présentent des singularités, aboutissant à une polarisabilité moléculaire qui tends vers l'infini.

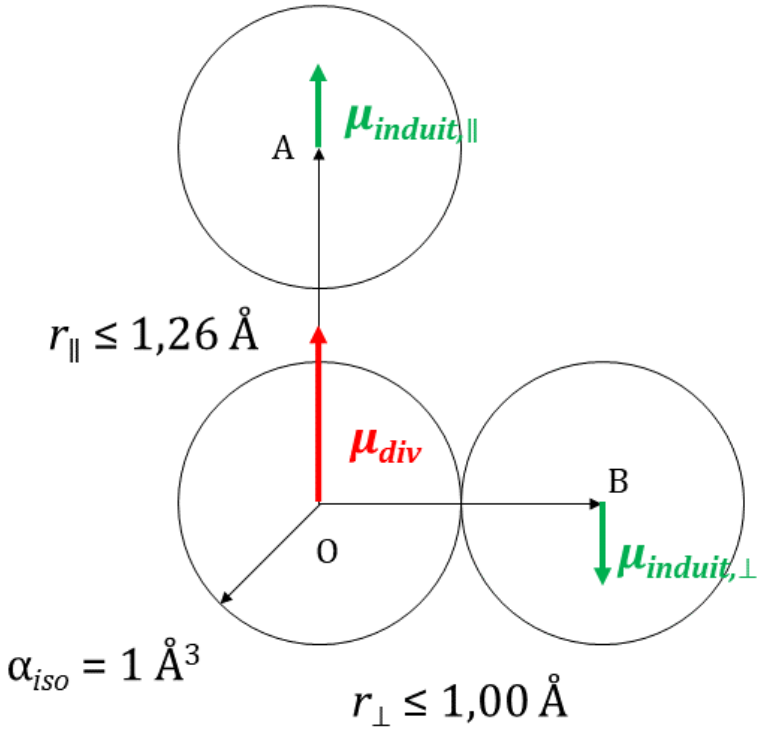


Figure 6. Rayons de divergence axiale (*head-to-tail*, ∥) et latérale (⊥), pour deux couples d'atomes de polarisabilité unitaire isotrope. Tout moment dipolaire induit sur l'atome A dans la direction parallèle '∥' ou induit sur l'atome B dans la direction perpendiculaire '⊥' va entrainer l'induction sur l'atome O d'un moment dipolaire $\boldsymbol{\mu}_{div}$, dont la norme est respectivement supérieure à $\boldsymbol{\mu}_{induit,\parallel}$ ou $\boldsymbol{\mu}_{induit,\perp}$.

La Figure 6 illustre deux situations où le moment dipolaire induit sur l'atome O a une norme supérieure à celle de l'atome A ou B. Dans un formalisme itératif de la polarisation (où A polarise O, puis O polarise A, etc. jusqu'à convergence), une boucle de polarisation infinie se crée et le moment induit sur chaque atome augmente à chaque cycle de calcul pour *in fine* tendre vers l'infini. Ces interactions non-physiques sont parfois appelées des « catastrophes de polarisation ». Ces situations correspondent à une valeur propre de $\boldsymbol{T}_{ij}(\mathbf{r}_{ij})$ inférieure à -1 (pour des polarisabilités unitaires). L'inclusion de polarisabilités anisotropes demande une vigilance particulière envers ces rayons de divergence, qui dépendent alors des projections parallèles et perpendiculaires des polarisabilités atomiques considérées (voir chapitre 4, partie 5).

10 - Modèles atténués (Thole, Birge)

Afin de réparer les déficiences du modèle d'Applequist, B.T. Thole[46,47] publie en 1981 une méthode pour limiter ces divergences. En analysant cette divergence pour une molécule diatomique modèle dont les deux atomes ont une polarisabilité isotrope, il est possible de déterminer le rayon de divergence de la polarisabilité moléculaire déterminée en utilisant $\boldsymbol{T}_{ij}$. Thole détermine que ce rayon est en dessous de :

$$|\mathbf{r}_{ij}| < (4\alpha_i\alpha_j)^{1/6} \quad (\mathbf{22})$$

Avec α_i et α_j les polarisabilités scalaires des atomes *i* et *j*. Ce rayon de divergence correspond à l'interaction entre les atomes A et O de la Figure 6, puisque ce rayon de divergence est supérieur au rayon de divergence entre les atomes O et B [qui est plus petit, et vaut $(\alpha_i\alpha_j)^{1/6}$]. Thole reformule le tenseur d'interaction dipolaire pour deux polarisabilités unitaires, en exprimant $\mathbf{r}_{ij}$ en fonction de α_i et α_j :

$$\mathbf{u}_{ij} = \mathbf{r}_{ij}(\alpha_i\alpha_j)^{-1/6} \quad (\mathbf{23})$$

$\mathbf{u}_{ij}$ est sans unité. $(\alpha_i\alpha_j)^{1/2}\boldsymbol{T}_{ij}$, le tenseur mis à l'échelle, est également sans unité. Si le tenseur mis à l'échelle a au moins une valeur propre inférieure à -1, alors la matrice de polarisabilité moléculaire de l'équation (21) n'est plus définie positive. L'application d'un champ électrique sur cette molécule génèrera une induction dipolaire infinie et éventuellement orientée à l'opposé du vecteur champ électrique (au moins pour certaines directions du champ) : on revient à la situation représentée Figure 6.

Afin d'empêcher la construction de tenseurs moléculaires avec ces caractéristiques, Thole assume que l'un des dipôles n'est plus ponctuel. Un dipôle ponctuel est constitué de deux charges infinitésimalement déplacées. Dans le modèle de Thole, le dipole est constitué de deux densités monopolaires $\rho_{Thole}(\mathbf{r})$ infinitésimalement déplacés. Ces densités électroniques radiales sont utilisées par Thole pour dériver analytiquement le champ électrique des dipôles induits. $\rho_{Thole}(\mathbf{r})$ est exprimée en fonction des polarisabilités α_i et α_j (via la variable normalisée $\mathbf{u}_{ij}$) :

$$\rho_{Thole}(\mathbf{u}_{ij}) = \frac{a^3}{8\pi} e^{-a|\mathbf{u}_{ij}|} \quad (\mathbf{24})$$

Où *a* est un paramètre ajustable. L'article original de Thole s'attache à reconstruire les polarisabilités moléculaires expérimentales d'un set de molécules. *a* est affiné en

parallèle des polarisabilités atomiques isotropes de son modèle par une méthode d'optimisation multifactorielle. En effet le tenseur moléculaire inversé de l'équation (21) peut être utilisé pour retrouver la polarisabilité moléculaire (sommation des sous-matrices puis diagonalisation). Thole utilise une valeur de *a* et 4 polarisabilités atomiques (H, C, N, O) pour calculer une quarantaine de polarisabilités moléculaires. À courte distance, le potentiel de ce dipôle est atténué par rapport à un dipôle ponctuel. Les modèles de densité de Thole sont très simples, le plus « physique » [équation (24)] utilise une simple exponentielle décroissante et deux paramètres (dont la polarisabilité, qui est utilisée dans ce modèle comme paramètre de la densité). Cette fonction décrit la densité de tout l'atome, contrairement au modèle HCM qui sépare les contributions de cœur, valence et déformation.

On commence par exprimer le tenseur $\boldsymbol{T}_{ij}$ en fonction de la distance $\mathbf{u}_{ij}$, mise à l'échelle des polarisabilités :

$$\boldsymbol{T}_{ij}(\mathbf{u}_{ij}) = (\alpha_i\alpha_j)^{1/2}.\boldsymbol{T}_{ij}(\mathbf{r}_{ij}) = \frac{1}{|\mathbf{u}_{ij}|^3}\delta_{ij} - \frac{3}{|\mathbf{u}_{ij}|^5}\mathbf{u}_{ij} \times \mathbf{u}_{ij}^t \qquad (25)$$

$\boldsymbol{T}_{ij}(\mathbf{u}_{ij})$ ne dépend donc plus de la polarisabilité des atomes *i* et *j*. Thole utilise ce tenseur mis à l'échelle pour trouver deux fonctions *f*' et *f*'', qui permettront d'atténuer suffisamment les termes hors diagonaux de la matrice de l'équation (21), et donc d'atténuer l'induction interatomique pour lever les divergences du modèle. Dans l'article original, et ceux reprenant le modèle de Thole, *f*' et *f*'' sont notées φ' et φ''. Cette notation sera évitée ici pour ne pas être confondues avec $\varphi(\mathbf{r})$ qui représente le potentiel électrostatique calculé à partir d'une distribution de charges. *f*' et *f*'' sont les dérivées par rapport à $|\mathbf{u}_{ij}|$ de la fonction *f*. $\boldsymbol{T}_{ij}(\mathbf{u}_{ij})$ est réécrit en fonction de *f*' et *f*'' :

$$\boldsymbol{T}_{ij}(\mathbf{u}_{ij}) = (-\frac{f'}{|\mathbf{u}_{ij}|})\delta_{ij} + (\frac{f'}{|\mathbf{u}_{ij}|} - f'')\frac{\mathbf{u}_{ij}\times\mathbf{u}_{ij}^t}{|\mathbf{u}_{ij}|^2} \qquad (26)$$

Lorsque :

$$f = \frac{1}{|\mathbf{u}_{ij}|} \qquad (27)$$

On retrouve l'équation (25) du modèle de Applequist, qui décrit le champ d'un dipôle ponctuel.

En effet, dans ce formalisme, les fonction *f'* et *f''* sont :

$$f' = \frac{-1}{|\mathbf{u}_{ij}|^2} \qquad \text{et} \qquad f'' = \frac{2}{|\mathbf{u}_{ij}|^3} \quad \textbf{(28)}$$

Le modèle de Thole utilise les fonctions *f'* et *f''* différentes pour atténuer l'interaction entre dipôles. Décrites aux équations (29) et (30), elles sont atténuées par un facteur compris entre 0 et 1. Dans ce travail, deux des fonctions d'atténuation originales de Thole ont été implémentées telles qu'elles, au sein de *MoProViewer*. Le paramètre *a* réapparaît dans ces fonctions d'atténuation. Voici ces fonctions, pour la densité électronique radiale de l'équation (24) :

$$\frac{f'}{|\mathbf{u}_{ij}|} = \frac{-1}{|\mathbf{u}_{ij}|^3}\left[1 - \left(\frac{1}{2}a^2\left|\mathbf{u}_{ij}\right|^2 + a\left|\mathbf{u}_{ij}\right| + 1\right)e^{-a|\mathbf{u}_{ij}|}\right] \qquad \textbf{(29)}$$

$$f'' = \frac{2}{|\mathbf{u}_{ij}|^3}\left[1 - \left(\frac{1}{4}a^3\left|\mathbf{u}_{ij}\right|^3 + \frac{1}{2}a^2\left|\mathbf{u}_{ij}\right|^2 + a\left|\mathbf{u}_{ij}\right| + 1\right)e^{-a|\mathbf{u}_{ij}|}\right] \quad \textbf{(30)}$$

Ces fonctions atténuées convergent vers le comportement d'un dipôle ponctuel pour des $|\mathbf{u}_{ij}|$ très supérieures à 1, quand le terme exponentiel tend vers 0. En dessous de cette distance mise à l'échelle, l'induction sera forcément inférieure à celle calculée pour un dipôle ponctuel :

$$\lim_{|\mathbf{u}_{ij}|\gg 1} f' = \frac{-1}{|\mathbf{u}_{ij}|^2} \qquad \text{et} \qquad \lim_{|\mathbf{u}_{ij}|\gg 1} f'' = \frac{2}{|\mathbf{u}_{ij}|^3} \qquad \textbf{(31)}$$

La densité électronique utilisée par Thole n'est pas comparable aux densités du modèle HCM. C'est à Birge *et al.*, 1983[48], que la fonction d'atténuation pour une densité radiale de type Slater est due, deux ans après le travail original de Thole. Birge *et al.* utilisent une densité radiale assez similaire à la densité radiale R_{nl} du modèle multipolaire. Pour des facilités de lecture, cette fonction $\rho_{Birge}(\mathbf{u}_{ij})$ est ici réécrite en utilisant les paramètres du modèle HCM :

$$\rho_{Birge}\left(\mathbf{u}_{ij}\right) = \frac{(\alpha^{1/3}\zeta)^{n_l+3}}{4\pi(n_l!)} \cdot |\mathbf{u}_{ij}|^{n_l} \cdot e^{-\alpha^{1/3}\zeta|\mathbf{u}_{ij}|} \qquad \textbf{(32)}$$

Où ζ est l'exposant de Slater, en Å^{-1}, et n_l le nombre quantique principal (écrit dans la convention de Hansen et Coppens). Le produit $\alpha^{1/3}\cdot\zeta$ permet d'exprimer l'exposant de Slater en fonction de la polarisabilité atomique. Cet exposant est donc mis à l'échelle de

$\mathbf{u}_{ij}$. Le tenseur $\boldsymbol{T}_{ij}$ du modèle de Birge est construit à partir de l'équation (26), exactement comme dans le modèle de Thole.

Bien entendu, de nouvelles fonctions *f'* et *f''* sont nécessaires :

$$\frac{f\prime}{|\mathbf{u}_{ij}|} = \frac{-1}{|\mathbf{u}_{ij}|^3}\left[1 - e^{-\alpha^{1/3}\zeta|\mathbf{u}_{ij}|} \cdot \sum_{i=0}^{n_l+2} \frac{\left(|\mathbf{u}_{ij}|\alpha^{1/3}\zeta\right)^{n_l-i+2}}{(n_l-i+2)!}\right] \qquad (33)$$

$$f'' = \frac{2}{|\mathbf{u}_{ij}|^3} - \left(\frac{2}{|\mathbf{u}_{ij}|^3} + \frac{\alpha^{1/3}\zeta}{|\mathbf{u}_{ij}|^2}\right) \cdot e^{-\alpha^{1/3}\zeta|\mathbf{u}_{ij}|} \cdot \sum_{i=0}^{n_l+2} \frac{\left(|\mathbf{u}_{ij}|\alpha^{1/3}\zeta\right)^{n_l-i+2}}{(n_l-i+2)!}$$

$$+ \frac{\alpha^{1/3}\zeta}{|\mathbf{u}_{ij}|^2} \cdot e^{-\alpha^{1/3}\zeta|\mathbf{u}_{ij}|} \cdot \sum_{i=0}^{n_l+1} \frac{\left(|\mathbf{u}_{ij}|\alpha^{1/3}\zeta\right)^{n_l-i+1}}{(n_l-i+1)!} \qquad (34)$$

Cette expression de $\boldsymbol{T}_{ij}$ – utilisant les fonction *f'* et *f''* de Birge *et al.* – a été implémentée dans l'outil de polarisation de *MoProViewer*. Son utilisation telle qu'elle, avec les polarisabilités isotropes déterminées par Birge *et al.*, est possible.

Il est important de noter que chez Thole et Birge, la densité électronique dépend de la polarisabilité. Ce raccourci est utile, car dans ce modèle la polarisabilité entre dans le coefficient de normalisation de la fonction de Slater, qui gouverne la convergence de *f'* et *f''* à longue distance, et sensé, car la polarisabilité est une propriété exclusivement électronique dans une molécule rigide. Dans le modèle HCM, cette propriété n'entre pas dans le calcul de $\rho(\mathbf{r})$, pourtant à la fois la densité électronique et la polarisabilité doivent être prises en compte dans l'atténuation de l'induction. Dans la partie 1.7, il sera vu que, indépendamment du modèle de Thole et Birge, certains paramètres HCMM des atomes transférables sont corrélés à des polarisabilités théoriques déterminées de manière indépendante. L'adaptation de ce modèle à des polarisabilités anisotropes sera également abordée dans la partie 4.5.

11 - Modèle topologique de la polarisabilité

La relation linéaire de l'équation (15) relie les moments dipolaires atomiques induits $\boldsymbol{\mu}_i$ et le champ électrique appliqué $\mathbf{E}_i(\mathbf{r})$ à la position de l'atome *i*. La partie précédente explicitait le traitement de l'induction de groupes de dipôles ponctuels. Ce traitement passe par des corrections qui atténuent l'induction et reproduisent le

comportement d'un atome non-ponctuel, plus réaliste d'un point de vue chimique. Dans ce travail, les dipôles atomiques induits par le processus de polarisation sont représentés en termes de fonctions de densités électroniques dipolaires, en utilisant les populations dipolaires correspondantes [$P_{1,-1}$, $P_{1,0}$ et $P_{1,+1}$, voir équation (16)] du développement multipolaire de l'équation (6).

Le présent travail possède un lien étroit avec la topologie des champs de densité électronique et de potentiel électrostatique. On souhaite reproduire à terme cette topologie à l'aide du modèle HCM. Il est donc naturel d'expliciter le concept de polarisabilités topologiques qui sera utilisée au cours des chapitres suivants.

Dans la partie concernant la topologie de la densité électronique, il a été vu que les surfaces de flux nul de la densité découpent l'espace en bassins qui ne laissent aucun vide : c'est une partition exhaustive, ou parfaite (contrairement à un empilement de sphères par exemple). Chaque bassin contient également un noyau, on tirera donc avantage du fait que les expansions multipolaires sont centrées sur les noyaux. On n'utilisera pas de dipôles au niveau des points critiques de liaison comme dans d'autres approches[49], et cette approximation sera explorée par la suite.

Les moments dipolaires induits sont additifs dans le modèle ponctuel. Pour un bassin atomique neutre, le moment dipolaire $\boldsymbol{\mu}_\Omega$ est indépendant de l'origine et est donné par :

$$\boldsymbol{\mu}_\Omega = \int_\Omega \rho_{total}(\mathbf{r})(\mathbf{r} - \mathbf{r}_\Omega)\, d\mathbf{r} \qquad (35)$$

Où ρ_{total} représente la densité de charge totale contenue dans le bassin (électrons et noyau). L'intégration s'étend sur le volume du ρ-bassin atomique. $\mathbf{r}_\Omega$ est le vecteur position du noyau associé au bassin (tous les ρ-bassins ont un noyau).

Pour obtenir les variations des dipôles de chaque bassin, des simulations théoriques ont été réalisées par des collaborateurs. Un jeu de structures théoriques de référence, le S66[50,51], a été servi de banc d'essai, en raison du très haut niveau de théorie utilisé pour optimiser les géométries moléculaires (MP2). Emmanuel Aubert (CRM²) a réalisé les calculs portant principalement sur le jeu de données S66. Anna Krawczuk (Groupe d'ingénierie cristalline et d'analyse structurale, *Jagiellonian University*, Cracovie) et Katarzyna Rzesikowska (*idem.*), alors doctorante en visite ERASMUS au CRM², ont

calculé les polarisabilités d'acides aminés et peptides tirés de la CSD[16]. Dans les deux cas le niveau de théorie est B3LYP, mais le *basis-set* diffère. Les détails des méthodes sont regroupés parties 1.7.b et 1.8. Dans les deux cas, la simulation est réalisée avec plusieurs champs électriques dans des directions variées. La partition QTAIM de la densité est réalisée à chaque valeur de champ, et les dipôles des bassins atomiques sont calculés à l'aide de l'équation (35). Les polarisabilités atomiques ont ensuite été déduites des dépendances linéaires de ces moments dipolaires atomiques par rapport au champ électrique externe appliqué, comme expliqué chapitre 1, parties 7.b et 8.

Comme la partition QTAIM de l'espace est exhaustive, la somme des moments dipolaires des bassins atomiques est le moment dipolaire moléculaire. De même, pour les polarisabilités, la sommation de toutes les polarisabilités atomiques $\boldsymbol{\alpha}_i$ conduit aux polarisabilités moléculaires $\boldsymbol{\alpha}_{\text{mol}}$.

$$\boldsymbol{\alpha}_{mol} = \sum_{atomes} \boldsymbol{\alpha}_i \qquad (36)$$

La partition de la polarisabilité moléculaire $\boldsymbol{\alpha}_{\text{mol}}$ en contributions atomiques $\boldsymbol{\alpha}_i$ est discutée dans les sections 8 et 9 de « l'État de l'Art », ainsi que dans la méthode utilisée pour modifier la densité électronique transférée en fonction de $\boldsymbol{\mu}_i$. L'erreur sur l'additivité des polarisabilités topologiques a été examinée dans l'article Leduc *et al.*[52] (parties 1.7 et 2.4).

La charge électrique d'un bassin topologique est calculée d'une manière homologue à dipôle, utilisant la formule (13), page 33.

12 - Calculs d'énergie de polarisation intermoléculaire E_{POL}

La méthode de polarisation proposée ici conduit à des modifications des paramètres atomiques (les populations dipolaires) utilisés pour représenter la densité électronique multipolaire de molécules en interaction. Par conséquent, l'application de cette procédure modifie l'énergie totale d'interaction électrostatique telle que calculée par EP/MM[27], puisque les molécules subissent une induction mutuelle. Il est donc possible de définir $E_{POL,DB}$, l'énergie de polarisation basée sur les densités de la base de données ELMAM2 polarisée, comme suit :

$$E_{POL,DB} = E_{ELEC,POLARIZED} - E_{ELEC,DB} \qquad (37)$$

où $E_{ELEC,DB}$ représente l'énergie d'interaction électrostatique entre les densités de charge des monomères transférées depuis la base de données, et $E_{ELEC,POLARIZED}$ représente l'énergie d'interaction électrostatique corrigée entre les densités de charge obtenues après application de la polarisation. $E_{ELEC,DB}$ et $E_{ELEC,POLARIZED}$ sont tous deux calculés à partir de l'équation (10), mais utilisent des densités électroniques légèrement différentes. Par définition, l'énergie de polarisation $E_{POL,DB}$ est toujours négative. La polarisabilité étant un tenseur défini positif, les moments dipolaires induits $\boldsymbol{\mu}_i$ s'alignent avec le champ électrique $\mathbf{E}(\mathbf{r})$ et sont toujours dans une orientation énergétiquement favorable. Le produit scalaire $\mathbf{E}^t(\mathbf{r})\cdot\boldsymbol{\mu}_i$ est toujours positif[46], voir équation (39). $E_{ELEC,DB}$ est obtenu en utilisant $\rho_{DB}(\mathbf{r})$, la densité de charge non perturbée, tandis que $E_{ELEC,POLARIZED}$ est calculée en utilisant $\rho_{total}(\mathbf{r})$, la densité de charge polarisée (perturbée). Ces densités électroniques sont reliées par :

$$\rho_{total}(\mathbf{r}) = \rho_{DB}(\mathbf{r}) + \rho_{pol}(\mathbf{r}) \qquad (38)$$

Où $\rho_{pol}(\mathbf{r})$ est la densité de déformation induite, construite à partir de composants dipolaires, associée à la polarisation *in situ* de la densité électronique transférée. Dans le présent travail, la méthode de polarisation est appliquée à des dimères (voir partie 1.7.a), il convient de souligner que $E_{POL,DB}$ est ici l'énergie de polarisation du dimère.

Spackman *et al.* ont récemment appliqué une autre méthode pour calculer les énergies de polarisation, utilisant des tenseurs isotropes. Cette méthode est incluse dans le programme *CrystalExplorer*[53,54]. Cette méthode est similaire à celle utilisée pour calculer l'énergie de polarisation d'un e-pixel dans le logiciel *PIXEL*[55,56]. Dans cette approche, l'énergie potentielle électrostatique de polarisation est calculée pour un atome de polarisabilité isotrope α_i sous l'influence d'un champ électrique externe $\mathbf{E}_i$ aux coordonnées nucléaires de l'atome *i*. Ainsi pour un système de *i* atomes (en général une molécule), l'énergie de polarisation s'écrit :

$$E_{POL} = -\frac{1}{2}\sum_i \alpha_i \cdot |\mathbf{E}_i|^2 \qquad (39)$$

où la somme porte sur les *i* atomes qui ressentent le champ électrique externe $\mathbf{E}_i$. Pour appliquer cette approche et la comparer à celle décrite dans ce manuscrit, une modification est nécessaire adapter la formule (39) polarisabilités anisotropes $\boldsymbol{\alpha}_i$:

$$E_{POL} = -\frac{1}{2}\sum_i(\mathbf{E}_i^t \cdot \boldsymbol{\alpha}_i \cdot \mathbf{E}_i) \qquad (\mathbf{40})$$

Les équations (39) et (40) peuvent être appliquées à un système complet (par exemple un dimère) ou une partie de ce système (par exemple l'un des monomères) et donner l'énergie de la partie de ce système. Il sera vu partie 4.3 que découper de cette manière l'énergie de polarisation calculée par EP/MM est moins direct.

Afin de comparer les calculs d'énergie de polarisation obtenus par les méthodes introduites ici, une référence théorique est nécessaire. Les méthodes DFT ou MP2 peuvent être utilisées pour obtenir les polarisabilités théoriques, ou pour optimiser les géométries des jeux de structures conçus pour le *benchmarking*, comme le S66[50,51]. Cependant ces méthodes ne calculent que des énergies totales. Une énergie de référence théorique est nécessaire et cette référence doit comporter des contributions énergétiques partitionnées de manière physiquement sensée.

C'est le cas de la méthode SAPT (Symmetry Adapted Perturbation Theory[57]) qui divise les énergies d'interaction totales en contributions énergétiques électrostatiques permanentes, d'induction, de dispersion et d'échange-corrélation. SAPT calcule les contributions d'induction à l'aide d'une méthode perturbative. Dans ce cadre, la perturbation représente l'interaction entre des monomères autrement isolés, sous la forme d'une série de puissances entières.

La perturbation de premier ordre est E_{ELEC}, l'interaction Coulombienne entre les densités de charge. E_{IND}, l'énergie d'induction SAPT, regroupe les perturbations du deuxième et du troisième ordre.

$$E_{SAPT} = E_{ELEC} + E_{IND} + E_{DISP} + E_{EX-CORR} \qquad (\mathbf{41})$$

Li *et al.*[58] ont comparé différentes versions de la théorie SAPT (ainsi que des simulations de type DFT et PMx) à l'aide du set S66x8 et d'autres jeux de données de la BEGDB[59]. Les jeux de données de la BEGDB utilisés ici et par Li *et al.* sont constitués de structures dont les géométries (et énergies totales) ont été optimisées par des simulations

de type MP2. Le présent travail utilise les valeurs d'énergie SAPT publiées par Li *et al.* en tant qu'énergie de référence du jeu de structures S66x8. Sa composition et son utilisation sont abordés en détails partie 2.4. Le présent travail se concentre sur la reproduction de la partie électrostatique et d'induction de la décomposition SAPT. Les deux sont accessibles par la densité électronique : la contribution électrostatique (E_{ELEC}) de SAPT peut être comparée aux énergies d'interaction électrostatiques calculées à partir de la densité électronique ELMAM2 transférée [équation (10)]. L'énergie d'induction issue de la décomposition SAPT (E_{IND}), qui inclut les effets d'induction dipolaires, est comparée à l'énergie de polarisation dérivée de la densité électronique ELMAM2 polarisée [équation (37)]. Les contributions E_{DISP} et $E_{EX\text{-}CORR}$ ne peuvent être dérivées de la densité électronique cristallographique.

13 - Bases de données

Afin de modéliser une densité électronique suffisamment précise pour réaliser une analyse topologique du potentiel qui en dérive, tout en conservant le même modèle pour étudier petites molécules et macromolécules, une approche basée sur les bases de données de paramètres décrivant la densité électronique est utilisée. Si le calcul d'une densité électronique théorique sur des édifices de milliers d'atomes est de plus en plus accessible par des méthodes de type diviser-et-conquérir[10], ces méthodes sont encore à un stade précoce de développement. Elles sont de plus déconnectées de l'observation expérimentale de la densité électronique en cristallographie. Les tentatives de lier densité théorique et expérience de cristallographie, notamment par le moyen de fonctions d'ondes contraintes par les rayons X, sont encore très novatrices[11].

La base de données de densité électronique ELMAM2 contient des paramètres expérimentaux qui permettent la reconstruction de densités électroniques moléculaires décrites dans le modèle HCM. La fondation de cette initiative est le principe de transférabilité, appliqué à la densité électronique. La première référence à ce principe est retrouvée dans Brock *et al.*, 1991[60], où la densité électronique du pérylène est utilisée pour reconstruire celles du naphtalène et de l'anthracène. La transférabilité d'une propriété donnée est définie comme la conservation de cette propriété entre deux objets similaires dans des contextes différents (ici, la densité électronique). On peut assumer que

la mesure de cette propriété sera raisonnablement proche entre ces objets similaires. Cette définition est souple, il faut choisir un seuil d'erreur tolérable sur la conservation de la propriété à transférer (variable en fonction de la propriété). Au départ, la librairie ELMAM2 a été développée pour la densité électronique. La première version fût publiée peu après le papier de Brock *et al.* (Pichon-Pesme *et al.*, 1995[21]). Pour ELMAM2, c'est le voisinage covalent des atomes qui est pris en compte pour quantifier la similarité de leur distribution électronique. Le présent travail étend l'applicabilité du principe de transférabilité aux polarisabilités.

La base de données ELMAM2 décrit principalement la densité électronique des types d'atomes qu'on trouve dans les molécules organiques (telles que les biomolécules et leurs ligands). Ces paramètres proviennent d'affinements cristallographiques par moindres carrés de petites molécules organiques contre des données de diffraction des rayons X à haute résolution. Les paramètres stockés dans la bibliothèque ELMAM2 ont été obtenus sur la base du principe de transférabilité énoncé précédemment. Pour créer une entrée (un type d'atome) dans ELMAM2, des atomes chimiquement similaires sont supposés avoir une densité électronique – et donc des paramètres HCMM – étroitement comparables. Malgré un environnement cristallin différent, les paramètres associés à ces atomes similaires peuvent être moyennés et symmétrisés[21,61] pour retrouver un pseudoatome raisonnablement transférable entre différents environnements cristallins ou moléculaires.

Ces paramètres et les densités électroniques qu'ils permettent de reconstruire ont été comparés à des densités théoriques à plusieurs reprises[62,63,64,65]. Voici, en guise d'exemple, les résultats obtenus par Domagała *et al.*, 2012[61]. Dans cette étude, les coordonnées cristallographiques d'une série de composés (de type aromatiques et tri-peptide) ont servi à l'obtention de densités théoriques avec CRYSTAL-06, utilisant la fonctionnelle B3LYP et le *basis set* 6-31(*d,p*). Cette densité a permis l'obtention de facteurs de structure théoriques. *MoPro* a été utilisé pour affiner un modèle HCM de cette densité électronique théorique. La comparaison entre modèles est évaluée par comparaison des moments électrostatiques et du potentiel électrostatique calculé sur une grille de pas 0,1 Å, en considérant les points à moins de 0,3 Å de la surface de van der Waals. Le *RMSD* [équation (A2), annexe page 291] sur les charges topologiques (QTAIM), entre ELMAM2 et B3LYP 6-31(*d,p*) est 0.061 e, et le coefficient de détermination R^2 est 98,4%. Pour les

moments dipolaires, le R^2 est supérieur à 79,2% pour tous les groupes de composés. L'erreur relative sur ces moments dipolaires est au maximum 19,8%, et encore supérieure pour les quadrupôles (maximum 36,8%). Pourtant, l'erreur relative sur le potentiel électrostatique à la surface de van der Waals est de 13,3% pour les peptides considérés et 30,1% pour les composés aromatiques[61]. Malgré ces différences de potentiels et de moments, la librairie ELMAM2 a permis sur ces deux jeux de structures d'obtenir une détermination R^2 de 97,0% et une erreur relative de 14,1% (peptides) et 16,9% (aromatiques), sur les énergies d'interaction électrostatiques.

Dans cet exemple, la comparaison expérience/théorie entre les énergies d'interaction électrostatiques portait sur les valeurs obtenues après transfert des paramètres ELMAM2, sans tenir compte de l'environnement non covalent des atomes propres aux cristaux considérés. En revanche, le présent travail comporte une réévaluation des énergies d'interaction ELMAM2 sur de nouveaux composés, et utilisant un modèle théorique plus évolué, développé plus bas. Cette réévaluation a été publiée au cours de la thèse et est reproduite partie 1.7 et 2.4. D'autres bases de données d'atomes multipolaires existent, basées sur des facteurs de structure théoriques, notamment UBDB[66,67,68,69] et la base de données Invariom[70].

14 - Estimation de la « distance » entre tenseurs de polarisabilité

En 1998, pour examiner la transférabilité des paramètres de déplacement atomique anisotropes, Jelsch *et al.*[71] ont publié une formule pour la norme d'un tenseur 3×3 symétrique. Cette norme (différente du déterminant et de la trace du tenseur) suit :

$$||\boldsymbol{A}|| = \sqrt{\frac{1}{3}\left(\sum_i \boldsymbol{A}_{ii}^2 + 2\sum_{i<j} \boldsymbol{A}_{ij}^2\right)} \qquad (\mathbf{42})$$

La norme $||\boldsymbol{A} - \boldsymbol{B}||$ est interprétée par la suite comme une distance entre les tenseurs $\boldsymbol{A}$ et $\boldsymbol{B}$. Elle sera bien entendu appliquée à des tenseurs de polarisabilité là où l'application originale visait les paramètres de déplacement atomiques anisotropes B_{ij}.

15 - Cinétique enzymatique et titration isotherme par calorimétrie

Idéalement, on souhaiterait comparer les calculs d'énergie issus de la méthode développée plus haut à des énergies d'interaction expérimentales. La présente partie regroupe donc les éléments méthodologiques pertinents pour obtenir ces énergies expérimentales en vue de la comparaison avec leurs homologues calculés. Fondamentalement, l'énergie de liaison entre les deux partenaires d'un complexe protéine-ligand non-covalent est liée à la constante d'équilibre chimique liée à cette réaction de complexation. Cette constante d'équilibre représente si le complexe peut exister dans un mélange réactionnel, et permet d'apprécier l'évolution du système vers un état donné. Le système considéré est une association réversible protéine-ligand, en solution aqueuse :

$$a \cdot P + b \cdot L \rightleftarrows c \cdot P_a L_b \qquad (\mathbf{43})$$

Où a est le coefficient stœchiométrique de la protéine P, b le coefficient stœchiométrique du ligand L et c celui du complexe non-covalent P_aL_b. Le quotient de réaction est le produit des activités chimiques des réactifs, divisées par l'activité chimique des produits de la réaction. Le cas présenté ici est fortement idéalisé (les partenaires sont totalement solubles, toutes les protéines sont correctement repliées, le nombre de site de fixation fonctionnels par protéine est constant, etc.). L'activité chimique d'une espèce est donc sa concentration, ici $[P]$, $[L]$ ou $[P_aL_b]$ (en M soit en $mol \cdot L^{-1}$), si les coefficients stœchiométriques a et b sont égaux à 1 et que la solution est diluée.

Une dose certaine de confusion peut entourer ces notions essentielles, surtout lorsqu'elles impliquent les « us et coutumes » des biochimistes ou des chimistes (qui ne rapportent pas forcément les mêmes grandeurs dans leurs publications). L'équation chimique (43) est réversible. Si on la lit dans le sens de l'association (de gauche à droite), le quotient de réaction Q_a associé à ce sens est le rapport des produits des activités chimiques :

$$Q_a = \frac{\prod activité(produits)}{\prod activité(réactifs)} = \frac{[P_aL_b]^c}{[P]^a[L]^b} \qquad (\mathbf{44})$$

Q_a peut être calculé à tout moment de la réaction, cependant, lorsque la réaction atteint l'équilibre, les concentrations $[P]$, $[L]$ et $[P_aL_b]$ cessent par définition d'évoluer et

prennent les valeurs $[P]_f$, $[L]_f$ et $[P_aL_b]_f$. Ces concentrations sont stables car la vitesse de formation de $[P_aL_b]$, notée v_{on}, est égale à la vitesse de formation de $[P]$ et de $[L]$, notée v_{off}. À tout moment, les vitesses de réaction v_{on} et v_{off} (en $M \cdot s^{-1}$) dépendent de constantes propres à la réaction :

$$v_{on} = k_{on} \cdot [P]^a \cdot [L]^b \qquad \text{et} \qquad v_{off} = k_{off} \cdot [P_aL_b]^c \qquad (\mathbf{45})$$

Les unités des constantes k_{on} et k_{off} dépendent de la stœchiométrie, là où v_{on} et v_{off} représentent toujours des moles d'évènements par seconde et par litre. Cependant, pour la plupart des applications biochimiques, k_{on} est exprimé en $M^{-1} \cdot s^{-1}$ et k_{off}, en s^{-1} (facilement achevé en pratique en considérant la concentration en site de fixation à la place de la concentration réelle en protéines). La diffusion de P et L limitent la constante k_{on} à une valeur théorique appelée limite de diffusion[72].

À l'équilibre v_{on} et v_{off} sont égales et $Q_a = K_a$, où K_a est la constante d'association. K_a est une constante d'équilibre mais la notation K_{eq} sera évitée (voir plus bas). L'unité de K_a est la M^{-1} dans le cas le plus simple (un site de fixation par protéine).

$$K_a = \frac{[P_aL_b]_f^c}{[P]_f^a[L]_f^b} = \frac{k_{on}}{k_{off}} \qquad (\mathbf{46})$$

K_a permet de déterminer l'évolution d'un mélange contenant P, L et/ou P_aL_b. Si, à un temps donné, Q_a est inférieur à K_a, les concentrations $[P]$, $[L]$ ou $[P_aL_b]$ vont progressivement évoluer jusqu'au moment où $Q_a = K_a$. K_a est peu reporté dans les publications traitant d'interactions biochimiques, où K_d, la constante de dissociation, lui est préférée :

$$K_d = \frac{[P]_f^a[L]_f^b}{[P_aL_b]_f^c} = \frac{k_{off}}{k_{on}} = {K_a}^{-1} \qquad (\mathbf{47})$$

On considère simplement la réaction dans le sens inverse, de P_aL_b vers P et L. Le quotient de réaction sera alors $Q_d = {Q_a}^{-1}$. Cette fois-ci, si Q_d est supérieur à K_d, la solution évoluera jusqu'au point où $Q_d = K_d$. K_d a la dimension d'une concentration et correspond à la moitié de la concentration $[L]$ pour obtenir la saturation de P (encore une fois, dans le cas où on considère un site de fixation par protéine). Cette préférence provient de la

facilité d'utilisation de K_d pour décrire le mécanisme classique de Michælis-Menten, qui modélise une réaction enzymatique irréversible à une seule étape :

$$a \cdot P + b \cdot L \underset{k_{off}}{\overset{k_{on}}{\rightleftarrows}} c \cdot P_a L_b \xrightarrow[k_{cat}]{} a \cdot P + b \cdot M \qquad (\mathbf{48})$$

Où M est le produit de la réaction, à partir du substrat L (on parlera de substrat si le ligand est impliqué dans une réaction chimique). Les constantes de vitesse associées à chaque réaction sont notées au-dessus des flèches matérialisant chaque réaction. Cette équation chimique représente la nécessité de contact moléculaire avant la réaction catalytique. Le complexe P_aL_b va être impliqué dans deux réactions en compétition : la dissociation et la réaction enzymatique à proprement parler. L'efficacité globale de ce processus est appréciée par la constante de Michælis-Menten, K_m :

$$K_m = \frac{k_{off}+k_{cat}}{k_{on}} \qquad (\mathbf{49})$$

En assumant [P] <<< [L] (la dissociation gouvernée par k_{off} devient négligeable, tout comme la concentration [P] en enzyme libre), K_m est interprétable comme la concentration [L] pour laquelle la vitesse de réaction est égale à la moitié de la vitesse maximale V_{max}.

K_a, K_d et K_m sont des constantes dites macroscopiques car elles peuvent être mesurées expérimentalement, directement par titration. Par opposition, les constantes k_{on}, k_{off} et k_{cat} sont dites microscopiques car elles ne peuvent être mesurées directement. Il est possible d'extrapoler leur valeur et d'en obtenir une mesure indirecte.

Une possibilité pour obtenir cette mesure est de réaliser une expérience de compétition. On dispose pour cet exemple de P, L et L^*, un homologue de L marqué et quantifiable (par exemple par un radio-isotope, tant que L et L^* ont le même site de fixation et le même K_d). A l'état initial, P est en excès par rapport à L^*. À ce point [L^*] = 0 M, et une certaine fraction de P sera sous forme $P_aL_b^*$ de concentration $[P_aL_b]_0$. On démarre la dissociation en ajoutant à cette solution un très large excès de L froid pour différents temps t d'incubation. Après ce temps t, on arrête la réaction (séparation) et on

dose la concentration $[L^*]_t$ libre, ou $[P_aL_b^*]_t$ (en fonction du système étudié). On s'intéresse à la fraction de partenaires encore liés à t, pour notre exemple :

$$F_B = \frac{[P_aL_b*]^c{}_t}{[P_aL_b*]^c{}_0} \text{ ou } \quad F_B = \frac{[P_aL_b*]^c{}_0 - [L*]^b{}_t}{[P_aL_b*]^c{}_0} \qquad (\mathbf{50})$$

Vis-à-vis de P, L et L^* ont le même k_{on}, par conséquent, $[L^*]$ libre, ou $[P_aL_b^*]$ en fonction, ne dépend que de k_{off} :

$$\ln(F_B) = -k_{off} \cdot t \qquad (\mathbf{51})$$

Ce type de détermination indirecte des constantes microscopiques est très fréquent en enzymologie. Souvent la méthodologie précise doit être adaptée au système (en fonction du type de marquage, poids moléculaire relatif, stabilité des partenaires, immobilisation éventuelle, valeur estimée de K_d ou K_m…).

Une importante méthode expérimentale qui permet déterminer l'affinité de deux partenaires est la titration isotherme par calorimétrie (ITC). L'ITC utilise deux cellules contenant l'un des deux partenaires à étudier. Pour un système protéine-ligand plusieurs critères sont pris en compte pour choisir le partenaire injecté et celui de la cellule (solubilité, disponibilité…). L'ITC est adaptés aux interactions protéine-protéine. La machine entretient un gradient de température constant entre ces deux solutions. On ajoute des petits volumes de l'autre partenaire dans l'une des cellules. Une fixation spontanée dégage de la chaleur, la machine devra compenser la chaleur générée pour maintenir le gradient constant. Ce travail est représenté sur la courbe de gauche de la Figure 7. En effet, chaque pic représente le travail nécessaire au spectromètre pour maintenir une température constante dans la cellule de test, suite à une injection de ligand dans la cellule†.

† *: Les pics situés au-dessus du 0 de la courbe de gauche (Figure 7) sont liés à la dilution du partenaire injecté, qui est ici endothermique et devient visible seulement à partir de la saturation.*

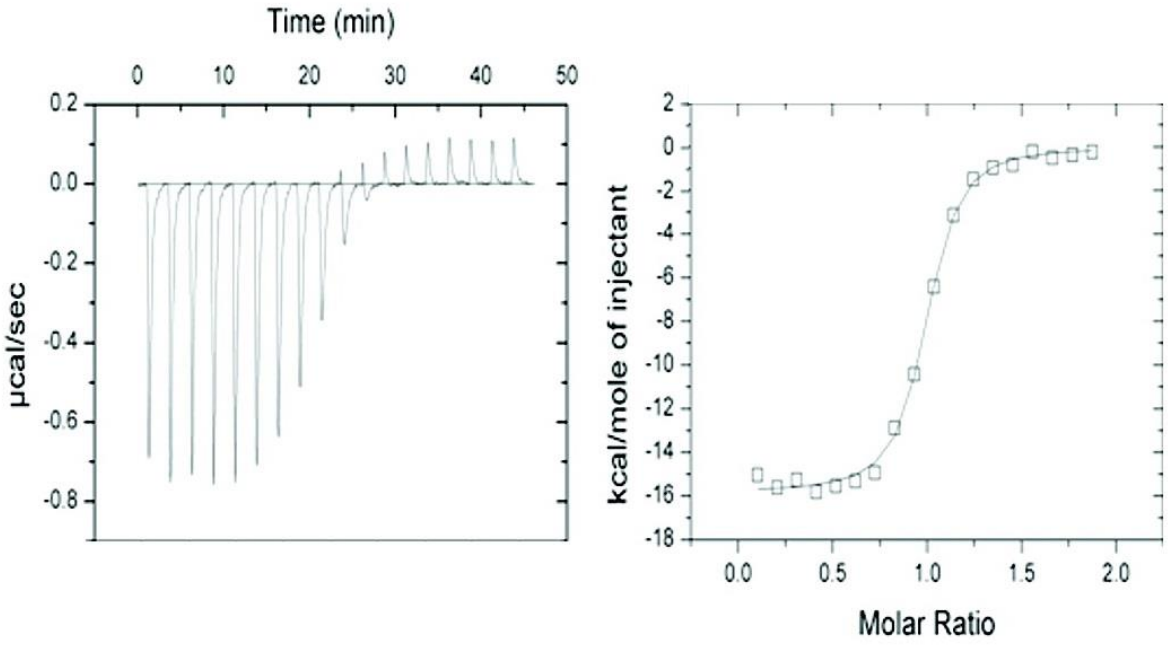


Figure 7. Un spectre d'ITC typique, à gauche, mesuré entre une entérotoxine C3 mutante de *Staphylococcus aureus* et la chaîne *β*8.2 du T-Cell Receptor de *Mus musculus* (reproduit de Cho *et al.*, 2010[73]). La courbe de droite relie l'amplitude des pics aux concentrations relatives des partenaires.

Après avoir pris en compte la concentration des partenaires, le point d'inflexion de la courbe à droite de la Figure 7 donne K_d. L'ITC est une technique consommant de grandes quantités de protéine, heureusement son intérêt n'est pas limité à la détermination de K_d. L'ITC permet d'accéder aux différentes grandeurs thermodynamiques de l'interaction étudiée, qu'il convient de définir.

Tout d'abord, la constante d'équilibre K_d d'une réaction chimique est liée à la différence relative d'énergie ΔG^0 entre les états initiaux et finaux par la relation de Gibbs-Helmholtz :

$$\Delta G^0 = -RT \cdot \ln(K_d) \qquad (\mathbf{52})$$

où T est la température en Kelvins, R est la constante des gaz parfaits ($J \cdot mol^{-1} \cdot K^{-1}$) et ΔG^0 est l'énergie libre de Gibbs. Cette relation exprime que plus la variation d'énergie liée à l'interaction est favorable, plus la constante de dissociation sera faible et l'équilibre sera orienté vers l'existence du complexe. L'énergie libre de Gibbs est :

$$\Delta G^0 = \Delta H^0 - T\Delta S^0 \qquad (\mathbf{53})$$

Dans cette relation, ΔH^0 est la variation standard d'enthalpie associée à la formation du complexe (en kJ·mol^{-1}) et ΔS^0 est la variation d'entropie standard (kJ·mol^{-1}·K^{-1}). L'enthalpie est la somme de l'énergie interne du système et du produit pression par volume. À pression et de volume constants, la variation ΔH^0 représente donc l'énergie interne échangée par le système. La relation (53) est applicable à une fixation réversible dans ces conditions physico-chimiques. Dans le cadre d'une interaction protéine-ligand, cette énergie dépend des interactions non-covalentes qui seront rompues ou créées par la fixation. La variation d'entropie dépend de la variation des degrés de liberté du système quand la fixation se produit. Dans le cadre des interactions considérées, cette variation est liée à la désolvatation du site de fixation, à l'immobilisation du ligand (diffusion, rotations, conformations internes), à l'immobilisation ou la libération de chaînes latérales ou à la modification de mouvements conformationnels plus larges.

L'ITC permet une mesure directe de ΔH^0. En effet, intégrer l'aire sous la courbe de travail Figure 7 (courbe de gauche) donne une énergie proportionnelle à ΔH^0. L'équation (53) permet de retrouver ensuite $-T\Delta S^0$. Réaliser plusieurs spectres ITC à différentes températures permet d'accéder à ΔC_p, la variation de capacité thermique isobare, qui décrit la dépendance de ΔH^0 à la température[74,75] (ΔC_p est exprimé en kJ·mol^{-1}·K^{-1}) :

$$\Delta C_p = \frac{d\Delta H^0}{dT} \qquad (\mathbf{54})$$

Les différentes grandeurs thermodynamiques sont représentées Figure 8. L'état de transition est un maximum local d'énergie que le système doit franchir pour passer de l'état lié *PL* à l'état libre *P*+*L*. Cet investissement énergétique nécessaire est $\Delta G^{\neq}$. La durée de vie de l'état de transition est trop courte pour être observable mais sa prise en compte est essentielle à la compréhension des cinétiques chimiques.

Plus l'énergie de l'état de transition sera basse, plus la réaction sera rapide. Abaisser $\Delta G^{\neq}$ entre substrat et état de transition est d'ailleurs un des mécanismes classiques pour expliquer l'efficacité de la catalyse enzymatique. Les constantes microscopiques k_{on} et k_{off} sont reliées à l'énergie de l'état de transition par des relations de type Arrhenius. L'équation d'Eyring[76] permet de relier k_{off} à l'énergie de l'état de transition :

$$\Delta G^{\neq} = -RT \cdot \ln(\frac{k_{off} \cdot h}{k_B \cdot T}) \qquad (55)$$

Les mesures des valeurs de k_{off} sont réalisées à différentes températures pour extraire $\Delta H^{\neq}$ de la dépendance entre $\Delta G^{\neq}$ et la température. Il faut cependant noter que le modèle de densité électronique présenté ici n'est applicable qu'aux molécules dans un état d'équilibre, on exclut donc *a priori* toute simulation de l'état de transition. Cependant les données thermodynamiques tirées de l'analyse de l'état de transition peuvent éclairer notre interprétation des interactions non-covalentes des protéines. Un exemple sera abordé au chapitre 3, partie 7.

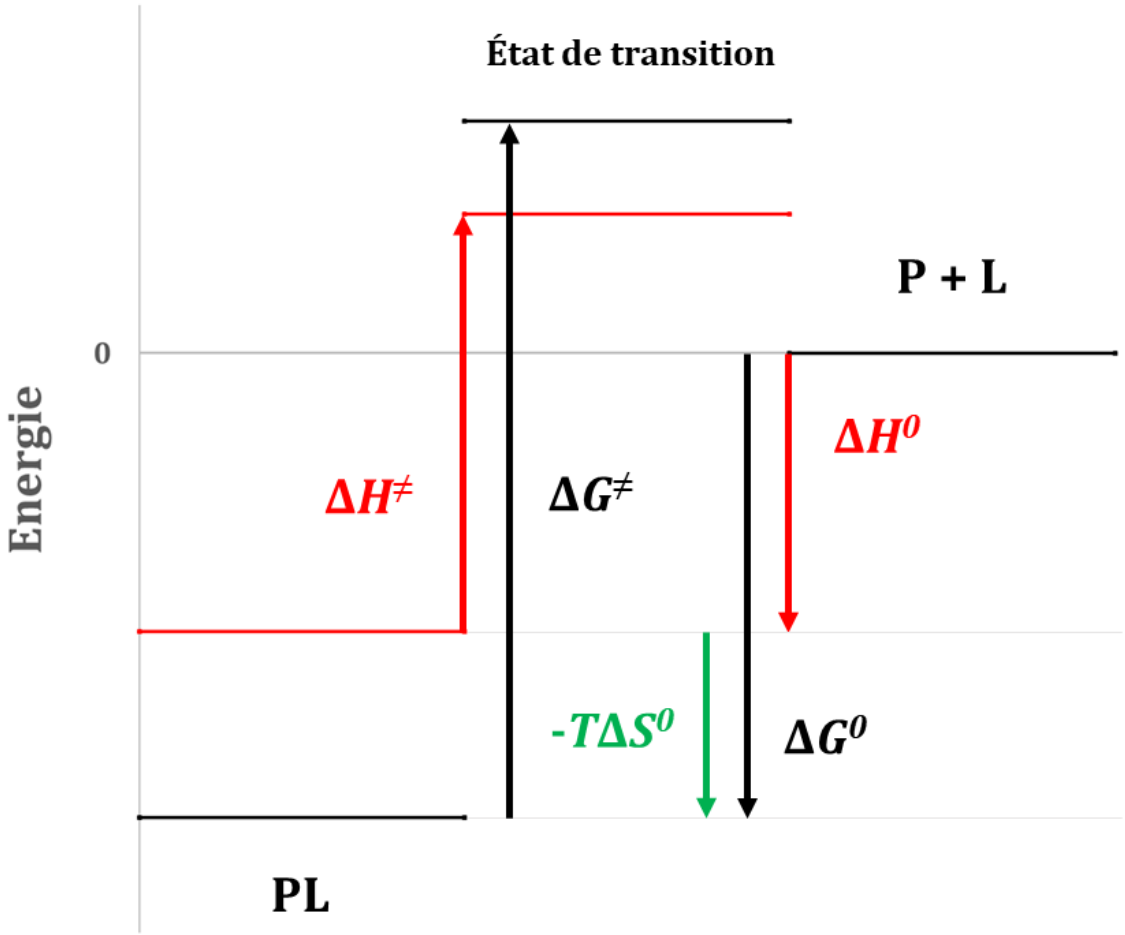


Figure 8. Les différentes grandeurs thermodynamiques utilisées. Les lignes noires représentent les niveaux d'énergie des trois états du système. Les lignes rouges représentent les niveaux d'enthalpie totale des états du système (l'énergie interne). Une valeur positive de $-T\Delta S^0$ est impossible si l'association *PL* est favorable, mais est possible quand on considère l'état de transition.

Chapitre 1 :

Polarisabilités

Atomiques

Topologiques

Introduction

Dans ce chapitre, les polarisabilités atomiques et moléculaires, issues de calculs *ab initio*, sont décrites en vue de leurs utilisations dans les chapitres suivants. Notablement, nos deux principaux objectifs sont de reproduire les polarisabilités moléculaires et d'utiliser des polarisabilités atomiques théoriques pour modifier (polariser) la densité électronique multipolaire transférée depuis la librairie ELMAM2. La polarisation à proprement parler est l'objet du chapitre suivant. Dans un premier temps, on souhaite – pour chaque polarisabilité atomique – accéder à sa variabilité entre différents contextes d'interaction. Évaluer ces variations interatomiques est crucial pour établir dans quelle proportion les polarisabilités atomiques sont transférables.

Dans le but général d'utiliser ces propriétés théoriques sur d'autres structures (y compris expérimentales), un outil adapté a été ajouté à *MoProViewer*. Pour éviter de biaiser cette évaluation, l'outil présenté ici a été conçu pour être indépendant des types atomiques de la librairie ELMAM2. Bien entendu, cet outil est cependant totalement compatible avec la librairie ELMAM2, en particulier pour permettre la construction de librairies de polarisabilités atomiques transférables.

Le contexte de développement logiciel de ce travail, le *package MoProSuite*, est décrit partie 1.1.

L'outil comparaison et de transfert des polarisabilités anisotropes est présenté partie 1.2.

Ce module a requis l'implémentation de trois méthodes particulières pour obtenir les comparaisons, elles sont décrites de la partie 1.3 à la partie 1.5.

La partie 1.6 traite de l'interface avec l'utilisateur et des résultats obtenus par l'utilisation de cet outil.

Cette description des polarisabilités atomiques du jeu de données S66 occupe une large place de l'article rédigé durant cette thèse. Ce jeu de données a été étendu par la suite. La partie 1.7 reprend les éléments publiés dans Leduc *et al.*, 2019[52].

La partie 1.8 détaille enfin les extensions de ce jeu de données aux acides aminés (isolés, dipeptides, tripeptides), essentiels aux applications présentées chapitre 2.

1.1 - *MoProSuite* et son module de polarisation

MoProSuite est un ensemble de logiciels[28,77,78,79] développés au laboratoire CRM^2 pour l'affinement cristallographique de la densité de charge à l'aide du modèle HCM. Le logiciel *MoPro* assure cette fonction liminaire d'affinement par la méthode des moindres carrés, ainsi que le transfert de bases de données de densité électronique (voir plus bas). *VMoPro* est un module dédié aux calculs de propriétés moléculaires et à la génération de figures 2D vectorielles (PostScript). *VMoPro* est interfacé via *MoPro* et *MoProViewer* ou peut être accédé directement par la console. *MoProViewer* est un logiciel de visualisation moléculaire capable de rendus 3D. Ce dernier dispose d'une riche interface graphique (GUI) permettant l'accès facile à de nombreuses capabilités de *MoProSuite*. Le travail de cette thèse est intégralement inscrit dans le développement de cette suite logicielle, d'où son introduction ici.

La quasi-totalité du développement logiciel décrit dans le présent manuscrit a été implémenté dans *MoProViewer*. Le langage de programmation utilisé est le C++, assisté par le *framework* Qt. Les opérations d'algèbre linéaire utilisent la librairie *Armadillo* (Sanderson & Curtin, 2016[80]). Pour résumer, à chaque fois qu'un calcul ou qu'une visualisation a dépassé les capacités basiques d'un tableur, un code C++ a été rédigé et intégré à *MoProViewer*. Au fur et à mesure, les fonctions inutiles ont été retirées à mesure que les autres ont été organisées en outils plus opérationnels.

Tous les outils logiciels décrit dans ce manuscrit sont intégrés à un module de *MoProViewer* nommé « *Beluga Multipole Polarizer* », accessible depuis le menu « *Tools* ». Beluga est le nom de développement et également le nom de la classe C++ du module. Ce module est à l'origine chargé des opérations de polarisation de la densité électronique, comme son nom l'indique, mais regroupe de nombreuses fonctionnalités auxiliaires, nécessaires à la polarisation. Le module « *Beluga Multipole Polarizer* » permet :

- Le transfert et la comparaison de polarisabilités (depuis le résultat d'un calcul théorique ou d'une base de données). Cette fonctionnalité est l'objet de la partie suivante.

- Le calcul de polarisation intermoléculaire et la visualisation des résultats correspondants. La chapitre 2 présente les résultats obtenus au cours de la thèse par application de cet outil.

• La génération de paramètres pour construire des bases de données d'atomes polarisables. Les détails méthodologiques et d'implémentations sont trouvés au chapitre 3.

• Le transfert de paramètres multipolaires et de polarisabilités atomiques depuis une base de données. Cette fonction permet l'application reportée partie 3.7.

• L'automatisation de calculs d'énergies d'interaction électrostatiques et deux méthodes de calcul de l'énergie de polarisation. Cette fonctionnalité a rendu plus facile les applications, reportées partie 2.4, à la base de données de référence S66.

• L'accès à quelques fonctions liées à la surface et aux volumes des isosurfaces de densité électronique et de potentiel électostatique. Des exemples sont regroupés au chapitre 4.

• Diverses fonctions logistiques facilitant la mise en œuvre des études présentées aux endroits pertinents du présent manuscrit (ajustement de charges électriques, orthogonalisation et normalisation des repères moléculaires, génération de cartes de densités pour le logiciel Bader…).

1.2 - Transfert et comparaison de polarisabilités atomiques anisotropes

Comme introduit précédemment, un outil de comparaison et de transfert de polarisabilités orientées a été implémenté dans le logiciel *MoProViewer*. On s'intéresse à la création de bases de données d'atomes polarisables. Dans un premier temps, on souhaitera apparier les polarisabilités atomiques d'une même molécule dans différents contextes d'interaction avec une correspondance 1 pour 1. En effet, ceci permettrait de tester si les variations des polarisabilités ($\boldsymbol{\alpha}_{dimère}$ - $\boldsymbol{\alpha}_{monomère}$) sont un obstacle à leur transférabilité ou, au contraire, une justification de ce principe. La méthode choisie doit être indépendante de la définition des types atomiques de la base de données ELMAM2, afin d'améliorer le jeu de types atomiques ELMAM2 et les rendre encore plus applicables. Ainsi l'objectif est de contourner automatiquement les différences d'orientation, de

référentiel et de numérotation des atomes entre les systèmes décrits, soit sous forme de monomères, soit sous forme de dimères.

L'application du code de cet outil permet d'apparier chaque atome des dimères S66 à un atome unique présent dans les systèmes isolés (monomères) correspondants. La méthode décrite ci-après a été implémentée pour traiter les polarisabilités rapportées dans les parties 1.7 (publiées dans Leduc *et al.,* 2019[52]) et 1.8. Plus précisément, le présent outil permet deux usages. Tout d'abord, il permet de comparer, atome par atome, les polarisabilités atomiques entre systèmes monomères et dimères. Enfin, cet outil permet le transfert des polarisabilités atomiques calculées dans des systèmes monomères vers les atomes des dimères, dans une orientation relative correcte, afin de comparer leur rôle énergétique. Cet outil est accédé via le bouton « *Transfert Theor. Polar.* » du module « *Beluga Multipole Polarizer* ». Une autre version de l'algorithme, adapté au transfert sur une série de fichiers, est accessible via le bouton « *Transf.Theor.Polar.(Group)* ». Cette seconde version n'offre pas de visualisation du résultat contrairement à la première (voir partie 1.5).

Au-delà de son utilisation sur les dimères S66, le même outil peut être utilisé pour comparer des polarisabilités transférées à des polarisabilités théoriques sans aucune modification. Les comparaisons générées sont évaluées par l'utilisateur à l'aide un mélange de représentation 3D, d'indicateurs numériques et d'un rapport de comparaison détaillé.

L'outil est composé de deux méthodes imbriquées. La première est chargée de trouver la correspondance entre les géométries moléculaires, la seconde retrouve la correspondance entre atomes. Cet outil utilise deux fichiers moléculaires : le fichier « cible », en général le dimère, vers lequel les polarisabilités doivent être transférées, est ouvert dans l'interface graphique de *MoProViewer*. Le fichier « source », lui, contient les polarisabilités atomiques qu'on veut transférer. Le fichier « source » est ouvert par l'outil de comparaison/transfert après définition des atomes à faire correspondre à l'aide des sélections de *MoProViewer* (bouton « *Transfert Theor. Polar.* »). Tous les atomes du fichier « source » doivent apparaître dans la sélection. Les comparaisons de polarisabilités atomiques sont toujours effectuées dans le sens ($\boldsymbol{\alpha}_{cible}$ - $\boldsymbol{\alpha}_{source}$).

1.3 - Méthode d'Umeyama

Cette première méthode permet au logiciel de retrouver une correspondance géométrique entre deux molécules représentées comme des nuages de points (les noyaux). Comme les distances interatomiques sont relativement conservées, le problème revient à trouver la matrice de rotation optimale pour faire coïncider deux nuages de points.

Comme le montre la Figure 9, reconnaître un objet et trouver son orientation sont des tâches triviales pour un humain. La figure ci-dessous met également l'emphase sur le fait que cette reconnaissance instinctive peut se faire malgré des représentations très différentes de l'objet. Par exemple la photo contient de nombreuses données inutiles à la reconnaissance du champignon (feuilles mortes, mousse, etc.) alors que presque tous les caractères de l'art ASCII contribuent à l'image du champignon. Malgré une représentation très simple de nos molécules en tant que nuages de points et l'absence de données inutiles, les méthodes informatiques pour obtenir cette reconnaissance d'orientation ne sont pas triviales.



Figure 9. La reconnaissance de la forme et l'orientation de ces champignons sont instantanés pour le lecteur, malgré des données très différentes. À gauche, un champignon en art ASCII (source : www.asciiart.eu). À droite, un cèpe de Bordeaux (*Boletus edulis*) après une rotation de 90° (source : *Wikimédia Commons*).

La méthode originale de Shinji Umeyama[81] permet de déduire cette matrice de rotation, mais également de retrouver les composantes de translation et de mise à l'échelle (*scaling*) entre les systèmes de coordonnées. L'implémentation de la méthode au sein de *MoProViewer* se limite à la construction de la matrice de rotation. Cette matrice de rotation est applicable à toute propriété du système, et permet le transfert de propriétés orientées d'un référentiel à l'autre. Deux présupposés sont utilisés : on suppose que le référentiel est orthonormé (pas de mise à l'échelle des coordonnées atomiques) et que les variations de conformation moléculaire sont faibles (molécules rigides). De plus, les coordonnées moléculaires sont exprimées par rapport au centre de masse de la molécule sélectionnée (pas de translation). Dans l'article original, Umeyama utilise également des coordonnées relatives pour trouver le vecteur de translation.

Le problème est posé de la manière suivante. Les coordonnées nucléaires $\mathbf{y}_i$ d'une molécule « cible » de n atomes sont stockées dans une matrice $\mathbf{A}$ de $3\times n$ éléments, les coordonnées nucléaires $\mathbf{x}_i$ de la molécule « source » de n atomes sont stockées dans une matrice $\mathbf{B}$ de même dimension. $\mathbf{x}_i$ et $\mathbf{y}_i$ sont des coordonnées relatives au centre de masse du nuage de points. On cherche la matrice $\mathbf{R}$ qui minimise l'écart au carré $\sum_i(||\mathbf{A}_i - \mathbf{R}\mathbf{B}_i||^2)$.

Seuls les éléments effectivement implémentés dans *MoProViewer* sont décrits ici. Les équations (56) à (60) sont directement extraites de la démonstration originale. Cette démonstration est longue et complexe mais énumère les conditions pour construire cette matrice de rotation à partir des coordonnées des points $\mathbf{x}_i$ et $\mathbf{y}_i$.

On cherche donc $\mathbf{R}$, la matrice de rotation de $\mathbf{B}$ vers $\mathbf{A}$. Cette matrice est liée la décomposition en valeurs singulières de la matrice $\mathbf{\Sigma}_{xy}$ définie telle que :

$$\mathbf{\Sigma}_{xy} = \frac{1}{n}\sum_n(\mathbf{y}_i \otimes \mathbf{x}_i^t) = \frac{1}{n}\mathbf{A}\mathbf{B}^t \qquad (\mathbf{56})$$

Où '⊗' représente le produit de Kronecker (*outer product*) entre les vecteurs $\mathbf{y}_i$ et $\mathbf{x}_i$. $\mathbf{\Sigma}_{xy}$ est décrite comme un produit tensoriel de distances à la moyenne. Sa décomposition en valeurs singulières est :

$$\mathbf{\Sigma}_{xy} = \mathbf{U}\mathbf{D}\mathbf{V}^t \qquad (\mathbf{57})$$

Où les matrices $\mathbf{U}$ et $\mathbf{V}$ sont orthogonales. La mise à l'échelle entre les deux nuages de points utilise les valeurs singulières contenues dans la matrice diagonale $\mathbf{D}$. On suppose $\mathbf{D} = \delta_{ij}$ (où δ_{ij} est le delta de Kronecker) car le coefficient de mise à l'échelle n'est pas calculé

ici. En effet, on ne veut pas modifier les distances interatomiques des molécules auxquelles on appliquera la méthode. L'implémentation dans *MoProViewer* utilise la décomposition en valeur singulières de $\mathbf{AB}^t$:

$$\mathbf{AB}^t = \mathbf{U}(n.\mathbf{D})\mathbf{V}^t \qquad (\mathbf{58})$$

Pour obtenir une matrice de rotation **R** avec la bonne chiralité (*handedness*), Umeyama introduit la matrice diagonale **S** définie (pour un nuage de points en 3 dimensions) telle que :

$$\mathbf{S} = \begin{cases} \delta_{ij} & si\ \det(\mathbf{U}).\det(\mathbf{V}) = 1 \\ diag(1,1,-1) & si\ \det(\mathbf{U}).\det(\mathbf{V}) = -1 \end{cases} \qquad (\mathbf{59})$$

Umeyama définit **R** par :

$$\mathbf{R} = \mathbf{USV}^t \qquad (\mathbf{60})$$

Ainsi, le déterminant de **R** est toujours +1 donc **R** est une base directe (*right-handed*). Il serait possible de modifier la condition de l'équation (59) pour forcer l'obtention d'une matrice représentant une réflexion (déterminant -1), ce qui pourrait être utile pour transférer une propriété associée à un atome chiral sur un atome de chiralité opposée.

La principale contrainte de cette méthode est d'utiliser des listes de points appariés entre les deux systèmes. En effet, la méthode trouve la matrice de rotation qui minimise l'écart au carré des distances entre les points $\mathbf{x}_i$ et $\mathbf{y}_i$ sans évaluer si $\mathbf{x}_i$ et $\mathbf{y}_i$ correspondent à des atomes (ou n'importe quel objet ponctuel) comparables de la molécule. Dans *MoProViewer*, les atomes sont chargés en mémoire dans l'ordre du fichier, ce qui amène au développement de la méthode suivante.

1.4 - Permutations : génération et composition

Toujours dans l'optique de transférer ou comparer des tenseurs de polarisabilité entre systèmes moléculaires d'orientation différentes, il a été constaté que la méthode précédente génère des rotations aberrantes si $\mathbf{x}_i$ et $\mathbf{y}_i$ ne décrivent pas les coordonnées relatives du même atome dans les deux molécules (molécules cible et source). Si l'utilisateur tente effectivement de comparer deux molécules comparables (même

nombre d'atomes, même géométrie par rapport au centre de masse), il existe au moins une permutation **P** des $\mathbf{y}_i$ (les colonnes de **B**) qui permet de fortement minimiser le résiduel au carré $S^2 = \Sigma_i|\mathbf{A}_i - \mathbf{R}\cdot\mathbf{B}_i|^2$. Un ensemble général de fonctions a été implémenté dans *MoProViewer* pour générer et composer les permutations **P** ainsi que les appliquer à un ensemble de données stockées dans un vecteur ou une liste.

On souhaite donc trouver **P**. Une approche de type force brute a d'abord été envisagée. En effet la méthode d'Umeyama est en pratique rapide à calculer. La génération de toutes les permutations d'un vecteur {0,1,2, …, *n*-1} utilise l'algorithme classique de Heap et génère les *n*! vecteurs de permutation. Les permutations sont représentées sous forme de vecteurs d'éléments entiers comprenant 0, …, *n*-1. Cette représentation a été préférée à une représentation matricielle. L'exemple suivant montre l'équivalence entre la présente notation et la notation matricielle. Dans l'exemple ci-dessous, la première ligne du vecteur de permutation matérialise simplement l'indice de l'élément du vecteur, pour la clarté de certaines questions d'implémentation discutées plus bas. La matrice de permutation **P** correspond à la permutation produisant la seconde ligne :

$$\mathbf{P} = \begin{matrix} 0 & 1 & 2 \\ 0 & 2 & 1 \end{matrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix} \qquad (\mathbf{61})$$

Pour appliquer une matrice de permutation **P** à un vecteur colonne, on pré-multiplie ce vecteur par **P**. D'une manière plus générale, l'opération « appliquer une permutation à un ensemble ordonné » s'appelle une composition. Une permutation est un ensemble ordonné. Comme le produit matriciel, elle n'est pas commutative. L'implémentation contient des méthodes pour composer les permutations – générées sous forme de vecteurs – entre elles ou pour les appliquer aux différents conteneurs utilisés dans le programme (principalement `vec`, `mat` pour *Armadillo* et `QList` pour Qt). Pour appliquer une permutation $\mathbf{P}_i$, on commence par créer un conteneur vide de taille *n*. On copie ensuite chaque élément du conteneur à permuter dans la case dont l'indice est stocké dans les éléments de $\mathbf{P}_i$. Des exemples de permutations sont disponibles Table 1.

La forme des vecteurs de permutation $\mathbf{P}_n$ et le résiduel S^2 sont maintenant définis. Avant de continuer à détailler le fonctionnement de l'algorithme implémenté dans ce module, un exemple simple est proposé. On souhaite superposer deux « molécules » formées de 4 points dont les coordonnées sont dans la dernière colonne de la Table 1. Ces

deux pseudo-molécules ont un atome à l'origine et un atome au bout de chaque vecteur de base du référentiel choisi. La seule et unique différence est une permutation dans l'ordre des atomes : l'atome 2 de la molécule 1 a les coordonnées de l'atome 3 de la molécule 2 et *vice versa*.

Table 1. Deux nuages de points superposables. Pour simplifier l'exemple, on choisira **R** = δ_{ij} comme matrice de rotation optimale entre les pseudo-molécules 1 et 2. Il y a 24 permutations à évaluer, l'indice indique la position de la permutation générée par l'algorithme de Heap.

	Ordre des points	Coordonnées, Å (x, y, z)
Pseudo-molécule 1	Atome 0	(0, 0, 0)
	Atome 1	(**1**, **0**, 0)
	Atome 2	(**0**, **1**, 0)
	Atome 3	(0, 0, 1)
Pseudo-molécule 2	Atome 0	(0, 0, 0)
	Atome 1	(**0**, **1**, 0)
	Atome 2	(**1**, **0**, 0)
	Atome 3	(0, 0, 1)

Quelques comparaisons évaluées		
Permutation utilisée		S, Å^2 (*RMSD*, Å)
P$_0$ (*identité*)	0 1 2 3	4 (*1*)
P$_2$	0 2 1 3	**0** (***0***)
P$_{10}$	0 1 3 2	6 (*1.23*)
P$_{17}$	3 2 1 0	2 (*0.71*)

Quatre permutations notables parmi les 4! = 24 permutations évaluées par le module de comparaison sont rapportées Table 1. La meilleure est bien entendu **P**$_2$, puisqu'elle aboutit à la superposition parfaite de nos nuages de points.

Il apparaît clairement que l'appariement des points des deux nuages à comparer affecte fortement le résiduel S^2 et le *RMSD* des positions des points. Dans cet exemple, la méthode d'Umeyama (chapitre précédent) n'a pas été appliquée. Il a été considéré que la

matrice de rotation optimale **R** était l'identité, afin de rendre plus simple la comparaison des permutations.

Dans l'outil de comparaison de polarisabilités de *MoProViewer*, la méthode d'Umeyama – qui détermine la rotation optimale **R** – est appliquée après la génération des $n!$ combinaisons d'atomes. Les *RMSD* effectivement comparés dans le programme dépendent donc de l'application de la méthode d'Umeyama à chaque molécule, dont l'ordre des n atomes a été permuté par chacune des $n!$ combinaisons possibles.

L'algorithme génère donc les $n!$ permutations possibles pour les n atomes de la molécule « source » et les stocke dans une matrice **M** de $n * (n!)$ éléments. Les vecteurs de permutation $\mathbf{P}_i$ sont les colonnes de **M**. Pour chaque colonne $\mathbf{P}_i$, l'algorithme permute les colonnes de la matrice **B**, qui contiennent les coordonnées des atomes de la molécule « cible ». La nouvelle matrice est $\mathbf{B}_i$:

$$\mathbf{B}_i = \mathbf{P}_{\boldsymbol{i}} \circ \mathbf{B} = \begin{bmatrix} \mathbf{P}_{\boldsymbol{i}} \circ \mathbf{B}_{\boldsymbol{x}} \\ \mathbf{P}_{\boldsymbol{i}} \circ \mathbf{B}_{\boldsymbol{y}} \\ \mathbf{P}_{\boldsymbol{i}} \circ \mathbf{B}_{\boldsymbol{z}} \end{bmatrix} \qquad (\mathbf{62})$$

Où '∘' symbolise l'opération de composition et $\mathbf{B}_x$, $\mathbf{B}_y$, $\mathbf{B}_z$ sont les lignes de la matrice **B**. Pour chaque matrice $\mathbf{B}_i$, la matrice de rotation optimale $\mathbf{R}_i$ est déterminée afin de calculer le résiduel S^2_i. La permutation $\mathbf{P}_i$ qui minimise le résiduel S^2_i est $\mathbf{P}_{best}$. $\mathbf{P}_{best}$ correspond à la matrice de rotation $\mathbf{R}_{best}$.

Le test de la méthode n'a pas fait apparaître le calcul – très rapide – des matrices $\mathbf{R}_i$ comme un élément limitant sur les molécules testées ($n_{atomes} \leq 17$)[50]. Par contre la fonction factorielle est renommée pour sa croissance explosive à mesure que n augmente, et *Armadillo* utilise une structure de donnée de type `unsigned int` (nommée `uword`) pour stocker les index des vecteurs et matrices. Pour une molécule de 9 atomes, la méthode force brute nécessiterait le stockage d'une matrice à 9! = 362880 colonnes. Ce nombre est supérieur à 65535, la limite pour un entier non-signé de 16 bits. La méthode de type force brute n'est donc pas applicable aux molécules de plus de 8 atomes (avec 8! = 40320 permutations à évaluer).

Il est possible de contourner cette limitation. L'application envisagée concerne des molécules organiques contenant plus de 8 atomes mais en général, la majorité sont des atomes d'hydrogène. L'assomption suivante a été faite : la géométrie de la molécule est

majoritairement déterminée par la position des atomes lourds. On considère un fragment de moins de 9 atomes lourds (strictement). On part d'un vecteur de taille *n*, qui contient les indices des atomes dans le fichier cible. Ce vecteur, en termes de permutations, est la permutation identité ($\mathbf{P}_0 \circ \mathbf{P} = \mathbf{P}$).

$$\mathbf{P}_0 = \begin{matrix} 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & \cdots & n-1 \\ \mathbf{0} & 1 & 2 & 3 & \mathbf{4} & 5 & 6 & \mathbf{7} & 8 & \cdots & n-1 \end{matrix} \qquad (\mathbf{63})$$

Pour cet exemple, les indices des atomes lourds sont en gras. Pour une molécule à *k* atomes d'hydrogène, le vecteur $\mathbf{P}_0$ est contracté en retirant les indices de ces atomes en $\mathbf{P}_{heavy,0}$ de taille *n* - *k*.

$$\mathbf{P}_{heavy,0} = \begin{matrix} 0 & 1 & 2 & \cdots & n-k-1 \\ \mathbf{0} & \mathbf{4} & \mathbf{7} & \cdots & \boldsymbol{l} \end{matrix} \qquad (\mathbf{64})$$

Où *l* est l'indice du dernier atome lourd de la molécule. Les éléments d'une permutation $\mathbf{P}_i$ de taille *n* sont tous inférieurs à *n*. Par contre une permutation contractée $\mathbf{P}_{heavy,i}$ de taille *n*-*k* peut contenir des éléments dont la valeur est supérieure ou égale à *n*-*k*. Ceci peut entrainer un appel hors-tableau lors de l'application d'une permutation, entraînant la terminaison prématurée du programme. Il n'est donc pas recommandable d'appliquer directement $\mathbf{P}_{heavy,i}$ à un conteneur de *n*-*k* éléments, et illogique d'appliquer une permutation de *n*-*k* éléments à un conteneur de *n* éléments.

Il est heureusement possible de contourner cette difficulté pour trouver $\mathbf{P}_{heavy,best}$, le vecteur contenant la permutation d'indices optimale pour les atomes lourds. Pour ce faire, l'algorithme de Heap est utilisé à nouveau pour générer la matrice $\mathbf{M}_{heavy}$, qui contient – en colonnes – toutes les (*n*-*k*)! permutations de *n*-*k* éléments. On les note de $\mathbf{Q}_0$ à $\mathbf{Q}_{(n-k)!}$:

$$\mathbf{M}_{heavy} = [\mathbf{Q}_0 \quad \mathbf{Q}_1 \quad \cdots \quad \mathbf{Q}_{(n-k)!}] \qquad (\mathbf{65})$$

Chaque colonne de $\mathbf{M}_{heavy}$ est utilisée pour générer une matrice de coordonnées relatives $\mathbf{B}_{heavy,i}$ qui est utilisée par la méthode d'Umeyama. La matrice $\mathbf{B}_{heavy,i}$ ne contient que les coordonnées des atomes lourds. La permutation $\mathbf{Q}_{best}$ est celle qui minimise l'écart au carré $S_{heavy,best}$ (calculé pour les atomes non-H seulement). $\mathbf{Q}_{best}$ est associée à la matrice de rotation $\mathbf{R}_{heavy,best}$. La composition $\mathbf{Q}_{best} \circ \mathbf{P}_{heavy,0}$ donne $\mathbf{P}_{heavy,best}$, le vecteur contenant la permutation d'indices optimale pour les atomes lourds de la molécule « source » vers la molécule cible.

On cherche à trouver la correspondance entre les *n* atomes des deux molécules. À cette étape, *MoProViewer* a trouvé la correspondance entre les *n-k* atomes non-hydrogène. On partira donc de $\mathbf{P}_{heavy,best}$ pour retrouver la permutation complète $\mathbf{P}_{best}$, à l'opposé de la contraction présentée aux équations (63) et (64). Il n'est pas nécessaire ici d'explorer les *k*! permutations possibles pour tous les *k* atomes d'hydrogène. Ces derniers n'ont qu'une liaison covalente, les seules permutations à évaluer sont les permutations parmi les atomes d'hydrogène portés par le même atome lourd. Si la permutation suivante est associée au minimum de S_{heavy} :

$$\mathbf{P}_{heavy,best} = \mathbf{Q}_{best} \circ \mathbf{P}_{heavy,0} = \begin{matrix} 0 & 1 & 2 & \cdots & n-k-1 \\ \mathbf{0} & \mathbf{7} & \mathbf{4} & \cdots & \boldsymbol{l} \end{matrix} \qquad (\mathbf{66})$$

Alors les candidats évalués pour trouver $\mathbf{P}_{best}$ auront cette forme :

$$\mathbf{P} = \begin{matrix} 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & \cdots & n-1 \\ \mathbf{0} & a & a & a & \mathbf{7} & \boldsymbol{b} & \boldsymbol{b} & \mathbf{4} & c & c & \cdots & n-1 \end{matrix} \qquad (\mathbf{67})$$

où les éléments représentés par les mêmes lettres désignent des atomes d'hydrogène liés au même atome lourd. À cette étape, l'outil utilise des permutations « nichées » (*nested*). Par exemple, dans la molécule « cible », l'atome avec l'indice 7 est lié aux hydrogènes d'indices 8 et 9. Dans la molécule « source », l'atome d'indice 4 est lié aux hydrogènes d'indices 5 et 6. Les équations (66) et (67) disent que l'atome d'indice 4 de la « source » correspond à l'atome d'indice 7 de la « cible ». Les positions 5 et 6 de la molécule « source » doivent correspondre aux positions 8 et 9 de la « cible ». Ces positions d'indice 8 et 9 sont notées *c* et seront évaluées indépendamment par l'algorithme.

Il y a donc deux possibilités pour compléter **P** aux positions 5 et 6 :

$$\mathbf{P}_{\boldsymbol{p}} = \begin{matrix} 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & \cdots & n-1 \\ 0 & a & a & a & 7 & \mathbf{8} & \mathbf{9} & 4 & c & c & \cdots & n-1 \end{matrix} \qquad (\mathbf{68})$$

$$\mathbf{P}_{\boldsymbol{q}} = \begin{matrix} 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & \cdots & n-1 \\ 0 & a & a & a & 7 & \mathbf{9} & \mathbf{8} & 4 & c & c & \cdots & n-1 \end{matrix} \qquad (\mathbf{69})$$

Ces deux permutations sont incomplètes, il manque les positions notées *a* et *c*. On utilisera un algorithme récursif pour obtenir toutes les permutations possibles pour les positions manquantes, sans perturber celle des atomes lourds. Si m_i est le nombre d'atomes d'hydrogène portés par l'atome lourd *i*, le nombre de permutations à évaluer à cette étape est :

$$p = \prod_{i=0}^{n-k} m_i! \qquad (70)$$

Pour la présente application à des molécules organiques, m_i est rarement supérieur 3, et est souvent égal à 2. Par conséquent, $p \ll k!$ dans les cas rencontrés ici. On suppose ici que $\mathbf{R}_{best} = \mathbf{R}_{heavy,best}$, de manière à éviter que les variations des positions des hydrogènes ne perturbent l'alignement des atomes lourds (ce choix est relatif aux options proposées à l'utilisateur, voir partie 1.5).

$\mathbf{P}_{best}$ sera alors la permutation complète qui minimise le résiduel au carré S^2 en utilisant la rotation $\mathbf{R}_{heavy,best}$ sur les coordonnées de chaque atome de la molécule « source ». $\mathbf{R}_{best}$ est retenue pour le transfert ou la comparaison de propriétés orientées.

Une optimisation évidente (non-implémentée) consisterait à reconnaître les types chimiques au-delà de la distinction hydrogène/non-hydrogène, pour n'autoriser que des permutations entre atomes de même numéro atomique *Z*. Les atomes lourds à un seul voisin (exemple : halogènes, oxygènes des groupes cétone ou aldéhyde…) pourraient être reconnus et traités de la même manière que les atomes d'hydrogène. Ces améliorations n'ont pas été implémentées pour conserver des fonctions très générales et réutilisables dans la partie du code dédié à la gestion des permutations.

Cette méthode pourrait éventuellement être étendue à des nuages de points de taille arbitraire, pour peu que l'on ajoute une méthode pour sélectionner des points clés de la molécule et de nouvelles méthodes pour compléter ce vecteur de points-clés. Cette méthode est bien entendu pensée pour l'utilisation sur de petites molécules rigides, des promolecules (par exemple : chaines latérales d'acides aminés) ou des synthons de conformation fixe. L'implémentation dans *MoProViewer* utilise la méthode présentée pour le transfert ou les comparaisons de polarisabilités (polarisabilités dans un dimère contre polarisabilités dans un monomère, polarisabilités théoriques contre polarisabilités transférables, etc.). Cette méthode permet dans de nombreux cas de s'abstraire de la définition d'axes locaux atomiques et de types atomiques, dans le but d'évaluer la transférabilité de propriétés orientées.

1.5 - Molécules Fantômes

À cette étape, il est donc possible d'obtenir le bon vecteur de permutations et la matrice de rotation optimale correspondante. Cependant, inspecter visuellement l'orientation de deux molécules par rapport à leurs référentiels ne permet pas de vérifier aisément si la matrice de rotation est correcte. Il faudrait pouvoir visualiser l'alignement en 3D. De plus, la principale limite de la méthode précédente est sa rigidité : on suppose que les molécules sont superposables, dépourvues de degrés de libertés conformationnels. Pour certains systèmes, présupposer de la rigidité de la molécule est une erreur. Pour d'autres, c'est une assomption utile. Il est donc nécessaire de pouvoir vérifier dans l'interface le résultat de la rotation, de préférence avant l'écriture d'un fichier transféré ou d'un long rapport de comparaison détaillé.

Pour obtenir ce contrôle du logiciel, un système de visualisation impliquant une molécule « fantôme » a été implémenté. Le fichier « source » est chargé, la matrice de rotation et la permutation optimale ont été déterminées par les méthodes précédentes. *MoProViewer* affiche à ce moment une image de la molécule « source » superposée à la molécule « cible », comme montré Figure 10, page 76. Le label des atomes de la molécule fantôme (la « source ») commencent tous par '___ghost_' suivi du nom original de l'atome (la répétition du tiret bas permet de laisser la place aux labels des atomes cibles). Les liaisons covalentes de la molécule « fantôme » sont reproduites depuis la molécule « source ». Le résultat peut apparaître contre-intuitif mais représente en réalité l'action de la méthode basée sur les permutations [partie 1.4, montré par exemple sur le panneau (b) de la Figure 10].

Les indicateurs numériques proposés sont basés sur les distances entre la géométrie « cible » et la géométrie « source » après rotation. Le *RMSD* calculé sur les positions atomiques de atomes lourds (non-hydrogène) permet de quantifier l'accord entre les géométries du squelette de la molécule. Le *RMSD* calculé sur toutes les positions atomiques informe sur la qualité globale de la superposition. Le ratio lourd/tous (des deux *RMSD* précédents) permet de diagnostiquer facilement la rotation d'un ou plusieurs atomes d'hydrogène d'un groupement terminal (méthyl, amine, alcool…). Dans le cas d'une molécule parfaitement rigide, le ratio lourd/tous est 1. Le ratio s'approche de zéro à mesure que le *RMSD* des positions de tous les atomes décroit du fait des différences de positions des atomes d'hydrogène.

Si l'utilisateur tente de transférer des polarisabilités depuis un système de conformation différente, la méthode ci-dessus peut générer des résultats plus ou moins aberrants (en fonction de si la flexibilité touche le squelette carboné ou uniquement la position des hydrogènes). La possibilité de représenter des « molécules fantômes » permet à l'utilisateur de visualiser la meilleure permutation des atomes de la molécule source, superposée à la molécule cible, affichée dans l'interface 3D. En fonction de quelle flexibilité doit être prise en compte, l'utilisateur dispose de 4 options :

• Utiliser la matrice de rotation $\mathbf{R}_{best}$ pour tous les atomes du système (conseillé si la superposition est bonne et l'écart au carré pour tous les atomes S_{all} < 0.03 Å).

• Utiliser la matrice de rotation $\mathbf{R}_{best}$ pour tous les atomes non-H, et la matrice de passage associée aux axes locaux des atomes d'hydrogène, tels que définis dans les fichiers « cible » et « source » (conseillé si le résiduel pour les atomes lourds, $S_{backbone}$ < 0.03 Å mais que $S_{all}/S_{backbone}$ > 4).

• Utiliser les axes locaux pour tous les atomes, tels que définis dans les fichiers « cible » et « source ». Ces systèmes d'axes sont destinés à l'orientation de l'expansion multipolaire. Cette méthode ne vérifie pas que ces axes soient comparables. Simplement, la polarisabilité en provenance du fichier « source » est exprimée dans son référentiel atomique local, par application de la matrice $\mathbf{L}_{source}$ ($\boldsymbol{\alpha}_{local} = \mathbf{L}_{source} \cdot \boldsymbol{\alpha}_{source} \cdot \mathbf{L}_{source}^{t}$). Comme on suppose que les axes locaux sont les mêmes, on exprime ensuite cette polarisabilité $\boldsymbol{\alpha}_{local}$ dans le référentiel de la molécule « cible » à l'aide de la matrice $\mathbf{L}_{cible}$ et de la relation suivante : $\boldsymbol{\alpha}_{cible} = \mathbf{L}_{cible}^{t} \cdot \boldsymbol{\alpha}_{local} \cdot \mathbf{L}_{cible}$. Dans l'implémentation, ces deux opérations séquentielles passent par l'utilisation d'une matrice $\mathbf{R}_{local} = \mathbf{L}_{cible}^{t} \cdot \mathbf{L}_{source}$. $\mathbf{R}_{local}$ est calculée pour chaque atome de la molécule en utilisant la permutation qui minimise S_{all}.

• Forcer $\mathbf{P}_{best} = \mathbf{P}_0$ (pour la maintenance, le débogage ou si l'utilisateur peut garantir l'ordre des atomes dans le fichier).

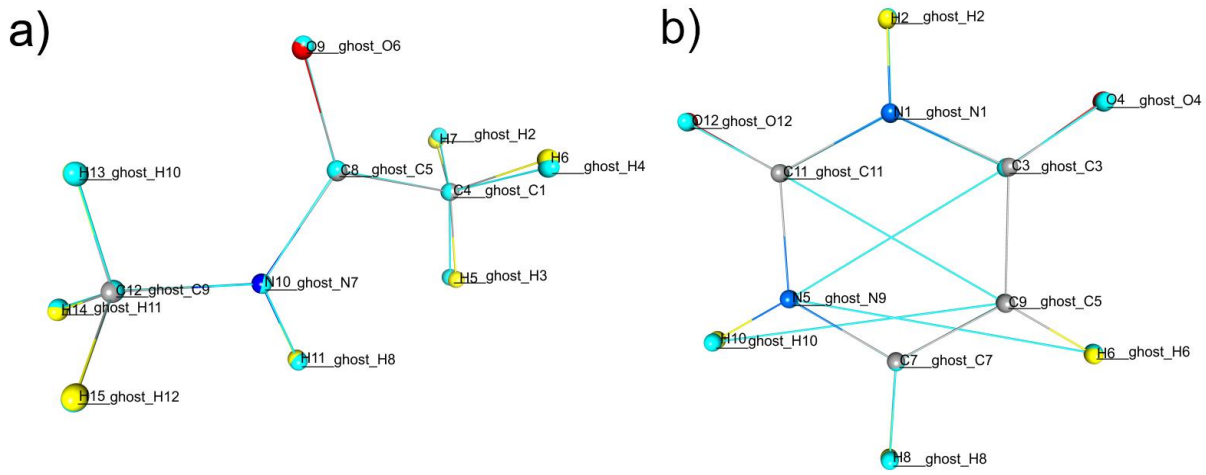


Figure 10. Visualisation d'alignements structuraux avec *MoProViewer*. (a) : Superposition des géométries du N-méthyl-acétamide en interaction avec l'eau (H_2O omis, H, jaune ; C, gris ; N, bleu ; O, rouge) et du N-méthyl-acétamide dans le vide (molécule cyan). (b) : Superposition des géométries de l'uracile en interaction π-π avec lui-même (partenaire omis, H, jaune ; C, gris ; N, bleu ; O, rouge) et de l'uracile dans le vide (molécule cyan). Les liaisons covalentes distordues de la molécule cyan montrent une permutation de l'ordre des atomes entre les deux fichiers, « cible » (dimère) et « source » (uracile isolé).

Les deux exemples de la Figure 10 génèrent des suggestions différentes de la part du module. Pour l'alignement du N-méthyl-acétamide, panneau (a), le *RMSD* sur toutes les positions des atomes non-H est 0.0099 Å, indiquant une très faible variation de conformation du squelette. Par contre le *RMSD* calculé sur tous les atomes est 0.04485 Å, ce qui donne un ratio (lourd)/(tous) de 0.222. Le module considèrera dans ce cas que l'accord entre la position des hydrogènes est insuffisant. Il sera suggéré à l'utilisateur d'utiliser l'option 2), soit d'utiliser les axes locaux pour les atomes d'hydrogène et la matrice de rotation optimale pour les autres atomes.

Sur le panneau (b), l'alignement de la molécule d'uracile présente deux anomalies. Tout d'abord l'ordre des atomes est différent entre le fichier contenant le monomère (molécule cyan) et le dimère (H, jaune ; C, gris ; N, bleu ; O, rouge). Cette permutation touche les atomes C5 et N9 du monomère‡, qui sont respectivement en position 9 et 5 du fichier « cible ». Les liaisons colorées en cyan montrent cette permutation. Une

‡ *: Pour les structures du S66, les chiffres désignent ici le numéro de ligne dans le fichier *.xyz accessible sur begdb.com.*

vérification manuelle de la comparaison ou du transfert par l'utilisateur est conseillée en cas de présence de ce type d'anomalie.

Enfin, les valeurs de *RMSD* des positions atomiques sont trop importantes pour l'alignement structural panneau (b) de la Figure 10. Le *RMSD* sur les 8 atomes lourds est de 0.0436 Å. Calculé sur tous les atomes, le *RMSD* passe à 0.0521 Å, soit une valeur comparable (ratio : 0.837). Ces deux *RMSD* sont trop faibles pour être causés par une erreur dans la détermination de la meilleure permutation. Dans le cas d'une erreur de permutation, le *RMSD* est au moins de 0.5 Å (du fait de la longueur des liaisons chimiques). Le seuil de tolérance pour un alignement décent a été choisi à 0.03 Å. Pour le dimère d'uraciles en interaction π-π, les deux RMSD sont supérieurs à 0.03 Å car la géométrie de chaque partenaire dévie du cycle aromatique plan présenté par l'uracile isolé. Dans ce cas, le module suggère à l'utilisateur d'utiliser les référentiels atomiques locaux pour la rotation de chaque polarisabilité atomique. Cette option est également recommandable en cas de permutations des positions atomiques entre les fichiers source et cible, si l'application de la matrice de rotation pourrait entraîner des résultats inattendus, faute de tests extensifs.

Les méthodes utilisant les axes locaux supposent que ces axes sont les mêmes entre les deux systèmes. Le choix de ces axes n'est pas unique et, dans le cadre des types atomiques ELMAM2, ces choix maximisent la symétrie locale de l'expansion multipolaire. L'utilisation d'axes locaux pour exprimer les polarisabilités est discuté en détails dans les chapitres 3 et 3-bis.

1.6 - Rapport de comparaison entre tenseurs de polarisabilité atomique

L'application de la méthode précédente permet de réécrire un fichier molécule au format *MoPro* avec les polarisabilités du fichier source stockées en lieu et place des paramètres de déplacement $\mathbf{U}_{ij}$ anisotropes du fichier cible (champ '`UANI`'). L'autre possibilité offerte à l'utilisateur est de comparer les polarisabilités déjà stockées dans les $\mathbf{U}_{ij}$ du fichier cible avec celles du fichier source. Le rapport de comparaison est affiché dans la fenêtre de l'outil de polarisation et dans le fichier `*.log` de *MoProViewer*.

Ce rapport permet l'examen des éléments suivants (dans l'ordre de sortie) :

• Le tenseur de polarisabilité de l'atome du fichier « cible » dans le référentiel moléculaire.

• Les trois invariants rotationnels du tenseur de l'atome « cible » : déterminant, trace et $0.5 \cdot (\text{trace}(\boldsymbol{\alpha}_{cible})^2 - \text{trace}(\boldsymbol{\alpha}_{cible}^2))$.

• Les trois valeurs propres suivi de la matrice contenant les trois vecteurs propres de la polarisabilité de l'atome « cible ».

Cette liste est répétée pour l'atome « source », mais en exprimant ces propriétés dans le référentiel moléculaire du fichier « cible ». Le module affiche ensuite les comparaisons suivantes :

• La matrice $\boldsymbol{\alpha}_{\text{cible}} - \boldsymbol{\alpha}_{\text{source}}$, exprimées dans le même référentiel moléculaire du fichier « cible ».

• La différence entre les valeurs propres de ces deux polarisabilités.

• La différence entre les matrices contenant les vecteurs propres de $\boldsymbol{\alpha}_{\text{cible}}$ et les vecteurs propres de $\boldsymbol{\alpha}_{\text{source}}$.

• La matrice de rotation qui a été utilisée pour la comparaison (si la comparaison utilise les axes locaux des atomes, cette matrice peut varier entre les atomes d'une même molécule).

• Et enfin la distance tensorielle entre $\boldsymbol{\alpha}_{\text{cible}}$ et $\boldsymbol{\alpha}_{\text{source}}$, calculée par la formule de Jelsch *et al.*, 1998[71].

Ces rapports de comparaison ont été compilés pour l'intégralité du jeu de données S66 (939 atomes distincts), afin de réaliser les statistiques présentées aux parties 1.7 et 2.4, ainsi que dans l'article Leduc *et al.*, 2019[52]. Lorsque des valeurs angulaires sont reportées [équation (A1) de l'annexe], celles-ci ont été calculées à partir des vecteurs propres documentés dans ces rapports de comparaison.

1.7 - « Polarization of electron density databases of transferable multipolar atoms. » *Partie 1 : Polarisabilités*

Cette partie a été intégralement extraite de l'article « Polarization of Electron Density Databases of Transferable Multipolar Atoms », parue le 11 juillet 2019 dans The Journal of Physical Chemistry A, par Théo Leduc, Emmanuel Aubert, Enrique Espinosa, Christian Jelsch, Cristian Iordache et Benoît Guillot. Cet article est parfaitement intégré au présent travail et sa publication représente un achèvement important pour le présent travail de thèse. Afin de faciliter la lecture, les éléments rapportés ont été traduits et réorganisés pour s'ajuster au plan choisi ici. Aucun élément n'a été ajouté afin que toutes les informations présentées dans cette partie restent telles que revues par les pairs.

1.7.a - Base de données de référence

La base de données S66 et ses extensions[50,51] ont été choisies pour tester et valider la procédure de polarisation développée ici. Cet ensemble est composé des structures théoriques de 14 petites molécules organiques neutres ($n_{atomes} \leq 17$) qui forment 66 dimères, dont les polarisabilités atomiques ont été calculées selon la méthode présentée page 44 et dans la sous-partie suivante (pour tous les dimères mais également pour les molécules isolées). Les auteurs originaux ont sélectionné les dimères pour imiter les interactions pertinentes pour les simulations biomoléculaires. Les énergies et les géométries moléculaires des monomères et des dimères (à distance d'équilibre ou dans une géométrie de transition. Exemple : empilement en T de cycles aromatiques) ont été optimisées par des calculs *ab initio* de niveau théorique MP2/cc-pVTZ CP[50]. L'extension S66x8 de cette base de données utilise les géométries S66 pour imiter une courbe de dissociation discrète. Tout en conservant les géométries moléculaires et l'orientation relative, la distance entre les centres de masses est multipliée par un facteur allant de 0.9 à 2, pour donner huit complexes par dimère. L'énergie d'interaction a été calculée au même niveau théorique à toutes les distances.

Pour pouvoir modéliser la densité électronique de toutes les molécules du jeu S66 avec la bibliothèque ELMAM2, quatre entrées (types atomiques) ont dû être ajoutées. L'hydrogène et le carbone de l'éthyne ont été ajoutés et leur densité multipolaire a été contrainte à une symétrie cylindrique. Les densités électroniques ont été obtenues à

partir d'un calcul théorique[82] réalisé avec le logiciel CRYSTAL-09[83]. L'hydrogène et le carbone de l'éthène ont été ajoutés par extension des types d'hybridation comparables existants (H101 et C301). Au total, la reconstruction de la densité électronique a utilisé 34 types atomiques ELMAM2 différents pour couvrir les atomes présents dans la base de données S66. La liste complète est disponible en annexe, Table A1, et le chapitre 3-bis est dédié aux types atomiques ELMAM2.

1.7.b - Calcul des polarisabilités moléculaires et atomiques

Les calculs ont été effectués sur les systèmes monomères et sur les systèmes dimères de l'ensemble de données S66[50,51] à leurs géométries d'équilibre. Le partage de la polarisabilité moléculaire en contributions atomiques a de nombreux points communs avec la méthode développée par Krawczuk, Pérez et Macchi[49] pour le logiciel *PolaBer*. Pour chaque molécule ou complexe, des calculs DFT ont été effectués au niveau théorique B3LYP/6-311++G(3df, 3pd) à l'aide du logiciel *Gaussian09*[84]. Des champs électriques uniformes sont appliqués dans les directions X, Y et Z avec des amplitudes croissantes allant de -0.005 a.u. à 0.005 a.u. (± 0.0014 e·Å^{-2}), avec un pas de 0.001 a.u. (2.8·10^{-4} e·Å^{-2}). La méthode *PolaBer* utilise généralement deux valeurs de champ électrique par direction de l'espace, et un set de fonctions de base différent (*basis set*, voir partie 1.8). Les densités électroniques résultantes des calculs à chaque valeur de champ électrique sont ensuite analysées avec le logiciel *AIMAll*[85]. Il en résulte une partition de Bader analytique qui permet l'intégration de la densité de charge dipolaire dans les bassins atomiques topologiques. L'application de l'équation (35) permet d'obtenir les moments dipolaires intégrés des bassins atomiques. Une isosurface à 0.002 e·a.u.$^{-3}$ ferme les ρ-bassins ouverts, qui, autrement, s'étendent formellement à l'infini. Pour un bassin atomique neutre, le moment dipolaire $\boldsymbol{\mu}_\Omega$ est indépendant de l'origine.

La sommation de toutes les polarisabilités atomiques $\boldsymbol{\alpha}_i$ conduit à des polarisabilités moléculaires $\boldsymbol{\alpha}_{mol}$ en accord avec les valeurs obtenues directement à partir du calcul DFT. Les déviations absolues moyennes et maximales calculées sur toutes les composantes moléculaires $\boldsymbol{\alpha}_{mol,i,j}$ sont respectivement 0.011 Å^3 et 0.015 Å^3.

Neuf dimères[§] sur les 66 du jeu de données S66 ont posé des problèmes de calcul et ont été exclus des résultats (par exemple, certaines polarisabilités atomiques n'étaient pas définies positives et/ou les moments dipolaires ne variaient pas linéairement par rapport au champ électrique appliqué).

Les volumes, charges et polarisabilités atomiques définis topologiquement ont été calculés pour les 14 monomères et les 66 dimères du jeu de données S66. 139 et 1103 polarisabilités atomiques, provenant respectivement des densités électroniques des monomères et des dimères, ont été produites (le détail des polarisabilités est disponible dans le matériel supplémentaire de l'article[52]). La comparaison des charges topologiques entre les monomères et les dimères permet de définir les transferts de charges topologiques, qui sont également décrits dans la partie 2.4.b.

Les variations des propriétés topologiques lors de la dimérisation ont été tabulées afin de comparer chaque atome de chaque dimère à son homologue dans le monomère. L'appariement des atomes entre monomères et dimères a été réalisé à l'aide du code décrit partie 1.6. Les critères de comparaison comprennent les polarisabilités isotropes et anisotropes orientées (sous forme de tenseur) ainsi que la charge et le volume du ρ-bassin atomique correspondant. Pour évaluer des variations plus fines des polarisabilités atomiques topologiques, l'ellipticité de polarisabilité ε, issue de l'équation (A2) et le rapport des valeurs propres transversales sur longitudinales η, issu de l'équation (A3) sont examinés[33] (équation en annexe, page 291).

1.7.c - Polarisabilités moléculaires théoriques

Les polarisabilités moléculaires ont été calculées, ainsi que la partition de cette propriété en contributions atomiques. La Table 2 contient les valeurs propres de la polarisabilité moléculaire diagonalisée, calculées à partir de la somme des polarisabilités atomiques topologiques. Le résultat est en bon accord avec les valeurs expérimentales de la littérature. Gussoni, dans un article de 1998[42], compile des polarisabilités moléculaires

[§] *: Les 9 dimères exclus sont : Pyridine⋯Pyridine (π-π, 25) ; Pyridine⋯Uracile (π-π, 29) ; Benzène⋯Cyclopentane (39) ; Uracile⋯Cyclopentane (42) ; N-méthyl-acétamide⋯Pentane (46) ; Benzène⋯Pyridine (TS, 49) ; Benzène⋯Acide acétique (52) ; Benzène⋯Acétamide (N–H-π, 53) et Benzène⋯Méthylamine (N–H-π, 56).*

électroniques mesurées en phase gazeuse. En comparant la polarisabilité moléculaire isotrope des 14 monomères à leurs valeurs expérimentales, on obtient un écart quadratique moyenne [*RMSD*, équation (A4), page 291] de 0.23 Å^3 et une erreur relative absolue moyenne [*MARE*, équation (A6), *idem.*] de 2.8 % seulement. Cette erreur se compare bien aux références historiques (3.5 % de Thole, 1981[46,47]) et aux travaux actuels sur le sujet (6 % de Dos Santos, Krawczuk & Macchi, 2015[86]). La régression linéaire entre les polarisabilités moléculaires isotropes expérimentales et théoriques donne une pente de 0.995, avec un coefficient de détermination R^2 de 0.994.

La plus grande différence absolue est obtenue pour le néopentane et le cyclopentane, pour lesquels les polarisabilités moléculaires sont sous-estimées respectivement de 0.55 Å^3 (-5.3%) et 0.34 Å^3 (-3.7%) par rapport aux valeurs expérimentales de Gussoni. Inversement, les polarisabilités théoriques entre éthène et éthyne concordent presque parfaitement, surestimant les valeurs expérimentales de respectivement 0.01 Å^3 (+0.2%) et 0.04 Å^3 (+1.1%). La polarisabilité théorique de l'acide acétique est également très proche de la valeur expérimentale, avec une sous-estimation de seulement 0.08 Å^3 (-0.8%).

Table 2. Polarisabilités théoriques moléculaires (diagonalisées et isotropes), calculées dans le vide pour les 14 molécules du jeu de données S66. Pour comparaison, les valeurs des polarisabilités isotropes expérimentales (Gussoni, 1998[42]) sont également présentées. * : *N*-méthyl-acétamide.

Molécule	α_{xx} (Å^3)	α_{yy} (Å^3)	α_{zz} (Å^3)	α_{iso} (Å^3)	α, Exp. (Å^3)	Erreur Rel. (%)
Acétamide	6.75	6.37	4.46	5.86	5.67	3.34
Acide Acétique	5.89	5.52	3.91	5.11	5.15	-0.81
Benzène	12.07	12.03	6.62	10.24	9.96	2.83
Cyclopentane	9.25	9.25	7.99	8.83	9.17	-3.73
Éthène	5.29	3.66	3.35	4.10	4.09	0.20
Éthyne	4.75	2.79	2.79	3.44	3.40	1.17
Méthylamine	4.19	3.66	3.55	3.80	3.97	-4.30
Méthanol	3.42	3.00	2.93	3.12	3.21	-2.88
Néopentane	9.70	9.69	9.69	9.69	10.24	-5.34
Pentane	11.45	9.19	8.56	9.73	9.88	-1.47
Peptide*	9.37	7.74	5.88	7.66	7.82	-2.03
Pyridine	11.48	10.82	6.03	9.44	9.17	2.96
Uracile	13.86	11.14	6.04	10.35	10.09	2.58
Eau	1.44	1.36	1.33	1.38	1.45	-5.17

1.7.d - Polarisabilités atomiques isotropes

Pour évaluer la transférabilité des polarisabilités atomiques, il est nécessaire d'évaluer si celles-ci ont des points communs quand leurs atomes partagent des caractéristiques chimiques communes. Le tout premier niveau de cette classification concerne évidemment la nature chimique des atomes disponibles dans le jeu de données

S66. Avant d'aborder l'anisotropie des polarisabilités, un certain nombre de remarques peuvent être tirées sur la base des polarisabilités scalaires α_{iso}.

La Figure 11 présente les valeurs moyennes de polarisabilité isotrope pour chacun des quatre éléments chimiques (H, C, N et O) présents dans le jeu de données S66. Les valeurs théoriques des atomes libres[87,88,89,90] sont données à titre de comparaison. Les atomes d'oxygène et d'azote sont les plus électronégatifs. Par conséquent, dans les molécules, les polarisabilités associées aux oxygènes et aux azotes augmentent, par rapport aux atomes libres, en raison de leur charge électronique accrue lorsque ces atomes sont intégrés dans un environnement moléculaire. Cette tendance est inversée pour les deux atomes relativement électropositifs (hydrogène, carbone). Il y a peu de variation des polarisabilités atomiques isotropes moyennes entre les calculs dans les systèmes monomères et dimères.

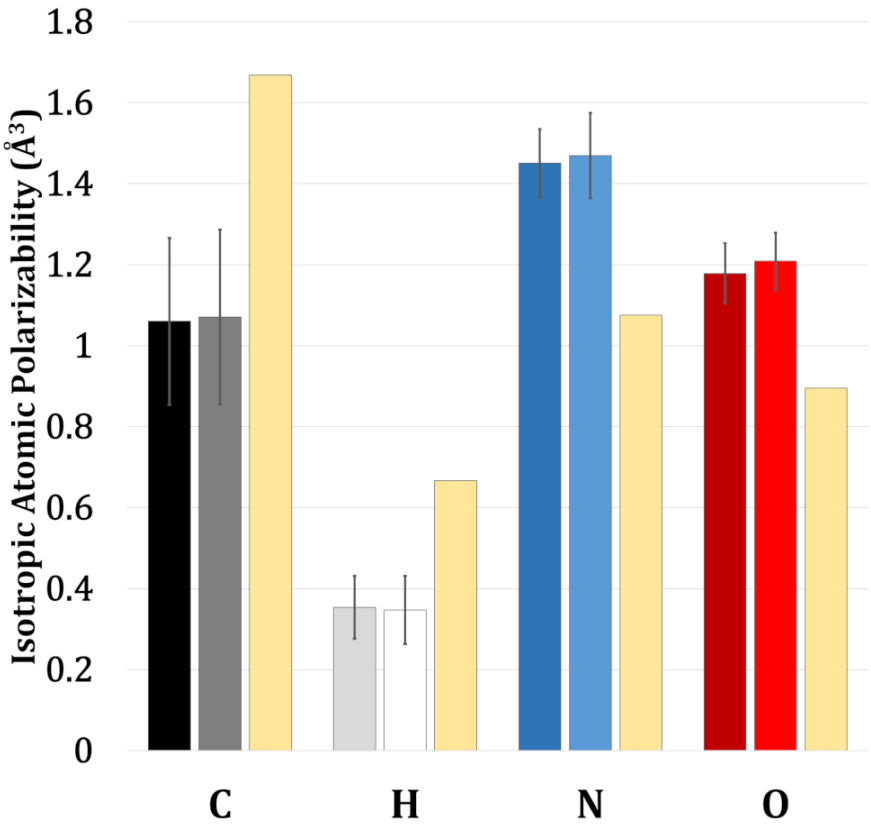


Figure 11. Valeurs moyennes des polarisabilités atomiques isotropes théoriques pour les quatre éléments chimiques des jeux de données S66 et S66x8, en Å³. Les barres d'erreur matérialisent le *SSD* de chaque groupe. Teintes plus sombres : molécules isolées, teintes plus claires : dimères. Jaune clair : valeurs théoriques pour un atome isolé.

Table 3. Maxima et minima des polarisabilités atomiques isotropes du S66, par éléments chimiques, ainsi que leurs charges topologiques. Pour information, les types ELMAM2 de ces atomes extrêmaux sont indiqués, avec les molécules du set S66 d'où ils proviennent.

Élement	α_{iso} (Å^3)	Type atomique ELMAM2	Molécule (partenaire)	Charge Topologique Atomique, Théorique (e)
H	0.1114	H, alcool	Acide acétique (Pentane)	0.659
	0.5821	H, aromatique	Benzène (Eau)	0.005
C	0.6514	C-ONN (sp^2)	Uracile (Uracile, π-π)	1.829
	1.526	C, éthyne (sp)	Éthyne (Ethyne, π-π)	-0.134
N	1.311	N, peptide	Uracile (Benzène, π-π)	-1.148
	1.697	N, aromatique	Pyridine (Pyridine, TS)	-1.087
O	1.090	O, H-carboxylique	Acide acétique (Pentane)	-1.129
	1.408	O, amide	Uracile (Uracile, BP)	-1.195

Afin de proposer une image légèrement plus détaillée des 4 échantillons de polarisabilités calculées, il est possible de chercher des tendances liées aux caractéristiques topologiques des bassins atomiques.

La Table 3 affiche les extrema des polarisabilités isotropes pour chaque élément, le nom complet des types d'atomes ELMAM2 qui leur correspondent, et la molécule à laquelle ils appartiennent. Les atomes H et C présentent les plus grandes variations de leurs polarisabilités. Pour ces deux espèces chimiques, une polarisabilité minimale est obtenue pour les atomes liés à un ou plusieurs atomes électronégatifs et la polarisabilité maximale est obtenue pour les atomes dans les molécules avec des liaisons π (éthyne et benzène). N et O ont une plus faible amplitude de variation quand on ne considère que leur représentation isotrope. Ici, les deux extrema (minimum et maximum) sont trouvés dans des atomes impliqués dans des liaisons π.

Pour les atomes d'oxygène et d'azote, les charges topologiques présentent une très faible amplitude de variation, de sorte que la charge seule peut ne pas expliquer la variation de polarisabilité au sein des atomes d'oxygène ou d'azote. Les polarisabilités isotropes minimales et maximales apparaissent lorsque ces atomes forment des liaisons conjuguées. On trouve les valeurs maximales lorsque les atomes d'oxygène ou d'azote sont les plus électronégatifs dans le groupement chimique conjugué.

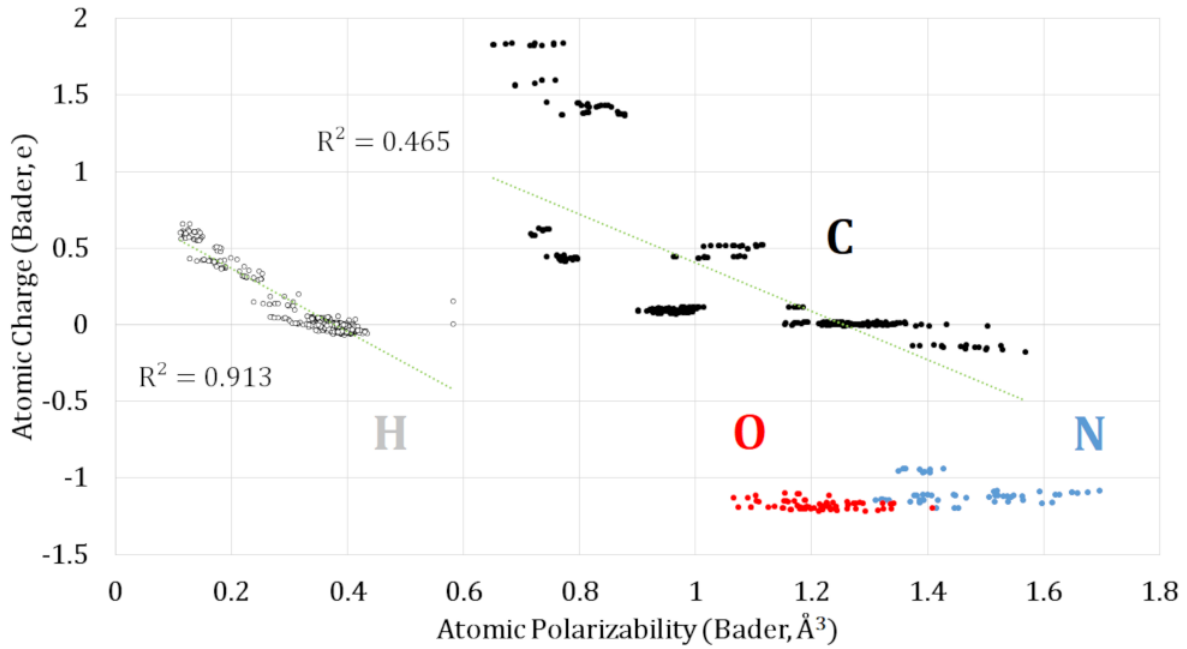


Figure 12. Corrélation entre charges atomiques topologiques et polarisabilités topologiques (provenant indifféremment des monomères et des dimères), classées par élément chimique. Carbone : noir ; hydrogène : blanc ; azote : bleu ; oxygène : rouge.

Une explication topologique de ces comportements peut être observée Figures 12 et 13. Il existe une corrélation inverse claire entre les charges topologiques des atomes d'hydrogène et leurs polarisabilités atomiques (R^2 = 0.913). Pour le carbone, la corrélation est plus faible mais une tendance apparaît (R^2 = 0.465), comme le montre la Figure 12. Les charges topologiques et les volumes topologiques sont intrinsèquement corrélés. La Figure 13 montre explicitement la corrélation entre les polarisabilités topologiques et les volumes atomiques, confirmant que les grands bassins atomiques contiennent plus d'électrons et sont de fait plus polarisables.

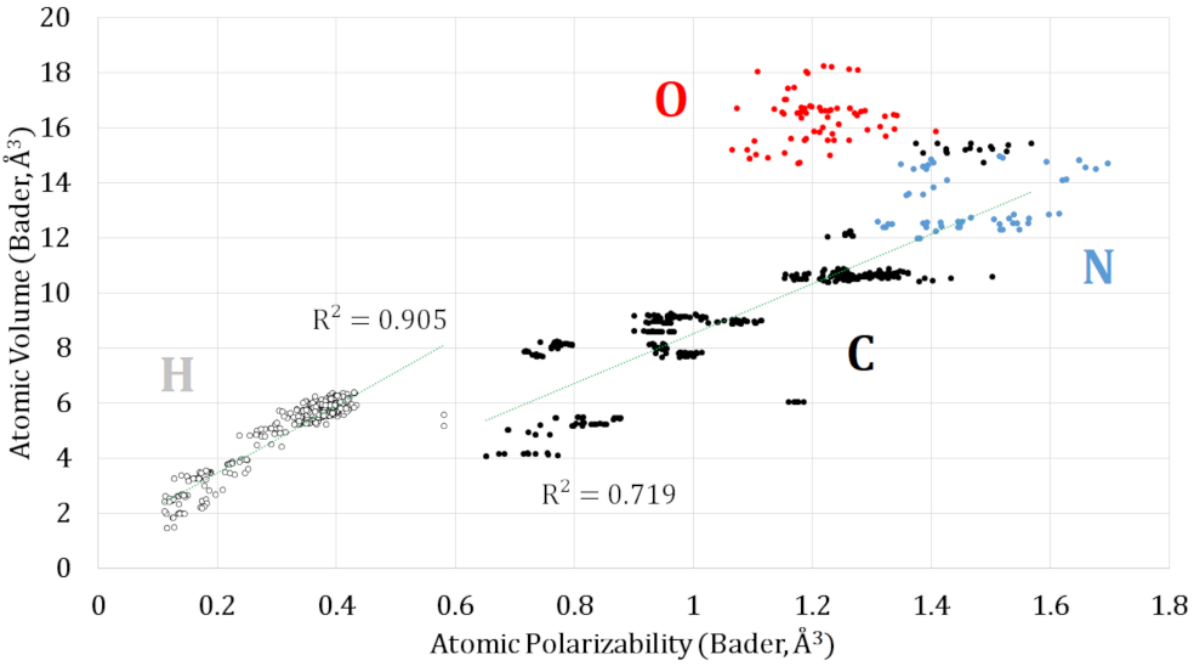


Figure 13. Corrélations entre volumes topologiques et polarisabilités topologiques (provenant indifféremment des monomères et des dimères), classées par élément chimique. Carbone : noir ; hydrogène : blanc ; azote : bleu ; oxygène : rouge.

1.7.e - Polarisabilités anisotropes par éléments chimiques

Bien entendu, la description isotrope proposée à la partie précédente est extrêmement sommaire par rapport à la description anisotrope de la polarisabilité, accessible par les calculs réalisés. La Figure 14 présente les valeurs propres des tenseurs de polarisabilité atomiques, calculées dans les dimères et les monomères S66, dont les moyennes portent sur les atomes H, C, N et O. Les *SSD* associés aux moyennes sont représentés [équation (A7), annexe page 292]. En utilisant cette classification atomique, il apparaît que la dispersion autour des moyennes est plutôt élevée, avec des *SSD* compris entre 11.1% et 37.7% de la moyenne de la valeur propre correspondante. Les magnitudes des valeurs propres parmi les atomes d'espèces chimiques identiques varient significativement dans le jeu de données S66. Cependant, les ellipsoïdes de polarisabilité sont notablement plus variables pour les atomes de carbone que pour les atomes de N, O et H. Le caractère systématiquement oblong de la polarisabilité des atomes d'hydrogène apparaît. Pour les polarisabilités associées aux hydrogènes, λ_1 a une valeur significativement plus grande par rapport à λ_2 et λ_3. Les valeurs moyennes λ_2 et λ_3 sont très similaires et sont associées à des valeurs de *SSD* importantes, avec des intervalles

d'erreur qui se chevauchent. Le cas des polarisabilités des atomes de carbone est particulièrement remarquable ici. Les invervalles d'erreur des trois moyennes des valeurs propres se chevauchent et ont de forts *SSD*. Par conséquent, une grande variété d'ellipsoïdes, de tailles et de formes variables, peut être trouvée parmi les polarisabilités des atomes de carbone du jeu de données S66. En conséquence, il apparaît que les polarisabilités atomiques moyennes anisotropes, calculées par espèce chimique, ne peuvent pas être considérées comme des entités transférables, en particulier celles des atomes de carbone.

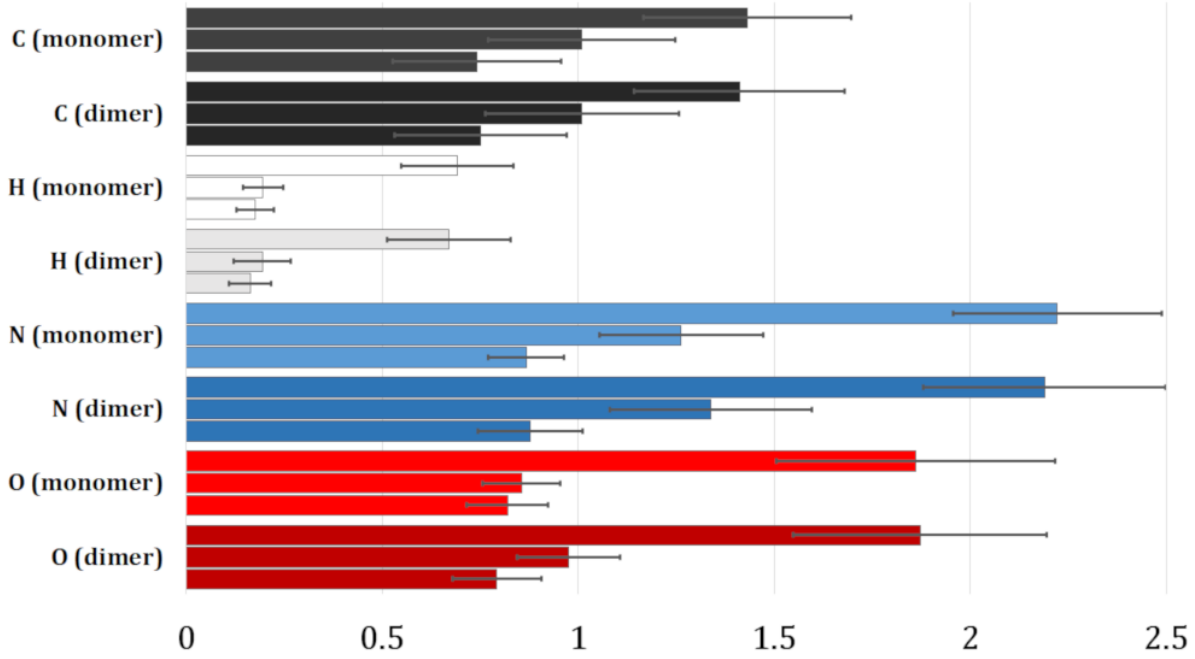


Figure 14. Moyennes des valeurs propres pour toutes les polarisabilités, classées par éléments chimiques, en Å^3. Les barres d'erreur symbolisent le *SSD* pour chaque échantillon.

Pour examiner les formes des tenseurs de polarisabilités avec plus de détails, on utilise des indicateurs basés sur leurs valeurs propres. Pour apprécier la forme globale de ces polarisabilités anisotropes théoriques, on peut comparer leur ellipticité ε [équation (A2)] et leurs ratios η [équation (A3)], page 291, en annexe. La Figure 15 regroupe les valeurs de ε et η, pour toutes les polarisabilités atomiques calculées à partir du jeu de données S66.

La ligne rouge ($\varepsilon = \eta - 1$, qui implique $\lambda_2 = \lambda_3$) indique la limite des deux paramètres tracés. Les polarisabilités le long de cette limite sont cylindriquement symétriques autour de leur premier vecteur propre, allant du cas sphérique (en bas, à gauche) à oblong (*prolate*, en haut, à droite). Considérant une valeur η constante, les ellipsoïdes aplaties (*oblate*) ont des valeurs ε plus faibles par rapport à la limite cylindrique (la valeur de λ_2 est plus proche de λ_1 que de λ_3). Ainsi, les ellipsoïdes aplaties s'écartent de la limite dans la direction du coin inférieur droit de la Figure 15. À l'exception d'un atome d'hydrogène excentré, les polarisabilités aplaties sont principalement associées à des atomes de carbone et d'azote aromatiques. Les tenseurs de polarisation les plus oblongs correspondent aux atomes d'hydrogène (la direction associée à leur plus grande valeur propre est presque toujours orientée le long de leur unique liaison covalente). Les atomes d'oxygène et de carbone occupent la région centrale du graphe, où les ellipsoïdes de polarisabilité présentent des caractères plus ou moins oblongs ou sphériques. L'oxygène de l'eau et le carbone central du néopentane ont les deux polarisabilités les plus sphériques trouvées dans le jeu de données S66. La Table 4 décrit quelques polarisabilités atomiques particulières, mises en évidence sur la Figure 15 par des croix vertes. Les paramètres de forme de ces polarisabilités particulières sont donnés pour aider le lecteur à saisir l'ampleur des variations présentées par les atomes de carbone et d'hydrogène.

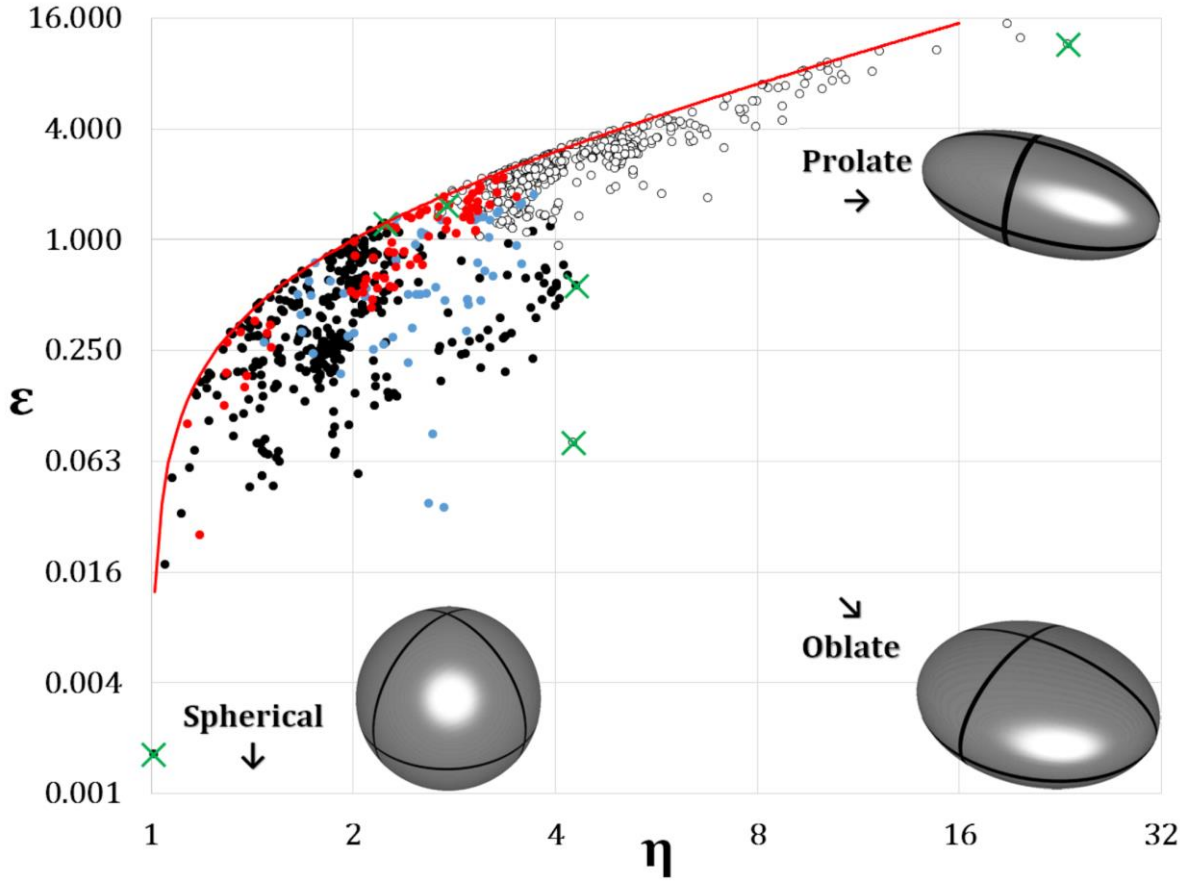
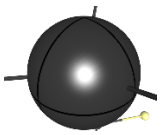
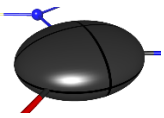
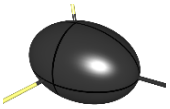


Figure 15. Paramètres topologiques des polarisabilités atomiques du jeu de données S66. L'ellipticité ε est décrite par l'équation (A2) et le rapport η, par l'équation (A3). Les deux axes utilisent une échelle logarithmique de base deux pour répartir le nuage de points. La ligne rouge ($\varepsilon = \eta - 1$) affiche la limite des deux paramètres tracés. Carbone : noir ; hydrogène : blanc ; azote : bleu ; oxygène : rouge. Les croix vertes désignent des atomes particuliers de la Table 4 (page suivante).

Table 4. Polarisabilités atomiques associées à des atomes de carbone et d'hydrogène présentant des ellipsoïdes de formes remarquables.

Nom de l'atome*	Molécule, Partenaire	Type atomique ELMAM2 (voir annexe)	ε	η	Image	Notes
C1	Néopentane (monomère)	C-C_4 (C405T)	0.001588	1.010		Carbone le plus sphérique, minimum de ε & η
C23	Uracile, Uracile (BP)	C-NNO (C315)	0.5604	4.304		Carbone le plus aplati
C19	Néopentane, Pentane	C-H_3 (C401)	1.219	2.233		Maximum de ε pour un carbone.
H14	Éthyne, Benzène	H-C≡CH (H118)	0.07820	4.254		Forme anormale**
H9	Pentane, Néopentane	H-C-HCC (H104)	1.528	2.761		Minimum de η pour un hydrogène
H20	Peptide, Benzène	H-N-CO (H112)	11.41	23.24		Maximum de η pour un hydrogène.

* : Symbole chimique suivi par le numéro de ligne de l'atome correspondant dans le fichier *.xyz original (www.begdb.com). ** : La forme de l'hydrogène éthyne est un artefact de la partition QTAIM : son bassin atomique est en contact avec six bassins atomiques des carbones du cycle benzénique voisin.

1.7.f - Polarisabilités anisotropes par types atomiques ELMAM2

Pour améliorer leur transférabilité, il faut regrouper les polarisabilités anisotropes de manière à saisir leur diversité. Une description plus détaillée des similitudes atomiques doit faire émerger de nouvelles tendances à étudier. À cette fin, les types atomiques ELMAM2 ont été envisagés puisque, par construction, ils partagent des caractéristiques chimiques similaires, l'une étant le même environnement covalent. Les valeurs propres moyennes des tenseurs de polarisabilité, calculées dans les systèmes monomères, peuvent être trouvées dans le panneau de gauche de la Figure 16, avec leur *SSD* associé. La moyenne court sur toutes les occurrences d'un type atomique ELMAM2 dans les molécules du S66. Le panneau de droite de la Figure 16 montre la même moyenne de valeurs propres sur des atomes partageant un type d'atome ELMAM2, mais cette fois calculée en dimères[**].

Figure 16 (ci-contre). Valeurs propres des ellipsoïdes de polarisabilité atomique trouvées dans le jeu de données S66, moyennées par atomes partageant un même type atomique ELMAM2, calculées en molécules isolées (à gauche) et calculées dans les systèmes dimères du S66 (à droite). Les barres d'erreur affichent le *SSD* de chaque échantillon (type atomique). Si le type d'atome n'est présent qu'une seule fois au sein du jeu de molécules, les barres d'erreur sont fixées à zéro. Panneau (a) : hydrogène ; panneau (b) : carbone ; panneau (c) : azote ; panneau (d) : oxygène.

[**] *: Le détail de l'intégralité des valeurs des polarisabilités atomiques anisotropes peut être retrouvé dans le matériel supplémentaire de l'article traduit ici (référence 52).*

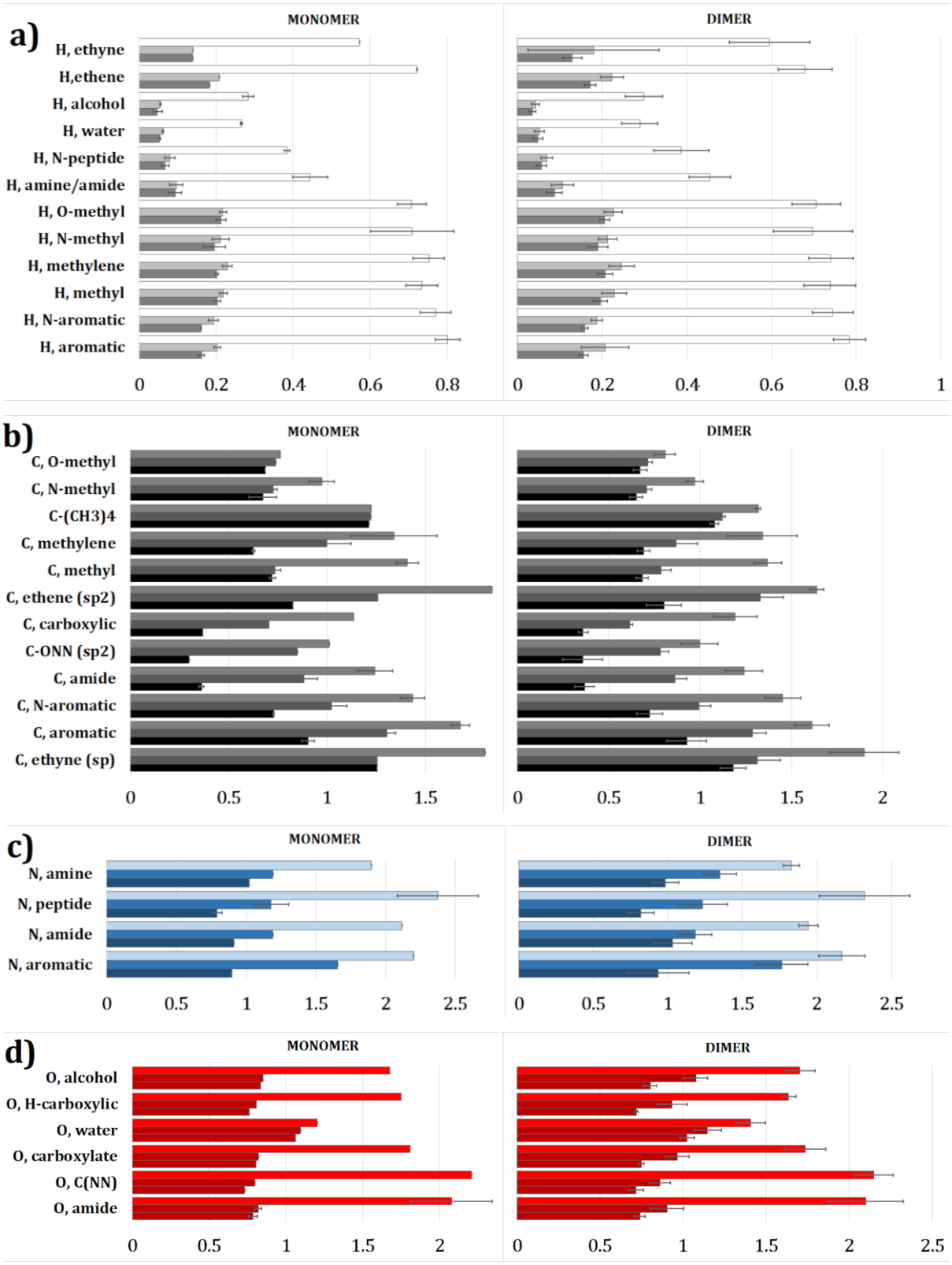

a)
MONOMER
DIMER
H, ethyne
H,ethene
H, alcohol
H, water
H, N-peptide
H, amine/amide
H, O-methyl
H, N-methyl
H, methylene
H, methyl
H, N-aromatic
H, aromatic
0
0.2
0.4
0.6
0.8
1
b)
MONOMER
DIMER
C, O-methyl
C, N-methyl
C-(CH3)4
C, methylene
C, methyl
C, ethene (sp2)
C, carboxylic
C-ONN (sp2)
C, amide
C, N-aromatic
C, aromatic
C, ethyne (sp)
0
0.5
1
1.5
2
c)
MONOMER
DIMER
N, amine
N, peptide
N, amide
N, aromatic
0
0.5
1
1.5
2
2.5
d)
MONOMER
DIMER
O, alcohol
O, H-carboxylic
O, water
O, carboxylate
O, C(NN)
O, amide
0
0.5
1
1.5
2
2.5

Figure 16.

Dans les deux cas (monomères ou dimères) présentés Figure 16, la tendance générale observée sur la Figure 15 est toujours visible : les atomes d'hydrogène sont les moins polarisables et affichent des valeurs similaires pour λ_2 et λ_3, qui sont petites comparées à λ_1. Tous les azotes ont une polarisabilité légèrement aplatie ($\lambda_2 > \lambda_3$). Les polarisabilités associées aux oxygènes (à l'exception de l'oxygène de l'eau) ont un caractère plus oblong, mais la différence relative entre λ_1 et les deux autres valeurs propres est plus petite. En utilisant la classification basée sur les types d'atomes ELMAM2, les polarisabilités des atomes d'oxygène font maintenant apparaître leurs différences : par exemple, l'oxygène de l'eau dévie fortement, comparé aux atomes d'oxygène des groupes hydroxyles. Encore une fois, le cas des atomes de carbone est remarquable. Ils présentent les variations les plus larges en termes de magnitude et de forme de leurs polarisabilités. Ces types atomiques sont associés à des polarisabilités sphériques, oblongues ou aplaties. Cependant, cette propriété apparaît sur la Figure 16 comme dépendante du type d'atome ELMAM2.

De plus, en considérant les moyennes sur tous les types d'atomes (pour les systèmes monomères et dimères), tous les *SSD* sont plus faibles que les *SSD* correspondants aux moyennes de la Figure 14 (par éléments chimiques). Par exemple, le *SSD* relatif sur la polarisabilité des atomes de type carbone méthyle est de 1.8 %, tandis que la même valeur pour tous les atomes de carbone est de 19.4 %. Pour le méthyle hydrogène, le *SSD* relatif est de 4.7 %, ce qui se compare aux 22.0 % obtenus pour tous les atomes d'hydrogène. Pour tous les types ELMAM2 représentant des atomes de carbone, si on exclut celui du carbone méthylène (voir ci-dessous), les intervalles d'erreur de λ_1 ne chevauchent pas les intervalles d'erreur associés aux deux autres valeurs propres. Ceci est interprété comme augmentation de spécificité, en forme et en magnitude, de la polarisabilité du type d'atome considéré, toujours par rapport à la moyenne des polarisabilités par éléments chimiques, présentée à la Figure 14. Cette manière de définir la transférabilité d'une propriété est également utilisée par les auteurs de la banque de données UBDB[68]. Dans leur approche, les paramètres d'un type d'atome donné (présents dans les petites molécules utilisées pour construire leur librairie) doivent présenter de petites variances, et ces variances doivent être suffisamment faibles pour que le type d'atome considéré puisse être distingué d'autres types d'atomes apparentés. C'est exactement ce qui est réalisé ici, en considérant les valeurs propres des tenseurs de polarisabilité, au lieu des paramètres de la densité électronique. Par conséquent, ces

observations suggèrent que les tenseurs de polarisabilité, lorsqu'associés à un type d'atome ELMAM2 donné (parmi ceux représentés dans le jeu de données S66), présentent des similitudes qui peuvent être exploitées pour construire une base de données de polarisabilités atomiques transférables.

Le cas du type atomique décrivant le carbone méthylène, évoqué précédemment, est un bon exemple de moyennation des polarisabilités à partir d'un échantillon hétérogène. En effet, nous avons inclus dans cette moyenne les atomes de carbone méthylène du pentane (de polarisabilité relativement oblongue) et ceux du cyclopentane (de polarisabilité relativement aplatie). Les carbones méthylènes des deux molécules ont le même voisinage covalent. Le caractère cyclique de la molécule apparaît comme une différence chimique pertinente pour la transférabilité des tenseurs de polarisabilité. Un seul type d'atome couvre les deux cas dans la base de données ELMAM2, malgré des angles C-C-C moyens différents (Pentane: 112.8 ° ± 0.4, Cyclopentane: 103.8 ° ± 1.3). Cela suggère que, dans ce cas, différencier ces deux types d'atomes ELMAM2 augmenterait la transférabilité des polarisabilités de ces pseudoatomes. Deux autres atomes ont un *SSD* élevé sur λ_1 (Figure 16): l'oxygène amide (type d'atome ELMAM2 noté O102) et l'azote de la liaison peptidique (N302). Ces types d'atomes sont retrouvés dans l'uracile, cyclique, et le *N*-méthyl-acétamide, linéaire (nommée « peptide » dans l'ensemble de données S66). Pour tenir compte de ces différences, les futures versions de la bibliothèque ELMAM2 pourraient comporter des types d'atomes séparés pour le carbone méthylène et l'azote peptidique, en fonction de leur implication dans une molécule cyclique ou non, afin d'accroître leur transférabilité. Le cas de l'atome d'oxygène du groupement amide est moins clair, car les *SSD* associés à λ_2 et λ_3 sont faibles, par conséquent ce type d'atome ne sera pas affecté, *a priori*, d'une duplication similaire.

1.7.g - Polarisabilités et incertitudes sur les populations dipolaires ELMAM2

Une autre manière d'appréhender la transférabilité de la polarisabilité atomique est disponible Figure 17. On compare ici les polarisabilités isotropes moyennes à la dispersion des paramètres dipolaires (du modèle HCM) de chaque type atomique, telle que tabulée dans le fichier ELMAM2. Cette dispersion est celle des populations multipolaires des atomes ayant servi à créer l'entrée en question dans la base de données

ELMAM2 (voir la liste en annexe, Table A1). Le *SSD* relatif d'une population multipolaire pour un ordre *l* donné, où *k* est le nombre de populations non-nulles d'ordre *l*, a été évalué par :

$$SSD(P_{l\pm m})_{relatif} = \frac{1}{k}\sum_{-l}^{+l}\frac{SSD(P_{l,m})}{|P_{l,m}|}, \quad \forall P_{l,m} \neq 0 \qquad \textbf{(71)}$$

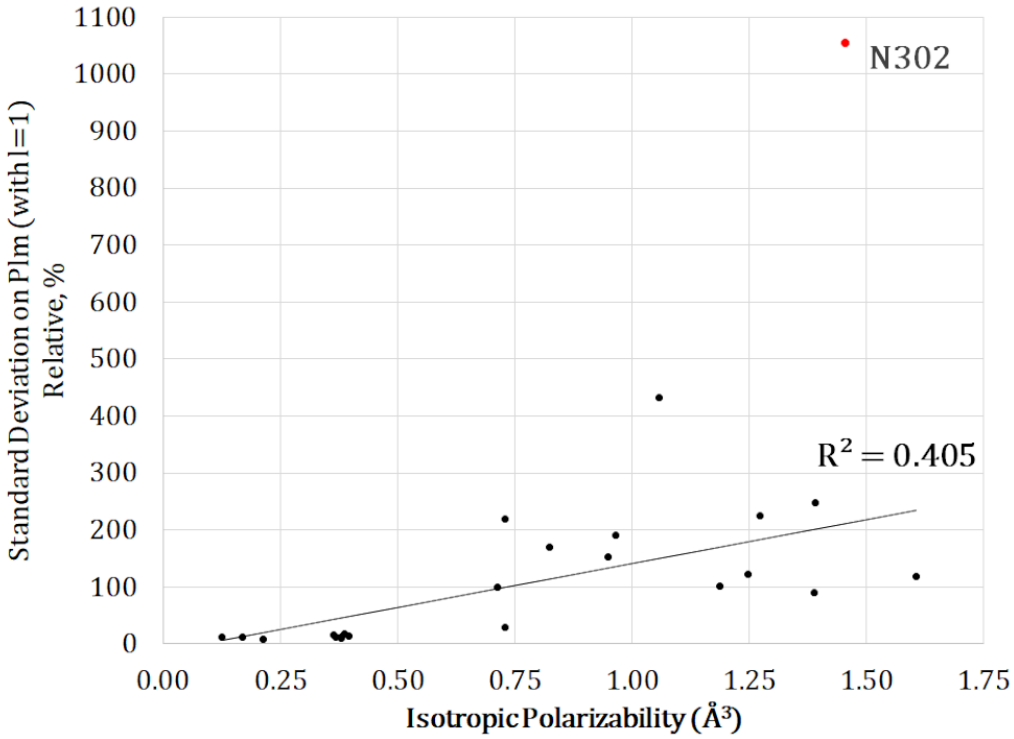


Figure 17. Corrélation entre les polarisabilités isotropes, moyennées par types d'atomes ELMAM2, et le *SSD* relatif [équation (71)] des populations multipolaires d'ordre l = 1 de ces mêmes types d'atomes. Les types atomiques construits à partir de moins de 6 observations (comme documenté dans ELMAM2) ont été exclus. L'azote peptidique (N302, en rouge) a un *SSD* relatif exceptionnellement haut et a été exclu de cette régression linéaire.

On compare ce *SSD* relatif, calculé à partir de $P_{1,-1}$, $P_{1,0}$ et $P_{1,+1}$, aux polarisabilités isotropes, moyennées par atomes de même type. Une tendance claire est observée pour les atomes les plus polarisables : ils appartiennent aux types atomiques avec un *SSD* relatif plus élevé. Au cours de la construction d'ELMAM2, les atomes multipolaires les plus polarisables présentaient donc un ensemble plus variable de populations dipolaires $P_{1,m}$ lors du calcul des paramètres multipolaires moyens. Ce résultat était attendu compte tenu

de notre modèle linéaire de polarisabilité et puisque les atomes utilisés pour construire ELMAM2 sont issus d'environnement cristallins variés.

L'azote de la liaison peptidique (N302, en rouge sur la Figure 17) présente une valeur aberrante de *SSD*. Il présente une petite population $P_{1,-1}$ associée à un *SSD* exceptionnellement grand. Ce type d'atome particulier, comme le type d'atome du carbone méthylène, est retrouvé à la fois dans les molécules cycliques (uracile) et linéaires (N-méthyl-acétamide) du S66. Ce *SSD* relatif élevé incite à prendre des précautions supplémentaires pour ce type d'atome afin d'améliorer le modèle (voir chapitre 3-bis).

1.7.h - Variations des valeurs propres de la polarisabilité

Le transfert et la comparaison automatique des polarisabilités, entre les systèmes monomères et les systèmes dimères ont permis d'effectuer un traitement cohérent de l'ensemble du jeu de données S66. Pour apprécier les variations globales des polarisabilités atomiques lors de la formation de dimères, les polarisabilités calculées dans les systèmes dimères ont été comparées à celles calculées dans les systèmes monomères. Le calcul des variations moyennes des valeurs propres de ($\boldsymbol{\alpha}_{dimère}$ - $\boldsymbol{\alpha}_{monomère}$), et de les *SSD* correspondants sont présentés la Figure 18. Pour 20 des 34 types d'atomes ELMAM2 trouvés dans le set S66, l'intervalle d'erreur englobe 0 pour les variations de chaque valeur propre. Le détail de ces petites variations de polarisabilité atomique (en ce qui concerne les valeurs propres des ellipsoïdes) sont montrés Figure 18.

La variation moyenne des polarisabilités atomiques topologiques, entre l'état dimère et l'état monomère, n'est que de 2.1 % (sur tous les types d'atomes). Ceci suggère que ces 34 types d'atomes représentés dans le jeu de données S66 sont associés à des polarisabilités avec de très faibles variations. Cette variation n'a pas de direction systématique (voir plus bas). Inversement, certains types d'atomes méritent plus d'attention. L'exemple le plus frappant est le carbone central du néopentane, qui passe systématiquement d'une ellipsoïde de polarisabilité sphérique à une ellipsoïde allongée lorsqu'il se lie à un autre monomère (provoquant une modification importante de ε et η).

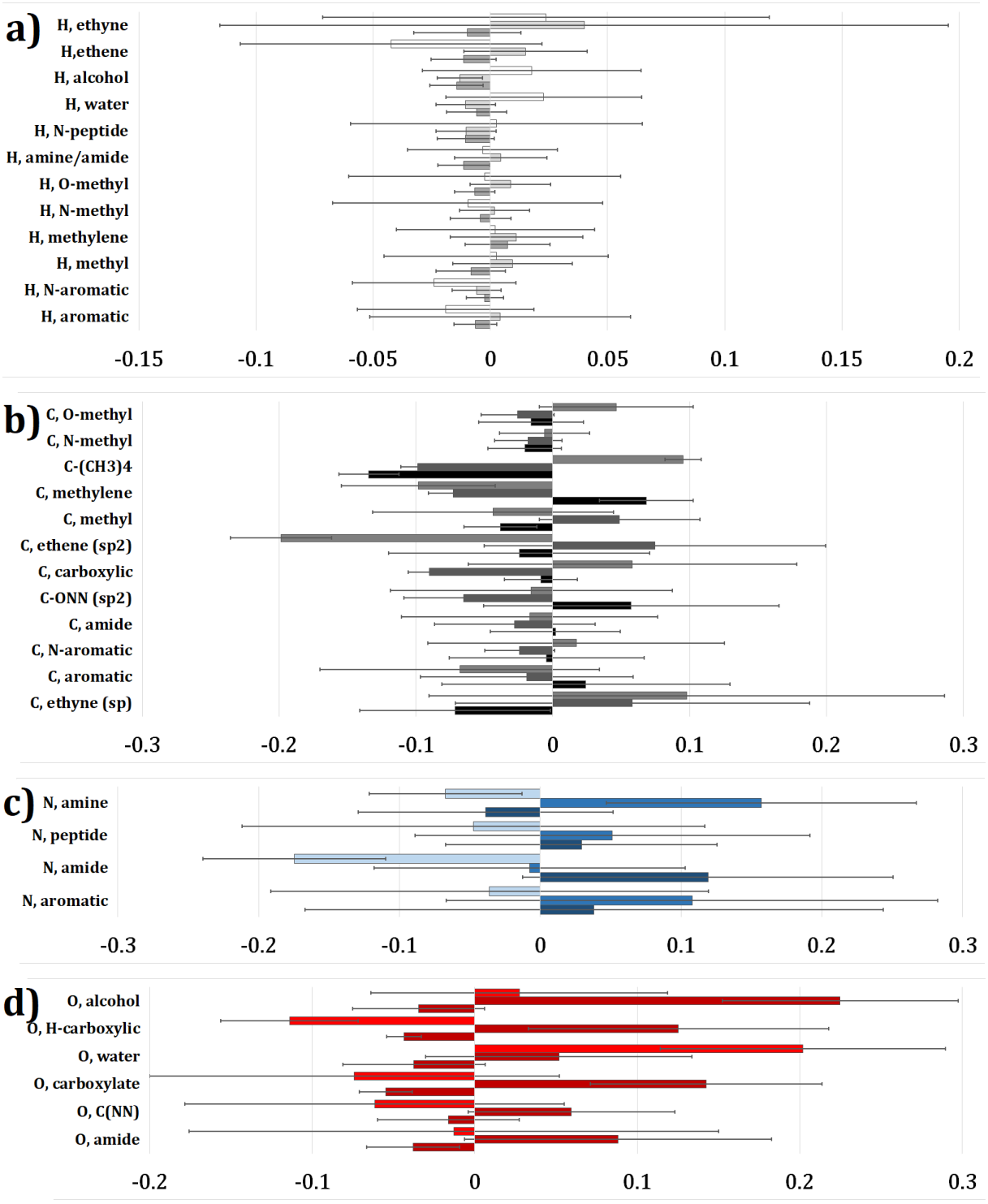


Figure 18. Variations moyennes des trois valeurs propres des ellipsoïdes de polarisabilité atomique, comparant dimères et molécules isolées. La moyenne a été calculée sur tous les atomes partageant un type ELMAM2. Les barres d'erreur montrent le *SSD* de chaque échantillon (type d'atome). a) hydrogène ; b) carbone ; c) azote ; d) oxygène.

Les polarisabilités de l'oxygène de l'eau et du carbone méthoxy- sont également presque sphériques dans les molécules isolées. Ces deux types atomiques acquièrent une polarisabilité allongée dans les dimères du S66. Pour l'azote du groupement amine et pour tous les atomes d'oxygène (sauf l'oxygène de l'eau), on observe une augmentation significative de λ_2. L'azote du groupement amide (N302), l'oxygène protoné de l'acide carboxylique et le carbone de l'éthène présentent une chute significative de λ_1. Pour ces 8 types d'atomes, le calcul en dimères donne des polarisabilités atomiques topologiques beaucoup plus aplaties que lors du calcul sur les molécules isolées.

1.7.i - Variations d'orientation des polarisabilités

Les variations d'orientation de chaque tenseur de polarisabilité entre l'état isolé et l'état dimère sont estimées à l'aide de la direction du premier vecteur propre. En effet, la plupart des polarisabilités calculées ont un vecteur propre univoque associé à λ_1. L'équation (A1), en annexe, permet d'obtenir l'angle entre cette direction à l'état isolé et cette direction à l'état dimère. La Figure 19 montre cette réorientation moyenne pour chaque type d'atome. La plupart des types atomiques présentent de faibles variations d'orientation (<15 °). Cependant, des réorientations importantes se produisent pour les ellipsoïdes des atomes de carbone de l'éthyne, l'éthène et le benzène. Les réorientations angulaires et leur *SSD* associé [équation (A7), page 292] pour le carbone central du néopentane, l'oxygène de l'eau et du carbone méthoxy- sont très élevés en raison du caractère quasi-sphérique de leur polarisabilité. Un changement moyen d'orientation au-dessus de 45 ° découle ici systématiquement du choix arbitraire de vecteurs propres pour des polarisabilités sphériques ($\lambda_1 \simeq \lambda_2 \simeq \lambda_3$). Ces valeurs sont hachurées sur la Figure 19 et ne seront pas interprétées.

Tous les atomes d'hydrogène ont des angles de réorientation comparables. Seulement 1.3 % (7 sur 533) des atomes d'hydrogène montrent une réorientation au-dessus d'un seuil – arbitraire – de 15 °. Les atomes d'hydrogène réorientent moins leurs polarisabilités, comparés aux atomes plus lourds : pour les atomes d'azote, cette proportion s'élève à 13.7 % (7 sur 51). En excluant les trois atomes (quasi-isotropes) susmentionnés, 14.0 % (8 sur 57) des atomes d'oxygène et 16.1 % (44 sur 273) des

atomes de carbone réorientent leurs polarisabilités anisotropes de plus de 15 ° lors de la dimérisation.

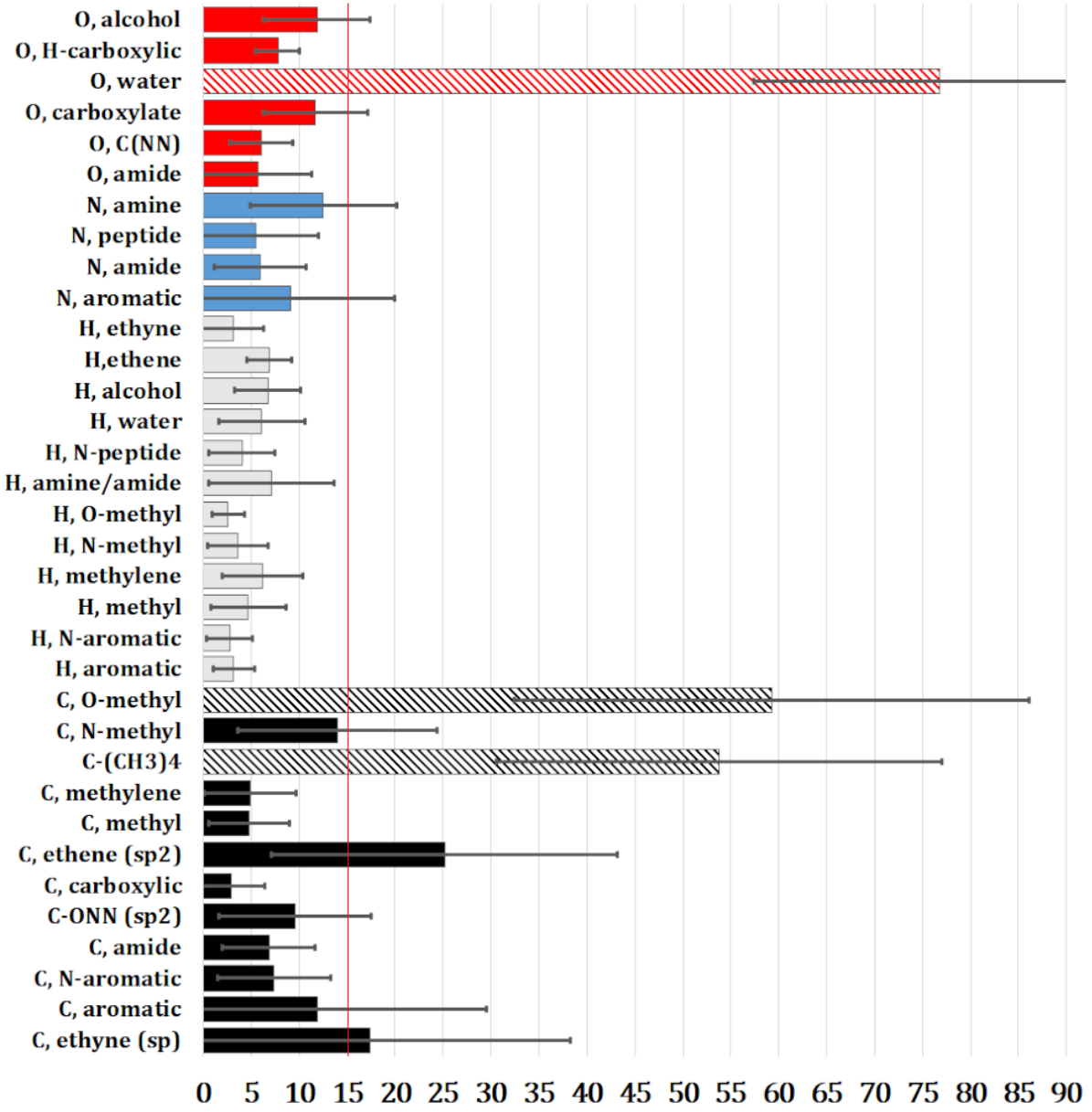


Figure 19. Angle de réorientation moyenne (en degrés) entre le vecteur propre $\mathbf{v}_1$ (associé à λ_1) de la polarisabilité atomique des dimères et le même vecteur propre de la polarisabilité atomique des monomères. Les barres d'erreur symbolisent le *SSD* pour chaque type atomique ELMAM2. Les barres hachurées sont artificiellement hautes et correspondent à un artefact de diagonalisation provoqué par des polarisabilités atomiques quasi-sphériques.

La comparaison des atomes de carbone, d'azote et d'oxygène est moins informative compte tenu des différences dans la taille des échantillons (entre C, N et O et entre les trois atomes lourds et les hydrogènes). Les polarisabilités des carbones de l'éthyne et de l'éthène apparaissent comme les plus susceptibles de se réorienter dans les dimères. Ces carbones sont les plus chargés négativement parmi les molécules du S66.

Expliquer les variations d'orientation de la polarisabilité, sur la seule base de la densité électronique multipolaire, est une tâche délicate. En utilisant les géométries moléculaires des dimères, l'angle entre le premier vecteur propre $\mathbf{v}_1$ de chaque polarisabilité atomique et le champ électrique ELMAM2 généré par l'autre monomère a été calculé. On compare cet angle calculé à partir des polarisabilités dans les dimères à l'angle calculé à partir des monomères (après transfert des polarisabilités des monomères aux systèmes dimères et polarisation jusqu'à l'équilibre, voir partie 2.4). Cet angle entre $\mathbf{v}_1$ et le champ électrique est plus faible pour 631 polarisabilités atomiques sur 939 calculées dans des dimères lorsque ceux-ci sont comparés aux angles issus des molécules isolées. Cette légère tendance – des polarisabilités à s'aligner avec le champ – a été étudiée et la séparation des données par élément, molécule, type d'atome ou groupe de types d'atome n'a révélé aucune corrélation satisfaisante (ces données ne seront présentées). Les variations sont larges, aucune autre corrélation avec l'orientation ou l'intensité du champ électrique, les polarisabilités, la distance au centre de masse de l'autre monomère ou les charges atomiques n'a permis d'expliquer l'intensité (ou la direction) de cette tendance ténue des polarisabilités atomiques à s'aligner sur les champs électriques ELMAM2.

1.8 - Amélioration du set de polarisabilités atomiques théoriques

La partie précédente suggère, selon divers arguments, la faisabilité d'une base de données d'atomes multipolaires polarisables, c'est-à-dire augmentés d'une polarisabilité anisotrope moyenne. En effet, catégoriser les polarisabilités atomiques par types ELMAM2 permet de réduire sensiblement la variabilité de leurs valeurs propres moyennes. De plus, pour de nombreux types atomiques, les variations – en magnitude et orientation – entre les systèmes dimères et monomères sont faibles pour une grande majorité de types atomiques (les exceptions observées nécessitent une vigilance

particulière). Si une base de données de polarisabilités atomiques à proprement parler n'a pas été construite dans l'article présenté, *MoProViewer* possède désormais la capacité de le faire. Cette capacité passe par l'implémentation de différents outils et méthodes qui seront détaillés dans le chapitre suivant.

Il faut maintenant s'intéresser au matériel de base pour construire cette librairie de polarisabilités transférables. En effet, si le set S66 est intéressant de par la possibilité d'apprécier des variations de polarisabilités atomiques liées aux interactions non-covalentes (dimères), ce jeu de données n'est constitué que de 14 molécules différentes. Leur taille (en nombre d'atomes) est très réduite. Pour comparaison, les densités électroniques d'ELMAM2 ont été moyennées à partir de 54 structures expérimentales différentes (voir Table A2 de l'annexe). Il est donc nécessaire d'obtenir un plus grand nombre de polarisabilités moléculaires et atomiques théoriques, afin que la librairie ait une bonne couverture des types différents atomiques utiles à la description d'un complexe protéine-ligand.

Des collaborateurs, au sein du laboratoire et au-delà, ont été mis à contribution pour générer un large jeu de polarisabilités moléculaires atomiques. Le Dr. Emmanuel Aubert a réalisé les calculs des polarisabilités atomiques portant sur les molécules du S66, ainsi que les polarisabilités atomiques du benzamide fluoré F2,25. Cette molécule a été ajoutée pour une éventuelle collaboration avec l'Université de Dublin (plus particulièrement en rapport avec les travaux de John Gallagher sur les benzamides substitués[91]). Afin d'obtenir un jeu de données cohérent avec les applications envisagées, reportées au chapitre 2, les collaboratrices de l'université de Cracovie ont fourni le reste des résultats. La professeure Anna Krawczuk a calculé les polarisabilités atomiques topologiques pour les 20 acides aminés protéinogènes, accompagnées de celles pour la *N*-acétyl-tyrosine-éthyl-ester, 3 di-peptides (Ala-Met, Gly-Ala et Gly-Thr) et 5 tri-peptide (Ala-Pro-Ala, Tyr-Gly-Gly et 3 Gly-Gly-Gly dans des contextes moléculaires différents). La docteure Katarzyna Rzesikowska a fourni les polarisabilités topologiques pour les autres di- et tri-peptides du jeu de données, pour un total de 127 oligopeptides (voir table A3 en annexe). 3 autres molécules ont été demandées pour permettre les applications envisagées aux chapitres 2 et 3 : le cofacteur biotine, l'inhibiteur de protéase benzamidine et le solvant diméthyléther.

Les 198 résultats de calculs retenus, et présentés ici, ont été réalisés *in vacuo* par les collaborateurs sus-cités. Certains résultats de calculs ont été rejetés, pour cause de polarisabilités atomiques non définies-positives (33 structures en tout ont été exclues). Ce problème a été également constaté avec le jeu de données S66, dont 9 dimères ont été exclus de l'analyse pour des raisons similaires (voir page 81). D'autres résultats, utilisant un modèle implicite de solvatation (de type *Polarizable Continuum Model*), ont été fournis mais restent à exploiter au moment de rédaction de ce manuscrit. Cette collaboration a été amorcée assez tard dans la progression de la thèse (ce qui explique que ces polarisabilités n'apparaissent pas dans l'article Leduc *et al.*, 2019[52]).

La méthodologie utilisée par Anna Krawczuk et Katarzyna Rzesikowska est celle recommandée pour l'utilisation du logiciel *PolaBer*[49]. Elle diffère de celle présentée précédemment et mise au point par Emmanuel Aubert. La fonctionnelle utilisée est CAM-B3LYP qui comprend des corrections à longue portée absentes de B3LYP, et le *basis set* est d-aug-cc-pVDZ, qui utilise des fonctions de type Slater plutôt que les fonctions gaussiennes du *basis set* 6-311++G(3df,3pd) pour représenter la densité électronique. L'utilisation d'une fonctionnelle plus complexe et de fonctions de Slater implique des temps de calculs supérieurs, ce qui justifie que – dans « l'approche *PolaBer* » – seules 2 valeurs non-nulles de champ électrique soient évaluées par direction de l'espace (±0.005 a.u., soit ±0.0014 e·Å^{-2}). Dans les deux cas, *Gaussian09*[84] a été utilisé pour les calculs DFT, et *AimAll*[85], pour obtenir la partition topologique de la densité électronique. La seconde différence majeure entre les deux méthodes théoriques employées réside dans l'évaluation du transfert de charge entre les bassins atomiques. Tel que défini dans Krawczuk, Pérez & Macchi[49], le moment dipolaire moléculaire est la somme de deux termes : le terme de polarisation atomique $\boldsymbol{\mu}_p(\boldsymbol{\Omega})$ et le terme de translation de charge, $\boldsymbol{\mu}_c(\boldsymbol{\Omega})$. Le terme de polarisation atomique est strictement homologue au moment dipolaire intégré sur un bassin de densité de l'équation (35). Quant à $\boldsymbol{\mu}_c(\boldsymbol{\Omega})$, il décrit le moment dipolaire généré par la différence de charge entre bassins atomiques voisins lors de l'application du champ électrique. L'évaluation des moments induits $\boldsymbol{\mu}_c(\boldsymbol{\Omega})$ permet de définir des polarisabilités de liaison (localisées au niveau du point critique de liaison covalente dans la densité électronique de la molécule considérée). Pour le traitement du S66, $\boldsymbol{\mu}_c(\boldsymbol{\Omega})$ a été supposé négligeable devant les faibles valeurs de champ appliquées. L'effet très faible de cette omission sur le S66 est examiné à la partie 2.4.b. Malgré l'omission des polarisabilités des liaisons liées à $\boldsymbol{\mu}_c(\boldsymbol{\Omega})$ et les différences dans les calculs

numériques, les polarisabilités obtenues par les deux méthodes apparaissent assez comparables entre elles et seront donc étudiées ensemble dans l'optique d'obtenir une base de données de polarisabilités transférables. On note également que l'échec éventuel des calculs est lié, pour les deux méthodes, à l'obtention de polarisabilités atomiques non définies-positives suite à la partition de la densité électronique.

Il résulte de cette compilation de structures un jeu de données foisonnant, provisoirement nommé « *Big Set* ». En effet, le « *Big Set* » couvre 81 types atomiques ELMAM2 à l'aide de 38 molécules (ou acides aminés). Ces 38 molécules sont retrouvées dans 198 structures, pour un décompte de 4967 atomes distincts (une polarisabilité anisotrope par atome). La liste complète des structures est disponible à la table A3 de l'annexe. Les polarisabilités des 38 monomères sont diponibles Table 5.

Table 5. Polarisabilités moléculaires théoriques (isotropes et valeurs propres) des 38 molécules isolées présentes dans le « *Big Set* ».

Nom	λ_1 (Å^3)	λ_2 (Å^3)	λ_3 (Å^3)	$\boldsymbol{\alpha}_{iso}$ (Å^3)
Acétamide	4.46	6.37	6.75	5.86
Acide acétique	3.91	5.52	5.89	5.11
Alanine	7.01	8.66	9.65	8.44
Arginine	14.78	17.17	20.53	17.49
Asparagine	10.44	11.49	13.32	11.75
Acide aspartique	8.75	11.19	12.85	10.93
Benzamidine	6.86	14.28	18.27	13.14
Benzène	6.62	12.03	12.07	10.24
Biotine	21.11	23.57	27.84	24.17
Cyclopentane	7.98	9.25	9.25	8.83
Cystéine	9.38	11.22	14.03	11.55
Diméthylether	3.42	3.62	4.24	3.76
Éthène	3.35	3.66	5.29	4.10
Éthyne	2.79	2.79	4.75	3.44
F2,25	13.97	24.62	38.89	25.83
Acide glutamique	10.98	13.31	14.13	12.81
Glutamine	12.23	13.82	14.14	13.40
Glycine	5.22	7.25	8.56	7.01
Histidine	10.84	15.45	18.91	15.07
Isoleucine	12.13	13.53	15.01	13.56
Leucine	12.77	13.59	14.44	13.60
Lysinium	12.09	14.06	17.52	14.55
N-méthyl-acétamide	5.88	7.74	9.37	7.66
Méthylamine	3.55	3.65	4.19	3.80
Méthanol	2.93	3.00	3.42	3.12
Méthionine	12.14	14.52	18.79	15.15
Néopentane	9.69	9.69	9.70	9.69
Pentane	8.56	9.19	11.45	9.73
Phénylalanine	12.98	17.11	18.78	16.29
Proline	9.58	10.90	12.10	10.86
Pyridine	6.03	10.81	11.48	9.44
Sérine	7.86	9.46	10.24	9.19
Thréonine	9.15	11.58	11.76	10.83
Tryptophane	14.94	22.33	27.82	21.70
Tyrosine	14.11	19.65	22.80	18.85
Uracile	6.044	11.14	13.86	10.35
Valine	10.70	11.89	12.75	11.78
Eau	1.33	1.36	1.44	1.37

Chapitre 2 :

Polarisation *in situ*

d'Atomes Multipolaires

Introduction

Un large jeu de polarisabilités atomiques anisotropes est désormais disponible. Déjà, le set S66 et son extension S66x8 forment déjà une masse de données délicate à manipuler manuellement. Afin de faciliter la prise en charge de vastes ensembles de propriétés atomiques, le module « *Beluga* » de *MoProViewer* propose de nombreuses fonctions relatives à la gestion des polarisabilités, à la construction automatique de bases de polarisabilités atomiques moyennes, à la détermination de la polarisation de la densité électronique et au calcul d'énergies de polarisation. L'objectif de ce chapitre est de détailler le fonctionnement et le contrôle de l'algorithme de polarisation, ainsi que les resultats de ce calcul, de leur affichage à leur analyse et interprétation.

Ce chapitre abordera d'abord les fonctions du module liées à la procédure de polarisation de la densité électronique, dans la partie 2.1.

L'emploi du module, et les options implémentées sont décrits partie 2.2.

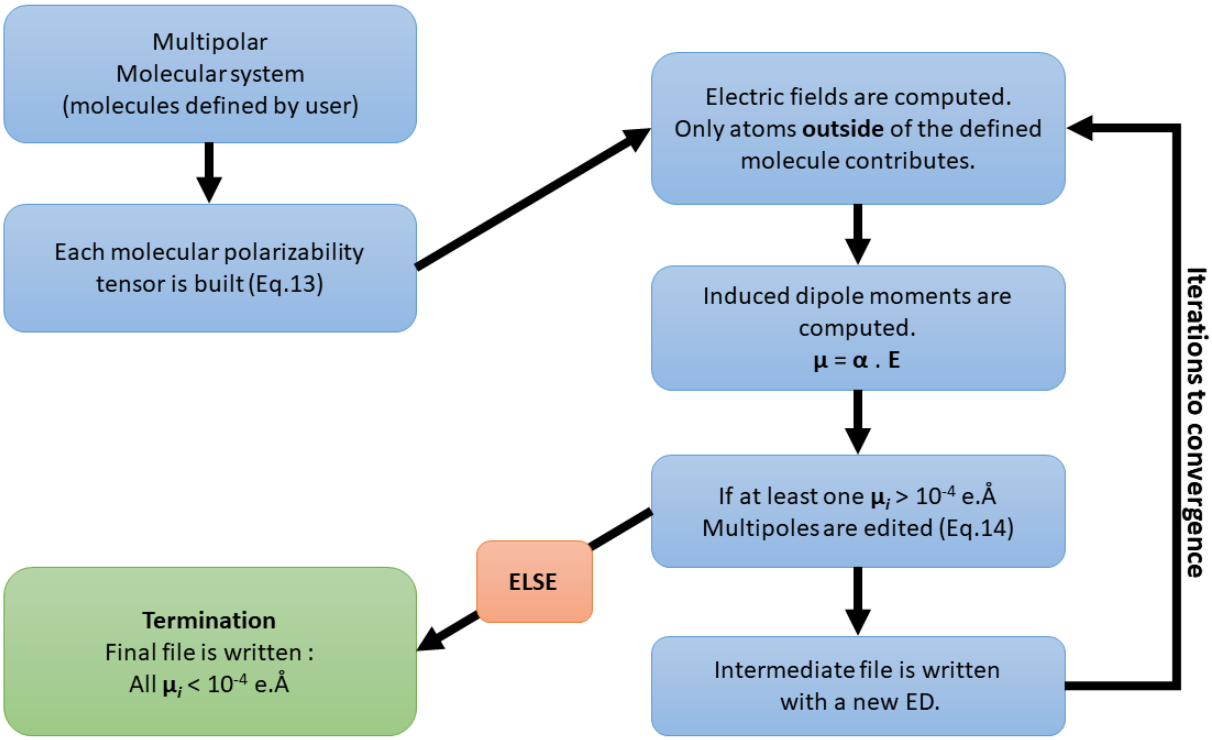
La partie 2.3 traite des résultats qu'il est possible d'obtenir, et des visualisations possibles.

Les résultats obtenus (et publiés[52]) sur la base du S66 informent sur l'induction calculée à l'aide de ces polarisabilités. Ils sont reportés partie 2.4. Dans cette partie, les performances de la librairie ELMAM2 pour reproduire des énergies d'interaction électrostatiques de référence sont reportées. L'induction dipolaire est caractérisée et les conséquences énergétiques de l'induction de moments électrostatiques occupent une large place de cette discussion. Enfin les performances globales de la méthode proposée (densités transférées-polarisées) sont reportées.

2.1 - Procédure de polarisation

L'outil développé prend en compte l'anisotropie des polarisabilités théoriques et inclut des méthodes pour les faire pivoter entre différents systèmes de coordonnées pour les atomes d'une même molécule, ou de différentes molécules, comme vu chapître 1. Ces méthodes ont pour finalité le transfert – ou la comparaison – de ces grandeurs atomiques et moléculaires. Une fois le système moléculaire préparé, on souhaite pouvoir le polariser. L'algorithme de polarisation implémenté suit les étapes présentées sur le Schéma 1.

Schéma 1. *Workflow* de l'algorithme de polarisation implémenté dans *MoProViewer*.



• L'utilisateur définit deux groupes d'atomes (molécules ou promolécules) qui se polarisent mutuellement. Chaque groupe sera donc considéré comme l'environnement chimique de l'autre.

• Un tenseur de polarisation moléculaire est construit pour chaque molécule du système. Ce tenseur inclut la polarisabilité anisotrope orientée pour chaque atome. Pour une (pro) molécule de n atomes, cette étape nécessite l'inversion d'une matrice carrée, symétrique et bloc-diagonale avec $(3 \cdot n)^2$ éléments [voir équation (17), page 38]. La matrice est constituée de tenseurs de polarisabilité atomiques, $\boldsymbol{\alpha}_i$, qui sont des tenseurs symétriques d'ordre 2 avec 3^2 éléments. L'équation (17) décrit la construction du tenseur, où $\boldsymbol{\mu}_i$ et $\mathbf{E}_i$ sont des vecteurs colonne (de $3 \cdot n$ éléments) concaténant respectivement les moments dipolaires et les champs électriques des différents atomes. Le vecteur des champs électriques ne contient que les champs externes, appliqués sur la molécule à polariser par les autres éléments du système. Toutes ces quantités doivent

impérativement être exprimées dans le même référentiel cartésien. Les zéros hors-diagonaux peuvent être remplacés par le tenseur de champ dipolaire Applequist-Thole[43,46,47,48] [$\mathbf{T}_{ij}$ ou $\mathbf{T}_{\alpha\beta}$, voir équation (21)]. Les tenseurs de $\mathbf{T}_{ij}$ modélisent l'induction dipolaire intramoléculaire [équation (20)] et sont discutés partie 4.5.

• Le champ électrique est calculé par différenciation numérique (en utilisant une expansion en série de Taylor d'ordre 4) du potentiel électrostatique EP/MM [équation (8)] à chaque position nucléaire d'un monomère. Seule la densité de charge de l'autre molécule (ou groupe d'atomes défini par l'utilisateur) contribue au champ électrique. Cette étape est répétée pour l'autre molécule. À chaque étape de la polarisation, le logiciel garde la trace de la séquence de champs électriques appliqués à chaque position nucléaire (voir la condition d'arrêt ci-dessous).

• La multiplication du vecteur contenant tous les champs électriques par la matrice de polarisabilité moléculaire génère un vecteur contenant tous les moments dipolaires induits (exprimées dans la même base cartésienne que les polarisabilités, les champs et les éventuels $\mathbf{T}_{ij}$).

• Si la magnitude d'un dipôle atomique induit est supérieure à un seuil déterminé, les dipôles sont convertis en populations $P_{l\pm m}$ correspondant à la magnitude calculée du moment dipolaire [ordre l = 1, voir l'équation (16), page 36]. Seules ces populations multipolaires contribuent au moment dipolaire atomique, comme abordé page 28. Les populations $P_{1,m}$ induites sont écrites dans l'expansion multipolaire de l'atome correspondant. Le logiciel enregistre ensuite les expansions modifiées dans un nouveau fichier moléculaire au format *MoPro* (*.par*). Si les magnitudes des dipôles atomiques induits sont tous inférieurs au seuil, la convergence est atteinte et le calcul s'arrête après l'écriture d'un fichier moléculaire final. Un seuil de terminaison de 10^{-4} e·Å ($0{,}48 \cdot 10^{-3}$ D) est choisi, pour différentes raisons, autres que les unités cristallographiques utilisées par *MoProSuite*. Premièrement, la plupart des approches de type champ de force utilisent un seuil de 10^{-3} D dans leurs procédures itératives. Deuxièmement, le système de fichiers *MoPro* stocke les population $P_{l,m}$ avec une précision de 10^{-3} e. Pour des valeurs habituelles de κ' et ζ, ce seuil de 10^{-4} e·Å garantit que les derniers dipôles calculés soient inférieurs à cette précision. Chaque expansion multipolaire est exprimée dans un référentiel local,

centré sur les coordonnées nucléaires. Dans ce référentiel, les relations (16) modifient les paramètres multipolaires en fonction de l'induction dipolaire calculée.

• Un fichier d'entrée spécifie tous les paramètres de calcul et, une fois l'algorithme terminé, un fichier de sortie final (`PolarizationResults.out`) est enregistré (en plus des fichiers moléculaires contenant la densité électronique polarisée à chaque cycle).

2.2 - Contrôle par l'utilisateur

Afin d'avoir un contrôle clair et reproductible sur les paramètres du calcul, un système de fichier d'entrée (`*.inp`) a été conçu et implémenté. C'est un fichier texte qui contient des commandes spécifiques au module de polarisation qui sont appliquées les unes à la suite des autres par un interpréteur. Une interface graphique permet de créer et d'éditer ce fichier *input*, en plus de vérifier la validité des commandes contenues dans un fichier préexistant. La synthaxe imite celle des autres fichiers `*.inp` de *MoProSuite*, mais cet interprêteur de commandes est complètement séparé des autres interprêteurs du package. L'interface mérite une révision au moment de la rédaction, la création d'un nouveau fichier à partir de zéro est relativement laborieuse. Pour cette raison, éditer un fichier d'entrée valide est pour l'instant plus pratique.

Les commandes suivantes sont supportées par le module de polarisation de *MoProViewer*, le signe '#' représente toujours le numéro du fichier à polariser. Les commandes notées « critiques » sont strictement nécessaires au lancement de l'algorithme de polarisation (lié uniquement au mot-clé '`POLARIZE`', décrit plus bas). Si l'une de ces commandes « critiques » est manquante, le mot-clé '`POLARIZE`' est sans effet car les paramètres de calculs ne sont pas suffisamment définis. Cette limitation est purement informatique et garantit seulement que la commande '`POLARIZE`' est exécutable, pas que les résultats auront un quelconque sens physico-chimique (et il est tout à fait possible d'utiliser ce parseur sans utiliser '`POLARIZE`').

• '`PATH # `' suivi du chemin absolu du fichier moléculaire, au format `*.par*` de *MoPro* (*critique*) : permet de désigner le fichier à ouvrir. L'interface du module de polarisation permet d'ouvrir plusieurs fichiers et de générer les commandes correctes automatiquement.

• 'CYCLE 0' et 'ENDCYCLE' (*critique*) : ces commandes étaient originellement définies pour le support de boucles au sein du parseur. La complexité de l'implémentation actuelle a rendu difficile l'implémentation des boucles en un temps raisonnable. À l'exception des commandes 'PATH #' et 'ENDBELUGA', toutes les autres commandes doivent être situées entre 'CYCLE 0' et 'ENDCYCLE' pour être prises en compte.

• 'OBJSELE # ' (*critique*) : permet la définition du premier groupe d'atomes à polariser pour le fichier en question. Il est possible de sélectionner des acides aminé (mot-clé 'RESIDUE') ou des atomes (mot-clé 'ATOM'). Plusieurs modes de sélection sont disponibles. Le mot-clé 'NAME' permet la sélection d'atomes par leur élément chimique ou la sélection de résidus par leur nom tel qu'écrit dans le fichier moléculaire. 'NUM' permet de sélectionner des atomes par leurs numéros respectifs (le compte commence toujours à 0). Chaque numéro à sélectionner est séparé par un espace. 'RANGE' fonctionne de la même manière que 'NUM', sauf que l'utilisateur définit deux numéros, séparés par un tiret. Le module sélectionne tous les atomes ou résidus dont le numéro est compris entre ces nombres (bornes incluses). 'LAST' sélectionne le dernier atome ou le dernier résidu du fichier. Dans de nombreux fichiers PDB, les ligands sont écrits après la chaîne peptidique avec laquelle ils interagissent. 'ALLBUTLAST' sélectionne tout ce qui n'est pas sélectionné par 'LAST'. L'interface graphique permet de générer des commandes 'OBJSELE # ' et 'ACTSELE # ' (voir ci-dessous) à partir des atomes sélectionnés à la souris dans *MoProViewer*. Ces sélections doivent être stockées explicitement dans le fichier d'entrée, car une instance unique de *MoProViewer* doit pouvoir polariser toute une base de données de molécules.

• 'ACTSELE # ' (*critique*) : permet la définition du second groupe d'atomes à polariser pour le fichier en question. Le reste des arguments disponibles est strictement identique à 'OBJSELE # ', à l'exception du mot-clé 'ALLBUTOBJ' qui sélectionne tous les atomes (ou résidus) non sélectionnés par la commande 'OBJSELE # '.

• 'OBJEXCLUDE # ' et 'ACTEXCLUDE # ' : prennent les mêmes arguments que les deux commandes ci-dessus mais permettent d'exclure de ces sélections certaines entités (atomes ou acides aminés).

• 'SETOBJCHARGE' et 'SETACTCHARGE' : suivis d'un nombre à virgule, permet de changer la charge totale d'une des sélections en modifiant les populations de valence P_{val}.

Il est conseillé de réaliser cette étape en amont du calcul de la polarisation (après le transfert de paramètres).

• 'SETISOPOL' (*spécial : le fichier d'entrée n'est valide que si l'une ou l'autre des commandes 'SETISOPOL' ou 'SETANISOPOL' est présente*) : permet de définir les polarisabilités isotropes à utiliser pour le calcul. Plusieurs options sont disponibles. L'option 'ATOM' permet de définir une polarisabilité isotrope pour chaque élément chimique (un nombre par élément). Utiliser l'option 'ATOM' demande de définir une polarisabilité par défaut ('ELSE') pour les éléments chimiques non-listés dans la commande 'SETISOPOL'. L'option 'UIJ' permet de considérer le champ réservé aux paramètres de déplacement atomique du fichier molécule ('UISO' ou 'UANI') comme une polarisabilité isotrope. Chaque atome aura donc une polarisabilité propre, si elle est anisotrope, le tiers de la trace est utilisée. Le transfert de polarisabilités utilise ce champ et est abordé au chapitre 3. 'FIXED' suivit d'un nombre définit la même polarisabilité pour tous les atomes du calcul. Un mot-clé inaccessible 'P-ELMAM' a été prévu pour utiliser les polarisabilités de la base de données mais n'est pas encore supporté.

• 'SETANISOPOL' (*spécial : le fichier d'entrée n'est valide que si l'une ou l'autre des commandes 'SETISOPOL' ou 'SETANISOPOL' est présente)* : permet de définir les polarisabilités anisotropes à utiliser. Le mot-clé 'UIJ' est conseillé pour utiliser le champ 'UANI' du fichier molécule en tant que polarisabilités anisotropes orientées (strictement équivalent à la même option du mot clé 'SETISOPOL'). Les mots-clés 'ATOM' et 'FIXED' sont ici inactifs. Le mot-clé 'P-ELMAM', pour transférer les polarisabilités anisotropes à la volée, a été implémenté mais n'a pas été testé. Il est conseillé d'utiliser l'utilitaire de transfert pour écrire au préalable les polarisabilités dans le champ 'UANI'.

• 'DAMP' (*critique*) : définit le type de couplage entre les dipôles d'une même sélection. 6 options sont possibles. 'NOTIJ' force tous les blocs hors diagonaux de la matrice de polarisabilité moléculaire à être nuls. Il n'y a aucun couplage, les polarisabilités sont additives [voir équation (13)]. Les autres options réfèrent à l'équation (14). 'NODAMP' permet d'utiliser le tenseur de couplage $\mathbf{T}_{ij}$ sans fonction d'atténuation, tel que défini par Applequist[43]. Cette option est déconseillée (fort risque de divergence) à moins d'utiliser le set de polarisabilités atomiques affinées par Applequist lui-même. Les options 'LINEAR' et 'EXPO-THOLE' donnent accès aux méthodes d'atténuation publiées par Thole[46] puis van Duijnen et Swart[47]. L'option 'EXPO-XIE' permet d'accéder à la fonction d'atténuation

développée par Xie *et al.* pour le champ de force CHARMM[92]. Cette méthode est similaire à 'EXPO-THOLE'. Ces méthodes sont reproduites telles que publiées et les trois dernières nécessitent la définition d'une intensité d'atténuation (voir l'item suivant). La dernière option, 'STOPDI' ne nécessite pas de paramètre supplémentaire. Elle donne accès à la fonction d'atténuation de $\mathbf{T}_{ij}$ dérivée par Birge[48] pour les densités électroniques de type Slater (utilisées ici pour modéliser la couche électronique de valence).

• 'DAMPSTR' (*critique*) : permet de définir un ou plusieurs coefficients d'atténuation pour les méthodes 'LINEAR', 'EXPO-THOLE' et 'EXPO-XIE'. Seul l'argument 'FIXED' est fonctionnel, le développement logiciel ayant été concentré sur la méthode 'STOPDI' de Birge. Cette ligne est nécessaire même pour les modes sans coefficient d'atténuation.

• 'EFIELD' (*critique*) : définit la méthode de calcul des champs électriques. Deux options sont disponibles. 'PT' instancie *VMoPro* pour le calcul du champ électrique par la méthode EP/MM. 'ANALYTICAL' permet d'accéder au champ électrique calculé dans *MoProViewer* comme en utilisant une sommation rapide de type Buckingham[24,25,26], homologue à la partie longue portée du calcul EP/MM.

• 'SOLV' (*critique*) : La seule option disponible est 'FULLINVMAT' qui construit un tenseur inversé tel que dans l'équation (17). À l'origine, l'implémentation de méthodes de polarisation itératives (à l'intérieur d'une molécule) étaient envisagées pour éviter l'inversion du tenseur moléculaire de polarisabilité car Thole insistait sur la lenteur de cette étape… En 1981. L'efficacité de la librairie *Armadillo* et des machines actuelles ont rendu cette option inutile (l'inversion est très rapide et sera donc toujours effectuée).

• 'READALLMOL' : lit la liste de fichiers et stocke les molécules en mémoire. La liste (mise à jour) doit être lue à chaque cycle de polarisation, notamment pour une interface correcte avec *VMoPro* (argument 'PT' de 'EFIELD').

• 'INVSELE' : échange les rôles des sélections 'Objects' et 'Actors' qui respectivement reçoivent la polarisation et génèrent le champ électrique. Cette fonction assure que la classe assurant la logistique du calcul est dans le bon état pour éviter des appels hors-tableau. Cette fonction, essentielle au schéma de polarisation intermoléculaire itérative, n'est pas « critique » d'un point informatique et peut être appelée (ou non) n'importe quel nombre de fois (tant que les sélections sont définies).

• 'POLARIZE' : Cette fonction-clé vérifie que le module est prêt pour le calcul (toutes les options sont définies), vérifie que le calcul n'est pas convergé, puis lance le calcul des champs électriques, des polarisabilités moléculaires (si elles n'ont pas été calculées à un cycle précédent) et des moments dipolaires induits.

• 'WRITEPARAM' : transforme les dipôles induits dans le repère global en populations multipolaires exprimées dans les référentiels atomiques locaux.

• 'SAVEMOL' : écrit les fichiers moléculaires édités par 'WRITEPARAM' et met à jour la liste de chemins initialisée par les commandes 'PATH # '. Comme 'READALLMOL', cette fonction est essentielle à l'interface avec *VMoPro* pour que le champ électrique calculé au cycle suivant prenne la densité électronique polarisée en compte.

• 'ENER' : permet d'automatiser les calculs d'énergie avec *VMoPro* depuis le module de polarisation, toujours entre les sélections 'Objects' et 'Actors'. Le seul argument valide est 'ELEC' pour obtenir l'énergie d'interaction électrostatique par EP/MM.

• 'ATMPOLENER' : Cette fonction permet de calculer l'énergie de polarisation par la méthode tensorielle décrite partie 4.3. Cette fonction prend en compte les corrections développées pour le présent travail. Le calcul est effectué pour chaque atome et la somme apparaît dans l'interface et le fichier MoProVierwer.log.

• 'PHIPRIMEOVERS' : permet de renvoyer les résultats numériques des fonctions $f'/|\mathbf{u}_{ij}|$ et f'' telles que décrites dans la littérature[46,47,48] – et aux équations (29), (30), (33) et (34) – pour les modèles d'atténuation 'NODAMP', 'EXPO-THOLE' et 'STOPDI'. Les paramètres sont fixés (les polarisabilités isotropes, les coefficients atténuation et coefficients κ' sont unitaires). Le calcul court de 0.01 Å à 3 Å avec un pas de 0.02 Å.

• 'XTALPOT' : si le calcul de polarisation n'utilise qu'un seul fichier, il est possible d'ajouter au calcul de polarisation une influence externe au système. Ce mot-clé permet par exemple l'étude d'un champ cristallin calculé par d'autres moyens. Le seul argument possible est le chemin absolu vers une carte de potentiel électrostatique au format XPLOR. Le gradient (négatif) de cette carte de potentiel électrostatique sera pris en compte au premier cycle de polarisation comme une contribution au champ électrique (si la carte est assez grande pour contenir toutes les positions atomiques à polariser). Voir chapitre 4.2.

• ‘ENDBELUGA’ (*spécial : il n’y a aucune garantie de stabilité du module de polarisation après le calcul si ce mot-clé est absent)* : permet la terminaison correcte du module en mettant ses paramètres dans l’état de départ. Déclenche la sauvegarde du fichier de sortie.

Le caractère ‘!’ désigne les commentaires, toute chaîne de caractère suivant ‘!’ est ignorée. Certaines fonctions expérimentales n’ont pas été citées car leur développement a été abandonnées, elles sont devenues redondantes ou ne concernent que le débogage du module. Un exemple de fichier d’entrée est disponible en annexe, Encart A1.

2.3 - *Output* et visualisations

Maintenant que le système de contrôle du calcul a été implémenté, on se concentrera sur la formalisation des résultats du calcul de polarisation. Le fichier de sortie `PolarizationResults.out` contient le détail complet des résultats des calculs de polarisation, pour chaque (pro)molécule, chaque atome et chaque cycle de calcul itératif. Le formatage de ce fichier est toujours le même et lorsque plusieurs molécules sont polarisées en série, le fichier de sortie contient les résultats concaténés de toute la série.

Pour chaque fichier contenant les molécules en interaction, le résultat du calcul est séparé par sélections, ‘`Objects`’ ou ‘`Actors`’ (voir plus haut). Les champs suivants décrivent la polarisation de chacun des deux partenaires.

• La mention ‘`Selection : Objects`’ ou ‘`Selection : Actors`’, suivie du chemin local vers le fichier contenant la structure.

• Le tenseur de polarisabilité moléculaire (inversé et exprimé dans le référentiel cartésien du fichier, généré par orthogonalisation du référentiel cristallographique). Les lignes et les colonnes du tenseur portent chacune la mention de l’atome concerné [3 colonnes et 3 lignes par atome, voir équations (17) et (21)].

Ensuite, pour chaque atome du partenaire :

• Une ligne caractérisant l’atome avec (dans l’ordre) son numéro, son résidu, son nom, son élément chimique, sa charge multipolaire (en e) et les 6 valeurs indépendantes de la polarisabilité anisotrope (en $Å^3$).

• Un tableau regroupant les composantes *x*, *y*, *z* et la norme du moment dipolaire atomique $\boldsymbol{\mu}_i$ (en e·Å). Chaque colonne correspond à un cycle de calcul, on peut constater la convergence du calcul à mesure que $\boldsymbol{\mu}_i$ tend vers 0.

• Un tableau regroupant les composantes *x*, *y*, *z* et la norme du champ électrique $\mathbf{E}_i$ (en $e \cdot Å^{-2}$) appliqué à cet atome. Une colonne par cycle de calcul. La convergence est constatée à mesure que $\mathbf{E}_i$ tend vers une valeur constante.

• La ligne suivante regroupe à chaque cycle le produit scalaire $\mathbf{E}^t{}_i \cdot \boldsymbol{\mu}_i$ (en $e^2 \cdot Å^{-1}$) pour ce cycle et cet atome. Cette quantité est proportionnelle à une énergie. Si une valeur négative apparaît dans cette ligne, il existe un risque sérieux de divergence du calcul. En effet, une valeur négative signe le calcul d'un moment dipolaire de direction opposée au champ appliqué. Ces valeurs négatives sont en théorie impossibles si le couplage dipolaire via le tenseur $\mathbf{T}_{ij}$ est désactivé.

• Un tableau regroupant les composantes *x*, *y*, *z* et la norme du vecteur ($\boldsymbol{\mu}_i$ - $\boldsymbol{\alpha}_i \cdot \mathbf{E}_i$), en e·Å. Ce vecteur est la partie de $\boldsymbol{\mu}_i$ induite par le couplage intramoléculaire. À la précision de la machine près, toutes les valeurs de cette table sont nulles si le couplage $\mathbf{T}_{ij}$ est désactivé.

Une fois toutes les quantités atomiques listées :

• Un tableau regroupe les composant *x*, *y*, *z* et la norme du moment dipolaire moléculaire $\boldsymbol{\mu}_{mol}$ (en e·Å), induit à chaque cycle de calcul. Encore une fois, on constate la convergence à mesure que $\boldsymbol{\mu}_{mol}$ tend vers 0 (Figure 20).

Une fois que le groupe '`Objects`' et le groupe '`Actors`' sont documentés pour une molécule, le fichier détaille l'induction du fichier molécule suivant (si le calcul a été réalisé sur plusieurs complexes). Le fichier de sortie contient de nombreuses variables destinées uniquement à vérifier que le calcul s'est correctement déroulé.

Afin de simplifier la représentation du résultat, il est possible de découper manuellement la partie du fichier correspondant à un complexe particulier, puis de l'ouvrir à l'aide du manager d'affichage du module de polarisation.

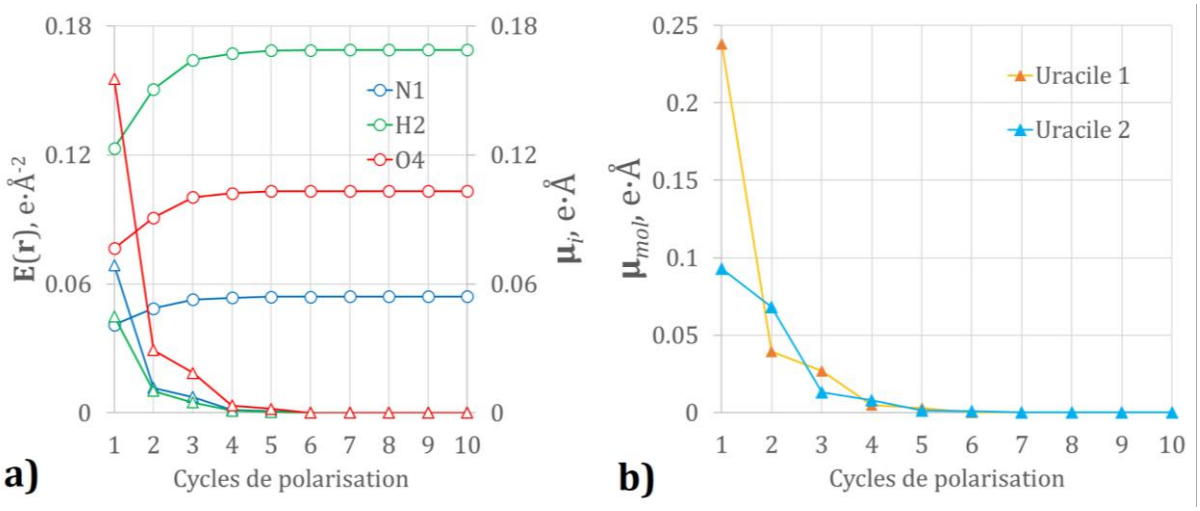


Figure 20. Au cours des cycles de polarisation, le champ électrique, calculé à la position **r** de chaque atome, se stabilise vers une valeur fixe. À chaque cycle, la contribution au moment dipolaire induit décroit vers 0. Panneau (a) : norme des champs électriques calculés sur trois atomes sélectionnés du dimère Uracile···Uracile en configuration paire de bases (cercles) et norme du moment dipolaire induit à chaque cycle sur ces atomes (triangles). Panneau (b) : moment dipolaire induit moléculaire ($\boldsymbol{\mu}_{mol} = \Sigma_i \boldsymbol{\mu}_i$) sur les deux uraciles du même système.

Le manager d'affichage – spécifique de l'outil Beluga – permet de représenter les différentes données vectorielles sous forme de flèches superposées à la représentation de la molécule dans *MoProViewer*. Le tenseur de polarisabilité moléculaire n'a pas de représentation évidente (à part sous sa forme réduite à une matrice de 3^2 éléments) mais les vecteurs $\boldsymbol{\mu}_i$, $\mathbf{E}_i$ et ($\boldsymbol{\mu}_i$ - $\boldsymbol{\alpha}_i \cdot \mathbf{E}_i$) sont facilement représentables par des flèches. Plusieurs présélections de représentations ont été implémentées. Il est possible de représenter la somme des vecteurs calculés à chaque cycle comme un seul vecteur, d'empiler les vecteurs pour représenter la somme vectorielle ou encore de les représenter en bouquet. Ces différents modes d'affichage sont illustrés Figure 21. Il est possible d'afficher tour les vecteurs associés à une molécule ou de sélectionner des atomes à la souris pour n'afficher que les moments et champs associés à la sélection (option « *Display only selected* »).

En plus du manager d'affichage dédié, il est possible – à l'aide de l'outil « *Map Operations* » préexistant dans *MoProViewer* – d'obtenir une de carte de $\rho_{pol}(\mathbf{r})$, la densité de déformation de polarisation, suivant l'équation (38), page 47. Pour ceci, on fait simplement la différence entre les densités électroniques totales du système avant et

après application de la procédure de polarisation (des exemples peuvent être trouvés Figure 21 et chapitre 3, partie 7.e).

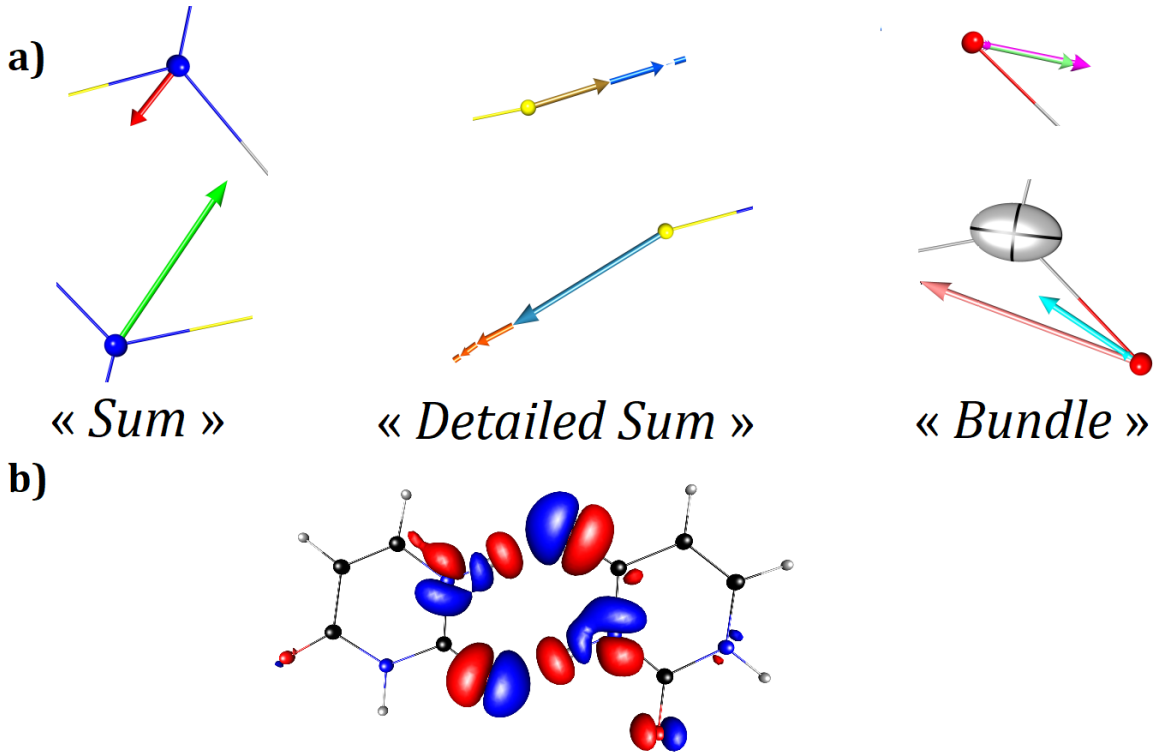


Figure 21. (a) : Exemples de représentations de grandeurs atomiques vectorielles avec *MoProViewer*. Différents atomes du dimère Uracile···Uracile en configuration paire de bases démontrent les affichages possibles pour les champs électriques et les moments induits calculés au cours de la polarisation. « *Sum* » utilise une seule flèche pour représenter la somme vectorielle totale – sur plusieurs cycles de calcul – de la grandeur à afficher ($\boldsymbol{\mu}_i$ ou $\mathbf{E}_i$). « *Detailed sum* » utillise une suite de vecteurs empilés pour représenter la grandeur représentée $\boldsymbol{\mu}_i$ ou $\mathbf{E}_i$ à chaque cycle. « *Bundle* » représente également le détail de $\boldsymbol{\mu}_i$ ou $\mathbf{E}_i$ à chaque cycle, sauf que tous les vecteurs partent du noyau de l'atome. Les option « *Detailed sum* » et « *Bundle* » utilisent la méthode de contraste automatique décrite par la suite pour souligner le vecteur associé au premier cycle de calcul. (b) : Exemple de densité déformation de polarisation $\rho_{pol}(\mathbf{r})$, obtenue à la convergence de l'algorithme sur le dimère Uracile···Uracile en configuration paire de bases (isosurfaces contourées à ±0.03 e·Å^{-3}).

La sélection de couleur personnalisées fait partie des options du manager d'affichage de Beluga. Une méthode automatique pour contraster le premier vecteur a été conçue (soit le moment dipolaire induit au premier cycle, soit le champ électrique ressenti au premier cycle sélectionné, option « *Contrast first vector* », voir plus bas). En effet, l'induction doit converger, ce qui signifie que si le champ appliqué au cycle n+1 est

supérieur au champ appliqué au cycle *n* sur tous les atomes, le système sera garanti de diverger (jusqu'à une induction infinie). Mettre en évidence le vecteur associé au premier cycle permet de vérifier que l'induction est maximale au premier cycle de calcul, puis décroit. Sur la Figure 21, cette option est utilisée sur représentations en bouquet (« *Bundle* ») et sur les représentations en vecteurs empilés (« *Detailed Sum* »).

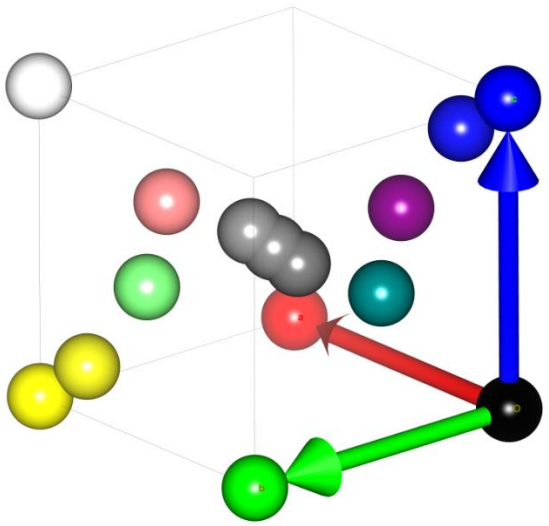


Figure 22. Le cube des couleurs, les sphères sont colorées selon leurs coordonnées. Chaque couleur présentée est en vis-à-vis de sa couleur complémentaire telle que calculée par *MoProViewer*. Le contraste est satisfaisant pour les couleurs au bord du cube. Perspective orthographique.

Les représentations 3D de *MoProViewer* utilisent *QGLViewer*[††], une librairie C++ qui permet d'interfacer le *framework* Qt pour la manipulation d'objets *OpenGL*. Les couleurs utilisent la norme RGBA. Chaque couleur primaire (rouge, vert, bleu) est représentée par un entier de 0 à 255. Les vecteurs sont ici tous opaques, on ignore la transparence (canal alpha, le 'A' de RGBA). Les couleurs accessibles peuvent donc être représentées dans un cube de côté 255 où chaque dimension représente une couleur primaire. Les coordonnées *x*, *y*, *z* de chaque point à l'intérieur du cube correspondent aux composants rouge, vert et bleu d'une couleur affichable à l'écran. On souhaite obtenir le maximum de contraste pour le premier, on cherche donc la couleur complémentaire à la

[††] : version 2.7.1, http://libqglviewer.com/

couleur sélectionnée. Pour la trouver, on a simplement choisi le symétrique de la couleur sélectionnée par rapport au centre du cube (rouge, 127.5 ; vert, 127.5 ; bleu, 127.5 : le centre du cube est gris). Le résultat est convaincant pour des couleurs assez éloignées du centre (au moins une coordonnée proche de 0 ou 255). Le contraste est moins bon si l'utilisateur sélectionne un gris car la couleur retournée sera très proche de la couleur sélectionnée. Chaque coin du cube est également éloigné du centre et offre un contraste maximal – une distance dans le cube – avec le gris. Cette méthode ne permet donc pas de trouver un contraste satisfaisant avec un gris.

2.4 - « Polarization of electron density databases of transferable multipolar atoms » *Partie 2 : Polarisation et Énergies*

Cette partie a été intégralement extraite de l'article « Polarization of Electron Density Databases of Transferable Multipolar Atoms », parue le 11 juillet 2019 dans The Journal of Physical Chemistry A, par Théo Leduc, Emmanuel Aubert, Enrique Espinosa, Christian Jelsch, Cristian Iordache et Benoît Guillot. Cet article est parfaitement intégré au présent travail et sa publication représente un achèvement important pour le présent travail de thèse. Afin de faciliter la lecture, les éléments rapportés ont été traduits et réorganisés pour s'ajuster au plan choisi. Aucun élément n'a été ajouté afin que toutes les informations présentées dans cette partie restent telles que revues par les pairs.

2.4.a - Interaction électrostatique permanente

Le module de polarisation de *MoProViewer* a donc été appliqué au jeu de données S66, présenté partie 1.7.a. Le but de la présente partie est d'évaluer les performances énergétiques du modèle de densité électronique transférée-polarisée. Comme indiqué par l'équation (41), page 48, la décomposition SAPT de l'énergie d'interaction totale entre deux densités de charges moléculaires donne accès à des contributions physiquement interprétables. Au sein de la décomposition SAPT, la contribution électrostatique permanente E_{ELEC} et la contribution d'induction E_{IND} peuvent être comparés aux grandeurs correspondantes calculées à partir de la densité de charge transférée-polarisée. Dans un premier temps, le terme SAPT E_{ELEC} de l'équation (41) est comparé à l'énergie d'interaction électrostatique [$E_{ELEC,DB}$, équation (10)] obtenue à partir de la

densité électronique ELMAM2 transférée sur les dimères des jeux de données S66 et S66x8 (Figure 22).

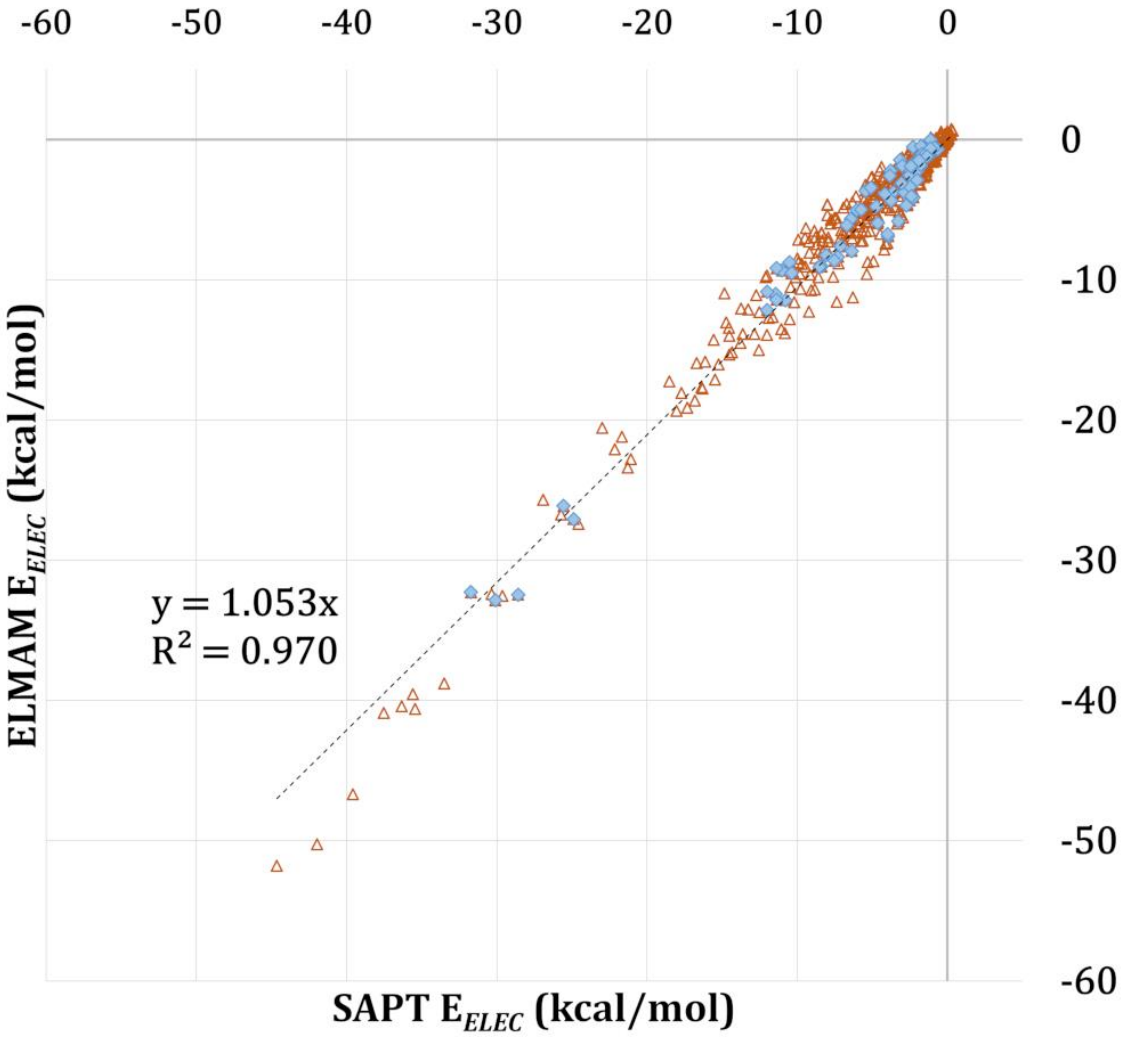


Figure 22. Corrélation entre SAPT E_{ELEC} et ELMAM $E_{ELEC,DB}$ sur le set S66 (losanges bleus, distances d'équilibre) et sur le jeu de données S66x8 (triangles oranges, toutes distances). La régression linéaire affichée porte sur toutes les distances (S66x8). Aux distances d'équilibre, la régression linéaire donne un coefficient de détermination de 0,973 et une pente de 1,038. Ces deux régressions linéaires ont été contraintes à avoir une ordonnée à l'origine nulle.

Avant toute chose, il convient de rappeler ici que ces deux énergies (E_{ELEC} de SAPT et E_{ELEC} d'ELMAM2) sont calculées par des moyens très différents : E_{ELEC} est un résultat purement théorique tandis que $E_{ELEC,DB}$ est obtenue en utilisant des paramètres de densité électronique, transférés et d'origine expérimentale. Les seuls points communs entre les

deux approches sont donc : les géométries des dimères considérés et l'utilisation, dans le formalisme multipolaire, de densités sphériques de type Hartree-Fock, pour représenter les électrons de cœur et de valence [équation (6), page 27].

Néanmoins, pour les dimères du S66 à distance d'équilibre, la régression linéaire entre les valeurs de E_{ELEC} et de $E_{ELEC,DB}$ montre une excellente concordance, avec un coefficient de détermination R^2 = 97.3% (coefficient de corrélation R de 98.64%) et une pente proche de l'unité, avec $E_{ELEC} = 1.04 \cdot E_{ELEC,DB}$ (voir Table 6, page 127). Une régression similaire a été effectuée sur l'ensemble des données S66x8, ce qui a conduit à un accord légèrement inférieur, ici R^2 = 97.0 % et la pente vaut 1.05.

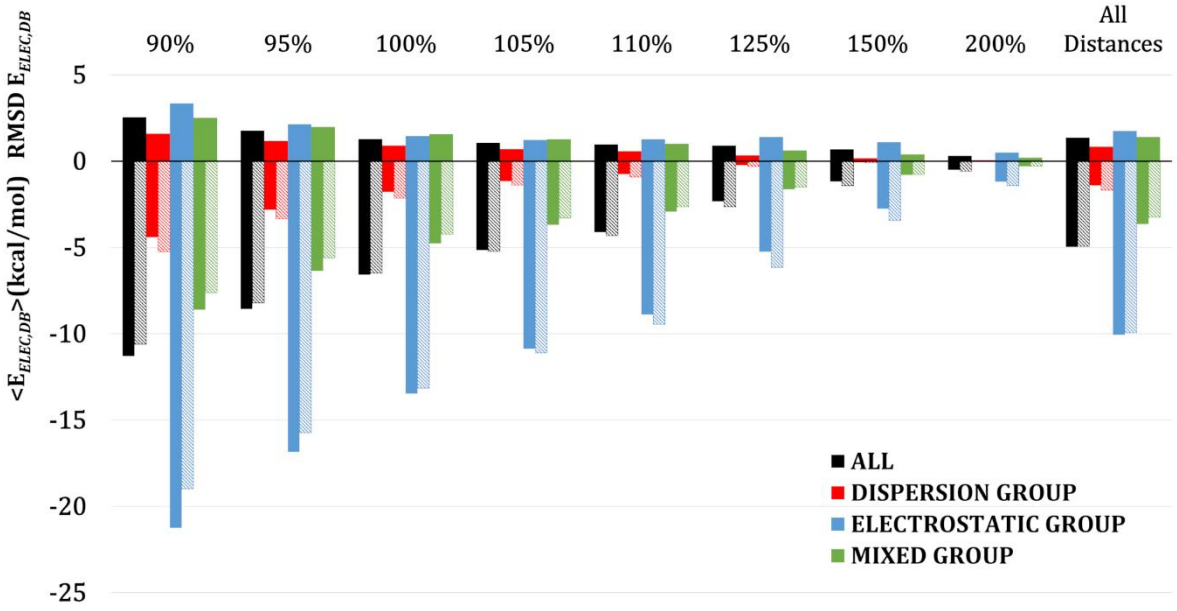


Figure 23. Energies d'interactions électrostatiques moyennes SAPT, $\langle E_{ELEC} \rangle$, et entre densités électroniques transférées, $\langle E_{ELEC,DB} \rangle$. Les dimères sont triés par distance relative à la distance d'équilibre et par groupes : électrostatique, dispersion et mixte. Les barres au-dessus de la ligne du zéro représentent le *RMSD* associés à ces énergies moyennes. En effet, toutes les valeurs $\langle E_{ELEC,DB} \rangle$ et $\langle E_{ELEC} \rangle$ sont négatives, et affichées sous cette ligne. Les barres de couleurs foncées représentent la valeur $\langle E_{ELEC,DB} \rangle$. Les barres de couleurs claires représentent les valeurs de référence $\langle E_{ELEC} \rangle$ de la partition SAPT.

Une erreur accrue dans l'estimation de E_{ELEC} pour les distances d'interaction les plus courtes explique ce comportement (Figure 23). En effet, en dessous de 110 % de la distance d'équilibre, il apparaît que le terme d'énergie $E_{ELEC,DB}$ surestime l'intensité de

l'interaction électrostatique permanente (par rapport à la référence SAPT). L'écart quadratique moyen (*RMSD*) atteint 2.6 kcal·mol^{-1} pour la plus courte distance (90 %, Figure 23). Tenant compte de toutes les distances d'interaction, le *RMSD* entre les énergies ELMAM2 et SAPT est de 1.4 kcal.mol^{-1}, ce qui peut être considéré comme un bon résultat étant donné les différences intrinsèques des deux méthodes et l'absence de toute paramétrisation *ad hoc*. Li et al. 2014[58] ont proposé une classification des dimères de l'ensemble S66 en fonction de la contribution la plus importante à l'énergie d'interaction totale (électrostatique, dispersive ou un mélange des deux), définissant ainsi trois groupes de dimères nommés : « électrostatique », « dispersion » et « mixte ». En considérant ces 3 groupes d'interaction, le *RMSD* à la distance d'équilibre est de 1.3 kcal·mol^{-1}. Le *RMSD* est inférieure à la précision "symbolique" de 1.0 kcal·mol^{-1} au-dessus de 110% de la distance d'équilibre.

Le groupe dispersion regroupe notamment des molécules présentant des cycles aromatiques en interaction π-π. Ce groupe présente le meilleur accord absolu entre les énergies ELMAM2 et SAPT (Figure 23). Pour les distances d'interaction les plus courtes, le *RMSD* est de 1.6 kcal·mol^{-1}, mais il chute à 0.9 kcal·mol^{-1} pour les géométries à l'équilibre (comparable au *RMSD* de 0.9 kcal·mol^{-1} calculé pour toutes les distances).

Inversement, le groupe électrostatique présente un accord plus faible en valeurs absolues : le *RMSD* à la distance la plus courte est de 3.4 kcal·mol^{-1} et de 1.5 kcal·mol^{-1} à distance d'équilibre. Calculé sur toutes les distances, le *RMSD* est 1.8 kcal·mol^{-1}. Comme attendu, le groupe mixte se comporte de manière intermédiaire. Si on considère les erreurs relatives cependant, il faut prendre en compte que le groupe électrostatique implique des énergies d'interaction globalement plus fortes ($\langle E_{ELEC} \rangle$ = -13.5 kcal·mol^{-1}) que le groupe dispersion ($\langle E_{ELEC} \rangle$ = -1.8 kcal·mol^{-1}) et le groupe mixte ($\langle E_{ELEC} \rangle$ = -4.8 kcal.mol^{-1}). Le groupe électrostatique est celui qui présente donc les erreurs relatives absolues moyennes (*MARE*) les plus basses : 10 % à distance d'équilibre. Le groupe dispersion et le groupe mixte, en interaction à distance d'équilibre, présentent des *MARE* significativement plus grandes, atteignant 33% dans les deux cas.

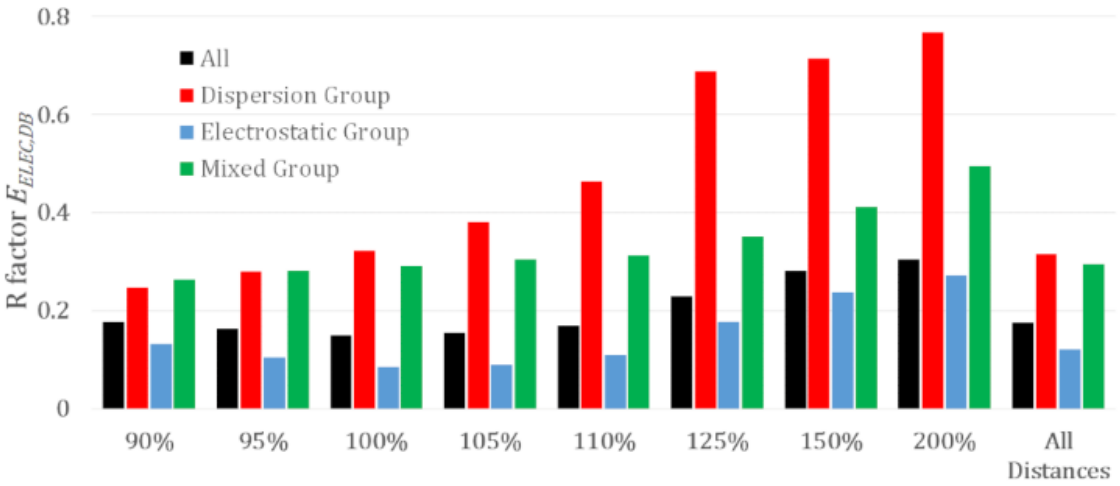


Figure 24. Répartition du facteur *R* [équation (A8), page 292 de l'annexe] calculé pour $E_{ELEC,DB}$ sur le jeu de données S66x8 (et comparé la contribution E_{ELEC} de SAPT), par distance relative à la distance d'équilibre et par groupes d'interactions. Barres noires : ensemble des trois groupes ; barres rouges : groupe dispersion ; barres bleues : groupe électrostatique ; barres vertes : groupe mixte.

Le *MARE* n'est pas le seul moyen d'évaluer les erreurs relatives, il est aussi possible d'utiliser le facteur *R* de l'équation (A8), page 292 de l'annexe. Dans la formule du *MARE* [équation (A6), page 291], l'erreur sur chaque calcul est relative à la magnitude du résultat attendu pour ce calcul. Dans la formule du facteur *R*, l'erreur cumulée est relative à la magnitude de la somme de tous les résultats de référence. Le facteur *R* est donc moins sensible aux valeurs extrêmes, en particulier aux grandes variations relatives associées à de petites valeurs de référence, là où le *MARE* est en revanche très sensible aux grandes variations relatives quelque soit la valeur de référence. Néanmoins, la Figure 24 permet de tirer des conclusions similaires à celles tirées de l'analyse du *MARE* : le groupe électrostatique, qui présente les plus fortes énergies d'interaction $E_{ELEC,DB}$, offre les meilleures prédictions relatives. Le groupe dispersion offre les moins bonnes prédictions relatives, et ceci est d'autant plus visible que les dimères interagissant à longue distance sont considérés ($R > 0.6$ pour ce groupe quand la distance d'interaction est au dessus de 110 % de la distance d'équilibre). On retrouve des performances intermédiaires pour le groupe mixte.

Table 6. Comparaison des régressions linéaires entre les énergies de référence SAPT et les énergies électrostatiques calculées à partir d'ELMAM2. Les dimères S66 (distances d'équilibre) ont été utilisés pour la détermination de $E_{ELEC,DB}$ et $E_{POL,DB}$ par EP/MM et en utilisant les polarisabilités théoriques des monomères du S66. L'énergie SAPT est ajustée linéairement à l'énergie ELMAM2, de sorte que E(ELMAM2) = pente × E(SAPT).

	SAPT, électrostatique (E_{ELEC})	SAPT électrostatique + SAPT induction (E_{ELEC} + E_{IND})
ELMAM2 électrostatique ($E_{ELEC,DB}$)	R^2 = **0.973** pente = **1.038** *RMSD* = **1.3** kcal.mol^{-1}	R^2 = 0.974 pente = 0.721 *RMSD* = 4.0 kcal.mol^{-1}
ELMAM2 électrostatique + induction ($E_{ELEC,DB}$ + $E_{POL,DB}$)	R^2 = 0.950 pente = 1.455 *RMSD* = 5.0 kcal.mol^{-1}	R^2 = **0.960** pente = **1.014** *RMSD* = **2.2** kcal.mol^{-1}

Il convient de mentionner que la même approche a été suivie par Kumar *et al.*[93] qui ont comparé l'énergie d'interaction électrostatique SAPT de premier ordre à l'énergie d'interaction E_{ELEC} obtenue par la méthode EP/MM sur des densités électroniques reconstituées à partir des pseudoatomes multipolaires de la librairie UBDB. Pour l'ensemble de données S66 (distances d'équilibre), ils ont obtenu des statistiques globales légèrement meilleures (*RMSD* = 1.1 kcal·mol^{-1}, R^2 = 98% avec une pente de 1.03). Mais il faut prendre en compte que Kumar *et al.* comparent les énergies d'interaction électrostatique théoriques SAPT à des valeurs E_{ELEC} dérivées de facteurs de structure théoriques. En effet, les paramètres des pseudoatomes de la bibliothèque UBDB ont été obtenus par calcul DFT (*basis set* 6-31G** + polarisation) sur de petites molécules extraites de la *Cambridge Structural Database*[16], en conservant la géométrie déterminée par cristallographie. Des facteurs de structure théoriques à haute résolution ont ensuite été obtenus par transformée de Fourier des densités électroniques théoriques, disposées en pseudo-cristaux cubiques avec une maille de 30 Å. Enfin, les affinements multipolaires contre ces facteurs de structure théoriques ont permis d'obtenir les paramètres

multipolaires moyens nécessaires à la construction de la bibliothèque UBDB. Cette procédure conduit à des paramètres de densité électronique transférables construits à partir de molécules pseudo-isolées. Les énergies d'interaction électrostatique EP/MM, calculées à partir des paramètres ELMAM2, sont fortement corrélées (coefficient de corrélation R = 98.8%) avec les E_{ELEC} calculées à partir des densités électroniques UBDB. Encore une fois, ce fait suggère que les pseudoatomes ELMAM2 sont dépourvus des effets environnementaux relatifs à l'environnement cristallin d'où les paramètres sont initialement extraits. C'est également le cas lorsque l'on compare $E_{ELEC,DB}$ à la somme E_{ELEC}+E_{IND} (l'énergie électrostatique totale SAPT plus l'énergie d'induction), ce qui conduit à un accord détérioré par rapport à la comparaison de $E_{ELEC,DB}$ avec la contribution SAPT de premier ordre, E_{ELEC} (voir Table 6). La reproduction de la contribution E_{IND} de SAPT est examinée à la partie 2.4.e.

2.4.b - Transfert de charge et autres contributions à l'induction de moments dipolaires

En utilisant la procédure de calcul décrite ici, huit cycles de polarisation ont permis d'atteindre la convergence des moments dipolaires induits, pour tous les dimères du jeu de données S66x8. Comparé au présent modèle d'induction dipolaire, il est évident que le terme d'induction totale de la méthode SAPT (utilisée comme énergie de référence, E_{IND}) couvre plus de contributions, telles que les effets de transfert de charge ou l'induction de moments électrostatiques d'ordre supérieur. Il a été récemment démontré, par exemple, que les contributions énergétiques liées au transfert de charge intermoléculaire et liées à la polarisation peuvent être modélisées à partir de l'énergie d'induction SAPT[57]. Cette étude porte cependant sur des distances d'interaction plus courtes que celles considérées ici (jusqu'à 70% des distances d'équilibre). La mise en évidence et la quantification du flux de charges électroniques à l'intérieur d'une molécule (transferts de charges intramoléculaires) dépendent de la manière dont les charges atomiques sont définies. Dans les calculs théoriques présentés ici, certaines charges topologiques peuvent varier significativement lorsque l'on compare les charges dérivées de la densité électronique du monomère à celles dérivées de la densité du dimère. En moyenne, dans les dimères, les atomes d'oxygène ont une charge topologique inférieure de 0.026 e, comparé à leurs charges dans la géométrie isolée (avec une densité électronique théorique). L'effet moyen

sur les autres espèces chimiques est nettement moins visible. Le déplacement de densité électronique d'un bassin atomique à un autre induira des moments électrostatiques, de sorte que cet effet contribue réellement au dipôle moléculaire induit total.

Le modèle proposé, de polarisation de la densité électronique transférée, se concentre uniquement sur l'induction de dipôles atomiques. La partition QTAIM des densités ELMAM2 transférées et des densités polarisées a été obtenue à l'aide du logiciel *BADER*[94,95,96]. Les charges et les volumes topologiques varient suite à la modification de la densité électronique lors de l'induction. L'induction dipolaire provoque le déplacement des surfaces de flux nul qui séparent la densité électronique des pseudo-atomes multipolaires. Paradoxalement, des effets de transfert de charge topologique existent alors que, dans la densité ELMAM2 tranférée polarisée, les charges multipolaires sont fixes.

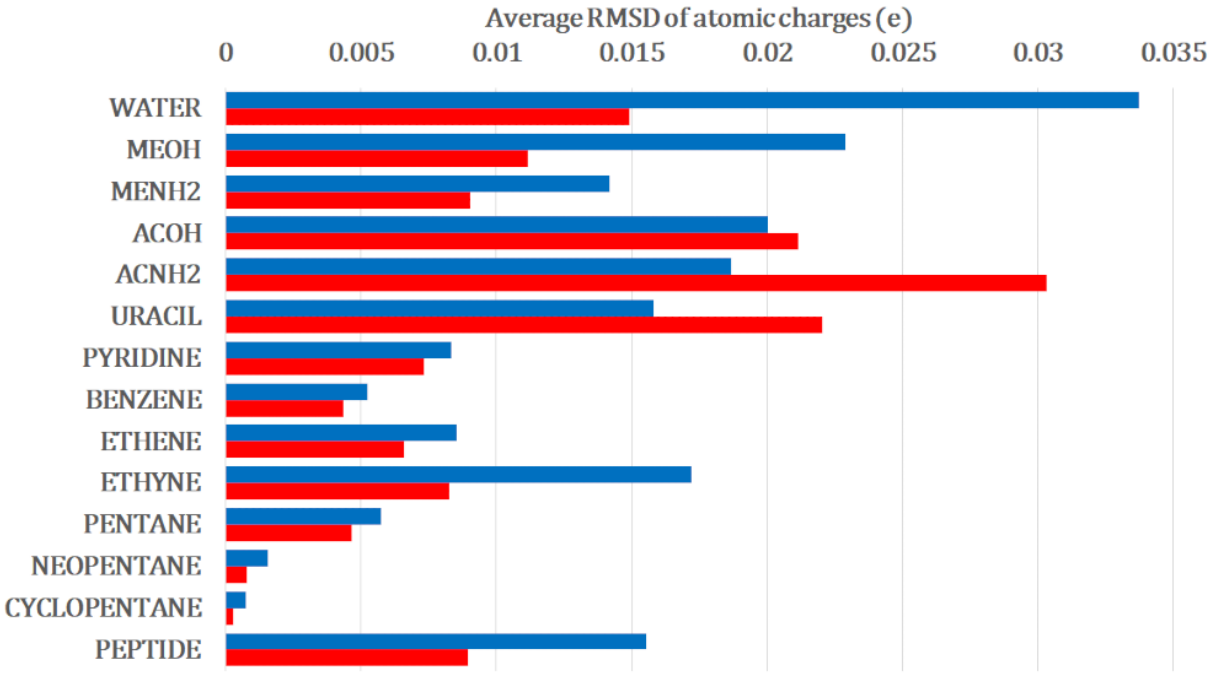


Figure 25. *RMSD* moyen des charges topologiques, entre les états dimère et monomère. Bleu: densité électronique théorique, partitionnée par le logiciel *AIMAll*. Rouge: densité électronique ELMAM2 polarisée, partitionnée par le logiciel *Bader*.

La Figure 25 détaille ce transfert de charge topologique pour les différents monomères présents dans le jeu de données S66. Il est en effet possible d'accéder à deux

« versions » de ce transfert de charge, celui tiré des partitions QTAIM de la densité électronique théorique et le transfert de charge, également tiré de la partition topologique de la densité, mais résultant uniquement de la modification des paramètres dipolaires transférés. La Figure 25 compare, molécule par molécule, le *RMSD* moyen [équation (A4), page 291] des différences de charges topologiques entre les systèmes monomères et dimères. Les densités électroniques ELMAM2 (polarisées) et théoriques sont considérées en parallèle. On cherche un indicateur de l'amplitude moyenne du flux de charge intramoléculaire lors de la formation des dimères. Les 14 molécules considérées dans le jeu de données S66 ont des amplitudes similaires de transfert de charge lors de la comparaison des densités électroniques ELMAM2 polarisées et théoriques. Dans les deux cas, le transfert de charge moyen est compris entre 0.001 e et 0.034 e. De manière prévisible, le transfert de charge le plus faible est observé pour les composés aliphatiques et le plus important pour les molécules polaires. Le coefficient de détermination global R^2 est 39.4 % pour les transferts de charges topologiques, comparant les densités transférées et théoriques. La pente est de 0,742. Cette tendance suggère que le modèle de polarisation proposé rend compte, au moins partiellement, de l'effet de transfert de charge tel que mis en évidence sur la base de charges topologiques, intégrées sur le ρ-bassins. La prise en compte explicite de cet effet dépasse le cadre de la méthode présentée ici. Il serait possible de modéliser cet effet en altérant les populations de valence pour représenter un transfert de charge intramoléculaire.

Dans le modèle HCM, des moments dipolaires moléculaires émergent de la somme des dipôles atomiques induits. La disposition de pseudoatomes non-neutres dans l'espace contribue aussi au dipôle moléculaire. Pour étudier l'ampleur de cet effet, en termes de moments dipolaires et au niveau de la partition topologique de $\rho(\mathbf{r})$, une approximation de charge ponctuelle est proposée. La charge nette topologique – intégrée sur chaque bassin atomique – est concentrée sur chaque position nucléaire et, pour chaque molécule, le dipôle de ces ensembles de charges a été calculé (au centre de masse). Pour la densité électronique théorique, les dipôles induits par la dimérisation – et calculés à partir de ces charges ponctuelles – vont de 10^{-4} D à 1,05 D, pour une moyenne de 0,26 D. La densité électronique ELMAM2, partitionné avec le logiciel Bader, subit le même traitement. En considérant ces charges topologiques concentrées en charges ponctuelles, les dipôles induits par la dimérisation varient de 10^{-3} D à 1,82 D, pour une moyenne de 0,15 D. Ces dipôles apparaissent importants comparés aux dipôles induits totaux présentés Figure 27

(plus bas) mais l'approximation du présent paragraphe n'est pas strictement comparable à un calcul de moment dipolaire qui prend en compte les paramètres de la densité électronique.

Le *RMSD* entre les magnitudes de ces dipôles – calculés à partir de charges ponctuelles, entre les densité théoriques et transférées-polarisées – est de 0,48 D, ce qui représente une forte dispersion comparée aux valeurs moyennes du paragraphe précédent. Dans l'acide acétique, l'acétamide et l'uracile, l'approximation des charges ponctuelles génère des dipôles induits comparables entre la partition théorique et ELMAM2. Cet agrément est beaucoup plus faible pour les autres monomères. L'acide acétique, l'uracile et l'acétamide ont tous des groupements chimiques à la fois polaires et polarisables. Une partie de l'effet de transfert de charges topologiques est donc correctement prise en compte. Cependant, une prise en compte explicite du transfert de charge (par modification des populations P_{val}) pourrait améliorer les performances topologiques du modèle de densité électronique transférée-polarisée. Le modèle d'induction de dipôles couplés ($\mathbf{T}_{ij}$) devrait modifier la magnitude des transferts de charge topologiques présentés Figure 25. Utiliser le tenseurs $\mathbf{T}_{ij}$ pour modéliser l'induction demandera à l'avenir de réévaluer la partition QTAIM de la densité polarisée et la comparer à nouveau avec la partition QTAIM de la densité électronique théorique.

Un autre moyen d'évaluer ces effets de polarisation au niveau des charges du modèle multipolaire est possible. Affiner une densité HCMM directement sur la densité électronique théorique de dimères en interaction refléterait évidemment les effets de la polarisation intermoléculaire. Ces effets seraient présents en termes de populations de valence atomiques modifiées (par rapport aux systèmes monomères) et également en termes de moments électrostatiques induits d'ordre supérieur, dont les dipôles. Cependant, il faut relativiser l'ampleur du phénomène : les amplitudes des transferts de charges topologiques observées ici sont faibles. Ils ont même ici une magnitude en général inférieure aux incertitudes statistiques, obtenues dans une étude récente sur les charges dérivées des populations topologiques de valence[97]. Le transfert de charge théorique maximum est de +0.165 e. Cela suggère que, même si les effets de transferts de charges intramoléculaires étaient explicitement modélisés, la perturbation requise par l'altération des populations de valence atomique transférées pourrait s'avérer peu significative par rapport aux incertitudes sur le paramètre correspondant. Néanmoins,

cette conclusion n'est relative qu'à la densité électronique et ne signifie pas que l'impact énergétique de ces transferts de charge serait négligeable. Par conséquent, nous ne nous attendons pas à ce que l'induction de moments dipolaires dans un modèle HCM reproduise parfaitement l'énergie d'induction SAPT. C'est pourquoi ces énergies théoriques ne sont utilisées qu'à des fins de comparaison : aucun paramètre du modèle de polarisation n'est ajusté pour reproduire les valeurs SAPT de référence.

La prise en compte explicite de la polarisation due au transfert de charge intra- ou intermoléculaire par modification des populations de valence atomique des pseudoatomes HCM n'entre pas dans le cadre de cette étude. Krawczuk, Macchi et collaborateurs[4,49,86] proposent une méthode d'évaluation de la translation des charges qui n'a pas été employée ici, puisqu'elle implique l'évaluation de contributions topologiques plus fines - et non atomiques – à savoir la polarisation au niveau des points critiques de liaison dans la densité électronique, comme décrit page 103. Si la densité électronique ELMAM2 a la capacité de reproduire ces caractéristiques[63], ce sujet mérite définitivement une étude dédiée.

2.4.c - Énergies d'induction : détermination de la meilleure méthode de calcul des polarisabilités

Les valeurs d'énergie de polarisation ELMAM2 [équation (37)] sont négatives pour tous les dimères. Cette énergie de polarisation $E_{POL,DB}$, dérivée des paramètres ELMAM2 et utilisant la méthode EP/MM, est comparée à l'énergie d'induction E_{IND} de SAPT dans les Figures 26 et 28. Pour calculer ces $E_{POL,DB}$, il est possible de choisir entre les polarisabilités des molécules isolées ou les polarisabilités calculées dans chaque système dimère. L'effet énergétique – à distances d'équilibre – des polarisabilités atomiques associées aux systèmes dimères est comparé sur la figure 26, par la corrélation entre $E_{POL,DB}$ et les énergies E_{IND} de référence (SAPT). Neuf dimères de la S66 ont été exclus de cette analyse en raison de l'absence de polarisabilité des dimères (voir page 81). En utilisant les polarisabilités des systèmes dimères, l'énergie $E_{POL,DB}$ surestime la référence SAPT, E_{IND}, comme montré par une pente de régression à 1.132. Les coefficients de détermination R^2 de la Figure 26 indiquent des performances similaires pour les polarisabilités des dimères (R^2 = 0,859) et des monomères (R^2 = 0,864). La pente de régression entre SAPT E_{IND} et

$E_{POL,DB}$ utilisant les polarisabilités calculées dans les systèmes monomères est de 0,953. Le *RMSD* associé est de 1.6 kcal·mol^{-1} pour les 57 dimères considérés. Pour les énergies $E_{POL,DB}$ utilisant des polarisabilités des dimères, le *RMSD* associé est légèrement plus élevé, à 1.8 kcal·mol^{-1}.

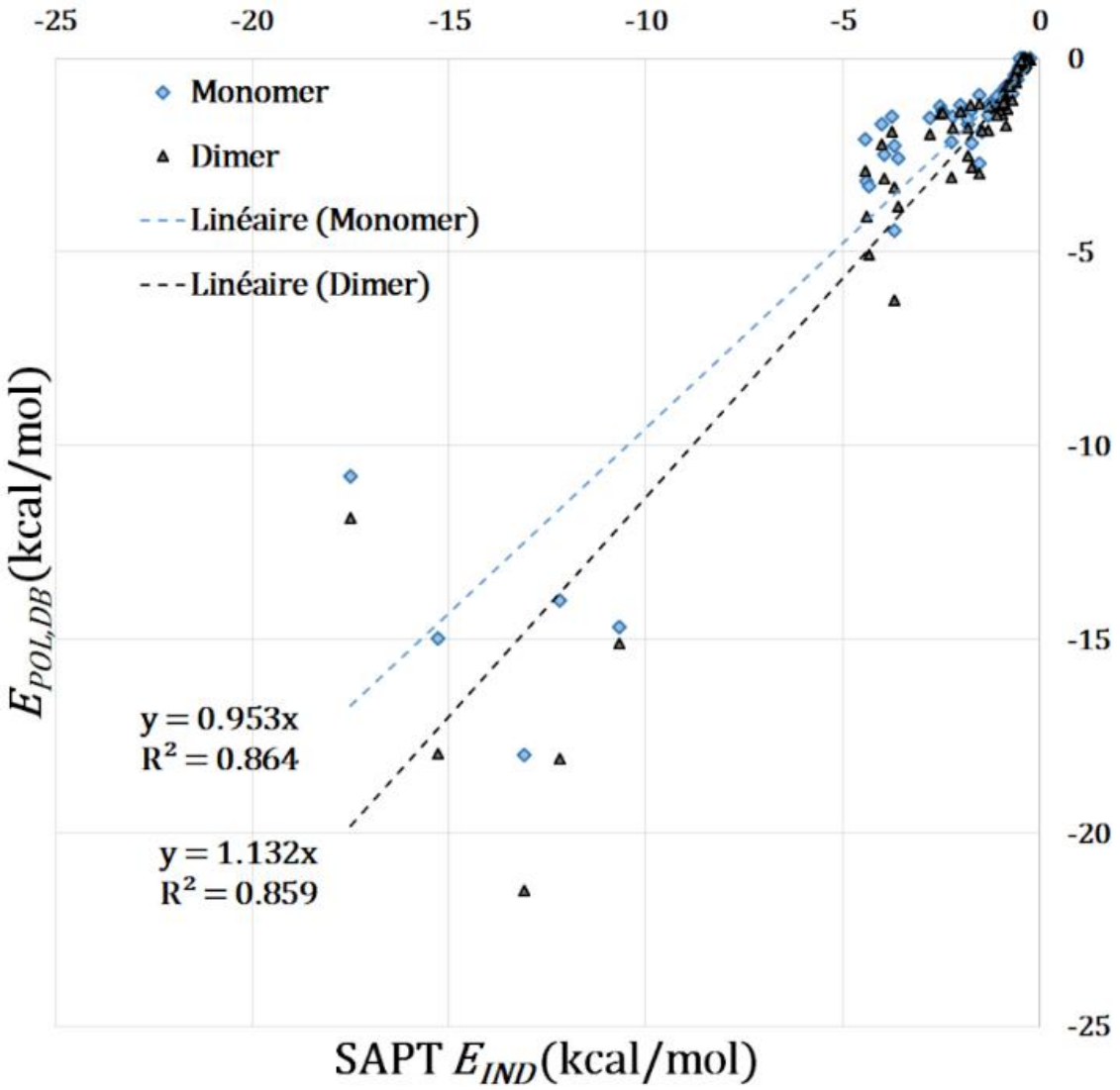


Figure 26. La corrélation entre E_{IND} (SAPT) et $E_{POL,DB}$, calculée aux distances d'équilibre avec les polarisabilités des systèmes monomères est représentée par les losanges bleus. Les triangles noirs matérialisent la corrélation entre E_{IND} (SAPT) et $E_{POL,DB}$, calculée en utilisant les polarisabilités des systèmes dimères. Les deux régressions n'incluent que les 57 dimères pour lesquels les polarisabilités des molécules isolées et des dimères sont disponibles et utilisent $E_{POL,DB}$ = pente × E_{IND}.

L'utilisation de polarisabilités calculées dans les systèmes dimères donne des énergies $E_{POL,DB}$ qui sont en moyenne 19 % plus intenses, par rapport aux $E_{POL,DB}$ obtenues avec les polarisabilités calculées dans les molécules isolées. Les différences entre les valeurs propres des polarisabilités des états monomère et dimère (Figure 18, partie 1.7.h, page 98) sont faibles. Pourtant, cette surestimation de E_{IND} est importante et donc, inattendue. Il est probable que les $E_{POL,DB}$ exagérées, obtenues avec les polarisabilité des dimères, indiquent une relation entre ces polarisabilités et une densité électronique déjà perturbée par l'interaction avec l'autre monomère. En conséquence, par la suite, les calculs de l'énergie de polarisation utiliseront uniquement les polarisabilités calculées à partir des monomères.

2.4.d - Dipôles induits moléculaires moyens

Avant de traiter le calcul des énergies de polarisation $E_{POL,DB}$ pour les 528 dimères du set S66x8, on s'intéresse aux dipôles induits à distance d'équilibre, après l'application de la méthode de polarisation. La Figure 27 présente, pour chaque molécule, le dipôle induit moléculaire moyen, obtenu à distance d'équilibre. Les polarisabilités atomiques des monomères (voir partie 1.7.b) sont utilisées pour calculer l'induction en vertu de leurs meilleures performances énergétiques. Contrairement à l'approximation basée sur des charges ponctuelles, proposée partie 2.4.b, ici, les dipôles induits sont calculés à l'aide de *VMoPro* et prennent donc en compte l'intégralité de l'expansion multipolaire de chaque molécule. Ces dipôles induits sont parfaitement concordants avec ceux obtenus dans le fichier de sortie du module de polarisation.

Le dipôle moléculaire induit le plus intense est obtenu pour l'uracile n°2 dans l'homodimère d'uracile (en conformation *BP*, « paire de bases »), avec 1.50 D (moment dipolaire permanent de l'uracile : 6.20 D). Le processus de polarisation produit une augmentation moyenne de +19.5% du moment dipolaire moléculaire de l'uracile sur tous les dimères qui en contiennent. L'acétamide est également affecté d'intenses dipôles induits, 0.97 D en moyenne (moment dipolaire permanent de l'acétamide : 5.09 D). L'enrichissement moyen en moment dipolaire sur tous les dimères contenant de l'acétamide est de +16.1%. L'acétamide est principalement impliqué dans les dimères avec lui-même ou l'uracile. Ces deux molécules ont de forts moments dipolaires et sont

hautement polarisables. Une autre molécule contenant un amide (*N*-méthyl-acétamide) présente également une induction élevée, comparable à celle calculée pour l'acide acétique.

L'induction dipolaire calculée la plus faible a été observée dans la molécule d'éthyne en interaction avec le pentane, avec $4.5 \cdot 10^{-3}$ D. Le dipôle moyen induit sur le cyclopentane est de $16.0 \cdot 10^{-3}$ D (le pentane et le cyclopentane ont tous les deux un dipôle permanent de 0,05 D). De même, on observe également une induction de faible intensité dans le néopentane. La molécule de pentane (quasi-apolaire) en raison de sa forme allongée, présente une magnitude moyenne de dipôle induit comparable à celle de petites molécules hydrophiles (eau, méthylamine et méthanol).

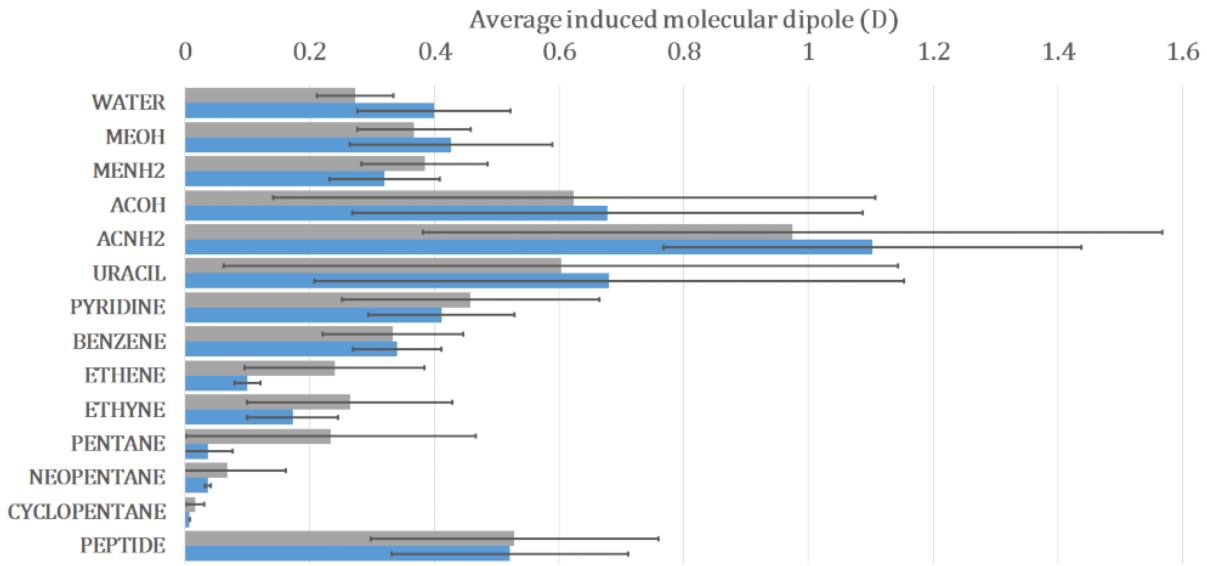


Figure 27. Dipôles moléculaire moyens induits sur la densité électronique ELMAM2, dans les dimères du jeu de données S66. Gris: norme moyenne des dipôles induits sur la molécule considérée (en D). Bleu: norme moyenne du dipôle induit sur le partenaire, par la molécule considérée (en D). Les polarisabilités utilisées sont celles des monomères. Les barres d'erreur représentent le *SSD* de chaque dipôle moyen.

Par contre, le pentane induit des dipôles beaucoup plus faibles sur ses partenaires, comparé à l'eau, la méthylamine ou le méthanol. Parmi les molécules aromatiques, le benzène porte et induit des dipôles plus faibles que la pyridine, qui porte et induit des dipôles plus faibles que l'uracile (qui contient des groupes polaires N-H et C=O). Il est

intéressant de constater que dans les systèmes dimériques étudiés, les moments induits sur une molécule $\boldsymbol{\mu}_{i,mol}$ et les moments induits par cette molécule sur ses partenaires, $\boldsymbol{\mu}_{i,partenaire}$, sont très corrélés. En effet, la régression $\boldsymbol{\mu}_{i,partenaire} = 1.038 \cdot \boldsymbol{\mu}_{i,mol}$ est associée à un R^2 de 0.906.

2.4.e - Calculs d'E_{POL} et E_{IND} sur les ensembles de données S66 & S66x8

La meilleure estimation de E_{IND} par $E_{POL,DB}$ est donc obtenue en polarisant la densité électronique transférée à l'aide des polarisabilités des monomères. Comme montré Figure 28, la régression linéaire entre ces deux quantités donne une pente de 0.955 et un coefficient de détermination R^2 = 0.865 (set S66 complet, distances d'équilibre uniquement). En utilisant cette méthode sur les 66 dimères, le *RMSD* entre E_{IND} et $E_{POL,DB}$ reste à 1.6 kcal·mol^{-1} (les statistiques de la partie précédente excluaient les 9 dimères dont les polarisabilités *in situ* n'étaient pas définies positives, voir page 81). L'énergie de polarisation moyenne ELMAM2 $\langle E_{POL,DB} \rangle$ est de -2.6 kcal·mol^{-1} aux distances d'équilibre, ce qui compare à une moyenne $\langle E_{IND} \rangle$ de -2.9 kcal·mol^{-1} pour la méthode SAPT. Ceci souligne la légère sous-estimation de E_{IND} par $E_{POL,DB}$, précédemment vue à l'examen de la pente. Néanmoins, le modèle de densité électronique transférée-polarisée reproduit une part importante des contributions énergétiques incluses dans le terme E_{IND}. Le plus grand écart entre $E_{POL,DB}$ et E_{IND}, en valeur absolue, est obtenu pour l'homodimère de l'acide acétique, avec une sous-estimation de 6,7 kcal·mol^{-1} de E_{IND} (-38.3 %). Le dimère *N*-méthyl-acétamide···Éthène donne quant à lui la meilleure prédiction, E_{IND} est sous-estimé de seulement 0.017 kcal·mol^{-1} (-2.3 %).

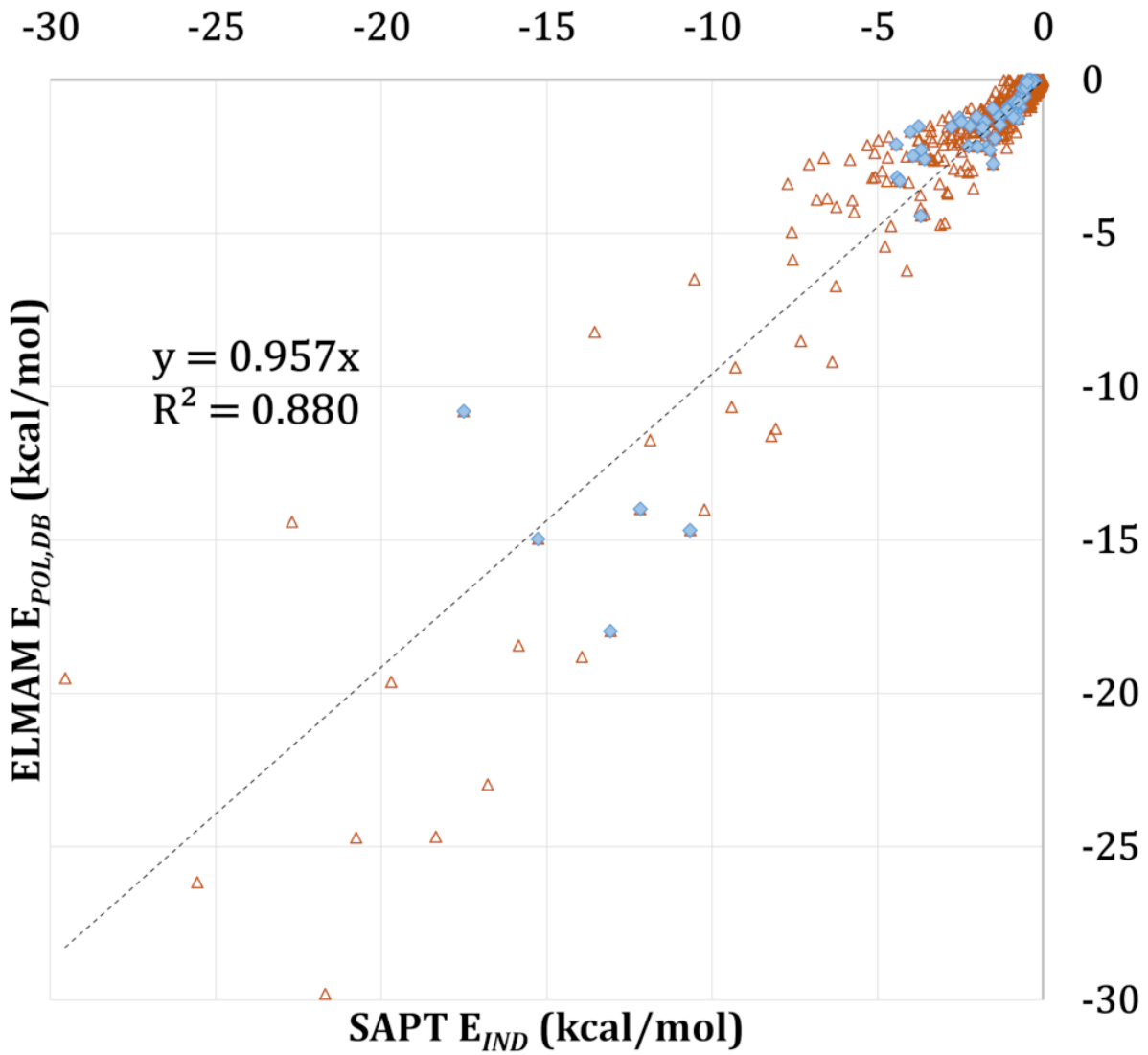


Figure 28. Corrélation entre E_{IND} (méthode SAPT) et $E_{POL,DB}$ pour le set S66 (distances d'équilibre, losanges bleus) et sur le set S66x8 (gamme complète de distances documentées, triangles oranges). Les polarisabilités calculées dans les systèmes monomères sont utilisées. La régression linéaire présentée a été effectuée sur l'ensemble de données S66x8, en utilisant l'équation $E_{POL,DB}$ = pente × E_{IND}. Pour les distances d'équilibre seulement, le coefficient de détermination est de 0.865 et la pente est de 0.955.

La Figure 28 résume les énergies de polarisation à distance d'équilibre, en utilisant les meilleurs paramètres de calcul préalablement déterminés (polarisabilités des monomères). Si l'on considère maintenant toutes les distances d'interaction incluses dans le jeu de données S66x8[68], $E_{POL,DB}$ et la référence SAPT, E_{IND}, sont corrélées avec un coefficient de détermination R^2 = 0.880 et une pente de régression de 0.957. Le *RMSD* sur toutes les distances est de 1.3 kcal·mol^{-1} pour un $E_{POL,DB}$ moyen de -1.6 kcal·mol^{-1} (pour ce

calcul, les énergies d'induction à longue distance sont très faibles). L'erreur maximale, en valeur absolue, est à nouveau calculée pour l'homodimère d'acide acétique, mais cette fois-ci pour le dimère en interaction à 90 % de la distance d'équilibre. E_{IND} est sous-estimé de 10.1 kcal·mol^{-1} pour ce dimère (-34.0 %). Les erreurs absolues minimales sont retrouvées dans les dimères où l'induction est minimale (typiquement : Cyclopentane···Néopentane à 200% de la distance d'équilibre). Pour de telles molécules apolaires interagissant à ces longues distances, tous les dipôles induits calculés sont inférieurs au seuil de convergence ($|\boldsymbol{\mu}_i| < 10^{-4}$ e·Å). Il n'y a alors pas de modification significative de la densité électronique, la densité polarisée est strictement la même que la densité transférée depuis ELMAM2. L'erreur relative la plus faible se produit de façon intéressante dans le dimère Acide acétique···Uracile, en interaction à 95% de la distance d'équilibre. $E_{POL,DB}$ dans ce cas est 0.068 kcal·mol^{-1} en dessous de E_{IND} (-0.35%), alors que ces deux molécules sont à la fois très polaires et polarisables.

En fonction des groupes de dimères proposés par Li *et al.*[58], la performance du modèle de polarisation varie. La Figure 29 présente les énergies de polarisation moyennes (ELMAM2 et SAPT) pour chaque classe d'interaction – électrostatique, dispersion et mixte – et pour chaque distance relative à l'équilibre. Aux distances d'équilibre, la moyenne $\langle E_{POL,DB} \rangle$ varie d'un groupe à l'autre. L'énergie de polarisation moyenne la plus faible a été calculée pour le groupe de dispersion, où $\langle E_{POL,DB} \rangle$ = -0,56 kcal·mol^{-1}. Les groupes mixtes et électrostatiques présentent des valeurs $\langle E_{POL,DB} \rangle$ plus élevées, respectivement -1.3 kcal·mol^{-1} et -4.8 kcal·mol^{-1}. Le modèle proposé reproduit bien la dépendance à la distance de l'énergie de polarisation, même si elle tend à sous-estimer E_{IND} aux courtes distances. La pente de régression aux distances d'équilibre est de 0.955, et des valeurs comparables sont obtenues à 90% et 95% de la distance d'équilibre (respectivement 0.951 et 0.950). Cela confirme la légère sous-estimation de E_{IND} à courte distance. A de plus grandes distances, l'amplitude des énergies d'interaction diminue drastiquement, par exemple, la référence $\langle E_{IND,200\%} \rangle$ est de 0.035 kcal·mol^{-1}. En effet, l'énergie de polarisation[46] diminue en r^{-3} de la distance interatomique. À 200% de la distance d'équilibre, la pente de régression monte jusqu'à 1.729 pour la régression $E_{POL,DB}$ = pente × E_{IND}.

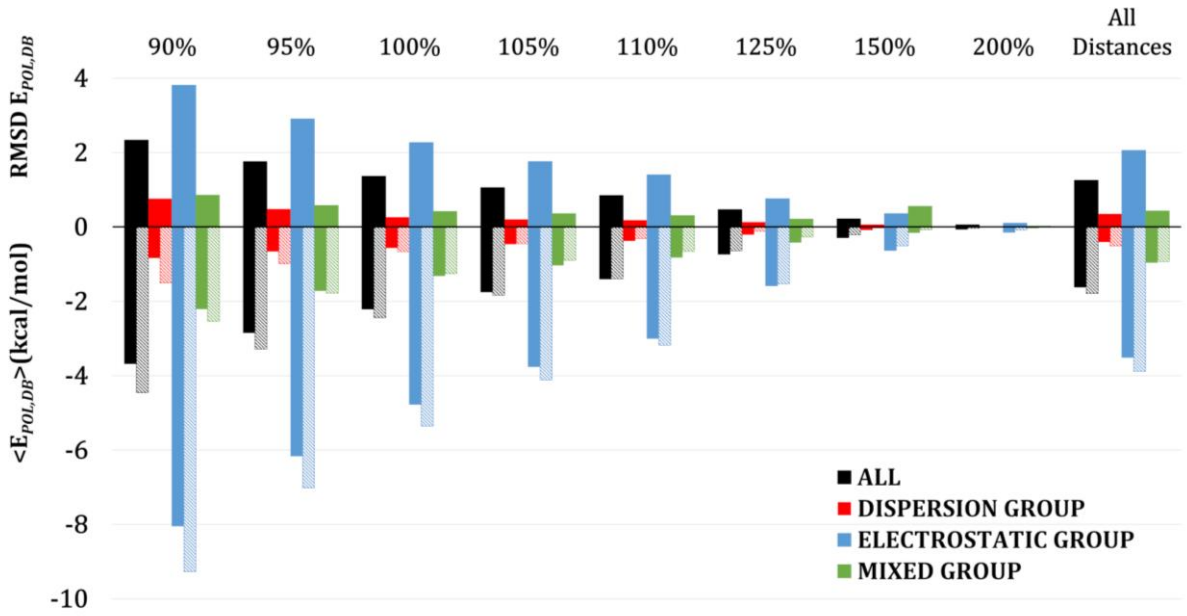


Figure 29. Energies de polarisation moyennes de la densité transférée-polarisée, $\langle E_{POL,DB} \rangle$, et SAPT, $\langle E_{IND} \rangle$. Les valeurs sont classées par distance relative à la distance d'équilibre et par groupes de dimères. Les valeurs RMSD entre $\langle E_{POL,DB} \rangle$ et $\langle E_{IND} \rangle$ sont au-dessus de la ligne du zéro. Toutes les valeurs $\langle E_{POL,DB} \rangle$ et $\langle E_{IND} \rangle$ sont négatives et représentées en dessous de cette ligne. Les barres de couleurs foncées représentent les valeurs dérivées de la polarisation des densités électronique ELMAM2, $\langle E_{POL,DB} \rangle$. Les barres de couleurs plus claires représentent les valeurs $\langle E_{IND} \rangle$ de la référence SAPT.

L'évaluation de la performance de la méthode est intéressante à la lumière des différents types d'interactions entre dimères. La Figure 29 regroupe également les valeurs de *RMSD* entre les $E_{POL,DB}$ en provenance d'ELMAM2 et les E_{IND} de SAPT, selon les trois mêmes classes d'interaction et les huit distances différentes. Le *RMSD* suit la même tendance que les énergies de polarisation. Les groupes présentant une induction élevée, comme les dimères du groupe électrostatique à une distance de 90 %, ont un *RMSD* élevé (3.8 kcal·mol^{-1}) tandis que les dimères du groupe mixte ou du groupe dispersion ont un *RMSD* inférieur au seuil symbolique de 1.0 kcal·mol^{-1}, et ce pour toutes distances. La Figure 29 montre également que les écarts absolus les plus élevés entre $\langle E_{POL,DB} \rangle$ et $\langle E_{IND} \rangle$ sont observés pour les dimères du groupe électrostatique, qui présentent également les énergies de polarisation les plus importantes. Par contre, l'erreur relative n'est pas la plus grande pour le groupe électrostatique : aux distances d'équilibre, la *MARE* pour ce groupe est de 32.6 % (comparable à la *MARE* calculée sur le S66 entier : 34.5 %). Le groupe mixte obtient les meilleurs résultats du point de vue des erreurs relatives (à distance

d'équilibre, la *MARE* de ce groupe mixte est de 22.2 %). Inversement, à distance d'équilibre, la moins bonne *MARE* (43.5%) est obtenue pour le groupe dispersion (où les énergies d'induction sont les plus faibles parmi les trois groupes).

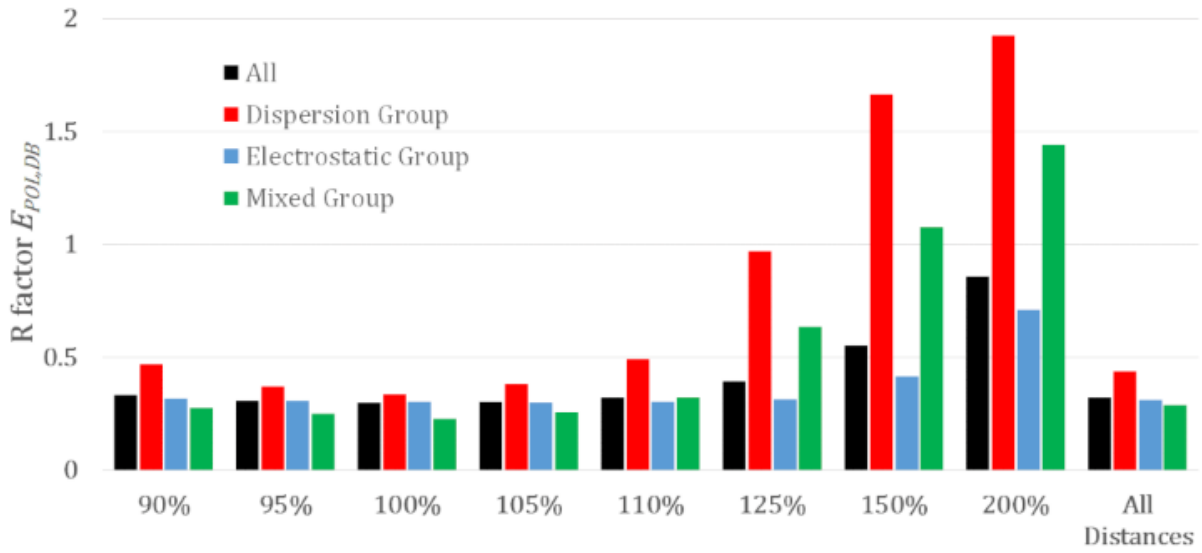


Figure 30. Répartition du facteur *R* [équation (A8), page 292] calculé entre $E_{POL,DB}$ sur le jeu de données S66x8 (comparé à E_{IND}), par distance relative à la distance d'équilibre et par groupes d'interactions. Barres noires : ensemble des trois groupes ; barres rouges : groupe dispersion ; barres bleues : groupe électrostatique ; barres vertes : groupe mixte.

Comme souligné page 126, le facteur *R* permet, tout comme le *MARE*, d'apprécier les performances relatives du modèle. La Figure 30 regroupe les valeurs de *R* entre les énergies d'induction $E_{POL,DB}$ et E_{IND}. Il apparaît immédiatement que toutes les valeurs R pour les énergies d'induction sont supérieures à celles de la figure S11 pour l'énergie d'interaction électrostatique. Ceci est dû à une magnitude moyenne inférieure des énergies de polarisation. Encore une fois, le facteur *R* augmente pour les longues distances intermoléculaires, à mesure que l'induction devient négligeable. Les performances des groupes apparaissent bien plus homogènes que lorsque les *MARE* sont analysé. On retrouve cependant le même ordre des groupes pour leurs performances relatives : le groupe mixte a les plus petites erreurs relatives, et le groupe dispersion les plus grandes. Le groupe électrostatique est associé à des performances intermédiaires.

2.4.f - Somme des contributions électrostatiques permanentes et induites

L'objectif est, à terme, de proposer une méthode permettant d'accéder à l'énergie d'interaction électrostatique totale (tenant compte des contributions permanentes et induites) à partir d'une distribution électronique transférée. Il est donc correct de comparer la somme E_{ELEC}+E_{IND} de la méthode SAPT à la quantité équivalente obtenue à partir du la densité électronique transférée-polarisée de ELMAM2. La Figure 31 montre cette comparaison, utilisant les polarisabilités des monomères pour calculer l'énergie de polarisation. Aux distances d'équilibre uniquement, la moyenne $\langle E_{ELEC,DB}+E_{POL,DB}\rangle$ est -8.8 kcal·mol^{-1} et le *RMSD* est de 2.2 kcal·mol^{-1}, comme le montrent les paramètres de régression reportés Table 6. La meilleure erreur absolue est obtenue pour le dimère Pentane···Uracile, avec une différence d'énergie électrostatique plus induction de 0.019 kcal·mol^{-1} comparé au calcul SAPT. Si l'on considère la même somme, uniquement aux distances d'équilibre, l'erreur la plus importante est obtenue pour le dimère Acétamide···Uracile, avec une énergie inférieure de 8.8 kcal·mol^{-1} à la référence SAPT.

Si on considère maintenant toutes les distances d'interaction, l'erreur la plus importante est à nouveau observée pour le dimère Acétamide···Uracile, cette fois à 90 % de la distance d'équilibre. La somme des deux énergies ELMAM2 est 15.2 kcal·mol^{-1} sous la référence SAPT (-67.5 kcal·mol^{-1}). La pente globale de l'ensemble des données S66x8 est proche de l'unité (1.027) et le coefficient de détermination est de 0.965, ce qui est comparable à l'accord entre $E_{ELEC,DB}$ et E_{ELEC} SAPT, et sensiblement meilleur que l'accord entre $E_{POL,DB}$ seule et E_{IND} seule, comme le montrent la Figure 31 et la Table 6.

Il est important de noter que l'accord entre $E_{ELEC,DB}$+$E_{POL,DB}$ et son homologue théorique est meilleur que celui obtenu lorsque la densité électronique transférée est utilisé pour reproduire l'énergie électrostatique théorique totale. Cette régression, $E_{ELEC,DB}$ = pente × (E_{ELEC}+E_{IND}), est documentée Table 6 et donne un R^2 similaire = 0.974. Cependant, la pente diffère avec 0.721 et le *RMSD* est très important : 4.0 kcal·mol^{-1}. Ceci indique que la procédure de polarisation proposée ici est effectivement nécessaire pour complémenter l'énergie d'interaction électrostatique calculée à l'aide la densité électronique ELMAM2 transférée mais non-perturbée. La densité transférée est insuffisante pour modéliser fidèlement tous les effets résultant des interactions entre les densités de charge moléculaire. En effet, à distance d'équilibre, $\langle E_{POL}\rangle$ représente 33.6 % de $\langle E_{ELEC}\rangle$. Cette proportion est retrouvée dans les trois groupes de dimères définis par Li

et al.[58] (groupe dispersion : $\langle E_{POL}\rangle/\langle E_{ELEC}\rangle$ = 31.4 % ; groupe électrostatique : $\langle E_{POL}\rangle/\langle E_{ELEC}\rangle$ = 35.5 % ; groupe mixte : $\langle E_{POL}\rangle/\langle E_{ELEC}\rangle$ = 27.6 %). Il faut noter que les paramètres ELMAM2, ainsi que ceux d'autres librairies de pseudoatomes multipolaires transférables, ont déjà été utilisées pour estimer des énergies d'interaction électrostatique, par exemple dans les complexes protéines-ligands[12,13,98,99,100]. Dans ces études, seule la contribution électrostatique permanente à l'énergie d'interaction totale a été prise en compte, négligeant la contribution due à la polarisation intermoléculaire. Cette dernière peut maintenant être estimée grâce à la méthode proposée ici. $E_{POL,DB}$, calculée à l'aide de la méthode présentée de polarisation de la densité électronique, sous-estime légèrement E_{IND} tandis que $E_{ELEC,DB}$ est associée à une faible surestimation des E_{ELEC} de référence.

Les pentes des régressions linéaires illustrées à la Figure 31 et à la Table 6 suggèrent que ces erreurs peuvent se compenser sur la somme $E_{ELEC,DB}$+$E_{POL,DB}$. Ceci pourrait être le résultat d'une polarisation reminiscente de la densité électronique ELMAM2 transférée. L'induction de dipôles sur la densité électronique ELMAM2 s'ajoute en effet à des moments électrostatiques préexistants. La présence de ces moments pourrait expliquer la légère surestimation de la contribution électrostatique permanente (ELMAM2 conserverait une partie des moments électrostatiques induits par le champ cristallin dans les études originales de densité de charge, voir Table A2). Cet effet reste à confirmer. S'il existe, il est trop faible pour être objectivable lors de l'analyse des erreurs sur les calculs présentés.

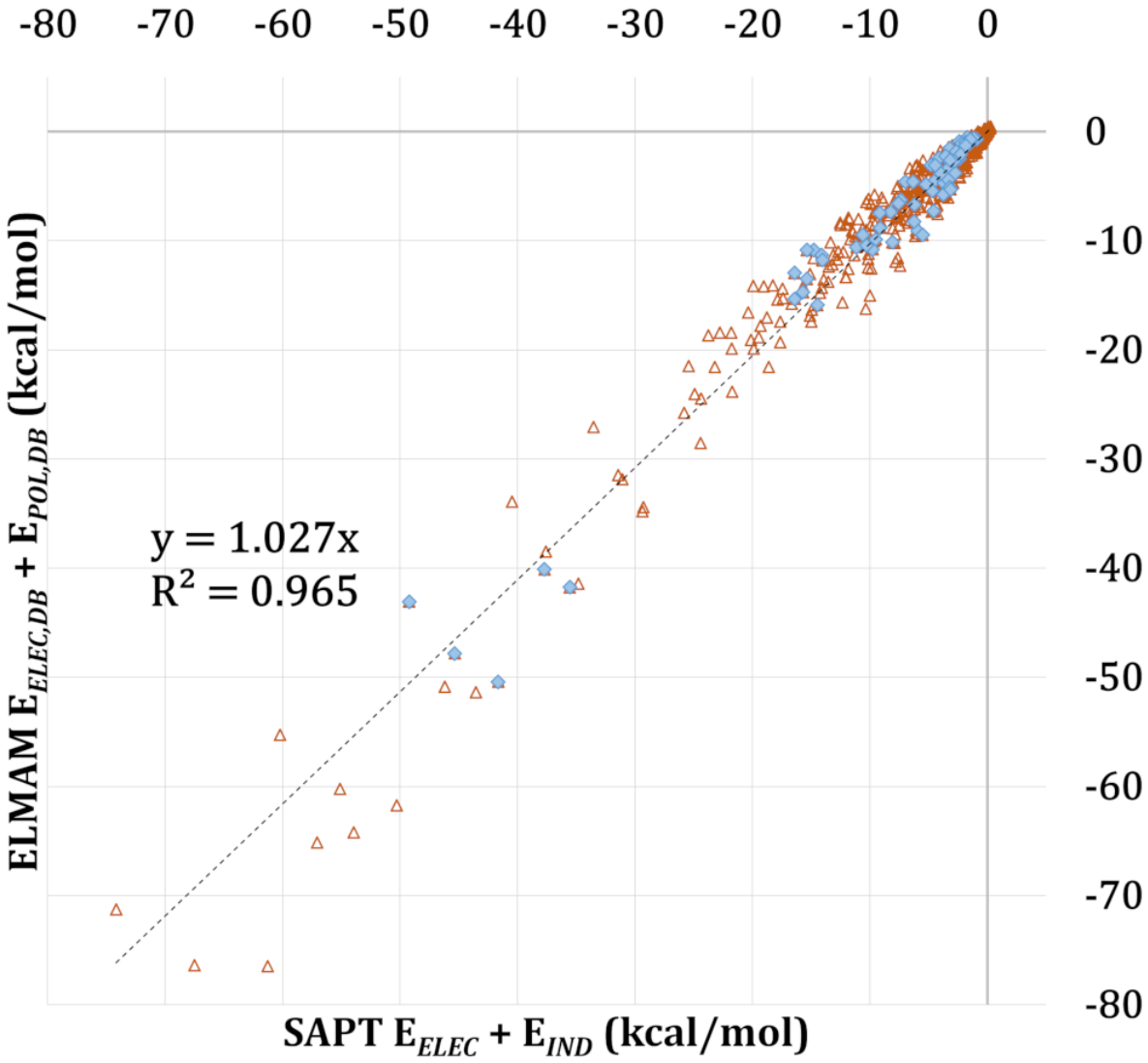


Figure 31. Corrélation entre la somme E_{ELEC}+E_{IND} (référence théorique SAPT) et $E_{ELEC,DB}$+$E_{POL,DB}$ (ELMAM2) sur le jeu de données S66. Les distances d'équilibre sont représentées par les losanges bleus. L'ensemble des distances représentées dans le jeu de données S66x8 sont matérialisées par des points orange. Les polarisations calculées dans les systèmes monomères ont été utilisées pour le calcul de la polarisation. La régression linéaire porte sur toutes les distances ; le coefficient de détermination basé uniquement sur les données à distance d'équilibre se trouve dans la Table 6.

Les prévisions les moins bonnes de l'ensemble S66 concernent l'acétamide, l'uracile et l'acide acétique. Ces molécules impliquent des fonctions chimiques polaires comportant des doubles liaisons conjuguées. Cependant, tous les dimères impliquant ces molécules ne sont pas associés à des erreurs importantes. Ces atomes présentent des polarisabilités relativement élevées, de sorte que l'on s'attendrait à une amélioration du

modèle avec la mise en œuvre et la paramétrisation du modèle des interactions dipolaires intramoléculaires atténuées[46], voir page 41 et partie 4.5.

2.4.g - Répartition des erreurs pour les calculs d'énergie

La figure 32 montre l'effectif de groupes présentant des erreurs absolues similaires, par tranches de 1.0 kcal·mol^{-1}. Des enveloppes gaussiennes de même moyenne et de même écart-type sont représentées pour servir de référence.

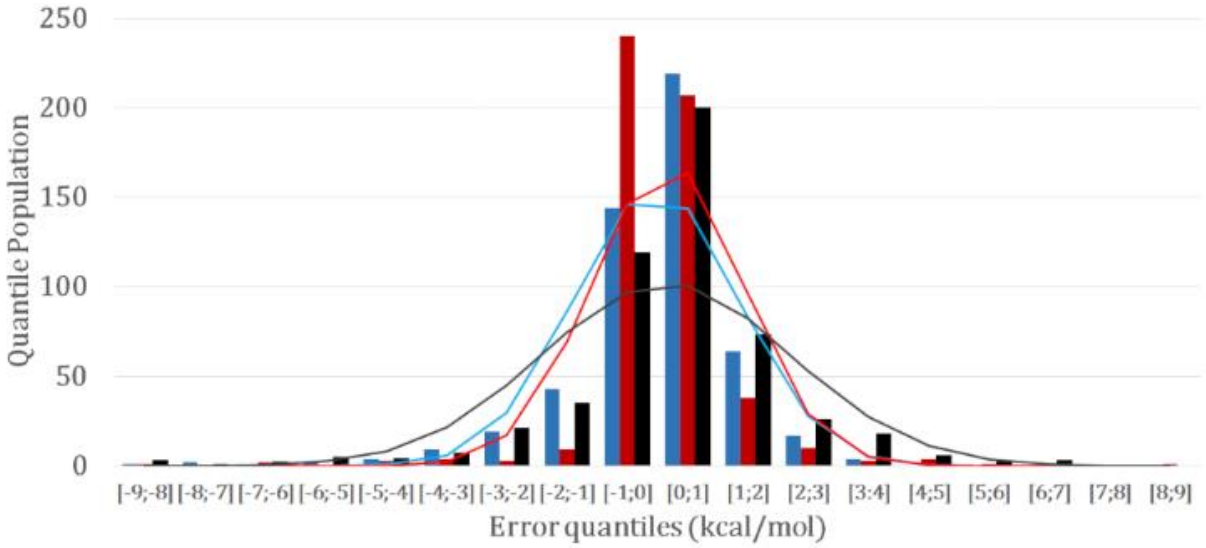


Figure 32. Distribution des erreurs sur les calculs d'énergie (jeu de données S66x8, n = 528, polarisabilités calculées dans les systèmes monomères), par quantiles de 1.0 kcal·mol^{-1}. Barres bleues : erreurs obtenues pour $E_{ELEC,DB}$; ligne bleue : enveloppe gaussienne correspondante aux erreurs pour $E_{ELEC,DB}$ (moyenne : μ = -0,026 kcal·mol^{-1} ; écart type : σ = 1,36 kcal·mol^{-1}). Barres rouges : erreurs obtenues pour $E_{POL,DB}$; ligne rouge : enveloppe gaussienne correspondante aux erreurs obtenues pour $E_{POL,DB}$ (μ = 0.168 kcal·mol^{-1} ; σ = 1.25 kcal·mol^{-1}). Barres noires : erreurs obtenues sur la somme $E_{ELEC,DB}$+$E_{POL,DB}$; ligne noire : enveloppe gaussienne correspondante aux erreurs pour $E_{ELEC,DB}$+$E_{POL,DB}$ (μ = 0,143 kcal·mol^{-1} ; σ = 2,07 kcal·mol^{-1}). Les queues éloignées de la distribution sont omises (aucun des quantiles omis n'a un effectif supérieur à 1).

Les erreurs sur de tels échantillons, de taille réduite, ne suivent l'enveloppe gaussienne qu'approximativement. Les erreurs dans la plage [-1;0] et [0;1] sont plus peuplées que l'approximation gaussienne pour chacune des énergies considérées

($E_{ELEC,DB}$, $E_{POL,DB}$ et $E_{ELEC,DB}$+$E_{POL,DB}$). Ceci suggère que notre modèle présente des observations aberrantes. Ces fortes erreurs exagèrent l'écart-type qui représente la distribution des erreurs. Par conséquent, bien que la plupart des facteurs déterminant l'énergie d'induction soient correctement modélisés, il est impossible d'affirmer que le modèle proposé soit totalement dépourvu d'erreurs systématiques. Il est possible que la source différente des énergies (méthode hybride expérimentale/théorique pour la base de données ELMAM2 et purement théorique pour SAPT) soit la cause de ces divergences. Une autre possibilité est que ces fortes erreurs sont liées à des molécules ou des groupements chimiques moins bien modélisés que les autres. En effet, il a été vu que les dimères impliquant les molécules d'acétamide, d'acide acétique ou d'uracile sont fréquemment associés à de fortes erreurs et ce à diverses distances. Des caractéristiques moléculaires semblent donc également impliquées dans les plus fortes erreurs obtenues pour les jeux de données S66 et S66x8.

2.5 - Conclusion de l'étude sur les jeux de données S66 et S66x8

Cette partie, ainsi que la partie 1.7, présente les polarisabilités atomiques et moléculaires théoriques calculées au niveau B3LYP/6-311++G(3df,3pd) pour les monomères, et pour 57 des 66 dimères, du jeu de données S66. La partition QTAIM de cette propriété en contributions atomiques est disponible dans son intégralité dans les informations supplémentaires de l'article publié[52] (non-reproduites dans le présent manuscrit). Les polarisabilités calculées ont été analysées en détail. Les tenseurs de polarisabilités atomiques topologiques présentent une grande amplitude de variation entre les atomes d'une même espèce chimique, autant du point de vue de leur forme que de leur intensité. Les atomes de carbone sont présents dans un large éventail de fonctionnalités chimiques, et leurs polarisabilités varient fortement. L'agrégation de cette propriété par types d'atomes ELMAM2 permet de mettre en évidence des similitudes. Notamment les moyennes des valeurs propres au sein de chaque groupe ont une variabilité réduite. Cette constatation laisse entrevoir un certain degré de transférabilité des polarisabilités atomiques. La transférabilité des bases de données de densité électronique a déjà de nombreuses applications, même en ignorant les effets de polarisation. L'utilisation d'une bibliothèque de polarisabilités atomiques est le sujet de l'étude de cas qui constitue une grande partie du chapitre suivant. En effet, la

transférabilité est une approximation qui n'est valable qu'à une certaine précision près et l'étude présentée possède des limites qu'on tentera de mettre en évidence, puis de lever.

Cette étude décrit également que l'application directe de polarisabilités atomiques topologiques à un ensemble de pseudoatomes HCM, rendue possible grâce à une nouvelle méthode implémentée dans le module *MoProViewer* du logiciel *MoProSuite*. Cette méthode est conçue pour donner accès à l'induction intermoléculaire de moments dipolaires en utilisant des densités électroniques HCM transférées depuis une base de données. Les fonctions dipolaires des pseudoatomes transférés sont modifiées itérativement pour représenter le phénomène d'induction mutuelle. Lorsque ce procédé arrive à convergence, la densité électronique transférée est complétée par des termes dipolaires induits sur chaque atome des molécules en interaction. Cette densité électronique modifiée donne accès à une estimation de la composante de polarisation dipolaire de l'énergie d'interaction électrostatique totale.

Les jeux de données S66 et S66x8 ont été utilisés comme banc de test pour les énergies d'interaction électrostatiques calculées à partir des densités électroniques multipolaires ELMAM2, transférées et polarisées. Il convient de souligner que la méthode de référence SAPT est indépendante à la fois de la méthode ELMAM2 de reconstitution de la densité électronique et de la méthode DFT utilisée pour calculer les polarisabilités. Les contributions énergétiques évaluées englobent les énergies électrostatiques permanentes et les énergies de polarisation liées à l'induction de dipôles atomiques. Les énergies d'interaction électrostatiques calculées à partir d'ELMAM2 présentent un bon accord avec celles obtenues par SAPT. Aucun paramètre ajustable n'est utilisé dans le modèle de densité électronique, et la pente de la régression linéaire est raisonnablement proche de l'unité. Il s'agit d'un accomplissement important pour confirmer la transférabilité de $\rho(\mathbf{r})$ à partir des pseudoatomes de la librairie ELMAM2. L'utilisation des polarisabilités calculées dans les molécules isolées, sur ces densités électroniques transférables, donne des prédictions encourageantes des énergies d'induction. Ce point suggère qu'une part significative de l'énergie totale d'induction est représentée par le modèle de polarisation proposé. La concordance avec la référence SAPT est étonnamment forte compte tenu de l'approche simple et directe présentée ici. Les meilleurs résultats, pour reconstruire cette induction intermoléculaire, sont obtenus en utilisant les polarisabilités atomiques dérivées des monomères isolés.

Par rapport aux contributions séparées $E_{ELEC,DB}$ et $E_{POL,DB}$, de meilleures prédictions énergétiques sont obtenues en calculant $E_{ELEC,DB}$+$E_{POL,DB}$. Encore une fois, seule la pente de la régression linéaire a été ajustée pour calculer le coefficient de détermination entre l'énergie de référence SAPT et $E_{ELEC,DB}$+$E_{POL,DB}$. Le modèle proposé ne possède aucun paramètre libre. La pente obtenue pour $E_{ELEC,DB}$+$E_{POL,DB}$ est plus proche de l'unité que celles obtenues pour $E_{ELEC,DB}$ seul (par rapport à SAPT E_{ELEC}) ou pour $E_{POL,DB}$ seul (par rapport à SAPT E_{IND}), ce qui suggère un degré de compensation des erreurs entre ces deux calculs. Une telle compensation d'erreur est attendue puisque les densités électroniques ELMAM2 ont été obtenues à partir d'observations cristallographiques et pourrait contenir une polarisation intermoléculaire résiduelle dans les paramètres HCM des pseudoatomes transférables.

Il a été constaté que l'utilisation de polarisabilités associées aux systèmes dimères génère des prédictions énergétiques surestimées. Ces polarisabilités à l'état dimérique sont calculées à partir de la partition QTAIM de la densité électronique théorique totale du dimère. Cette densité théorique est déjà perturbée (polarisée) par l'autre monomère lors de la formation du dimère. L'objectif est de reproduire cette perturbation sur la densité électronique ELMAM2. Les calculs d'énergie de polarisation ont montré que les polarisabilités calculées à partir des monomères conviennent mieux à cette fin, comparées aux polarisabilités calculées à partir des dimères.

Le but de la publication de cette étude a été double. Premièrement, les polarisabilités atomiques et moléculaires, deux propriétés importantes, ont été calculées pour la base de données de référence S66. Deuxièmement, un outil permettant d'exploiter ces propriétés dans un logiciel spécialisé de modélisation multipolaire a été introduit publiquement. L'amélioration du modèle de polarisation pour les densités électroniques HCMM nécessitera de lever plusieurs interrogations. La relation entre la transférabilité et la polarisabilité atomique de la densité électronique doit être explorée plus systématiquement. Le jeu de données S66 ne recouvre qu'un sous-ensemble des types atomiques ELMAM2 (34/154), et ceux-ci ne représentent pas tous les types atomiques nécessaires à la reconstitution de la densité électronique des acides aminés protéinogènes. Par conséquent, l'application à des complexes protéine-ligand nécessite les calculs supplémentaires présentés dans la suite de ce chapitre.

La partition de la polarisabilité dans les groupes supra-atomiques (synthons ou promolecules) a également été négligée dans cette étude. Dos Santos *et al.*, 2015[86], et Krawczuk *et al.*, 2009[49], ont observé que les tenseurs de polarisabilité sont plus transférables pour des groupes d'atomes (méthyle, phényle, COOH, etc.) que pour les atomes individuels. La partition de la polarisabilité par groupes d'atomes doit être explorée à l'avenir, ainsi que les propriétés des liaisons chimiques (points critiques selles, translations de charges, etc.). Ceci implique un examen plus approfondi de la différence entre la partition HCM et QTAIM des propriétés moléculaires de la densité électronique. Ce sujet séduisant mérite sa propre étude et le présent travail esquisse sa faisabilité. L'application des contraintes de symétrie locale d'ELMAM2 à la polarisabilité atomique sera abordée par la suite. L'effet du couplage dipolaire intramoléculaire (voir page 37) doit être évalué en profondeur. Cette atténuation intramoléculaire de l'induction devrait améliorer les énergies d'induction, mais aussi produire une décomposition différente des polarisabilités moléculaires en contributions atomiques, comparé à la partition QTAIM. De plus, ce modèle nécessite des fonctions d'atténuation de l'induction intramoléculaire éventuellement paramétrables. Le choix de cette méthode d'atténuation modifiera nécessairement la répartition de la polarisabilité moléculaire en contributions atomiques[43,46,47,48]. Cet axe d'étude peut être particulièrement pertinent pour définir la polarisabilité de synthons ou de promolecules à partir de pseudoatomes du HCMM. Les quatre derniers axes évoqués – compléter la base de données, évaluer la transférabilité supra-atomique des polarisabilités, appliquer des contraintes de symétrie et coupler l'induction dipolaire – constituent une stratégie cohérente pour lever les limites du modèle présenté ici. Cette stratégie peut mener à une compréhension fructueuse pour appliquer le modèle de densité transférée-polarisée aux grands édifices moléculaires que sont les complexes protéine-ligand, comme il sera tenté à la partie 3.7.

Chapitre 3 :

Construction et Application

d'une Base de Données

de Polarisabilités Atomiques Transférables

Introduction

Le module « *Beluga* » de *MoProViewer* est donc capable de caractériser et comparer de larges jeux de polarisabilités atomiques, comme vu chapitre 1. Le chapitre 2 montre que ce module est capable de retrouver l'effet des polarisabilités sur la densité électronique tranférée, et de générer des prédictions utiles grâce aux fonctions de polarisation. Le présent chapitre est donc consacré au dernier outil – plus ou moins – complet qui a été développé au cours de ce travail de thèse. Ce dernier outil permet la création d'une base de données de polarisabilités atomiques à partir d'un ensemble de structures moléculaires.

L'outil a été penser pour compléter l'effort de développement de la librairie ELMAM2 plutôt que de créer une nouvelle librairie indépendante (ce qui aurait encore allongé le temps de développement). L'outil nécessite donc une base de données atomiques, formatée comme ELMAM2, et attribue une polarisabilité moyenne à chaque type atomique, à partir des polarisabilités atomiques des structures initiales.

Les considérations générales liées à l'implémentation des types atomiques ELMAM2 dans *MoProViewer* ouvrent ce chapitre. En effet cette fonction était auparavant uniquement incluse dans le logiciel d'affinement *MoPro*. Les critères de définition des types atomiques seront discutés.

La définition d'une moyenne pour un ensemble de tenseurs définis-positifs n'est pas mathématiquement unique. Un problème posé par cet état de fait est ensuite exposé, ainsi qu'une méthode permettant de contourner celui-ci.

L'outil de construction de banque de polarisabilités étant complètement introduit, il est appliqué au « *Big Set* » de 198 molécules introduit à la partie 1.8. Le transfert de la densité électronique et des polarisabilités atomiques sur 3 structures, modèles d'interactions protéine-ligand, est décrit.

Ce chapitre est clos par l'étude de cas des 3 modèles préparés, différentes streptavidines en interaction avec le ligand biotine. L'induction dipolaire, les énergies d'interaction électrostatique et de polarisation sont interprétées à la lumière d'informations expérimentales biochimiques. Une analyse topologique des interactions entre biotine et streptavidine est également proposée.

3.1 - L'approche des bases de données : considérations générales liées à la librairie ELMAM2

Le principe de transférabilité est la fondation méthodologique qui permet de reconstruire une densité électronique à partir de fragments : promolécules, pseudoatomes multipolaires[61,68,70] ou non[82], charges ou moments ponctuels[29,30,101] etc.

Le principe de transférabilité est défini ainsi : la densité électronique associée à un fragment moléculaire est sensiblement la même que celle du même fragment dans un contexte différent. Le terme 'sensiblement' indique que le principe de transférabilité n'est vrai qu'à un certain niveau de tolérance. Dans le cadre d'ELMAM2, les fragments sont des pseudoatomes multipolaires moyens. Chaque paramètre d'un pseudoatome donné est associé à sa déviation standard permettant de quantifier la dispersion des paramètres moyennées.

On s'intéresse ici à la transférabilité de la polarisabilité, une propriété liée à la densité électronique. La transférabilité des polarisabilités atomiques a été argumentée aux chapitres 1 et 2, on souhaite maintenant pouvoir construire des bases de données de polarisabilités, associées pour le présent travail aux types d'atomes définis dans ELMAM2. Le travail de développement du module de polarisation de *MoProViewer* a poussé à porter ELMAM2 dans ce module de *MoProSuite* plutôt que modifier *MoPro* (chargé du transfert) pour supporter les polarisabilités. Le portage d'ELMAM2 dans *MoProViewer* a nécessité de rendre possible la lecture du fichier 'ELMAM2.txt', de la réaction d'une classe C++ nommée '`atomType`', représentant les types atomiques ELMAM2, de fonctions de reconnaissance de ces types atomiques et enfin de fonctions de transfert.

Le fichier de base de données est lisible par un humain. Chaque ligne commence par un mot-clé qui définit le nombre et la signification des nombres et chaines de caractères de cette ligne. Chaque type atomique utilise au maximum une seule fois chaque mot-clé, rendant la rédaction d'une fonction de lecture triviale. Les fonctions de transfert ont été largement réimplémentées à partir des fonctions permettant le transfert de polarisabilités atomiques orientées. Ces deux aspects du développement ne seront pas plus explicités, à l'exception des méthodes permettant de trouver l'orientation correcte des pseudoatomes.

La classe C++ '`atomType`' a été basée sur la classe '`Atome`' avec laquelle on souhaite pouvoir la comparer. Cependant un type atomique est, sur de nombreux points, plus simple qu'un '`Atome`' de *MoProViewer*. Un type atomique n'a pas de coordonnées propres, ni de représentation graphique par exemple. Par contre un type atomique possède une expansion multipolaire (comprenant κ et κ'), une polarisabilité, une symétrie locale, etc. Un type atomique ne possède pas d'axes locaux, mais il possède toutes les informations pour construire ce référentiel local une fois un atome reconnu (type d'axes, éléments chimiques des directions de définition *dir1* et *dir2*).

La similitude entre atomes est évaluée par la nature chimique de leurs voisins covalents (voisins au 1er degré) et des voisins de ces derniers. Certains types ELMAM2 sont définis uniquement par leurs voisins covalents, d'autres utilisent jusqu'au 3ème degré de voisinage. Le voisinage d'un type atomique est décrit par la chaîne atomique (*atom neighbor string,* voir figure 33), qui possède toujours la même structure :

• Le symbole de l'espèce chimique du type, avec sa première lettre capitalisée. Exemple (a) : **C**onn.

• Suivent les symboles des espèces chimiques voisines de l'atome, toujours en minuscules. Exemple (b) : C**fcc**.

• Certaines chaînes atomiques comportent un nombre en amont des symboles des voisins. Ce nombre représente l'ordre de liaison entre l'atome du type ses voisins. Exemple : C**1.5**c**1.5**c**1**h.

• Dans les exemples précédents, les voisins au 2ème degré sont ignorés. Certains atomes du voisinage sont suivis de crochets pour noter les espèces chimiques des atomes voisins des voisins. Exemple (c) : Cc**[chh]**sc**[nch]**h. Dans l'exemple précédent, le carbone du type atomique est lié à 4 atomes dont 2 carbones tétravalents (le premier voisin est lié au carbone du type atomique et deux autres carbones). En effet, les symboles entre crochets ne reprennent pas les voisins déjà évalués.

• Pour définir des voisins au 3ème degré, on utilise des parenthèses. Les parenthèses sont ouvertes et fermées toujours entre 2 crochets. Exemple (d) : Ho[p**(ooo)**]. Dans cet exemple de l'hydrogène de l'acide phosphorique, le phosphore est lié à 4 oxygènes. Encore une fois, les atomes déjà visités ne sont pas notés. Les voisins au 3ème degré sont utiles en particuliers pour les atomes à un voisin.

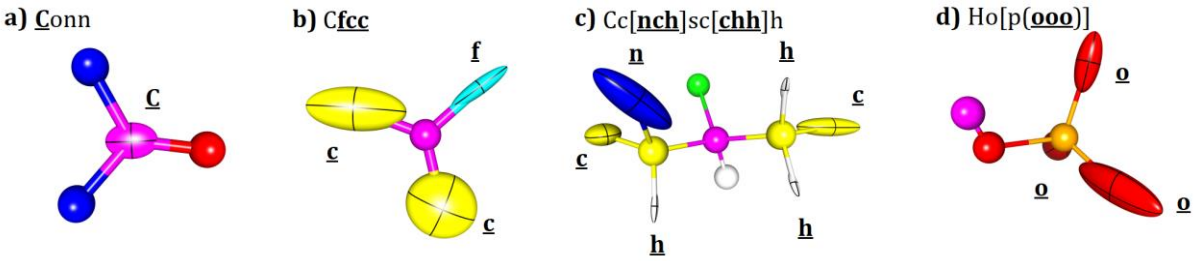


Figure 33. Quelques chaines de voisinage atomique en images. L'atome associé au type atomique – dont les chaines sont indiquées – est de couleur rose. Seuls les atomes de la chaine atomique sont représentés. Les ellipsoïdes mettent simplement en évidence les atomes soulignés-gras des chaines atomiques.

L'interprétation de la chaîne atomique est réalisée à l'instanciation des types atomiques à partir des données du fichier 'ELMAM2.txt'. Une fois cette liste de types atomiques initialisée à partir d'une base de données, *MoProViewer* est prêt à transférer les atomes sélectionnés via l'interface graphique.

Pour identifier un type atomique pertinent pour décrire un atome donné, il faut que son nombre de liaisons covalentes corresponde au nombre de voisins du type atomique. La nature chimique des voisins telle que décrite par la chaîne atomique doit être respectée. Enfin, l'atome doit appartenir, ou non, à un cycle de taille définie.

MoProViewer parcourt la liste et réalise 3 tests pour chaque type atomique. La fonction '`matchBonds`' vérifie que l'atome a bien autant de voisins que l'attribut '`NBOND`' indiqué dans le fichier de base de données. La fonction '`matchCycles`' utilise l'algorithme décrit partie 3.3 (voir plus bas) pour vérifier l'appartenance à un cycle selon la règle n°4 ci-après (attribut '`CYCLE`'). Enfin la fonction '`matchNeighborChem`' teste la correspondance entre l'élément chimique de l'atome et du type atomique testé, puis explore le graphe moléculaire pour déterminer si la chaîne atomique décrit bien le voisinage de l'atome considéré.

La chiralité d'un atome devrait être également évaluée mais l'algorithme n'a pas été terminé car tous les types atomiques utilisés comportent des éléments de symétrie de type miroir. En dehors des types additionnels, seul C412T, un type inutilisé dans le présent jeu de données, ne possède pas de symétrie miroir. Le système d'axes n'est pas évalué pour reconnaître un type atomique. On ne suppose pas que les axes locaux sont au préalable dans la bonne configuration. Une fois que la reconnaissance des types

atomiques d'une molécule est terminée, *MoProViewer* vérifie que les atomes associés avec succès à un type ELMAM2 ont bien le référentiel local documenté dans le fichier de base de données.

Pour cela, la méthode '`matchAxisSystem`' vérifie d'abord si les types d'axes correspondent[102] entre l'atome et le type atomique. Si le type d'axes correspond, la méthode vérifie si les directions de définition (*dir1* et *dir2*) correspondent à celles documentées pour le type atomique reconnu (on vérifie uniquement la nature chimique des atomes vers qui pointent les vecteurs *dir1* et *dir2*). Si l'utilisateur coche la case correspondante dans l'interface, chaque atome qui ne remplit pas ces conditions sera réparé par la méthode '`repairAxis`' de *MoProViewer*, qui utilise les nouvelles règles décrites ci-dessous – et les modifications correspondantes de la base de données – pour attribuer à chaque atome un système d'axes d'orientation univoque. En fonction des erreurs d'orientation détectées, un second passage de l'algorithme de transfert peut être nécessaire. Des messages d'erreur détaillés invitent l'utilisateur à vérifier le fichier corrigé automatiquement.

3.2 - Création d'une base pELMAM3 (polarizable Experimental Library of Multipolar Atom Models)

Une base de données de propriétés atomiques d'origine expérimentale peut difficilement être considérée comme terminée, par exemple du fait de la constante disponibilité de nouvelles sources de données ou de l'évolution des critères de qualité des structures éligibles à l'intégration dans la banque. Le set de molécules choisi pour construire la banque gouverne également son application raisonnable et la qualité de la densité reconstruite.

Le présent manuscrit est donc une occasion idéale pour faire le point sur les évolutions de la base ELMAM2 depuis sa publication en 2012. Si la plupart des modifications sont liées au présent travail et à l'utilisation des types atomiques dans *MoProViewer*, d'autres collaborateurs ont apporté leurs propres modifications. En l'attente d'une publication, le récapitulatif de l'intégralité de ces modifications est utile et naturel, d'autant plus que c'est la version modifiée d'ELMAM2 qui a été utilisée pour construire la base de données de polarisabilités qui fait l'objet de ce chapitre.

La présente version de la banque compte 153 types atomiques. Afin de conserver une compatibilité maximale avec le transfert par *MoPro*, tout en levant certaines ambiguïtés (par exemple sur l'orientation des types atomiques), le transfert avec *MoProViewer* utilise des règles légèrement différentes pour catégoriser les atomes en types atomiques.

Améliorations contextuelles, nouvelles règles de lecture :

Règle 1 : *dir1* et *dir2*, les deux vecteurs relatifs servant à la construction de la base locale, pointent vers des atomes voisins au 1er degré. Suivant Domagała et Jelsch, 2008[102], ces deux vecteurs doivent être choisis de manière à maximiser la symétrie locale de l'atome multipolaire transféré. Cette évaluation passe uniquement par l'élément chimique, ce qui entraine des indéfinitions d'orientation quand *dir1* et *dir2* peuvent pointer vers différentes combinaisons de voisins. Pour réparer cette indéfinition, on place *dir1* toujours avant *dir2* dans la chaîne atomique. Ce qui permet de vérifier dans l'ordre de la chaîne atomique la présence de *dir1* puis de *dir2*. *MoProViewer* construit des listes de voisinage pour représenter la connexion des atomes dans une molécule. Elles ne sont pas exploitables ici car l'ordre des connexions diffère entre les fichiers. Certaines chaînes ont dû être modifiées pour faire apparaître des différences entre *dir1* et *dir2* (exemple : C408). Si un type est reconnu pendant le transfert, mais que son système d'axes ne correspond pas à la définition ELMAM2, *dir1* et *dir2* sont mises à jour. Avec la présente implémentation, un seul passage de l'outil de réparation permet, pour tous les systèmes considérés ici (protéines-ligands ou petites molécules), d'obtenir des axes corrects. Le transfert sur la structure de l'arginase 1HQG (5023 atomes, voir partie 4.2) a laissé 4 atomes à transférer manuellement (les deux ions Mn^{2+}, modélisés par des monopôles, et l'ion OH^-, pour l'instant absent de la base). Pour les atomes à un seul voisin, ce voisin est toujours pointé par *dir1* et *dir2* est cherché parmi les voisins au 2ème degré.

Règle 2 : Définition de la symétrie '`cy`' (cylindrique autour de l'axe Z) pour réserver le '`1`' pour l'identité, appliqué à tous les hydrogènes. Les opérateurs de symétrie des atomes ELMAM2 ont été implémentés sous forme de matrices de rotation permettant de retrouver les positions équivalentes d'un point symétrique par simple produit matriciel. Cette solution élégante n'est pas adaptée à la symétrie '`cy`' : une symétrie cylindrique

implique un nombre infini de positions équivalentes (correspondant à un groupe limite de Curie). Pour moyenner des tenseurs de polarisabilité, la symétrie 'cy' est approximée par 'mxmy' (voir partie 3.5 pour la justification de cette approximation, dans ce cas précis).

Règle 3 : L'utilisation de « jokers » (*wildcards*) dans une chaîne atomique était déjà possible avec les versions précédentes de la base de données. Afin de contrôler plus finement le transfert de paramètres, désormais un caractère en majuscule correspond à n'importe quel nombre d'atomes vérifiant le symbole « joker ». Les atomes utilisant des « jokers » majuscule peuvent avoir n'importe quel nombre de voisins. Inversement, un caractère minuscule correspond à un seul atome par symbole « joker », ce qui implique que le nombre de voisins de l'atome est fixé par le nombre de caractères « joker ». La ligne 'WILDCARD' définit, type par type, quels éléments vérifient les caractères « jokers ».

*Exemple 1 : Le type atomique H113 (ammonium) répond à la chaîne atomique 'H**n**[chx]'. La ligne 'WILDCARD' fait correspondre x à C ou H. L'azote souligné aura ici exactement 4 voisins dont : un carbone, un hydrogène et un « joker ». Cet hydrogène correspond donc à une amine protonée primaire ou secondaire.*

*Exemple 2 : Le type atomique C311 (carbone cétone) répond à la chaine atomique 'Co**c**[X]**c**[X]'. La ligne 'WILDCARD' fait correspondre X à C, H ou O. Les deux carbones soulignés peuvent avoir n'importe quel nombre de voisins C, H ou O, mais aucun autre élément chimique n'est autorisé.*

Règle 4 : Évaluation des cycles : Auparavant, les attributs '0' et '-' avaient la même signification dans la base de données ELMAM2. L'écriture d'un algorithme de détection de cycle a permis de différencier ces deux symboles. Désormais, les atomes des molécules linéaires portent l'attribut '0'. La recherche de cycle s'arrête au 7ème voisin (le cyclo-octane serait par exemple considéré linéaire par le programme). Le symbole '-' (indifférent) entraine la non-évaluation du caractère cyclique. Sinon le nombre indiqué indique le nombre d'atomes dans le cycle, à l'exception de la valeur spéciale '56' qui signifie que l'atome est membre d'un cycle à 5 ou à 6 membres.

Règle 5 (tacite) : En général, placer les types atomiques les plus spécifiques en amont des moins spécifiques, quand cela permet de lever une ambiguïté sur certains atomes. Le premier type atomique valide du fichier est retenu en cas d'ambiguïté. Pour cette raison, les modifications sont présentées dans l'ordre où elles apparaissent dans le

fichier ELMAM2.txt, d'où certaines discontinuités dans la numérotation des types atomiques.

Ajout des éléments des tenseurs de polarisabilité :

Tous les types atomiques ont reçu 2 nouvelles lignes, '`ALPHA`', contenant les 6 valeurs indépendantes de la polarisabilité atomique (*xx*, *yy*, *zz*, *xy*, *xz* et enfin *yz*). Ce tenseur est toujours exprimé dans le système d'axes locaux. La ligne suivante, '`ALPHA_ESD`', contient la déviation standard de chaque élément, calculée par rapport aux éléments indépendants des polarisabilités moyennées, exprimées dans le même référentiel. Si des contraintes de symétrie sont appliquées à la moyenne, celles-ci ne sont pas prises en compte pour le calcul de cette déviation (afin d'avoir des déviations standard comparables entre les moyennes symétrisées ou non). Les moyennes tensorielles de polarisabilités sont toujours calculées dans le référentiel local du type atomique.

3.3 - Détection de boucles

MoPro est capable de transférer la librairie de paramètres multipolaires ELMAM2. Afin de pouvoir construire une version polarisable de cette base de données, la possibilité de reconnaître des types atomiques et de lire le fichier de base de données (`ELMAM2.txt`) ont été réimplémentées dans *MoProViewer*. Seule l'évaluation de la géométrie des types atomiques n'a pas été implémentée (les distances et les angles de liaisons avec les premiers voisins sont documentés dans le fichier ELMAM2).

Cependant, les types atomiques ELMAM2 possèdent un attribut non-évalué dans *MoPro*, le champ « `CYCLE` ». Cet attribut décrit l'appartenance du type atomique à une molécule cyclique. Il a été vu partie 1.7.f, que cette évaluation de l'appartenance à un cycle est pertinente pour la transférabilité de la polarisabilité.

Dans le format original des types atomiques ELMAM2, le champ « `CYCLE` » peut avoir les valeurs synonymes '`0`' et '–', signifiant l'appartenance à une molécule linéaire, ou un chiffre entier qualifiant l'appartenance à un cycle de taille définie. Par exemple, '5' représente l'appartenance à un cycle à 5 membres (cyclopentane), '6', l'appartenance à un cycle à 6 membres (benzène, cyclohexane). L'attribut spécial '56' signale l'appartenance à

un cycle à 5 ou 6 membres indifféremment. Le champ « CYCLE » n'est pas utilisé dans *MoPro* pour la reconnaissance automatique des types atomiques.

Le fichier ELMAM2 a été modifié pour stocker des polarisabilités transférables, exprimées dans le référentiel local. Pour ces versions modifiées (par exemple, pELMAM3.txt), le champ « CYCLE » est interprété par *MoProViewer* de manière légèrement différente : '0' signifie désormais l'appartenance à une molécule linéaire (évaluée jusqu'aux cycles de 7 membres), alors que '–' désigne l'absence de vérification du caractère cyclique. Le type atomique « CYCLE - » pourra donc appartenir à une molécule linéaire ou cyclique. Les types existants ont été modifiés, cependant la syntaxe originale a été conservée sauf si l'absence de modification a entraîné des problèmes de reconnaissance.

Voici la méthode retenue pour déterminer l'appartenance d'une molécule à un cycle. Pour chercher un cycle à *n* membres, le but sera de constituer une liste de *n* atomes parmi la molécule, liés deux-par-deux et dont le dernier est lié au premier. Cette liste est appelée list_visited_atoms dans la suite. La méthode considère le graphe moléculaire comme un arbre à parcourir, à partir d'un atome de départ. L'architecture est donc récursive, compte tenu du fait que l'opération à réaliser pour chaque atome (nœud de l'arbre) est la même : l'ajouter ou non à cette liste. Cette opération pour chaque nœud se résume à la question suivante : « *Sans passer par une liaison déjà visitée, quand il ne reste plus qu'une seule liaison covalente à parcourir, l'atome de départ fait-il partie des voisins l'atome considéré ?* » La réponse à cette question dépend de la réponse à la même question pour les nœuds de l'arbre situés en aval, impliquant la nécessité d'employer une approche récursive dans l'implémentation informatique. Cette architecture demande nécessairement une portée de recherche maximale en nombre de liaisons covalentes pour pouvoir se terminer, d'où une limite de l'implémentation actuelle : 'CYCLE 0' donne un faux positif en cas d'atome membre d'un cycle de plus de 7 membres (par exemple : cyclo-octane).

La règle « *ne pas passer par un atome déjà visité* » dispose d'une exception si et seulement si l'atome appartient aussi à un cycle potentiellement plus grand que celui déjà évalué. Cette condition évite l'échec de la méthode lorsqu'appliquée à une molécule comportant plusieurs cycles partageant deux atomes au moins (ex. : adénine, guanine, tryptophane, etc.). Dans ce cas, en fonction de l'ordre des voisins dans la liste stockée en

mémoire, un des carbones liant un cycle à 5 et un cycle à 6 membres pourra être évalué de manière erronée comme non-cyclique alors qu'il appartient à un cycle trop petit (déjà parcouru par la fonction) et un cycle de la taille requise. Cet exemple contre-intuitif est expliqué sur la Figure 34.

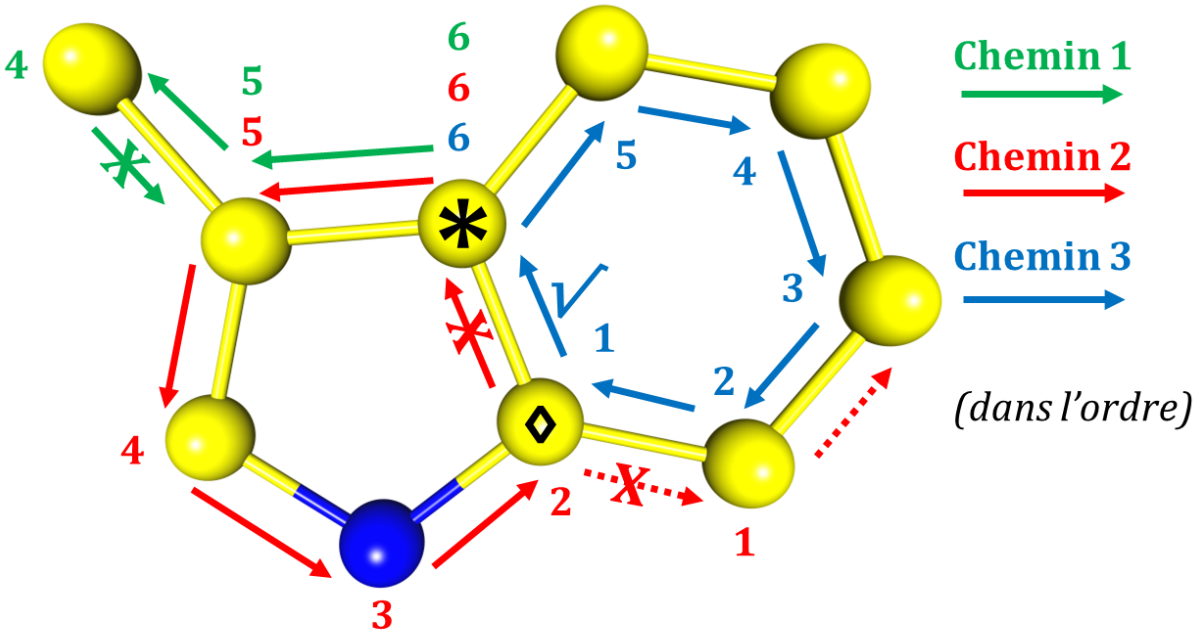


Figure 34. Le noyau indole du tryptophane peut piéger un algorithme incapable de retourner sur ses pas. On souhaite savoir si l'atome marqué d'un astérisque est membre d'un cycle à 6 membres. Les chemins 1, 2 et 3 sont évalués par *MoProViewer* dans l'ordre indiqué. Les chiffres de couleur indiquent le nombre de liaisons restantes à parcourir à chaque étape de chaque chemin. Le chemin 1 termine après 2 étapes car l'algorithme ne peut pas revenir directement sur ses pas. Le chemin 2 termine après 5 étapes, car l'atome de départ est trouvé trop tôt. À ce moment, l'atome marqué d'un '⋄' fait partie des atomes visité. Le chemin 3 est le bon, mais doit passer par l'atome marqué '⋄'. Le chemin 3 a déjà parcouru 5 liaisons pour arriver à cet atome. Le chemin 2 l'a visité après 4 liaisons, le chemin 3 peut donc évaluer si cet atome possède l'atome '*' de départ parmi ses voisins‡‡.

Cette version de la méthode ne retourne donc vrai que si le nombre exact de liaisons a été parcouru. L'algorithme garde donc en mémoire, pour chaque atome visité,

‡‡ *: L'atome qui termine le chemin en pointillés n'est pas noté comme visité car évaluer ses voisins termine le chemin considéré.*

la longueur du chemin parcouru en nombre de liaisons covalentes, au cas où un chemin plus long serait exploré par la suite.

Pour plus de clarté, le pseudocode ci-dessous omet les détails d'implémentation non-pertinents pour la compréhension de l'algorithme, en particulier les spécificités des classes de *MoProViewer*, quelques optimisations liées au langage C++ et la condition, triviale, permettant d'éviter les cycles à un ou deux membres.

L'argument `atom_start` de cette fonction '`testCycle()`' pointe vers l'atome dont on cherche l'appartenance à un cycle. L'argument `evaluated_atom` est utilisé pour l'architecture récursive de l'algorithme. Au premier appel de la fonction, `evaluated_atom` est `atom_start`. L'identité des atomes est évaluée par la surcharge de l'opérateur de comparaison (« == »). Comparer deux atomes signifie évaluer si leurs pointeurs sont stockés au même endroit au sein de la classe molécule [cet opérateur est utilisé par la fonction `contains()`]. `evaluated_atom.neighbor[]` représente la liste des voisins covalents de `evaluated_atom` (cette liste est construite au chargement de la molécule par *MoProViewer*). Une fois un voisin visité, il est stocké au sein de `list_visited_atoms`, une liste d'atomes destinée à éviter que l'algorithme ne visite plusieurs fois la même position. `list_remaining_nbonds` est une liste d'entiers qui stocke `max_bonds` au moment où un nouvel atome est ajouté à `list_visited_atoms` (les entiers de la liste sont dans le même ordre que les atomes). `list_remaining_nbonds` autorise une seconde visite d'une position de `list_visited_atoms` seulement si un chemin plus court a été exploré précédemment. Cette liste permet à l'algorithme d'éviter de retourner des faux négatifs dans les molécules polycycliques. L'argument `max_bonds` doit être supérieur ou égal à 1. Il empêche toute boucle infinie. D'un point de vue chimique, `max_bonds` doit être supérieur ou égal à 3. Un *wrapper* est disponible dans le code source de *MoProViewer* pour éviter au programmeur de passer des arguments incohérents. Tous les arguments (sauf `max_bonds`) de la fonction suivante sont passés sous forme de pointeurs.

Pour les types d'atomes avec un cycle de type "56", le résultat correspond donc à :

```
testCycle (atom_start, atom_start, empty_list1, empty_list2, 5)
OR testCycle (atom_start, atom_start, empty_list1, empty_list2, 6)
```

L'implémentation d'une version retournant vrai pour les cycles de taille inférieure ou égale à `max_bonds` a été implémentée. Cette version est utilisée pour évaluer si une

molécule est linéaire (jusqu'à la septième liaison covalente, champ « `CYCLE 0` »). Aucunes de ces méthodes n'est appelée si le type atomique stipule « `CYCLE -` ».

Encart 1. Pseudocode décrivant la méthode d'identification de cycles dans la procédure de transfert implémentée dans *MoProViewer*

```
bool testCycle (atom atom_start, atom evaluated_atom, list list_visited_atoms,
list list_remaining_nbonds, int max_bonds)
{
bool result = false
list_visited_atoms.append(evaluated_atom)
list_remaining_nbonds.append(max_bonds)
for i in evaluated_atom.neighbor[] {
      //Cas 1: 'atom_start' appartient à un cycle de taille 'max_bonds' :
L'atome suivant est 'atom_start' et le nombre de liaisons parcourues est
'max_bonds'.
      if (max_bonds == 1
          AND evaluated_atom.neighbors.contains(atom_start) ){
            result = true
            break loop
      }
      //Cas 2: l'atome suivant a déjà été visité en utilisant un chemin plus
court. Il est toujours susceptible de faire partie d'un cycle de taille
requise.
      else if(list_visited_atoms.contains(evaluated_atom.neighbor[i])
              AND max_bonds > list_remaining_nbonds[i] ){
            list_remaining_nbonds[i] = max_bonds
            result = testCycle(atom_start, evaluated_atom.neighbor[i],
                               list_visited_atoms, list_remaining_nbonds,
                               max_bonds-1)
      }
      //Cas 3 : l'atome n'a pas été évalué et est susceptible de faire partie
d'un cycle de taille requise.
      else if(NOTlist_visited_atoms.contains(evaluated_atom.neighbor[i])
              AND max_bonds > 1 ){
            result = testCycle(atom_start, evaluated_atom.neighbor[i],
                               list_visited_atoms, list_remaining_nbonds,
                               max_bonds-1)
      }
      //Cas 4 : l'atome appartient aux atomes visités ou max_bonds ont été
parcourrus sans succès : 'result' reste faux.
} return (result) }
```

3.4 - Moyennes de tenseurs définis positifs

Pour obtenir des polarisabilités moyennes pour chaque type ELMAM2, il convient d'examiner les méthodes pour obtenir la moyenne d'un tenseur de second ordre. En effet toutes les polarisabilités sont ici représentées par des tenseurs 3·3 définis positifs et symétriques. La méthode la plus simple consiste à moyenner les matrices élément par élément.

Cependant, cette première méthode ne permet pas de conserver les paramètres topologiques du tenseur (ε et η, examinés à la partie 1.7.e) qui représentent son anisotropie. Les tenseurs moyennés sont systématiquement plus isotropes que la moyenne des échantillons dont ils sont issus. La Figure 36, page 166, met ce problème en évidence.

Une technique basée sur la décomposition de chaque polarisabilité en valeurs propres et vecteurs propres a été conçue pour éviter ce problème. Ensuite cette méthode sera comparée à une variante de la même méthode incorporant la symétrie locale des types atomiques ELMAM2.

Chaque type atomique définit un système d'axes locaux qui définit à son tour l'orientation des propriétés (polarisabilité ou densité électronique) lors du transfert sur une structure moléculaire. Il est naturel d'utiliser ce référentiel local pour exprimer les polarisabilités atomiques et y calculer nos moyennes. En effet, ces référentiels ont été choisis[102] pour augmenter la symétrie de la densité électronique, et donc réduire le nombre des paramètres multipolaires par atomes. Il est nécessaire de comparer des atomes dont les propriétés sont similaires. Il est aussi important que la définition des axes locaux soit la moins ambigüe, la plus consistante possible, afin que des atomes effectivement similaires voient leurs propriétés moyennées dans la même orientation. Les nouvelles règles de la partie 3.2, ainsi que les modifications listées au chapitre 3-bis, rendent univoque l'orientation des référentiels locaux. Lever ce type d'ambiguïté a par ailleurs motivé un grand nombre des modifications présentées dans ce chapitre 3-bis.

Pour construire une base de polarisabilités dans *MoProViewer*, une collection de molécules dont les polarisabilités atomiques sont documentées en lieu et place des paramètres de déplacement atomique (*ADP*) est chargée. Chaque atome est associé à son type atomique et si le référentiel local est incorrect, celui-ci est bien entendu corrigé. Dans

notre système de fichiers, les polarisabilités sont exprimées dans le référentiel moléculaire global[§§]. Les valeurs numériques des polarisabilités initiales sont donc inchangées par cette réparation des référentiels atomiques locaux.

Ensuite, pour chaque type atomique présent dans le set de molécules, on liste toutes les polarisabilités atomiques associées à ce type. Les polarisabilités initiales, exprimées dans le référentiel global, sont exprimées dans le référentiel local de chaque atome. On décompose ensuite chaque tenseur en valeurs et vecteurs propres. Les trois valeurs propres sont des scalaires indépendants et sont moyennées par ordre décroissant. Cette partie du procédé est celle qui permet la meilleure conservation de l'anisotropie des polarisabilités moyennées. Puis les vecteurs propres associés à ces mêmes valeurs propres sont moyennés à leur tour (simple somme vectorielle divisée par le nombre de polarisabilités à moyenner). Ceci suppose que la définition des types atomiques soit correctement réalisée, la figure 35 montre un exemple d'échantillon hétérogène de 2 polarisabilités atomiques. Les conséquences de ces hétérogénéités des propriétés moyennées sont discutées à chaque étape de la méthode.

On admet à ce point que les atomes sont comparables et que la définition de leur orientation est reproductible. L'échec du calcul est encore possible quand, au cours du calcul de la moyenne des vecteurs propres, la somme vectorielle est nulle ou très petite. En effet notre librairie d'algèbre linéaire peut indifféremment retourner **v** et –**v** où **v** est un vecteur propre unitaire. L'association entre vecteurs et valeurs propres est cruciale, on considère le cas où les trois valeurs propres sont significativement différentes ($\lambda_1 > \lambda_2 > \lambda_3$). Dans ce cas, l'association entre vecteurs propres et valeurs propres est univoque. Les polarisabilités atomiques ont presque toutes un caractère oblong (*prolate*), en pratique, l'association du premier vecteur à λ_1 est univoque.

[§§] *: Dans l'implémentation actuelle du système de fichiers MoPro, les ADP sont exprimées dans le référentiel moléculaire global cristallographique (coordonnées fractionnaires). Ce choix est tout à fait sensé pour les ADP, pas pour les polarisabilités qui utilisent le même champ du fichier. Pour la présente application, les paramètres de maille ont été orthonormalisés par une routine accessible dans le module de polarisation afin que les valeurs numériques des polarisabilités inscrites dans le fichier moléculaire soient égales aux valeurs en Å³.*

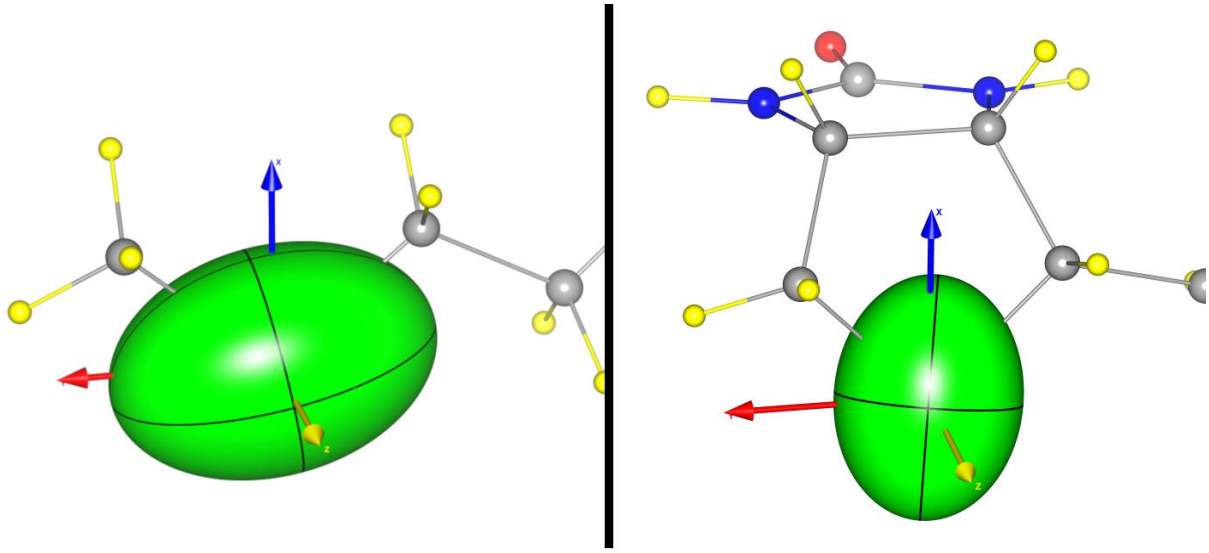


Figure 35. Polarisabilités théoriques (CAM-B3LYP/d-aug-cc-pVDZ) pour le soufre de la Méthionine (type S202, à gauche) et pour le soufre de la Biotine (type S202T, à droite). Échelle 0.4 Å^3/Å. Les flèches symbolisent la direction des axes locaux (système *bXY*, construit à partir des deux carbones voisins). Dans les deux cas, les premiers voisins sont des carbones sp^3, pourtant le vecteur propre associé à lambda_1 est dans un cas proche de **y** et dans l'autre plus proche de **x**. Ce constat a motivé la création du type S202T, spécifique des cycles à 5 membres, comme décrit partie 3.a, pour éviter une moyenne de polarisabilités *a priori* non-comparables.

On souhaite obtenir un jeu de 3 vecteurs propres moyens, chacun associé à une valeur propre différente. La procédure suivante est répétée pour chacun des 3 vecteurs propres. La suite vectorielle suivante est utilisée pour obtenir un vecteur moyen, elle est appliquée à chacun des trois vecteurs propres à moyenner. Cette méthode demande de définir un vecteur de référence $\mathbf{v}_{ref}$, on choisira le vecteur propre $\mathbf{v}_0$ du premier tenseur de la liste des n tenseurs comme référence :

$$\langle \mathbf{v} \rangle_n = \frac{1}{n} \sum_n \begin{cases} \mathbf{v}_i, & \mathbf{v}_i \cdot \mathbf{v}_{ref} \geq 0 \\ -\mathbf{v}_i, & \mathbf{v}_i \cdot \mathbf{v}_{ref} < 0 \end{cases} \qquad (72)$$

Où '·' symbolise le produit scalaire et où le premier vecteur $\mathbf{v}_i$ de l'échantillon à moyenner est considéré comme vecteur de référence : $\mathbf{v}_0 = \mathbf{v}_{ref}$. C'est une contrainte assez lâche, la suite contraint seulement les vecteurs sommés à se situer du même côté du plan normal défini par $\mathbf{v}_0$. L'effet de ce choix est discuté par la suite. Le facteur $1/n$ pourrait être ignoré puisque cette somme concerne des vecteurs propres.

La Figure 35 souligne l'importance d'utiliser la méthode ci-dessus sur des tenseurs comparables. En effet moyenner l'orientation des deux ellipsoïdes présentées Figure 35 aboutirait à un vecteur propre moyen – associé à λ_1 – situé quelque part, proche du vecteur local (**x** – **y**), dans la direction de la liaison covalente au carbone sp^3 à droite du soufre. Cette caractéristique n'est présente dans aucune des deux polarisabilités moyennées et n'est donc pas souhaitable. Toutefois, une hétérogénéité mineure présente dans l'échantillon doit affecter nos moyennes et ne peut être complètement ignorée. Par exemple si un type atomique est lié à des atomes de charge différente, on souhaite avoir une orientation moyenne qui tienne compte des différences d'orientation dans les molécules initiales (cette considération est liée aux observations de la partie 1.7.d et 1.7.i, page 86 et 99).

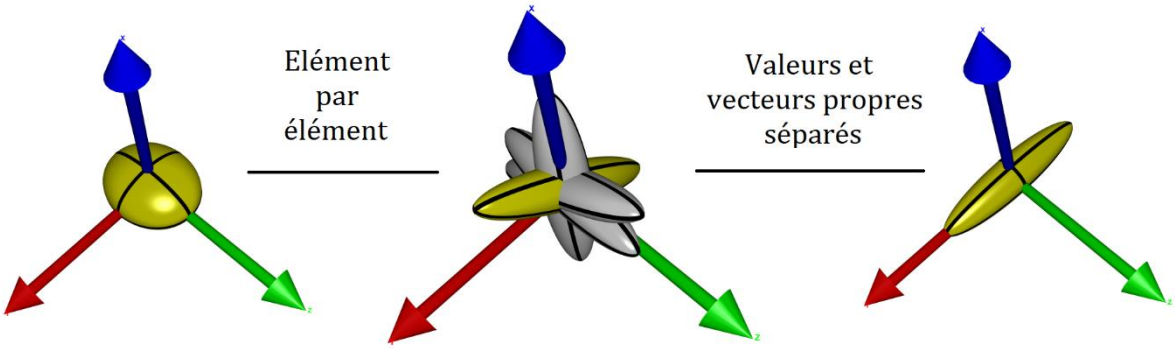


Figure 36. Moyennes d'un échantillon de tenseurs de polarisabilité (aucune contrainte de symétrie). Au centre, les 4 tenseurs de l'échantillon ont été moyennés dans le référentiel représenté par les flèches. Les valeurs propres de chaque tenseur sont identiques, mais leurs orientations sont hétérogènes. L'ellipsoïde jaune de l'échantillon est celle dont les vecteurs propres sont choisis comme référence $\mathbf{v}_{ref}$ [équation (72)]. À gauche, le tenseur obtenu en moyennant chaque composant $\boldsymbol{\alpha}_{ij}$ indépendamment. À droite, le tenseur obtenu en utilisant la méthode conçue pour *MoProViewer*.

Choisir le référentiel atomique local comme base de trois vecteurs de référence pour la moyenne de l'équation (72) aurait été possible. Sans être certain que ce choix introduise un biais dans la moyenne présentée, il a été préféré de tenter de concevoir une méthode indépendante. La figure 36 compare le résultat de la méthode présentée à une moyenne élément par élément.

Tout l'enjeu de la méthode est de conserver les caractères des polarisabilités originales dans la moyenne calculée. Il y a trois cas possibles :

• La polarisabilité est parfaitement aplatie (*oblate*) : la valeur propre $\lambda_1 = \lambda_2$, par conséquent les premiers et seconds vecteurs propres seront une paire de vecteurs orthogonaux dans le plan normal au troisième vecteur propre.

• La polarisabilité est parfaitement oblongue (*prolate*) : la valeur propre $\lambda_2 = \lambda_3$, par conséquent les seconds et troisièmes vecteurs propres seront une paire de vecteurs orthogonaux dans le plan normal au premier vecteur propre.

• la polarisabilité est sphérique, $\lambda_1 = \lambda_2 = \lambda_3$, chaque vecteur de l'espace est un vecteur propre.

Dans la figure 36, la moyenne élément par élément à un caractère aplati (*oblate*) qu'on ne retrouve dans aucun membre de l'échantillon à moyenner. L'échantillon de tenseurs a bien entendu été choisi pour mettre ceci en évidence. Lorsqu'on compare ce résultat à celui de la méthode des valeurs et vecteurs propres séparés, le tenseur moyenné a un caractère allongé. Cependant, l'orientation du premier tenseur de l'échantillon (en jaune) est critique pour l'orientation de la moyenne obtenue. Les vecteurs propres de ce tenseur sont les $\mathbf{v}_{ref}$ de l'équation (72). Aucune contrainte sur les paramètres hors diagonaux n'a été appliquée (les contraintes de symétrie sont introduites après). Pourtant l'orientation de la moyenne apparait fortement biaisée vers l'orientation du tenseur dont on utilise les vecteurs propres comme référence. Une possible interprétation de ce résultat est que l'on « déplace » le problème d'inhomogénéité de l'échantillon : pour éviter d'avoir une moyenne aberrante sur un échantillon hétérogène, on suppose que l'un des membres de l'échantillon est homogène au reste de l'échantillon. La conséquence de ce choix est que, dans cette méthode, l'orientation du tenseur de référence prime sur l'orientation de tous les autres. Pour l'exemple de la figure 36, la polarisabilité de référence (en jaune) a été délibérément choisie avec une orientation très déviante, pour montrer l'importance de cette sélection (ce choix n'influe pas la moyenne élément par éléments). Dans la plupart des cas pratiques, les anomalies sont rares et cette méthode permet d'améliorer efficacement la conservation des paramètres topologiques de la moyenne.

Il existe deux limitations à la méthode présentée ici :

• Cette méthode ne garantit pas que les vecteurs moyennés soient orthogonaux entre eux. Si plusieurs vecteurs de directions très différentes sont moyennés entre eux, il n'y a aucune garantie que la base de vecteurs propres soit orthogonale. À cette étape, pour un type atomique donné, on obtient donc $\langle \mathbf{v}_1 \rangle$, $\langle \mathbf{v}_2 \rangle$ et $\langle \mathbf{v}_3 \rangle$ associés à $\langle \lambda_1 \rangle$, $\langle \lambda_2 \rangle$ et $\langle \lambda_3 \rangle$. Comme la plupart des polarisabilités atomiques ont un caractère oblong (voir Figure 15 page 90), et pour obtenir une erreur plus faible sur les plus grands éléments du tenseur moyenné, il est naturel de modifier $\langle \mathbf{v}_3 \rangle$ en premier pour orthogonaliser la base ($\langle \mathbf{v}_1 \rangle$, $\langle \mathbf{v}_2 \rangle$, $\langle \mathbf{v}_3 \rangle$), en suivant la procédure ci-dessous :

$$\mathbf{v}_{3,orth} = \langle \mathbf{v}_1 \rangle \times \langle \mathbf{v}_2 \rangle$$

$$\mathbf{v}_{2,orth} = \langle \mathbf{v}_2 \rangle - 0.5 \left(\langle \mathbf{v}_1 \rangle \cdot \langle \mathbf{v}_2 \rangle \right) \langle \mathbf{v}_1 \rangle \qquad \mathbf{(73)}$$

$$\mathbf{v}_{1,orth} = \langle \mathbf{v}_1 \rangle - 0.5 \left(\langle \mathbf{v}_1 \rangle \cdot \langle \mathbf{v}_2 \rangle \right) \langle \mathbf{v}_2 \rangle$$

Le terme $0.5(\langle \mathbf{v}_1 \rangle \cdot \langle \mathbf{v}_2 \rangle)$ permet de répartir l'erreur sur l'orthogonalité équitablement entre $\mathbf{v}_{1,orth}$ et $\mathbf{v}_{2,orth}$. Les trois vecteurs propres sont ensuite normalisés et si $\mathbf{v}_{1,orth} \cdot \mathbf{v}_{2,orth} > 0.0002$ (seuil arbitraire), la procédure est répétée de manière récursive jusqu'à orthogonalité. À ce moment, on recompose le tenseur moyen de polarisabilité dans le référentiel atomique, prêt à être utilisé pour le transfert sur un système cible.

• Si, par malchance, la liste de tenseurs à moyenner commence par une anomalie (*outlier*), l'orientation de cette anomalie sera considérée comme l'orientation de référence. Il reste donc nécessaire de vérifier manuellement si les polarisabilités moyennes ont une orientation sensée. Un critère de qualité qui peut être considéré est si le vecteur propre associé à λ_1 est bien orienté vers le groupement voisin le plus électronégatif. Pour l'instant, cette vérification manuelle est fortement recommandable.

Pour construire la base de données présentée partie 3.6, on part d'un set de molécules – associées aux polarisabilités à moyenner – contenant à la fois des molécules isolées et d'autres en interaction (dimères, dipeptides, etc., voir partie 1.8). Pour tenter d'obtenir l'orientation la plus transférable possible, on s'est assuré que les molécules isolées ont été chargées en mémoire avant les autres, afin que, dans la mesure du possible, les $\mathbf{v}_{\text{ref}}$ utilisés pour chaque type atomique correspondent à ceux des polarisabilités des molécules isolées.

Considérant les caractéristiques mathématiques de notre moyenne, en particulier sa sensibilité au choix de l'orientation de référence, il faut une méthode pour quantifier la dispersion de l'échantillon qui soit indépendante de la présente procédure. Pour cela, on quantifiera toujours la dispersion de chacun des 6 paramètres indépendants de la polarisabilité moyenne par sa déviation standard [*SSD*, équation (A7), apge 292], plutôt que de quantifier la dispersion des valeurs propres ou des vecteurs propres. Utiliser le *SSD*, calculé indépendamment pour chaque composant du tenseur, permet de comparer différentes procédures de manière consistante.

Pour améliorer encore cette méthode, rendre totalement univoque le choix des $\mathbf{v}_{ref}$ serait désirable. Sur la Figure 36, on peut constater que si la forme du tenseur moyenné élément par élément diffère des tenseurs originaux, son orientation apparaît comme un bon compromis entre celles des tenseurs à moyenner. Il est donc possible d'utiliser la base propre de la moyenne élément par élément comme orientation de référence pour la moyenne avec vecteurs et valeurs propre séparés. La matrice $\mathbf{A}$ contient les vecteurs propres de la matrice suivante :

$$\langle \boldsymbol{\alpha} \rangle_{\text{élément}-\text{élément}} = \begin{bmatrix} \langle \boldsymbol{\alpha}_{xx} \rangle & \langle \boldsymbol{\alpha}_{xy} \rangle & \langle \boldsymbol{\alpha}_{xz} \rangle \\ \langle \boldsymbol{\alpha}_{xy} \rangle & \langle \boldsymbol{\alpha}_{yy} \rangle & \langle \boldsymbol{\alpha}_{yz} \rangle \\ \langle \boldsymbol{\alpha}_{xz} \rangle & \langle \boldsymbol{\alpha}_{yz} \rangle & \langle \boldsymbol{\alpha}_{zz} \rangle \end{bmatrix} = \mathbf{A}^{-1} \begin{bmatrix} \lambda_{1,\langle \boldsymbol{\alpha} \rangle} & & \\ & \lambda_{2,\langle \boldsymbol{\alpha} \rangle} & \\ & & \lambda_{3,\langle \boldsymbol{\alpha} \rangle} \end{bmatrix} \mathbf{A}$$

$$\mathbf{A} = [\mathbf{v}_{ref,x} \quad \mathbf{v}_{ref,y} \quad \mathbf{v}_{ref,z}] \qquad (\mathbf{74})$$

Les résultats de cette méthode pour trouver les trois $\mathbf{v}_{\text{ref}}$ devraient être très comparables avec la méthode présentée pour des échantillons homogènes. Dans la limite d'un très grand nombre de tenseurs aléatoires moyennés, le référentiel de référence obtenu paur l'équation (72) devrait converger vers celui de l'équation (74) (à une permutation d'axes près). L'effet des hétérogénéités et anomalies (*outliers*) est quant à lui moins prédictible.

Cette méthode a été testée pour la construction de la base de données de la partie 3.6. Les polarisabilités obtenues diffèrent de 10^{-4} Å^3 au maximum de celles obtenues en utilisant la méthode précédente telle que décrite plus tôt (la base de données est construite à partir de 4967 polarisabilités atomiques, voir partie 1.8). Devant cet accord surprenant, cette méthode est devenue la méthode utilisée par défaut par le module « *Beluga* » lors de la moyennation de tenseurs de polarisabilités. D'autre part, cet accord peut être interprété comme un signe d'homogénéité des échantillons étudiés.

3.5 - Ajout de contraintes de symétrie à la méthode précédente

La partie suivante détaille l'adaptation de l'algorithme aux symétries locales des types atomiques. En effet, le « *Big Set* » contient de nombreuses molécules en interaction (dipeptides et dimères essentiellement). Les polarisabilités atomiques topologiques ne sont pas sensibles uniquement à l'environnement covalent. Une base de données construite à partir du « *Big Set* » pourrait gagner en transférabilité en incluant les contraintes de symétries locales des types atomiques. Comme détaillé la parties 1.7.h et 1.7.i, les polarisabilités calculées pour le S66 peuvent varier entre l'état isolé et le cas où deux molécules sont en interaction sous forme dimérique. On souhaite, pour notre modèle transférable, que ces influences extra-moléculaires soient représentées de manière statistique. Il a été choisi d'utiliser les symétries locales des atomes ELMAM2 pour contraindre ces influences sans les exclure complètement. Si la polarisabilité est une propriété électronique, il semble naturel qu'elle hérite de la symétrie de la densité électronique, au moins dans une certaine mesure. On s'attend naturellement à des performances supérieures des polarisabilités transférables calculées en intégrant les contraintes de symétrie du type atomique.

Pour obtenir des moyennes tensorielles incluant des contraintes de symétrie héritées du type atomique, une méthode basée sur des matrices représentatives de l'opération de symétrie a été implémentée dans *MoProViewer*. Chaque opérateur de symétrie locale, défini dans le format ELMAM2, est associé avec une liste de matrices représentatives des opérations associées. Chaque matrice de cette liste permet d'obtenir l'image d'un vecteur ou d'un tenseur par l'une des opérations de symétrie du groupe. Les matrices implémentées sont disponibles en annexe, par opérateur de symétrie local ELMAM2.

Le principe général est le suivant : un type atomique est associé à plusieurs polarisabilités à moyenner ; on génère toutes les images de chaque polarisabilité initiale par les opérations de symétrie du type atomique. On moyenne ensuite cette nouvelle liste de tenseurs selon la procédure décrite plus haut (valeurs propres moyennes affectées aux vecteurs propres de la moyenne éléments par éléments). Chaque polarisabilité symétrique apparaîtra *n*-fois dans la moyenne où *n* est le nombre d'opérations du groupe (comptant l'identité). Le poids relatif de chaque polarisabilité par rapport aux autres est inchangé comparé à la moyenne sans contrainte de symétrie.

L'opération de symétrie 'cy' est un cas particulier. Elle représente une symétrie cylindrique autour de l'axe local **z** de l'atome. Pour moyenner les tenseurs de polarisabilité, la symétrie 'cy' est approximée par '`mxmy`'. Un second algorithme moins élégant « répare » ensuite le tenseur en forçant :

$$\lambda_2 = \lambda_3 = \frac{\lambda_2+\lambda_3}{2} \qquad (75)$$

À la page 51, une méthode pour obtenir une estimation de la distance entre deux tenseurs est décrite. Elle est utilisée pour quantifier la transférabilité les *ADP* dans l'article où elle a été publiée[71]. L'idée de moyenne repose sur celle de distance entre objets, la moyenne minimise la distance entre elle-même et chaque membre de l'échantillon. Il est important de noter qu'en utilisant la distance de l'équation (42), seule la moyenne élément par élément respecte cette propriété. La moyenne par vecteurs et valeurs propres séparées est systématiquement plus éloignée de chaque membre de l'échantillon à moyenner. L'ajout de contraintes de symétrie autres que l'opération identité augmente également cette distance à tous les membres de l'échantillon à moyenner. Cette augmentation reste marginale en pratique mais nécessite d'être notée.

3.6 - Construction de librairie de polarisabilités indépendantes

La littérature[86] ainsi que le travail présenté[52] ont montré la faisabilité de construire ce type de base de données. Leur application est pertinente pour la prédiction des polarisabilités moléculaires et des énergies de polarisation.

L'implémentation actuelle de l'outil décrit dans ce chapitre écrit le résultat préformaté pour être copié dans la base de données, ce qui rend aisé la création d'une base de polarisabilités atomiques utilisant les types atomiques ELMAM2. L'outil est maintenant appliqué à la création d'une base de données à partir du « *Big Set* » décrit partie 1.8. La liste des 198 structures moléculaires est disponible en annexe (Table A3). 38 structures parmi ces 198 représentent des acides aminés isolés ou des petites molécules isolées. Ces 38 structures de molécules isolées seront désignées par la suite comme « monomères » du « *Big Set* » (ce groupe contient les monomères du S66, voir partie 1.7.a). La méthode décrite précédemment, incluant les contraintes de symétrie, est

utilisée pour obtenir les moyennes des tenseurs de polarisabilité (pour comparaison, une base de données utilisant la méthode élément par élément a également été générée).

Table 7. Vue d'ensemble des polarisabilités transférables de la base pELMAM3. $n(\boldsymbol{\alpha}_i)$ désigne le nombre de polarisabilités atomiques moyennées, par élément chimique ou par type atomiques. * : tous types atomiques confondus.

Élément chimique	$n(\boldsymbol{\alpha}_i)$, total	Nombre de types atomiques	Maximum de $n(\boldsymbol{\alpha}_i)$ par type atomique	Nombre de types atomiques avec $n(\boldsymbol{\alpha}_i)$=1	$\langle\boldsymbol{\alpha}_{iso}\rangle$ (Å^3), *	⟨charge⟩ (e), *
H	2623	21	615 (H, méthyle)	2	0.263	+0.112
C	1417	32	241 (C, aromatique)	2	1.027	-0.008
N	336	9	122 (N, ammonium primaire)	1	1.328	-0.172
O	576	12	248 (O, carboxylate)	1	1.100	-0.258
F	2	1	2 (F, fluoro-)	0	0.741	-0.012
P	1	1	1 (P, phosphate)	1	0.932	-0.026
S	10	4	7 (S, méthionine)	3	2.651	-0.100
Cl	2	1	2 (Cl, chlorure)	0	3.295	-1.000

Les types atomiques utilisés pour catégoriser les atomes sont ceux de la librairie ELMAM2, complémentés des modifications listées au chapitre 3-bis. En tout cette version de la librairie contient 153 types atomiques. Pour refléter ces différences avec les versions précédentes, le nom 'pELMAM3' est proposé pour la librairie, afin de souligner le caractère polarisable des pseudoatomes transférables. Pour décrire les 4967 atomes représentés dans les structures de départ, 81 types atomiques de la base de données ont été nécessaires.

Tous les types d'atomes ELMAM2 ne sont donc pas couverts. Certains types de la version utilisée sont cependant des types temporaires, laissés là par d'autres utilisateurs

pour des applications spécifiques. La nombre de polarisabilités moyennées par types atomiques (la *couverture* de la librairie) varie également : certains types sont associés à de grands nombres de polarisabilités. D'autres n'apparaissent qu'une seule fois dans le « *Big Set* ». Si la couverture de la base ELMAM2 est incomplète, il y a une nuance à apporter cependant : tous les types atomiques ne sont pas aussi fréquents les uns que les autres dans les molécules étudiées. Le « *Big Set* » a été conçu pour couvrir les types utiles à la modélisation biomoléculaire et tous les types atomiques des acides aminés biogéniques sont largement couverts.

La Table 7 montre ces différences entre types atomiques en mettant en avant le type atomique le plus représenté, et le nombre types atomiques représentés une seule fois dans l'ensemble de structures considérées. Premièrement, les atomes d'hydrogène représentent plus de la moitié des atomes du « *Big Set* ». Les hydrogènes des groupements méthyle, méthylène et ammonium primaire sont représentés plus de 300 fois. Aucun autre élément chimique ne possède de type atomique associés avec autant de polarisabilités atomiques théoriques. Les types atomiques associées à une seule polarisabilité*** concernent l'hydrogène thiol de la cystéine et un hydrogène méthylène de la biotine.

Le carbone est très représenté, mais est associé à un plus grand nombre de types atomiques. C'est l'atome avec les polarisabilités théoriques les plus variables (voir partie 1.7.e). Les deux types atomiques les plus représentés sont le carbone aromatique ($n(\boldsymbol{\alpha}_i)$ = 241) et le carbone méthyle ($n(\boldsymbol{\alpha}_i)$ = 201). Le carbone est lui retrouvé 162 fois dans le « *Big Set* ». Les deux types atomiques associés à une seule polarisabilité sont le carbone guanidium de l'arginine et le carbone imidamine de la benzamidine.

L'oxygène est le troisième atome le plus représenté. L'oxygène carboxylate et l'oxygène amide sont associés avec le plus grand nombre de polarisabilités théoriques, respectivement, 248 et 146. Le « *Big Set* » comporte 78 molécules d'eau dont les polarisabilités ont été calculées. Le type d'oxygène associé à une seule polarisabilité est l'oxygène sp^3 de la fonction ester.

*** *: Ces types atomiques associés à une seule polarisabilité incluent uniquement les contraintes de symétrie du type atomique correspondant pour obtenir la polarisabilité transférable.*

L'azote est le dernier élément suffisamment représenté dans le jeu de données considéré. Les types d'azote amide et ammonium primaire sont les plus représentés (respectivement $n(\alpha_i)$ = 119 et $n(\alpha_i)$ = 122). L'ammonium secondaire de la proline correspond au seul type d'azote associé à une seule polarisabilité théorique.

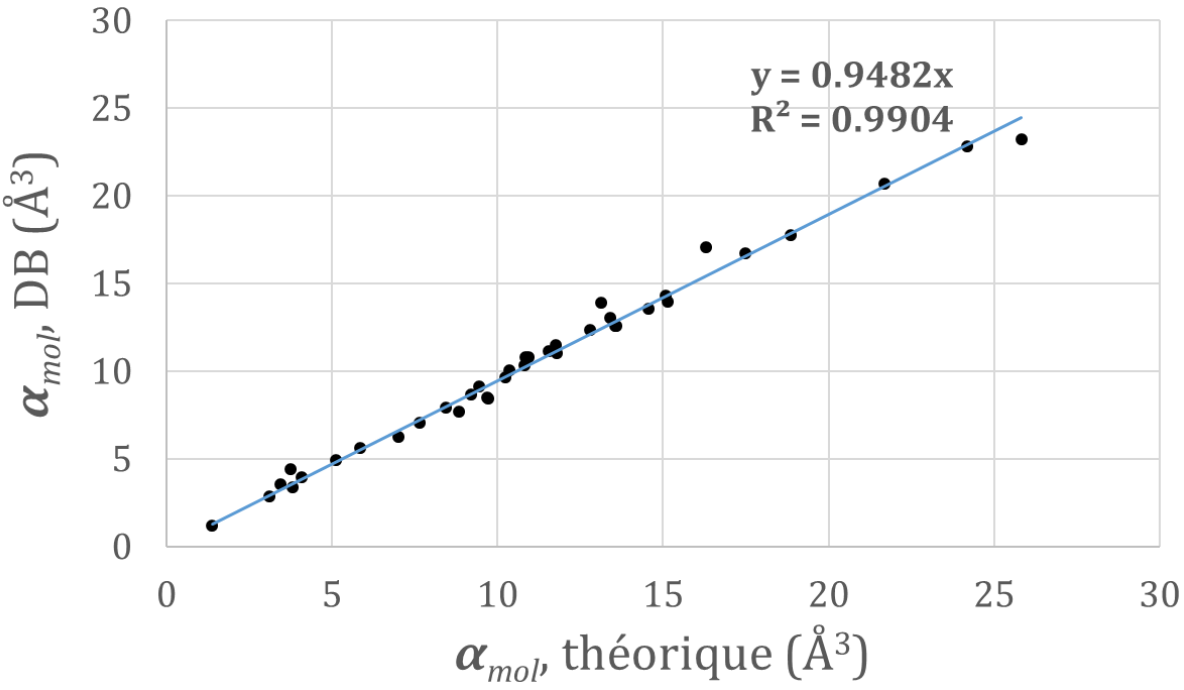


Figure 37. Corrélation entre les polarisabilités moléculaires théoriques des 38 molécules isolées du « *Big Set* » (α_{mol}, théoriques, voir Table 5, page 105) et les polarisabilités moléculaires reconstruites à partir de la base de données pELMAM3, générée à partir des 198 structures de la Table A3.

Les polarisabilités transférables doivent être évaluées pour leur capacité à reproduire les polarisabilités moléculaires. Du fait que la plupart des conformations moléculaires du « *Big Set* » sont issues de structures cristallines, il est impossible – dans le temps de la thèse – d'obtenir un banc de test expérimental adapté. On procède donc à un contrôle interne : les polarisabilités transférables, obtenues à partir du « *Big Set* » sont transférées sur les 38 monomères évoqués plus haut. On compare, Figure 37, la polarisabilité moléculaire isotrope théorique avec la polarisabilité moléculaire isotrope transférée. L'accord est excellent, la pente de régression est proche de l'unité et le coefficient de détermination vaut 0.990. La légère sous-estimation (≈ -5 %) est tout de

même notable. Elle peut être relative aux différences d'orientation entre les polarisabilités théoriques et transférées. Une autre possibilité serait un effet indirect d'échantillonnage. En effet toutes les molécules de la Figure 37 sont isolées, et le « *Big Set* » contient une majorité de molécules en interaction (dimères et dipeptides).

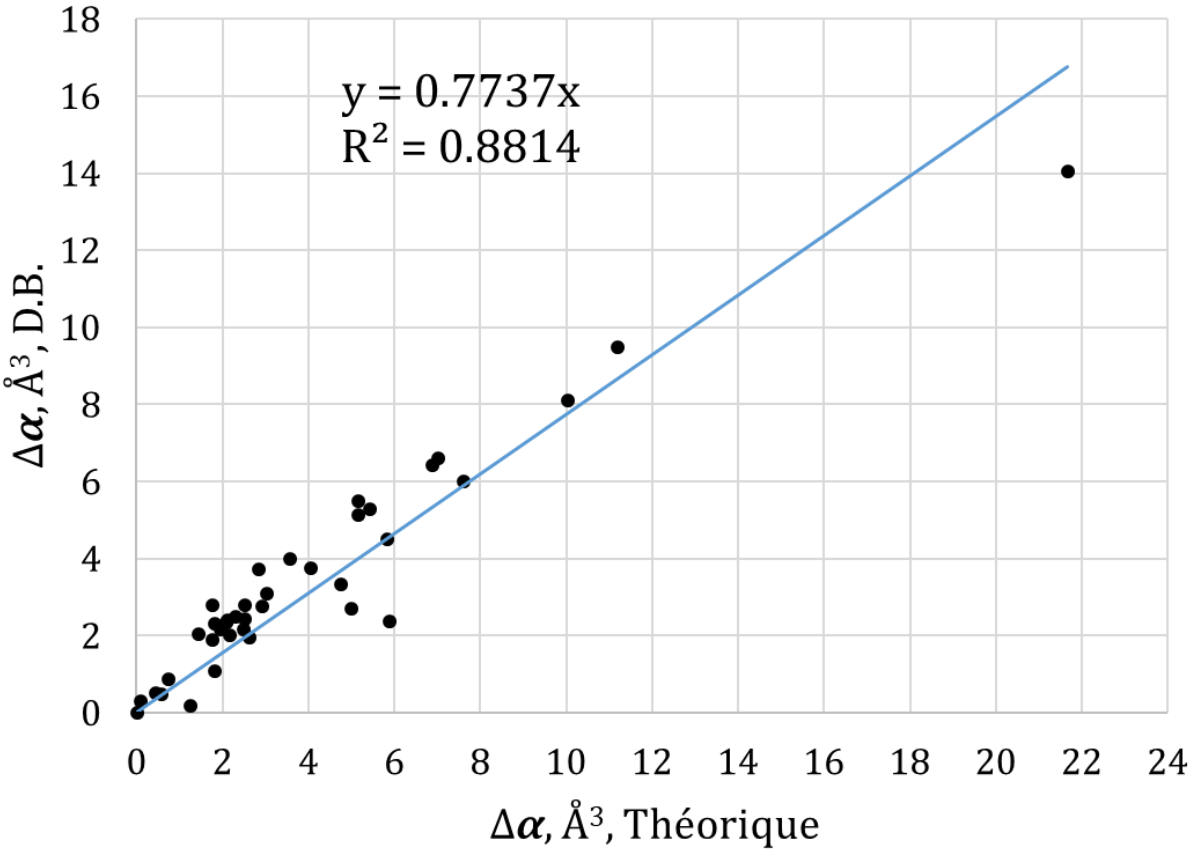


Figure 38. Corrélation entre l'anisotropie $\Delta\alpha$ (Å^3) des polarisabilités moléculaires théoriques et des polarisabilités moléculaires reconstruites à l'aide de la base de données pELMAM3. $\Delta\alpha$ vaut 0 dans le cas d'une polarisabilité parfaitement sphérique.

Tout l'intérêt de générer une base de données de polarisabilités anisotropes additives est de représenter l'anisotropie des polarisabilités moléculaires. Dos Santos, Krawczuk et Macchi, 2015[86], utilisent l'indicateur $\Delta\alpha$ suivant pour caractériser l'anisotropie des tenseurs de polarisabilité :

$$\Delta\alpha = \left\{\frac{3\cdot\mathrm{trace}(\alpha^2)-\mathrm{trace}(\alpha)^2}{2}\right\}^{\frac{1}{2}} \quad (76)$$

Pour les 38 monomères du « *Big Set* », les valeurs de $\Delta\boldsymbol{\alpha}$ ont été calculées pour les polarisabilités moléculaires théoriques et transférées. La régression linéaire de la figure 38 montre que cette propriété est globalement sous-estimée, pourtant un grand nombre de points sont présents à proximité de la diagonale $x = y$. La pente de régression à 0.774 est donc due aux quelques monomères pour lesquels l'anisotropie est la moins bien reproduite. Par exemple, on sous-estime l'anisotropie du cyclopentane ($\Delta\boldsymbol{\alpha}_{théorique}$ = 1.26 Å^3 ; $\Delta\boldsymbol{\alpha}_{transféré}$ = 0.17 Å^3), et les polarisabilités de l'eau sont trop anisotropes[†††] ($\Delta\boldsymbol{\alpha}_{théorique}$ = 0.10 Å^3 ; $\Delta\boldsymbol{\alpha}_{transféré}$ = 0.29 Å^3).

Utiliser les moyennes élément par élément permet d'avoir un bon accord sur les polarisabilités isotropes (R^2 = 0.991, données non-communiquées), mais, comme attendu, l'anisotropie de la polarisabilité est moins bien reproduite (pente = 0.752, données non-communiquées). Cette différence est principalement due aux molécules dont l'anisotropie théorique est sous-estimée. Cette sous-estimation est plus importante avec la méthode de moyennation élément par élément, ce qui est un argument pour l'utilisation des moyennes par valeurs propres et vecteurs propres séparées (plus contraintes de symétrie).

Les polarisabilités atomiques obtenues sont associées aux paramètres de densité électronique ELMAM2. Il est donc tentant de chercher des points communs entre les types atomiques qui ont des polarisabilités proches. Deux paramètres étudiés dans ce contexte sont κ et κ', les coefficients d'expansion-contraction de la densité de valence et de déformation (respectivement). Puisqu'ils décrivent l'étalement de la densité électronique dans l'espace, on s'attend à une polarisabilité plus grande pour les atomes dont la densité électronique est la plus diffuse (κ et $\kappa' < 1$). Malheureusement, κ et κ' n'apparaissent pas corrélés avec les polarisabilités isotropes transférables. Les variations relatives de κ et κ' sont en effet infimes par rapport aux variations relatives de $\boldsymbol{\alpha}_{iso}$.

À la partie 1.7.d, il a été noté que les polarisabilités théoriques du S66 sont corrélées aux charges intégrées et aux volumes des bassins topologiques de densité électronique. Si les charges HCM diffèrent fortement des charges topologiques, on sait que ces deux définitions de la charge atomique donnent des résultats corrélés. La Figure 39

[†††] *: On rappellera que la plupart (77/78) des molécules d'eau du « Big Set » forment des liaisons hydrogène avec d'autres molécules.*

montre que cette corrélation charge-polarisabilité est retrouvée dans la base de données qui a été construite.

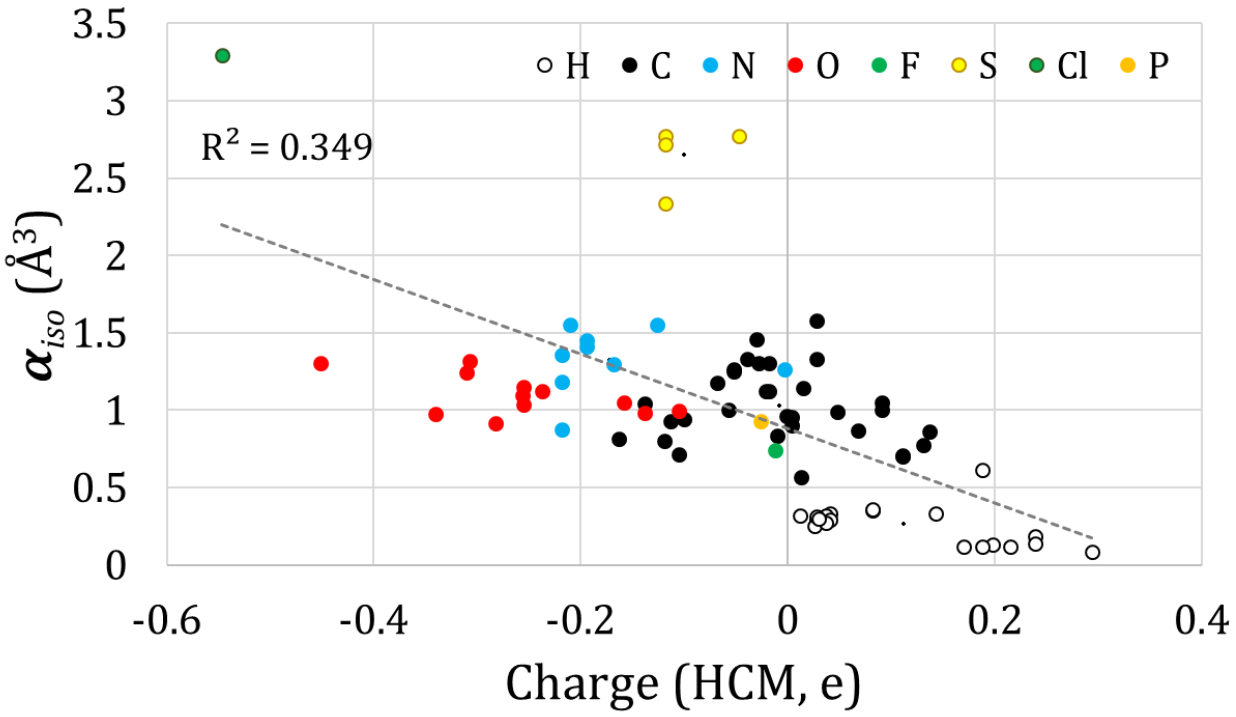


Figure 39. Corrélation entre charges multipolaires transférables et polarisabilités isotropes transférables. La régression linéaire utilise les 81 types atomiques pour lesquels une polarisabilité moyenne est obtenue.

Il s'agit cependant d'une tendance globale, légèrement moins marquée que celle obtenue pour les charges topologiques. Une figure similaire à la Figure 39 est trouvée page 86, mais en comparant cette fois la polarisabilité aux charges intégrées. Cette tendance et moins bonne car les charges multipolaires sont beaucoup plus variables que les charges topologiques. On observe tout de même des régions où l'on retrouve de nombreux types atomiques associés aux mêmes éléments chimiques.

Une autre possibilité serait d'étudier à nouveau la relation entre les incertitudes sur les moments dipolaires et les polarisabilités des types atomiques, comme vu partie 1.7.g, page 95. En vue d'une publication des résultats présentés, cette comparaison sera effectuée à l'avenir.

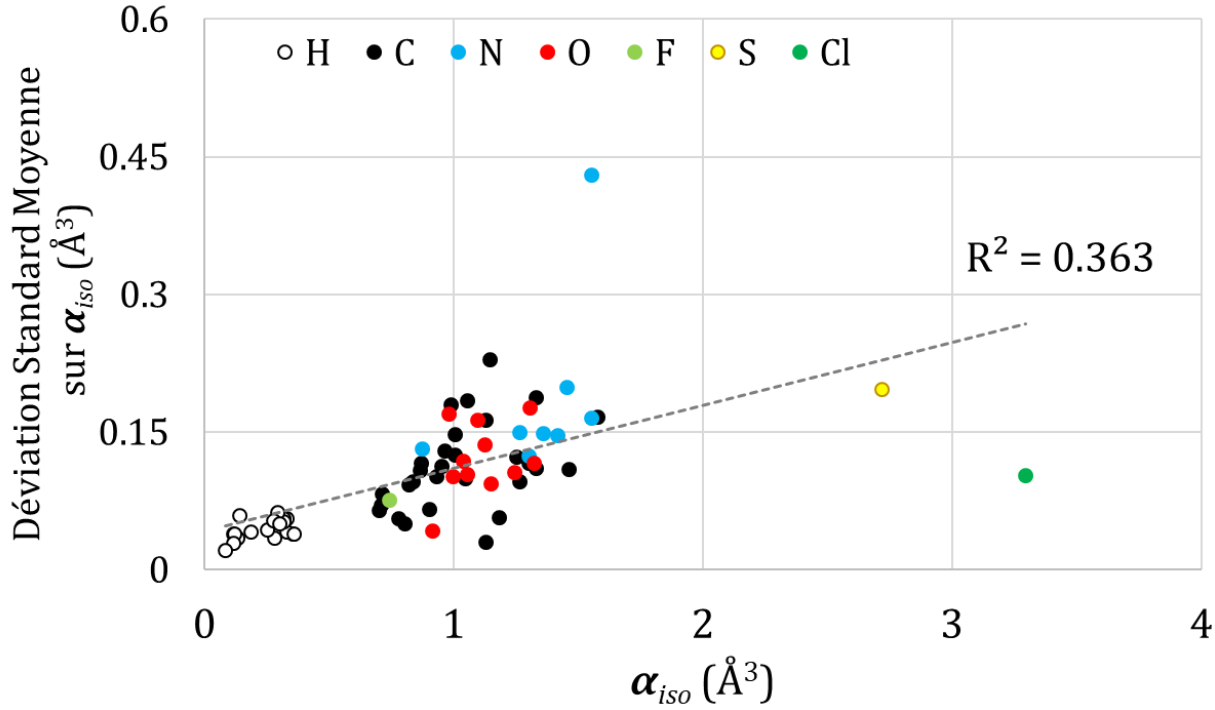


Figure 40. Corrélation entre polarisabilités isotropes de la base de données et la moyenne des *SSD* associés aux 6 éléments indépendants du tenseur $\boldsymbol{\alpha}$ (noté Déviation Standard Moyenne). Seuls les 70 types atomiques associés à plusieurs polarisabilités théoriques sont montrés ici [voir équation (A7) en annexe].

Pour quantifier la dispersion des polarisabilités moyennées, la moyenne des *SSD*, sur les 6 composants indépendants des polarisabilités transférables, ont été examinés (et désignées Figure 40 par le nom déviation standard moyenne). On observe une tendance relative. Les plus fortes polarisabilités sont associées à de plus fortes déviations. Cependant, cette déviation moyenne est inférieure à 0.25 Å^3 pour tous les types atomiques considérés, sauf un, associé à l'azote de la proline lié en N-ter. Cet azote n'est retrouvé que fois dans le « *Big Set* » et est associé à un *SSD* moyen de 0.43 Å^3. La dispersion des polarisabilités moyennées sur les atomes d'hydrogène est relativement forte, en effet, s'ils sont associés aux plus faibles *SSD* moyens, leur polarisabilité est aussi faible. Ces *SSD* moyens représentent de 10 à 40 % des $\boldsymbol{\alpha}_{iso}$ des types d'hydrogènes de la base de données.

La répartition des erreurs moyennes par rapport au nombre de polarisabilités atomiques moyennes $n(\boldsymbol{\alpha}_i)$ doit être examinée. Il est bien entendu possible de réaliser une étude bien plus approfondie de cette base de données de polarisabilités atomiques. On souhaite cependant sonder son applicabilité à un système biochimique.

3.7 - Applications biochimiques

3.7.a - Interactions protéine-ligand : généralités et modélisations.

Le développement de base de données SILICA (*Structure and ITC Linked Index for Computational Applications*) a été amorcée durant mon stage de Master 2. L'un des objectifs originaux de la présente thèse était de tester extensivement la méthode de polarisation proposée sur un ou plusieurs systèmes d'intérêt biochimique. Assembler un jeu de données expérimentales pour le test de méthodes computationnelles est un projet attirant mais chronophage. La construction de cette base de données a été suspendue au terme du stage mais c'est une ressource très utile pour choisir une étude de cas. Sa mise à jour pourrait tout à fait être reprise à tout moment.

La méthode de l'ITC a été décrite dans l'état de l'art, page 52. Considérant que les mesures ITC sont de nos jours moins routinières que celles de biocristallographie, la base de données SILICA a été construite sur le principe suivant : à partir de l'analyse de revues de la littérature sur cette technique, les systèmes testés par ITC ont été recherchés sur la PDB afin de trouver les structures de cocristaux protéine-ligand correspondants. Pour qu'une entrée soit retenue, il faut que le ligand testé soit le même dans la titration et le cocristal. Les mesures d'ITC peuvent utiliser sans restriction des protéines multidomaines, complètes et biologiquement actives. Seule la présence du domaine de fixation dans le cocristal a été retenu comme critère de sélection, les domaines globulaires isolés étant notablement plus faciles à cristalliser.

88 structures ont pu être retenues, mais ne correspondent pas toutes à des données ITC ou cristallographiques de très haute qualité. La Table A4 regroupe une partie des couples ITC – structures PDB (ceux dont les références bibliographiques sont connues). Quatre complexes de SILICA ont été envisagés : 3RY2, 1SWK, 1NDJ et 1TIO. Les structures 3RY2, 1SWK et 1NDJ représentent le complexe streptavidine biotine. 1TIO est une structure de trypsine inhibée par la benzamidine et est abordée partie 4.1. Le chapitre 4 aborde aussi la structure de l'arginase mutante 1HQG, complexée aux produits de réaction urée et ornithine, envisagée pour d'autres comparaisons que celles basées sur l'ITC.

Préparation des structures :

La densité électronique ELMAM2 a été transférée sur chaque complexe mentionné ci-dessus. Pour que cette opération soit possible, les positions des atomes d'hydrogène doivent être déterminées à partir des positions des atomes lourds. Le logiciel *Reduce*[103] (disponible sur le serveur *MolProbity*[104]) permet d'ajouter les atomes d'hydrogène à une structure protéique expérimentale et a été utilisé sur chaque complexe. Après le transfert, la neutralisation de la charge des fragments transférés est nécessaire. La neutralisation passe, dans *MoProViewer*, par une méthode qui divise et partage l'écart de charge permettant d'obtenir un fragment neutre entre les populations P_v des atomes du fragment. P_v est en effet un attribut transférable, les atomes qui partagent un type atomique ont donc la même P_v. Empiriquement, il a été constaté que le transfert d'atomes ELMAM2 tend à exagérer légèrement le nombre d'électrons d'une molécule. La neutralisation des fragments est également possible avec *MoPro* depuis les débuts d'ELMAM[21,61] et a été réimplémenté dans *MoProViewer*. L'électroneutralité de l'unité asymétrique est bien entendu cruciale en cristallographie.

Lorsque la qualité des données cristallographiques le permet, des conformations alternatives de chaines latérales (ou du squelette peptidique) sont modélisées dans une structure cristalline de protéine. Pour les simulations présentées, seule la conformation majoritaire de chaque structure a été étudiée. Elle est définie comme regroupant tous les atomes dont l'occupation atomique est supérieure à 0.5.

Le passage à une charge formelle d'un groupement chimique donné est géré de la même manière. Par exemple, on modifiera la P_v de chaque atome d'un Aspartate pour amener la charge du fragment à -1 e. En général chaque atome reçoit ou perd entre 10^{-2} et 10^{-3} électrons, indifféremment pour les espèces neutres ou chargées. La procédure aboutit usuellement à des charges moléculaires à moins de 10^{-3} électrons de la charge formelle pour 1NDJ, 1SWK et 1HQG (respectivement 1723, 1720 et 5023 atomes).

Cette méthode permet toujours d'obtenir une charge très proche de la charge formelle pour un complexe de la taille d'une protéine. Par contre, si les acides aminés ont une composition atomique répétitive, leur fréquence varie de protéine en protéine.

Pour l'exemple, on souhaite neutraliser deux pro-peptides non-chargés : Gly-Gly-Ala et Gly-Ala-Ala. Ces deux pro-peptides sont composés respectivement de 24 et 27

atomes. Pour l'exemple, la charge après transfert est +0.3 électrons dans les deux cas. Cette déviation de la charge formelle dépend autant du nombre d'atomes que de la combinaison de types atomiques ELMAM2 utilisés. Il est donc aisé de comprendre que la charge des résidus Glycine sera différente dans les deux pro-peptides. Les glycines de GGA recevront 0.0875 électrons et la glycine de GAA recevra 0.0778 électrons, alors que dans les deux cas, les P_v transférées sont identiques.

Pour éviter que les charges des résidus dépendent de la séquence de la protéine, la procédure suivante est employée. Elle est basée sur le transfert et la neutralisation séquentielle de chaque type de résidu de la chaîne protéique.

On commence par transférer séparément les extrémités N-ter et C-ter de la protéine. Lors de l'étape d'ajustement de la charge, ils sont respectivement fixés à une charge de +1 e et de -1 e.

Ensuite on sélectionne tous les acides aminés identiques (considérant leur état de protonation). Ils seront transférés ensemble et fixés à leur charge formelle. Un groupe de 6 lysines réparties sur la séquence sera fixé à une charge de +6 e. Ainsi, la fréquence de chaque type atomique est identique quel que soit le nombre d'acides aminés sélectionnés, ce qui assure que chaque atome reçoit une charge identique dans les mêmes acides aminés. Les éventuels ligands de la protéine reçoivent le même traitement et sont fixés à leur propre charge formelle.

De protéine en protéine, cette méthode garantit que le même acide aminé aura les mêmes P_v, tout en garantissant que la structure totale est raisonnablement proche de sa charge formelle. Le principal intérêt de cette approche est de rendre consistante (du point de vue de la charge) l'étude parallèle de plusieurs protéines et donc de rendre les comparaisons de propriétés électrostatiques plus solides et pertinentes.

Certaines des protéines préparées contiennent des ions (trypsine 1TIO et arginase 1HQG, voir chapitre 4). Certains ont un rôle structural comme le Ca^{2+} dans la trypsine, d'autres ont un rôle catalytique (Mn^{2+} dans l'arginase) et certains sont présents du fait de l'expérience de cristallisation (SO_4^{2-} dans la trypsine). Tous ces ions inorganiques sont considérés non-polarisables. Les anions (Cl^-, PO_4^{3-}, SO_4^{2-}…) ont en général des paramètres multipolaires disponibles dans ELMAM2. Les autres sont modélisés par des monopoles. Les P_v de ces atomes sont ajustées de la même manière que les acides aminés chargés.

3.7.b - L'interaction streptavidine-biotine

La biotine est aussi appelée vitamine B_8 (en France, vitamine B_7 dans les pays anglo-saxons). C'est un important cofacteur enzymatique, notamment pour le cycle de Krebs (cofacteur de la pyruvate carboxylase) et dans diverses carboxylases impliquées par exemple dans le métabolisme de la valine, la leucine et l'isoleucine. Le système streptavidine-biotine a une importance majeure en biotechnologie et son utilisation a fait l'objet de différents brevets. Le 13 septembre 2019, une recherche *Google Patent* pour les termes « *streptavidin biotin* » retourne 117000 résultats (concernant l'utilisation, la production, la quantification, etc., de ces molécules). En effet, la famille de protéine des avidines possède une très forte affinité pour la biotine, la constante de dissociation K_d de ce complexe est parmi les plus favorables mesurées ($K_d \approx 10^{-13}$ M). La bactérie *Streptomyces avidinii* produit la streptavidine utilisée pour la plupart des applications biotechnologiques. L'application majeure – qui rend ce système technologiquement crucial – est la purification de protéines d'intérêt, qu'il est possible d'étiqueter avec un *tag* streptavidine : on ajoute en C-ter ou N-ter du gène à produire la séquence de la streptavidine. On obtient une protéine de fusion avec un domaine streptavidine attaché à la protéine à étudier. Une fois la protéine produite, sa séparation du reste des protéines est possible à l'aide d'une colonne biotinylée. À l'inverse mais plus rarement, il est possible de fonctionnaliser la molécule à purifier avec la biotine pour la purifier sur une colonne où la streptavidine est immobilisée. La streptavidine est un tétramère et est structurée en tonneau β. Elle a fait l'objet de différentes études, structurales et thermodynamiques[105,106] afin d'investiguer les déterminants moléculaires de la très haute affinité de ce complexe. Il s'agit clairement d'un système modèle pour les interactions biologiques de très haute affinité. Cette affinité est majoritairement gouvernée par la constante de vitesse associée à de dissociation du complexe, k_{off}, de l'ordre de 10^{-6} s^{-1} [équation (45)]. Fait notable, des mesures de la force de détachement spécifique entre biotine et streptavidine sont disponibles[105]. Ces mesures ont été réalisées à l'aide d'un microscope à force atomique dont la pointe est biotinylée. En corrigeant la mesure pour les interactions non-spécifiques, la force de détachement moyenne a été déterminée à 253 ± 20 pN ou 393 ± 10 pN pour la streptavidine sauvage (en fonction de la pointe de microscope utilisée).

Table 8. Paramètres cristallographiques des 3 structures du complexe biotine-streptavidine (reproduit de Freitag *et al.*, 1998[107], Freitag *et al.*, 1999[108] & Le Trong *et al.*, 2011[109]). Toutes les données ont été collectées à 100 K.

Code PDB	3RY2 (*WT*)	1SWK (*W79F*)	1NDJ (*Y43F*)	3RY1 (*forme libre*, WT)
Résolution (Å)	50.0 – 0.88	10.0 – 2.0	10.0 – 1.8	47.5 – 1.03
Complétude (%)	95.5	97.0	80	97.1
R_{cryst}	0.116	0.159	0.195	0.118
R_{free}	0.131	0.221	0.312	0.135
Groupe d'espace	*I*222	$P2_1$	$P2_1$	$P2_1$
Paramètres de maille	a = 46.42 Å b = 94.06 Å c = 104.61 Å	a = 59.3 Å b = 64.6 Å c = 76.0 Å β = 93.0°	a = 51.3 Å b = 98.4 Å c = 52.8 Å β = 112.9°	a = 58.0 Å b = 84.4 Å c = 46.0 Å β = 98.8°
Nombre d'atomes (protéine)	2143 *(1814)*	3773 *(3488)*	3789 *(3496)*	4281 *(3625)*
B_{eq} *moyen (protéine, Å²)*	*14.7 (±8.4)*	*30.8 (±15.2)*	*24.2 (±11.9)*	*12.2 (±5.5)*
B_{eq} *moyen (biotine, Å²)*	*10.5 (±0.9)*	*23.7 (±6.2)*	*26.3 (±10.6)*	*non-applicable*

Quatre structures de la PDB seront explorées ici (voir table 8). Deux structures présentent la streptavidine sauvage, soit liée à la biotine (code PDB : 3RY2), soit libre (code PDB : 3RY1)[109]. Les deux autres structures (code PDB : 1SWK et 1NDJ) présentent des mutants de la streptavidine dont l'affinité pour la biotine est altérée. La biotine est présente dans ces deux structures mutantes. La figure 41 montre le détail de l'interaction entre biotine et streptavidine sauvage. Toutes ces protéines (recombinantes) sont issues de *Streptomyces avidinii*. La structure 1SWK[107] présente la mutation W79F. Ce mutant a une enthalpie standard de liaison (ΔH^0) améliorée. Cette mutation se situe au niveau de la queue carboxylate de la biotine. La structure 1NDJ[110], quant à elle, possède une mutation Y43F logée également au sein du site de fixation. Cette mutation se situe au niveau de

l'oxygène du double cycle de la biotine (l'oxygène uréide). L'enthalpie standard de liaison ΔH^0 entre la biotine et le mutant Y43F est diminuée. Le reste du site de fixation de la biotine dans les mutants est identique, d'un point de vue séquence, à celle de la streptavidine sauvage. La littérature rapporte de nombreuses interprétations pour lier les observations cristallographiques et les mesures thermodynamiques. Ces conclusions seront comparées aux résultats des calculs d'énergie basés sur les méthodes décrites dans les chapitres précédents et réalisés avec *MoProViewer*.

La neutralisation des résidus utilise la méthode décrite partie 3.7.a, pendant le transfert de la densité électronique ELMAM2 et des polarisabilités symétriques et moyennées, suivant l'approche décrite partie 3.6. Les énergies d'interactions électrostatiques E_{ELEC} sont calculées à cette étape.

La biotine est ensuite polarisée par tous les acides aminés dont le centre de gravité est situé à moins de 3.5 Å du ligand. Dans le même temps, le champ électrique généré par la biotine polarise ces acides aminés de contact. La procédure de polarisation a été effectuée et à convergé en 7 cycles pour le complexe sauvage, en 8 cycles pour le complexe d'enthalpie augmentée (W79F) et 6 cycles pour le complexe d'enthalpie diminuée (Y43F). Il apparaît que plus l'interaction est intense, plus le nombre de cycles pour obtenir une polarisation convergée augmente.

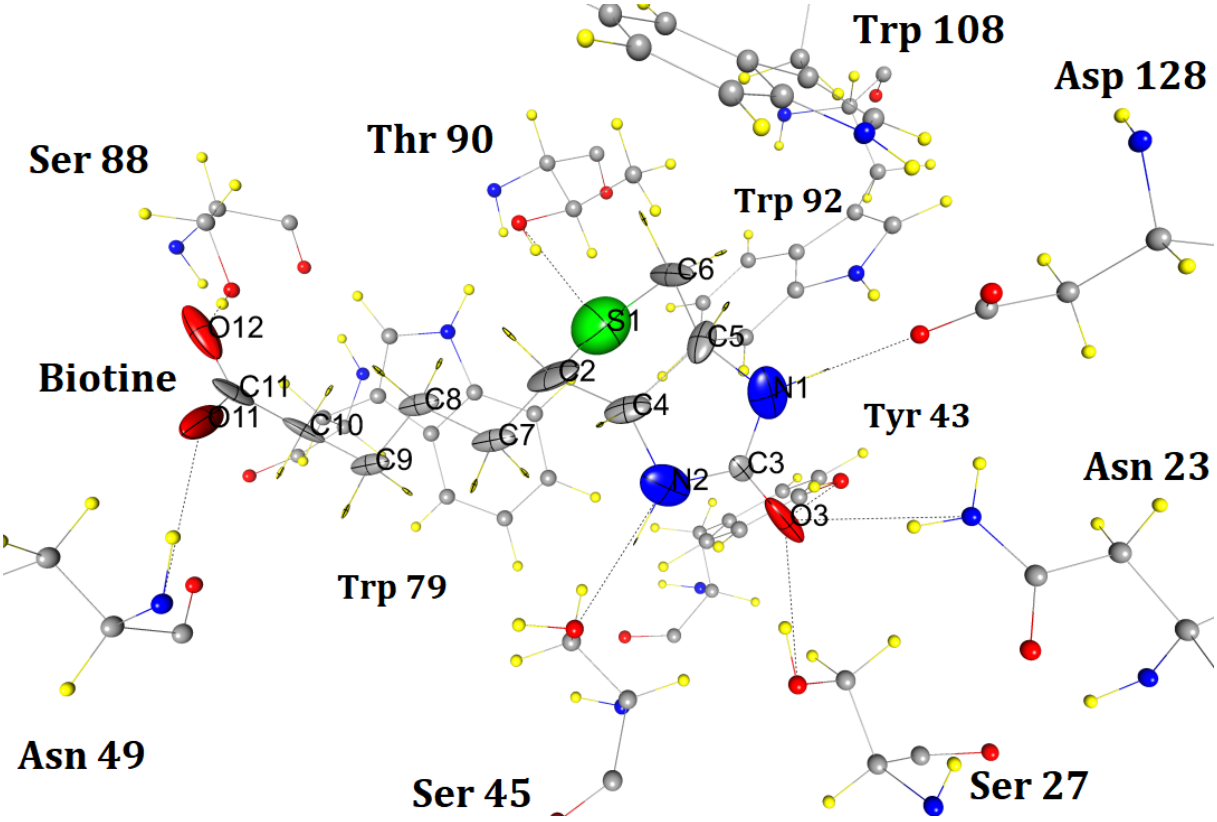


Figure 41. Vue d'ensemble des interactions de la molécule de biotine au sein de la streptavidine sauvage (code PDB : 3RY2). Les traits pointillés représentent les 8 liaisons hydrogène discutées dans le texte. Les 11 acides aminés représentés sont en interaction étroite avec la biotine (leucines 25 et 110, valine 47, glycine 48, alanines 50 et 86 ont été omises pour la clarté). Les acides aminés écrit dans une police plus petite sont en arrière-plan. La partie de droite de la biotine est la tête qui contient le cycle uréide et le soufre thiophène. La partie gauche est la queue carboxylate. La biotine et l'aspartate 128 portent une charge de -1 e, les autres résidus sont neutres. Les ellipsoïdes de la biotine représentent les polarisabilités atomiques transférées (échelle 0.25 Å/Å^3).

Les calculs préliminaires ont montré une énergie de polarisation E_{POL} très forte pour les complexes convergés, même à l'échelle des très fortes interactions électrostatiques E_{ELEC} qui ont été calculées. Dans le modèle présenté, les dipôles d'une même molécule n'interagissent pas entre eux. Par contre ils interagissent avec les dipôles induits de l'autre molécule, et ce dès que le second cycle de polarisation. Au premier cycle de polarisation, les seuls champs électriques calculés sont issus des densités transférées. On peut supposer que l'induction réelle de dipôles est plus faible que l'induction calculée à convergence dans notre modèle [voir équation (19), page 38], ce qui exagère les énergies de polarisation calculées à convergence. L'induction calculée au premier cycle,

elle, correspond à un effet nécessairement présent au sein de l'interaction streptavidine-biotine, celui de la densité transférée (voir Figure 20, page 119). Pour cette raison, on analysera dans la suite de cette application biochimique uniquement l'induction, et les énergies d'interaction, calculées au premier cycle de polarisation. Les résultats retenus sont présentés en annexe (Table A5). Sur la Figure 42, les acides aminés dont les ellipsoïdes sont omises sont exclus de l'algorithme de polarisation, mais pas des calculs de E_{ELEC} et E_{POL}. Leur contribution aux énergies d'interaction est notée dans la ligne « autres A.A. » de la Table A5.

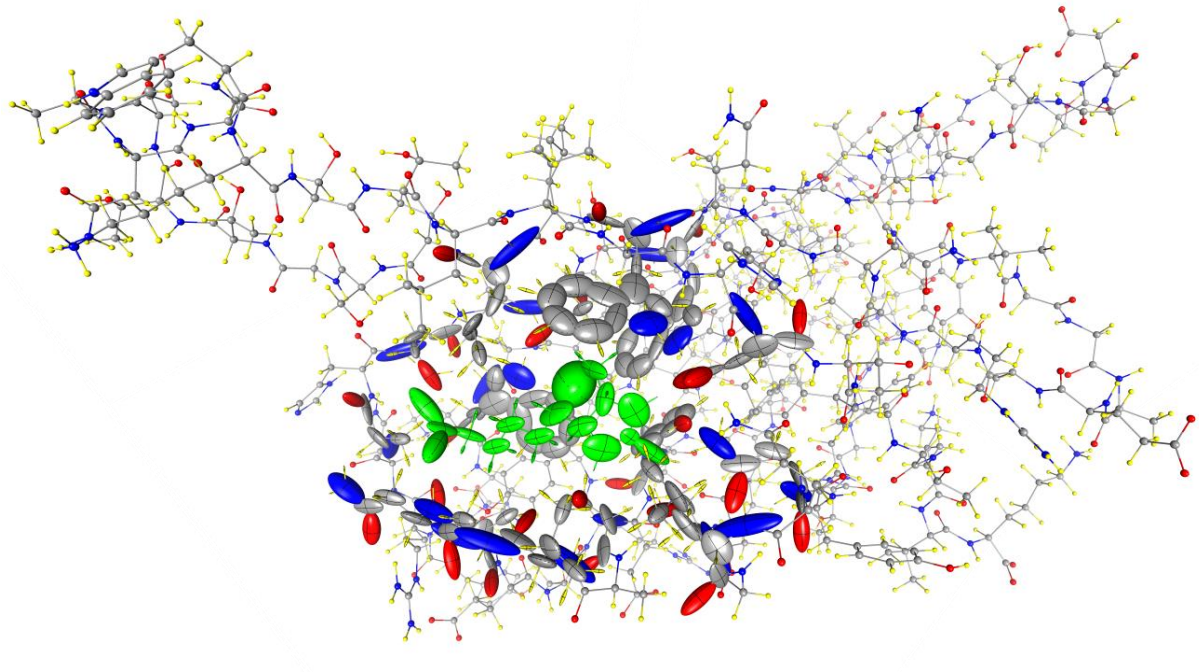


Figure 42. Vue d'ensemble du site de fixation de la biotine (en vert) sur son récepteur streptavidine. Les ellipsoïdes de polarisabilité sont montrées uniquement sur la zone du complexe ayant subi une polarisation de la densité électronique (échelle 0.5 Å/Å^3). Pour chaque variant, un seul site de fixation a été étudié (celui de la chaîne A du fichier obtenu de la PDB).

La conformation de la streptavidine libre (code PDB : 3RY1) et de la streptavidine liée à la biotine (code PDB : 3RY2) ont été comparées. Ces deux cristaux de streptavidine sauvage ont des paramètres cristallographiques proches et ont été publiées ensemble. Le RMSD brut entre les positions des atomes non-hydrogène de la protéine est 1.841 Å. Ce chiffre a été obtenu à l'aide de *Pymol*, en prenant en compte la totalité des 1557 atomes

non-hydrogène des deux protéines (seules les chaines A et B de la structure 3RY1 ont été alignées sur les deux chaines de la structure 3RY2 de symétrie supérieure). En tout, deux chaînes de 121 acides aminés, résolus dans les deux structures cristallographiques, ont été alignés. *Pymol* propose une fonction d'alignement structurale basée sur le rejet itératif de paires d'atomes dont distance est supérieure à 2.0 Å (par défaut). Cela permet de filtrer les chaines latérales flexibles et de focaliser l'alignement sur les parties dont la structure est conservée.

En utilisant cet outil, le RMSD entre les 1184 atomes dont la position est la mieux conservée est 0.232 Å, entre la forme libre et liée. La figure 43 montre la qualité de cet alignement au niveau du site de fixation de la tête de la biotine. La base du tétraèdre complexant l'oxygène 3 de la biotine (composé de Asn 23, Ser 27 et Tyr 43) est présent dans la forme libre (cette fixation sera traitée en détails ci-dessous). Dans la structure libre, le site de fixation est bien entendu hydraté, mais accueille également des molécules de précipitant MPD (4S-2-méthyl-2,4-pentanediol et 4R-2-méthyl-2,4-pentanediol). Les molécules d'eau et les molécules de MPD forment avec les résidus du site de fixation des interactions tout à fait distinctes de celles réalisées par la biotine. Malgré ces interactions différentes, la position des résidus du site de fixation est très conservée. Cette conservation s'étend également au cœur hydrophobe du tonneau β de la streptavidine car les paires d'atomes rejetées de l'alignement concernent des chaines latérales exposées au solvant (et une boucle nommée L3,4 dans la chaîne B de la streptavidine libre[110]). Ces observations sont largement détaillées par les auteurs des structures cristallographiques. La flexibilité de la boucle L3,4 ne sera pas étudiée ici (voir Le Trong *et al.*, 2011[109]).

Les mutants ont été observés dans des structures de qualité cristallographique inférieure (voir Table 8). La conservation globale de la structure de la streptavidine a été vérifiée, les atomes du mutant W79F génèrent un *RMSD* brut de 0.677 Å avec ceux de la protéine sauvage (1588 atomes). Cette déviation baisse à 0.220 Å lorsqu'on considère les 1302 paires d'atomes les plus comparables. Dans ce mutant, le *RMSD* pour les atomes de la biotine seulement est de 0.090 Å, ce qui indique une pose très conservée.

Le mutant Y43F a également une conformation très comparable celle de la protéine sauvage : le *RMSD* pour tous les atomes est 0.633 Å (1596 atomes comparés). Ce *RMSD* tombe à 0.193 Å lorsqu'on considère les 1322 meilleures paires d'atomes. Pour les atomes de la biotine seulement, l'alignement est relativement moins bon (*RMSD* de 0.222 Å) mais

il convient de lier ceci à une différence d'orientation du carboxylate terminant la queue de la biotine. Cette orientation différente est liée aux différences de fixation détaillées plus tard.

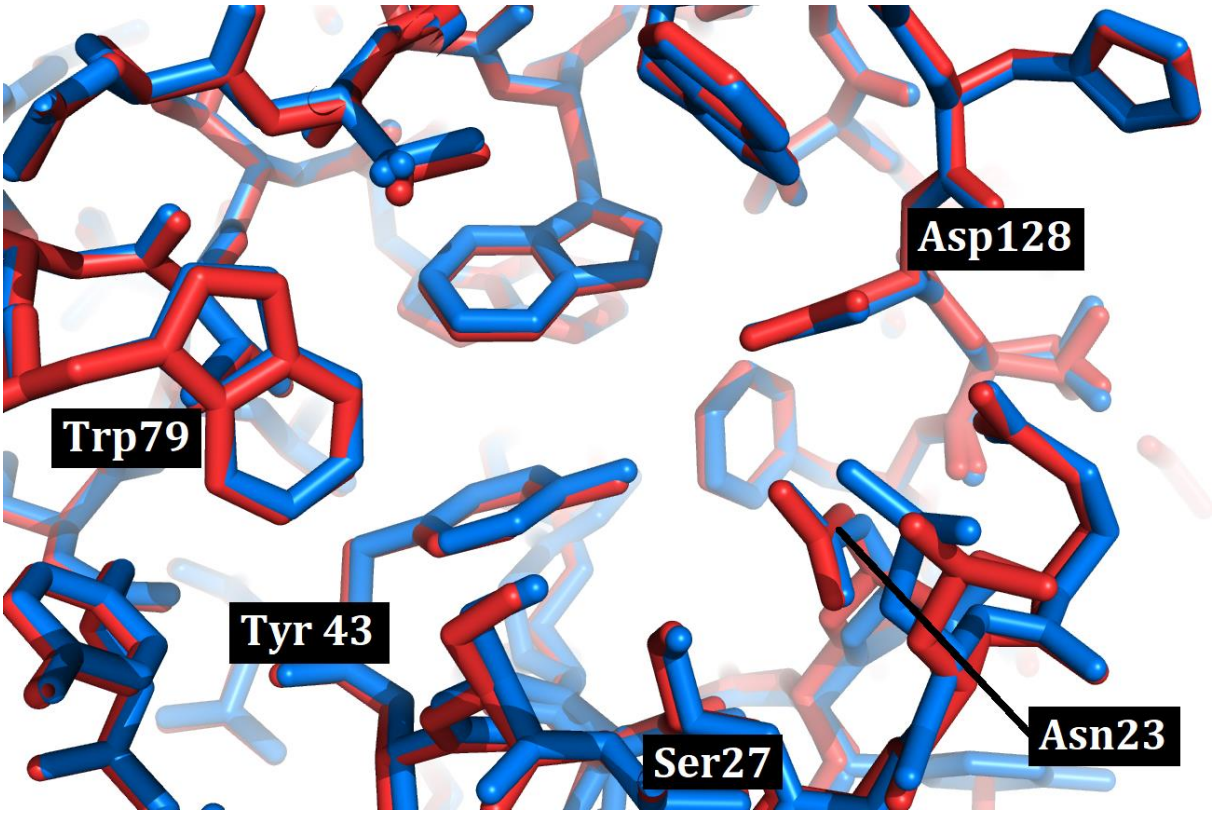


Figure 43. La conservation de la conformation du site de fixation de la streptavidine, entre la forme libre et liée à la biotine, est exceptionnelle. Forme libre en rouge, forme liée en bleu (la biotine est omise). Image *Pymol*.

3.7.c - Comparaison des énergies d'interactions des streptavidines mutantes

Dans le but d'évaluer notre méthode, on se basera sur les mesures de titration calorimétrique isotherme publiées par Chilkoti et Stayton, 1995[105], et par Klumb *et al.*, 1998[106]. Ces mesures sont reportées Table 9.

La première constatation est que les énergies calculées sont numériquement extrêmement élevées (voir Table 9). La somme $E_{ELEC}+E_{POL}$ est 7.8 fois plus forte que l'enthalpie expérimentale, pour l'interaction entre la biotine et la streptavidine *wild-type*, 8.22 fois supérieure pour le mutant W79F et 5.5 fois supérieure pour le mutant Y43F. Ces

facteurs ont le même ordre de grandeur. Aucune régression linéaire ne sera faite ici pour corréler les valeurs numériques expérimentales et modélisées, du fait du faible nombre de points, de l'absence de potentiel répulsif et de la prise en compte d'une seule conformation (pose), pour expliquer des mesures ITC à différentes températures (25°C/298 K et 37°C/310 K, voir équation (54), les données cristallographiques ont, elles, toutes été collectées à une température de -173°C/100 K). Ces différentes limites de notre approche étant posées, une analyse bien plus fine reste possible.

Pour comparer les variations relatives (ΔE) des énergies d'interaction biotine-streptavidine entre les mutants considérés, on utilise la formule modifiée suivante :

$$\Delta E_{(\%)} = \frac{E_{WT} - E_{mutant}}{|E_{WT}|} \cdot 100 \qquad (77)$$

En effet, si le total des enthalpies standard expérimentales et des énergies d'interaction calculées sont négatives, elles peuvent être exprimées comme une somme de termes positifs (répulsifs) et négatifs (attractifs). On souhaite avoir des pourcentages négatifs lorsque l'interaction est moins favorable, quel que soit le signe de E_{mutant} et de E_{WT}. La formule ci-dessus a en effet la vertu de donner des pourcentages négatifs quand E_{mutant} est positif et supérieur à E_{WT}, c'est-à-dire quand une des contributions du mutant entraîne une affinité diminuée pour le ligand. Et des pourcentages positifs quand E_{mutant} est négatif et inférieur à E_{WT}.

Tout d'abord il est nécessaire de comparer les différences entre mutants pour l'interaction électrostatique E_{ELEC}, donc uniquement entre densités électroniques transférées, sans polarisation. Le mutant W79F a une enthalpie ΔH^0 expérimentale augmentée par rapport à la forme sauvage passant de -104.2 ± 1.7 à -108.8 ± 2.1 kJ·mol^{-1} (soit une interaction favorisée de +4.4 % à 25°C). Sur E_{ELEC}, l'énergie d'interaction électrostatique, on constate en effet une différence de +6.5 % entre W79F et la streptavidine *wild-type*.

L'autre mutant, Y43F, a quant à lui une enthalpie diminuée à 37 °C (ici, -4.1 %). Si cette interaction moins favorable n'est pas retrouvée à 25 °C, il faut remarquer que la différence d'enthalpie entre 25 °C et 37 °C est beaucoup plus faible pour Y43F que pour la streptavidine sauvage. Klumb *et al.*, 1998[106], ont exploré cette différence d'enthalpie, et l'ont reliée à une augmentation de k_{off}. Le comportement à 25°C est expliqué par les auteurs comme étant lié à un k_{off} trop lent pour obtenir l'équilibre thermodynamique au

cours de l'expérience (la mesure de ΔG^0 à 25°C est reportée impossible), et ils recommandent donc d'interpréter la mesure à 37 °C, température à laquelle ΔG^0 a été mesurée. La mesure à 25 °C ne sera pas interprétée. On trouve une différence relative de -4.1 % entre les ΔH^0 expérimentales mesurées à 37 °C pour la streptavidine *wild-type* et Y43F. Le calcul de l'énergie d'interaction électrostatique E_{ELEC} entre le mutant Y43F et la biotine montre une différence relative de -27.8 % avec $E_{ELEC,WT}$. Cette différence relative est exagérée (mais cohérente) par rapport au -4.1% retrouvés par ITC à 37 °C.

D'un côté, la variation relative de E_{ELEC} reproduit mieux la variation relative d'enthalpie pour le mutant avec un ΔH^0 amélioré (W79F, +4.4 % expérimentaux contre +6.5 % pour E_{ELEC}). De l'autre côté, le mutant d'enthalpie diminuée (Y43F) montre une E_{ELEC} diminuée mais, cette fois, dans une proportion beaucoup plus forte que la diminution de ΔH^0 observée (-4.1% expérimentaux contre -27.8 % pour E_{ELEC}).

La polarisation n'a pas encore été prise en compte, mais on s'attend évidemment à une interaction encore plus forte entre la biotine et la streptavidine. Conformément au modèle de polarisation appliqué, toutes les E_{POL} sont effectivement négatives. Ces énergies ont toutes une magnitude plus faible que E_{ELEC}.

Pour le mutant W79F, E_{POL} est +19.0 % plus forte que $E_{POL,WT}$. En comparant la somme des deux contributions, la différence avec le sauvage (ΔE) est de +10.7 %, montrant comme attendu une interaction plus stable entre ligand et streptavidine. La variation de ΔH^0 est cependant toujours surestimée.

Le mutant Y43F montre une énergie de polarisation E_{POL} à -11.5 % de $E_{POL,WT}$. Considérant, pour ce mutant, la forte différence entre E_{ELEC} et $E_{ELEC,WT}$, cette variation est surprenante. En considérant la somme E_{ELEC}+E_{POL}, la prise en compte de la polarisation de la densité électronique permet de faire passer ΔE à -22.2 % (à partir de -27.8 % pour E_{ELEC} seule).

Il apparaît donc que prendre en compte de la contribution de polarisation améliore globalement la capacité de notre méthode à appréhender la variabilité des résultats expérimentaux mais sans les reproduire entièrement. Cette amélioration existe malgré une forme très simple de la modélisation de la polarisation utilisées ici (polarisabilités indépendantes et un seul cycle de polarisation).

Les deux forces principales de notre modèle sont l'utilisation de potentiels multipolaires de haute qualité (à la fois pour les calculs de champs électriques et d'énergies d'interaction), et l'utilisation de polarisabilités atomiques anisotropes dérivées de calculs *ab initio*. Ces deux particularités devraient permettre une bonne modélisation d'interactions directionnelles. Il est donc logique d'étudier comparativement l'interaction de la biotine avec les résidus des différentes structures de la streptavidine retenues ici. Le détail des énergies d'interaction (électrostatique et de polarisation) entre la biotine et les différents acides aminés de contact est disponible en annexe (Table A5).

Table 9. Énergies d'interaction entre streptavidine et biotine. Les valeurs d'enthalpie totale ΔH^0 sont expérimentales, les valeurs E_{ELEC} et E_{POL} ont été calculées avec *MoProSuite* en utilisant la méthode présentée ici. La variation d'énergie relative ΔE, calculée par rapport au type sauvage est évaluée sur la somme E_{ELEC} + E_{POL}. Les valeurs marquées par des astérisques seront comparées entre elles.

Variant *Code PDB*	ΔH^0 (25 °C, kJ·mol^{-1})	ΔH^0 (37 °C, kJ·mol^{-1})	E_{ELEC}, EP/MM (kJ·mol^{-1})	E_{POL}, cycle 1, EP/MM, (kJ·mol^{-1})	ΔE (kJ·mol^{-1})
WT *3RY2*	-104.2 ± 1.7	-123.0	-532.2	-276.2	0
			Somme : -808.4		
W79F *1SWK*	-108.8 ± 2.1 (***+4.4 %***)*	*non-déterminé*	-566.7 (*+6.5 %*)	-328.5 (*+19.0 %*)	-86.8 (***+10.7 %***)*
			Somme : -895.2		
Y43F *1NDJ*	-115.1 ± 3.8 (*+10.5 %*)	-118.0 (***-4.1 %***)**	-384.2 (*-27.2 %*)	-244.5 (*-11.5 %*)	+179.8 (***-22.2 %***)**
			Somme : -628.6		

Afin de détailler les effets des mutations, on s'intéressera d'abord aux liaisons hydrogène. Ces liaisons impliquent traditionnellement de fortes interactions électrostatiques. Ensuite, les deux mutants sont des substitutions entre acides aminés

aromatiques, dont les chaînes latérales sont très polarisables. Ce type d'interaction demande donc une attention supplémentaire. Ces éléments seront enfin comparés aux interprétations de la littérature. Pour chaque site d'interaction, le complexe sauvage sera d'abord examiné, puis les effets des mutations sur ces sites d'interaction seront explicités.

Table 10. La fixation de la biotine implique 8 liaisons hydrogène, pour la streptavidine sauvage et W79F. Le variant Y43F implique 7 liaisons hydrogène. Les énergies d'interaction E_{ELEC}+E_{POL} sont calculées entre tous les atomes des résidus considérés et tous les atomes de la biotine.

		3RY2 (*WT*)	1SWK (*W79F*)	1NDJ (*Y43F*)
Asn 23	*d*(Nδ-O3) (Å)	*3.01*	*3.05*	*2.95*
	E_{ELEC}+E_{POL} (kJ·mol^{-1})	-69.9	-72.0	-87.8
Ser 27	*d*(Oγ-O3) (Å)	*2.70*	*2.38*	*2.72*
	E_{ELEC}+E_{POL} (kJ·mol^{-1})	-141.4	-235.0	-138.0
Tyr 43	*d*(Oη-O3) (Å)	*2.70*	*2.66*	***non applicable***
	E_{ELEC}+E_{POL} (kJ·mol^{-1})	-119.0	-126.4	-29.5
Ser 45	*d*(Oγ-N2) (Å)	*2.98*	*2.84*	*3.07*
	E_{ELEC}+E_{POL} (kJ·mol^{-1})	-48.6	-51.2	-24.7
Asn 49	*d*(N-O11) (Å)	*2.84*	*2.87*	*2.96*
	E_{ELEC}+E_{POL} (kJ·mol^{-1})	-130.9	-101.1	-135.8
Ser 88	*d*(Oγ-O12) (Å)	*2.81*	*2.56*	*2.9*
	E_{ELEC}+E_{POL} (kJ·mol^{-1})	-120.8	-181.3	-91.4
Thr 90	*d*(Oγ-S) (Å)	*3.40*	*3.46*	*3.08*
	E_{ELEC}+E_{POL} (kJ·mol^{-1})	-44.0	-41.9	-63.3
Asp 128	*d*(Oδ2-N1) (Å)	*2.81*	*2.87*	*3.13*
	E_{ELEC}+E_{POL} (kJ·mol^{-1})	+28.3	+27.5	+82.4

Il existe 8 liaisons hydrogène entre la biotine et la streptavidine sauvage (*idem* pour le mutant W79F). La mutation Y43F altère une tyrosine réalisant une de ces liaisons, ce mutant n'en présente donc que 7. Les énergies d'interaction calculées entre ces huit acides aminés sont reportées Table 10.

Trois parties différentes de la biotine réalisent ces liaisons. Le groupement carboxylate qui termine la queue aliphatique fait deux liaisons hydrogène (avec l'asparagine 49 et la sérine 88). La tête de la biotine est constituée du cycle uréide et du cycle soufré. Ce dernier lie la partie aliphatique de la molécule et réalise une liaison hydrogène (avec la thréonine 90). Enfin, le cycle uréide est impliqué dans 5 liaisons hydrogène.

Parmi les 5 liaisons hydrogène réalisées par le cycle uréide avec la streptavidine sauvage, trois ont le même atome de la biotine pour accepteur. Cet atome – O3 – accepte les hydrogènes de l'asparagine 23, de la sérine 27 et de la tyrosine 43. Les liaisons ont une organisation tétraédrique, montrée Figure 44 [panneau (a) pour le sauvage]. O3 est au sommet et la base comprend Nδ (Asn23), Oγ (Ser27) et Oη (Tyr43).

Oxygène uréide, wild-type

Dans le complexe sauvage, la sérine 27 est le résidu avec la plus forte interaction avec la biotine (indifféremment pour E_{ELEC}, E_{POL} et E_{ELEC}+E_{POL}). La distance d'interaction entre O3 et Oγ de la sérine 27 est quasi-identique à celle de la liaison entre O3 et Oη de la tyrosine 43. Pourtant l'interaction E_{ELEC}+E_{POL} de la liaison O3···Ser27 est 21.4 kJ·mol^{-1} plus intense que la liaison O3···Tyr43. Oγ et Oη ont le même type atomique (même polarisabilité et mêmes paramètres multipolaires). La partie électrostatique de cette différence d'énergie peut s'expliquer en rappelant que les doublets non-liants de O3 sont situés dans le plan du cycle uréide. Si Oγ (Ser27) et Oη (Tyr43) sont à la même distance de O3, Oγ est bien mieux aligné avec le plan uréide que Oη. On peut supposer que la différence d'énergie serait moins marquée en utilisant une densité électronique sphérique pour l'atome accepteur O3. Ceci peut expliquer la différence de 18.6 kJ·mol^{-1} entre les E_{ELEC} associées à ces deux liaisons. Ceci illustre l'effet de la prise en compte de densités électroniques asphériques pour modéliser les interactions électrostatiques permanentes des liaisons hydrogène (cette conclusion n'est pas généralisable à E_{POL}, puisque dans le complexe sauvage, la polarisation de O3 dépend en même temps des deux résidus comparés ici : tyrosine 43 et sérine 27).

La dernière liaison hydrogène impliquant O3 est O3···Asn23 (qui implique le Nδ de l'asparagine 23). Elle est associée à une énergie d'interaction environ 2 fois inférieure à

l'interaction O3···Ser27 (ce facteur 2 est observé pour E_{ELEC}, E_{POL} et E_{ELEC}+E_{POL}). Cette liaison est plus longue (+0.3 Å) que O3···Ser27 et O3···Tyr43. Elle implique un azote qui plus polarisable mais moins électriquement chargé que les groupements hydroxyles. Les fortes différences entre les énergies calculées incitent à toujours mettre en parallèle ces résultats avec les différences de géométrie entre les liaisons hydrogène comparées.

La Figure 44 montre une forte conservation de la géométrie des liaisons hydrogène de O3 dans les trois complexes présentés. On retrouve une organisation tétraédrique et des distances interatomiques comparables dans la streptavidine sauvage, W79F et Y43F. Il existe cependant de légères différences structurales et il s'avère que celles-ci ont un impact énergétique.

Oxygène uréide, W79F

Le mutant W79F [panneau (b), Figure 44] présente qualitativement les mêmes liaisons hydrogène que le sauvage au niveau de O3. La biotine y est dans la même conformation que dans le complexe sauvage. Les caractéristiques des liaisons hydrogène sont différentes : les liaisons sont plus courtes que dans le variant sauvage, particulièrement pour la sérine 27 [-0.32 Å entre O3 et Oγ(Ser27)]. La mutation W79F est au niveau du site de fixation de la biotine, mais se situe loin de l'atome O3. Il est possible que cette mutation réduise la gêne stérique et permette à la biotine d'accéder à une conformation plus favorable au niveau des liaisons hydrogène formées par O3. L'interaction E_{ELEC}+E_{POL} pour O3···Ser27 passe de -141.1 kJ·mol^{-1} (sauvage) à 235.0 kJ·mol^{-1} (W79F), soit +66.2%. Dans ce mutant, la sérine 27 est encore le résidu avec les plus fortes E_{ELEC}, E_{POL} et E_{ELEC}+E_{POL}. Pourtant ici, c'est l'énergie de polarisation qui explique la plus grande partie de cette différence d'énergie (E_{POL} est 92.9 % plus fort que $E_{POL,WT}$, alors que E_{ELEC} n'augmente 'que' de 55.6 %).

Ce rapprochement de O3 et de la sérine 27 entraine également un raccourcissement – certes moindre – de la liaison hydrogène avec la tyrosine 43 [-0.04 Å entre O3 et Oη(Tyr43) comparé au sauvage]. Considérant ces variations, la variation d'énergie d'interaction E_{ELEC}+E_{POL} est bien moins dramatique que pour la sérine 27. E_{ELEC}+E_{POL} passe de -119.0 à -126.4 kJ·mol^{-1}, soit une augmentation de +6.2 %. De manière intéressante, pour ce mutant W79F et ce résidu tyrosine 43, E_{ELEC} est plus faible que dans le complexe sauvage (-8.3%), mais E_{POL} est bien plus forte (+39.3 %). Autrement dit, cette

mutation semble favoriser la stabilisation de l'interaction via la polarisation mutuelle de la biotine et de la tyrosine 43.

La dernière liaison de O3, dans le complexe W79F, implique l'asparagine 23. Cette liaison hydrogène est extrêmement similaire à celle du complexe sauvage en termes de longueur (+0.04 Å), de géométrie et d'énergie (+3.0 % comparé au sauvage, toutes contributions prises en compte).

Oxygène uréide, Y43F

Pour terminer avec ce locus d'interactions centré sur O3, il reste à s'intéresser au dernier mutant. Le mutant Y43F touche directement le site de fixation de O3. La mutation est présentée panneau (c) de la Figure 44. Cependant la longueur des liaisons hydrogène entre l'oxygène uréide et l'asparagine 23 d'un côté et la sérine 27 de l'autre côté sont très comparables à celles de la protéine sauvage. La liaison avec l'asparagine 23 est 0.06 Å plus courte dans le mutant Y43F. Malgré cette géométrie similaire, l'énergie d'interaction de ce résidu avec la biotine est 25.6 % plus intense que dans le sauvage. Pour l'asparagine 23, l'augmentation de E_{POL} (+39.8 % par rapport à $E_{POL,WT}$) est plus forte que celle de E_{ELEC} (+20.2 % par rapport à $E_{ELEC,WT}$). La liaison hydrogène avec la sérine 27 est plus longue de 0.02 Å. L'interaction de la sérine 27 avec la biotine est sensiblement de même intensité comparée au sauvage, avec une variation de -2.7%. E_{ELEC} est légèrement diminuée, E_{POL} légèrement augmentée. Cette interaction conserve la magnitude la plus forte au sein du site de fixation du mutant Y43F, parmi tous les résidus de contact.

La phénylalanine 43 ne permet pas de liaison hydrogène avec sa chaîne latérale, pourtant, l'hydrogène ζ (à 2.92 Å de O3) est remarquablement proche de la position de l'oxygène η de la tyrosine 43 dans la streptavidine sauvage (l'oxygène η est à 2.70 Å de O3). En l'absence de cette liaison hydrogène (Hζ étant trop loin), l'intensité de l'énergie d'interaction électrostatique E_{ELEC} varie de -81.7 % par rapport au sauvage (+67.6 kJ·mol^{-1}). L'énergie de polarisation pour ce résidu montre une variation de -60.2 % par rapport au sauvage (+21.8 kJ·mol^{-1}). Considérant la conservation de structure entre les modèles, ces résultats appuient encore l'importance d'examiner les interactions les plus proches du ligand.

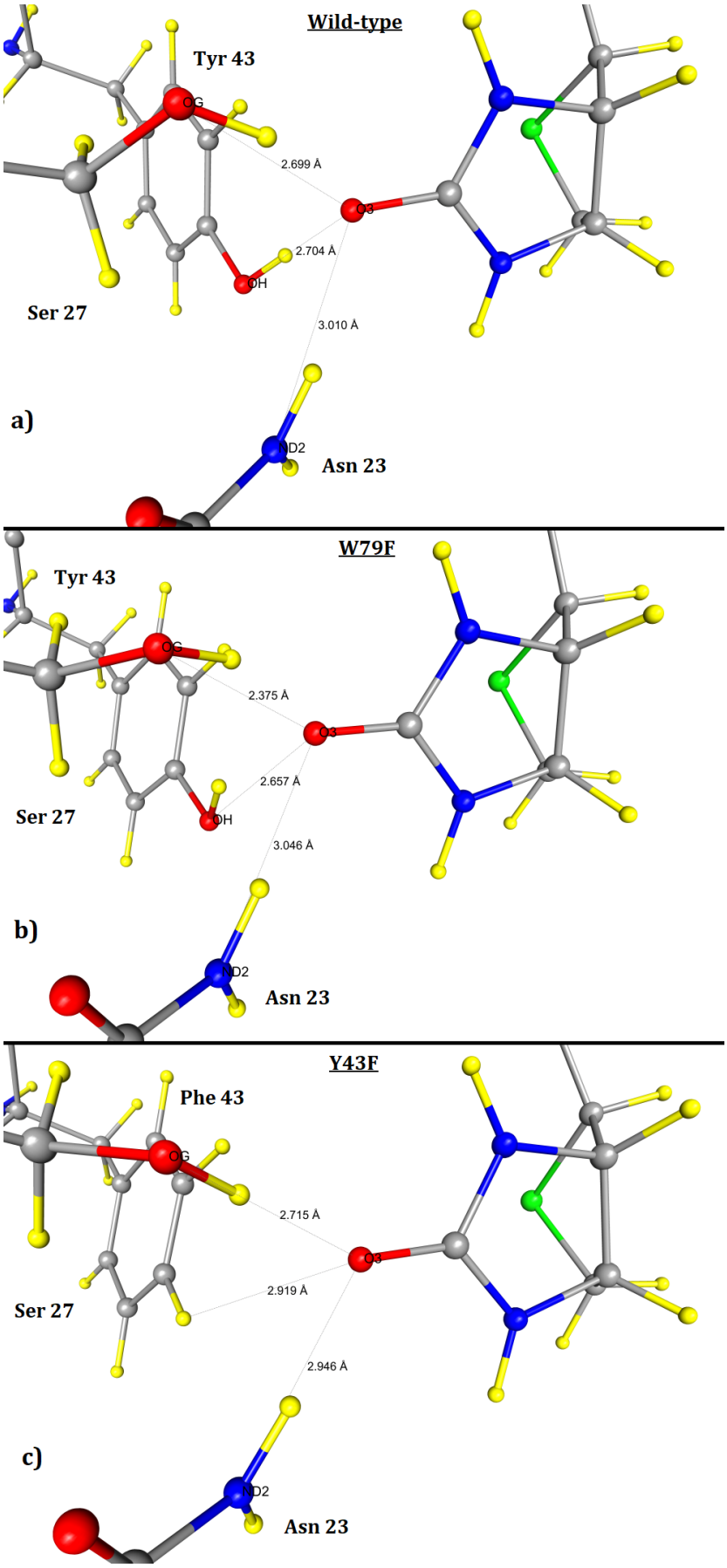

Wild-type
Tyr 43
OG
2.699 Å
O3
2.704 Å
OH
Ser 27
3.010 Å
a)
ND2
Asn 23
W79F
Tyr 43
OG
2.375 Å
O3
2.657 Å
OH
Ser 27
3.046 Å
b)
ND2
Asn 23
Y43F
Phe 43
OG
2.715 Å
O3
2.919 Å
Ser 27
2.946 Å
c)
ND2
Asn 23

Figure 44.

Figure 44 (ci-contre). L'atome O3 (oxygène uréide) de la biotine est stabilisé par un tétraèdre de liaisons hydrogène, dont les distances sont affichées. Panneau (a) : streptavidine sauvage (code PDB : 3RY2). Panneau (b) : mutant d'enthalpie améliorée (W79F, code PDB : 1SWK) montrant au niveau de l'oxygène le même réseau d'interaction que le variant sauvage. Panneau (c) : mutant d'enthalpie diminuée (Y43F, code PDB : 1NDJ). La distance entre le Hζ de la phénylalanine 43 mutante et l'oxygène uréide est montrée pour appuyer la conservation de la conformation du site de liaison à la biotine malgré l'absence de liaison hydrogène entre O3 et Hζ.

Azotes uréides, wild-type

Les trois liaisons hydrogène précédentes impliquaient des donneurs de liaison à l'oxygène uréide O3. Deux autres liaisons, avec des résidus accepteurs cette fois, existent avec le cycle uréide. La première implique l'oxygène Oγ de la sérine 45 et l'azote N2 de la biotine. La liaison hydrogène N2···Ser45 est plus longue que les autres liaisons impliquant des sérines (sérines 27 et sérine 88, traitée plus bas). Cependant ces deux autres liaisons sont de type O–H···O. La liaison N2···Ser45 est une liaison de type N–H···O, comparable à la liaison entre O3···Asn23 qui implique, elle aussi, un azote sp^2. La liaison N2···Ser45 est plus courte que O3···Asn23. Dans le complexe sauvage, son énergie $E_{ELEC,WT}$ est légèrement inférieure à celle de cette dernière liaison. L'énergie de polarisation de la liaison formée par la sérine 45 est largement inférieure à celle de la liaison avec l'asparagine 23. E_{ELEC} représente 90.3 % de l'interaction totale E_{ELEC}+E_{POL} calculée.

La seconde liaison hydrogène acceptant un proton de la biotine implique l'azote N1 de la biotine avec l'oxygène Oδ de l'aspartate 128 (N1···Asp128). Cette contribution est énergétiquement défavorable, impliquant deux groupements chargés négativement. La contribution $E_{ELEC,WT}$ est dominante et largement positive (62.8 kJ·mol^{-1}). La composante de polarisation $E_{POL,WT}$ est la quatrième plus intense de tout le site de fixation de la streptavidine sauvage. À -34.5 kJ·mol^{-1}, $E_{POL,WT}$ compense 54.9 % de l'énergie d'interaction électrostatique, sans pour autant la rendre favorable.

Azotes uréides, W79F

Ces deux liaisons hydrogène avec la sérine 45 et l'aspartate 128 sont en vis-à-vis, donc un déplacement relatif du cycle uréide dans les complexes mutants (par rapport à la pose observée dans la forme sauvage) pourra être apprécié en les examinant ensemble. Pour les deux liaisons précédentes, l'effet de la mutation W79F est un rapprochement de

la biotine par rapport à la sérine 45 et un éloignement par rapport l'aspartate 128. La liaison N2···Ser45 est 0.14 Å plus courte que dans le sauvage. On attend un fort effet sur l'énergie d'interaction, pourtant la somme E_{ELEC}+E_{POL} n'est que 2.6 kJ·mol^{-1} plus favorable (+5.3 %). La liaison avec l'aspartate 128 est légèrement plus longue (+0.06 Å) dans le mutant W79F (interaction plus forte). Ce très faible allongement de la liaison est lié à une baisse de 2.8 % de E_{ELEC}+E_{POL} (on rappelle que cette contribution est répulsive). Ce raccourcissement induit une augmentation de +4.3 kJ·mol^{-1} de E_{ELEC} par rapport au complexe sauvage. Cette augmentation est plus que compensée par une variation favorable de E_{POL} de -5.1 kJ·mol^{-1}.

Azotes uréides, Y43F

Dans le complexe Y43F, la liaison N2···Ser45 s'allonge de +0.09 Å. La somme E_{ELEC}+E_{POL} est environ divisée par 2 par rapport au sauvage (E_{POL} est quant à elle divisée par 4 et devient inférieure à 2 kJ·mol^{-1}). Le mutant Y43F (interaction moins forte) présente la plus longue distance d'interaction entre Oδ(Asp128) et N2, +0.32 Å plus loin que le sauvage. L'énergie la plus haute (défavorable) est associée à cette liaison allongée. Dans le mutant Y43F, la magnitude de E_{ELEC} augmente de 66.1% et celle de E_{POL} est 36.6 % plus faible (moins favorable) que dans le sauvage.

Figure 45 (ci-contre). Les interactions de la queue carboxylate de la biotine avec la streptavidine impliquent des interactions hydrophobes et électrostatiques. Le tryptophane 79 (WT et mutant Y43F) et la phénylalanine 79 (mutant W79F) sont impliqués dans l'interaction avec la partie aliphatique de la biotine. Pour chaque variant, la distance mesurée avec l'acide aminé 79 est la plus courte distance interatomique entre ce résidu et la biotine. L'interactions de O12 avec la sérine 88 est une liaison hydrogène importante d'un point de vue électrostatique. La position de l'hydrogène γ de la sérine 88 a été manuellement corrigée (voir l'interprétation page 201). Panneau (a) : streptavidine sauvage (code PDB : 3RY2). Panneau (b) : mutant d'enthalpie diminuée (Y43F, code PDB : 1NDJ). Panneau (c) : mutant d'enthalpie améliorée (W79F, code PDB : 1SWK).

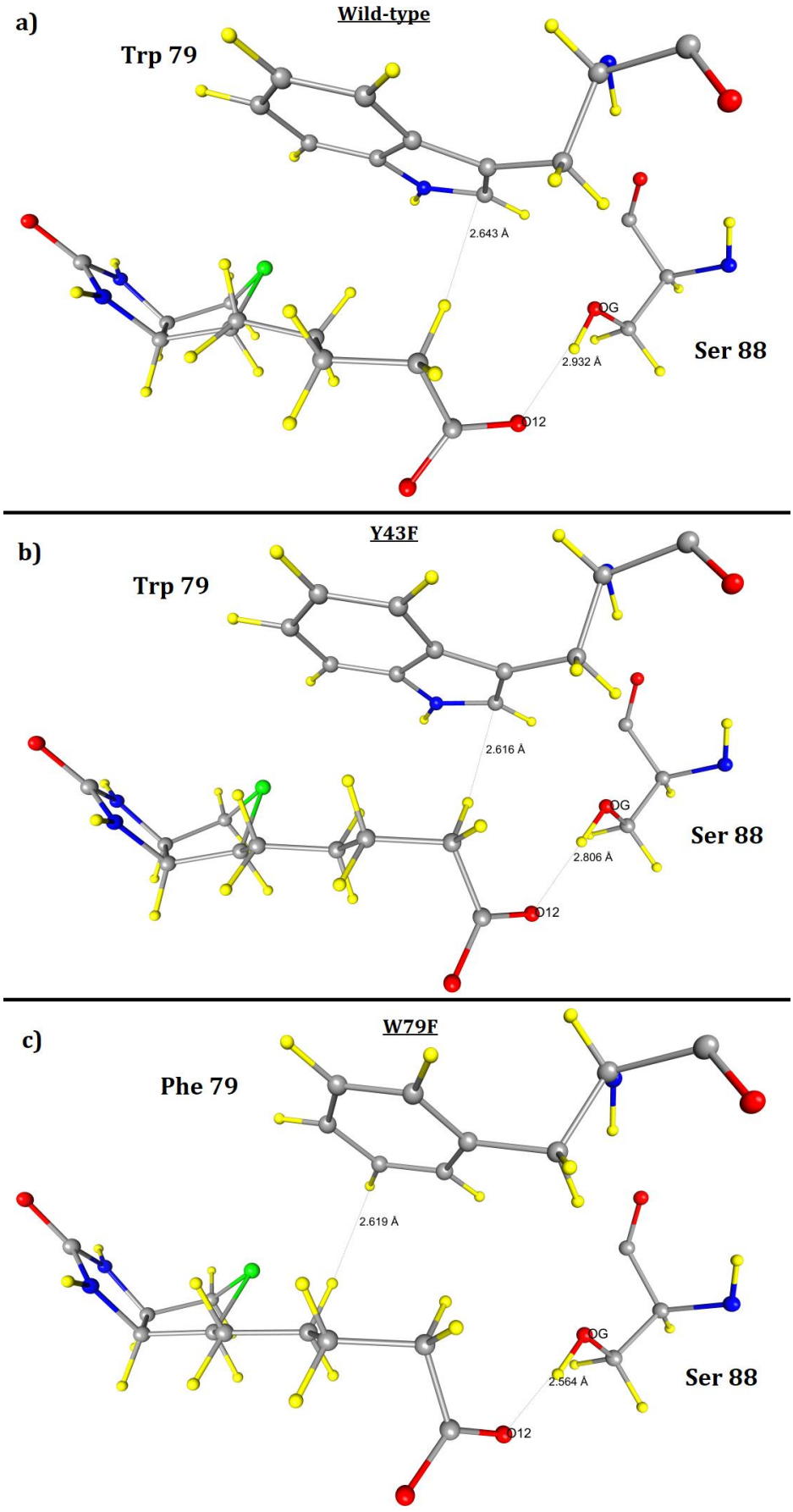

a)
Wild-type
Trp 79
2.643 Å
OG
Ser 88
2.932 Å
O12
b)
Y43F
Trp 79
2.616 Å
OG
Ser 88
2.806 Å
O12
c)
W79F
Phe 79
2.619 Å
OG
Ser 88
2.564 Å
O12

Figure 45.

Ces observations sont contrintuitives : malgré les charges négatives des deux groupes d'atomes en interaction (biotine et aspartate 128), plus la distance d'interaction est courte, plus la somme E_{ELEC}+E_{POL} apparaît favorable. Cependant, l'énergie de polarisation n'est pas la seule responsable de cette compensation de la répulsion charge-charge. L'interaction entre multipôles permanents influe également, car le minimum de E_{ELEC} – pour la liaison N1···Asp128 – est retrouvé dans le complexe sauvage, où la distance est minimale (2.81 Å). Si l'induction dipolaire calculée ne permet pas de vaincre la répulsion charge-charge des carboxylates de la biotine et de l'aspartate 128, l'intensité des différentes E_{ELEC} et E_{POL} calculées entre ces deux résidus montre que la liaison hydrogène entre N1···Asp128 est définitivement importante pour mitiger la répulsion.

Il est tentant de considérer que l'interaction avec l'aspartate 128 soit partiellement responsable de la spécificité d'orientation de l'interaction biotine-streptavidine. Pourtant, Baugh *et al.*, 2016[111], ont publié une étude cristallographique d'un mutant D128N. Non seulement la conformation du mutant en complexe avec la biotine est identique au sauvage, conservant la liaison hydrogène avec N1, mais le k_{off} de ce complexe est similaire à celui du mutant Y43F. La répulsion charge-charge semble donc avoir un effet mineur sur la fixation de la biotine, alors que la présence de la liaison hydrogène semble cruciale.

Soufre thiophène

Une liaison hydrogène isolée existe avec l'autre cycle de la tête de la biotine (le cycle contenant un soufre thiophène). Elle implique les doublets non-liants du soufre et l'hydrogène γ de la thréonine 90. La liaison hydrogène S···Thr90 est la plus longue du complexe sauvage, avec 3.40 Å, et la seconde énergie d'interaction la plus faible parmi les liaisons hydrogène de ce complexe. L'interaction électrostatique permanente E_{ELEC} représente environ 73 % de la somme E_{ELEC}+E_{POL}.

Dans le mutant W79F, la distance entre le soufre et Oγ augmente de 0.06 Å. L'énergie E_{ELEC}+E_{POL} diminue de 4.8 % (+2.09 kJ·mol^{-1}). Cette variation est imputable à une variation de E_{POL} seule de +2.6 kJ·mol^{-1} (défavorable).

Dans le mutant Y43F, la liaison se raccourcit significativement, -0.32 Å, et la magnitude de E_{ELEC}+E_{POL} augmente en conséquence avec une variation de -19.4 kJ·mol^{-1} comparée au sauvage (+44.0 % en valeur relative). E_{ELEC} et E_{POL} contribuent à cette affinité améliorée, pour le complexe Y43F, E_{ELEC} représente une proportion de 77% de E_{ELEC}+E_{POL}.

Chaîne aliphatique et carboxylate, wild-type

Le dernier locus d'interaction entre la biotine et la streptavidine concerne la queue carboxylate du ligand. On y retrouve deux liaisons hydrogène, avec l'asparagine 49 (O11···Asn49) et avec la sérine 88 (O12···Ser88), ainsi que le site de la mutation W79F [panneau (c) de la Figure 45]. Cette liaison avec l'acide aminé 79 semble plutôt hydrophobe ou dispersive, on s'attend donc à des énergies E_{ELEC} ou E_{POL} plus faibles.

Chaque oxygène carboxylate (O11 et O12) de la biotine est impliqué dans une liaison hydrogène. L'une d'entre elles, entre O11 et l'azote de l'asparagine 49, ne peut être mutée car elle implique le groupement amide de squelette peptidique (*backbone*). Dans le complexe sauvage (et également dans le mutant Y43F), l'interaction E_{ELEC}+E_{POL} entre biotine et asparagine 49 est la deuxième interaction la plus intense. Cette liaison O···H–N est comparable aux liaisons O3···Asn23 et N2···Ser45. Elle est sensiblement plus courte [d(O12-N) = 2.84 Å] et son énergie d'interaction est sensiblement plus intense que celle des deux autres liaisons O···H–N : +87.3 % par rapport à O3···Asn23 et +169.2 % par rapport à N2···Ser45. Elle est également comparable à la liaison N1···Asp128, au vu de l'implication d'un oxygène chargé et d'un azote sp^2. Si la répulsion électrostatique entre l'aspartate 128 et la biotine rend la comparaison de E_{ELEC} difficile (tous deux chargés négativement), force est de constater que pour des distances d'interaction très similaires (2.84 Å et 2.81 Å), les énergies de polarisation de ces deux liaisons sont très proches (E_{POL}(O11···Asn49) = -37.0 kJ·mol^{-1} ; E_{POL}(N1···Asp128) = -34.5 kJ·mol^{-1}).

La liaison hydrogène de la sérine 88 est très importante d'un point de vue énergétique (3e plus importante valeur de E_{ELEC}+E_{POL} dans le complexe sauvage). Son traitement dans le modèle présenté peut être débattu. En effet, Les atomes d'hydrogène des structures 3RY2, 1SWK et 1NDJ ont été ajoutés à l'aide de *Reduce*[103]). Ce logiciel est très satisfaisant pour compléter les hydrogènes manquant sur des structures de peptides ou de protéines, y compris pour déterminer quel azote de l'histidine est protoné. Le traitement d'autres molécules organiques (par exemple l'ornithine, biologique mais non-protéinogène) peut laisser à désirer. Ce problème peut apparaître pour les acides aminés en interaction directe avec le ligand. En effet, pour les structures 3RY2 (*wild-type*) et 1SWK (mutant W79F), cet hydrogène γ de la sérine 88 était orienté loin du groupement carboxylate, alors que l'espace entre les oxygènes est suffisant pour réaliser une liaison hydrogène (voir Table 10 pour les distances interatomiques). La structure 1NDJ (mutant

Y43F) montrait une configuration plausible qui a été imitée. Une rotation de 101.18° de cet hydrogène autour de l'axe Cβ-Oγ de la sérine 88 a permis d'aligner Hγ avec la liaison hydrogène. La matrice de rotation a été calculée en utilisant la formule de Rodrigues [équation (A9), page 292 de l'annexe].

Les calculs d'énergies d'interaction électrostatique et de polarisation entre la biotine et la sérine 88 montraient en effet une anomalie. Pour le complexe sauvage, la somme E_{ELEC}+E_{POL} est -16.3 kJ·mol^{-1} avec l'hydrogène mal positionné. Après correction, E_{ELEC}+E_{POL} passe à -120.8 kJ·mol^{-1}. Dans le complexe W79F, la somme E_{ELEC}+E_{POL} est -24.0 kJ·mol^{-1} avec l'hydrogène mal positionné. Après correction, E_{ELEC}+E_{POL} passe à -181.3 kJ·mol^{-1}. La différence est énorme et ne peut être ignorée. Comme l'espace entre O12 et Oγ est suffisant, il sera supposé par la suite que les positions corrigées sont meilleures.

La liaison hydrogène entre sérine 88 et l'oxygène carboxylate O12 de la biotine est de longueur intermédiaire, par rapport aux autres liaisons hydrogène impliquant des sérines. L'énergie E_{ELEC}+E_{POL} de O12···Ser88 est moins favorable que celle la liaison O3···Ser27 avec une différence de 20.6 kJ·mol^{-1}, cependant O3···Ser27 est plus courte de 0.11 Å. La composante électrostatique E_{ELEC} représente 74.9 % de E_{ELEC}+E_{POL} pour O12···Ser88 et 71.7 % de E_{ELEC}+E_{POL} pour O3···Ser27, soit des proportions similaires. L'autre liaison hydrogène impliquant une sérine concerne la sérine 45. Si on compare maintenant O12···Ser88 et N2···Ser45, Oγ de la sérine 45 joue le rôle d'accepteur, la distance d'interaction est plus longue (+0.17 Å) et l'énergie beaucoup moins favorable pour N2···Ser45 (+72.2 kJ·mol^{-1}). La composante de polarisation est bien plus faible pour cette liaison N2···Ser45, où E_{ELEC} correspond à 90.3 % de E_{ELEC}+E_{POL}.

La dernière interaction qui sera détaillée est celle qui concerne le site de mutation W79F. Sur la Figure 45, l'interaction avec le tryptophane 79 apparaît de nature hydrophobe. La distance affichée sur la figure entre cet acide aminé et la biotine correspond à la plus petite distance interatomique entre ces résidus. Comme cette interaction n'est pas une liaison hydrogène, elle est absente de la Table 10 et est présentée Table 11 (et en annexe, Table A5). La biotine possède quatre carbones aliphatiques. Les trois carbones aliphatiques les plus proches du carboxylate sont en vis-à-vis de la chaîne latérale indole du tryptophane 79. Le tryptophane 79 a une énergie d'interaction avec la biotine plus intense que tous les autres tryptophanes de la protéine. Pour ce résidu 79, la somme E_{ELEC}+E_{POL} est comparable aux énergies les plus faibles trouvées dans des liaisons

hydrogène (excluant l'aspartate 128 dont la contribution globale est déstabilisante). Parmi les autres interactions, non-hydrogène et plus hydrophobes, seule la leucine 110 montre une somme E_{ELEC}+E_{POL} comparable. Toutes les autres interactions non-hydrogène ont une magnitude inférieure. Entre la biotine et Trp79, E_{ELEC} représente 67.1 % de E_{ELEC}+E_{POL}. Cette proportion est réduite comparée aux autres interactions tryptophane-biotine (> 75%), le tryptophane 79 est donc – par sa position – plus exposé au champ électrique du ligand et plus polarisé. Les deux autres tryptophanes (92 et 108) n'interagissent pas de manière comparable avec la biotine, ils sont situés à proximité du double cycle et interagissent avec la moitié thiophène. Leur surface d'interaction avec le ligand semble plus faible comparée au tryptophane 79.

Table 11. Interactions entre l'acide aminé 79 (Trp ou Phe) et la biotine.

Interaction : *Biotine – A.A.*	3RY2 (*WT*) *Trp 79*	1NDJ (*Y43F*) *Trp 79*	1SWK (*W79F*) *Phe 79*
E_{ELEC} (kJ·mol^{-1})	-33.19	-33.62 (*+1.3%*)	-20.80 (*-37.3%*)
E_{POL} (kJ·mol^{-1})	-16.28	-16.70 (*+2.6%*)	-10.44 (*-35.9%*)
E_{ELEC}+E_{POL} (kJ·mol^{-1})	-49.47	-50.32 (*+1.7%*)	-31.23 (*-36.9%*)

Chaîne aliphatique et carboxylate, W79F

La fixation de la queue carboxylate évolue quand les mutants sont examinés. Dans le mutant W79F (panneau (c), Figure 45), l'interaction avec le squelette peptidique (asparagine 49) s'allonge très légèrement (+0.03 Å) mais la somme E_{ELEC}+E_{POL} varie (fortement) de +29.8 kJ·mol^{-1} (soit -22.8 %). Cette interaction est liée à une augmentation de E_{ELEC} de +23.6 kJ·mol^{-1}. Par conséquent, l'interaction Asn49-biotine devient la troisième plus intense, dans le mutant W79F. L'interaction entre la sérine 88 et le carboxylate de la biotine devient la seconde plus intense calculée. La magnitude de E_{ELEC}+E_{POL} est 50.1 % plus forte (-60.5 kJ·mol^{-1}), mais la proportion de E_{ELEC} dans cette somme est constante entre W79F et le sauvage (environ 75%). Cette variation d'énergie est associée à une forte variation de géométrie : la distance entre O12 et le Oγ de la sérine 88 diminue de 0.25 Å. En considérant seulement ces deux liaisons hydrogène, la position du carboxylate de la biotine apparaît comme plus favorable dans le complexe mutant W79F par rapport au

sauvage. Dans ce complexe, la phénylalanine 79 est à une distance comparable de la biotine, lorsque comparée au tryptophane 79 du sauvage. La chaîne latérale de la phénylalanine est moins encombrante et peut expliquer le rapprochement entre la biotine et la sérine 88. L'interaction de la biotine avec la phénylalanine est 36.9 % moins intense, mais la proportion de E_{ELEC} dans cette somme est très proche du sauvage (66.6 % comparé à 67.1 %), malgré des chaines latérales de nature différente.

Chaîne aliphatique et carboxylate, Y43F

La structure générale de l'interaction à ce locus est relativement conservée par le mutant Y43F, panneau (b) de la Figure 45. Pour ce mutant d'affinité diminuée, la distance entre O12 et l'oxygène γ de la sérine 88 est raccourcie de 0.13 Å. La distance entre O11 et l'azote peptidique de l'asparagine 49 augmente de 0.12 Å. Considérant l'importance énergétique de ces deux interactions dans notre modèle, on s'attend à un effet énergétique non-négligeable. En effet, l'interaction E_{ELEC}+E_{POL} avec la sérine 88 perd 24.4 % d'intensité (+29.4 kJ·mol^{-1}). Cette variation est importante pour expliquer la variation d'enthalpie de ce mutant. Cet effet est imputable à une baisse drastique de la contribution électrostatique, plus défavorable de 25.5 kJ·mol^{-1}. Par contre, l'interaction O11···Asn49 du mutant Y43F est à -4.9 kJ·mol^{-1} de l'interaction au sein de la streptavidine sauvage, elle varie peu pour un éloignement comparable à celui observé dans l'autre mutant W79F. Dans le complexe Y43F, E_{ELEC} varie de -12.6 kJ·mol^{-1} (soit +13.4 % par rapport au sauvage) quand E_{POL} augmente de +7.6 kJ·mol^{-1} (soit -20.7 % de $E_{POL,WT}$). Concernant le tryptophane 79, l'atome de la biotine le plus proche est le même dans le mutant Y43F et la protéine sauvage. La distance est identique à 0.03 Å près. La Table 11 montre que l'interaction avec le tryptophane 79 de ce mutant est sensiblement identique à l'interaction sauvage.

Autres résidus

L'interaction de la biotine avec neuf acides aminés a été examinée en détails. Ces acides aminés ont été choisis car impliqués dans des liaisons hydrogène et/ou dans une mutation à étudier. Il y a en tout 17 acides aminés de contact dans ce complexe. Les huit derniers sont : leucine 25, valine 47, glycine 48, alanines 50 et 86, tryptophanes 92 et 108, et enfin la leucine 110. Ces résidus ne forment pas d'interactions « canoniques » avec la biotine (ponts salins, liaisons hydrogène, *stacking*). Cependant l'interaction (pour E_{ELEC}, E_{POL} et E_{ELEC}+E_{POL}) entre la leucine 110 et la biotine est comparable à l'interaction entre

biotine et tryptophane 79 (dans les complexes *wild-type* et Y43F). La valine 47 et la glycine 48 font partie de la boucle L3,4, dont la mobilité peut être constatée sur la structure de la streptavidine sauvage isolée (code PDB : 3RY1). Dans le complexe sauvage, ces 8 acides aminés expliquent 15.8 % de l'énergie E_{ELEC}+E_{POL} totale. Ces 8 acides aminés présentent une contribution collective à l'interaction avec la biotine de -131.4 kJ·mol^{-1}, dont 63.2 % de ce total est attribuable à E_{ELEC}. Cette proportion est proche de la moyenne pour tous les résidus, 73.0 % (en excluant la valine 47 et l'aspartate 128 dont les E_{ELEC} sont positives).

Dans le complexe d'enthalpie améliorée (W79F), la variation de conformation est faible, de même que la variation d'énergie associée à ces 8 acides aminés restants. Leur E_{ELEC}+E_{POL} est de -133.5 kJ·mol^{-1}, soit -2.1 kJ·mol^{-1} plus favorable comparé à la forme sauvage. Cette variation est causée par une augmentation de E_{POL} de +3.9 kJ·mol^{-1} (défavorable), compensée par une variation de +6.0 kJ·mol^{-1} de E_{ELEC}.

Dans le complexe d'enthalpie diminuée (Y43F), la contribution de ces 8 acides aminés est modifiée dans une plus grande proportion (Le *RMSD* des positions atomiques de la biotine est plus grand pour Y43F, comparé à l'autre mutant). La somme E_{ELEC}+E_{POL} est de -148.6 kJ·mol^{-1}, c'est-à-dire -17.2 kJ·mol^{-1} plus favorable que pour la forme sauvage. Ici, à la fois E_{ELEC} et E_{POL} contribuent à une augmentation de l'affinité calculée entre ces 8 résidus et la biotine, par rapport au complexe sauvage. E_{ELEC} est plus favorable de -11.3 kJ·mol^{-1} et E_{POL} est abaissée de -5.9 kJ·mol^{-1}. Ces variations compensent donc légèrement la baisse globale d'énergie d'interaction calculée entre toute la streptavidine et la biotine. Pour ces 8 acides aminés dans les 3 complexes, la proportion $E_{ELEC}/(E_{ELEC}+E_{POL})$ est 65 %, soit très proche de la moyenne de tous les résidus (73 %).

3.7.d - Conclusions partielles

Les énergies d'interaction calculées présentent d'importantes variations entre les mutants. Une partie de l'effet observable par l'expérience d'ITC semble reproduit par notre modèle, principalement en considérant l'interaction des acides aminés mutés avec la biotine (tyrosine/phénylalanine 43 et tryptophane 79), mais également en considérant l'effet de ces mutations sur les autres liaisons hydrogène du site actif, en particulier O3···Ser27, N2···Ser45, O11···Asn49, O12···Ser88 et N1···Asp128. Le calcul de l'énergie de polarisation permet d'enrichir l'analyse des interactions non-covalentes au sein du site de

fixation. La suite du texte récapitule et met en perspective l'effet énergétique de chaque mutation sur la fixation de la biotine. Une vue d'ensemble des contributions est disponible Figure 46 (une vue plus détaillée est disponible en annexe, Figure A1).

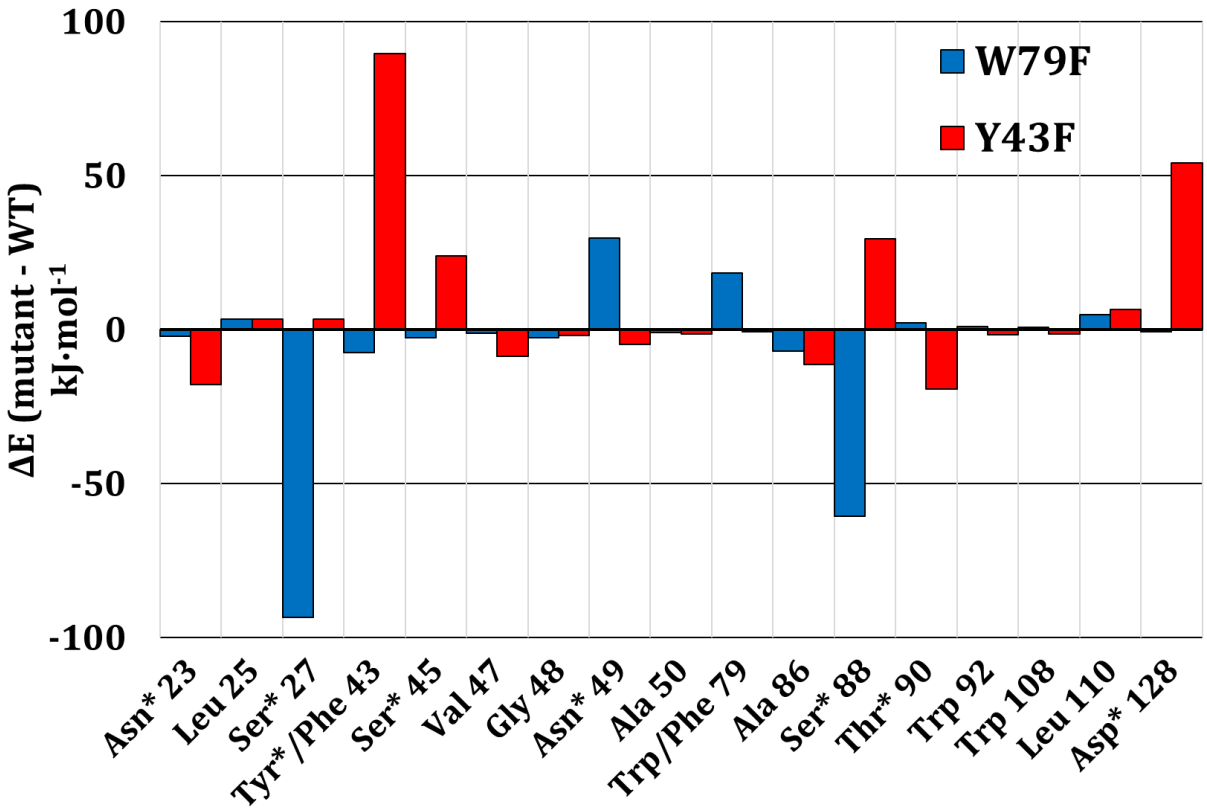


Figure 46. Variation de l'énergie E_{ELEC}+E_{POL} (kJ·mol^{-1}) entre streptavidine mutante et sauvage, par acide aminé de contact. Les acides aminés marqués d'un astérisque sont impliqués dans une liaison hydrogène.

Comme vu partie 3.7.b, la détermination des causes structurales de l'affinité exceptionnelle du complexe streptavidine-biotine a déjà été largement explorée expérimentalement. Freitag *et al.*, 1998[107], rapportent que des études théoriques ont été menées pour éclairer le rôle des différents tryptophanes dans cette interaction.

Tout d'abord, il convient de rappeler qu'une affinité aussi forte sera limitée par la diffusion du ligand dans l'eau (Zhou et Zhong, 1982[72], estiment que le plafond théorique de la constante de vitesse de diffusion est environ 10^{-10} M^{-1}·s^{-1} à 25°C). k_{on}, la constante de vitesse liée à l'association du complexe ne peut dépasser la limite de diffusion. Par conséquent, le principal facteur qui influence la constante de dissociation K_d (= k_{off}/k_{on} ≈

10^{-13} M) – entre les différents mutants – est nécessairement k_{off}, la constante de vitesse associée à la dissociation du complexe (voir page 52). Cette constante dépend de la barrière énergétique entre les streptavidine libre et liée à la biotine, et donc de la stabilité de l'état de transition qui sépare ces deux états (notée $\Delta G^{\neq}$ dans la littérature[105]). Cette énergie n'est pas accessible par notre méthode, seule une partie des variations enthalpiques (ΔH^0) peut être explorée‡‡‡.

Dans la littérature, la forte enthalpie du complexe streptavidine-biotine est décrite comme découlant d'abord d'interactions hydrophobes et ensuite de liaisons hydrogène[105,107]. L'investigation de ces interactions hydrophobes a mené à l'étude du complexe sauvage et (entre autres) du mutant W79F par microscopie de force atomique[105] et par ITC. Cette étude de microscopie pointe déjà l'importance de facteurs entropiques dans l'affinité de la streptavidine pour la biotine et dans l'énergie $\Delta G^{\neq}$. L'étude des mutants touchant le réseau de liaisons hydrogène, dont Y43F, est légèrement postérieure[106,110]. Les études cristallographiques de ces complexes ont ensuite relié les variations d'entropie conformationnelle (pour ΔG^0 et $\Delta G^{\neq}$) à la mobilité de la boucle L3,4[107,109].

Du point de vue des interactions électrostatique et de polarisation calculées ici, l'importance des résidus tryptophanes n'apparaît pas comme plus importante par rapport aux liaisons hydrogène. L'estimation des composantes de désolvatation et de van der Waals apparait nécessaire. Le tryptophane 79 est une exception, il affiche la plus forte interaction E_{ELEC}+E_{POL} avec la biotine parmi tous les résidus non-engagés dans une liaison hydrogène et son importance électrostatique ne doit pas être négligée.

Dans chaque complexe, les énergies d'interaction les plus intenses sont retrouvées au niveau de l'oxygène uréide O3 et de la queue carboxylate. Pour le mutant W79F, on cherche à expliquer l'augmentation d'enthalpie standard pour la biotine. Sur la Figure 46, on observe clairement que la fixation de l'oxygène uréide (par la sérine 27) et de la queue carboxylate (par la sérine 88) impliquent des interactions plus énergétiques. La déstabilisation de la liaison de la queue carboxylate avec l'asparagine 49 est compensée

‡‡‡ *: Il serait techniquement possible d'appliquer notre méthode à la structure hypothétique d'un état de transition et approximer la contribution électrostatique à $\Delta H^{\neq}$, mais il n'existe aucune garantie d'applicabilité des densités électroniques ELMAM2 dans ce contexte hors-équilibre.*

largement par les deux interactions précédentes. Cette amélioration de la fixation de la biotine est également mitigée par l'interaction atténuée avec la phénylalanine 79. La mutation W79F échange un acide aminé aromatique contre un autre moins volumineux. La conformation de la biotine en interaction avec le mutant est très proche mais il apparaît que cette levée de contrainte stérique permet au cycle uréide d'atteindre une position plus stable, plus proche de la sérine 27. Cette nouvelle position induit un effet intéressant sur la tyrosine 43 : la position de la liaison hydrogène est moins optimale, comme le montre l'augmentation de E_{ELEC} pour ce résidu, mais cet effet est plus que compensé par la polarisation. On observe une augmentation plus ou moins légère de l'énergie d'interaction à mesure que les liaisons hydrogène de l'autre côté de la biotine s'allongent (asparagine 49 et thréonine 90).

Les calculs présentés montrent des variations d'énergie parfois importantes pour des distances d'interaction similaires. L'exemple le plus flagrant est l'interaction avec l'asparagine 49. Dans ce complexe W79F, entre cet acide aminé et la biotine, une variation de 0.03 Å de la distance d'interaction, par rapport au complexe sauvage, entraine une variation de +29.8 kJ·mol^{-1}, dont 79 % sont liées à une augmentation de E_{ELEC} (+23.6 kJ·mol^{-1}). L'énergie de polarisation ne diminue en magnitude que marginalement (+6.2 kJ·mol^{-1}). Un exemple opposé est celui de la tyrosine 43, où cette fois l'énergie de polarisation compense une interaction électrostatique moins favorable. On voit ici les conséquences flagrantes de la capacité de notre modèle à modéliser la directionnalité des interactions, autant au niveau de la densité électronique (HCM), que de l'anisotropie des polarisabilités. En effet, l'anisotropie des polarisabilités génère des moments dipolaires induits qui ne sont pas forcément colinéaires au champ électrique, comme il sera vu dans la partie suivante. Ces dipôles induits de la protéine et de la biotine peuvent interagir favorablement ou non en fonction de leur orientation et donc de la géométrie de l'interaction. Il faut aussi prendre en compte que ces dipôles sont représentés par une densité électronique distribuée, qui fait que E_{POL} ne dépend pas uniquement du potentiel et du champ électrique au niveau du noyau. Cette capacité à modéliser des interactions directionnelles pourrait être objectivée systématiquement par des comparaisons

théoriques supplémentaires (par exemple en utilisant le set de *benchmarking* S66a8[68§§§]), ce qui est hors de portée de cette étude de cas.

Le mutant W79F a clairement une enthalpie standard améliorée pour la biotine. Chilkoti *et al.*, 1995[105], observent une barrière de dissociation augmentée ($\Delta H^{\neq}$ augmente de 12.1 kJ·mol^{-1} par rapport au complexe sauvage), une enthalpie de liaison abaissée (ΔH^{0} varie de -6.3 kJ·mol^{-1} par rapport au complexe sauvage) et une force de rappel augmentée (mesurée entre une streptavidine immobilisée et une pointe de microscope à force atomique biotinylée). k_{off} pour ce mutant est diminué, la libération de la biotine est plus lente que dans le sauvage, bien que ΔG^{0} soit moins favorable au complexe. En effet l'acide aminé 79 est situé proche de la boucle L3,4 et la mutation W79F affecte sa mobilité[105], possiblement en rigidifiant L3,4. L'article rapportant la structure cristallographique de ce mutant (1SWK)[107] confirme cette interprétation sur la base de mesures ITC supplémentaires : la variation de capacité calorifique ΔC_p est diminuée pour le mutant [voir équation (54)], suggérant une streptavidine W79F libre moins flexible que la protéine sauvage (l'état « ouvert » de la boucle L3,4 semble plus énergétique pour W79F que pour WT).

Le comportement expérimental de la streptavidine W79F dévie d'une fixation rigide idéale (modèle clé-serrure classique), du fait des éléments cités au-dessus. Pourtant, sur la base des structures cristallographiques et de nos calculs d'énergie, il est possible de relier l'augmentation d'enthalpie aux liaisons hydrogène plus favorables avec les sérines 27 et 88. La levée de contraintes stériques joue un rôle important dans ces interactions serrées. Si la mobilité de la boucle est un facteur incontestable à la baisse de k_{off}, il faut garder en mémoire que la différence d'énergie $\Delta H^{\neq}$ est calculée entre l'état initial (lié) et l'état de transition. Une enthalpie standard ΔH^{0} plus favorable – entre le complexe sauvage et W79F – sera une contribution à une barrière énergétique $\Delta H^{\neq}$ plus importante. En effet la barrière à la dissociation du complexe est définie depuis l'enthalpie de l'état lié ($\Delta H^{\neq} = H^{\neq} - H^{0}$, voir figure 8 page 58).

La mutation Y43F entraine la perte d'une liaison hydrogène au niveau d'un des deux sites clés pour l'interaction avec la biotine. Cette perte d'une seule liaison hydrogène

§§§ *: Le set S66a8 contient les dimères du S66 à distance d'équilibre, affectées de rotations de ±30° autour des axes* **x**, **y** *et* **z**, *là où, dans le S66x8, les orientations sont conservées et les distances d'interaction sont étirées ou contractées.*

représente 49.8 % de l'effet total calculé entre le mutant Y43F et la forme sauvage. La perte d'énergie d'interaction au niveau de O3 est légèrement compensée par l'asparagine 23. Le soufre de la biotine se rapproche de la thréonine 90, et l'énergie de cette liaison baisse. La position du double cycle de la biotine au sein du site de fixation change : les interactions avec la sérine 45 et l'aspartate 128 sont plus faibles car rallongées. La présence de la phénylalanine 43 à la place d'une tyrosine implique une déstabilisation des interactions avec la sérine 88. Ce site de fixation est en effet très rigide, et la nouvelle position du cycle implique une position moins stable de la queue carboxylate par rapport à la protéine.

Les paramètres thermodynamiques de l'interaction entre le mutant Y43F et la biotine sont plus complexes à interpréter, du fait des différences d'enthalpie avec le sauvage : ΔH^0 est plus intense à 25°C et moins intense à 37°C. La mesure à 25°C ne permet pas de déterminer ΔG^0 avec fiabilité (et donc accéder à ΔS^0). Ceci pousse les auteurs[106] à supposer que l'état lié du complexe Y43F est déstabilisé à 25°C par des contributions entropiques. La boucle L3,4 recouvre les résidus 45 à 52 et est donc très proche du site de mutation. À 37°C, les variations d'énergie libre ΔG^0 et d'enthalpie libre ΔH^0 impliquent une affinité diminuée. La barrière énergétique liée à l'énergie de l'état de transition est abaissée par une contribution $T\Delta S^{\neq}$ augmentée par rapport au sauvage (l'entropie de l'état de transition est supérieure dans le mutant Y43F comparé au *wild-type* donc son énergie diminue). Le Trong et *al.*, 2003[110], notent – sur la base de structures cristallographiques libres et en interaction avec la biotine – que cette augmentation de la contribution entropique peut être liée à la libération d'une molécule d'eau supplémentaire du site de fixation du mutant Y43F. Ces différentes influences aboutissent à un k_{off} multiplié par 4. Par contre les auteurs de cet article[110] avancent dans leur conclusion que ΔH^0 est plus favorable dans le mutant que le complexe sauvage. Cette conclusion contredit les valeurs numériques publiées et reproduites de Klumb *et al.*, 1998[106].

Les calculs d'énergie d'interaction et de polarisation montrent que la perte d'une liaison hydrogène est un facteur clé de la diminution d'énergie d'interaction entre streptavidine Y43F et biotine. Cependant la différence relative entre les E_{ELEC}+E_{POL} du *wild-type* et de Y43F est exagérée. La performance de notre modélisation est moins bonne pour ce mutant que pour le mutant Y79F. S'il manque toujours une contribution répulsive

au modèle présenté, on peut supposer que d'autres facteurs de stabilisation manquent, par exemple les phénomènes de dispersion (notamment les forces de London).

Comme vu dans la partie 2.4, les énergies E_{ELEC} et E_{POL} augmentent en magnitude quand la distance diminue. Dans le cas présenté, certains atomes se rapprochent, induisant des interactions plus fortes. Or on cherche à sonder la profondeur du puit de potentiel où se trouve la biotine dans les différents cristaux. Si le site est concave et que ces contributions s'équilibrent en partie, la prise en compte d'effets répulsifs est pertinente. Il est impossible de trouver la profondeur exacte de ce puit, sans estimer les contraintes stériques. Une forte interaction électrostatique peut exister dans une région relativement vide ou au contraire permettre des contacts très étroits entre les atomes. Une solution simple serait d'adapter un potentiel de type Lennard-Jones[112] (ou potentiel de type *m*-*n*) au modèle d'interaction présenté mais ce potentiel devrait techniquement être réévalué à chaque nouvelle version du modèle. Cette solution pose aussi le problème de quelles énergies d'interaction utiliser pour affiner ce potentiel de Lennard-Jones. Une autre solution serait d'estimer empiriquement les énergies d'interaction liées aux phénomènes d'échange-corrélation (et aux phénomènes de dispersion). Cette dernière stratégie a l'avantage d'être indépendante du modèle choisi pour représenter l'induction *in situ* de moments multipolaires et pourrait continuer à utiliser le *benchmark* SAPT utilisé partie 2.4.

L'estimation de ces forces de van der Waals est importante pour évaluer les interactions dites hydrophobes. Ces interactions sont pertinentes pour l'étude de chaque streptavidine. Une partie de l'énergie de désolvatation est liée à l'enthalpie de liaison des molécules d'eau au sein du site de fixation de la streptavidine libre. Évaluer précisément les énergies de dispersion, de répulsion (échange-corrélation) et de désolvatation (c'est-à-dire l'énergie de liaison des molécules de solvant avec le site de fixation de la streptavidine libre) permettrait d'accéder indirectement à une estimation de certaines composantes de la contribution entropique, en particulier celles permettant d'apréhender les interactions hydrophobes. Comparer cette partie manquante avec des variations expérimentales d'entropie de liaison pourrait à l'avenir se révéler riche d'enseignement, mais est encore hors de portée. Des perspectives concernant le traitement des molécules d'eau sont disponibles partie 4.1.

3.7.e - L'interaction streptavidine-biotine du point de vue topologique

La partie précédente s'attachait aux conséquences énergétiques de la polarisation du complexe streptavidine-biotine. La description de cette polarisation à proprement parler n'a pas encore été discutée. La polarisation entraîne une modification de la densité électronique qui doit par conséquent modifier la topologie de ce champ ; les aspects liés à la topologie de la densité électronique et du potentiel électrostatique ont été volontairement laissés à part, du fait de leur caractère plus novateur. En effet, l'objectif original de ce travail de thèse était de développer des outils basés sur la topologie du potentiel électrostatique (et de la densité électronique) pour étudier, et comprendre, les interactions protéine-ligand. Cette partie est une tentative visant à dégager un cadre d'interprétation adapté à cette optique.

Il est proposé de relier les conclusions précédentes – et la déformation de densité électronique liée à la polarisation – à des éléments topologiques, ici des faisceaux de ligne de champ électrique (pour le champ de potentiel électrostatique) et des propriétés au point critique des liaisons hydrogène (pour le champ de densité électronique). Les trois locus d'interaction impliquant des liaisons hydrogène sont présentés. Plusieurs études ont déjà été publiées sur la topologie de $\rho(\mathbf{r})$ et $\varphi(\mathbf{r})$ dans les régions impliquées dans des liaisons hydrogène[37,41,113], reliant la densité au point critique, la trace et les valeurs propres du Laplacien $\nabla^2\rho(\mathbf{r})$ avec la stabilité et les distances de ces liaisons.

Oxygène uréide

L'oxygène uréide interagit avec les résidus dont la contribution électrostatique est la plus intense. Les liaisons hydrogène ont été décrites comme les arrêtes d'un tétraèdre, dont la base est constituée des atomes donneurs dans la sérine 27, la tyrosine 43 et l'asparagine 23. L'oxygène uréide (O3) est le sommet de ce tétraèdre. La Table 12 regroupe les caractéristiques topologiques des trois liaisons non-covalentes de O3 pour les trois systèmes, aux côtés des caractéristiques des deux autres liaisons hydrogène du groupement uréide (impliquant N1 et N2, respectivement avec l'aspartate 128 et la sérine 45). La Figure **48** montre l'organisation du champ électrique de la streptavidine sauvage et met en évidence que ces cinq liaisons hydrogène appartiennent à un même système électrostatique. La densité de polarisation de la biotine – conséquente à l'application du champ – est présentée Figure 47.

Table 12. Caractéristiques topologiques de la densité électronique au points critiques associé à l'oxygène uréide O3 et aux azotes uréides N1 et N2. Les valeurs reportées concernent la densité totale polarisée, les valeurs entre parenthèses matérialisent la différence ($x_{POLARISÉ} - x_{TRANSFÉRÉ}$). *BPL* : longueur du chemin d'interaction.

	Liaison	$\rho_{BCP}(\mathbf{r})$ (e·Å^{-3})	$\nabla^2\rho_{BCP}(\mathbf{r})$ (e·Å^{-5})	λ_1 (e·Å^{-5})	λ_2 (e·Å^{-5})	λ_3 (e·Å^{-5})	*BPL* (Å)
WT	O3 Ser27	0.247 *(-0.040)*	1.29 *(+0.04)*	4.20 *(-0.47)*	-1.51 *(+0.23)*	-1.41 *(+0.28)*	1.766 *(+0.002)*
	O3 Asn 23	0.103 *(-0.018)*	1.19 *(< +0.01)*	2.02 *(-0.20)*	-0.46 *(+0.09)*	-0.37 *(+0.11)*	2.182 *(+0.083)*
	O3 Tyr 43	0.250 *(-0.041)*	2.26 *(-0.02)*	4.17 *(-0.51)*	-1.50 *(+0.23)*	-1.40 *(+0.25)*	1.768 *(+0.005)*
	N1 Asp128	0.230 *(-0.042)*	1.82 *(+0.10)*	4.41 *(-0.39)*	-1.30 *(+0.25)*	-1.29 *(+0.25)*	1.797 *(> -0.001)*
	N2 Ser45	0.165 *(-0.010)*	1.23 *(-0.05)*	2.86 *(-0.14)*	-0.83 *(+0.04)*	-0.80 *(+0.05)*	1.985 *(> -0.001)*
W79F	O3 Ser27	0.447 *(-0.048)*	5.48 *(+2.30)*	11.07 *(+1.25)*	-3.18 *(+0.44)*	-2.42 *(+0.61)*	1.678 *(+0.175)*
	O3 Asn 23	0.102 *(-0.018)*	1.10 *(-0.01)*	1.96 *(-0.20)*	-0.46 *(+0.09)*	-0.40 *(+0.10)*	2.111 *(+0.008)*
	O3 Tyr 43	0.232 *(-0.026)*	1.85 *(+0.13)*	4.25 *(-0.21)*	-1.34 *(+0.13)*	-1.06 *(+0.21)*	1.945 *(+0.147)*
	N1 Asp128	0.202 *(-0.037)*	1.53 *(-0.01)*	3.73 *(-0.42)*	-1.11 *(+0.20)*	-1.10 *(+0.20)*	1.852 *(< +0.001)*
	N2 Ser45	0.197 *(-0.006)*	1.73 *(+0.01)*	3.63 *(-0.04)*	-0.99 *(+0.02)*	-0.91 *(+0.03)*	1.894 *(> -0.001)*
Y43F	O3 Ser27	0.242 *(-0.041)*	1.29 *(+0.01)*	4.15 *(-0.43)*	-1.47 *(+0.21)*	-1.39 *(+0.23)*	1.768 *(+0.002)*
	O3 Asn 23	0.110 *(-0.020)*	1.44 *(+0.07)*	2.29 *(-0.15)*	-0.48 *(+0.10)*	-0.37 *(+0.13)*	2.244 *(+0.183)*
	N1 Asp128	0.124 *(-0.013)*	0.87 *(-0.12)*	1.86 *(-0.24)*	-0.50 *(+0.06)*	-0.48 *(+0.06)*	2.160 *(+0.002)*
	N2 Ser45	0.125 *(-0.002)*	1.03 *(-0.01)*	2.08 *(-0.02)*	-0.54 *(< +0.01)*	-0.51 *(< +0.01)*	2.121 *(+0.005)*

La première constatation concerne la densité électronique des liaisons : lors de l'application de la polarisation, $\rho_{BCP}(\mathbf{r})$ diminue pour chaque liaison hydrogène considérée (y compris celles traitées plus bas). Cette diminution n'est pas homogène. Pour certaines liaisons, par exemple N2···Ser45 du mutant Y43F, cette diminution est négligeable. Au

maximum, cette déplétion relative de charge atteint 18.3 % pour la liaison N1···Asp128 du mutant W79F. On s'attend à une polarisation plus importante du fait du fort champ électrique émis par le groupe carboxylate de l'aspartate, et la liaison N2···Ser45 est relativement faible énergétiquement.

Cependant, il faut prendre en compte les différences de géométrie entre les différents mutants. Plus une liaison hydrogène est courte, plus son énergie est basse et plus elle aura tendance à avoir une forte densité électronique au point critique. La liaison hydrogène la plus forte est O3···Ser27 dans chaque complexe. Pour les mutants W79F et Y43F, c'est également la plus riche en électrons. Pour le sauvage, $\rho_{BCP}(\mathbf{r})$ de cette liaison est la deuxième plus riche du complexe, juste derrière O3···Tyr43. Dans le mutant W79F, O3···Ser27 est la plus forte liaison hydrogène étudiée ici, la plus riche en électrons, et celle qui présente la plus forte déplétion absolue de $\rho_{BCP}(\mathbf{r})$ lorsqu'on compare les densités transférées et polarisées. Toujours dans ce mutant W79F, la plus forte valeur du Laplacien et la plus forte valeur propre λ_1 correspondent à O3···Ser27. La liaison O3···Tyr43 est très comparable avec la liaison O3···Ser27 de la streptavidine sauvage du point de vue de l'énergie et des distances d'interaction, mais aussi du point de vue de $\rho_{BCP}(\mathbf{r})$, de $\nabla^2\rho_{BCP}(\mathbf{r})$ et de λ_1.

Du point de vue topologique, il est aussi possible d'appréhender les différences de fixation en analysant les positions des points critique de liaisons. Comme vu précédemment, la polarisation diminue la valeur de $\rho_{BCP}(\mathbf{r})$ mais aussi déplace le point critique associé. Ce déplacement au cours de la polarisation peut s'apprécier par la variation de longueur du chemin d'interaction (*BPL*). La *BPL* ne peut être inférieure à la distance interatomique. Si la *BPL* augmente, le point critique s'éloigne de l'axe d'interaction. Au cours de la polarisation, la variation de *BPL* est toujours négligeable ou positive. Trois chemins d'interaction gagnent plus de 0.15 Å après polarisation : O3···Ser27 (W79F), O3···Tyr43(W79F) et O3···Asn23 (Y43F). La *BPL* de O3···Asn23 dans le variant sauvage montre un allongement intermédiaire à +0.08 Å. La forte variation de *BPL* peut s'expliquer par l'intensité des liaisons O3···Ser27 (W79F) et O3···Tyr43(W79F). La liaison O3···Asn23 (Y43F) n'est pas aussi énergétique. Pourtant la polarisation de O3 (voir Figure 47) est quasiment identique dans Y43F et dans le variant sauvage, malgré la mutation de la tyrosine. Dans le complexe sauvage le moment induit sur O3 est 0.215 e·Å et dans le mutant Y43F, 0.212 e·Å (l'orientation relative du champ électrique de la

protéine est similaire). Ceci appuie le rôle prépondérant de la sérine 27 dans la polarisation de l'oxygène uréide et relie des allongements similaires de la *BPL* de la liaison O3···Asn23 avec une polarisation similaire. Dans le mutant W79F, les liaisons O3···Ser27 et O3···Tyr43 s'allongent. La polarisation dans ce mutant est plus forte (0.323 e·Å), à la fois du fait d'un champ électrique plus intense et d'un meilleur alignement entre la direction du premier vecteur propre de la polarisabilité et le champ électrique.

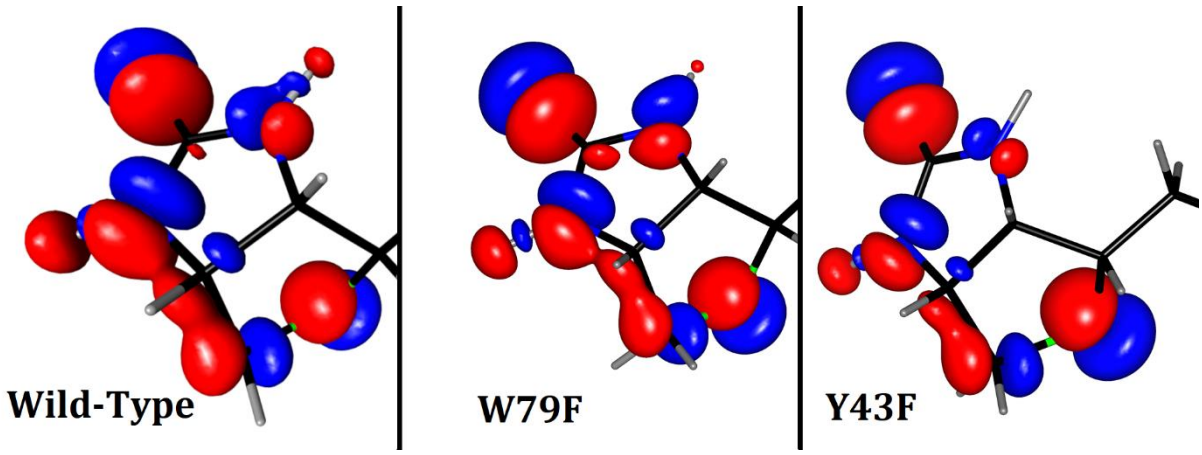


Figure 47. Densité électronique de polarisation $\rho_{pol}(\mathbf{r}) = \rho_{POLARISÉ}(\mathbf{r}) - \rho_{TRANSFÉRÉ}(\mathbf{r})$ de la tête de la biotine. Le contour des isosurfaces est à ±0.03 e·Å^3.

Sur la Figure 48, si on compare la direction du moment induit – quasi colinéaire à la liaison covalente de O3 – à la direction des liaisons hydrogène, on observe que le cortège électronique de O3 va se rapprocher de la sérine 27 et de la tyrosine 43, déplaçant le point critique entre ces deux résidus.

La plus grande valeur propre du Laplacien, λ_1, est toujours positive et orientée dans la direction globale de l'interaction. Elle décrit la courbure de la vallée de densité électronique qui sépare les atomes. O3···Ser27 (W79F) a encore une fois la courbure la plus forte. Pour les liaisons O–H···O de ce locus, plus l'interaction est énergétique, plus la courbure est forte. À l'application de la polarisation, cette même liaison hydrogène O3···Ser27 (W79F) montre un comportement opposé aux autres liaisons hydrogène. Pour toutes les autres, λ_1 diminue à l'application de la polarisation ; pour la liaison la plus courte et intense, O3···Ser27 (W79F), λ_1 augmente à l'application de la polarisation. Pour les liaisons de type N–H···O, dans le complexe sauvage, il apparaît que la liaison

N1···Asp128 a une valeur de λ_1 largement supérieure à N2···Ser45 ou O3···Asn23. Cette liaison est allongée dans les mutants W79F et Y43F, et on ne retrouve pas le même constat dans ces complexes. Le lien entre λ_1, la densité d'énergie cinétique et la répulsion électronique est discutée plus bas.

Sur la Figure 48, on constate que la proximité de cet oxygène est à la confluence de 3 faisceaux de lignes de champ, au départ de chaque hydrogène. Au voisinage du noyau de O3, les lignes de champ paraissent presque parallèles. La sérine 27 a la plus forte interaction électrostatique avec la biotine. Le faisceau de lignes de champ électrique le plus large part de la sérine 27. Les lignes de champ suivent la direction de la liaison O=C de la biotine puis se divisent en deux au voisinage du carbone uréide. La plus grande partie des lignes se termine au niveau des doublets libres de l'aspartate 128 (où le potentiel électrostatique présente un minimum local). Le second bouquet de lignes, plus fin, bifurque et se termine au niveau des doublets non-liants de l'oxygène γ. Le faisceau au départ de Hζ de la tyrosine 43 suit le même trajet que le faisceau précédent : au niveau de l'oxygène uréide, il suit la direction de la liaison O=C, puis se sépare en deux au niveau du carbone. La plus grande partie des lignes de champ au départ de la tyrosine se terminent sur l'aspartate 128, la plus petite partie sur la sérine 45. Parmi ces trois liaisons hydrogène, l'asparagine 23 est associée à l'énergie d'interaction $E_{ELEC}+E_{POL}$ la plus faible. Elle est aussi le point de départ du faisceau de lignes de champ le plus ténu. L'intégralité des lignes de champ qui partent du groupement NH_2– de l'asparagine 23, en direction de la biotine, convergent vers l'aspartate 128 (l'oxygène amide de la Figure 48 est à l'avant-plan). Ce faisceau est légèrement excentré par rapport à O3, ce qui est cohérent avec une importance électrostatique moindre (comparé à sérine 27 et tyrosine 43). On observe également clairement, au niveau de l'oxygène uréide, que les lignes de champ suivent le grand axe de la polarisabilité transférée. Cette position est bien entendu la plus favorable du point de vue de l'énergie de polarisation, autorisant la maximisation du dipôle atomique de O3 dans cette conformation.

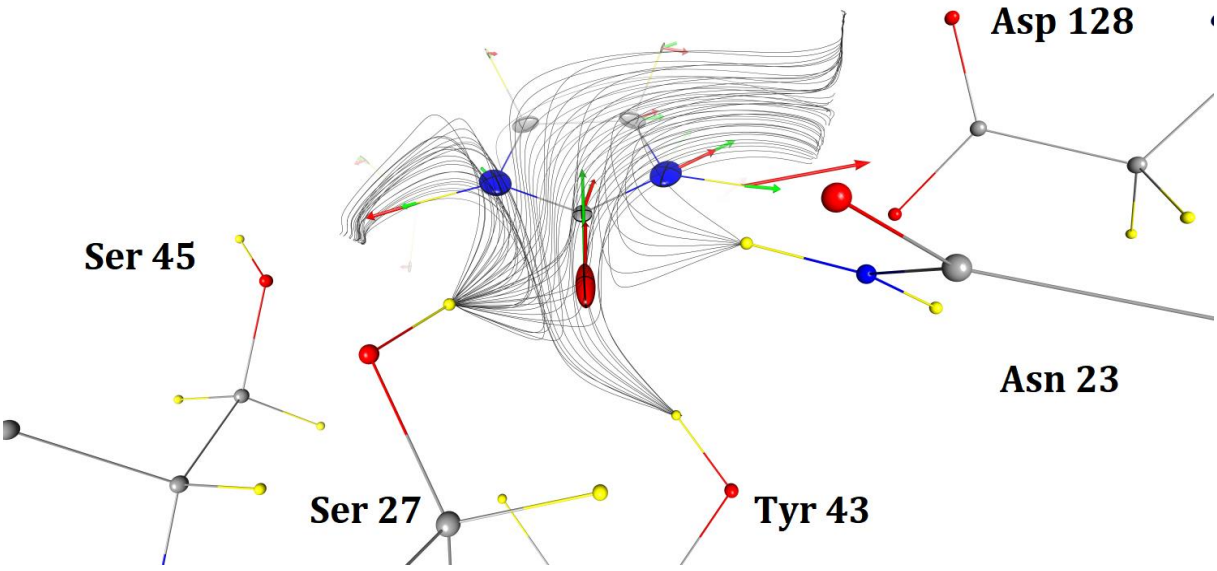


Figure 48. Les lignes de champ électrique permettent de trouver les acides aminés qui influencent la polarisation du groupement uréide de la biotine. Seul le champ électrique des acides aminés est représenté. Les flèches vertes symbolisent les moments dipolaires induits sur la biotine, les flèches rouges, le champ électrique dû aux résidus de contact. Échelles : 0.15 Å/Å^3 (polarisabilités) ; 10 Å/e·Å (moments dipolaires) ; 10 Å/e·Å^2 (champs électriques).

Soufre thiophène

Le second locus d'interaction implique une seule liaison hydrogène de bien plus faible intensité : celle qui relie le soufre thiophène à la thréonine 90. Ce soufre est modélisé avec 16.14 électrons, on s'attend donc à des valeurs de densité au point critique plus fortes que pour le locus précédent. Pourtant ces trois liaisons O–H···S – présentées Table 13 – ont une densité électronique au point critique proche de la borne inférieure des valeurs obtenues au précédent locus d'interaction. Les liaisons O–H···S sont longues : les distances O-S sont supérieures à 3 Å et les BPL sont supérieures à 2 Å (la détermination de la *BPL* par *MoProViewer* a généré des valeurs incohérentes pour 3 mesures, ces *BPL* sont notées n.d.). La déplétion de $\rho_{BCP}(\mathbf{r})$ suite à la polarisation est très faible (< 4%).

Table 13. Caractéristiques topologiques de la densité électronique au points critiques associé au soufre thiophène (*BPL* : longueur du chemin d'interaction). Les valeurs reportées concernent la densité polarisée, les valeurs entre parenthèses matérialisent la différence ($x_{POLARISÉ} - x_{TRANSFÉRÉ}$).

	Liaison	$\rho_{BCP}(\mathbf{r})$ (e·Å^{-3})	$\nabla^2\rho_{BCP}(\mathbf{r})$ (e·Å^{-5})	λ_1 (e·Å^{-5})	λ_2 (e·Å^{-5})	λ_3 (e·Å^{-5})	*BPL* (Å)	Dist. H···S (Å)
WT	Thr 90	0.128 *(-0.004)*	0.57 *(-0.06)*	1.60 *(-0.08)*	-0.52 *(+0.01)*	-0.50 *(+0.01)*	2.548 *(n.d.)*	2.433
W79F	Thr 90	0.113 *(-0.003)*	0.56 *(-0.05)*	1.43 *(-0.06)*	-0.45 *(+0.01)*	-0.43 *(> -0.01)*	*n.d* *(n.d.)*	2.497
Y43F	Thr 90	0.212 *(-0.004)*	0.66 *(-0.04)*	2.61 *(-0.05)*	-1.01 *(+0.01)*	-0.94 *(> -0.01)*	2.144 *(0.00)*	2.139

La trace de $\nabla^2\rho_{BCP}(\mathbf{r})$ est la plus faible des trois complexes. Cette densité électronique plus plate est sensée considérant la longue distance d'interaction et la densité diffuse du soufre. Encore une fois, la polarisation entraine une diminution de la courbure de $\rho_{BCP}(\mathbf{r})$, même si cet effet est très faible sur les courbures négatives (transversales). Ainsi, pour les complexes sauvage et W79F, la courbure positive λ_1 associée à cette liaison hydrogène S···Thr90 est la plus faible du complexe. Pour le mutant Y43F, cette liaison est raccourcie. La magnitude de l'interaction E_{ELEC}+E_{POL} augmente de moitié. Ce raccourcissement lié à la position légèrement différente de la tête de la biotine dans le mutant Y43F entraine une valeur de λ_1 bien supérieure aux autres liaisons S···Thr90 (2e courbure plus forte du mutant Y43F). Au sein de ce mutant, la liaison S···Thr90 est la 5e liaison hydrogène la plus intense. La capacité à extrapoler l'intensité de l'interaction électrostatique sur la base des caractéristiques du point critique (densité / Laplacien) apparaît comme valable au sein de liaisons hydrogène impliquant les mêmes éléments chimiques.

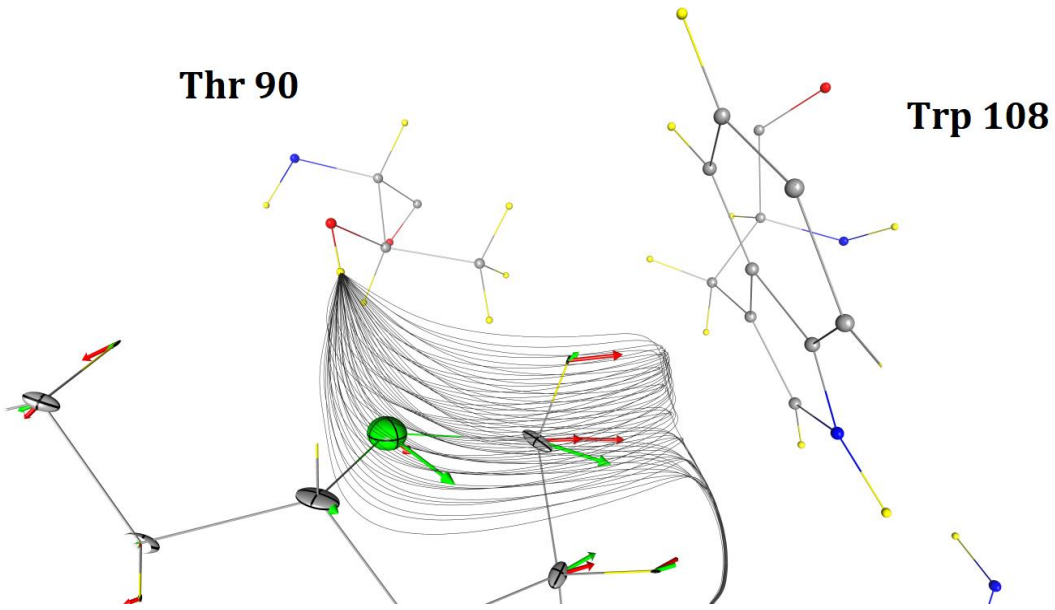


Figure 49. Les lignes de champ électrique permettent de trouver les acides aminés qui influencent la polarisation du soufre de la biotine. Seul le champ généré par la protéine est représenté. Les flèches vertes symbolisent les moments dipolaires induits sur la biotine, les flèches rouges, le champ électrique des résidus de contact. Le faisceau de ligne de champ partant vers le bas de la figure se termine au contact du carboxylate de Asp 128. Échelles : 0.10 Å/Å^3 (polarisabilités) ; 10 Å/e·Å (moments dipolaires) ; 10 Å/e·Å^2 (champs électriques).

Cet atome de soufre est très polarisable et la Figure 47 montre une densité de déformation – due à la polarisation – importante malgré l'intensité relativement faible de cette interaction. La position de la biotine amène le grand axe de la polarisabilité (associé à λ_1) dans une direction proche de celle du champ électrique. Cette configuration permet le développement d'un fort moment dipolaire induit au niveau du soufre : 0.103 e·Å soit 0.49 D (α_{ISO} = 2.31 Å^3). Cette induction est comparable au dipôle permanent du soufre (0.43 D) et pour cette interaction, E_{POL} représente 27% de E_{ELEC}+E_{POL}. Le faisceau de lignes de champ associé à cette liaison est montré Figure 49. Ici, l'organisation spatiale du faisceau est beaucoup plus simple que le précédent : toute les lignes partent de l'hydrogène (de l'hydroxyle γ) vers le soufre thiophène, puis bifurquent dans la direction du tryptophane 108, en suivant la liaison covalente S–C(H_2). Dans notre modèle, le noyau indole du tryptophane porte une charge légèrement négative sur les atomes de carbone et d'azote, représentant les électrons π des cycles aromatiques, ce qui ferme les bassins topologiques du potentiel électrostatique associés à ces noyaux. Les lignes de champ en provenance de la thréonine 90 se terminent au niveau de la surface de flux nul qui entoure

le noyau indole du tryptophane 108. Une petite partie de ces lignes de champ continue son chemin jusqu'au carboxylate de l'aspartate 128, mais un artefact numérique peut causer ce type de représentation.

Table 14. Caractéristiques topologiques de la densité électronique au points critiques associé aux oxygènes carboxylates O11 et O12 (*BPL* : longueur du chemin d'interaction). Les valeurs reportées concernent la densité polarisée, les valeurs entre parenthèses matérialisent la différence ($x_{POLARISÉ} - x_{TRANSFÉRÉ}$).

	Liaison	$\rho_{BCP}(\mathbf{r})$ (e·Å^{-3})	$\nabla^2\rho_{BCP}(\mathbf{r})$ (e·Å^{-5})	λ_1 (e·Å^{-5})	λ_2 (e·Å^{-5})	λ_3 (e·Å^{-5})	*BPL* (Å)
WT	Asn 49	0.205 *(-0.040)*	1.56 *(+0.03)*	3.79 *(-0.42)*	-1.12 *(+0.22)*	-1.11 *(+0.23)*	1.828 *(+0.002)*
	Ser 88	0.217 *(-0.044)*	0.72 *(-0.22)*	3.33 *(-0.71)*	-1.31 *(+0.25)*	-1.30 *(+0.24)*	1.826 *(> -0.001)*
W79F	Asn 49	0.171 *(-0.028)*	1.50 *(-0.02)*	3.15 *(-0.33)*	-0.88 *(+0.15)*	-0.77 *(+0.16)*	1.915 *(+0.006)*
	Ser 88	0.334 *(-0.084)*	2.40 *(+1.08)*	6.91 *(-0.21)*	-2.28 *(+0.64)*	-2.24 *(+0.66)*	1.592 *(-0.001)*
Y43F	Asn 49	0.154 *(-0.026)*	1.19 *(-0.10)*	2.63 *(-0.36)*	-0.73 *(+0.13)*	-0.71 *(+0.13)*	1.976 *(< +0.001)*
	Ser 88	0.162 *(-0.025)*	0.64 *(-0.32)*	2.35 *(-0.56)*	-0.87 *(+0.12)*	-0.85 *(+0.12)*	1.975 *(+0.004)*

Queue aliphatique et carboxylate

Le dernier locus examiné concerne la queue carboxylate de la biotine, qui interagit avec la protéine par deux fortes liaisons hydrogène. Les variations d'interaction électrostatique à cet endroit sont importantes pour expliquer la différence d'enthalpie entre le complexe sauvage et les deux mutants. Par rapport au complexe sauvage, le complexe W79F présente une liaison O12···Ser88 raccourcie et fortement stabilisée. Dans le mutant Y43F, ces deux liaisons hydrogène, O12···Ser88 et O11···Asn49, sont allongées et contribuent à l'enthalpie réduite entre biotine et streptavidine.

Dans le complexe sauvage, $\rho_{BCP}(\mathbf{r})$ est relativement élevé pour les liaisons hydrogène de la queue carboxylate comparé aux liaisons précédentes. O12···Ser88 et O11···Asn49 sont associés à des énergies d'interaction proches et ont des densités

électroniques comparables au point critique. Le raccourcissement de la liaison O12···Ser88 dans le complexe W79F est visible sur la $\rho_{BCP}(\mathbf{r})$ de cette interaction. Seule la liaison O3···Ser27 a une $\rho_{BCP}(\mathbf{r})$ supérieure dans ce mutant W79F. Ces liaisons raccourcies sont associées à des énergies d'interaction plus favorables. La densité de déformation présentée Figure 50 souligne l'asymétrie de la polarisation du carboxylate dans le mutant W79F (la polarisation dans le complexe sauvage est la plus symétrique des trois). Cependant les densités de déformation des oxygènes restent globalement orientées dans l'axe des liaisons covalentes. Pour les deux liaisons du mutant Y43F, $\rho_{BCP}(\mathbf{r})$ diminue par rapport au sauvage, dans une mesure comparable à la diminution liée à la liaison N1···Asp128, qui est également allongée dans le complexe Y43F par rapport au complexe sauvage. L'énergie de N1···Asp128 est diminuée dans le mutant Y43F mais dans une mesure plus importante que l'énergie associée à O12···Ser88 et O11···Asn49. La différence de fixation dans le complexe Y43F est soulignée par la densité de déformation de la Figure 50, visuellement différente de celle des deux autres variant, sauvage et W79F. Les valeurs de $\rho_{BCP}(\mathbf{r})$ retrouvées dans ces complexes sont de l'ordre de grandeur de celles trouvées par Howard *et al.*, 2016[12], pour l'étude topologique des liaisons hydrogène entre la fonction carboxylate de l'acide oléique et la FABP (*Fatty-Acid Binding Protein*).

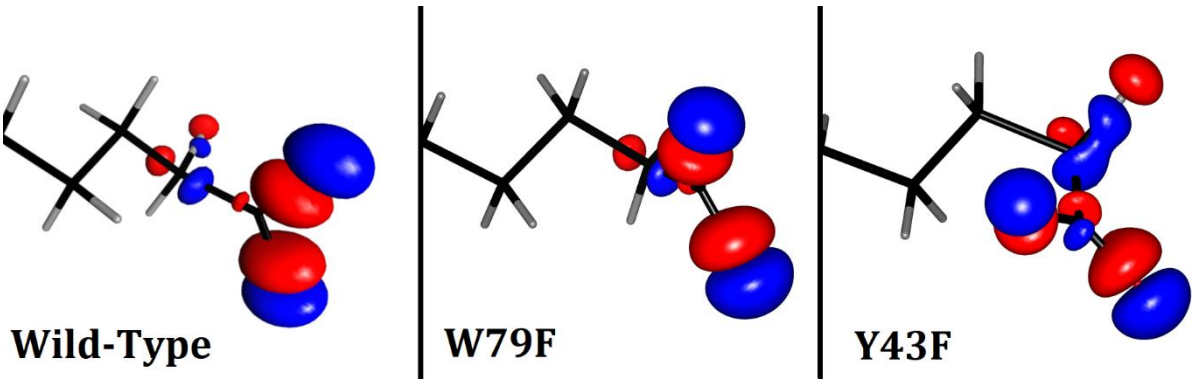


Figure 50. Densité de polarisation $\rho_{POLARISÉ}(\mathbf{r})$ - $\rho_{TRANSFÉRÉ}(\mathbf{r})$ de la queue de la biotine. Le contour des isosurfaces est à ±0.03 e·Å^3.

La liaison N1···Asp128 est un point de comparaison pour la liaison O11···Asn49, car elle implique un carboxylate chargé négativement, un azote donneur et une distance d'interaction comparable. Dans le complexe sauvage, $\nabla^2\rho_{BCP}(\mathbf{r})$ de O11···Asn49 est comparable à la liaison avec N1···Asp128. La même constatation peut être tirée sur le

complexe Y43F, mais les valeurs numériques sont atténuées compte tenu de l'allongement des liaisons. Dans ces deux complexes (sauvage et Y43F), le Laplacien de la liaison O12···Ser88 est très faible comparé aux autres liaisons impliquant une sérine (y compris pour la liaison N2···Ser45, de faible énergie). Ce comportement n'est pas attendu mais ne semble pas lié à la correction manuelle de la position de Hγ de la sérine 88, puisque le mutant Y43F n'a pas nécessité cette correction. Dans le mutant W79F, la fonction carboxylate se rapproche de la sérine 88. Il en découle une forte augmentation de $\nabla^2\rho_{BCP}(\mathbf{r})$, et comme pour $\rho_{BCP}(\mathbf{r})$, seule O3···Ser27 a un $\nabla^2\rho_{BCP}(\mathbf{r})$ plus élevé dans ce mutant. Les caractéristiques de la liaison O11···Asn49 dans le complexe W79F sont comparables à celles du complexe sauvage.

Examiner uniquement λ_1 ne donne pas tout à fait le même portrait des interactions. Pour le complexe sauvage et Y43F, $\nabla^2\rho_{BCP}(\mathbf{r})$ est très différent pour les liaisons O11···Asn49 et O12···Ser88. Par contre ces deux liaisons ont des λ_1 et des distances d'interaction très comparables. Dans le mutant Y43F, λ_1 est diminué par rapport au complexe sauvage, tout comme $\nabla^2\rho_{BCP}(\mathbf{r})$, pour les deux liaisons du groupement carboxylate. Dans le complexe W79F, la liaison O11···Asn49 est encore très proche du complexe sauvage, avec une courbure λ_1 comparable à celle retrouvée dans la liaison N1···Asp128. λ_1 est légèrement supérieure dans la liaison N1···Asp128 du complexe sauvage. Dans le mutant W79F, la liaison O12···Ser88 présente une forte λ_1, 3e plus forte de tout le complexe W79F, derrière O3···Ser27 et O3···Tyr43. Encore une fois, les valeurs de $\nabla^2\rho_{BCP}(\mathbf{r})$ sont comparables à celles reportées dans la littérature (Howard *et al.*[12], Nassour *et al.*[82])

L'application de la polarisation induit une diminution de $\rho_{BCP}(\mathbf{r})$ qui est comparable à la diminution obtenue pour d'autres liaisons hydrogène de forte énergie (typiquement O3···Ser27). Ces liaisons de la queue carboxylate impliquent de fortes E_{POL} et de forts champs électriques. Comme dans d'autres liaisons comparables en termes d'énergie (O3···Ser27 et O3···Tyr43), on constate un aplatissement similaire des courbures individuelles avec la polarisation, pour les deux liaisons et les trois protéines. La polarisation allonge très peu (voire pas du tout) les chemins d'interaction, comme constaté au niveau du cycle uréide.

Les lignes de champ électrique sont présentées Figure 51. Deux faisceaux de lignes de champ baignent la fonction carboxylate. Le premier part de la sérine 88 (la liaison

hydrogène la plus forte) et bifurque entre les oxygènes du carboxylate vers le carbone méthylène. Les lignes de champ sont globalement orientées dans la direction de la liaison carbone-carbone (verticale sur la figure). Environ 1 Å après ce carbone méthylène, le faisceau se sépare en deux. Une partie se termine au niveau de la surface de flux nul entourant l'oxygène peptidique de l'asparagine 49. Cependant, la majorité se termine au niveau de la surface de flux nul entourant la chaîne latérale du tryptophane 79. Le second faisceau, plus petit, part de l'hydrogène peptidique de l'asparagine 49. Il rejoint le premier faisceau au niveau du carbone carboxylate. Presque toutes les lignes de champ qui partent de cet atome forment une boucle pour terminer au niveau de l'oxygène peptidique du même acide aminé.

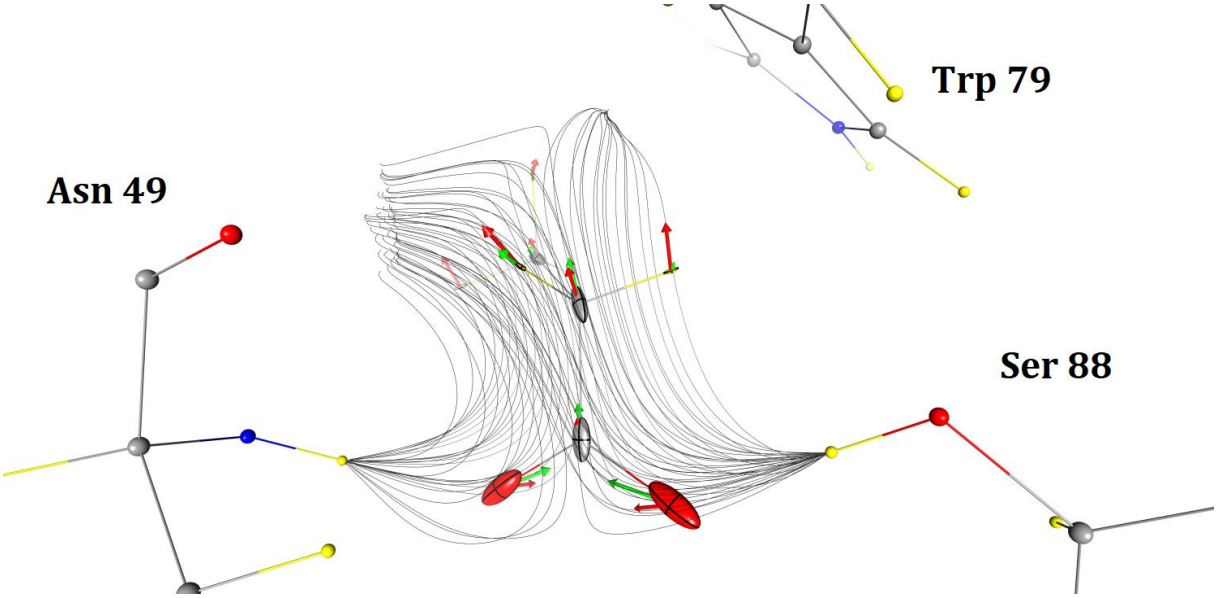


Figure 51. Les lignes de champ électrique permettent de trouver les acides aminés qui influencent la polarisation de la queue carboxylate de la biotine. Seul le champ électrique des acides aminés est représenté. Les flèches vertes symbolisent les moments dipolaires induits sur la biotine, les flèches rouges, le champ électrique des résidus de contact. Échelles : 0.15 Å/Å^3 (polarisabilités) ; 10 Å/e·Å (moments dipolaires) ; 10 Å/e·Å^2 (champs électriques).

Au niveau de ce groupement, on trouve une géométrie tétraédrique des lignes de champ (toutefois différente de celle observée au niveau de l'oxygène uréide, qui concernait la géométrie des liaisons hydrogène). Deux sommets de la pyramide sont des émetteurs de lignes de champ : Hγ de la sérine 88 et H peptidique de l'asparagine 49. Les

deux autres sommets sont des attracteurs de lignes de champ : le noyau indole du tryptophane 79 et l'oxygène peptidique de l'asparagine 49.

3.7.f - Conclusions sur l'approche topologique

Les liaisons hydrogène des trois complexes sont relativement diverses, impliquant différentes distances et différents atomes, mais l'échantillon étudié ne regroupe que 23 liaisons hydrogène. Il est difficile de réaliser une comparaison systématique qui donnerait une idée de la performance du modèle uniquement sur la base de l'étude du complexe streptavidine-biotine. Pourtant plusieurs observations générales sont possibles.

Les valeurs de $\rho_{BCP}(\mathbf{r})$ sont toutes comprises entre 0.1 et 0.3 e·Å^{-3}, à deux exceptions près : la liaison O3···Ser27 et 012···Ser88, dans le complexe d'enthalpie améliorée W79F. Toutes les valeurs de $\rho_{BCP}(\mathbf{r})$ sont relativement plus fortes que celles trouvées par Mata *et al.*, 2010, sur la base de densités électroniques de type MP2/6-311++G(d,p)[113]. Ils trouvent une corrélation exponentielle inverse entre $\rho_{BCP}(\mathbf{r})$ et la distance (R^2 = 0.979). La régression sur les données présentées ici est disponible panneau (a) de la Figure 52. Une tendance générale se dégage. Les deux valeurs les plus hautes de $\rho_{BCP}(\mathbf{r})$ (complexe W79F ; O3···Ser27 et 012···Ser88) concernent les deux liaisons hydrogène les plus courtes et les plus fortes. Le processus de polarisation induit un appauvrissement de la densité au point critique, d'autant plus fort dans les liaisons hydrogène les plus stables, et contribue donc à légèrement rapprocher $\rho_{BCP}(\mathbf{r})$ de la fourchette de valeurs obtenues par Mata *et al.*. Le panneau (b) de la Figure 52 montre la corrélation entre $\rho_{BCP}(\mathbf{r})$ et E_{ELEC}+E_{POL}. La corrélation était attendue car ces deux variables sont toutes deux liées à la distance, cette tendance est cependant la plus faible de cette figure (R^2 = 0.627).

Le signe du Laplacien $\nabla^2\rho_{BCP}(\mathbf{r})$ est toujours positif. Il aurait été dramatique de trouver des valeurs négatives pour le Laplacien, car celles-ci sont caractéristiques des liaisons covalentes[33]. Les liaisons hydrogène de la streptavidine ont donc un caractère « ionique », ou plutôt électrostatique dans le contexte présent (mais dans tous les cas, c'est une interaction dite *closed-shell*). Les $\rho_{BCP}(\mathbf{r})$ du modèle présenté sont environ un ordre de grandeur en dessous des valeurs attendues pour des liaisons covalentes. $\nabla^2\rho_{BCP}(\mathbf{r})$ est légèrement corrélé avec E_{ELEC}+E_{POL} (R^2 = 0.544, données non-incluses), cette tendance était attendue puisque les valeurs propres λ_1, λ_2 et λ_3 sont corrélées avec les densités

énergétiques ci-dessous. La variation de $\nabla^2\rho_{BCP}(\mathbf{r})$ avec la polarisation peut être vue comme une augmentation ou une baisse du caractère ionique d'une liaison hydrogène ; à trois exceptions près, $\nabla^2\rho_{BCP}(\mathbf{r})$ diminue avec la polarisation. Les liaisons où $\nabla^2\rho_{BCP}(\mathbf{r})$ augmente avec la polarisation sont très courtes et énergétiques (O3···Tyr43, O3···Ser27 et O12···Ser88 du complexe W79F).

Dans le formalisme QTAIM, il est possible d'exprimer la densité d'énergie $H(\mathbf{r})$ au point critique[33] à l'aide du théorème du viriel local :

$$\frac{h^2}{(4\pi)^2 \cdot m}\nabla^2\rho_{BCP}(\mathbf{r}) = 2 \cdot G(\mathbf{r}) + V(\mathbf{r})$$

$$\text{Où :} \quad H(\mathbf{r}) = G(\mathbf{r}) + V(\mathbf{r}) \qquad (\mathbf{78})$$

où m est la masse de l'électron, $V(\mathbf{r})$ est la densité d'énergie potentielle (toujours négative) et $G(\mathbf{r})$ la densité d'énergie cinétique des électrons (toujours positive).

Mata *et al.*, 2010[113], trouvent une forte corrélation exponentielle inverse entre ces énergies $G(\mathbf{r})$ et $V(\mathbf{r})$ avec la distance d'interaction [$V(\mathbf{r})$ augmente plus vite que $G(\mathbf{r})$ de manière à ce que les interactions restent favorables]. λ_1 est très fortement corrélée à $G(\mathbf{r})$ et cette observation est à mettre en parallèle avec le signe de $G(\mathbf{r})$: une forte énergie cinétique tend à diluer les électrons, ils fuient le point critique et la courbure de $\rho(\mathbf{r})$ augmente. Sur la base de notre modèle, il est possible d'estimer $G(\mathbf{r})$ à partir de λ_1. Encore une fois, on retrouve la tendance de la littérature sur le panneau (c) de la Figure 52. La régression de type $a \cdot x^{-b}$ donne un R^2 notablement plus élevé que les autres régressions effectuées.

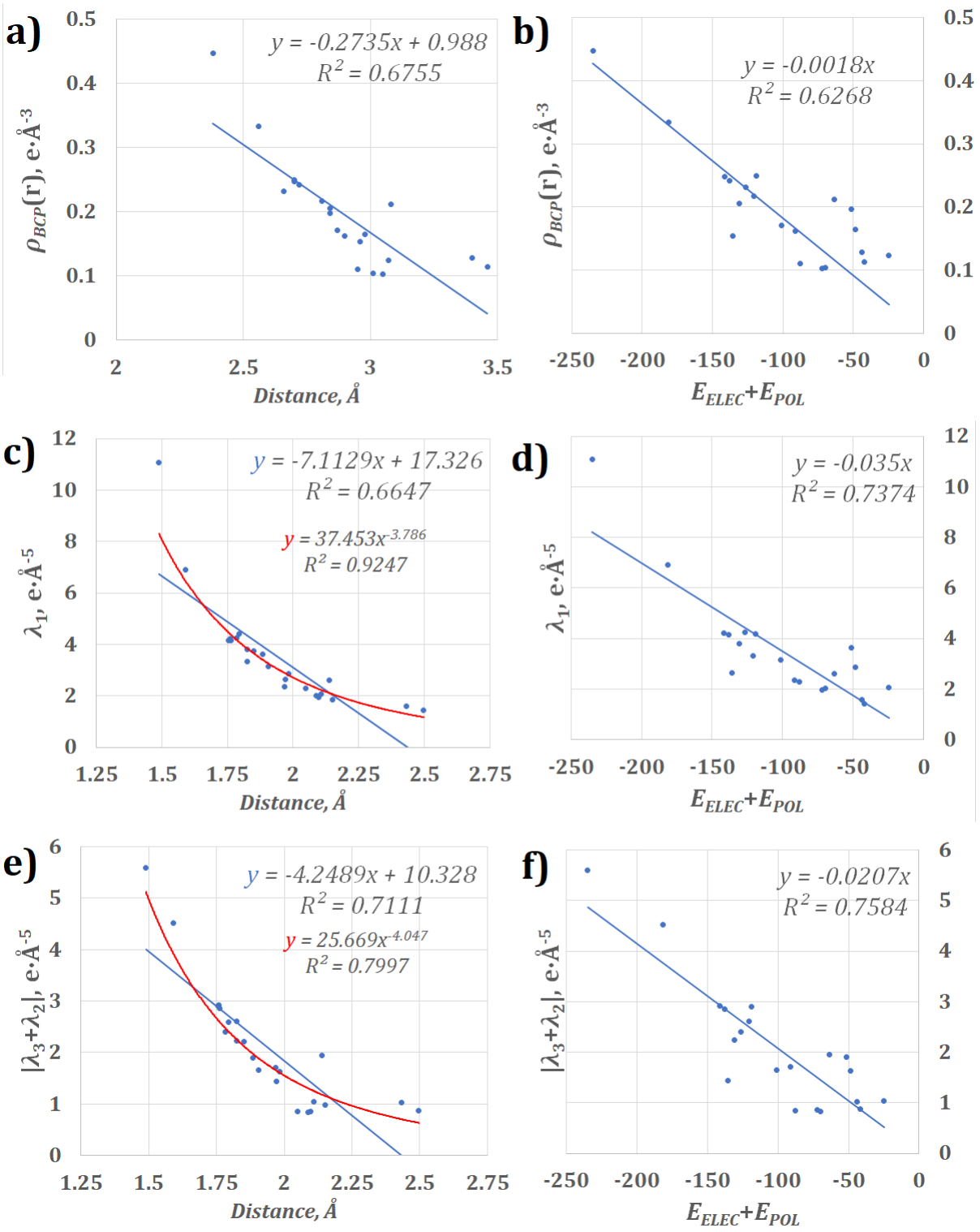


Figure 52. Corrélations entre $\rho_{BCP}(\mathbf{r})$, λ_1 ou $|\lambda_3+\lambda_2|$ et les distances d'interaction ou l'énergie $E_{ELEC}+E_{POL}$ (kJ·mol^{-1}). Les distances sont ici exprimées entre l'atome accepteur et l'hydrogène impliqué. La liaison N1···Asp128 a été exclue des régressions impliquant $E_{ELEC}+E_{POL}$, du fait de son énergie positive.

La densité d'énergie potentielle $V(\mathbf{r})$ est la seule contribution capable de rassembler les électrons au niveau de la liaison chimique. $|V(\mathbf{r})|$ est corrélé à $|\lambda_3+\lambda_2|$. En effet, plus $V(\mathbf{r})$ est négatif, plus les électrons de la liaison sont rassemblés le long de l'axe de la liaison. $|\lambda_3+\lambda_2|$ corrèle avec la distance, en exponentielle inverse chez Mata *et al.*, et cette tendance est retrouvée dans la présente étude. Comme pour λ_1, la régression de forme $a \cdot x^{-b}$ génère un meilleur R^2 (bien que pour $|\lambda_3+\lambda_2|$, l'amélioration est marginale). Par contre le paramètre b de ces deux régressions contre la distance d'interaction est proche de 4 dans les deux cas. Dans Mata *et al.*, la corrélation entre $|\lambda_3+\lambda_2|$ et λ_1 a presque les mêmes paramètres que la régression en $|V(\mathbf{r})|$ et $G(\mathbf{r})$. Dans notre modèle comme dans leur étude des liaisons hydrogène, $|\lambda_3+\lambda_2|$ est inférieur à λ_1.

Le panneau (d) de la Figure 52 confirme que de λ_1 et $E_{ELEC}+E_{POL}$ sont corrélées. De même, le panneau (f) souligne la corrélation entre $|\lambda_3+\lambda_2|$ et $E_{ELEC}+E_{POL}$, suggérant que les interactions les plus fortes impliquent de forts $V(\mathbf{r})$ et $G(\mathbf{r})$ à courte distance. La corrélation entre $|\lambda_3+\lambda_2|$ et $E_{ELEC}+E_{POL}$ est légèrement plus forte, mais son coefficient directeur est plus faible, ce qui est contre-intuitif avec une plus grande augmentation de $V(\mathbf{r})$ par rapport à $G(\mathbf{r})$ à mesure que la distance diminue.

La procédure de polarisation transforme la densité au point critique. D'abord, un déplacement du point critique est possible : dans certains cas, le point critique de la liaison se déplace légèrement plus loin de l'axe de l'interaction, à plus forte raison que le moment induit n'est pas dans l'axe de l'interaction (typiquement O3···Tyr43 dans le complexe W79F). D'autres liaisons montrent des déplacements négligeables (exemple : S···Thr90 dans le mutant Y43F, où le champ électrique est dans l'axe de la liaison hydrogène).

La valeur du Laplacien diminue marginalement au cours de la polarisation dans la plupart des cas, sauf pour les liaisons hydrogène les plus courtes du complexe W79F (O3···Tyr43, O3···Ser27, O12···Ser88), ce qui correspond à une augmentation du caractère ionique de ces liaisons hydrogène. Si ce n'est pas étonnant pour la liaison O12···Ser88 de la queue carboxylate, ce caractère ionique peut être rapproché d'une remarque trouvée dans Klumb *et al.*, 1998[106], qui mentionne une étude théorique où une forme oxyanion de O3 aurait été retrouvée en résonnance avec la forme sp^2 modélisée pour l'oxygène uréide. Cette interprétation n'est cependant pas absolue car dans la liaison N1···Asp128, $\nabla^2\rho_{BCP}(\mathbf{r})$

diminue malgré l'implication d'un anion. $\nabla^2\rho_{BCP}(\mathbf{r})$ est tout de même particulièrement faible pour la liaison N1···Asp128 dans le complexe Y43F, où cette liaison est étirée.

La courbure positive λ_1 de la densité est réduite par la procédure de polarisation. Les variations des courbures négatives λ_2 et λ_3 suite à la polarisation sont en général plus faibles, mais l'effet est systématiquement négligeable ou dans le sens d'une baisse de magnitude. En revanche, la liaison O3···Ser27 du mutant W79F est la plus forte et courte de l'échantillon étudié, cette liaison montre une augmentation de λ_1 et de $\nabla^2\rho_{BCP}(\mathbf{r})$ suite à la polarisation. Dans la direction de la liaison hydrogène, la polarisation entraîne en général une vallée de densité plus profonde et moins courbée, sauf dans le cas des liaisons les plus fortes où la courbure augmente.

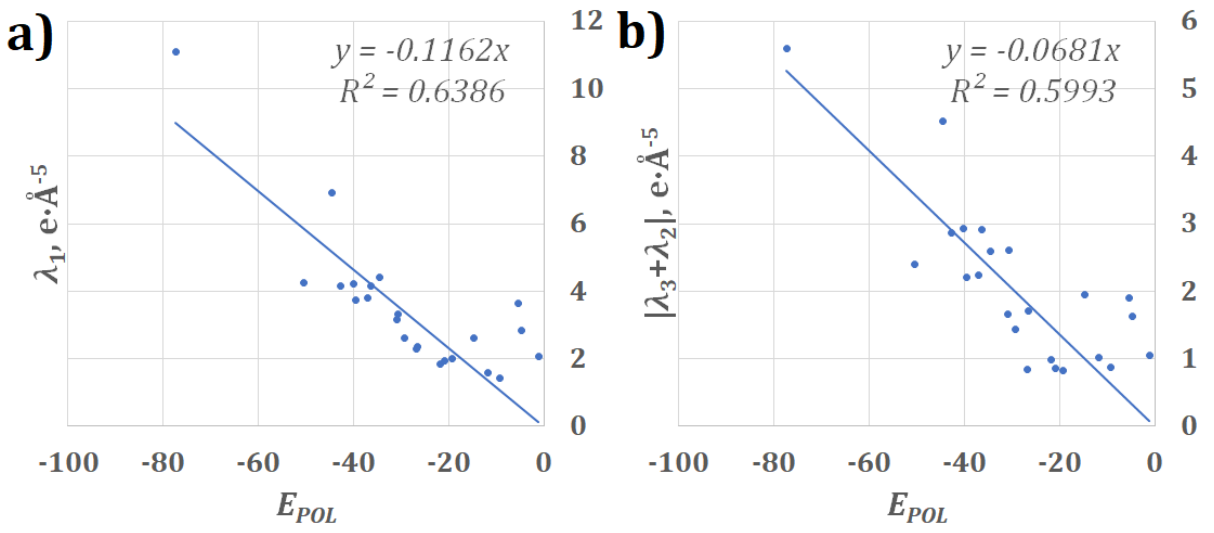


Figure 53. Corrélation entre les énergies de polarisation (kJ·mol^{-1}) et les valeurs propres λ_1 [corrélée à $G(\mathbf{r})$] et $|\lambda_2+\lambda_3|$ [corrélée à $V(\mathbf{r})$].

La Figure 53 montre les corrélations entre les courbures de la densité électronique polarisée et les énergies électrostatiques de polarisation seulement. Il existe définitivement un lien entre ces courbures et l'énergie de polarisation, elles sont cependant inférieures à celles obtenues pour E_{ELEC}+E_{POL}. Le fait que la corrélation avec E_{POL} du panneau (a) (λ_1) soit légèrement meilleure que celle du panneau (b) ($|\lambda_2+\lambda_3|$) est cohérent avec l'hypothèse que la polarisation mène une partie des électrons à quitter l'espace entre les atomes impliqués [baisse de $\rho_{BCP}(\mathbf{r})$]. Ces corrélations sont extrêmement

intéressantes pour qualifier l'effet de l'induction dipolaire sur la topologie de la densité électronique et méritent une étude plus systématique.

L'ellipticité d'une liaison covalente, exprimée comme (λ_3/λ_2)-1, est liée à la symétrie de la densité électronique autour du point critique. Ce paramètre topologique a été laissé de côté pour notre interprétation (les valeurs ont été omises pour la clarté des Tables 12 à 14). Pour des liaisons covalentes cette symétrie renseigne sur l'ordre de liaison : (λ_3/λ_2)-1 sera proche de 0 dans le cas d'une liaison simple ou triple, mais déviera dans le cas d'une liaison double (bien entendu des cas intermédiaires doivent être attendus). Les liaisons hydrogène traitées ont une ellipticité (λ_3/λ_2)-1 très faible, en moyenne 0.093 avec un écart-type de 0.095. Leur symétrie est sphérique à 10% près, ce qui est compréhensible quand on prend en compte qu'une liaison hydrogène implique obligatoirement une orbitale *s* pour l'atome H. L'interprétation de (λ_3/λ_2)-1 est globalement peu claire pour les liaisons non-covalentes. Seules 4 liaisons hydrogène sur les 23 ont une ellipticité supérieure à la moyenne de 0.093 par plus d'un écart type : O3···Asn23 (WT et Y43F), O3···Ser27 (W79F) et O3···Tyr43 (W79F). Ces liaisons impliquent des champs électriques qui ne sont pas dans l'axe de l'interaction et une forte polarisation. La liaison hydrogène S···Thr90 est la seule pour laquelle la densité électronique devient systématiquement moins elliptique dans les directions de λ_2 et λ_3 après application de la polarisation. C'est-à-dire que les valeurs de λ_2 et λ_3 se rapprochent l'une de l'autre, et la densité se rapproche d'une symétrie cylindrique autour de l'axe H···S. Cependant, cette liaison est déjà très symétrique et les variations de λ_2 et λ_3 sont très faibles. Deux autres liaisons présentent des baisses d'ellipticité à l'application de la polarisation, mais ces baisses sont de l'ordre du pourcent.

Il a été vu que notre modèle permet d'obtenir une topologie plausible du champ de densité électronique. Il est pour l'instant difficile d'objectiver la qualité de cette topologie, bien que des tendances connues soient retrouvées. Il semble raisonnable d'affirmer que le modèle de densité électronique transférée-polarisée pourrait être exploré à l'aide d'un banc de test plus systématique. Pour le système streptavidine-biotine, la partie 3.7.d montrait que les calculs d'énergie résidus par résidus expliquaient une partie de la variabilité expérimentale, en partie grâce à l'extraordinaire rigidité du site de fixation. Il a été vu ici que ces interactions hydrogènes, capitales pour la fixation de la biotine, répondent à une taxonomie différente en termes de topologie du champ de densité

électronique. La distance d'interaction est un facteur capital à prendre en compte pour l'interprétation, mais l'étroitesse de certains contacts à l'énergie intense engendre une topologie distincte de contacts énergétiquement plus faibles. Ceci constitue un lien indirect entre topologie de la densité électronique et propriétés expérimentales d'un complexe protéine-ligand.

Plusieurs axes d'études sont possibles à l'avenir. En effet, cette partie était focalisée sur les liaisons hydrogène, afin de tenter de relier leurs paramètres topologiques à la polarisation relativement forte impliquée dans ces liaisons. Les contacts plus hydrophobes auraient pu donner une perspective différente [on s'attend par exemple à des $\rho_{BCP}(\mathbf{r})$ plus faibles pour ce type de contact H···H]. Si l'intégralité des contacts du site actif sont topologiquement décrits, une autre analyse devient possible : relier l'additivité des contributions énergétiques à la liaison de la biotine, avec la somme des λ_1 et la somme des $|\lambda_2+\lambda_3|$.

La topologie du champ de potentiel électrostatique a été évoquée par l'inspection de l'organisation des lignes de champ et cette organisation supporte les liens qui ont été tirés sur la base des différents calculs d'énergie d'interaction électrostatique et d'énergie de polarisation. Ceci représente une incitation claire à continuer d'étudier des interactions protéine-ligand à l'aide de ce paradigme.

Chapitre 3-bis :

Améliorations Contextuelles

de la

Base de Données ELMAM2

Avertissement :

Pour une première lecture, l'auteur conseille fortement de sauter ce chapitre pour aller directement en page 253. En effet, ce chapitre a été pensé pour servir de référence pour de futurs travaux sur la base de données ELMAM2. Il s'agit donc d'un catalogue répétitif et aride de toutes les modifications apportées à la base de données depuis sa dernière publication dans une revue à comité de lecture[61]. En plus de sa valeur pour le suivi des modifications, le présent chapitre regroupe de nombreux exemples de l'application de la philosophie de construction des atomes ELMAM à des cas pratiques. Le temps investi, la volonté d'exhaustivité et la potentielle utilité – pour de futurs utilisateurs – de ce travail justifient sa place ici. Cependant, les éléments méthodologiques et applications scientifiques qui ont guidé ces modifications sont tous détaillés, respectivement, au chapitre 2 et au chapitre 3, pour des raisons évidentes de clarté.

Rappel des nouvelles règles de lecture :

Règle 1 : *dir1* et *dir2*, les deux vecteurs relatifs servant à la construction de la base locale, pointent vers des atomes voisins au 1er degré. Suivant Domagała et Jelsch, 2008[102], ces deux vecteurs doivent être choisis de manière à maximiser la symétrie locale de l'atome multipolaire transféré. Cette évaluation passe uniquement par l'élément chimique, ce qui entraine des indéfinitions d'orientation quand *dir1* et *dir2* peuvent pointer vers différentes combinaisons de voisins. Pour réparer cette indéfinition, on place *dir1* toujours avant *dir2* dans la chaîne atomique. Ce qui permet de vérifier dans l'ordre de la chaîne atomique la présence de *dir1* puis de *dir2*. *MoProViewer* construit des listes de voisinage pour représenter la connexion des atomes dans une molécule. Elles ne sont pas exploitables ici car l'ordre des connexions diffère entre les fichiers. Certaines chaînes ont dû être modifiées pour faire apparaître des différences entre *dir1* et *dir2* (exemple : C408). Si un type est reconnu pendant le transfert, mais que son système d'axes ne correspond pas à la définition ELMAM2, *dir1* et *dir2* sont mises à jour. Avec la présente implémentation, un seul passage de l'outil de réparation permet, pour tous les systèmes considérés ici (protéines-ligands ou petites molécules), d'obtenir des axes corrects. Le transfert sur la structure de l'arginase 1HQG (5023 atomes, voir partie 4.1) a laissé 4 atomes à transférer manuellement (les deux ions Mn^{2+}, modélisés par des monopôles, et l'ion OH^-, pour l'instant absent de la base). Pour les atomes à un seul voisin, ce voisin est toujours pointé par *dir1* et *dir2* est cherché parmi les voisins au 2ème degré.

Règle 2 : Définition de la symétrie 'cy' (cylindrique autour de l'axe Z) pour réserver le '1' pour l'identité, appliqué à tous les hydrogènes. Les opérateurs de symétrie des atomes ELMAM2 ont été implémentés sous forme de matrices de rotation permettant de retrouver les positions équivalentes d'un point symétrique par simple produit matriciel. Cette solution élégante n'est pas adaptée à la symétrie 'cy' : une symétrie cylindrique implique un nombre infini de positions équivalentes (correspondant à un groupe limite de Curie). Pour moyenner des tenseurs de polarisabilité, la symétrie 'cy' est approximée par 'mxmy' (voir partie 3.5 pour la justification de cette approximation, dans ce cas précis).

Règle 3 : L'utilisation de « jokers » (*wildcards*) dans une chaîne atomique était déjà possible avec les versions précédentes de la base de données. Afin de contrôler plus finement le transfert de paramètres, désormais un caractère en majuscule correspond à

n'importe quel nombre d'atomes appartenant vérifiant le symbole « joker ». Les atomes utilisant des « jokers » majuscule peuvent avoir n'importe quel nombre de voisins. Inversement, un caractère minuscule correspond à un seul atome par symbole « joker », ce qui implique que le nombre de voisins de l'atome est fixé. La ligne '`WILDCARD`' définit, type par type, quels éléments vérifient les caractères « jokers ».

*Exemple 1 : Le type atomique H113 (ammonium) répond à la chaîne atomique 'H**n**[chx]'. La ligne 'WILDCARD' fait correspondre x à C ou H. L'azote souligné aura ici exactement 4 voisins dont : un carbone, un hydrogène et un « joker ». Cet hydrogène correspond donc à une amine protonée primaire ou secondaire.*

*Exemple 2 : Le type atomique C311 (carbone cétone) répond à la chaine atomique 'Co**c**[X]**c**[X]'. La ligne 'WILDCARD' fait correspondre X à C, H ou O. Les deux carbones soulignés peuvent avoir n'importe quel nombre de voisins C, H ou O, mais aucun autre élément chimique n'est autorisé.*

Règle 4 : Évaluation des cycles : Auparavant, les attributs '`0`' et '-' avaient la même signification. L'écriture d'un algorithme de détection de cycle a permis de différencier ces deux symboles. Désormais, les atomes des molécules linéaires portent l'attribut '`0`'. La recherche de cycle s'arrête au 7ème voisin (le cyclo-octane serait par exemple considéré linéaire par le programme). Le symbole '-' (indifférent) entraine la non-évaluation du caractère cyclique. Sinon le nombre indiqué indique le nombre d'atomes dans le cycle, à l'exception de la valeur spéciale '56' qui signifie que l'atome est membre d'un cycle à 5 ou à 6 membres.

Règle 5 (tacite) : En général, placer les types atomiques les plus spécifiques en amont des moins spécifiques, quand cela permet de lever une ambiguïté sur certains atomes. Le premier type atomique valide du fichier est retenu en cas d'ambiguïté. Pour cette raison, les modifications sont présentées dans l'ordre où elles apparaissent dans le fichier ELMAM2.txt (et pELMAM3.txt), d'où certaines discontinuités dans la numérotation des types atomiques.

Ajout des éléments des tenseurs de polarisabilité :

Tous les types atomiques ont reçu 2 nouvelles lignes, 'ALPHA', contenant les 6 valeurs indépendantes de la polarisabilité atomique (*xx*, *yy*, *zz*, *xy*, *xz* et enfin *yz*). Ce tenseur est toujours exprimé dans le système d'axes locaux. La ligne suivante, 'ALPHA_ESD', contient la déviation standard de chaque élément tensoriel, calculée par rapport aux éléments indépendants des polarisabilités moyennées, exprimées dans le même référentiel. Si des contraintes de symétrie sont appliquées à la moyenne, celles-ci ne sont pas prises en compte pour le calcul de cette déviation (afin d'avoir des déviations standard comparables entre les moyennes symétrisées ou non). Les moyennes tensorielles de polarisabilité sont toujours calculées dans le référentiel local du type atomique.

Types atomiques modifiés ou supprimés en accord aux nouvelles règles :

Tous les hydrogènes ont reçu la symétrie 'cy' à la place de '1'.

H106 est l'hydrogène lié à un carbone aminé sp^3, il est utilisé pour le Cα des acides aminés. Ce type atomique a eu sa chaîne atomique modifiée. Hc[nxh] devient Hc[nhx] car la nouvelle implémentation peut, dans certains cas, manquer de reconnaître un atome si les caractères « joker » sont évalués avant les caractères normaux, plus stricts.

H113, hydrogène ammonium, est nécessaire au transfert de la lysine et des acides aminés N-terminaux. La chaîne atomique passe de Hn[chh] à Hn[chx], ou x = c|h. Cette modification accommode ce type atomique pour les hydrogènes ammonium d'une proline N-terminale ou isolée.

C303, carbone aromatique substitué par 3 carbones, était présent aux côtés d'un type d'atome non-différenciable, C304. En effet leurs chaînes atomiques originales sont identiques ('Cc[xx]c[yy]c[ch]', où x est S, O ou C et y est O, N, C ou H). Puisque C304 a été moyenné à partir d'un plus grand nombre d'atomes que C303 (14 contre 11). Ce dernier a été désactivé au profit de C304.

C304 est également un carbone aromatique substitué par 3 carbones. Il est retrouvé au niveau du Cδ2 du tryptophane dans les protéines. Ce type a été mis à jour pour maximiser la symétrie locale à l'aide d'une chaîne atomique bien plus spécifique : 'Cc[nx]c[xx]c[ch]' où x signifie C, O, N ou S. Cette chaîne exclut le Cε3 du tryptophane de la

définition du système d'axes. Comme présenté Figure 54, *dir1* et *dir2* sont alors dans une position de symétrie maximale au sein du noyau indole du tryptophane (l'expansion a deux plans miroir, '*mymz*'). Cette spécification de la chaîne atomique empêche la reconnaissance d'un des carbones de la benzamidine (voir C304T). Son attribut CYCLE passe de '6' à '56' pour vérifier plus rapidement la géométrie du noyau indole.

C305, le carbone Cγ de Trp, Phe et Tyr avait une chaîne atomique non-spécifique pour l'axe *x*. C305 possède un dipôle significatif en *x* (sa symétrie est *mymz*). Pour accommoder Phe, Trp et Tyr en suivant la nouvelle règle n°1, la chaîne atomique Cc[X]c[X]c[X] devient Cc[xx]c[xx]c[yyy]. x correspond à N, C ou H. X et y correspondent tous les deux à O, N, C ou H. La nouvelle chaîne interdit d'utiliser Cβ comme *dir1* ou *dir2*. L'attribut cycle passe de '6' à '56' pour accommoder le Cγ de Trp. L'orientation de ce carbone au sein du noyau indole est présentée Figure 54.

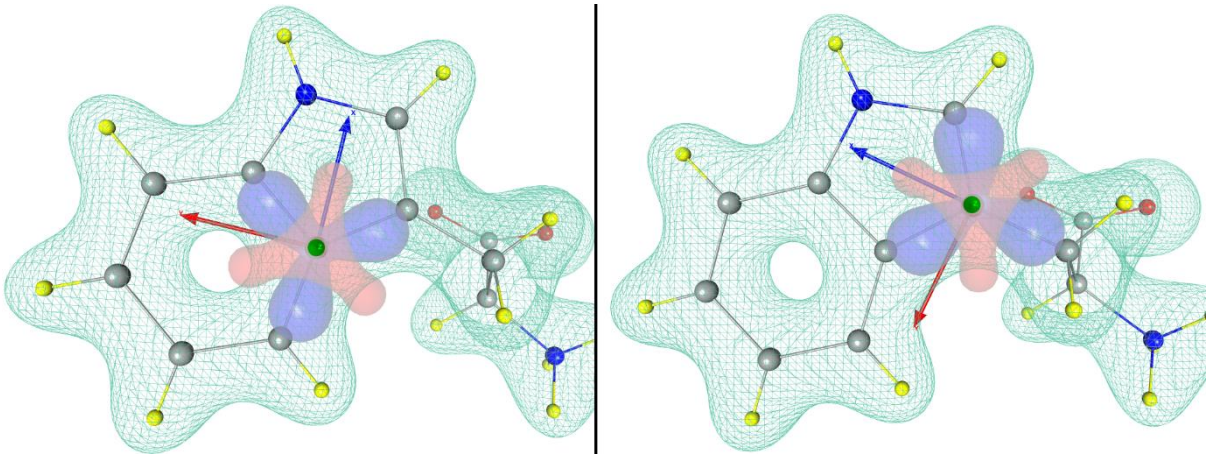


Figure 54. Orientation des carbones Cδ2 (C304) et Cγ (C305) dans le tryptophane. Les trois vecteurs figurent le référentiel atomique local (bleu : *X*, rouge : *Y*, vert : *Z*). Les surfaces bleues et rouges sont des isosurfaces à ±0.04 e·Å^{-3} de la densité électronique de déformation du carbone transféré. Le maillage cyan sombre est l'isosurface à 0.4 e·Å^{-3} de la densité moléculaire transférée totale.

C306, le Cδ2 de l'histidine, est voisin de N, C et H. Son cycle, auparavant non-évalué, passe de '6' à '56' pour continuer à être transféré avec la nouvelle règle n°4.

C308T, le Cε de l'histidine, possède une chaîne atomique non spécifique alors que la plupart des histidines sont protonées en Nδ ou Nε. L'expansion multipolaire documente des dipôles positifs en *x* et en *y* (symétrie *mz*, système *bXY*) ce qui suppose une relocalisation des électrons vers l'azote déprotoné. La chaîne atomique passe donc de Cnnh à Cn[xx]nh où x=c|n|o|h. Cette modification assure que le N trivalent est associé à *dir1*. Dans cette orientation, le dipôle atomique de Cε (0.19 D) est orienté loin de *dir2*, c'est-à-dire l'azote divalent (Nδ) car on s'attend à ce que l'azote divalent soit plus riche en électrons que l'azote trivalent (Nε), voir Figure 55. Le transfert sur un ion histidinium pourrait aboutir à une orientation aléatoire et pourrait nécessiter une réévaluation de la densité multipolaire. Un seul ion histidinium a été utilisé dans ce travail, au sein de la trypsine 1TIO (voir partie 4.1). Le transfert n'a pas montré ici d'anomalie d'orientation.

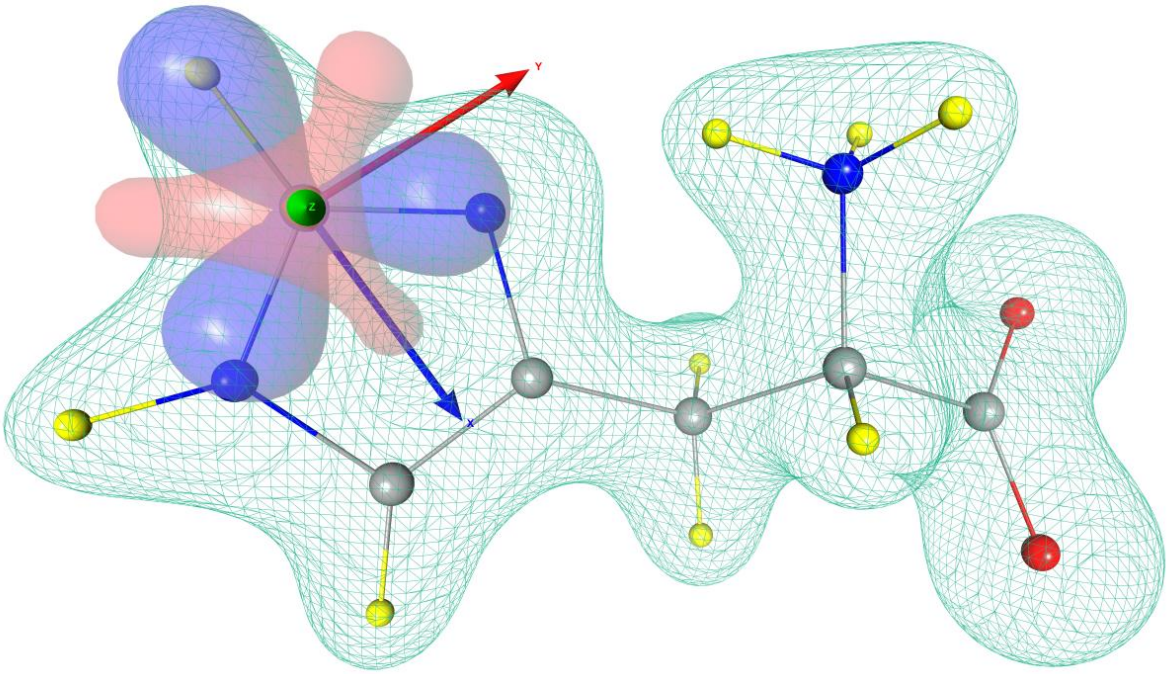


Figure 55. Orientation transférée du Cε (C308T) de l'histidine (protonée en Nε). Les trois vecteurs figurent le référentiel atomique local (bleu : *X*, rouge : *Y*, vert : *Z*). Les surfaces bleues et rouges sont des isosurfaces à ±0.04 e·Å^{-3} de la densité déformation du carbone transféré. Le maillage cyan sombre est l'isosurface à 0.4 e·Å^{-3} de la densité moléculaire transférée totale.

C312, le carbone phénolique a été déplacé avant C311 et Ckes (voir plus bas, « *types atomiques ajoutés depuis 2012* ») en vertu de la règle n°5. Sa chaîne atomique ne peut pas valider un carbone cétonique ou aldéhyde, mais les chaînes de C311 et Ckes peuvent accommoder un carbone phénolique (C311 et Ckes sont tous deux des carbones cétone ou aldéhyde). Afin d'éviter de comparer des atomes différents, placer C312 en amont (définition plus spécifique) permet d'obtenir une reconnaissance correcte sans modification des définitions (le premier type atomique valide du fichier est retenu en cas d'ambiguïté, en vertu de la nouvelle règle n°5).

C314, le carbone amide est retrouvé dans le squelette peptidique des protéines mais aussi dans l'uracile (aromatique). Son attribut cycle est passé de '0' à '-' de manière à être compatible avec le carbone amide de l'uracile (nécessaire au traitement du S66, voir partie 1.7.f).

C315, est le carbone uréide retrouvé dans la biotine et l'uracile. Le type original appartient à un cycle à 6 membres. Dans la biotine, il appartient à un cycle saturé de 5 membres. L'attribut CYCLE passe donc de '6' à '56'.

C402, le carbone méthylène. Il a été observé que, sur la base de la polarisabilité, ce carbone se comporte différemment en fonction de son appartenance à un cycle ou à une molécule linéaire (comparaison entre pentane et cyclopentane, voir partie 1.7.f). Ce type passe de cycle '-' à '0' pour exclure les molécules cycliques.

C405T, carbone tétragonal substitué par 4 autres carbones (néopentane). Sa symétrie originale est *mxmy*, alors que ses voisins suivent une symétrie *$\bar{4}$3m*. Ceci est important pour apporter des contraintes de symétrie aux moyennes de polarisabilités (voir plus bas). L'attribut SYMPLM passe donc à '-43m'.

C408, carbone alpha des acides aminés (sauf Glycine). Cet atome est un centre chiral notable. Son expansion multipolaire a beau avoir une symétrie '*mx*', inverser *dir1* et *dir2* amène une densité électronique déraisonnable du côté de l'hydrogène *α* (voir Figure 56). La densité n'est pas chirale mais peut tout de même être utilisée à l'envers (système d'axes de type *bZX*, le plan miroir est dans le plan C*α*-N-H*α*). Il n'est pas possible d'utiliser correctement ce type atomique si les deux carbones définissant ses axes locaux ne sont pas distinguables. La chaîne atomique 'C1n1c1c1h' devient 'C1n1c[xx]1c[xxx]1h', où x est

parmi C, O, H, S, N, Se[****]. Cette modification assure la correcte identification de *dir1* (C=O) et *dir2* (C*β*) par l'outil de réparation d'axes.

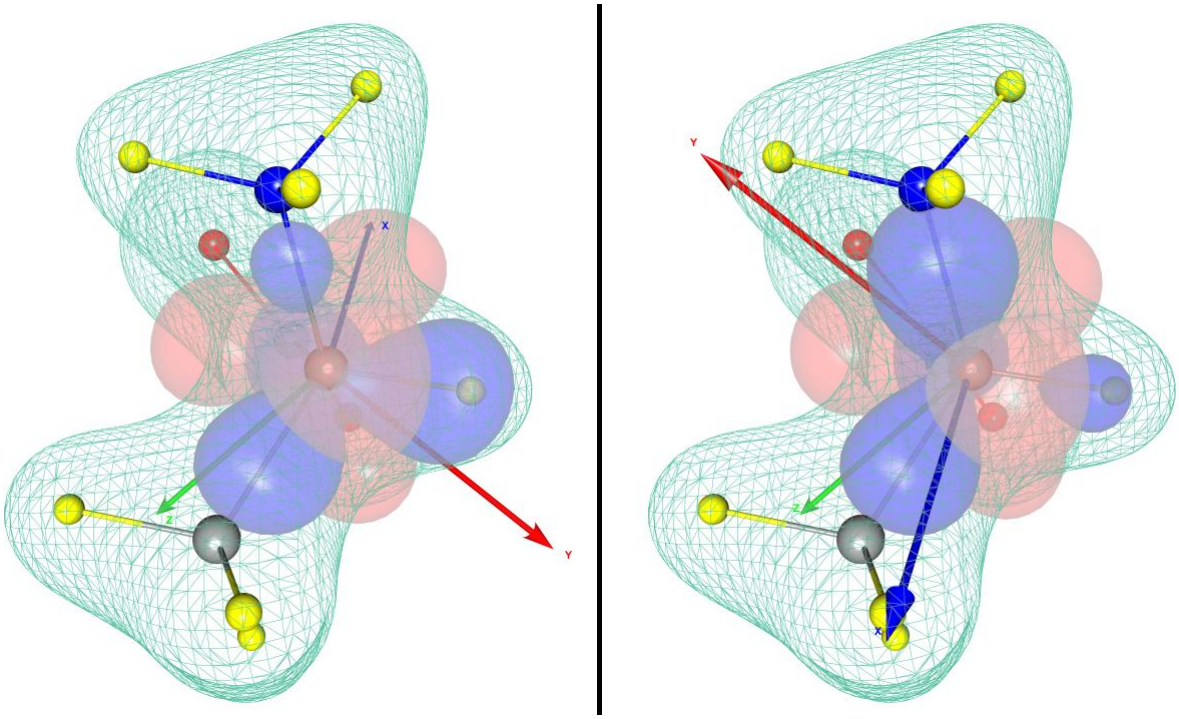


Figure 56. Orientation non-spécifique du C*α* (C408). À gauche, l'axe *X* (normal au plan miroir) est tourné vers le carboxylate, à droite, l'axe *X* est tourné vers le C*β*. On constate que l'orientation de gauche génère une densité électronique relativement trop forte au niveau de la liaison C*α*-H*α* comparé à la liaison C*α*-N. La nouvelle règle n°1 – avec la nouvelle chaîne atomique – permet de forcer l'orientation de droite. Les trois vecteurs figurent le référentiel atomique local (bleu : *X*, rouge : *Y*, vert : *Z*). Les surfaces bleues et rouge sont des isosurfaces à ±0.04 e·Å^{-3} de la densité électronique de déformation du carbone transféré. Le maillage cyan sombre est l'isosurface à 0.4 e·Å^{-3} de la densité moléculaire transférée totale.

C411, C*β* de la Thréonine, souffre de la même indéfinition que C408 (voir ci-dessus) et est examiné Figure 57. Le dipôle de cet atome doit pointer vers l'atome d'hydrogène afin de relocaliser les électrons vers les atomes plus électronégatifs. La chaîne atomique

[****] : *Il est possible de retrouver du sélénium dans les structures biologiques sous forme de sélénométhionine ou de sélénocystéine. La sélénométhionine est incorporée dans les protéines à la place de la méthionine lors d'une production sur un milieu pauvre en soufre et est utilisée pour le phasage expérimental des biocristaux (méthode SAD et MAD). La sélénocystéine est naturellement présente dans le site actif de certaines oxydoréductases.*

‘C1o1c1c1h’ devient ‘C1o1c[hhx]1c[xxx]1h’, où x signifie C, O, H ou N, ce qui contraint *dir1* à pointer vers le C*γ* et *dir2* vers C*α*.

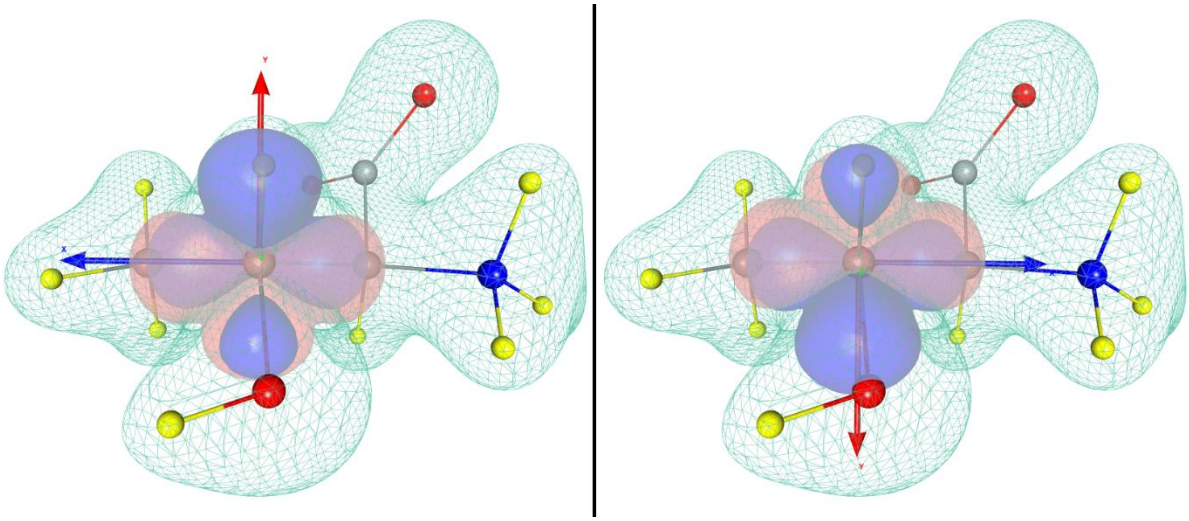


Figure 57. Orientation non-spécifique du C*β* de la Thréonine (C411). À gauche, l’axe *X* pointe vers le C*γ*, à droite, vers le C*α* (système *bZX*). La densité de la liaison C*β*-H doit être plus faible que celle de la liaison C*β*-O. La nouvelle règle n°1 – avec la nouvelle chaîne atomique – permet de forcer l’orientation de droite. Les trois vecteurs figurent le référentiel atomique local (bleu : X, rouge : Y, vert : Z). Les surfaces bleues et rouges sont des isosurfaces à ±0.04 e·Å^{-3} de la densité électronique de déformation du carbone transféré. Le maillage cyan sombre est l’isosurface à 0.4 e·Å^{-3} de la densité moléculaire transférée totale.

C414, C*β* de la cystéine, porte originellement un attribut CYCLE à ‘0’, ce qui correspond à une molécule linéaire dans les nouvelles règles présentées ici. Un atome très similaire est présent dans l’un des cycles de la biotine, l’un de nos systèmes modèles. L’attribut cycle de ce type passe de ‘0’ à ‘-’ (non-évalué) pour correspondre à la fois au C*β* (linéaire) de la cystéine et au carbone cyclique de la biotine.

O102, oxygène amide de chaîne atomique ‘O1.5c[1.5n(cx)1c(X)]’. x représente C ou H, X représente O, N, C ou H. Afin de retirer un doublon dans les atomes supplémentaires (nommé OCOX, voir ci-dessous), ‘S’ a été ajouté comme signification possible de X.

OCOX, oxygène carboxylate, est un duplicata de O104 qui contraint le carbone portant le groupement –CO_2^- à être d’hybridation sp^2. La position finale de ce type

atomique (en aval de O104G) rend impossible son utilisation en pratique (règle n°5). OCOX a été retiré *a priori* au profit de O104G.

Types atomiques ajoutés pour le présent travail :

H111T est l'hydrogène ammonium. C'est un simple duplicata de l'hydrogène amine (H111) dont la chaîne atomique a été adaptée pour l'ion NH_4^+ ('Hn[hhh]').

H117 est l'hydrogène de l'éthène C_2H_4. Sa chaîne atomique est 'Hc[c(hh)h]'. En l'absence de densité de charge expérimentale pour ce composé, les paramètres multipolaires de ce type ont été copiés à partir de H101, l'hydrogène aromatique.

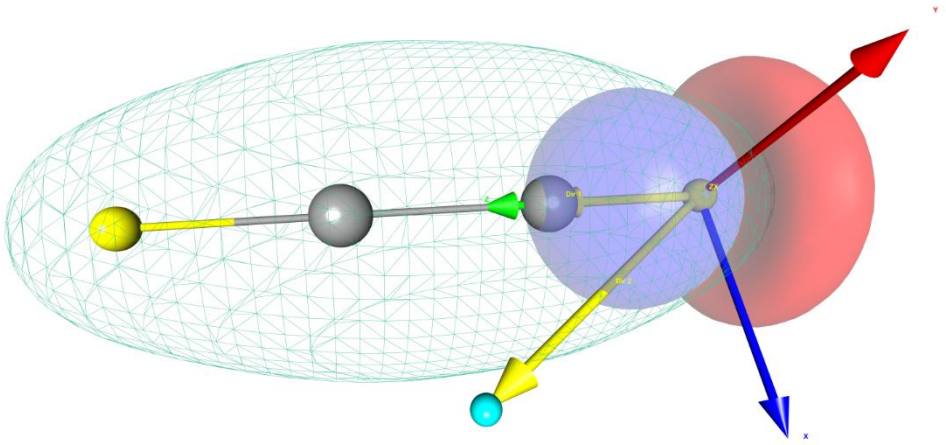


Figure 58. Définition de l'orientation de l'hydrogène éthyne (H118) par un atome hors de l'axe de la molécule. Les vecteurs jaunes représentent les directions de définition permettant de construire le référentiel atomique local (système *ZX*). Ce référentiel local est représenté par les trois autres vecteurs (bleu : *X*, rouge : *Y*, vert : *Z*). Les surfaces bleues et rouges sont des isosurfaces à ±0.04 e·Å^{-3} de la densité électronique de déformation de l'hydrogène transféré. Le maillage cyan sombre est l'isosurface à 0.4 e·Å^{-3} de la densité moléculaire transférée totale. L'orientation des carbones de l'éthyne (type C209, ci-dessous) utilise la même convention.

H118 est l'hydrogène de l'éthyne C_2H_2, représenté Figure 58. Sa chaîne atomique est 'Hc[c(h)]'. Il a été obtenu à partir de calculs théoriques[82] utilisant le logiciel CRYSTAL-09[83]. Sa symétrie est de type 'cy'. H118 est issu d'une molécule linéaire. Il utilise donc un « joker » pour la définition des directions principales, comme les ions monoatomiques de ELMAM2. Ici, seule *dir2* est un « joker » qui doit être de préférence loin de l'axe de

symétrie de l'éthyne pour une orientation correcte de cette densité multipolaire de symétrie cylindrique (pour éviter une imprécision numérique lors de la construction des axes locaux).

H105T, est l'hydrogène chiral de la biotine (dupliqué de H105). Sa chaîne atomique devient Hc[scc] à la place de Hc[ccc]. Cette extension a été réalisée à partir du type atomique le plus proche en termes de chaîne atomique.

C304T, carbone aromatique retrouvé dans la benzamidine, il lie le cycle au groupement C(NH2)2. Ce type a été dupliqué à partir de C304 et est représenté Figure 59. En effet, avant de contraindre l'orientation de C304 (voir plus haut), ce type d'atome pouvait servir à la fois pour la benzamidine et le tryptophane. Il a été choisi a priori de distinguer ces deux types. Ici, la position de symétrie maximale est différente comparé au noyau indole. C304T utilise la chaîne 'Cc[ch]c[ch]c[xx]' pour décrire ses voisins, ce qui contraint *dir1* et *dir2* à pointer vers le cycle aromatique. Ici, x signifie N ou O, ce qui permet le traitement de la molécule « F2_25 ». C304T n'accommode que les cycles à 6 membres.

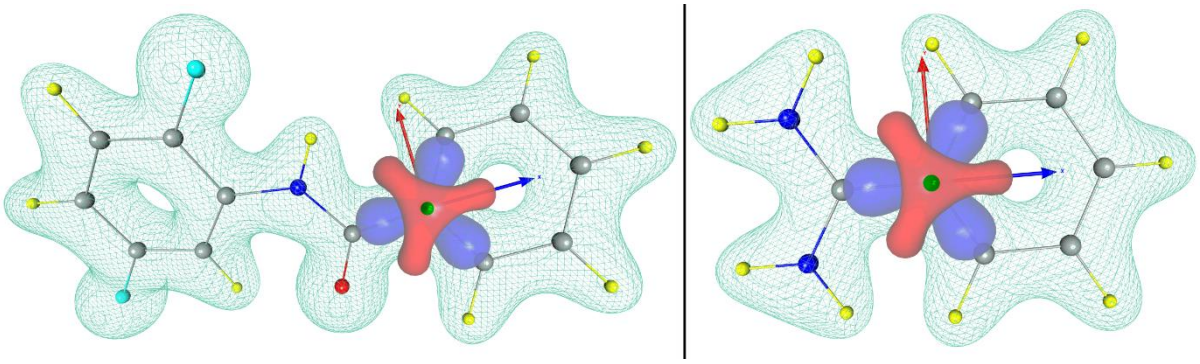


Figure 59. Atomes de type C304T. À gauche, la molécule « F2_25 » [*N*-(2,5-fluorophényl)-benzamide] soulignant l'atome C11. À droite, la molécule de benzamidine soulignant l'atome C1. Les trois vecteurs figurent le référentiel atomique local (bleu : *X*, rouge : *Y*, vert : *Z*). Les surfaces bleues et rouges sont des isosurfaces à ±0.04 e·Å^{-3} de la densité électronique de déformation du carbone transféré. Le maillage cyan sombre est l'isosurface à 0.4 e·Å^{-3} de la densité moléculaire transférée totale.

C315T, retrouvé dans l'urée (chaîne atomique : Conn). L'évaluation de l'appartenance à un cycle empêche C315, le carbone uréide, cyclique, d'être applicable à l'urée (voir plus bas). Le type C315 a donc été dupliqué, mais C315T n'accommode que les

atomes non-cycliques (cycle '0'). C315T est absent de notre jeu de polarisabilités théoriques. Sa polarisabilité est une copie de celle de C315.

C318 est carbone sp^2 de l'éthène. Sa chaîne atomique est 'Cc[hh]hh'. Cet atome est un duplicata de C301, mais dont le dipôle a été réorienté le long de la double liaison C=C.

C209 est le carbone sp de l'éthyne. Il a été obtenu à partir du même calcul théorique que H118[82]. C209 est le seul carbone de symétrie 'cy' dans la librairie. Comme pour H118, *dir2* est un « joker » qui doit être de préférence loin de l'axe de symétrie de l'éthyne pour une orientation correcte de cette densité multipolaire, voir Figure 58.

C402T est le carbone méthylène cyclique. Dupliqué de C402, il en possède la densité électronique. Il est cependant contraint d'appartenir à un cycle à 5 membres (cycle '5'). L'extension aux cycles à 6 membres est discutable dans le futur, en fonction des molécules qui seront étudiées. Cette possibilité d'extension peut être pertinente pour discuter la transférabilité des carbones entre hexoses (par exemple glucose) et pentoses (par exemple ribose).

C408T, dupliqué de C*α* (C408), correspond aux deux atomes de carbone de la biotine (C4 et C5) qui joignent les deux cycles à 5 membres de cette molécule, voir Figure 60. C408 a été choisi comme type atomique à copier car il présente des voisins au premier degré qui ont les mêmes types chimiques. L'attribut CYCLE passe de '0' à '56'. La chaîne atomique est 'C1n1c[sxx]1c[xxx]1h'. Ici, x peut signifier H, C ou N seulement. Cette chaîne définit un voisinage plus spécifique de cet atome pour éviter une reconnaissance erronée dans un acide aminé protéinogène. Ceci permet que le moment dipolaire de ce type atomique pointe vers H (système d'axes *bZX*, dont les deux directions principales pointe vers les carbones voisins). Les deux carbones concernés de la biotine sont C4 et C5. Ces deux atomes sont de chiralité opposée, indifférenciés par ailleurs. Inverser manuellement *dir1* et *dir2* pour C5 permet d'obtenir une meilleure orientation de la densité de C5. Des ébauches de code pour la détection automatique de centres chiraux ont été testés mais n'ont pas été terminés. Cette expansion multipolaire n'est pas chirale (symétrie '*mx*'), comme de nombreuses autres attribuées à des atomes chiraux.

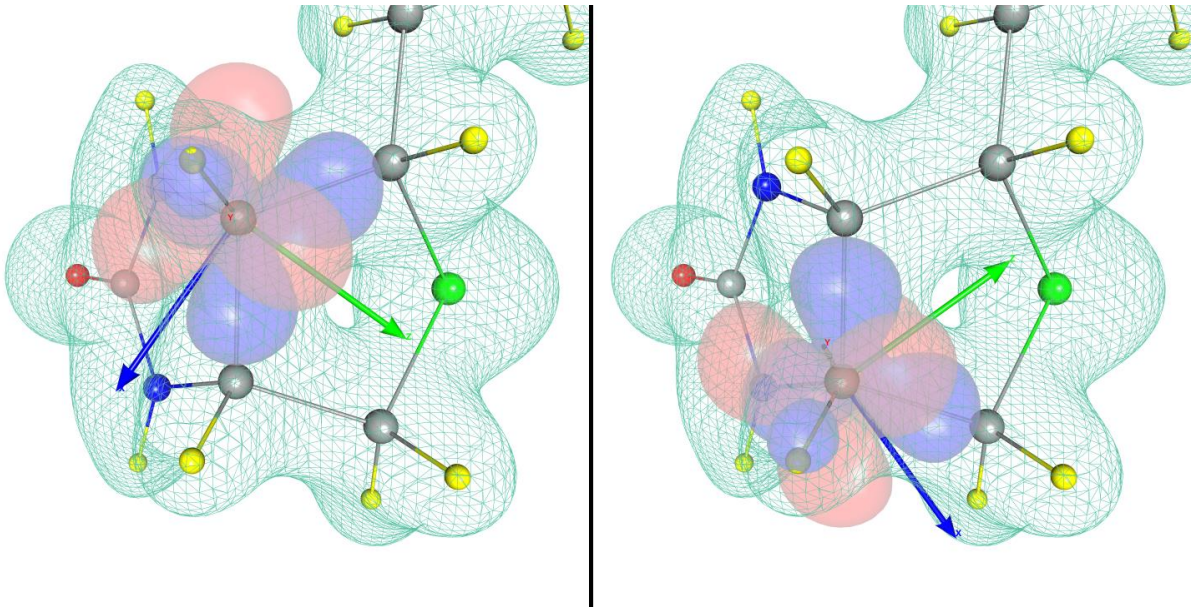


Figure 60. Le carbone C4 (C408T) de la biotine (à gauche, en configuration *S*) et le carbone C5 de la biotine (à droite, en configuration *R*) ont une orientation définie par les carbones voisins. L'orientation transférée est celle de C4, celle de C5 a été manuellement corrigée en inversant le rôle de *dir1* et *dir2* afin d'obtenir une densité correcte au niveau de la liaison C-H. Les trois vecteurs figurent le référentiel atomique local (bleu : X, rouge : Y, vert : Z). Les surfaces bleues et rouges sont des isosurfaces à ±0.04 e·Å^{-3} de la densité électronique de déformation du carbone transféré. Le maillage cyan sombre est l'isosurface à 0.4 e·Å^{-3} de la densité moléculaire transférée totale.

Cohhh, le carbone méthoxy-, a été dupliqué à partir du type le plus proche, C410, un carbone lié à O (éther ou alcool), C et deux H. Sa chaîne atomique est C1o1h1h1h. Son ajout a été motivé par la présence du méthanol et du diméthyl-éther dans notre jeu de structures.

C415, permet de transférer le dernier carbone manquant pour la biotine, le carbone C2 reliant la partie cyclique à la partie aliphatique de cette vitamine. Il est dupliqué de C414, sa chaîne atomique est 'Cc[chh]sc[nch]h', cet ordre d'écriture place le premier axe (le soufre) avant le second axe (l'hydrogène). Le carbone concerné est le carbone chiral C2 de la biotine. La biotine d'origine biologique est *S* pour ce centre chiral mais certaines structures cristallographiques (par exemple, 2F01) montrent la fixation d'un mélange racémique. Cette orientation maximise la symétrie de la densité électronique (voir Figure 61) et devrait permettre le transfert sur une *R*-biotine.

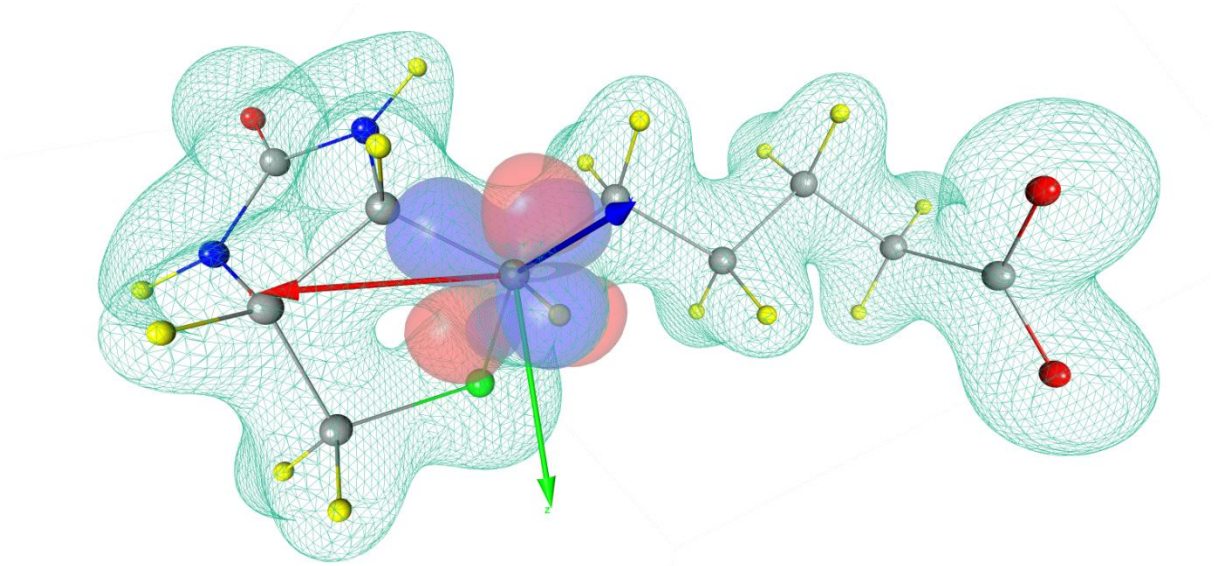


Figure 61. Carbone C2 de la biotine (C415, configuration *S*, l'hydrogène lié à C2 est au premier plan par rapport au soufre). La densité électronique du carbone C2 contribue peu à la densité de la liaison C-S. Les trois vecteurs figurent le référentiel atomique local (bleu : X, rouge : Y, vert : Z). Les surfaces bleues et rouges sont des isosurfaces à ±0.04 e·Å^{-3} de la densité électronique de déformation du carbone transféré. Le maillage cyan sombre est l'isosurface à 0.4 e·Å^{-3} de la densité moléculaire transférée totale.

N306, amine neutre hybridée sp^3 de géométrie tétragonale, est présentée Figure 62. Cet azote existe notamment pour le traitement de l'arginine (son amine α est déprotonée dans les structures théoriques dans le vide). Son expansion multipolaire a été copié de l'ammonium N401 (symétrie et système d'axes compris). Les premiers voisins de cette amine sont identiques à N301, l'amine NH2 plane, qui représente un atome d'hybridation sp^2. Pour différencier l'amine sp^3 (N306) et l'amine sp^2 (N301), la chaîne atomique de N306 est 'Nc[hxx]hh', ce qui contraint le carbone qui porte l'amine à être hybridé sp^3. Elle est placée en amont de N301 afin d'être retenue en cas d'ambiguïté. Ce choix est justifié par le fait que notre jeu de données ne comporte pas d'amine sp^3 déprotonée lié à un carbone sp^2 (exemple : aniline sous forme neutre), qui nécessiterait alors un nouveau type atomique.

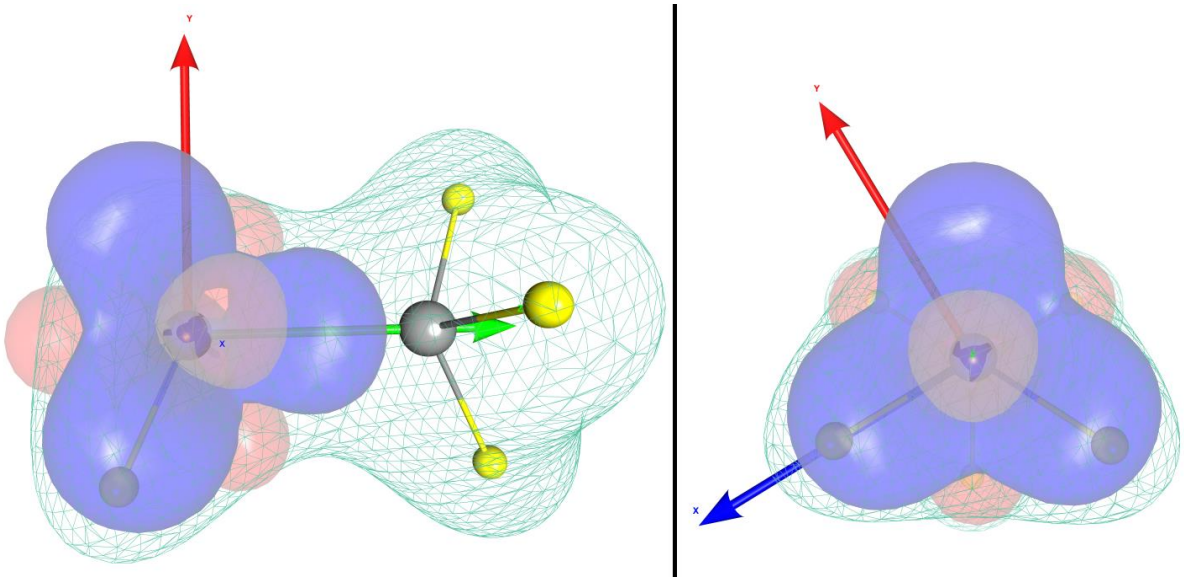


Figure 62. Densité reconstruite pour l'amine neutre sp^3 (N306), vue le long des directions locales *X* et *Z*. Ce pseudoatome possède une symétrie *'3m'* (comme pour l'ammonium N401, les 3 plans miroir se coupent sur l'axe *Z*). Cette symétrie permet d'utiliser indifféremment chacun des deux hydrogènes pour la définition de l'axe *X* (cette symétrie devient invisible à l'inspection de la densité totale). Les trois vecteurs figurent le référentiel atomique (bleu : *X*, rouge : *Y*, vert : *Z*). Les surfaces bleues et rouges sont des isosurfaces à ±0.05 $e \cdot Å^{-3}$ de la densité électronique de déformation de l'azote transféré. Le maillage cyan sombre est l'isosurface à 0.4 $e \cdot Å^{-3}$ de la densité moléculaire transférée totale.

N302T est un duplicata de N302. Ces deux types atomiques décrivent un azote sp^2 amide. N302 n'accommode que les molécules linéaires alors que N302T est spécifiques des cycles à 6 membres (exemple : uracile). La duplication de ce type a été décidée pour augmenter la transférabilité de la polarisabilité de cet atome (réduction drastique de la déviation standard sur λ_1 de la polarisabilité moyennée[52]).

N401T, est un nouveau type d'atome ajouté pour traiter l'azote des prolines isolées ou N-terminales. Il est basé sur N401 mais plusieurs différences sont présentes. N401T est contraint d'appartenir à un cycle à 5 membres. Sa chaîne atomique est 'Ncchh'. Son système d'axes est *bYZ*, défini à partir des deux carbones voisins. Sa symétrie passe de *3m* à *mxmz*. Le cycle de la proline n'est pas plan donc ce type ne respecte pas la convention normale de ELMAM2 où les vecteurs Z sont perpendiculaires au plan du cycle. Ces importantes modifications d'orientation ont nécessité des changements appropriés de l'expansion multipolaire. Les populations décrivant le moment quadrupolaire n'ont pas

été modifiées, le nouveau système d'axe les positionnant en accord avec la nouvelle symétrie. Les paramètres dipolaires ont été tournés de 90° de manière à amener le moment dipolaire à l'opposé de l'axe Y (les électrons sont déplacés vers l'intérieur du cycle). Les populations octupolaires ont été tournées par essai-erreur jusqu'à alignement des lobes positifs de la densité sur les liaisons chimiques (dans le nouveau système d'axes, voir Figure 63).

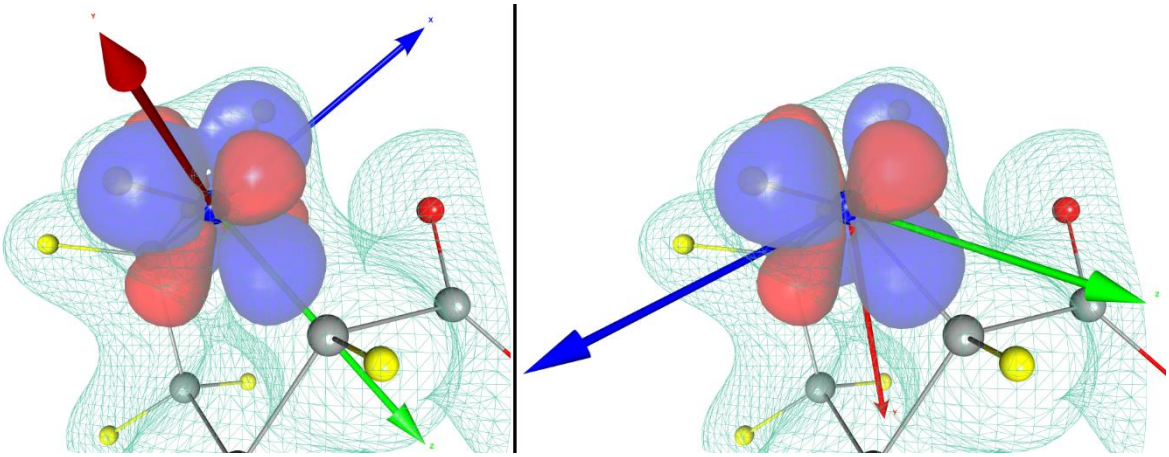


Figure 63. Comparaison de densité électronique entre l'ancienne amine primaire (N401, à gauche) et la nouvelle amine secondaire (N401T, à droite). Les trois vecteurs figurent le référentiel atomique local (bleu : *X*, rouge : *Y*, vert : *Z*). Les surfaces bleues et rouges sont des isosurfaces à ±0.05 e·Å^{-3} de la densité électronique de déformation de l'azote transféré. Le maillage cyan sombre est l'isosurface à 0.4 e·Å^{-3} de la densité moléculaire transférée totale.

N402T est un duplicata de N401. Il est destiné au transfert de l'ion ammonium présent dans certains calculs de polarisabilités. Ce type atomique diffère de N401 par sa chaîne atomique ('Nhhhh' au lieu de 'Nchhh'), sa symétrie ('-43m' au lieu de '3m') et ses populations dipolaires contraintes à 0.

O102T est un duplicata de O102. Les calculs théoriques montrent que la polarisabilité de l'oxygène amide peut avoir deux orientations différentes au sein des amides libres ou impliquées dans des liaisons peptidiques. Cette différence est dans le plan *xy* du référentiel atomique local. O102T utilise la chaîne atomique 'O1.5c[1.5n(hh)1c(X)]' afin de forcer l'azote de l'amide à porter 2 hydrogènes.

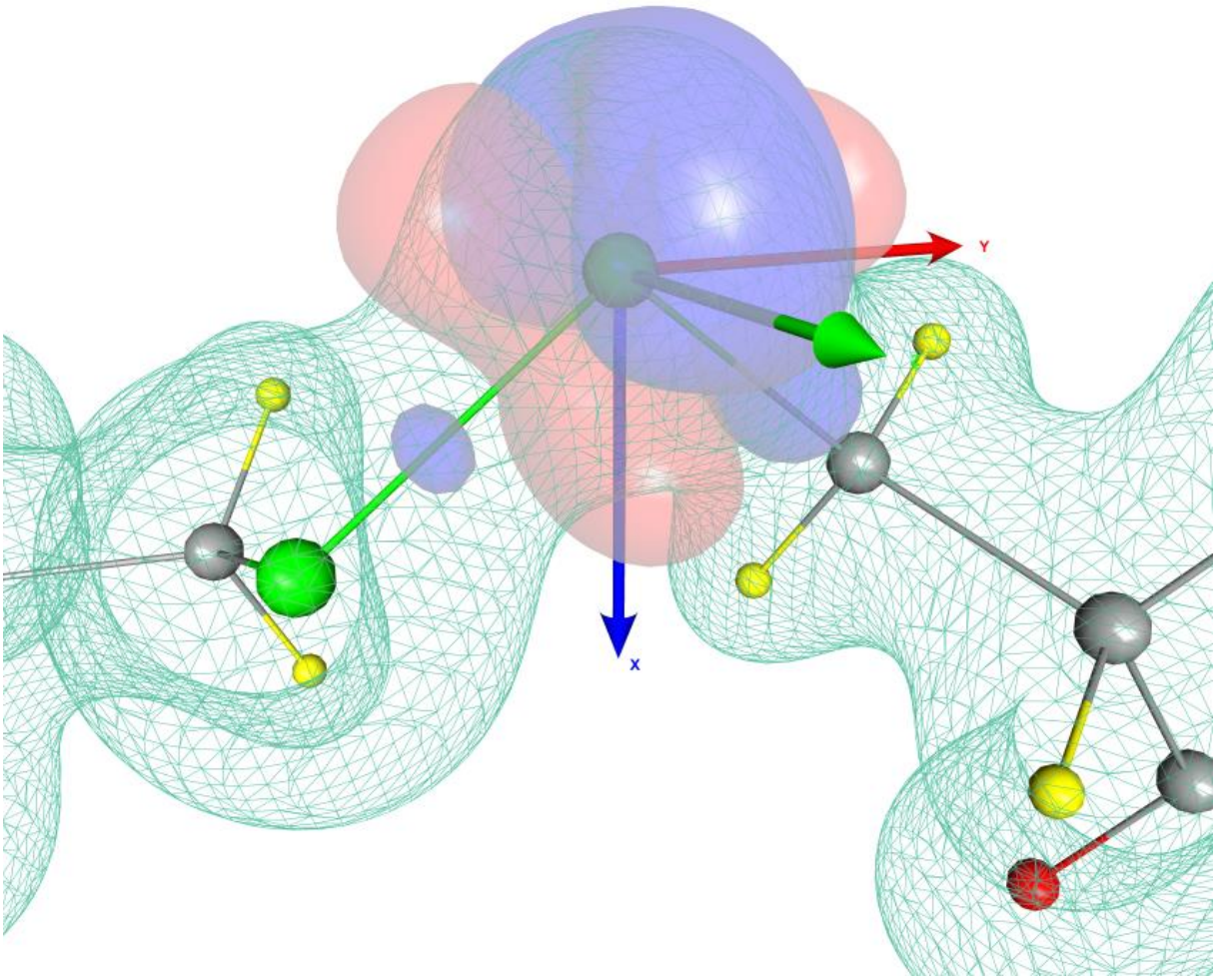


Figure 64. Nouvelle densité transférable pour les ponts disulfures réduits (Ssc), présentée sur le pont disulfure Cys25-Cys41 de la trypsine (code PDB : 1TIO). Les trois vecteurs figurent le référentiel atomique local (bleu : *X*, rouge : *Y*, vert : *Z*). Les surfaces bleues et rouges sont des isosurfaces à ±0.03 e·Å^{-3} de la densité électronique de déformation du soufre transféré. Le maillage cyan sombre est l'isosurface à 0.4 e·Å^{-3} de la densité moléculaire transférée totale.

S202T est un duplicata de S202. Ces deux types ont pour chaîne atomique 'Scc', S202 est utilisé pour le transfert du soufre de la Méthionine. S202 n'accommode que les molécules linéaires. Leur définition est identique sauf que S202T doit appartenir à un cycle à 5 membres. La mise en place de S202 a été faite *a priori* sur la base de l'inspection visuelle des calculs de polarisabilité sur la biotine : l'axe principal de la polarisabilité de ce souffre pointe vers le centre du cycle, là où la polarisabilité du souffre dans la

méthionine pointe dans l'axe de la chaine latérale. L'angle entre les premiers vecteurs propres des deux polarisabilités approche 90°, comme présenté Figure 35 (page 165).

Ssc est un type atomique inédit permettant le transfert de paramètres sur les cystines (ponts disulfures réduits) fréquemment rencontrés dans les structures protéiques (voir Figure 64). La cystine étant absente de notre jeu de polarisabilités théoriques, la polarisabilité du type atomique S201 (le plus proche) a été utilisée telle qu'elle pour le type atomique Ssc. Paramètres fournis par Christian Jelsch.

Autres types atomiques ajoutés depuis 2012 :

HnNC: dupliqué de l'hydrogène H112 porté par le N de la liaison peptidique, pour le transfert sur des diazoles (molécules contenant des liaisons N-N).

Ckes: dupliqué du carbone aldéhyde ou cétone C311. Les deux types sont équivalents en tout point à l'exception des « jokers ». C311 tolère n'importe quel nombre de C, H et O comme seconds voisins (X=c|h|o) alors que Ckes ajoute S, P, N, F et Cl comme seconds voisins possibles (X=c|h|o|s|p|n|f|cl). Utilisé pour le projet benzamides[91].

Oest est un type atomique dupliqué à partir de O106, l'oxygène ester hybridé sp^2 (1 voisin). Les deux types atomiques sont strictement identiques à l'exception de leur chaîne atomique. Celle de Oest est 'O2c[1o(*)1c(*)]', qui comporte des astérisques, non-définies dans la procédure de transfert de *MoProViewer*. Oest n'est pas reconnu dans le présent jeu de données et est le seul type atomique à utiliser cette syntaxe. Les acides aminés méthylés en C-ter utilisent O106.

Section : Atomes additionnels.

CL-, permet le transfert sur l'ion chlorure. Ce type atomique est purement sphérique et est un modèle utile pour ajouter d'autres types sphériques à la banque.

Clc est le second type chlore de la base de données[91]. Il utilise la chaîne atomique 'Clc[cc]' pour transférer sur des molécules de type chlorobenzène. Ce type porte

originellement un attribut CYCLE '0', passé en CYCLE '-' pour accommoder la règle n°4. Note : les paramètre κ et un κ' de ce type atomique sont égaux à 1.0.

Carbonate : Cooo (originellement C1) et Oc (originellement O1) sont des types atomiques dérivés de l'extension des deux types suivants (Nooo et On) permettant le transfert de l'ion carbonate. Ces deux types ont reçu des noms explicites et non-redondants.

Nitrate : Nooo (originellement N1) et On (originellement O1) sont des types atomiques fournis par Lutz permettant le transfert de l'ion nitrate. Ces deux types ont reçu des noms explicites et non-redondants. L'oxygène nitrate avait disparu de la base de données (remplacé par un duplicata de l'azote nitrate) et a été restauré à partir de l'oxygène carbonate (duplicata du type manquant).

O104G est un duplicata (généralisation) de O104 qui ne contraint pas le carbone porteur du groupement $-CO_2^-$ à être sp^3 (chaîne atomique : 'O1.5c[1.5o1c(X)]' à la place de 'O1.5c[1.5o1c(xxx)]' pour O104).

Phosphate : Les trois types P1 sont des duplicatas les uns des autres, permettant chacun le transfert de paramètre sur le phosphore de l'acide phosphorique et ses sels (H_2PO_4, HPO_4^- et PO_4^{2-}). Ces 3 types atomiques sont suivis de 2 types appelés O1 permettant le transfert des oxygènes (protonés ou non) du phosphate. Ces 2 types atomiques n'ont pas la même expansion, l'oxygène protoné a été renommé Oh1. O1B est un duplicata de l'oxygène phosphate non-protoné qui accommode les dérivés d'acide phosphonique. Le type H1 suivant décrit l'hydrogène acide du phosphate. P1bis est un duplicata du phosphate du Nicotinamide-Adénosine Dinucléotide, étendu aux acides phosphoniques.

Composés nitrilés : ils sont couverts par les types O1, N1 et Cncc. Ils permettent le transfert des 3 atomes spécifiques aux composés organiques de type nitro-. Cncc est passé en CYCLE '-' pour suivre les nouvelles règles. Note : ces 3 types atomiques ont un κ et un κ' égaux à 1.0.

Cclcc et son duplicata (chaîne atomique moins spécifique) permettent le transfert sur des composés de type chloro-. Afin de suivre les nouvelles règles, et d'être applicables aux benzamides, les deux types 'Cclcc' passent en CYCLE '-'. Note : ce type atomique a un κ et un κ' égaux à 1.0.

HnSC est un duplicata HnNC et de H112. Il est utilisé pour le transfert des molécules contenant des liaisons N-S, par exemple les antibiotiques de la classe des sulfamides.

HcSC est un duplicata de H101 à l'exception de P_{val} qui est à 0.94 e pour ce type atomique.

S1 est le souffre de l'ion sulfate généré à partir de P1 (voir phosphates). Présent à deux autres endroits dans la banque. S1 a été retiré de ELMAM2 du fait de sa redondance avec Soooo.

Snn est un duplicata de S202 qui permet de transférer un soufre substitué par deux azotes.

Deux types atomiques portent le nom Cscc. Le premier porte la chaîne atomique 'Cs[X]c[xx]c[xx]', où x est C, H ou N et X est C, H, N ou O. Le second porte la chaîne atomique non-spécifique suivante : 'Cscc'. Ces deux types sont des duplicatas de C312 (Cζ de la Tyrosine).

Csch est un duplicata de C301, le carbone aromatique (type benzène), pour les molécules soufrées.

Nsc est un duplicata de N201, utile pour transférer les sulfamides.

O=so est l'oxygène de l'ion sulfate et des différents groupements sulfonates. Il utilise la chaîne 'Os[oxx]' où x représente O, C ou N. Ce type est un duplicata de O1 et O1B (voir phosphates).

Soooo est le souffre de l'ion phosphate généré à partir de P1 (voir phosphates). Soooo est un duplicata de S1. Soooo est retrouvé dans le présent travail au sein de la structure protéique 1TIO (voir partie 4.1).

Soonc est un autre duplicata de S1 accommodant les groupements sulfamides de type C-$\underline{S}O_2$-N.

Nsch est un duplicata de l'azote peptidique N302, où un carbone est substitué par un souffre.

Chapitre 4 :

Au-delà du Modèle des Polarisabilités Indépendantes, et Autres Perspectives

Introduction

En préambule de la conclusion, ce chapitre présente diverses pistes d'amélioration dont la mise en œuvre a effectivement commencé. Elles pourront être explorées au CRM², par exemple à des fins pédagogiques, ou dans d'autres contextes de recherche. Il semble essentiel de préciser que le temps a été le facteur principal d'arrêt de ces axes, parfois pour des raisons de mise en place tardive (modèles protéiques supplémentaires), parfois à cause de questions scientifiques profondes qui n'auraient pu être explorées qu'au détriment des résultats présentés auparavant (énergie de pénétration). La poursuite de l'adaptation du modèle de Thole au modèle HCM est dans une situation intermédiaire : elle nécessite à la fois les résultats présentés, en termes de *benchmarking* et de développement logiciel, mais requiert la finalisation – et le test – d'autres outils informatiques, ébauchés dans ce chapitre.

Cette partie décrit deux possibilités d'études d'interaction protéine-ligand supplémentaires, dont la préparation a été suspendue au moment de la rédaction. Ensuite la thématique du calcul de l'énergie dite de pénétration, liée au phénomène de polarisation sera abordée. L'expression « pénétration » réfère au recouvrement de la densité électronique d'atomes multipolaires proches, et sa représentation 3D sera discutée. Diverses méthodes et outils liés au modèle d'induction de Thole suivront, concernant d'une part l'utilisation de polarisabilités anisotropes dans un modèle d'induction atténué et d'autre part l'algorithme de Gauss-Newton pour traduire des polarisabilités entre un modèle ponctuel et un modèle topologique. Enfin, la présentation d'une tentative de nouvelle représentation à visée topologique clôturera ce chapitre.

4.1 - Trypsine et arginase

Trypsine

La caractérisation électrostatique du complexe streptavidine-biotine est la seule réellement approfondie dans ces pages (partie 3.7). Le système est idéal pour diverses raisons déjà évoquées. La streptavidine n'est pas une enzyme. Elle offre donc la possibilité d'observer une interaction biologique sans mutation inactivatrice, ni blocage chimique de la réaction. On peut tout de même se demander si les sites actifs des enzymes n'ont pas

des particularités électrostatiques qui pourraient être mises en évidence à l'aide des outils développés.

Encore une fois, étudier un système classique et bien caractérisé du point de vue fonctionnel, aide fortement à l'interprétation des résultats éventuels. La trypsine est pour cela un parfait modèle. La trypsine est une protéase, son mécanisme enzymatique est basé sur une sérine nucléophile qui est activée uniquement par des interactions entre les chaines latérales de trois acides aminés du site actif, la triade catalytique, caractéristique des protéases à sérine (Figure 65). L'histidine 40 voisine arrache l'hydrogène γ de la sérine 177 grâce à son interaction avec l'aspartate 84.

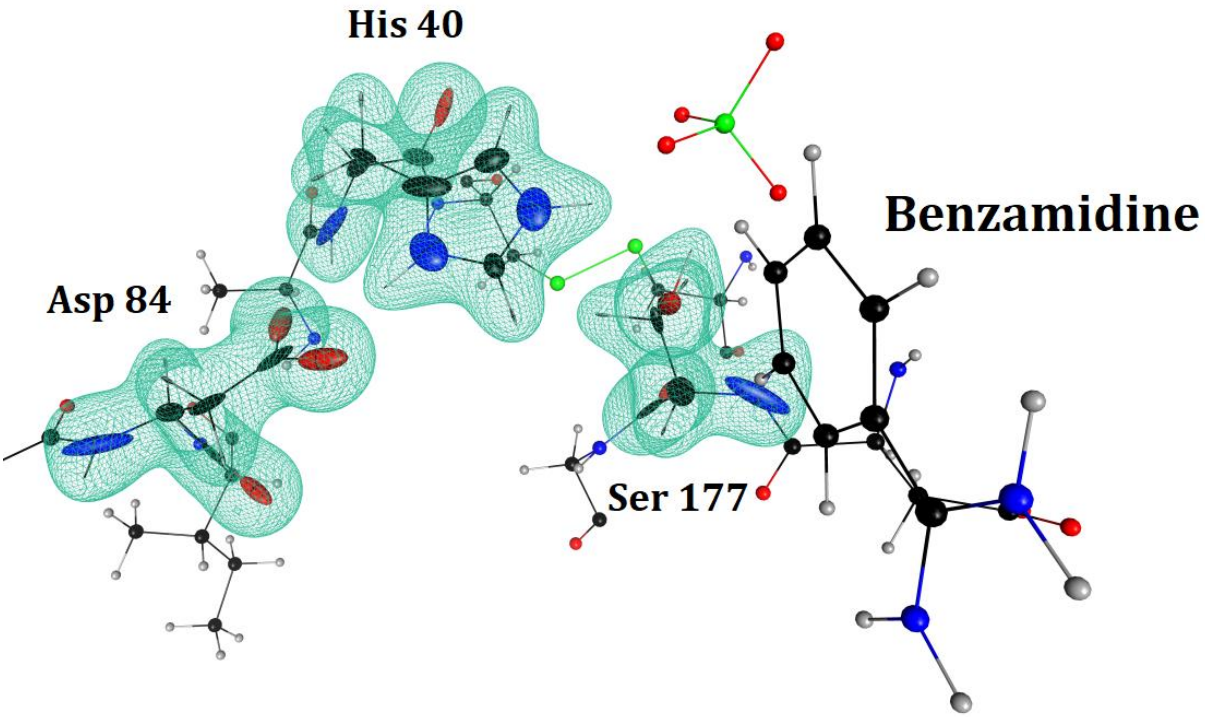


Figure 65. La triade catalytique (sérine 177, histidine 40, aspartate 84) de la structure 1TIO est inactive du fait de la forme histidinium du résidu 40. Densité électronique et polarisabilités transférées (isosurface : 0.3 e·Å^{-3} ; ellipsoïdes : 0.25 Å^3/Å)

La trypsine 1TIO est inhibée par la benzamidine. La protéine interagit également avec deux ions : un sulfate et un calcium. La benzamidine, protonée, se fixe au sein du site actif aux côtés du sulfate. Elle interagit avec l'aspartate 171, acide aminé capital pour la

reconnaissance des lysines et arginines du substrat natif de l'activité protéase. L'ion calcium a un rôle structural uniquement.

La trypsine 1TIO a été transférée et neutralisée de la même manière que la streptavidine (voir chapitre 3.7.a). Cependant, *MoProViewer* liste les liaisons de coordination qui lient le calcium parmi les liaisons covalentes. L'outil « *Bond* » du logiciel a été utilisé pour retirer ces liaisons de coordination. Cette manipulation permet la reconnaissance correcte des types atomiques ELMAM2 des glutamates 51 et 61 (au niveau des Oε) ainsi que des oxygènes amides (peptidiques) de l'asparagine 53 et de la valine 56.

La modélisation du Ca^{2+} utilise uniquement sa densité électronique de cœur, ρ_{core} de l'équation (6), page 27. En effet on lui attribue une charge nette de +2 e, ce qui implique une population de valence P_{val} nulle. La densité électronique de l'anion sulfate a été modélisée à l'aide de la librairie ELMAM2. Au cas où une estimation de l'énergie de polarisation de cet ion aurait été nécessaire, des polarisabilités isotropes additives ont été attribuées aux différents atomes. La polarisabilité expérimentale[114] de l'ion SO_4^{2-} en milieu aqueux et 4.432±0.110 Å^3. Cette polarisabilité moléculaire a été partagée proportionnellement aux P_{val} du soufre et des oxygènes. Le soufre transféré a une P_{val} de 6.026 e, soit 18.9 % des électrons de valence, il reçoit une polarisabilité $\boldsymbol{\alpha}_{iso}$ de 0.839 Å^3. Chaque oxygène est associé à 20.3 % des électrons de valence (P_{val} = 6.452 e), ils se voient attribuer une polarisabilité de 0.898 Å^3. Idéalement le sulfate ne participerait à la polarisation de la benzamidine et des acides aminés du site actif que par la contribution de ses multipoles.

Ce modèle protéique a donc été préparé pour la polarisation et l'étude topologique, mais celles-ci ont été suspendues, faute de temps. Le modèle suivant est plus complexe et sa préparation n'est pas terminée, mais il pourrait permettre des comparaisons bien plus intéressantes que le système trypsine.

Arginase

L'arginase mutante 1HQG est venue compléter les protéines étudiées pour comparaison avec le système urée-éther-couronne, décrit partie 4.2 ci-après. L'arginase est une protéine d'intérêt physiologique[115], car impliquée dans l'excrétion rénale de

l'azote du métabolisme sous forme d'urée. L'urée est transportée par le sérum depuis les tissus vers le rein sous forme d'arginine libre. Cette structure a été choisie particulièrement car elle contient de l'urée. Cette molécule est extraordinairement rare dans la PDB. Au 8 octobre, 118676 structures parmi les 156587 déposées possèdent des ligands. 71 seulement comportent une molécule d'urée dans leur cristal. La raison est simple, avant la généralisation des techniques thermodynamiques et spectroscopiques pour étudier la dénaturation des protéines, leur stabilité était fréquemment mesurée par leur résistance à… l'urée. L'arginase 1HQG fait donc partie des rares élues à pouvoir accepter l'urée chaotrope au sein de son cristal. Mais plus encore, cette structure a été choisie car l'urée fait partie des produits biologiques de sa réaction catalytique. C'est donc une interaction spécifique, comme celles recherchées dans la conception d'un ligand.

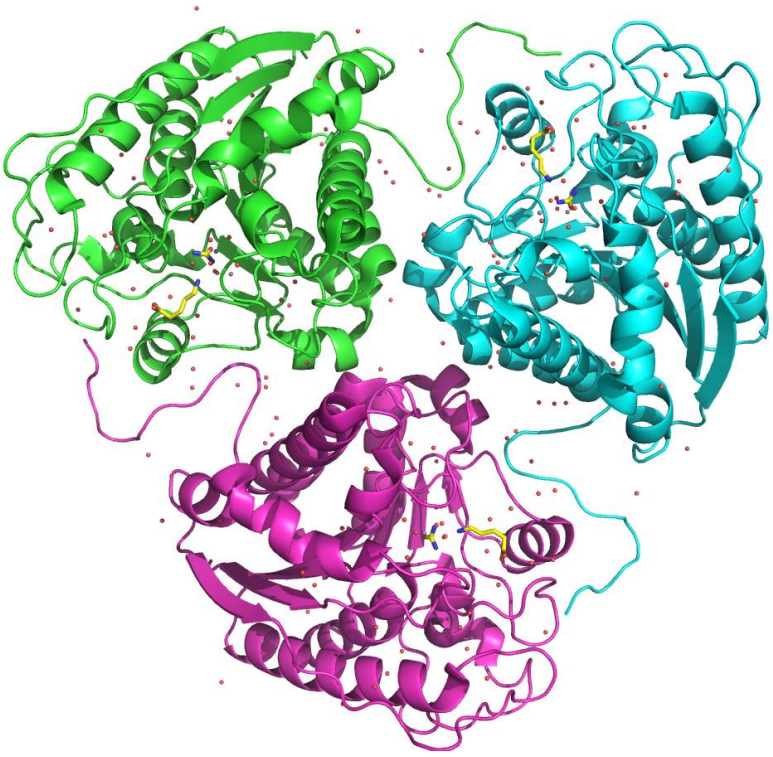


Figure 66. L'arginase 1HQG est un homo-trimère. La chaîne A, préparée pour cette partie comporte 5023 atomes après ajout des atomes d'hydrogène. Les produits de réaction urée et ornithine sont retrouvés au niveau des sites actifs (en jaune). Cette structure comporte un grand nombre de molécules d'eau cristallographiques représentées par des sphères rouges.

L'arginase 1HQG possède une mutation H141C située au niveau du site actif. Cette cystéine est réduite et impliquée dans un pont disulfure avec un β-mercaptoéthanol. La mutation retrouvée dans la structure PDB n'est pas inhibitrice, mais réduit l'activité catalytique. En effet, k_{off} apparait réduit ce qui favorise la fixation des produits.

La chaine A de l'arginase, avec ses produits et molécules d'eau associées, a été importée dans *MoProViewer* après ajout des atomes d'hydrogène. Le logiciel *Reduce*[103] (disponible sur le serveur *MolProbity*[104]) permet d'ajouter ces atomes à une structure protéique expérimentale. *Reduce* ne donne cependant pas de position pour les atomes d'hydrogène de l'eau. Or il y a un ion hydroxyde et 11 autres molécules d'eau à proximité du site actif de l'arginase. La présence de leurs atomes d'hydrogène, à des positions réalistes, est cruciale pour modéliser le potentiel électrostatique au niveau du site actif de l'arginase 1HQG. L'étude Cox *et al.*, 2001[115], a été consultée pour connaître l'état de protonation des différents partenaires du site actif.

Les hydrogènes des molécules d'eau ont été ajoutés par *Pymol*. Leur position laisse à désirer au premier regard. Cependant leur présence est suffisante pour permettre le transfert de paramètres à partir de *MoProViewer*. Les 2 ions manganèse Mn^{2+} du site actif ont été traités comme le Ca^{2+} de la trypsine.

L'idée originale était d'optimiser après le transfert la position de ces hydrogènes, sans toucher aux positions des oxygènes des molécules d'eau, déterminées expérimentalement. Comme abordé page 201 à propos de la sérine 88, la position de ces atomes d'hydrogène peut avoir un effet énergétique crucial. Garder les oxygènes à des positions fixes devrait pouvoir alléger les conflits stériques (qui ne sont pas encore évaluables de fait de l'absence d'énergie de van der Waals adaptée à notre modélisation). Choisir une position optimale pour les hydrogènes d'une molécule d'eau dans une structure de protéine n'est pas un problème trivial, cependant on peut utiliser le potentiel électrostatique pour générer une estimation de première intention.

On souhaite minimiser l'énergie des positions des hydrogènes autour de l'oxygène par rapport au potentiel électrostatique multipolaire des résidus et ligands voisins (les hydrogènes de l'ammonium ε de l'ornithine ont également une position douteuse). Pour cela on calcule une carte de potentiel électrostatique autour de la molécule d'eau à optimiser. Une fonction d'échantillonnage de *MoProViewer* permet de générer des points

uniformément répartis sur une sphère (centrée sur l'oxygène, de rayon r = 0.925 Å). La position avec le potentiel électrostatique le plus faible est recherchée en suivant le gradient du champ de potentiel le long de la surface de la sphère. Cette position est la position candidate pour le premier hydrogène. La position du second hydrogène est recherchée sur la même sphère de 0.925 Å de rayon. On veut cependant un angle constant entre les trois atomes de H_2O (107.5 °). Les positions candidates pour le second hydrogène seront donc réparties sur un cercle, généré à l'aide de la formule de Rodrigues [équation (A9), pae 292 de l'annexe].

Ce code pour optimiser les positions des hydrogènes n'est pas terminé. Il a été pensé pour tester la possibilité future d'implémentation de simulations de type Monte Carlo à *MoProViewer* (et aussi pour l'auto-formation de l'auteur).

La préparation des modèles d'interaction protéine-ligand reste un facteur très limitant à l'emploi des méthodes telles que présentées à la partie 3.7. Par exemple, les paramètres multipolaires et la polarisabilité topologique de l'anion hydroxyde sont absent de la librairie développée ici et doivent être ajoutés. Tous ces facteurs ont repoussé l'étude de ce système à une date ultérieure.

4.2 - Application à un cluster de complexe urée-éther-couronne

Détermination des interactions électrostatiques dans un cristal d'urée

En 2017, l'étude « *A theoretical-electron-density databank using a model of real and virtual spherical atoms* »[82] comparait les performances de différentes bases de données d'atomes transférables pour leur capacité à reproduire des densités électroniques expérimentales et théoriques, ainsi que le champ de potentiel électrostatique. Cet article introduit une méthode basée sur des atomes sphériques virtuels pour modéliser les déformations de la densité électronique habituellement modélisées par le HCMM.

Pour cette étude, les différents modèles de densité électronique de cristaux d'urée ont été préparés (les coordonnées sont tirées de Birkedal *et al.*, 2004[116]). L'urée est en effet un système cristallographique modèle, elle est présente dans de très nombreux cristaux dont certains d'excellente résolution (voir plus bas). Les énergies d'interactions électrostatiques (permanentes) entre les différents dimères dans ce cristal ont été

calculées, utilisant les différents modèles de densité électronique abordés dans l'étude (*IAM*, ELMAM2[61], UBDB[68], RVSA[82] – le modèle utilisant des atomes sphériques virtuels – ainsi que des contrôles affinés contre des facteurs de structure théoriques).

Cette étude impliquait également une autre structure du complexe streptavidine-biotine, tirée des structures PDB préparées pour la base SILICA (code 2F01 dans la Table A4). Elle a été fournie au professeur Benoît Guillot pour l'analyse du potentiel et de la topologie de la densité électronique. Les résultats sont numériquement comparables à ceux reportés partie 3.7.e.

L'arginase a été choisie dans la partie précédente entre autres car le cristal d'urée avait déjà été étudié. Ainsi, différents modèles de densité électronique de l'urée sont déjà disponibles pour étudier les liaisons hydrogène dans ce cristal. La polarisation de ces liaisons reste à évaluer. L'objectif de la préparation des structures de l'arginase-urée et du cristal d'urée était la comparaison au système urée-éther-couronne décrit ci-après.

Une mesure expérimentale du champ électrique dans le complexe urée-éther-couronne

En février 2019, une étude détaillée par Shi *et al.*[117] d'un potentiel système modèle est parue dans le *Journal of the American Chemical Society*. Cette étude de diffraction de rayons X et de neutrons à haute résolution porte sur un co-cristal d'urée et d'un éther couronne $C_{12}H_{24}O_6$. L'éther couronne est composé de 6 unités à deux carbones liés par des groupements éther en un cycle saturé de 18 atomes (Figure 67). Les auteurs ont réalisé une étude détaillée de la densité électronique du complexe ainsi que de son potentiel électrostatique, des champs électriques et moments dipolaires, en complémentant l'approche expérimentale par des calculs de densité électronique théorique. Considérant l'approche hybride du présent travail, semi-expérimentale et semi-théorique, l'étude poussée de Shi *et al.* constitue un parfait point de comparaison.

Le co-cristal contient deux demi-molécules d'éther-couronne cristallographiquement indépendantes qui comportent chacune un centre d'inversion (groupe d'espace $P2_1/c$). Le cristal est organisé en couches d'urée séparant les couches d'éther couronne, comme montré Figure 67. En tout, l'unité asymétrique comporte 5 molécules d'urée. Chaque éther couronne réalise 4 liaisons hydrogène avec deux molécules d'urée reliées par le centre d'inversion du cycle. Ces molécules d'urée – les

invitées du système hôte-invité – sont clairement situées en dehors de la couche d'urée. On s'intéresse particulièrement à la composante électrostatique et de polarisation de l'énergie d'interaction entre les éthers couronne (les hôtes) et les molécules d'urée invitées.

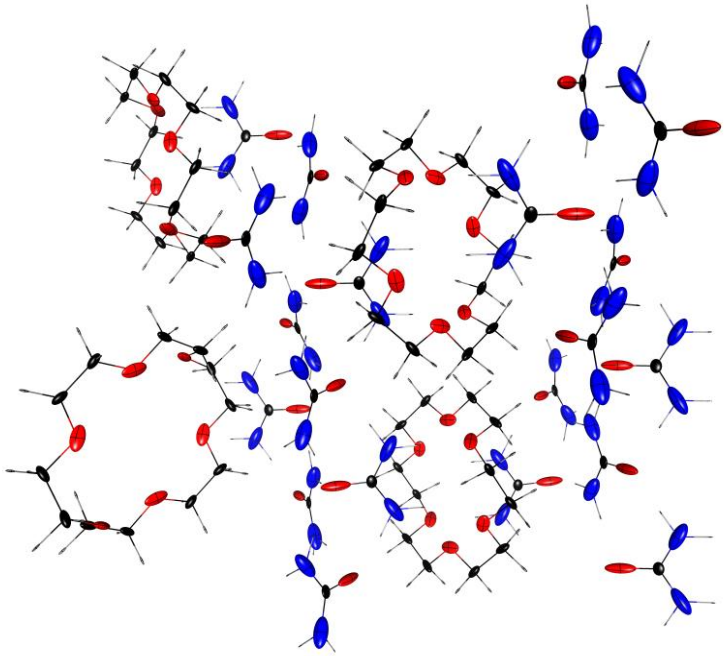


Figure 67. La maille du cristal urée-éther couronne est composée de 4 unités asymétriques. Les ellipsoïdes des polarisabilités transférées de la base pELMAM3 sont représentées à une échelle de 0.25 $Å^3/Å$.

Afin d'analyser cette interaction, plusieurs étapes de préparation ont été nécessaires. Considérant le centre d'inversion trouvé au centre de chaque éther couronne, ces molécules ne sont pas représentées dans leur totalité au sein de l'unité asymétrique. La réduction de la symétrie du système a été réalisée, par une description de la structure dans le groupe d'espace P1, pour faciliter le transfert des paramètres multipolaires et des polarisabilités (voir chapitre 3.2). Les molécules incomplètes ont été translatées pour reconstituer les interactions à étudier au sein de la maille centrale.

Le but de simuler un nano-cristal était de pouvoir contrôler si les contributions à longue distance peuvent ou non être ignorées et sinon, si celles-ci sont fiables dans un système aussi massif que la chaine A de l'arginase (5023 atomes). La maille centrale s'est

donc vue adjointe des 18 mailles les plus proches afin de générer le système de 6232 atomes présenté Figure 68, d'une taille comparable à celle de l'arginase. Les interactions au sein de la maille centrale de ce nano-cristal virtuel doivent servir de contrôle à celles obtenues lors de l'étude de l'arginase.

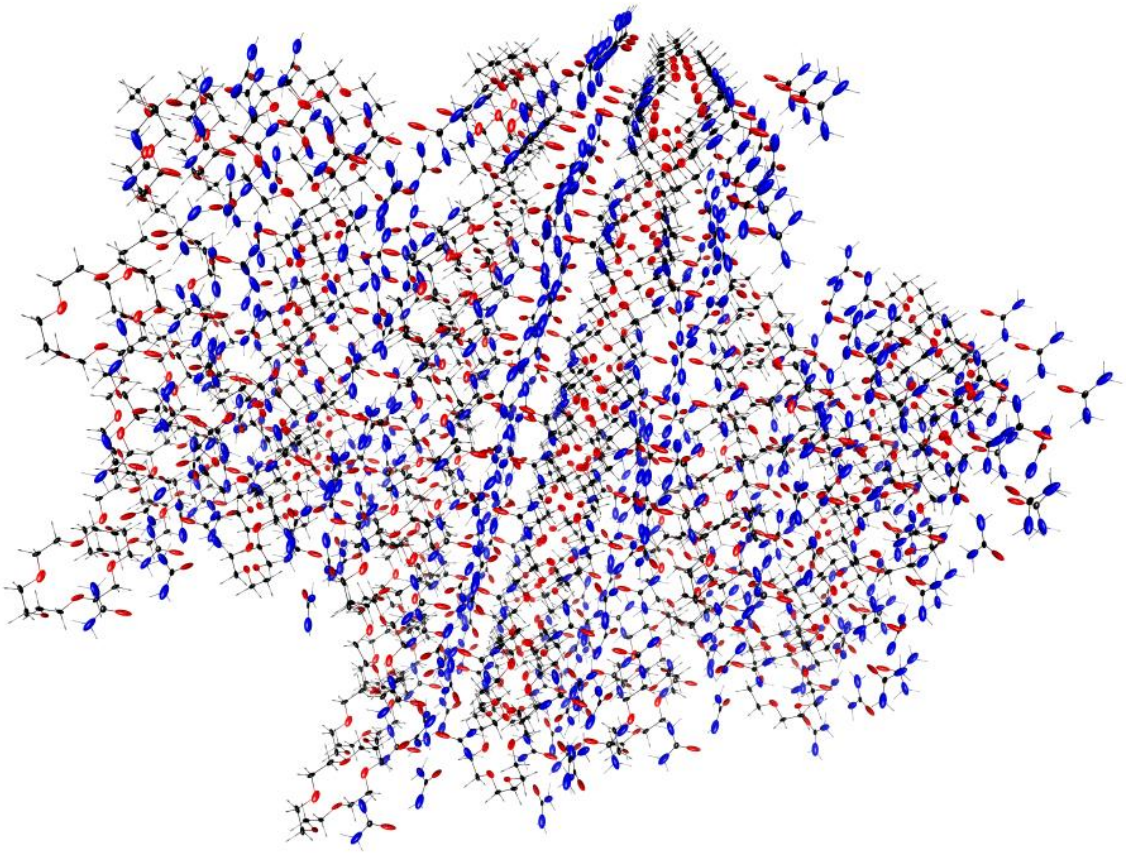


Figure 68. Le nano-cristal préparé à partir de la structure de Shi *et al.*[117] réunit 6232 atomes. Les ellipsoïdes de polarisabilités transférées sont à une échelle de 0.25 $Å^3/Å$.

Les contributions électrostatiques à longue distance que ce système permettra d'estimer sont souvent désignées sous le terme de champ cristallin (ce champ pourrait être simulé par une méthode de type sommation d'Ewald). Cette contribution électrostatique est d'intérêt pour déterminer les effets de l'environnement cristallin sur les atomes multipolaires. Cet effet est supposé perdu dans la procédure de moyennation et de symétrisation des paramètres multipolaires. Étudier cet effet trouvé en phase

cristalline est important pour améliorer la transférabilité des pseudoatomes ELMAM2 mais aussi pour l'ingénierie cristalline.

Seule la préparation de la structure moléculaire a été réalisée. L'étude est à un stade préliminaire, il est nécessaire de vérifier si son design permet d'obtenir les résultats escomptés. L'estimation du champ cristallin n'a pas été possible dans le temps imparti. Une fois celui-ci obtenu, une approche de type *Divide-and-Conquer* est envisagée pour étudier les interactions entre urée et éther-couronne.

4.3 - Énergie de pénétration liée au phénomène de polarisation

Bojarowski *et al.*, 2018[29,30,101], ont récemment publié une méthode de détermination des énergies d'interaction électrostatiques basée sur le schéma EP/MM [équation (10)]. Le schéma EP/MM utilise à courte portée une intégration explicite du la densité électronique et du potentiel électrostatique (EP). Une sommation d'interaction entre moments électrostatiques ponctuels est utilisée pour les contributions à longue portée (MM). La partie « EP » prend en compte l'interpénétration des nuages électroniques associés aux molécules en interaction. Leur article définit donc l'énergie de pénétration (électrostatique) comme la différence entre le calcul d'énergie d'interaction électrostatique par la méthode « EP » et le calcul par la méthode « MM ». L'énergie de pénétration représente donc la contribution énergétique liée au recouvrement de la densité électronique à courte portée.

Les énergies de polarisation E_{POL} présentées dans ce manuscrit sont toutes calculées par EP/MM, utilisant les équations (10) et (37), pages 30 et 47. Ces énergies décroissent très vite avec la distance, et à courte distance, les densités électroniques moléculaires en interaction sont potentiellement dans une situation d'interpénétration. Pour obtenir la partie de E_{POL} relative à cette pénétration des nuages électroniques, il faut obtenir une énergie de polarisation à partir de moments ponctuels. Suivant la définition de Bojarowski *et al.*, l'énergie de pénétration associée à la polarisation sera ainsi écrite comme la différence entre l'énergie de polarisation calculée par la méthode d'intégration (« EP ») et celle obtenue par sommation de moments ponctuels (« MM »).

E_{POL} avec EP/MM, partage des contributions moléculaires :

La page 48 fournit l'équation (40) – homogène au produit scalaire ½·**E**·**μ** – pour calculer cette énergie. Cependant, la densité électronique est modifiée par la polarisation, il se pose immédiatement la question de quel champ électrique utiliser dans cette relation. Pour cela, il faut récapituler les contributions à l'énergie d'interaction électrostatique totale entre les densités polarisées des deux molécules en interaction, notées ici *i* et *j*.

Cette énergie d'interaction électrostatique totale est (en omettant la contribution nucléaire) :

$$E_{ELEC,POLARIZED} = \int\int \frac{\rho_{TOTAL,i}(\mathbf{r})\rho_{TOTAL,j}(\mathbf{r'})}{|\mathbf{r}-\mathbf{r'}|} d\mathbf{r'}\, d\mathbf{r}$$

$$E_{ELEC,POLARIZED} = \int \rho_{TOTAL,i}(\mathbf{r}) \cdot \varphi_{TOTAL,j}(\mathbf{r'}) d\mathbf{r'} d\mathbf{r} \qquad (79)$$

Où $\rho_{TOTAL,i}(\mathbf{r})$ et $\rho_{TOTAL,j}(\mathbf{r})$ sont les densités électroniques transférées-polarisées [soit la somme de $\rho_{DB}(\mathbf{r})$ et $\rho_{pol}(\mathbf{r})$, voir page 47]. Cette relation est aussi écrite en fonction du potentiel électrostatique total, $\varphi_{TOTAL,j}$ généré par la molécule *j*.

Puisque $\rho_{TOTAL} = \rho_{DB} + \rho_{pol}$, cette énergie totale $E_{ELEC,POLARIZED}$ contient :

• L'interaction entre $\rho_{DB,i}(\mathbf{r})$, la densité électronique transférée, avec $\varphi_{DB,j}(\mathbf{r})$, le potentiel électrostatique de la densité électronique transférée de l'autre molécule. Cette interaction est simplement E_{ELEC}.

• L'interaction entre $\rho_{pol,i}(\mathbf{r})$, la densité électronique de déformation liée à l'induction et et $\varphi_{pol,j}(\mathbf{r})$, le potentiel électrostatique généré par $\rho_{pol,j}(\mathbf{r})$ de l'autre molécule. Il s'agit donc d'une interaction dipôle induit – dipôle induit, contenue dans E_{POL}.

• Les termes croisés qui représentent l'interaction entre $\rho_{DB,i}(\mathbf{r})$ et $\varphi_{pol,j}(\mathbf{r})$, ainsi que, réciproquement l'interaction entre $\rho_{pol,i}(\mathbf{r})$ et $\varphi_{DB,j}(\mathbf{r})$. Ces termes sont contenus également dans E_{POL}.

Il serait désirable de séparer les contributions moléculaires contenues dans l'énergie de polarisation E_{POL} par EP/MM. Cette séparation en contributions moléculaires est cruciale pour pouvoir déterminer E_{POL} dans un système cristallin[53,54,118] à partir de l'interaction entre densités électroniques.

Si on réécrit E_{POL} (par la méthode « EP ») en utilisant les contributions listées à la page précédente, l'énergie d'interaction électrostatique EP/MM devient, pour une paire de molécules *i* et *j* :

$$E_{POL,EP} = \int\int \frac{\rho_{pol,i}(\boldsymbol{r})\rho_{DB,j}(\boldsymbol{r'})}{|\boldsymbol{r}-\boldsymbol{r'}|} d\boldsymbol{r'}\, d\boldsymbol{r}$$

$$+ \int\int \frac{\rho_{DB,i}(\boldsymbol{r})\rho_{pol,j}(\boldsymbol{r'})}{|\boldsymbol{r}-\boldsymbol{r'}|} d\boldsymbol{r'}\, d\boldsymbol{r} \qquad (\mathbf{80})$$

$$+ \int\int \frac{\rho_{pol,i}(\boldsymbol{r})\rho_{pol,j}(\boldsymbol{r'})}{|\boldsymbol{r}-\boldsymbol{r'}|} d\boldsymbol{r'}\, d\boldsymbol{r}$$

Pour obtenir $E_{POL,EP,i}$, l'énergie de polarisation de la molécule *i*, on suppose que l'interaction électrostatique entre $\rho_{pol,i}(\mathbf{r})$ et $\rho_{pol,j}(\mathbf{r})$ est partagée équitablement entre les molécules *i* et *j*. L'énergie de polarisation de la molécule *i* n'est pas explicitement calculée dans les procédures présentées, mais peut formellement être définie comme :

$$E_{POL,EP,i} = \int\int \frac{\rho_{pol,i}(\boldsymbol{r})\rho_{DB,j}(\boldsymbol{r'})}{|\boldsymbol{r}-\boldsymbol{r'}|} d\boldsymbol{r'}\, d\boldsymbol{r} + \frac{1}{2}\int\int \frac{\rho_{pol,i}(\boldsymbol{r})\rho_{pol,j}(\boldsymbol{r'})}{|\boldsymbol{r}-\boldsymbol{r'}|} d\boldsymbol{r'}\, d\boldsymbol{r} \quad (\mathbf{81})$$

Où le premier terme symbolise l'interaction entre $\rho_{pol,i}(\boldsymbol{r})$ et le potentiel calculé à partir de la densité électronique transférée sur la molécule *j*. Le second terme représente l'interaction électrostatique entre les déformations induites par le processus de polarisation dans les deux molécules en interaction. Maintenant que l'énergie de polarisation EP/MM est définie pour un monomère, il est possible de la comparer à l'énergie de polarisation – de type ponctuelle – de l'équation (40).

Énergie de pénétration liée à l'induction

Dans l'équation (40), page 48, le champ électrique $\mathbf{E}_{total}(\mathbf{r})$ généré par $\rho_{total}(\mathbf{r})$ a deux composantes additives : $\mathbf{E}_{DB}(\mathbf{r})$ [généré par $\rho_{DB}(\mathbf{r})$] et $\mathbf{E}_{pol}(\mathbf{r})$ [généré par $\rho_{pol}(\mathbf{r})$]. Bien entendu, $\boldsymbol{\alpha}_k \cdot \mathbf{E}_{total}(\mathbf{r}_k)$ est le dipôle induit total (convergé) sur l'atome *k* (situé au point $\mathbf{r}_k$). De même, $\boldsymbol{\alpha}_k \cdot \mathbf{E}_{DB}(\mathbf{r}_k)$ représente le dipôle induit au premier cycle de polarisation. Donc, l'énergie calculée par $0.5 \cdot [\mathbf{E}_{DB}(\mathbf{r}_k)^t \cdot \boldsymbol{\alpha}_k \cdot \mathbf{E}_{DB}(\mathbf{r}_k)]$ ne compte que l'énergie du dipôle induit par $\mathbf{E}_{DB}(\mathbf{r}_k)$ au premier cycle en interaction avec le champ électrique permanent $\mathbf{E}_{DB}(\mathbf{r}_k)$, régnant au noyau de l'atome *k*.

Par conséquent, la formule tensorielle de l'équation (40) pour E_{POL} doit être corrigée comme suit, dès lors que $\rho_{DB}(\mathbf{r})$ est modifiée par $\rho_{pol}(\mathbf{r})$ et que par conséquent $\mathbf{E}_{DB}(\mathbf{r}_k)$ est différent de $\mathbf{E}_{total}(\mathbf{r}_k)$:

$$E_{POL,MM} = -\frac{1}{2}\sum_k \left(\left(\mathbf{E}^t_{DB}(\mathbf{r}_k) + \mathbf{E}^t_{pol}(\mathbf{r}_k) \right) \cdot \boldsymbol{\alpha}_k \cdot \left(\mathbf{E}_{DB}(\mathbf{r}_k) + \mathbf{E}_{pol}(\mathbf{r}_k) \right) \right) \quad \mathbf{(82)}$$

$$\begin{aligned} E_{POL,MM} = -\frac{1}{2}\sum_k [& \mathbf{E}^t_{DB}(\mathbf{r}_k) \cdot \boldsymbol{\alpha}_k \cdot \mathbf{E}_{DB}(\mathbf{r}_k) \\ & + 2 \cdot \mathbf{E}^t_{DB}(\mathbf{r}_k) \cdot \boldsymbol{\alpha}_k \cdot \mathbf{E}_{pol}(\mathbf{r}_k) \\ & + \mathbf{E}^t_{pol}(\mathbf{r}_k) \cdot \boldsymbol{\alpha}_k \cdot \mathbf{E}_{pol}(\mathbf{r}_k)] \end{aligned} \quad \mathbf{(83)}$$

Il apparait dans cette expression qu'il manque l'interaction du dipôle induit total $\boldsymbol{\alpha}_k \cdot [\mathbf{E}_{DB}(\mathbf{r}_k) + \mathbf{E}_{pol}(\mathbf{r}_k)]$ avec le champ électrique généré par les multipôles permanents. On propose donc la correction suivante, pour estimer E_{POL} par la méthode « MM » :

$$E_{POL,MM} = -\frac{1}{2}\sum_k (\mathbf{E}^t_{total}(\mathbf{r}_k) \cdot \boldsymbol{\alpha}_k \cdot \mathbf{E}_{total}(\mathbf{r}_k) + \mathbf{E}^t_{DB}(\mathbf{r}_k) \cdot \boldsymbol{\alpha}_k \cdot \mathbf{E}_{total}(\mathbf{r}_k)) \quad \mathbf{(84)}$$

La sommation prend en compte les k atomes du dimère composé des molécules i et j. Si l'atome est contenu dans la molécule i, le terme $\mathbf{E}_{DB}(\mathbf{r}_k)^t \cdot \boldsymbol{\alpha}_k \cdot \mathbf{E}_{total}(\mathbf{r}_k)$ représente l'interaction entre la densité de déformation induite associée à la molécule i et le champ électrique généré par la densité transférée de la molécule j. Inversement, si l'atome k appartient à la molécule j, il ressent le champ $\mathbf{E}_{DB}(\mathbf{r}_k)$ qui est généré par densité électronique transférée de la molécule i. Cette expression est implémentée dans le module de polarisation du logiciel *MoProViewer*. Lors d'un calcul de polarisation, les énergies calculées à l'aide de l'équation (84) peuvent être accédées à l'aide du mot-clé '`ATMPOLENER`', qui utilisera le champ électrique du cycle en cours comme $\mathbf{E}_{total}(\mathbf{r}_k)$.

La méthode EP/MM donne une énergie d'interaction électrostatique qui inclut forcément les deux molécules du système dimère. D'ailleurs ce calcul est symétrique, inverser le rôle des densités électroniques de i et de j dans l'équation (79) n'a pas d'effet numérique (aux erreurs d'intégration près††††). À l'inverse de la méthode EP/MM, appliquer l'équation (84) à tous les atomes d'une molécule donne l'énergie de polarisation

†††† : *Pour les calculs reportés, ces erreurs d'intégration ne dépassent pas 1% des énergies calculées. Par conséquent, deux calculs d'E_{POL} réciproques sont moyennés pour obtenir les E_{POL} telles que communiquées ici.*

de cette molécule, tandis qu'appliquer l'équation (84) à tous les atomes d'un dimère donne l'énergie de polarisation de ce dimère. Des discussions pertinentes sur les interactions électrostatiques peuvent être trouvées dans Suponitsky *et al.*, 1999[118], et les références qu'il contient.

La contribution de pénétration à l'énergie d'induction, soit la différence entre l'équation (84) et E_{POL} calculée par EP/MM [équations (79) pour un dimère ou (81) pour un monomère], n'a pas été évaluée systématiquement pour le moment. Le calcul de l'équation 84 a été implémenté, mais pas celui le calcul de l'équation (81). Au moment de la rédaction, ces méthodes n'ont pas fait l'objet d'un banc d'essai systématique. L'analyse de ce type de calculs d'énergies pourrait tout à fait rentrer dans le cadre d'une publication courte (type lettre).

4.4 - Cartes de recouvrement de densité électronique, méthodes de calcul des volumes des isosurfaces

L'objectif ici est d'associer une représentation instinctive aux calculs d'énergie de pénétration ébauchés à la partie précédente. Si l'interpénétration des nuages électroniques est importante, la différence entre les énergies calculées par EP et les énergies calculées par l'interaction de moments ponctuels va croitre, et ce pour les contributions E_{ELEC} et E_{POL}. Il est possible de matérialiser cette pénétration des nuages électroniques par une fonction scalaire très simple, $\rho_{OVERLAP}(\mathbf{r})$, la densité de recouvrement entre les densités électroniques $\rho_i(\mathbf{r})$ et $\rho_j(\mathbf{r})$:

$$\rho_{OVERLAP,i,j}(\mathbf{r}) = min\left(\rho_i(\mathbf{r}), \rho_j(\mathbf{r})\right) \qquad (\mathbf{85})$$

Cette fonction tend vers zéro partout si les entités *i* et *j* sont très distantes. Par exemple, cette fonction peut être appliquée à deux atomes pour observer le recouvrement lié à une liaison covalente. Ici, on s'intéresse à des interactions intermoléculaires.

Les figures 69 et 70 représentent la fonction $\rho_{OVERLAP}(\mathbf{r})$ pour deux dimères du jeu de données S66, respectivement Uracile···Uracile (π-π) et *N*-méthyl-acétamide···Eau.

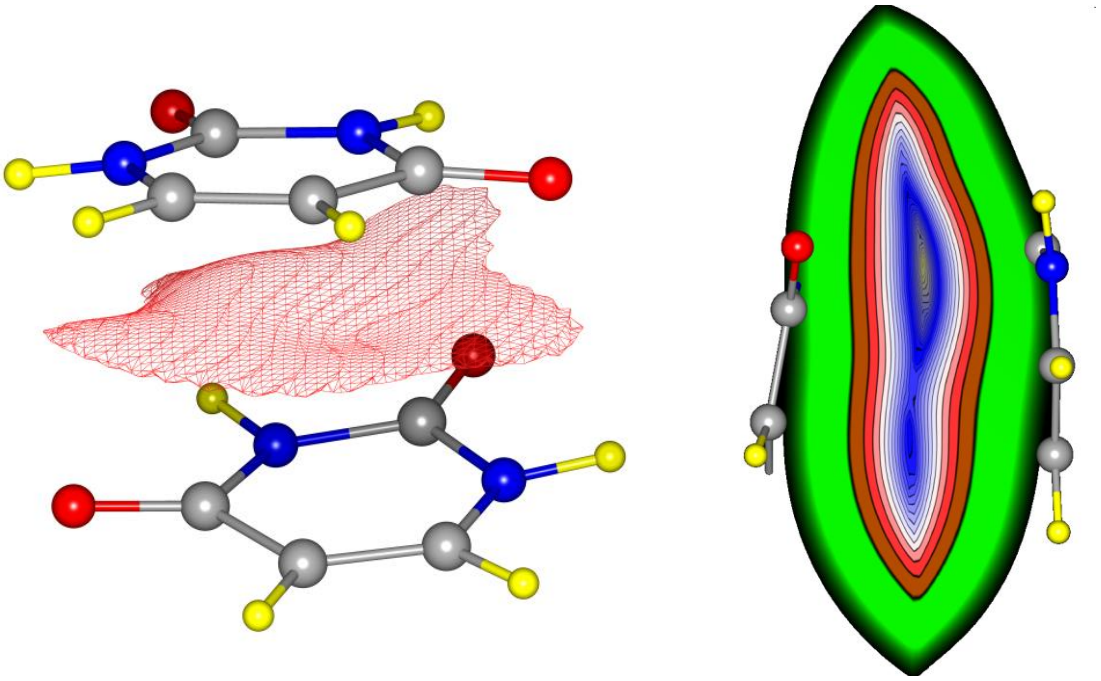


Figure 69. Sur ce dimère d'uracile tiré du S66 (distance d'équilibre), la fonction $\rho_{OVERLAP}(\mathbf{r})$ est calculée à partir des densités électroniques transférées et est matérialisée par une isosurface rouge contourée à 0.01 e·Å^{-3} (à gauche). À droite, une coupe de la densité de recouvrement montre un maximum entre les atomes N (en bleu) et O (en rouge).

D'un côté, on compare un système avec une large surface d'interaction (Figure 69) avec un système lié par une liaison hydrogène (Figure 70). Il apparaît clairement que l'interaction entre les uraciles implique une plus grande « surface » moléculaire que celle entre l'amide et l'eau. Cependant, le pic de densité de recouvrement trouvé dans le complexe *N*-méthyl-acétamide···Eau semble plus intense. L'isosurface de la Figure 70 est également bien plus convexe que celle de la Figure 69.

Il serait intéressant de mesurer la distance entre les points critiques (3, -3) de $\rho_{OVERLAP}(\mathbf{r})$ et les points critiques de liaison de $\rho(\mathbf{r})$. Ils sont probablement très proches. Au moment de la rédaction, seul la méthode de calcul de $\rho_{OVERLAP}(\mathbf{r})$ a été implémentée, dans l'outil « *Map Operations* » de *MoProViewer*, afin de réaliser les figures présentées.

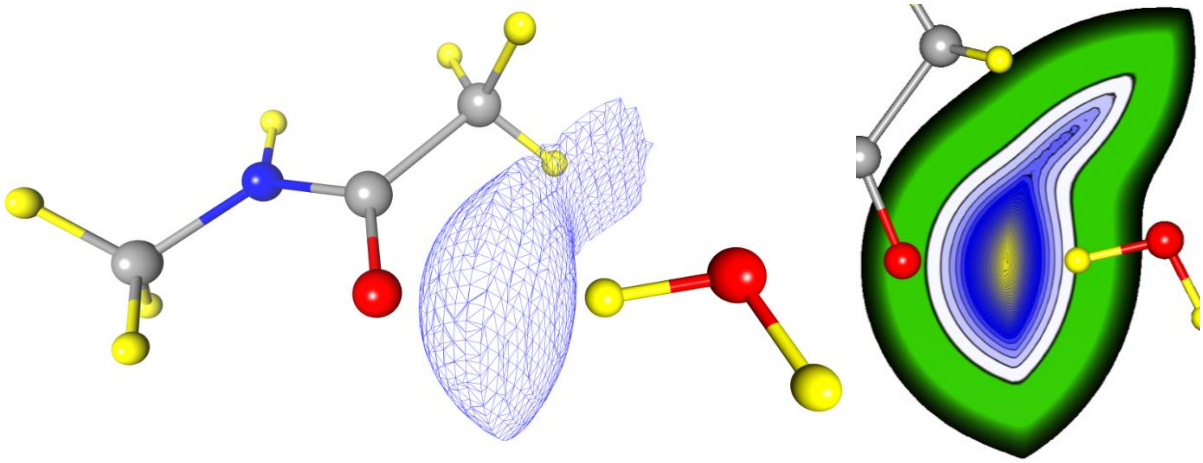


Figure 70. Sur ce dimère *N*-méthyl-acétamide···Eau tiré du S66 (distance d'équilibre), la fonction $\rho_{OVERLAP}(\mathbf{r})$ est calculée à partir des densités électroniques transférées et est matérialisée par une isosurface bleue contourée à 0.01 e·Å^{-3} (à gauche). À droite, une coupe de la densité de recouvrement montre un maximum le long de la liaison hydrogène.

Méthodes de calcul des surfaces et des volumes des isosurfaces :

Afin de faciliter l'investigation des isosurfaces avec *MoProViewer*, des fonctions pour en calculer la surface ou le volume (fermé par la surface) ont été implémentées. Elles ont été initiées pour l'étude de $\rho_{OVERLAP}(\mathbf{r})$. Plus tard, elles pourront tout à fait être réutilisées pour étudier – par exemple – des surfaces de flux nul. Les deux méthodes implémentées utilisent directement le *mesh*, maillage de triangles affiché à l'écran.

Pour déterminer le volume d'un polyèdre convexe, on utilise une astuce très simple, trouvée dans une communication de conférence par Zhang et Chen[119]. Informatiquement chaque triangle de la surface est représenté par 3 vecteurs qui relient ses sommets à l'origine du repère utilisé. Chaque triangle forme donc un tétraèdre dont un des sommets est l'origine du repère. On utilise la formule suivante pour calculer le volume de ce tétraèdre.

$$V = \frac{1}{6}(-x_3y_2z_1 + x_2y_3z_1 + x_3y_1z_2 - x_1y_3z_2 - x_2y_1z_3 + x_1y_2z_3) \qquad \mathbf{(86)}$$

Où (x_1,y_1,z_1), (x_2,y_2,z_2) et (x_3,y_3,z_3) sont les vecteurs position des 3 *vertex* de la face qui sert de base au tétraèdre. Ce volume est un volume orienté qui suppose que le produit vectoriel $(x_3\text{-}x_1, y_3\text{-}y_1, z_3\text{-}z_1) \wedge (x_2\text{-}x_1, y_2\text{-}y_1, z_2\text{-}z_1)$ ait la même direction que la normale à la

face. La structure de données du triangle contient aussi 3 autres vecteurs, unitaires, qui définissent la normale à la surface (pour l'illumination graphique avec *OpenGL*). L'implémentation est cohérente entre la définition de la normale par Zhang et Chen[119] et la génération des vecteurs normaux par *MoProViewer*. L'extérieur et l'intérieur du polyèdre sont donc définis. On fera donc la somme des volumes orientés de ces tétraèdres. Les volumes de même signe seront tous associés à des triangles dont l'origine est du même côté. La valeur absolue de la somme des volumes orientés est le volume du polyèdre.

La méthode de calcul des surfaces utilise une méthode similaire, où la somme de l'aire de chaque triangle aboutit à celle de la surface la surface. Une fonctionalité permettant le calcul des surfaces et volumes des iso-surfaces, calculées sur une gamme de valeurs d'une carte scalaire quelconque, a été implémentée dans le « *3D Map Manager* » de *MoProViewer*, mais seul le test du code a été réalisé.

4.5 - Atténuation anisotrope et adaptation du tenseur de Birge au modèle HCM

Sections perpendiculaires d'ellipsoïdes.

Une différence majeure entre les modèles de Thole et Birge, et le modèle des densités transférées polarisées, est l'utilisation de polarisabilités anisotropes. Avant la publication de Thole, l'utilisation de polarisabilités anisotropes non-atténuées avait déjà été explorée par Birge, 1980[120]. Pour utiliser la fonction d'atténuation de Birge avec ces polarisabilités non-sphériques, tout en conservant la convergence de l'induction, une méthode basée sur le concept de polarisabilité apparente (en fonction des axes du système) est proposée.

Pour projeter la polarisabilité sur le vecteur interatomique $\mathbf{r}_{ij}$ on utilise une relation similaire à l'équation (40), page 48‡‡‡‡. En effet pour utiliser les fonctions d'atténuation, il est nécessaire d'obtenir des polarisabilités scalaires et de mettre les distances interatomiques à l'échelle.

‡‡‡‡ *: Thole[46] et Birge[48] utilisent la symétrie des interactions entre les atomes i et j pour prendre en compte le cas où* $\boldsymbol{\alpha}_i \neq \boldsymbol{\alpha}_j$.

Les polarisabilités atomiques scalaires, utilisées pour mettre à l'échelle $\mathbf{r}_{ij}$, répondent à la relation de projection suivante :

$$\alpha_{i,\parallel} = \mathbf{r}_{ij}{}^{t}\boldsymbol{\alpha}_i\mathbf{r}_{ij} \qquad (87)$$

$\alpha_{j,\parallel}$ est déterminé de même. La notation $\alpha_{i,\parallel}$ revoie aux rayons de divergence de la Figure 6, page 40. Dans le cas d'une polarisabilité très anisotrope il est possible que $\alpha_{i,\perp}$, qui est une projection de $\boldsymbol{\alpha}_i$ sur un vecteur normal à $\mathbf{r}_{ij}$, puisse être responsable de catastrophes de polarisation. Ce risque est encouru si $\alpha_{i,\perp} \geq 4 \cdot \alpha_{i,\parallel}$. Dans ce cas le rayon de divergence pour les composantes perpendiculaires est plus grand que pour les composantes axiales.

Le but est l'utilisation du tenseur de Birge avec des polarisabilités anisotropes. On va donc chercher les vecteurs normaux à $\mathbf{r}_{ij}$ qui maximisent la projection de $\boldsymbol{\alpha}_i$ en utilisant la relation (87). La théorie de la représentation associe à un tenseur défini positif de rang 2 une ellipsoïde en trois dimensions. Les vecteurs normaux à $\mathbf{r}_{ij}$ forment un plan qui coupe l'ellipsoïde. La Figure 71 montre le plan de section, normal au vecteur interatomique noté **x**. Exprimer la polarisabilité en termes de polarisabilité axiale et perpendiculaire revient à la rotation du tenseur $\boldsymbol{\alpha}_i$ dans un nouveau référentiel, local à l'interaction.

Le problème peut se réduire à comment trouver ce nouveau référentiel, dont l'orientation autour de l'axe de l'interaction (axe **x**), permet de maximiser les deux composants perpendiculaires à l'axe (minimiser le composant *yz* de $\boldsymbol{\alpha}$). Pour résoudre cette question directement et éviter un grand nombre de rotations matricielles inutiles (l'inversion du tenseur moléculaire nécessite un référentiel commun à tous les centres atomiques), l'implémentation utilise la méthode de Klein, 2012[121] pour déterminer les valeurs propres d'une ellipse en deux dimensions, définie comme étant la section de l'ellipsoïde représentant $\boldsymbol{\alpha}_i$ par le plan normal à la direction d'interaction, en vert sur la Figure 71.

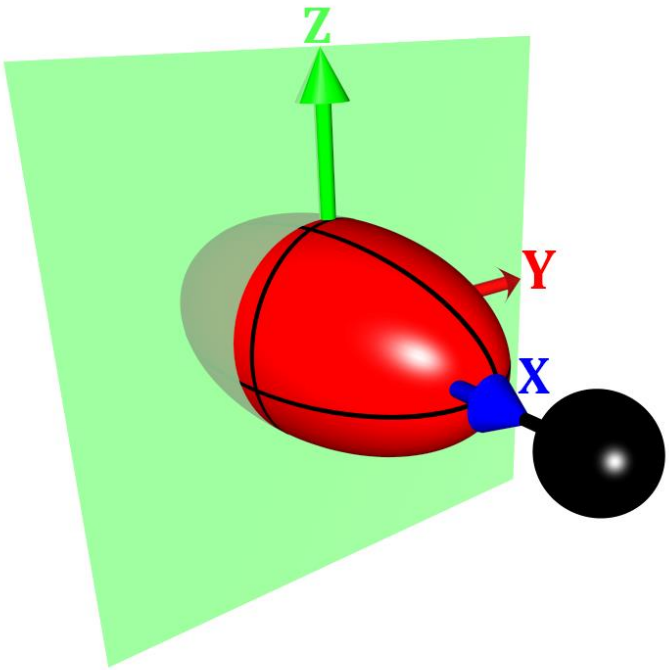


Figure 71. La méthode de Klein[121] permet d'obtenir la section d'une ellipsoïde par un plan, qui est une ellipse. Cette méthode permet de trouver les valeurs propres de cette ellipse de section. La valeur propre maximale est utilisée en tant que $\alpha_{i,\perp}$. Le vecteur bleu est le vecteur **x** (les deux autres vecteurs de la base orthonormée sont dans une position quelconque).

La méthode de Klein est générale et permet de trouver toutes les caractéristiques de la section de l'ellipsoïde par le plan (centre, valeurs propres, vecteurs propres, équation de la ligne, etc.). Ici le plan sécant passe par le centre de l'ellipsoïde et on ne s'intéresse qu'aux valeurs propres, ce qui constitue un cas particulier de la méthode de Klein.

Les deux valeurs propres (notées $\lambda_{\pm}$) de l'ellipse d'intersection répondent à l'équation suivante :

$$\lambda_{\pm} = \left(\frac{2(\mathbf{n}.\mathbf{n})}{-\left(\frac{n_2+n_3}{a^2}+\frac{n_1+n_3}{b^2}+\frac{n_1+n_2}{c^2}\right)} \pm \frac{2(\mathbf{n}.\mathbf{n})}{\sqrt{\left(\frac{n_2+n_3}{a^2}+\frac{n_1+n_3}{b^2}+\frac{n_1+n_2}{c^2}\right)^2 + 4(\mathbf{n}.\mathbf{n})\left(\frac{{n_1}^2}{bc}+\frac{{n_2}^2}{ac}+\frac{{n_3}^2}{ab}\right)}} \right)^{1/2} \quad (88)$$

Où $\mathbf{n} = \begin{pmatrix} n_1 \\ n_2 \\ n_3 \end{pmatrix}$ est un vecteur normal au plan sécant, exprimé dans la base propre de l'ellipsoïde $\boldsymbol{\alpha}_i$ dont *a*, *b*, *c* sont les valeurs propres. Dans cette base :

$$\boldsymbol{\alpha}_i = \begin{pmatrix} a & 0 & 0 \\ 0 & b & 0 \\ 0 & 0 & c \end{pmatrix} \quad (\mathbf{89})$$

L'implémentation choisie diagonalise la polarisabilité de l'atome considéré et utilise les vecteurs propres de la polarisabilité pour exprimer **n**, le vecteur interatomique, dans cette base propre. Les deux valeurs de λ sont calculées à l'aide de l'équation (88) et la valeur maximale est considérée comme pertinente pour éventuellement atténuer le tenseur $\mathbf{T}_{ij}$, si cette valeur de lambda est supérieure à la valeur $\alpha_{i,\parallel}$ de l'équation (87). Ce modèle conserve donc une atténuation isotrope du tenseur $\mathbf{T}_{ij}$ mais prend en compte l'orientation relative des tenseurs. Ce calcul de $\alpha_{i,\perp}$ à l'aide de la formule de Klein est intégré à la construction du tenseur moléculaire par le module Beluga lorsque l'utilisateur souhaite utiliser le modèle STOPDI de Birge avec des polarisabilités anisotropes. Cette méthode d'induction n'a pas bénéficié d'un banc d'essai systématique.

Adaptation du tenseur de Birge au modèle HCM

Les fonctions d'atténuation dérivées par Birge *et al.* pour les fonctions radiales de Slater ne sont pas directement compatibles avec le modèle HCM. En effet, Birge et collaborateurs, à l'instar de Thole, expriment les densités électroniques des atomes en fonction de α_i et α_j. Or, non seulement le modèle HCM n'utilise pas originellement de polarisabilité atomique, mais il utilise en plus le coefficient d'expansion-contraction κ' [on ignore κ car on polarise la densité électronique de déformation seulement, voir équation (6)].

Fondamentalement κ' est un facteur de mise à l'échelle de la partie radiale de la densité électronique. Les formules (32) à (34) des pages 43 et 44 sont donc adaptées au paramètre κ' comme suit :

$$\rho(\mathbf{u}_{ij}) = \kappa'^3 \frac{(\alpha^{1/3}\zeta)^{n_l+3}}{4\pi(n_l!)} (\kappa'|\mathbf{u}_{ij}|)^{n_l} e^{-\alpha^{1/3}\zeta\kappa'|\mathbf{u}_{ij}|} \quad (\mathbf{90})$$

$$\frac{f\prime}{|\mathbf{u}_{ij}|} = \frac{-1}{(\kappa\prime|\mathbf{u}_{ij}|)^3}\left[1 - e^{-\alpha^{1/3}\zeta\kappa\prime|\mathbf{u}_{ij}|}\sum_{i=0}^{n_l+2}\frac{\left(\kappa\prime|\mathbf{u}_{ij}|\alpha^{1/3}\zeta\right)^{n_l-i+2}}{(n_l-i+2)!}\right] \qquad (\mathbf{91})$$

$$f'' = \frac{2}{(\kappa'|\mathbf{u}_{ij}|)^3} - \left(\frac{2}{(\kappa'|\mathbf{u}_{ij}|)^3} + \frac{\alpha^{1/3}\zeta}{(\kappa'|\mathbf{u}_{ij}|)^2}\right)e^{-\alpha^{1/3}\zeta\kappa'|\mathbf{u}_{ij}|}\sum_{i=0}^{n_l+2}\frac{\left(\kappa'|\mathbf{u}_{ij}|\alpha^{1/3}\zeta\right)^{n_l-i+2}}{(n_l-i+2)!} +$$
$$\frac{\alpha^{1/3}\zeta}{(\kappa'|\mathbf{u}_{ij}|)^2}e^{-\alpha^{1/3}\zeta\kappa'|\mathbf{u}_{ij}|}\sum_{i=0}^{n_l+1}\frac{\left(\kappa'|\mathbf{u}_{ij}|\alpha^{1/3}\zeta\right)^{n_l-i+1}}{(n_l-i+1)!} \qquad (\mathbf{92})$$

κ' est affecté au vecteur interatomique mis à l'échelle et à la constante de normalisation de la densité. Ces formules n'ont pas pu être testées efficacement, en effet les polarisabilités topologiques sont trop intenses pour que l'atténuation gouvernée par f' et f'' suffise à obtenir une matrice de polarisabilité moléculaire définie-positive. Ce résultat n'est pas surprenant dans la mesure où la partition topologique de la polarisabilité moléculaire diffère de celle qu'il est possible d'obtenir à l'aide d'un formalisme de type Thole ou Birge. De même que pour la détermination de $\alpha_{i,\perp}$, la mise à l'échelle de par f' et f'' par κ' est calculée par Beluga lorsque l'utilisateur souhaite utiliser le tenseur de polarisation moléculaire de Birge.

Traduction entre modèles de polarisabilités atomiques

Un module de « traduction » entre ces deux modèles de partition de la polarisabilité moléculaire (en polarisabilités atomiques ponctuelles ou topologiques) est en phase de conception. On veut trouver un set de polarisabilités atomiques, anisotropes, moyennées par types d'atomes, qui permettent de reconstituer la polarisation théorique de bassins atomiques en utilisant le modèle de Thole ou Birge (*Point Dipole Interaction*) pour prendre en compte l'interaction électrostatique entre dipôles induits de la même molécule. L'implémentation utilise l'algorithme de Gauss-Newton pour générer ces polarisabilités « ponctuelles et interactives » à partir de polarisabilités topologiques.

Pour cette application, l'algorithme de Gauss-Newton, implémenté en C++ par Maxime Bunel alors étudiant de M2 au CRM², a été porté dans *MoProViewer* et le code a été adapté pour la comparaison de données non-scalaires (3-vecteurs représentant les moments dipolaires induits à chaque site atomique), calculés en utilisant le formalisme de Applequist-Thole-Birge.

Comme l'atténuateur de Birge n'a pas de paramètre libre, les seules variables à affiner serait les valeurs indépendantes – et non-contraintes par symétrie – des polarisabilités atomiques transférables (exprimées dans le référentiel local).

On utilise, comme au chapitre 3.6, un jeu de molécules dont les polarisabilités atomiques topologiques théoriques sont connues, et dont les types atomiques sont bien définis.

L'algorithme utilise les polarisabilités topologiques de départ pour générer un large set de moments induits, moléculaires et atomiques, dans des directions variées. Ensuite, à partir des polarisabilités qu'on cherche à affiner, on construit un tenseur moléculaire [équation (21), page 39] pour calculer la polarisation en réponse aux même champs électriques utilisés pour obtenir les moments de référence.

L'algorithme de Gauss-Newton fait ensuite varier les différents paramètres des polarisabilités transférables pour tenter d'obtenir un jeu capable de reproduire au mieux la polarisation théorique. Ce module demande une logistique considérable pour gérer les données. Les preuves de concept demanderont encore un temps de développement considérable malgré le temps déjà investit pour modifier l'algorithme de Gauss-Newton dans ce but. Un code prototype est cependant déjà implémenté dans le module Beluga mais les polarisabilités obtenues par son application n'ont pas été examinées.

4.6 - Molécules velues

Pour finir, voici un dernier outil développé durant la thèse. Il fait plus figure de curiosité, mais peut avoir un réel intérêt une fois finalisé. Ici, l'objectif est d'enrichir la représentation traditionnelle d'un champ scalaire par des isosurfaces. Les isosurfaces sont un maillage (*mesh*) de points (*vertex*) associés à la même valeur du champ.

Pour décrire le champ de potentiel électrostatique en 3 dimensions, les iso-surfaces sont utilisées depuis longtemps. La méthode présentée est inspirée par une représentation très classique, où une isosurface de densité électronique est coloré selon les valeurs du potentiel. Cette représentation permet d'identifier d'un regard les zones électrophiles et nucléophiles d'une molécule. On souhaite créer une représentation qui pourrait renseigner sur le caractère directionnel des interactions entre molécules.

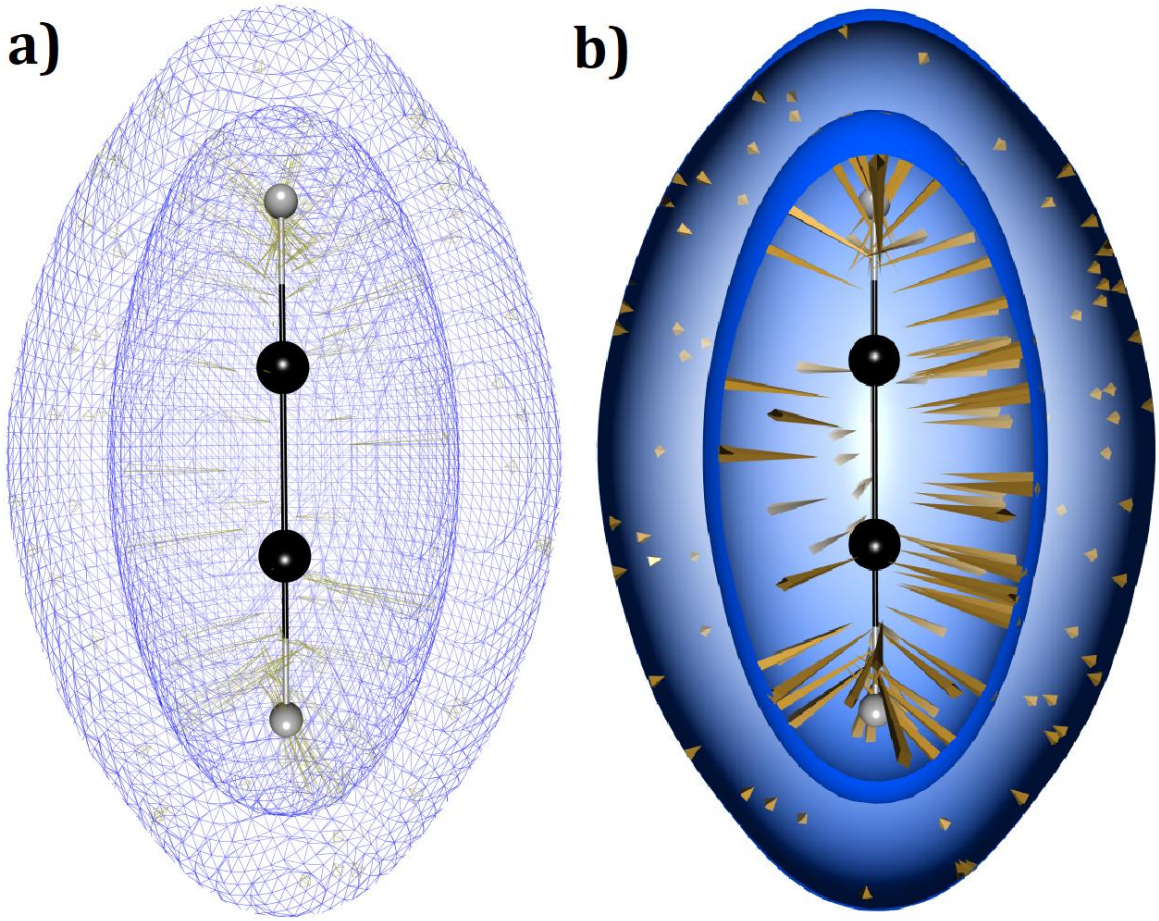


Figure 72. Isosurfaces « velues » de densité électronique transférée de l'éthyne, en représentation « fil de fer », panneau (a) et remplie, panneau (b). Les pyramides représentant le gradient sont en jaune. Isosurface extérieure : +0.02 e·Å^{-3} ; isosurface intérieure : +0.2 e·Å^{-3}.

Pour montrer cette information, une méthode implémentée dans *MoProViewer* est capable de rajouter des « poils » aux isosurfaces. Ces excroissances – en réalité des pyramides à base carrée – représentent le gradient du champ scalaire, calculé au niveau des *vertex* de l'isosurface. La base de la pyramide est centrée au point où le gradient est calculé. La pointe de la pyramide représente la pointe du vecteur gradient. Pour l'instant, l'outil implémenté est limité au gradient du champ représenté par l'isosurface [$\nabla\rho(\mathbf{r})$ pour une carte de $\rho(\mathbf{r})$; $\nabla\varphi(\mathbf{r})$ pour une carte de $\varphi(\mathbf{r})$]. Idéalement, l'outil permettrait d'accoler une représentation vectorielle de $-\nabla\varphi(\mathbf{r})$ (le champ électrique) sur des iso-surfaces de densité électronique.

Un exemple, impliquant la densité électronique transférée de l'éthyne est visible Figure 72. Le gradient de $\rho(\mathbf{r})$ est orienté vers les noyaux et pour que les pyramides soient visibles sur la Figure 72, l'isosurface à l'avant-plan a été coupée. La coloration automatique des pyramides utilise les couleurs complémentaires de l'algorithme présenté page 121.

Les iso-surfaces représentées Figure 72 sont constituées de 10^3 à 10^4 triangles. Chaque pyramide à base carrée comporte 6 triangles. Dessiner une pyramide par polygone rendrait la figure illisible et multiplierait par 7 le nombre de triangles à afficher. *MoProViewer* tire au sort les triangles où le gradient doit être représenté (tirage aléatoire sans replacement, un bouton à cliquer génère un nouveau tirage pour que l'utilisateur ait un contrôle sur celui-ci). Pour l'instant, un polygone sur 100 porte une pyramide mais cette option pourrait aisément être réglable.

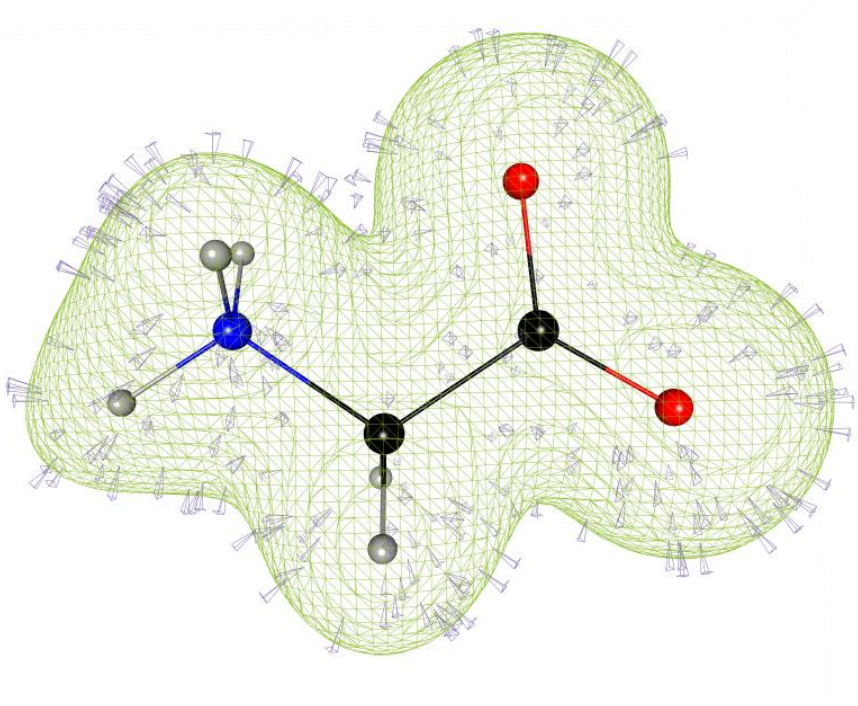


Figure 73. Isosurface « cloutée » de la densité électronique transférée de la glycine, isosurface contourée à 0.1 $e{\cdot}Å^{-3}$. Le gradient de la densité pointe vers l'intérieur de l'isosurface, mais les pyramides sont représentées à l'extérieur de la surface. *Image d'intention.*

Une autre amélioration possible serait d'ajouter une option pour représenter les pyramides à l'extérieur de la surface, la pointe touchant le point ou le gradient est calculé, et la base représentant l'autre extrémité du vecteur. Cette option passerait par la simple évaluation du signe du produit scalaire $\nabla\rho(\mathbf{r})\cdot\mathbf{n}$ où $\mathbf{n}$ est le vecteur normal au polygone lié à la pyramide. Si cette possibilité n'est pas encore implémentée, elle devrait donner une représentation proche de celle montrée sur la molécule de glycine, Figure 73.

Appliquer cette représentation au potentiel électrostatique permet de mettre en évidence les zones où le champ électrique est le plus intense. Cette représentation est démontrée Figure 74, où les isosurfaces de potentiel sont adjointes de pyramides représentant le gradient de $\varphi(\mathbf{r})$.

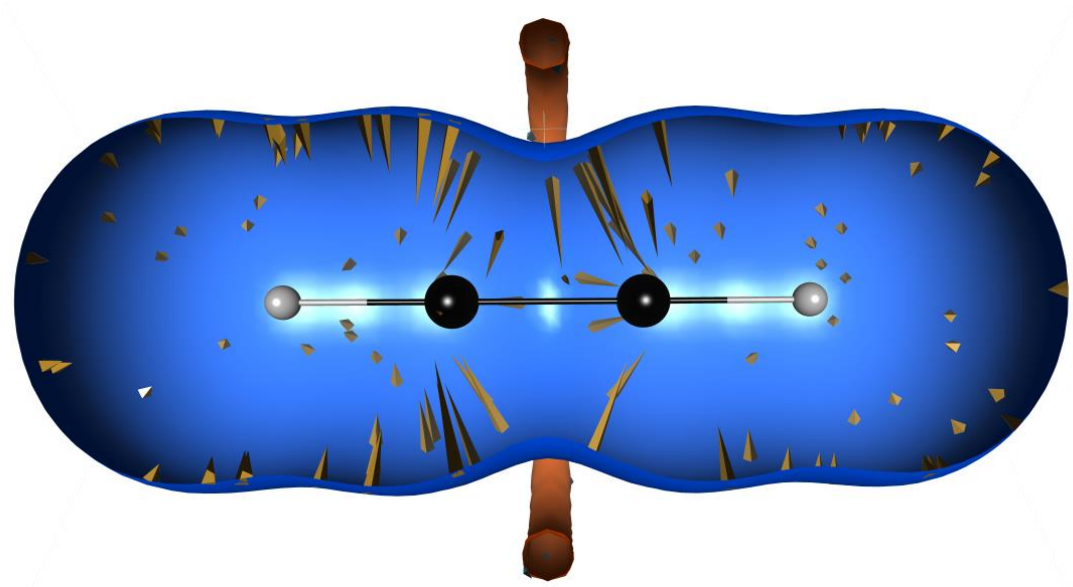


Figure 74. Isosurfaces du potentiel électrostatique transféré de l'éthyne. Les isocontours sont fixés à +/- 0.06 e·Å^{-1}.

On note immédiatement la position équatoriale des plus forts gradients du potentiel (autour de la liaison C≡C). Pour l'instant, le gradient est représenté à l'échelle 1:1, donc sur cette image, 1 Å représente un gradient de 1 e·Å^{-2}. C'est encore un réglage auquel l'utilisateur devrait pouvoir accéder : sur la figure 74, les pyramides adjointes au lobe négatif du potentiel, en orange, sont invisibles. Choisir une isosurface du potentiel

avec une valeur plus négative permet de révéler des gradients de norme supérieure, comme montré Figure 75.

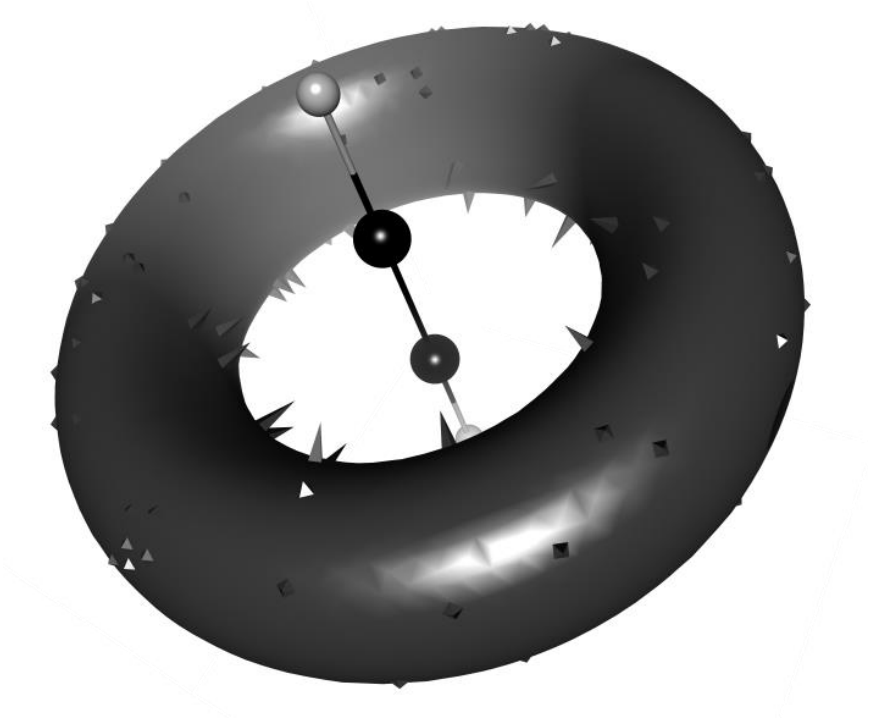


Figure 75. Le lobe négatif du potentiel électrostatique transféré de l'éthyne. Isocoutour à -0.35 e·Å^{-1}.

Un dernier exemple, du lobe négatif du potentiel électrostatique des doublets non-liants du carboxylate de la glycine (affublé de pyramides de gradient) est visible Figure 76. Le choix d'un gris anthracite pour colorer les isosurfaces des Figures 75 et 76 n'a pas de signification scientifique particulière, mais rappelle l'apparence d'un ferrofluide. Ces suspensions colloïdales de nanoparticules magnétiques sont un moyen très connu pour mettre en évidence l'effet des champs magnétiques (pour la communication scientifique destinée au grand public).

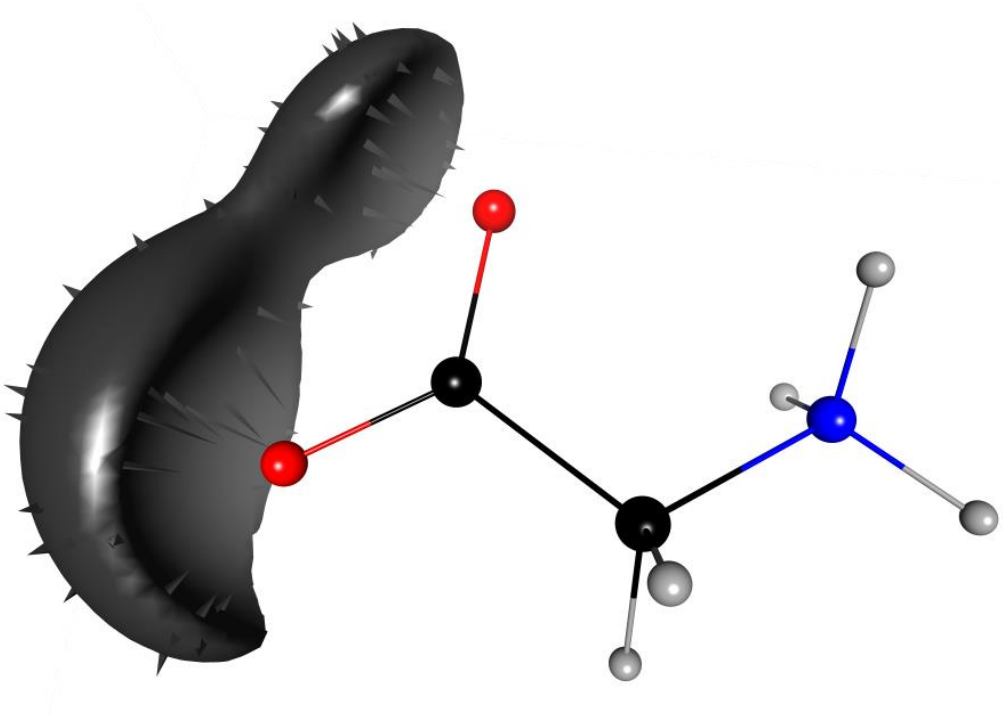


Figure 76. Les doublets non-liants du carboxylate de la glycine génèrent un potentiel négatif. Le lobe négatif du potentiel électrostatique transféré est matérialisé par une isosurface à -0.21 e·Å^{-1}.

Conclusions

Générales

Ce travail de thèse constitue une tentative d'offrir une vue d'ensemble des effets de la polarisabilité sur la densité électronique, dans le contexte particulier de sa reconstruction à partir de pseudoatomes tirés d'une librairie de paramètres expérimentaux. Les outils présentés constituent une base solide pour continuer l'étude de la polarisabilité et de la polarisation de la densité électronique transférée, et exprimée dans le formalisme multipolaire de Hansen et Coppens. Plus que cela, les trois principaux outils logiciels développés (comparaison, polarisation, construction de librairies), décrits dans leurs chapitres respectifs, représentent l'épine dorsale de ce travail. L'application de ces outils a en effet guidé leur développement, d'une manière itérative, jusqu'à permettre l'obtention des résultats présentés et l'intégration à *MoProViewer*.

L'obtention de polarisabilités atomiques topologiques théoriques pour le jeu de structures S66, de référence, est un résultat important en soi. Leur publication dans le *Journal of Physical Chemistry, section A* en tant qu'article sélectionné (« *ACS Editors' Choice* ») montre l'intérêt actuel de la communauté pour cette propriété, dont l'importance pour déterminer les énergies d'interactions électrostatiques à l'échelle moléculaire a été vue dans ces pages. Cet article va cependant bien plus loin qu'un simple catalogue de 1242 tenseurs de polarisabilités atomiques de haute qualité, dérivés de la partition topologique de la densité électronique.

En effet, l'outil de comparaison implémenté dans *MoProViewer* a permis une classification détaillée des polarisabilités, comme détaillé au chapitre 1. Leurs variations ont été étudiées en détails, en fonction de leurs éléments chimiques, en fonction des types atomiques ELMAM2 associées, mais aussi en fonction du contexte supra-moléculaire (dimères de la base de données S66). Il a été trouvé que les types atomiques ELMAM2 constituent une classification satisfaisante – à quelques améliorations près – des polarisabilités. Les déviations statistiques sur les valeurs propres qui les caractérisent sont drastiquement réduites quand on les compare associées à des atomes partageant un type ELMAM2. Les types atomiques ELMAM2 sont conçus pour reconstruire la densité électronique à l'aide de pseudoatomes transférables. La polarisabilité étant une propriété électronique, trouver une variabilité réduite des polarisabilités associées aux mêmes types atomiques permet de postuler leur transférabilité. Ce postulat est étayé par de nombreux arguments vus au cours de ce manuscrit : variations relativement faibles en réponse à l'environnement non-covalent (en magnitude et en orientation, des *outliers*

existent cependant), alignement des polarisabilités atomiques avec les liaisons covalentes des atomes associés[86], relations entre polarisabilités et charges électriques (multipolaires ou intégrées sur les bassins topologiques de densité électronique), etc.

Ces arguments en faveur de la transférabilité des polarisabilités atomiques topologiques ont clairement motivé la constitution du « *Big Set* », un jeu complet de polarisabilités atomiques adapté à l'étude des protéines et des peptides. Ce jeu de données inédit, décrivant ces polarisabilités atomiques topologiques théoriques de grande qualité, est le résultat d'une collaboration internationale fructueuse avec l'université de Cracovie. Cette collaboration avec la professeure Anna Krawczuk et la docteure Katarzyna Rzesikowska doit aboutir à une publication dont les limites exactes restent à définir. Les critères de transférabilité étudiés sur la base du jeu de données S66 seront probablement réutilisés (et complémentés de nouveaux tests).

L'article[52] reproduit ici comprend la publication d'un important banc de test théorique, et systématique, pour les énergies d'interaction électrostatiques permanentes calculées à partir des densités électroniques ELMAM2 transférées. L'analyse phénoménologique des polarisabilités atomiques du jeu de données S66 (chapitre 1) est donc complémentée par le calcul d'énergies d'induction électrostatique, ou « énergies de polarisation ». Ce *benchmarking* démontre de très bonnes performances, malgré le fait que la méthode testée est totalement indépendante de la méthode de référence SAPT. Cette détermination est permise grâce à l'outil de polarisation implémenté dans *MoProViewer* et présenté au chapitre 2. Cet outil permet d'utiliser les polarisabilités pour donner accès à la déformation de la densité électronique transférée, en réponse aux champs électriques extramoléculaires. De bonnes capacités prédictives, pour reproduire les énergies d'induction de référence de la méthode SAPT, sont obtenues à l'aide de ces densités modifiées.

L'utilisation de polarisabilités obtenues lors des calculs sur les monomères isolés du S66 donne les meilleures prédictions pour les énergies d'induction. Outre les performances brutes de la méthode présentée, cet élément peut être discuté comme un argument pour la transférabilité de polarisabilités atomiques topologiques. En effet, les molécules isolées sont associées à des polarisabilités vierges d'influences extramoléculaires, là où des polarisabilités idéalement transférables seraient dépourvues

de ces influences extramoléculaires de manière statistique (en moyennant un grand nombre d'interactions variées).

Ce faisceau cohérent d'arguments en faveur de la transférabilité des polarisabilités atomiques topologiques appelle naturellement la création d'une base de données, pour permettre justement le transfert de cette propriété. L'outil présenté au chapitre 3 permet à la fois la création de cette base de données de polarisabilités – ici à partir du « *Big Set* » – et son transfert sur une ou plusieurs structures moléculaires d'intérêt. La méthodologie utilisée comporte sa part d'innovation (moyennes de tenseurs symétriques définis positifs) et a entrainé le portage dans *MoProViewer* de la librairie et des types atomiques ELMAM2. Cette nouvelle compatibilité implique l'implémentation de classes C++ générales qui – à l'avenir – pourront être utilisées pour des applications de la base de données sans rapport avec la polarisabilité.

La construction de cette base de données 'pELMAM3' à partir des 198 structures moléculaires du « *Big Set* » est un jalon important pour la continuation du développement de la librairie de densités électroniques expérimentales développée depuis presque 25 ans au CRM^2. En effet, l'utilisation de l'outil de polarisation ajoute aux atomes transférables une flexibilité supplémentaire : la sensibilité aux perturbations de l'environnement chimique (au travers du calcul EP/MM du champ électrique externe et des densités HCM d'induction dipolaire). Le modèle de polarisation pourrait encore gagner en sophistication, comme ébauché au chapitre 4, mais le précédent est là, et l'outil présenté permet déjà des applications prometteuses.

En effet, l'étude de cas biochimique présentée, traitant de l'interaction entre streptavidine et biotine, représente à la fois une démonstration technique de faisabilité mais également une preuve de la pertinence de la prise en compte des effets de la polarisation pour modéliser les influences électrostatiques au sein des systèmes protéine-ligand. L'utilisation de densités électroniques transférées-polarisées pourrait être à l'avenir utilisée en conjonction avec d'autres outils, par exemple pour ré-évaluer des conformations issues d'expériences d'amarrage moléculaire (via une fonction de *scoring*).

La topologie du champ de densité électronique (et celle, esquissée, du champ de potentiel électrostatique) offre un point de vue inédit sur les interactions électrostatiques au sein du site de fixation de la biotine. D'une part, l'étude topologique étaye les

conclusions réalisées sur la base des calculs d'énergie seuls. D'autre part, les liaisons hydrogène étudiées et comparées entre les mutants révèlent une diversité d'influences déterminantes pour l'interaction globale entre biotine et le récepteur streptavidine. Au sein d'un site de fixation concave, le ligand est en effet soumis à une variété d'interactions plus ou moins favorables, des contacts étroits se réalisant au profit d'autres, plus lâches.

D'un autre côté, l'étude topologique offre un regard neuf sur le processus de polarisation de la densité électronique. L'exemple le plus frappant étant la déplétion électronique au point critique des liaisons hydrogène. Une forte densité électronique au point critique est habituellement associée à d'intenses interaction, alors qu'ici, un processus strictement favorable énergétiquement, la polarisation de la densité électronique, provoque la fuite des électrons de l'espace interatomique. Mais plus encore, l'étude topologique des densités transférées et polarisées s'est révélée sensible à de très faibles variations de densité électronique, rencontrées par exemple dans les liaisons hydrogène les plus faibles du complexe.

Cette étude a été pensée dans la continuité méthodologique de celle réalisée sur le site actif de la FABP[12], où l'orientation de l'acide oléique, et des nombreuses molécules d'eau cristallographiques, sont reliées au potentiel électrostatique et aux champs électriques calculés à partir des densités ELMAM2. Cette approche de la description des interactions dans un contexte biochimique est encore novatrice et requiert de nouvelles études pour établir un cadre d'interprétation précis. En effet, les précédents sont rares mais les outils reportés ici aideront à remédier à cet état des connaissances.

Ici, les calculs d'énergie ne sont pas une fonction-objectif à atteindre (comme dans une approche *top-down*), mais un contrôle-qualité *a posteriori*. En effet, le présent travail utilise une approche *bottom-up* : l'évaluation de propriétés microscopiques (densité électronique, polarisabilités atomiques) permet la prédiction de propriétés macroscopiques (ici les enthalpies de liaison). Dans les approches de type champ de force, c'est habituellement la philosophie *top-down* qui prime : un ensemble de paramètres microscopique (charges, raideur des liaisons…) est ajusté pour reproduire des paramètres macroscopiques (températures de transition de phase, constante diélectrique…). La construction d'un champ de force complet par l'approche *bottom-up* est un sujet de recherche ouvert depuis de nombreuses années : le développement d'un

potentiel de van der Waals adapté aux densités électroniques transférables ELMAM2 est en effet en cours au laboratoire CRM^2. Ce potentiel de van der Waals serait l'étape suivante dans l'objectif de développer un potentiel d'interaction totale, quantitatif et basé sur des densités électroniques multipolaires, expérimentales et transférables.

Annexes

Formules diverses :

L'angle Θ (en radians) entre les vecteurs unitaires $\boldsymbol{u}$ et $\boldsymbol{v}$ est calculée comme suit, où "·" est le produit scalaire. Ce calcul est plus pertinent pour les ellipsoïdes allongées (*prolate*) où le premier vecteur propre $\boldsymbol{v}_1$ (associé avec λ_1) est distinct du second vecteur propre $\boldsymbol{v}_2$ (associé avec λ_2) :

$$\Theta(\boldsymbol{u}, \boldsymbol{v}) = |\cos^{-1}(\boldsymbol{u} \cdot \boldsymbol{v})| \qquad \textbf{(A1)}$$

L'ellipticité[33], où λ_1 est la plus grande valeur propre du tenseur de polarisabilité et λ_2 is la seconde valeur propre :

$$\varepsilon = \frac{\lambda_1}{\lambda_2} - 1 \quad \textbf{(A2)}$$

Le ratio transverse à longitudinal[33], où λ_1 est la plus grande valeur propre du tenseur de polarisabilité et λ_3 est la plus petite (la polarisabilité est toujours définie-positive) :

$$\eta = \frac{\lambda_1}{|\lambda_3|} \qquad \textbf{(A3)}$$

L'écart quadratique moyen (*RMSD*, *Root Mean Square Deviation*), où $\langle x_i \rangle$ symbolise la moyenne arithmétique sur les i variables x (les formules utilisant une moyenne corrigée pour un échantillon sont écrites explicitement) :

$$RMSD\left(E_{ELEC,TRANSFERRED}\right) = \sqrt{\langle |E_{i,ELEC,TRANSFERRED} - E_{i,ELEC,SAPT}|^2 \rangle} \qquad \textbf{(A4)}$$

L'erreur relative moyenne (*MRE*, *Mean Relative Error*) :

$$MRE\left(E_{ELEC,TRANSFERRED}\right) = \left\langle \frac{E_{i,ELEC,TRANSFERRED} - E_{i,ELEC,SAPT}}{E_{i,ELEC,SAPT}} \right\rangle \qquad \textbf{(A5)}$$

L'erreur relative moyenne absolue (*MARE*, *Mean Absolute Relative Error*) :

$$MARE\left(E_{ELEC,TRANSFERRED}\right) = \left\langle \left| \frac{E_{i,ELEC,TRANSFERRED} - E_{i,ELEC,SAPT}}{E_{i,ELEC,SAPT}} \right| \right\rangle \qquad \textbf{(A6)}$$

La Déviation Standard de l'Échantillon (*SSD*, *Sample Standard Deviation*), où *n* est la population de l'échantillon. Le *SSD* est supposé 0 pour les échantillons dont la population est *n* = 1 :

$$\sigma\left(E_{ELEC,TRANSFERRED}\right) = \sqrt{\frac{\sum_i\left(E_{i,ELEC,TRANSFERRED} - \langle E_{ELEC,TRANSFERRED}\rangle\right)^2}{n-1}} \quad \textbf{(A7)}$$

Le facteur R :

$$R\left(E_{ELEC,TRANSFERRED}\right) = \frac{\sum_i |E_{i,ELEC,TRANSFERRED} - E_{i,ELEC,SAPT}|}{\sum_i |E_{i,ELEC,SAPT}|} \quad \textbf{(A8)}$$

La formule de rotation de Rodrigues donne la matrice de rotation, pour un angle θ, autour d'un vecteur unitaire $\mathbf{v}$:

$$\mathbf{R}_{\mathbf{v}}(\theta) =$$

$$\begin{bmatrix} \cos(\theta) + \mathbf{v}_x^2[1-\cos(\theta)] & \mathbf{v}_x\mathbf{v}_y[1-\cos(\theta)] - \mathbf{v}_z\sin(\theta) & \mathbf{v}_x\mathbf{v}_z[1-\cos(\theta)] + \mathbf{v}_y\sin(\theta) \\ \mathbf{v}_x\mathbf{v}_y[1-\cos(\theta)] + \mathbf{v}_z\sin(\theta) & \cos(\theta) + \mathbf{v}_y^2[1-\cos(\theta)] & \mathbf{v}_y\mathbf{v}_z[1-\cos(\theta)] - \mathbf{v}_x\sin(\theta) \\ \mathbf{v}_x\mathbf{v}_z[1-\cos(\theta)] - \mathbf{v}_y\sin(\theta) & \mathbf{v}_y\mathbf{v}_z[1-\cos(\theta)] + \mathbf{v}_x\sin(\theta) & \cos(\theta) + \mathbf{v}_z^2[1-\cos(\theta)] \end{bmatrix}$$

(A9)

Table A1. Les 38 types atomiques ELMAM2 utilisés pour la description du set S66 (les atomes grisés sont discutés partie 1.7.f et au chapitre 3-bis).

Description	ELMAM2 name	Chaîne atomique ELMAM2	Nombre d'occurrences (monomères)	Molécules
C, éthyne (sp)	C209	Cc[h]h	2	Ethyne(2)
C, aromatique	C301	C1.5c1.5c1h	10	Benzene(6), Pyridine(3), Uracil(1)
C, N-aromatique	C306	Cnch	3	Pyridine(2), Uracil(1)
C, amide	C314	C2o1.5n1.5c	3	Acetamide(1), Peptide(1), Uracil(1)
C-ONN (sp2)	C315	Conn	1	Uracil(1)
C, carboxylique	C316	C1.5o1.5o1c	1	Acetic acid(1)
C, éthène (sp2)	C318	Cc[hh]hh	2	Ethene(2)
C, méthyle	C401	Cchhh	9	Acetamide(1), Acetic acid(1), Neopentane(4), Pentane(2), Peptide(1)
C, methylène	C402	Ccchh	3	Pentane(3)
C, methylene cycl.	C402T	Ccchh	5	Cyclopentane(5)
C-(CH3)4	C405T	Ccccc	1	Neopentane(1)
C, N-methyle	C406T	Cnhhh	2	Methylamine(1), Peptide(1)
C, O-methyle	Cohhh	Cohhh	1	Methanol(1)
H, aromatique	H101	Hc[cc]	10	Benzene(6), Pyridine(3), Uracil(1)
H, N-aromatique	H102	Hc[nx]	3	Pyridine(2), Uracil(1)
H, methyle	H103	Hc[chh]	27	Acetamide(3), Acetic acid(3), Neopentane(12), Pentane(6), Peptide(3)
H, methylène	H104	Hc[cch]	16	Cyclopentane(10), Pentane(6)
H, N-methyle	H106	Hc[nxh]	6	Methylamine(3), Peptide(3)
H, O-methyle	H108	Hc[oxx]	3	Methanol(3)
H, amine/amide	H111	Hn[ch]	4	Acetamide(2), Methylamine(2)
H, N-peptide	H112	Hn[cc]	3	Peptide(1), Uracil(2)
H, eau	H114T	Ho[h]	2	Water(2)
H, alcool	H115	Ho[c]	2	Acetic acid(1), Methanol(1)
H,éthène	H117	Hc[c(hh)h]	4	Ethene(4)
H, éthyne	H118	Hc[c(h)]	2	Ethyne(2)
N, aromatique	N201	Ncc	1	Pyridine(1)
N, amide	N301	Nchh	1	Acetamide(1)
N, peptide	N302	N1.5c1.5c1h	1	Peptide(1)
N, amide cyclique	N302T	N1.5c1.5c1h	2	Uracil(2)
N, amine	N306	Nc[hxx]hh	1	Methylamine(1)
O, amide	O102	O1.5c[1.5n(xx)1c(X)]	3	Acetamide(1), Peptide(1), Uracil(1)
O, C(NN)	O103T	Oc[nn]	1	Uracil(1)
O, carboxylate	O105	O2c[1o1c(X)]	1	Acetic acid(1)
O, eau	O201T	Ohh	1	Water(1)
O, H-carboxylique	O202	Oc[oc]h	1	Acetic acid(1)
O, alcool	O203	Oc[X]h	1	Methanol(1)

Table A2. Liste des 54 structures cristallographiques utilisées pour construire les densités électroniques de la banque ELMAM2.

Reproduit de Domagała et al., 2012[61], les références renvoient à celles tirées de cet article.

Abréviation	Nom, formule	Radiation, Température	Références
Ala	L-Alanine, $C_3H_7NO_2$	Mo, 23 K	Destro et al., 1988
Enképhaline	Leucine-Enképhaline Trihydrate, $C_{28}H_{37}N_5O_7$, 3 H_2O	Mo, 100 K	Wiest et al., 1994
Argphos	L-Arginine Phosphate hydrate, $C_6H_{15}N_4O_2^+$, $H_2PO_4^-$, H_2O	Mo, 130 K	Espinosa et al., 1996
Oxalic	Oxalic acid, $C_2H_2O_4$	Mo, 100 K	Martin et al., 1998
His	DL-Histidine, $C_6H_9N_3O_2$	Mo, 110 K	Coppens et al., 1999
AcTyr	N-acétyl-L-tyrosine éthyl ester hydrate, $C_{13}H_{19}NO_5$, H_2O	Mo, 110 K	Dahaoui et al., 1999
GlyThr	Glycyl-L-thréonine dihydrate, $C_6H_{12}N_2O_4$, 2 H_2O	Mo, 110 K	Benabicha et al., 2000
GlyAsp	Glycyl-aspartic acid dihydrate, $C_6H_{10}N_2O_5$, 2 H_2O	Mo, 123 K	Pichon-Pesme et al., 2000
TyrGlyGly	Tyrosyl-glycyl-glycine hydrate, $C_{13}H_{17}N_3O_5$, H_2O	Mo, 123 K	Pichon-Pesme et al., 2000
Panb	p-amino-p'-nitrobiphenyl, $C_{12}H_{10}N_2O_2$	Mo, 20 K	Volkov et al., 2000
Ureaphos	Urea - phosphoric acid, $CH_7N_2O_5P$	Mo, 100 K	Rodrigues et al., 2001
NTO	5-nitro-2,4-dihydro-3H-1,2,4-triazol-3-one, α-NTO, $C_2H_2N_4O_3$	Ag, 100 K	Zhurova et al., 2001
CycloProAla	cyclo-(D,L-Pro)$_2$-(L-Ala)$_4$ hydrate, $C_{22}H_{34}N_6O_6$, H_2O	Synchrotron, 0.5583 Å, 100K	Dittrich et al., 2002
Thiocoumarin 2	2H-chromène-2-thione, (2-thiocoumarine), C_9H_6OS	Mo, 90K	Munshi et al., 2002

HPPM	Hydrogèno-[(2,4-diaminopyrimidine-1-io) méthyl]-phosphonate hydrate, $C_5H_9N_4O_3P$, H_2O	Mo, 105K	Slouf et al., 2002
BIGH	Biguanidinium Dinitramide, $C_2H_8O_4N_8$	Mo, 90K	Zhurova et al., 2002
BIGH2	Biguanidinium Bis-Dinitramide, $C_2H_9O_8N_{11}$	Mo, 90K	Zhurova et al., 2002
Schiff1	Dianil de 2-hydroxy-5-methylisophth-aldéhyde, $C_{21}H_{18}O_1N_2$	Mo, 100 K	Dominiak et al., 2003
Schiff2	3,5-dinitro-*N*-salicylidènoéthylamine, $C_9H_9O_5N_3$	Mo, 100 K	Dominiak et al., 2003
Schiff3	3-nitro-N-salicylidènocyclohéxylamine, $C_{13}H_{16}O_3N_2$	Mo, 100 K	Dominiak et al., 2003
NAD	NAD+, β-Nicotinamide Adénine Dinucléotide, $C_{21}H_{27}N_7O_{14}P_2$, 4 H_2O	Mo, 100 K	Guillot et al., 2003
TFT	Tétrafluorotérephthalonitrile, $C_8N_2F_4$	Mo, 122.4K	Sørensen et al., 2003
Urée	Urée, CH_4N_2O	Synchrotron 0.5996 Å, 123K	Birkedal et al., 2004
Busulfan	1,4-butanediol-diméthylsulfonate, $C_6H_{14}O_6S_2$	Mo, 100K	Ghermani et al., 2004
Barba	2,4,6,8-tétraphénylbarbaralane, $C_{33}H_{26}$	Synchrotron 0.56 Å, 110K	Luger et al., 2004
Xylitol	Xylitol, $C_5H_{12}O_5$	Mo, 122.4 K	Madsen et al., 2004
FamotidineA	Famotidine (polymorphe A), $C_8H_{15}N_7O_2S_3$	Mo, 100 K	Overgaard et al., 2004
FamotidineB	Famotidine (polymorphe B), $C_8H_{15}N_7O_2S_3$	Mo, 100 K	Overgaard et al., 2004
Trp	L-tryptophane, acide formique, $C_{11}H_{12}N_2O_2$, CH_2O_2	Synchrotron 0.54 Å, 100K	Scheins et al., 2004
Coumarine	2H-chromène-2-one, $C_9H_6O_2$	Mo, 90K	Munshi et al., 2005
Thiocoumarine 1	1-thiocoumarine, C_9H_6OS	Mo, 90K	Munshi et al., 2005

Pentaerythritol	Pentaérythritol, $C_5H_{12}O_4$	Mo, 15 K	Zhurov et al., 2005
AccoumarinA	3-Acétylcoumarine, (polymorphe A), $C_{11}H_8O_3$	Mo, 90K	Munshi et al., 2006a
AccoumarinB	3-Acétylcoumarine, (polymorphe B), $C_{11}H_8O_3$	Mo, 90K	Munshi et al., 2006a
Cytosine_hyd	Cytosine hydrate, $C_4H_5N_3O$, H_2O	Mo, 90K	Munshi et al., 2006b
Salicylic	Acide salicylique, $C_7H_6O_3$	Mo, 90K	Munshi et al., 2006b
Thiosemicarb	1-formyl-3-thiosemicarbazide, $C_2H_5N_3OS$	Mo, 90K	Munshi et al., 2006c
Thiouracil	2-thiouracile, $C_4H_4N_2OS$	Mo, 90K	Munshi et al., 2006b
Stilbene	Diméthylstilbène, $C_{16}H_{16}$	non reporté, 16K	Ogawa et al., 2006
Estradiol	17β-œstradiol Urée, $C_{18}H_{24}O_2$, CH_4N_2O	Mo, 100 K	Parrish et al., 2006
Estrone	3-hydroxy-1,3,5-(10)-œstratrièn-17-one, CHO	Mo, 20 K	Zhurova et al., 2006
Diazapentalene	1,3,4-trinitro-7,8-diazapentalène, $C_6H_3N_5O_6$	Ag, 90 K	Chen et al., 2007
Sarcosine	Sarcosine, $C_3H_7NO_2$	Mo, 100 K	Dittrich et al., 2007
Alaproala	L-alanyl-L-prolyl-L-alanine hydrate, $C_{11}H_{19}N_3O_4$, H_2O	Synchrotron, 100K	Kalinowski et al., 2007
Ornithine	L-ornithine hydrochloride, $C_5H_{13}N_2O_2^+$, Cl^-	Mo, 100K	Dittrich et al., 2007
Thymidine	Thymidine, $C_{10}H_{14}N_2O_5$	Mo, 20K	Hübschle et al., 2008
Thiouronium	2-(indol-3-yl)-1,1,3,3-tétraméthylthiouronium nitrate, $C_{13}H_{18}N_3S^+$, NO_3^-	Mo, 110K	Lutz et al., 2008
FOX-7	1,1-diamino-2,2-dinitroéthylène, $C_2H_4N_4O_4$	Synchrotron, 0.503 Å, 100K	Meents et al., 2008
RhPCl	[Rh(C_7H_8)(PPh_3)Cl], $C_{25}H_{23}ClPRh$	Mo, 100K	Sparkes et al., 2008
Paracetamol	Paracetamol, $C_8H_9NO_2$	Mo, 100K	Bouhmaida et al., 2009

16Ni	[3,11-Bis(méthoxycarbonyl)-1,5,9,13-tétraazacyclohéxadéca-1,3,9,11-tétraènato-(2-)-κ^4N]nickel(II), $C_{16}H_{22}N_4O_4Ni$	Mo, 100K	Domagała et al., 2009
Fidarestat	Fidarestat, $C_{12}H_{10}N_3O_4F$	Synchrotron, 0.450 Å, 100K	Fournier et al., 2009
Coumcarb	Acide coumarin-3-carboxylique, $C_{10}H_6O_4$	Mo, 100K	Howard et al., 2009
Transcin	Acide trans-cinnamique, $C_9H_8O_2$	Mo, 100K	Howard et al., 2009

Table A3. Liste complète des 198 molécules dont les polarisabilités théoriques ont été calculées et partitionnées en contributions atomiques anisotropes dans le but de construire une base de polarisabilités atomiques transférables. La chiralité des acides aminés est omise si seule la série L (protéinogénique) est présente dans la structure. La colonne « Formule » recense les atomes inclus dans la détermination des polarisabilités théoriques. Les structures marquées « théoriques » sont retrouvées sur le site web de la BEGDB (www.begdb.com). ***** : code PDB de la structure dont les coordonnées de la molécule ont été extraites. ** : fournie par Emmanuel Aubert, optimisée par DFT (*Gaussian09*[84]), fonctionnelle : B3LYP-D3/6-611G(d,p). *** : fournies par Anna Krawczuk.

Nom	Formule	Code CSD* ou Référence	Set
Acétamide	C_2H_5NO	théorique	S66 (monomère)
Acide acétique	$C_2H_4O_2$	théorique	S66 (monomère)
Alanine	$C_3H_7NO_2$	***	A.A. isolés
Arginine	$C_6H_{14}N_4O_2$	***	A.A. isolés
Asparagine	$C_4H_8N_2O_3$	***	A.A. isolés
Acide aspartique	$C_4H_7NO_4$	***	A.A. isolés
Benzamidine	$C_7H_9N_2$	1TIO*	*supplémentaire*
Benzène	C_6H_6	théorique	S66 (monomère)
Biotine	$C_{10}H_{15}N_2O_3S$	3RY2*	*supplémentaire*
Cyclopentane	C_5H_{10}	théorique	S66 (monomère)
D-Cystéine	$C_3H_8NO_2S$	***	A.A. isolés
Diméthylether	C_2H_6O	ABEWAG	*supplémentaire*
Éthène	C_2H_4	théorique	S66 (monomère)
Éthyne	C_2H_2	théorique	S66 (monomère)
F2,25	$C_{13}H_9NOF_2$	théorique**	*supplémentaire*
Acide glutamique	$C_5H_9NO_4$	***	A.A. isolés
Glutamine	$C_5H_{10}N_2O_3$	***	A.A. isolés
Glycine	$C_2H_5NO_2$	***	A.A. isolés
Histidine	$C_6H_9N_3O_2$	***	A.A. isolés

Isoleucine	$C_6H_{13}NO_2$	***	A.A. isolés
Leucine	$C_6H_{13}NO_2$	***	A.A. isolés
Lysinium	$C_6H_{15}N_2O_2$	***	A.A. isolés
N-méthyl-acétamide	C_3H_7NO	théorique	S66 (monomère)
Méthylamine	CH_5N	théorique	S66 (monomère)
Méthanol	CH_4O	théorique	S66 (monomère)
Méthionine	$C_5H_{11}NO_2S$	***	A.A. isolés
Néopentane	C_5H_{12}	théorique	S66 (monomère)
Pentane	C_5H_{12}	théorique	S66 (monomère)
Phénylalanine	$C_9H_{11}NO_2$	***	A.A. isolés
Proline	$C_5H_9NO_2$	***	A.A. isolés
Pyridine	C_5H_5N	théorique	S66 (monomère)
Sérine	$C_3H_7NO_3$	***	A.A. isolés
Thréonine	$C_4H_9NO_3$	***	A.A. isolés
Tryptophane	$C_{11}H_{12}N_2O_2$	***	A.A. isolés
Tyrosine	$C_9H_{11}NO_3$	***	A.A. isolés
Uracile	$C_4H_4N_2O_2$	théorique	S66 (monomère)
Valine	$C_5H_{11}NO_2$	***	A.A. isolés
Eau	H_2O	théorique	S66 (monomère)
Eau ··· Eau (01)	H_2O, H_2O	théorique	S66 (électrostatique)
Eau ··· Méthanol (02)	H_2O, CH_4O	théorique	S66 (électrostatique)
Eau ··· Méthylamine (03)	H_2O, CH_5N	théorique	S66 (électrostatique)
Eau ··· *N*-méthyl-acétamide (04)	H_2O, C_3H_7NO	théorique	S66 (électrostatique)
Méthanol ··· Méthanol (05)	CH_4O, CH_4O	théorique	S66 (électrostatique)
Méthanol ··· Méthylamine (06)	CH_4O, CH_5N	théorique	S66 (électrostatique)
Méthanol ··· *N*-méthyl-acétamide (07)	CH_4O, C_3H_7NO	théorique	S66 (électrostatique)
Méthanol ··· Eau (08)	CH_4O, H_2O	théorique	S66 (électrostatique)
Méthylamine ··· Méthanol (09)	CH_5N, CH_4O	théorique	S66 (mixte)
Méthylamine ··· Méthylamine (10)	CH_5N, CH_5N	théorique	S66 (mixte)
Méthylamine ··· *N*-méthyl-acétamide (11)	CH_5N, C_3H_7NO	théorique	S66 (mixte)
Méthylamine ··· Eau (12)	CH_5N, H_2O	théorique	S66 (électrostatique)
N-méthyl-acétamide ··· Méthanol (13)	C_3H_7NO, CH_4O	théorique	S66 (électrostatique)
N-méthyl-acétamide ··· Méthylamine (14)	C_3H_7NO, CH_5N	théorique	S66 (électrostatique)
N-méthyl-acétamide ··· *N*-méthyl-acétamide (15)	C_3H_7NO, C_3H_7NO	théorique	S66 (électrostatique)
N-méthyl-acétamide ··· Eau (16)	C_3H_7NO, H_2O	théorique	S66 (électrostatique)
Uracile ··· Uracile (BP, 17)	$C_4H_4N_2O_2$, $C_4H_4N_2O_2$	théorique	S66 (électrostatique)
Eau ··· Pyridine (18)	H_2O, C_5H_5N	théorique	S66 (électrostatique)
Méthanol ··· Pyridine (19)	CH_4O, C_5H_5N	théorique	S66 (électrostatique)
Acide acétique ··· Acide acétique (20)	$C_2H_4O_2$, $C_2H_4O_2$	théorique	S66 (électrostatique)
Acétamide ··· Acétamide (21)	C_2H_5NO, C_2H_5NO	théorique	S66 (électrostatique)
Acide acétique ··· Uracile (22)	$C_2H_4O_2$, $C_4H_4N_2O_2$	théorique	S66 (électrostatique)
Acétamide ··· Uracile (23)	C_2H_5NO, $C_4H_4N_2O_2$	théorique	S66 (électrostatique)
Benzène ··· Benzène (π-π, 24)	C_6H_6, C_6H_6	théorique	S66 (dispersion)
Uracile ··· Uracile (π-π, 26)	$C_4H_4N_2O_2$, $C_4H_4N_2O_2$	théorique	S66 (mixte)
Benzène ··· Pyridine (π-π, 27)	C_6H_6, C_5H_5N	théorique	S66 (dispersion)
Benzène ··· Uracile (π-π, 28)	C_6H_6, $C_4H_4N_2O_2$	théorique	S66 (dispersion)

Benzène ··· Éthène (30)	C_6H_6, C_2H_4	théorique	S66 (dispersion)
Uracile ··· Éthène (31)	$C_4H_4N_2O_2$, C_2H_4	théorique	S66 (dispersion)
Uracile ··· Éthyne (32)	$C_4H_4N_2O_2$, C_2H_2	théorique	S66 (mixte)
Pyridine ··· Éthène (33)	C_5H_5N, C_2H_4	théorique	S66 (dispersion)
Pentane ··· Pentane (34)	C_5H_{12}, C_5H_{12}	théorique	S66 (dispersion)
Néopentane ··· Pentane (35)	C_5H_{12}, C_5H_{12}	théorique	S66 (dispersion)
Néopentane ··· Néopentane (36)	C_5H_{12}, C_5H_{12}	théorique	S66 (dispersion)
Cyclopentane ··· Néopentane (37)	C_5H_{10}, C_5H_{12}	théorique	S66 (dispersion)
Cyclopentane ··· Cyclopentane (38)	C_5H_{10}, C_5H_{10}	théorique	S66 (dispersion)
Benzène ··· Néopentane (40)	C_6H_6, C_5H_{12}	théorique	S66 (dispersion)
Uracile ··· Pentane (41)	$C_4H_4N_2O_2$, C_5H_{12}	théorique	S66 (dispersion)
Uracile ··· Néopentane (43)	$C_4H_4N_2O_2$, C_5H_{12}	théorique	S66 (dispersion)
Éthène ··· Pentane (44)	C_2H_4, C_5H_{12}	théorique	S66 (dispersion)
Éthyne ··· Pentane (45)	C_2H_2, C_5H_{12}	théorique	S66 (dispersion)
Benzène ··· Benzène (TS, 47)	C_6H_6, C_6H_6	théorique	S66 (dispersion)
Pyridine ··· Pyridine (TS, 48)	C_5H_5N, C_5H_5N	théorique	S66 (dispersion)
Benzène ··· Éthyne (50)	C_6H_6, C_2H_2	théorique	S66 (mixte)
Éthyne ··· Éthyne (TS, 51)	C_2H_2, C_2H_2	théorique	S66 (mixte)
Benzène ··· Eau (54)	C_6H_6, H_2O	théorique	S66 (mixte)
Benzène ··· Méthanol (55)	C_6H_6, CH_4O	théorique	S66 (mixte)
Benzène ··· *N*-méthyl-acétamide (57)	C_6H_6, C_3H_7NO	théorique	S66 (mixte)
Pyridine ··· Pyridine (CH···N, 58)	C_5H_5N, C_5H_5N	théorique	S66 (mixte)
Éthyne ··· Eau (59)	C_2H_2, H_2O	théorique	S66 (électrostatique)
Éthyne ··· Acide acétique (60)	C_2H_2, $C_2H_4O_2$	théorique	S66 (électrostatique)
Pentane ··· Acide acétique (61)	C_5H_{12} , $C_2H_4O_2$	théorique	S66 (dispersion)
Pentane ··· Acétamide (62)	C_5H_{12}, C_2H_5NO	théorique	S66 (dispersion)
Benzène ··· Acide acétique (63)	C_6H_6, $C_2H_4O_2$	théorique	S66 (dispersion)
N-méthyl-acétamide ··· Éthène (64)	C_3H_7NO, C_2H_4	théorique	S66 (mixte)
Pyridine ··· Éthyne (65)	C_5H_5N, C_2H_2	théorique	S66 (électrostatique)
Méthylamine ··· Pyridine (66)	CH_5N, C_5H_5N	théorique	S66 (mixte)
N-acétyl-tyrosine éthyl ester, hydrate	$C_{13}H_{17}NO_4$, H_2O	ATYREE03	A.A. isolés
Ala-Ala, chlorure	$C_6H_{12}N_2O_3$, Cl^-	ALALHC	Dipeptides
D-Ala-L-Met	$C_7H_{16}N_2O_3S$	ALAMET01	Dipeptides
L-Ala-D-Met	$C_7H_{16}N_2O_3S$	ALAMET10	Dipeptides
Ala-Met	$C_7H_{16}N_2O_3S$	ALAMET11	Dipeptides
Ile-Ala	$C_9H_{18}N_2O_3$	AQARUF01	Dipeptides
Ile-Ala	$C_9H_{18}N_2O_3$	AQARUF	Dipeptides
Gly-Glu, dihydrate	$C_7H_{12}N_2O_5$, 2 H_2O	BEVXEF01	Dipeptides
Ala-Tyr, ammonium	$C_{12}H_{15}N_2O_4$, NH_4^+	BIHWEV	Dipeptides
Leu-Glu	$C_{11}H_{20}N_2O_5$	BOFZOL	Dipeptides
Ala-Asp	$C_7H_{12}N_2O_5$	BURLIJ	Dipeptides
Glu-Asp	$C_9H_{14}N_2O_7$	BURLOP	Dipeptides
Tyr-Val, hydrate	$C_{14}H_{22}N_2O_4$, H_2O	CIHNUC	Dipeptides
Val-Glu	$C_{10}H_{18}N_2O_5$	CIJGUX	Dipeptides
Leu-Phe, propan-2-ol, hydrate	$C_{15}H_{24}N_2O_3$, 2 C_3H_8O, H_2O	COCGOQ	Dipeptides
Leu-Phe, propan-2-ol, hydrate	$C_{15}H_{24}N_2O_3$, 2 C_3H_8O, H_2O	COCGOQ01	Dipeptides
Ala-Thr	$C_7H_{14}N_2O_4$	EWOVAN	Dipeptides
Asn-Val, hydrate (molécule 1)	$C_9H_{17}N_3O_4$, 4 H_2O	FOXBAW	Dipeptides
Asn-Val, hydrate (molécule 3)	$C_9H_{17}N_3O_4$, 4 H_2O	FOXBAW	Dipeptides

Trp-Gly, hydrate	$C_{13}H_{15}N_3O_3$, H_2O	FULGEY	Dipeptides
Asp-Ala	$C_7H_{12}N_2O_5$	FUMTAI	Dipeptides
Gly-Phe	$C_{11}H_{14}N_2O_3$	GLDLPA	Dipeptides
Gly-D-Asn	$C_6H_{11}N_3O_4$	GLLASP	Dipeptides
Gly-Thr, hydrate (molécule 1)	$C_6H_{12}N_2O_4$, 2 H_2O	GLTHRE	Dipeptides
Gly-Thr, hydrate (molécule 2)	$C_6H_{12}N_2O_4$, 2 H_2O	GLTHRE	Dipeptides
Gly-Trp, dihydrate	$C_{13}H_{15}N_3O_3$, 2 H_2O	GLTRDH	Dipeptides
Gly-Trp, dihydrate	$C_{13}H_{15}N_3O_3$, 2 H_2O	GLTRDH01	Dipeptides
Glu-Gly	$C_7H_{12}N_2O_5$	GLUGLY	Dipeptides
Gly-Ala, chlorure	$C_5H_{10}N_2O_3$, Cl^-	GLYALA	Dipeptides
Gly-Ala	$C_5H_{10}N_2O_3$	GLYALA	Dipeptides
Gly-Ala	$C_5H_{10}N_2O_3$	GLYALB	Dipeptides
Gly-Ala	$C_5H_{10}N_2O_3$	GLYALB01	Dipeptides
Gly-Ala	$C_5H_{10}N_2O_3$	GLYALB02	Dipeptides
Gly-Ala	$C_5H_{10}N_2O_3$	GLYALB03	Dipeptides
Gly-Ala	$C_5H_{10}N_2O_3$	GLYALB04	Dipeptides
Gly-Ala	$C_5H_{10}N_2O_3$	GLYALB05	Dipeptides
Gly-Ala	$C_5H_{10}N_2O_3$	GLYALB06	Dipeptides
Gly-Ala	$C_5H_{10}N_2O_3$	GLYALB07	Dipeptides
Gly-Ala	$C_5H_{10}N_2O_3$	GLYALB08	Dipeptides
Gly-Ala	$C_5H_{10}N_2O_3$	GLYALB09	Dipeptides
Gly-Ala	$C_5H_{10}N_2O_3$	GLYALB10	Dipeptides
Gly-Ala	$C_5H_{10}N_2O_3$	GLYDLA	Dipeptides
Gly-Gly, phosphate, hydrate	$C_4H_8N_2O_3$, HPO_4^{2-}, H_2O	GLYGLP	Dipeptides
Gly-Gly	$C_4H_8N_2O_3$	GLYGLY01	Dipeptides
Gly-Gly (octodeutérée)	$C_4D_8N_2O_3$	GLYGLY10	Dipeptides
Gly-Gly (octodeutérée)	$C_4D_8N_2O_3$	GLYGLY12	Dipeptides
Gly-Gly	$C_4H_8N_2O_3$	GLYGLY14	Dipeptides
Gly-Gly	$C_4H_8N_2O_3$	GLYGLY15	Dipeptides
Gly-Gly	$C_4H_8N_2O_3$	GLYGLY16	Dipeptides
Gly-Gly	$C_4H_8N_2O_3$	GLYGLY17	Dipeptides
Gly-Gly	$C_4H_8N_2O_3$	GLYGLY18	Dipeptides
Gly-Thr, dihydrate	$C_6H_{12}N_2O_4$, 2 H_2O	GLYTRE	Dipeptides
Gly-Thr, dihydrate	$C_6H_{12}N_2O_4$, 2 H_2O	GLYTRE05	Dipeptides
Gly-Thr, dihydrate	$C_6H_{12}N_2O_4$, 2 H_2O	GLYTRE03	Dipeptides
Ala-Ile	$C_9H_{18}N_2O_3$	HEGLOV	Dipeptides
Val-Val	$C_{10}H_{20}N_2O_3$	HEGLUB	Dipeptides
Ile-Val	$C_{11}H_{22}N_2O_3$	HEGMAI	Dipeptides
Leu-Leu, 2-méthyl-propan-2-ol (molécule 1)	$C_{12}H_{24}N_2O_3$, 2 $C_4H_{10}O$	HIQWAF	Dipeptides
Leu-Leu, hydrate (molécule 1)	$C_{12}H_{24}N_2O_3$, 2 H_2O	IDUZOW	Dipeptides
Phe-Pro, hydrate	$C_{14}H_{18}N_2O_3$, H_2O	JADVAL	Dipeptides
Leu-Ser	$C_9H_{18}N_2O_4$	JAZBOC	Dipeptides
Leu-Ser	$C_9H_{18}N_2O_4$	JAZBOC01	Dipeptides
Leu-Ser	$C_9H_{18}N_2O_4$	JAZBOC02	Dipeptides
Leu-Tyr	$C_{15}H_{22}N_2O_4$	JUKMEH	Dipeptides
His-Leu	$C_{12}H_{20}N_4O_3$	JUKMOR	Dipeptides
Ser-Asn, hydrate	$C_7H_{13}N_3O_5$, H_2O	KIXBID	Dipeptides
Ser-Tyr, hydrate	$C_{12}H_{16}N_2O_5$, H_2O	KIXBOJ	Dipeptides
Trp-Ser, hydrate	$C_{14}H_{17}N_3O_4$, H_2O	KIXBUP	Dipeptides

Thr-Ala	$C_7H_{14}N_2O_4$	MAPKOE	Dipeptides
Ile-Ser, hydrate	$C_9H_{18}N_2O_4$, H_2O	MAZXUH	Dipeptides
Phe-Ser	$C_{12}H_{16}N_2O_4$	MAZYAO	Dipeptides
Met-Ser, hydrate (molécule 2)	$C_8H_{16}N_2O_4S$, 2 H_2O	MAZYES	Dipeptides
Met-Met	$C_{10}H_{20}N_2O_3S_2$	METMET	Dipeptides
Val-Phe, dihydrate	$C_{14}H_{20}N_2O_3$, 2 H_2O	MOBYAD	Dipeptides
His-Gly, hémi-hydrate (molécule 1)	$C_8H_{13}N_4O_3$, H_2O	NAYJOM	Dipeptides
His-Gly, hémi-hydrate (molécule 2)	$C_8H_{13}N_4O_3$, H_2O	NAYJOM	Dipeptides
Gly-Gly, sesquihydrate	$C_4H_8N_2O_3$, 2 H_2O	POTPET	Dipeptides
Gly-Gly, sesquihydrate	$C_4H_8N_2O_3$, 2 H_2O	POTPET01	Dipeptides
Gly-Gly, sesquihydrate	$C_4H_8N_2O_3$, 2 H_2O	POTPET02	Dipeptides
Gly-Gly, sesquihydrate	$C_4H_8N_2O_3$, 2 H_2O	POTPET03	Dipeptides
Phe-Ala, dihydrate	$C12H_{16}N_2O_3$, 2 H_2O	QIMBUJ	Dipeptides
Leu-Ala, tétrahydrate	$C_9H_{18}N_2O_3$, 4 H_2O	RAVMOQ	Dipeptides
Gly-His, dihydrate	$C_8H_{13}N_4O_3$, 2 H_2O	RAVZEU	Dipeptides
Leu-His, hydrate	$C_{12}H_{20}N_4O_3$, H_2O	RAVZIY	Dipeptides
His-Glu	$C_{11}H_{16}N_4O_5$	RAVZUK	Dipeptides
Gly-Pro, hémihydrate (molécule 1)	$C_7H_{12}N_2O_3$, H_2O	SEHGES	Dipeptides
Gly-Pro, hémihydrate (molécule 2)	$C_7H_{12}N_2O_3$, H_2O	SEHGES	Dipeptides
Ser-Gly	$C_5H_{10}N_2O_4$	SERGLY	Dipeptides
His-Ala, dihydrate	$C_9H_{14}N_4O_3$, 2 H_2O	TEKNAY	Dipeptides
Val-Gln	$C_{10}H_{19}N_3O_4$	TIPTOB	Dipeptides
Glu-Val	$C_{10}H_{18}N_2O_5$	TIPTUH	Dipeptides
Gly-Gly-Gly	$C_6H_{11}N_3O_4$	TGLYCY10	Tripeptides
2 Gly-Gly-Gly	$C_6H_{11}N_3O_4$	TGLYCY10	Tripeptides
Tyr-Gly-Gly, hydrate	$C_{13}H_{17}N_3O_5$, H_2O	LTYRGG01	Tripeptides
His-Gly (molécule 2 uniquement)	$C_8H_{13}N_4O_3$	WASPOV	Dipeptides
Asp-Phe, His-Gly, hydrate	$C_{13}H_{15}N_2O_5$, $C_8H_{13}N_4O_3$, H_2O	WASPOV	Dipeptides
Val-Ala	$C_8H_{16}N_2O_3$	WIRYEB	Dipeptides
Val-Ala	$C_8H_{16}N_2O_3$	WIRYEB02	Dipeptides
Phe-Val	$C_{14}H_{20}N_2O_3$	XEGNAY	Dipeptides
Ala-Val	$C_8H_{16}N_2O_3$	XUDVOH	Dipeptides
Ala-Val	$C_8H_{16}N_2O_3$	XUDVOH04	Dipeptides
Ala-Val	$C_8H_{16}N_2O_3$	XUDVOH05	Dipeptides
Ala-Pro-Ala, hydrate	$C_{11}H_{19}N_3O_4$, H_2O	SILTUD01	Tripeptides

Encart A1. Exemple de fichier d'entrée du module de polarisation « *Beluga* »

```
!Beluga, a polarization tool for MoProViewer, by Leduc T., Jelsch C.,
Iordache C. & Guillot B.
!Comments begins by '!'
!
PATH  0     C:\...\01_2701_01WaterWater100_02rot_theorPolMon.par00
PATH  1     C:\...\02_2709_02WaterMeOH100_03neu_theorPolMon.par00
PATH  2     C:\...\03_2717_03WaterMeNH2100_03trfrotneu_theorPolMon.par00
PATH  3     C:\...\04WaterPeptide100_02_theorPolMon.par00
PATH  4     C:\...\05MeOHMeOH100_03neu_theorPolMon.par00
PATH  5     C:\...\06_MeOHMeNH2100_06neu_theorPolMon.par00
PATH  6     C:\...\07_MeOHPeptide100_03neu_theorPolMon.par00
PATH  7     C:\...\08_MeOHWater100_06neu_theorPolMon.par00
PATH  8     C:\...\12_MeNH2Water100_03neu_theorPolMon.par00
PATH  9     C:\...\13_PeptideMeOH100_03neu_theorPolMon.par00
PATH  10    C:\...\14_PeptideMeNH2100_03neu_theorPolMon.par00
PATH  11    C:\...\15_PeptidePeptide100_02_theorPolMon.par00
CYCLE 0
!The following 20 lines divide each .par00 file in two molecules
(residues) that polarize each other.
OBJSELE     0     RESIDUE NUM 0
ACTSELE     0     RESIDUE NUM 1
OBJSELE     1     RESIDUE NUM 0
ACTSELE     1     RESIDUE NUM 1
OBJSELE     2     RESIDUE NUM 0
ACTSELE     2     RESIDUE NUM 1
OBJSELE     3     RESIDUE NUM 0
ACTSELE     3     RESIDUE NUM 1
OBJSELE     4     RESIDUE NUM 0
ACTSELE     4     RESIDUE NUM 1
OBJSELE     5     RESIDUE NUM 0
ACTSELE     5     RESIDUE NUM 1
OBJSELE     6     RESIDUE NUM 0
ACTSELE     6     RESIDUE NUM 1
OBJSELE     7     RESIDUE NUM 0
ACTSELE     7     RESIDUE NUM 1
OBJSELE     8     RESIDUE NUM 0
ACTSELE     8     RESIDUE NUM 1
OBJSELE     9     RESIDUE NUM 0
ACTSELE     9     RESIDUE NUM 1
OBJSELE     10    RESIDUE NUM 0
ACTSELE     10    RESIDUE NUM 1
OBJSELE     11    RESIDUE NUM 0
ACTSELE     11    RESIDUE NUM 1
SETANISOPOL UIJ
DAMP  NOTIJ
DAMPSTR     FIXED 1.0
EFIELD      PT
SOLV  FULLINVMAT
```

Exemple de fichier d'entrée du module de polarisation « *Beluga* » (suite)

```
!The following block repeats for each polarization cycle, from here...
READALLMOL
POLARIZE
INVSELE
POLARIZE
WRITEPARAM
SAVEMOL
!...to here.
!The previous commands are only executed if the calculation is not
converged.
!10 repeats of this block are enougth to obtain dipole convergence on
every dimer tested.
ENDCYCLE
ENDBELUGA
```

Liste des matrices de rotation par élément de symétrie ELMAM2 (omettant l'identité pour toutes les opérations) :

Identité ('I' et '-') :

aucune

Inversion ('i') :

$$\begin{bmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \end{bmatrix}$$

Miroir simple *x* ('mx') :

$$\begin{bmatrix} -1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Miroir simple *y* ('my') :

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Miroir simple *z* ('mz') :

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -1 \end{bmatrix}$$

Axe 2, *x* ('2x') :

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \end{bmatrix}$$

Axe 2, *y* ('2y') :

$$\begin{bmatrix} -1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -1 \end{bmatrix}$$

Axe 2, *z* ('2z') :

$$\begin{bmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Axe 3, *x* ('3x') :

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & -0.5 & \sin(\frac{2\pi}{3}) \\ 0 & -\sin(\frac{2\pi}{3}) & -0.5 \end{bmatrix} \qquad \begin{bmatrix} 1 & 0 & 0 \\ 0 & -0.5 & -\sin(\frac{2\pi}{3}) \\ 0 & \sin(\frac{2\pi}{3}) & -0.5 \end{bmatrix}$$

Axe 3, *y* (‘3y’) :

$$\begin{bmatrix} -0.5 & 0 & \sin(\frac{2\pi}{3}) \\ 0 & 1 & 0 \\ -\sin(\frac{2\pi}{3}) & 0 & -0.5 \end{bmatrix} \qquad \begin{bmatrix} -0.5 & 0 & -\sin(\frac{2\pi}{3}) \\ 0 & 1 & 0 \\ \sin(\frac{2\pi}{3}) & 0 & -0.5 \end{bmatrix}$$

Axe 3, *z* (‘3z’) :

$$\begin{bmatrix} -0.5 & \sin(\frac{2\pi}{3}) & 0 \\ -\sin(\frac{2\pi}{3}) & -0.5 & 0 \\ 0 & 0 & 1 \end{bmatrix} \qquad \begin{bmatrix} -0.5 & -\sin(\frac{2\pi}{3}) & 0 \\ \sin(\frac{2\pi}{3}) & -0.5 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Miroir, *y* et Axe 3, *z* (‘3m’) :

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \qquad \begin{bmatrix} -0.5 & \sin\left(\frac{2\pi}{3}\right) & 0 \\ -\sin\left(\frac{2\pi}{3}\right) & -0.5 & 0 \\ 0 & 0 & 1 \end{bmatrix} \qquad \begin{bmatrix} -0.5 & -\sin\left(\frac{2\pi}{3}\right) & 0 \\ \sin\left(\frac{2\pi}{3}\right) & -0.5 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

$$\begin{bmatrix} -0.5 & -\sin(\frac{2\pi}{3}) & 0 \\ -\sin(\frac{2\pi}{3}) & 0.5 & 0 \\ 0 & 0 & 1 \end{bmatrix} \qquad \begin{bmatrix} -0.5 & \sin(\frac{2\pi}{3}) & 0 \\ \sin(\frac{2\pi}{3}) & 0.5 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Miroirs *x*, *y* et *z* (‘mmm’) :

$$\begin{bmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \end{bmatrix} \quad \begin{bmatrix} -1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad \begin{bmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -1 \end{bmatrix}$$

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \end{bmatrix} \quad \begin{bmatrix} -1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -1 \end{bmatrix} \quad \begin{bmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Axe 2, *x*, *y* et *z* (‘222’ ou ‘2x2y’ ou ‘2x2z’ ou ‘2y2z’) :

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \end{bmatrix} \qquad \begin{bmatrix} -1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -1 \end{bmatrix} \qquad \begin{bmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Miroirs *x* et *y* (‘mxmy’) :

$$\begin{bmatrix} -1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \qquad \begin{bmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \qquad \begin{bmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Miroirs x et z (‘mxmz’) :

$$\begin{bmatrix} -1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \qquad \begin{bmatrix} -1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -1 \end{bmatrix} \qquad \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -1 \end{bmatrix}$$

Miroirs y, et z (‘mymz’) :

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \end{bmatrix} \qquad \begin{bmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \qquad \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -1 \end{bmatrix}$$

Miroirs x, y, xy, $x\bar{y}$ (‘4m’) :

$$\begin{bmatrix} -1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \qquad \begin{bmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \qquad \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \qquad \begin{bmatrix} 0 & -1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

$$\begin{bmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \qquad \begin{bmatrix} 0 & 1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \qquad \begin{bmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Axe 4, z (‘4z’) :

$$\begin{bmatrix} 0 & 1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \qquad \begin{bmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \qquad \begin{bmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Miroirs, y et z plus Axe 3, z (‘3mmz’) :

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \qquad \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -1 \end{bmatrix} \qquad \begin{bmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \end{bmatrix}$$

$$\begin{bmatrix} -0.5 & \sin\left(\frac{2\pi}{3}\right) & 0 \\ -\sin\left(\frac{2\pi}{3}\right) & -0.5 & 0 \\ 0 & 0 & 1 \end{bmatrix} \qquad \begin{bmatrix} -0.5 & -\sin\left(\frac{2\pi}{3}\right) & 0 \\ \sin\left(\frac{2\pi}{3}\right) & -0.5 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

$$\begin{bmatrix} -0.5 & -\sin(\frac{2\pi}{3}) & 0 \\ \sin(\frac{2\pi}{3}) & -0.5 & 0 \\ 0 & 0 & -1 \end{bmatrix} \qquad \begin{bmatrix} -0.5 & \sin(\frac{2\pi}{3}) & 0 \\ -\sin(\frac{2\pi}{3}) & -0.5 & 0 \\ 0 & 0 & -1 \end{bmatrix}$$

$$\begin{bmatrix} -0.5 & -\sin(\frac{2\pi}{3}) & 0 \\ -\sin(\frac{2\pi}{3}) & 0.5 & 0 \\ 0 & 0 & 1 \end{bmatrix} \qquad \begin{bmatrix} -0.5 & -\sin(\frac{2\pi}{3}) & 0 \\ -\sin(\frac{2\pi}{3}) & 0.5 & 0 \\ 0 & 0 & -1 \end{bmatrix}$$

$$\begin{bmatrix} -0.5 & \sin(\frac{2\pi}{3}) & 0 \\ \sin(\frac{2\pi}{3}) & 0.5 & 0 \\ 0 & 0 & 1 \end{bmatrix} \qquad \begin{bmatrix} -0.5 & \sin(\frac{2\pi}{3}) & 0 \\ \sin(\frac{2\pi}{3}) & 0.5 & 0 \\ 0 & 0 & -1 \end{bmatrix}$$

Tétragonal *43m* (‘-43m’)

$$\begin{bmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad \begin{bmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad \begin{bmatrix} 0 & 1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

$$\begin{bmatrix} 0 & -1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad \begin{bmatrix} -1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad \begin{bmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

$$\begin{bmatrix} 0 & 0 & -1 \\ 0 & 1 & 0 \\ -1 & 0 & 0 \end{bmatrix} \quad \begin{bmatrix} 0 & 0 & -1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \quad \begin{bmatrix} 0 & 0 & -1 \\ 0 & -1 & 0 \\ 1 & 0 & 0 \end{bmatrix} \quad \begin{bmatrix} 0 & 0 & -1 \\ -1 & 0 & 0 \\ 0 & -1 & 0 \end{bmatrix}$$

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix} \quad \begin{bmatrix} 0 & -1 & 0 \\ 0 & 0 & -1 \\ 1 & 0 & 0 \end{bmatrix} \quad \begin{bmatrix} -1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & -1 & 0 \end{bmatrix} \quad \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -1 & 0 & 0 \end{bmatrix}$$

$$\begin{bmatrix} 0 & 0 & -1 \\ -1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \quad \begin{bmatrix} 0 & 0 & -1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix} \quad \begin{bmatrix} 0 & 0 & -1 \\ 1 & 0 & 0 \\ 0 & -1 & 0 \end{bmatrix} \quad \begin{bmatrix} 0 & 0 & -1 \\ 0 & -1 & 0 \\ -1 & 0 & 0 \end{bmatrix}$$

$$\begin{bmatrix} 0 & -1 & 0 \\ 0 & 0 & 1 \\ -1 & 0 & 0 \end{bmatrix} \quad \begin{bmatrix} -1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix} \quad \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix} \quad \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & -1 & 0 \end{bmatrix}$$

Table A4. Couples ITC – structures PDB de la base de données SILICA.

Code PDB	Description	ΔH^0	ΔG^0	Références (ITC)
3CCZ	HMG-CoA réductase – 5HI	-61.03	-42.22	Sarver et al., J. Med. Chem. 2008, 51, 3804–3813
3CCW	HMG-CoA réductase – 4HI	-73.99	-45.56	
3CD0	HMG-CoA réductase – 6HI	-74.06	-43.93	
3CCT	HMG-CoA réductase – 3HI	-48.91	-43.89	
2Q6B	HMG-CoA réductase – HR2	-33.02	-44.31	
3CD5	HMG-CoA réductase – 7HI	-34.73	-45.61	
2O9A	Acetate Operon Regulator - Pyruvate	-5.73	30.14	Lorca et al., J. Biol. Chem. 2006, 282, 22, 16476-16491
2IKJ	Aldose réductase – « 393 »	-79.1	-42.2	Steuber et al., J. Mol. Biol. 368: 618-38 (2007)
2IKI	Aldose réductase – « 388 »	-62.1	-42.7	
2IKH	Aldose réductase – LIT	-8.7	-31.3	
2IKG	Aldose réductase – BTO	-25.6	-35.4	
2AGT	Aldose réductase – Fiderestat	-68.6	-38.9	Petrova et al., J. Med. Chem. 2005, 48, 18, 5659-5665
1PWM	Aldose réductase (mutant) – Fiderestat	-75.5	-46.7	
1EF3	Aldose réductase – Fiderestat			
1AH0	Aldose réductase – SBI	-51.8	-37.9	
1Z8A	Aldose réductase – 62P	n.d	-34.5	Steuber et al., 2006, J. Mol. Biol., 356, 45-56
1Z89				
1BIB	BirA – Biotine	-81.09	-41.13	- Brown & Beckett, Biochemistry, 44, 8, 2005 - Naganathan & Beckett, J. Mol. Biol. (2007) 373, 96–111,
1HXD				
2EWN	BirA – Biotinol-5'-AMP	-49.32	-41.38	
1GFW	Caspase-3 - Sulfonamide	-45.17	-42.64	Alabaugh et al., Biochemistry 2007, 46, 9462-9471
2Q6N	CytP450 – 1CI	-45.56	-31.76	Zhao et al., Biochemistry, 2007; 46(41) 11559–11567
1SUO	CytP450 – CPZ	-45.56	-31.76	
1T69	CytP450 – SHH	-37.51	-35	Singh et al., Biochemistry 2014, 53, 7445–7458
1T64	CytP450 – TSN	-36.86	-36	
3F0R				
2UXU	TtgR – CLM	-45.14	-27.59	Alguel et al., (2007) J. Mol. Biol. 369: 829
2UXI	TtgR – G50	-63.12	-42.64	
1SG0	Quinone réductase – STL	-61.03	-40.96	Calamini et al., Biochem J. 2008, 413(1): 81–91
1KV1	MAPk – BMU	-23.41	-33.96	Kroe et al., J. Med. Chem. 2003, 46, 4669-4675

1C1N	Trypsine – Benzamidine	-18.9	-26.6	Talhout & Engberts, 2001, Eur. J. Biochem. 268, 1554-1560
1TIO				
1C5P				
2IL2	Rénine – LIX	-39.08	-40.88	Sarver et al., 2007, Analytical Biochemistry, 360, 30–40
2IKO	Rénine – 7IG	-39.71	-31.35	
2IKU	Rénine – LIY	-41.8	-40.84	
2G1O	Rénine – 2IG	-54.34	-39.67	
2G1R	Rénine – 3IG	-44.73	-40.30	
2G1S	Rénine – 4IG	-28.42	-36.66	
2G24	Rénine – 7IG	-39.71	-31.35	
2I4Q	Rénine – UA4	-37.20	-38.75	
2QWX	Quinone réductase – Mélatonine 1	-43.68	-33.73	Calamini et al., Biochem J. 413(1): 81–91
2QX4				
2QX6				
2QX8	Quinone réductase – Mélatonine 2	-32.81	-35.53	
2QX9				
1SWN	Streptavidine W108F – Biotine	-98.23	n.d	- Klumb et al., Biochemistry, Vol, 37, No, 21, 1998 - Freitag et al., J. Mol. Biol. (1998) 279, 211-221
1SWP	Streptavidine W120F – Biotine	-81.09		
1SWK	Streptavidine W79F – Biotine	-108.68		
1SWR	Streptavidine W120A – Biotine	-53.56		
3RY2	Streptavidine – Biotine	-104.18		
2RTG				
2RTF				
2RTE				
2RTD				
2IZJ				
2IZI				
2IZH				
2IZG				
2IZF				
1STP				
1MK5				
1SWE				
1SWD				
2F01	Streptavidine – R-Biotine / S-Biotine			
2GH7				
1N43	Streptavidine N23A – Biotine	-75.31		
1N9M	Streptavidine S27A – Biotine	-97.49		
1NDJ	Streptavidine Y43F – Biotine	-115.06		
1NC9	Streptavidine – Iminobiotine	-66.94		
2ZC9	Thrombine – 22U	-37.1	-35.4	Baum et al., J. Mol. Biol. (2009) 391, 552–564
2ZDA	Thrombine – 32U	-44.2	n.d	
2ZFP	Thrombine – 19U	-33.5	-31.3	

2ZGX	Thrombine – 29U	-38.7	-40.1	
2ZO3	Thrombine – 33U	-40.7	n.d	
3DHK	Thrombine – 23U	-45.6	-39.6	
3DUX	Thrombine – 64U	-30.5	n.d	
3F68	Thrombine – 95U	-19.4	-32.8	
1P13	v-Src – CdpYANFK	-3.3	-21.4	Taylor et al., 2007, PROTEINS: Structure, Function, and Bioinformatics 67:981–990
1NZL	v-Src – PQpYEpYIPA			
1NZV	v-Src – PQpYIpYVPA			
1SHA	v-Src – pYVPML			
1SHB	v-Src – pYLRVA			
2TIO	Trypsine – Benzamidine	-18.9	-26.6	
1SPS	v-Src – EPQpYEEIPIYL	-34.8	-39.9	
1TP3	PDZ – KKETPV	-17.99	-34.31	Saro et al., Biochemistry, 2007, 46, 6340-6352
1TP5	PDZ – KKETWV	-14.64	-31.80	
1BE9	PDZ – KQTSV	-19.66	-28.03	

Table A5. Énergies d'interaction électrostatiques et de polarisation pour les complexes streptavidine-biotine. Voir équation (77) pour les valeurs relatives.

A.A.	#	3RY2 (*WT*)		1SWK (*W79F*)		1NDJ (*Y43F*)	
		E_{ELEC} (kJ/mol)	E_{POL} (kJ/mol)	E_{ELEC} (kJ/mol)	E_{POL} (kJ/mol)	E_{ELEC} (kJ/mol)	E_{POL} (kJ/mol)
Asn	*23*	-50.72	-19.18	-51.13 *(+0.8%)*	-20.90 *(+8.9%)*	-60.94 *(+20.2%)*	-26.82 *(+39.8%)*
Leu	*25*	-10.08	-1.83	-7.92 *(-21.5%)*	-0.66 *(-64.2%)*	-6.66 *(-33.9%)*	-1.99 *(+8.4%)*
Ser	*27*	-101.33	-40.08	-157.67 *(+55.6%)*	-77.31 *(+92.9%)*	-95.23 *(-6.0%)*	-42.74 *(+6.6%)*
Tyr (*Phe*)	*43*	**-82.76**	**-36.27**	**-75.90 *(-8.3%)***	**-50.51 *(+39.3%)***	**-15.12 *(-81.7%)***	**-14.42 *(-60.2%)***
Ser	*45*	-43.89	-4.74	-45.77 *(+4.3%)*	-5.46 *(+15.2%)*	-23.54 *(-46.4%)*	-1.14 *(-76.0%)*
Val	*47*	14.33	-2.82	+13.17 *(+8.1%)*	-2.79 *(-1.1%)*	+8.32 *(+41.9%)*	-5.61 *(+98.8%)*
Gly	*48*	-17.72	-10.77	-19.85 *(+12.1%)*	-11.40 *(+5.8%)*	-18.86 *(+6.5%)*	-11.69 *(+8.5%)*
Asn	*49*	-93.92	-36.99	-70.29 *(-25.2%)*	-30.82 *(-16.7%)*	-106.49 *(+13.4%)*	-29.34 *(-20.7%)*
Ala	*50*	-6.79	-1.47	-7.79 *(+14.7%)*	-1.55 *(+5.0%)*	-8.72 *(+28.4%)*	-0.99 *(-32.8%)*
Trp (*Phe*)	*79*	**-33.19**	**-16.28**	**-20.80 *(-37.3%)***	**-10.44 *(-35.9%)***	**-33.62 *(+1.3%)***	**-16.70 *(+2.6%)***
Ala	*86*	-8.54	-7.66	-12.98 *(+52.0%)*	-10.16 *(+32.6%)*	-14.06 *(+64.7%)*	-13.62 *(+77.9%)*
Ser	*88*	-90.25	-30.57	-136.92 *(+51.7%)*	-44.43 *(+45.3%)*	-64.75 *(-28.3%)*	-26.63 *(-12.9%)*
Thr	*90*	-32.14	-11.82	-32.67 *(+1.6%)*	-9.20 *(-22.1%)*	-48.69 *(+51.5%)*	-14.63 *(+23.8%)*
Trp	*92*	-14.65	-2.86	-16.15 *(+10.2%)*	-0.34 *(-88.0%)*	-16.06 *(+9.6%)*	-3.10 *(+8.4%)*
Trp	*108*	-9.03	-2.83	-8.34 *(-7.6%)*	-2.73 *(-3.3%)*	-10.16 *(+12.5%)*	-3.16 *(+11.8%)*
Leu	*110*	-30.52	-18.18	-29.09 *(-4.7%)*	-14.92 *(-18.0%)*	-28.07 *(-8.0%)*	-14.19 *(-21.9%)*
Asp	*128*	62.77	-34.45	+67.03 *(-6.8%)*	-39.50 *(+14.7%)*	+104.26 *(-66.1%)*	-21.83 *(-36.6%)*
autres A.A.*		16.18	1.44	+46.37 *(-186.5%)*	+4.59 *(-72.2%)*	+54.23 *(-235.1%)*	+4.16 *(-56.0%)*
Total		***-532.24***	***-276.16***	***-566.70 (+6.5%)***	***-328.51 (+19.0%)***	***-384.17 (-27.8%)***	***-244.45 (-11.5%)***
		-808.40		***-895.22 (+10.7%)***		***-628.61 (-22.2%)***	

***** : N-ter (+Glu, +Ala) et C-ter (+Val, +Lys) sont différents pour la protéine sauvage.

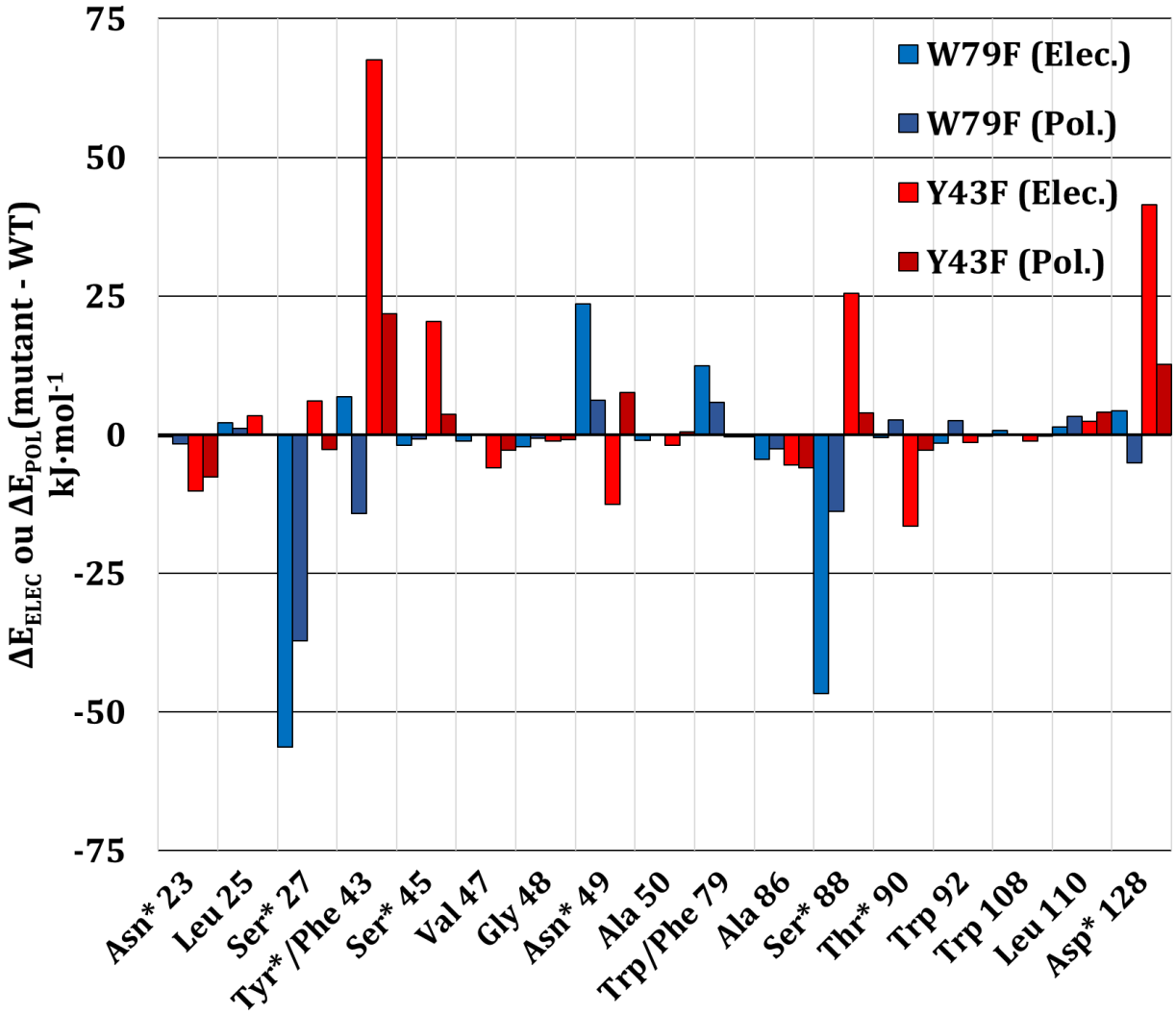


Figure A1. Variation des contributions E_{ELEC} et E_{POL}, par acides aminés, entre la streptavidine sauvage et les mutants W79F ou Y43F.

Références

(1) Mandal, S., Moudgil, M., Mandal, S.K., Rational Drug Design, Eur. J. Pharmacol. 2009, 625, 90-100.

(2) Eibes, G., Arca-Ramos, A., Lema, J.M., Moreira, M.T., Enzymatic technologies for remediation of hydrophobic organic pollutants in soil, Appl. Microbiol. Biotechnol. 2015, 99, 21, 8815-8829.

(3) Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., Bourne, P.E., The Protein Data Bank, Nucleic Acids Res. 2000, 28, 235-242.

(4) Krawczuk, A., Macchi, P., Charge density analysis for crystal engineering, Chem. Cent. J., 2014, 8, 68-83.

(5) Dos Santos, L. H. R., Macchi, P., The Role of Hydrogen Bond in Designing Molecular Optical Materials, Crystals, 2016, 6, 43-57.

(6) Matta, C. F., Arabi, A. A., Electron-density descriptors as predictors in quantitative structure–activity/property relationships and drug design, Future Med. Chem., 2011, 3(8), 969-994.

(7) Dittrich, B., Matta, C. F., Contributions of charge-density research to medicinal chemistry, IUCrJ, 2014, 1, 457-469.

(8) Farrugia, L. J., Evans, C., Tegel, M., Chemical Bonds without "Chemical Bonding"? A Combined Experimental and Theoretical Charge Density Study on an Iron Trimethylenemethane Complex, J. Phys. Chem. A, 2006, 110, 7952-7961.

(9) Bader, R. F. W., Carroll, M. T., Cheeseman, J. R., Chang, C., Properties of Atoms in Molecules: Atomic Volumes, J. Am. Chem. Soc., 1987, 109, 7968-7979.

(10) Gadre, S.R., Jovan Jose, K.V., Rahalkar, A.P., Molecular tailoring approach for exploring structures, energetics and properties of clusters, J. Chem. Sci. 2010, 122, 1, 47-56.

(11) Grabowsky, S., Genoni, A., Bürgi, H.-B., Quantum Crystallography, Chem. Sci., 2017, 8, 4159-4176.

(12) Howard, E.I., Guillot, B., Blakeley, M. P., Haertlein, M., Moulin, M., Mitschler, A., Cousido-Siah, A., Fadel, F., Valsecchi, W. M., Tomizaki, T., Petrova T., Claudot, J., Podjarny, A., High-resolution neutron and X-ray diffraction room-temperature studies of an H-FABP–oleic acid complex: study of the internal water cluster and ligand binding by a transferred multipolar electron-density distribution, IUCrJ, 2015, 3, 115-126.

(13) Guillot, B., Jelsch, C., Podjarny, A., Lecomte, C., Charge-density analysis of a protein structure at subatomic resolution: the human aldose reductase case, Acta Cryst., 2008, D64, 567-588.

(14) Logsdon, B.C., Vickrey, J.F., Martin, P., Proteasa, G., Koepke, J.I., Terlecky, S.R., Wawrzak, Z., Winters, M.A., Merigan, T.C., Kovari, L.C., Crystal structures of a multidrug-resistant human immunodeficiency virus type 1 protease reveal an expanded active-site cavity, J. Virol 2004, 78, 3123-3132.REF10

(15) Coleman, J.A., Green, E.M., Gouaux, E., X-ray structures and mechanism of the human serotonin transporter, Nature 2016, 532, 334-339.

(16) Groom, C. R., Bruno, I. J., Lightfoot, M. P., Ward, S. C., The Cambridge Structural Database, Acta Cryst., 2016, B72, 171-179.

(17) Reilly, A. M., Cooper, R. I., Adjiman, C. S., Bhattacharya, S., Boese, A. D., Brandenburg, J. G., Bygrave, P. J., Bylsma, R., Campbell, J. E., Car, R., Case, D. H., Chadha, R., Cole, J. C., Cosburn, K., Cuppen, H. M., Curtis, F., Day, G. M., DiStasio Jr, R. A., Dzyabchenko, A., van Eijck, B. P., Elking, D. M., van den Ende, J. A., Facelli, J. C., Ferraro, M. B., Fusti-Molnar, L., Gatsiou, C.-A., Gee, T. S., de Gelder, R., Ghiringhelli, L. M., Goto, H., Grimme, S., Guo, R., Hofmann, D. W. M., Hoja, J., Hylton, R. K., Iuzzolino, L., Jankiewicz, W., de Jong, D. T., Kendrick, J., de Klerk, N. J. J., Ko, H.-Y., Kuleshova, L. N., Li, X., Lohani, S., Leusen, F. J. J., Lund, A. M., Lv, J., Ma, Y., Marom, N., Masunov, A. E., McCabe, P., McMahon, D. P., Meekes, H., Metz, M. P., Misquitta, A. J., Mohamed, S., Monserrat, B., Needs, R. J., Neumann, M. A., Nyman, J., Obata, S., Oberhofer, H., Oganov, A. R., Orendt, A. M., Pagola, G. I., Pantelides, C. C., Pickard, C. J., Podeszwa, R., Price, L. S., Price, S. L., Pulido, A., Read, M. G., Reuter, K., Schneider, E., Schober, C., Shields, G. P., Singh, P., Sugden, I. J., Szalewicz, K., Taylor, C. R., Tkatchenko, A., Tuckerman, M. E., Vacarro, F., Vasileiadis, M., Vazquez-Mayagoitia, A., Vogt, L., Wang, Y., Watson, R. E., de Wijs, G. A., Yang, J., Zhu, Q. & Groom, C. R., Report on the sixth blind test of organic crystal structure prediction methods, Acta Cryst. 2016, B72, 439-459.

(18) Ben Nasr, M., Influence de l'environnement cristallin sur les propriétés moléculaires du kétoprofène dans des co-cristaux, Thèse, Université de Lorraine, Nancy, 2016.

(19) Gilmore, C.J., Maximum entropy and Bayesian statistics in crystallography: a review of practical applications, Acta Cryst. 1996, A52, 561-589.

(20) Coppens, P., X-Ray Charge Densities and Chemical Bonding, 1st ed.; Oxford Science Publications: Oxford, 1997.

(21) Pichon-Pesme, V., Lecomte, C., Lachekar, H., On Building a Data Bank of Transferable Experimental Electron Density Parameters: Application to Polypeptides, J. Phys. Chem. 1995, 99, 6242-6250.

(22) Hansen, N. K., Coppens, P., Testing Aspherical Atom Refinements on Small-Molecule Data Sets, Acta Cryst., 1978, A34, 909–921.

(23) Roetti, C., Clementi, E., Simple basis sets for molecular wavefunctions containing atoms from Z = 2 to Z = 54, J. Chem. Phys., 1974, 60, 4725-4729.

(24) Buckingham, A.D., Permanent and Induced Molecular Moments and Long-Range Intermolecular Forces, Adv. Chem. Phys. 1967, 12, 107-142.

(25) Buckingham, A.D., Utting, B.D., Intermolecular Forces, Annu. Rev. Phys. Chem., 1970, 21, 287-316.

(26) Buckingham, A.D., Fowler, P.W., Hutson, J.M., Theoretical Studies of van der Waals Molecules and Intermolecular, Chem. Rev., 1988, 88, 963-988.

(27) Volkov, A., Coppens, P., Calculation of Electrostatic Interaction Energies in Molecular Dimers from Atomic Multipole Moments Obtained by Different Methods of Electron Density Partitioning, J. Comput. Chem. 2004, 25, 921-934.

(28) Bouhmaida, N., Bonhomme, F., Guillot, B., Jelsch, C., Ghermani, N. E., Charge density and electrostatic potential analyses in paracetamol, Acta Cryst., 2009, B65, 363-374.

(29) Bojarowski, S. A., Kumar, P., Dominiak, P. M., A Universal and Straightforward Approach to Include Penetration Effects in Electrostatic Interaction Energy Estimation, ChemPhysChem, 2016, 17, 2455-2460.

(30) Bojarowski, S. A., Kumar, P., Wandtke, C., Bittrich, B., Dominiak, P. M., Universal Method for Electrostatic Interaction Energies Estimation with Charge Penetration and Easily Attainable Point Charges, J. Chem. Theory Comput., 2018, 14, 6336-6345.

(31) Vandenbrande, S., Waroquier, M., Van Speybroeck, V., Verstraelen, T., The Monomer Electron Density Force Field (MEDFF): A Physically Inspired Model for Noncovalent Interactions, J. Chem. Theory Comput. 2017, 13, 161-179.

(32) Stone, A., The Theory of Intermolecular Forces, 1st ed.; Oxford Scholarship, 1996.

(33) Souhassou, M., Blessing, R. H., Topological analysis of experimental electron densities, J. Appl. Cryst., 1991, 32, 210-217.

(34) Platts, J.A., Overgaard, J., Jones, C., Iversen, B.B., Stasch, A., First Experimental Characterization of a Non-nuclear Attractor in a Dimeric Magnesium(I) Compound, J. Phys. Chem. A 2011, 115, 194-200.

(35) Gatti, C., Chemical bonding in crystals: new directions, Z. Kristallogr. 2005, 220, 399-457.

(36) Cukrowski, I., de Lange, J.H., Adeyinka, A.S., Mangondo, P., Evaluating common QTAIM and NCI interpretations of the electron density concentration through IQA interaction energies and 1D cross-sections of the electron and deformation density distributions, Comput. Theor. Chem. 2015, 1053, 60-76.

(37) Espinosa, E., Souhassou, M., Lachekar, H., Lecomte, C., Topological analysis of the electron density in hydrogen bonds, Acta Cryst. 1999, B55, 563-572.

(38) Thomas, S.P., Jayatilaka, D., Guru Row, T.N., S···O chalcogen bonding in sulfa drugs: insights from multipole charge density and X-ray wavefunction of acetazolamide, Phys. Chem. Chem. Phys., 2015, 17, 25411.

(39) Li, J., Xu, T., Ping, Y., van Mourik, T., Früchtl, H., Krik, S.R., Consequences of theory level choice evaluated with new tools from QTAIM and the stress tensor for a dipeptide conformer, Chem. Phys. Lett. 2018, 696, 42-47.

(40) Mata, I., Molins, E., Espinosa, E., Zero-Flux Surfaces of the Electrostatic Potential: The Border of Influence Zones of Nucleophilic and Electrophilic Sites in Crystalline Environment, J. Phys. Chem. A 2007, 111, 9859-9870.

(41) Mata, I., Molins, E., Alkorta, I., Espinosa, E., The Paradox of Hydrogen-Bonded Anion−Anion Aggregates in Oxoanions: A Fundamental Electrostatic Problem Explained in Terms of Electrophilic···Nucleophilic Interactions, J. Phys. Chem. A 2015, 119, 183−194.

(42) Gussoni, M., Rui, M., Zerbi, G., Electronic and relaxation contribution to linear molecular polarizability. An analysis of experimental values, J. Mol. Struct., 1998, 447, 163-215.

(43) Applequist, J., Carl, J.R., Fung, K.K., An Atom Dipole Interaction Model for Molecular Polarizability. Application to Polyatomic Molecules and Determination of Atom Polarizabilities, J. Amer. Chem. Soc., 1972, 94, 2952-2960.

(44) Silberstein, L., VII. Molecular refractivity and atomic interaction, Philosophical Magazine Series 6, 1917, 33(193), 92-128.

(45) Silberstein, L., L. Molecular refractivity and atomic interaction. II, Philosophical Magazine Series 6, 1917, 33(198), 521-533.

(46) Thole, B.T., Molecular Polarizabilities Calculated with a Modified Dipole Interaction, Chem. Phys., 1981, 59, 341-350.

(47) van Duijnen, P.T., Swart, M., Molecular and Atomic Polarizabilities: Thole's Model Revisited, J. Phys. Chem. A, 1998, 102, 2399-2407.

(48) Birge, R.R., Schick, G.A., Bocian, D.F., Calculation of molecular polarizabilities using a semiclassical Slater-type orbital-point dipole interaction (STOPDI) model, J. Chem. Phys. 1983, 79, 2256-2264.

(49) Krawczuk, A., Pérez, D., Macchi, P., *Polaber*: a program to calculate and visualize distributed atomic polarizabilities based on electron density partitioning, J. Appl. Cryst., 2014, 47, 1452-1458.

(50) 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, 2427-2438.

(51) 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, 3466-3470.

(52) Leduc, T., Aubert, E., Espinosa, E., Jelsch, C., Iordache, I., Guillot, B., Polarization of Electron Density Databases of Transferable Multipolar Atoms, J. Phys. Chem. A 2019, 123, 32, 7156-7170.

(53) Mackenzie, C. F., Spackman, P. R., Jayatilaka, D., Spackman, M. A., CrystalExplorer model energies and energy frameworks: extension to metal coordination compounds, organic salts, solvates and open-shell systems, IUCrJ, 2017, 4, 575-587.

(54) Spackman, M. A., Towards the use of experimental electron densities to estimate reliable lattice energies, CrystEngComm, 2018, 20, 5340-5347.

(55) Gavezzotti, A., Calculation of Intermolecular Interaction Energies by Direct Numerical Integration over Electron Densities. I. Electrostatic and Polarization Energies in Molecular Crystals, J. Phys. Chem. B, 2002, 106, 4145-4154.

(56) Gavezzotti, A., Calculation of Intermolecular Interaction Energies by Direct Numerical Integration over Electron Densities. 2. An Improved Polarization Model and the Evaluation of Dispersion and Repulsion Energies, J. Phys. Chem. B, 2003, 107, 2344-2353.

(57) Jeziorski, B., Moszynski, R., Szalewicz, K., Perturbation Theory Approach to Intermolecular Potential Energy Surfaces of van der Waals Complexes, Chem. Rev., 1994, 94(7), 1887-1930.

(58) 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, 1563−1575.

(59) Řezáč, J., Jurečka, P., Riley, K. E., Černý, J., Valdes, H., Pluháčková, K., Berka, K., Řezáč, T., Pitoňák, M., Vondrášek, J., Hobza, P., Quantum Chemical Benchmark Energy and Geometry Database for Molecular Clusters and Complex Molecular Systems (www.begdb.com): A Users' Manual and Examples, Collect. Czech. Chem. Commun. 2008, 73, 1261-1270.

(60) Brock, C. P., Dunitz, J. D., Hirshfeld, F. L., Transferability of Deformation Densities among Related Molecules: Atomic Multipole Parameters from Perylene for Improved Estimation of Molecular Vibrations in Naphthalene and Anthracene, Acta Cryst., 1991, B47, 789-797.

(61) 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 Cryst., 2012, A68, 337-351.

(62) Meyer, B., Guillot, B., Ruiz-Lopez, M. F., Genoni, A., Libraries of Extremely Localized Molecular Orbitals. 1. Model Molecules Approximation and Molecular Orbitals Transferability, J. Chem. Theory Comput., 2016, 12, 1052-1067.

(63) Meyer, B., Guillot, B., Ruiz-Lopez, M. F., Jelsch, C., Genoni, A., Libraries of Extremely Localized Molecular Orbitals. 2. Comparison with the Pseudoatoms Transferability, J. Chem. Theory Comput., 2016, 12, 1068-1081.

(64) Meyer, B., Genoni. A., Libraries of Extremely Localized Molecular Orbitals. 3. Construction and Preliminary Assessment of the New Databanks, J. Phys. Chem. A, 2018, 122, 8965-8981.

(65) Domagała, S., Munshi, P., Ahmed, M., Guillot, B., Jelsch C., Structural analysis and multipole modelling of quercetin monohydrate – a quantitative and comparative study, Acta Cryst., 2011, B67, 63-78.

(66) Volkov, A., Li, X., Koritsanszky, T., Coppens, P., *Ab Initio* Quality Electrostatic Atomic and Molecular Properties Including Intermolecular Energies from a Transferable Theoretical Pseudoatom Databank, J. Phys. Chem. A, 2004, 108, 4283-4300.

(67) Volkov, A., Messerschmidt, M., Coppens, P., Improving the scattering-factor formalism in protein refinement: application of the University at Buffalo Aspherical-Atom Databank to polypeptide structures, Acta Cryst., 2007, D63, 160-170.

(68) Jarzembska, K. N., Dominiak, P. M., New version of the theoretical databank of transferable aspherical pseudoatoms, UBDB2011 – towards nucleic acid modelling, Acta Cryst., 2012, A68, 139-147.

(69) Dominiak, P. M., Volkov, A., Li, X., Messerschmidt, M., Coppens, P., A Theoretical Databank of Transferable Aspherical Atoms and Its Application to Electrostatic Interaction Energy Calculations of Macromolecules, J. Chem. Theory Comput., 2007, 3, 232-247.

(70) Dittrich, B., Hübschle, C. B., Pröpper, K., Dietrich, F., Stolper, T., Holstein, J. J., The generalized invariom database (GID), Acta Cryst., 2013, B69, 91-104.

(71) 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 Cryst. 1998, D54, 1306-1318.

(72) Zhou, G.-Q., Zhong, W.-Z., Diffusion-Controled Reaction of Enzymes A Comparision of Chou's Model and Alberty-Hammes-Eigen's Model, Eur. J. Biochem. 1982, 128, 383-387.

(73) Cho, S., Chittoor, P. S., Bonsor, D. A., Kerzic, M. C., Guan, R., Yang, J., Kieke, M.C., Andersen, P.S., Kranz, D.M., Mariuzza, R.A., Sundberg, E.J., Assessing Energetic Contributions to Binding from a Disordered Region in a Protein−Protein Interaction, Biochemistry 2010, 49, 9256-9268.

(74) Bjelić, S., Jelesarov, I., A survey of the year 2007 literature on application of isothermal titration calorimetry, J. Mol. Recognit. 2008, 21, 289–311.

(75) Ghai, R., Falconer, R .J., Collins, B.R., Applications of isothermal titration calorimetry in pure and applied research—survey of the literature from 2010, J. Mol. Recognit. 2012, 25, 32–52.

(76) Eyring, H., The Activated Complex and the Absolute Rate of Reaction, J. Chem. Phys. 1935, 3, 107-115.

(77) Jelsch, C., Guillot, B., Lagoutte, A., Lecomte, C., Advances in protein and small-molecule charge density refinement methods using *MoPro*, J. Appl. Cryst., 2005, 38, 38-54.

(78) Zarychta, B., Pichon-Pesme, V., Guillot, B., Lecomte, C., Jelsch, C., On the application of an experimental multipolar pseudo-atom library for accurate refinement of small-molecule and protein crystal structures, Acta Cryst., 2007, A63, 108-125.

(79) Guillot, B., Espinosa, E., Huder, L., Jelsch, C., MoProViewer: a tool to study proteins from a charge density science perspective, Acta Cryst., 2014, A70, C279.

(80) Sanderson, C., Curtin, R., Armadillo: a template-based C++ library for linear algebra, J. Open Source Softw. 2016, 1, 26.

(81) Umeyama, S., Least-Squares Estimation of Transformation Parameters Between Two Point Patterns, IEEE Trans. Pattern Anal. Mach. Intell. 1991, 13, 376-380.

(82) Nassour, A., Domagala, S., Guillot, B., Leduc, T., Lecomte, C., Jelsch, C., A theoretical-electron-density databank using a model of real and virtual spherical atoms, Acta Cryst., 2017, B73, 610-625.

(83) Dovesi, R., Orlando, R., Civalleri, B., Roetti, C., Saunders, V.R., Zicovich-Wilson, C.M., Z. Kristallogr. 2005, 220, 571.

(84) Gaussian 09, Revision D.01, M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, G. A. Petersson, H. Nakatsuji, X. Li, M. Caricato, A. Marenich, J. Bloino, B. G. Janesko, R. Gomperts, B. Mennucci, H. P. Hratchian, J. V. Ortiz, A. F. Izmaylov, J. L. Sonnenberg, D. Williams-Young, F. Ding, F. Lipparini, F. Egidi, J. Goings, B. Peng, A. Petrone, T. Henderson, D. Ranasinghe, V. G. Zakrzewski, J. Gao, N. Rega, G. Zheng, W. Liang, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, K. Throssell, J. A. Montgomery, Jr., J. E. Peralta, F. Ogliaro, M. Bearpark, J. J. Heyd, E. Brothers, K. N. Kudin, V. N. Staroverov, T. Keith, R. Kobayashi, J. Normand, K. Raghavachari, A. Rendell, J. C. Burant, S. S. Iyengar, J.

Tomasi, M. Cossi, J. M. Millam, M. Klene, C. Adamo, R. Cammi, J. W. Ochterski, R. L. Martin, K. Morokuma, O. Farkas, J. B. Foresman, and D. J. Fox, Gaussian, Inc., Wallingford CT, 2016.

(85) AIMAll (Version 17.11.14), Todd A. Keith, TK Gristmill Software, Overland Park KS, USA, 2017 (aim.tkgristmill.com).

(86) Dos Santos, L.H.R., Krawczuk, A., Macchi, P., Distributed Atomic Polarizabilities of Amino Acids and their Hydrogen-Bonded Aggregates, J. Phys. Chem. A, 2015, 119, 3285-3298.

(87) Thierfelder, C., Assadollahzadeh, B., Schwerdtfeger, P., Schäfer, S., Schäfer, R., Relativistic and Electron Correlation Effects in Static Dipole Polarizabilities for the Group 14 Elements from Carbon to Element 114, Phys. Rev. A, 2008, 78, 052506-1-7.

(88) S. P. Goldman, Gauge-invariance method for accurate atomic-physics calculations: Application to relativistic polarizabilities, Phys. Rev. A, 1989, 39, 976-980.

(89) Das, A. K., Thakkar, A. J., Static response properties of second-period atoms: coupled cluster calculations, J. Phys. B, At. Mol. Opt. Phys., 1998, 31, 2215-2223.

(90) Werner, H.-J., Meyer, W., Finite perturbation calculations for the static dipole polarizabilities of the first-row atoms, Phys. Rev. A, 1976, 13, 13-16.

(91) Mocilac, P., Donnelly, K., Gallagher, J.F., Structural systematics and conformational analyses of a 3*3 isomer grid of fluoro-N-(pyridyl)benzamides: physicochemical correlations, polymorphism and isomorphous relationships, Acta Cryst., 2012, B68, 189–203.

(92) Xie, W., Pu, J., Gao, J., A Coupled Polarization-Matrix Inversion and Iteration Approach for Accelerating the Dipole Convergence in a Polarizable Potential Function, J. Phys. Chem. A 2009, 113, 2109–2116.

(93) Kumar, P., Sławomir A. Bojarowski, Katarzyna N. Jarzembska, Sławomir Domagała, Kenno Vanommeslaeghe, Alexander D. MacKerell, Jr., Dominiak., P.M. A Comparative Study of Transferable Aspherical Pseudoatom Databank and Classical Force Fields for Predicting Electrostatic Interactions in Molecular Dimers, J. Chem. Theory Comput., 2014, 10, 1652−1664.

(94) Henkelman, G., Arnaldsson, A., Jónsson, H., A fast and robust algorithm for Bader decomposition of charge density, Comput. Mater. Sci., 2006, 36, 354-360.

(95) Sanville, E., Kenny, S. D., Smith, R., Henkelman, G., Improved Grid-Based Algorithm for Bader Charge Allocation, J. Comput. Chem., 2007, 28, 899-908.

(96) Tang, W., Sanville, E., Henkelman, G., A grid-based Bader analysis algorithm without lattice bias, J. Phys.: Condens. Matter, 2009, 21, 084204-084211.

(97) Fournier, B., Guillot, B., Lecomte, C., Escudero-Adán, E. C., Jelsch, C., A method to estimate statistical errors of properties derived from charge-density modelling, Acta Cryst., 2018, A74, 170-183.

(98) Fournier, B., Bendeif, E.-E., Guillot, B., Podjarny, A., Lecomte, C., Jelsch, C., Charge Density and Electrostatic Interactions of Fidarestat, an Inhibitor of Human Aldose Reductase J. Am. Chem. Soc., 2009, 131, 10929-10941.

(99) Dominiak, P. M., Volkov, A., Dominiak, A. P., Jarzembska, K. N., Coppens, P., Combining crystallographic information and an aspherical-atom data bank in the evaluation of the electrostatic interaction energy in an enzyme–substrate complex: influenza neuraminidase inhibition, Acta Cryst., 2009, D65, 485-499.

(100) Liebschner, D., Elias, M., Moniot, S., Fournier, F., Scott, K., Jelsch, C., Guillot, B., Lecomte, C., Chabrière, E., Elucidation of the Phosphate Binding Mode of DING Proteins Revealed by Subangstrom X-ray Crystallography, J. Am. Chem. Soc. 2009, 131, 7879-7886.

(101) Bojarowski, S. A., Kumar, P., Dominiak, P. M., Interplay of point multipole moments and charge penetration for intermolecular electrostatic interaction energies from the University at Buffalo pseudoatom databank model of electron density, Acta Cryst., 2017, B73, 598-609.

(102) Domagała, S., Jelsch, C., Optimal local axes and symmetry assignment for charge-density refinement, J. Appl. Cryst., 2008, 41, 1140-1149.

(103) Word, J.M., Lovell, S.C., Richardson, J.S., Richardson, D.C., Asparagine and Glutamine: Using Hydrogen Atom Contacts in the Choice of Side-chain Amide Orientation, J. Mol. Biol. 1999, 285, 1735-1747.

(104) Chen, V.B., Arendal III, W.B., Headd, J.J., Keedy, D.A., Immormino, R.M., Kapral, G.J., Murray, L.W., Richardson, J.S., Richardson, D.C., MolProbity: all-atom structure validation for macromolecular crystallography, Acta Cryst 2010, D66, 12-21.

(105) Chilkoti, A., Stayton, P.S., Molecular Origins of the Slow Streptavidin-Biotin Dissociation Kinetics, J. Am. Chem. Soc. 1995, 117, 10622-10628.

(106) Klumb, L.A., Chu, V., Stayton, P.S., Energetic Roles of Hydrogen Bonds at the Ureido Oxygen Binding Pocket in the Streptavidin-Biotin Complex, Biochemistry 1998, 37, 7657-7663.

(107) Freitag, S., Le Trong, I., Chilkoti, A., Klumb, L.A., Stayton, P.S., Stenkamp, R.E., Structural Studies of Binding Site Tryptophan Mutants in the High-affinity Streptavidin-Biotin Complex, J. Mol. Biol. 1998, 279, 211-221.

(108) Freitag, S., Le Trong, I., Klumb, L.A., Chu, V., Chilkoti, A., Stayton, P.S., Stenkamp, R.E., X-ray crystallographic studies of streptavidin mutants binding to biotin, Biomol. Eng. 1999, 16, 13-19.

(109) Le Trong, I., Wang, Z., Hyre, D.E., Lybrand, T.P., Stayton, P.S., Stenkamp, R.E., Streptavidin and its biotin complex at atomic resolution, Acta Cryst. 2011, D67, 813-821.

(110) Le Trong, I., Freitag, S., Klumb, L.A., Chu, V., Stayton, P.S., Stenkamp, R.E., Structural studies of hydrogen bonds in the high-affinity streptavidin-biotin complex: mutations of amino acids interacting with the ureido oxygen of biotin, Acta Cryst. 2003, D59, 1567-1573.

(111) Baugh, L., Le Trong, I., Stayton, P.S., Stenkamp, R.E., Lybrand, T.P., A Streptavidin Binding Site Mutation Yields an Unexpected Result: An Ionized Asp128 Residue Is Not Essential for Strong Biotin Binding, Biochemistry 2016, 55, 5201-5203.

(112) Frenkel, D., Smit, B., Understanding Molecular Simulation, 2nd ed.; San Diego Academic Press: San Diego, 2002.

(113) Mata, I., Alkorta, I., Molins, E., Espinosa, E., Universal Features of the Electron Density Distribution in Hydrogen-Bonding Regions: A Comprehensive Study Involving H···X (X=H, C, N, O, F, S, Cl, p) Interactions, Chem. Eur. J. 2010, 16, 2442-2452.

(114) Li, M., Zhuang, B., Lu, Y., Wang, Z.-G., An, L., Accurate Determination of Ion Polarizabilities in Aqueous Solutions, J. Phys. Chem. B 2017, 121, 26, 6416-6424.

(115) Cox, J.D., Cama, E., Colleluori, D.M., Pethe, S., Boucher, J.-L., Mansuy, D., Ash, D .E., Christianson, D.W., Mechanistic and Metabolic Inferences from the Binding of Substrate Analogues andProducts to Arginase, Biochemistry, 2001, 40, 2689-2701.

(116) Birkedal, H., Madsen, D., Mathiesen, R.H., Knudsen, K., Weber, H.-P., Pattison, P., Schwarzenbach, D., The Charge Density Of Urea From Synchrotron Diffraction Data, Acta Cryst., 2004, A60, 371-381.

(117) Shi, M.W., Thomas, S.P., Hathwar, V.R., Edwards, A.J., Piltz, R.O., Jayatilaka, D., Koutsantonis, G.A., Overgaard, J., Nishibori, E., Iversen, B.B., Spackman, M.A., Measurement of Electric Fields Experienced by Urea Guest Molecules in the 18–Crown–6:Urea (1:5) Host–Guest Complex: An Experimental Reference Point for Electric-Field-Assisted Catalysis, J. Am. Chem. Soc. 2019, 141, 9, 3965-3976.

(118) Suponitsky, K. Y., Tsirelson, V. G., Feil, D., Electron-density-based calculations of intermolecular energy: case of urea, Acta Cryst., 1999, A55, 821-827.

(119) Zhang, C., Tsuhan C., Efficient Feature Extraction for 2D/3D Objects In Mesh Representation, Proceedings : 2001 International Conference on Image Processing, IEEE, 2001, 3, 935-938.

(120) Birge, R.R., Calculation of molecular polarizabilities using an anisotropic atom point dipole interaction model which includes the effect of electron repulsion, J. Chem. Phys. 1980, 72, 5312-5319.

(121) Klein, P.P., On the Ellipsoid and Plane Intersection Equation, Appl. Math. 2012, 3, 1634-1640.

Mots-clés : Cristallographie, Relation Structure-Fonction, Topologie, Protéines

Résumé. L'objectif de ce travail est d'améliorer la compréhension des interactions non-covalentes, à partir de l'étude de la densité électronique, notamment appliquée aux complexes protéines-ligands. La densité électronique est relativement conservée pour des atomes dans des contextes moléculaires similaires. Cette observation préalable fonde le principe de transférabilité. Cette approximation est à la racine des approches basées sur les bases de données, telles que la librairie de types atomiques ELMAM2, d'origine expérimentale. La base de données permet de reconstruire la densité électronique de molécules biologiques à partir de fragments observés lors d'expériences de cristallographie à haute résolution (exprimés dans le modèle multipolaire de Hansen et Coppens). Une approche novatrice est appliquée ici pour déterminer un type spécifique d'interaction électrostatique : l'interaction entre moments dipolaires induits. Cette interaction classique est représentée à l'aide d'un modèle de densité électronique transféré auquel est adjoint un ensemble de polarisabilités atomiques théoriques. Cette propriété décrit la réponse de la densité à l'application d'un champ électrique. Le développement de nouveaux outils informatiques, intégrés au logiciel *MoProSuite*, permet plusieurs axes d'étude présentés dans ce manuscrit. La transférabilité des polarisabilités atomiques est discutée pour évaluer leur conservation entre différents contextes moléculaires. Les interactions à courte portée entre molécules impliquent de très forts champs électriques. Leurs effets sont pris en compte en modélisant le phénomène de polarisation. L'adjonction de la polarisation à la densité électronique transférée est évaluée pour des systèmes biologiques et des petites molécules organiques. Ces densités transférées puis polarisées permettent de calculer des énergies d'interactions électrostatiques précises qui sont comparées à un banc de test théorique et appliquées à une étude de cas expérimental. L'étude de la topologie des champs scalaires, dans le cas du potentiel électrostatique au sein des complexes protéine-ligand, est encore à un stade de développement précoce. Le modèle et les nouveaux outils développés dans cette thèse permettent d'envisager l'application de cette approche, qui décrit la forme des champs de densité et de potentiel électrostatique, à des systèmes biologiques.

Keywords: Crystallography, Structure-Function Relationship, Topology, Proteins

Abstract. The goal of this work is to improve the understanding of non-covalent interactions, from the study of electron density, notably applied to protein-ligand complexes. The electron density is a relatively conserved property of atoms in similar molecular environments. This preliminary observation establishes the transferability principle. This approximation is at the root of database-based approaches, such as the ELMAM2 atom type library. This database makes it possible to reconstruct the electron density of biological molecules from fragments observed during high-resolution crystallographic experiments (and expressed in the Hansen and Coppens multipolar model). Determination of a specific type of electrostatic interaction, the interaction between induced dipole moments, profits of an innovative approach. Transferable experimental electronic densities are complemented by theoretical polarizabilities. Polarizabilities describe the response of the electron density to the application of an electric field, in order to model the classical interaction between induced dipoles. The development of new software tools, integrated into the *MoProSuite* software, allows several focal area presented in this manuscript. The transferability of atomic polarizabilities is discussed to evaluate their conservation between different molecular environments. Short-range interactions between molecules involve very strong electric fields. Their effect is taken into account by modelling the polarization phenomenon. The adjunction of polarization to the transferred electron density is evaluated on biological systems and small organic molecules. These transferred, then polarized, densities make it possible to calculate accurate electrostatic interaction energies, which are compared to a theoretical benchmark and an experimental case study. The study of scalar field topology, in the case of the electrostatic potential within protein-ligand complexes, is at an early stage of development. The model and the new tools developed in this thesis are steps toward the application of this approach, which describes the shape of the fields of density and electrostatic potential, to biological systems.