



**UNIVERSITÉ
DE LORRAINE**

**BIBLIOTHÈQUES
UNIVERSITAIRES**

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 bibliothèque : ddoc-theses-contact@univ-lorraine.fr
(Cette adresse ne permet pas de contacter les auteurs)

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>

Développement de nouveaux descripteurs
électrostatiques issus de méthodes de cristallographie
quantique appliqués au domaine de la biologie
structurale

THESE

présentée et soutenue publiquement le 14 septembre 2023
pour l'obtention du titre de

Docteur de l'Université de Lorraine

Mention Physique

par

Eva Mocchetti

Composition du jury :

Rapporteurs :

Julia Contreras-García,	Directrice de recherche, LCT, Sorbonne Université
Dominique Housset,	Chercheur HDR, IBS, CEA

Examineurs :

Claudine Mayer,	Professeure, ICube, Universités de Starbourg et Paris Cité
Enrique Espinosa,	Professeur, CRM2, Université de Lorraine (Président du jury)

Directeurs de thèse :

Benoît Guillot,	Professeur, CRM2, Université de Lorraine
Claude Didierjean,	Maître de conférence, CRM2, Université de Lorraine

Table des matières

Table des matières	i
Remerciements	v
Table des figures	vii
Liste des tableaux	ix
1 Introduction	1
1.1 Introduction générale	1
1.2 Cristallographie quantique et transférabilité des paramètres du modèle multipolaire de la densité électronique	5
1.2.1 Cristallographie quantique	5
1.2.2 Modèle multipolaire de Hansen et Coppens	7
1.2.3 Transférabilité des paramètres de densité électronique atomique du modèle multipolaire	10
1.3 Champs scalaires moléculaires et analyse topologique	14
1.3.1 Densité électronique moléculaire	15
1.3.2 Potentiel électrostatique et champ électrique moléculaires	21
1.4 Modélisation de l'énergie d'interaction intermoléculaire totale	27
1.4.1 A partir des densités de charge expérimentales et transférées	28
1.4.2 Méthodes empiriques	31
1.4.3 Méthodes de chimie quantique	36
1.5 Résumé de l'introduction et objectifs de la thèse	41
2 Descripteurs issus de la topologie du potentiel électrostatique	45
2.1 Points critiques	46
2.1.1 Définition des points critiques du potentiel électrostatique	46
2.1.2 Algorithme de recherche des points critiques	48
2.2 Topographie des lignes de champ électrique et faisceaux primaires	52
2.2.1 Lignes de champ électrique	52
2.2.2 Faisceaux primaires et surfaces de flux nul	52
2.2.3 Algorithme de détermination des surfaces entourant les faisceaux primaires	54
2.3 Partition de l'espace en zones d'influence électrophile et en zones d'influence nucléophile	56

2.3.1	Définition des zones d'influence électrophile et nucléophile	56
2.3.2	Partitions de l'espace moléculaire	58
2.3.3	Surface de flux nul du potentiel et vérification du théorème de Gauss . . .	60
2.3.4	Interprétation des descripteurs électrostatiques de partitions en zones d'influence électrophile et nucléophile	62
2.4	Implémentation dans MoProViewer	64
2.4.1	La suite logiciel MoPro et le logiciel MoProViewer	64
2.4.2	Implémentation de la détermination des points critiques du potentiel électrostatique	66
2.4.3	Implémentation de la détermination des faisceaux primaires et des zones d'influence	68
2.4.4	Perspectives de développements	70
2.5	Conclusion partielle de chapitre	71
3	Potentiel d'interaction basé sur la densité électronique expérimentale transférée ELMAM	75
3.1	Modèle basé sur les données de la librairie ELMAM2	76
3.1.1	Décomposition du potentiel d'interaction total ELMAM	76
3.1.2	Jeu de données d'entraînement et de validation du modèle	77
3.1.3	Validation des contributions électrostatique et d'induction	79
3.2	Potentiel de dispersion	83
3.2.1	Approximation de London de la dispersion	83
3.2.2	Polarisabilités atomiques	84
3.2.3	Introduction de paramètres atomiques empiriques	86
3.2.4	Influence de la composition du dimère	87
3.2.5	Autres modèles d'énergie de dispersion	88
3.3	Potentiel d'échange-répulsion	89
3.3.1	Modèle de recouvrement des densités électroniques	89
3.3.2	Intégration du produit des densités électroniques	91
3.3.3	Modèle dépendant de la distance interatomique	96
3.3.4	Introduction de paramètres atomiques empiriques	100
3.3.5	Influence de la composition du dimère	101
3.3.6	Autres modèles d'énergie de répulsion	102
3.4	Potentiel d'interaction van der Waals ELMAM	103
3.4.1	Sommation des contributions ELMAM de dispersion et d'échange-répulsion	103
3.4.2	Combinaison linéaire	105
3.4.3	Influence de la composition du dimère	108
3.5	Conclusion partielle de chapitre	109
4	Applications aux complexes protéine-ligand	113
4.1	Intérêts de l'étude des zones d'influence : complexe Trypsine Bovine - Inhibiteur SGPI	114
4.1.1	Contexte de l'étude	114

4.1.2	Caractérisation de la trypsine du point de vue des zones d'influence . . .	114
4.1.3	Draft de l'article décrivant les descripteurs issus de la topologie du potentiel électrostatique et leur application à la trypsine	116
4.2	Application des zones d'influence en biologie structurale : complexe Neuropiline 1 - peptide KDKPPR	139
4.2.1	Contexte de l'étude	139
4.2.2	Caractérisation de la fixation de l'inhibiteur du point de vue des zones d'influence nucléophile	139
4.2.3	Article publié décrivant la structure de NRP1	141
4.3	Energies d'interaction électrostatique ELMAM pour l'étude des complexes protéine-ligand : complexe Glutathion Transférase - Glutathion	159
4.3.1	Etude structurale de l'enzyme SynGSTC1	159
4.3.2	Caractérisation de l'interaction protéine-ligand par les énergies électrostatiques ELMAM	159
4.3.3	Article publié décrivant la structure de SynGSTC1	160
4.4	Perspective d'application à un système dynamique : pompage de l'ion chlorure par l'halorhodopsine	180
4.4.1	L'halorhodopsine NmHR	180
4.4.2	Barrière moléculaire stérique	182
4.4.3	Barrière moléculaire électrostatique	187
4.5	Conclusion partielle de chapitre	190
5	Conclusion globale et perspectives	193
	Bibliographie	223
	Abréviations	225
	Résumé / Abstract	229

Remerciements

Tout d'abord, je tiens à remercier Dr. Julia Contreras-García et Dr. Dominique Housset d'avoir accepté d'être rapporteurs de ce travail. Je remercie également Pr. Claudine Mayer et Pr. Enrique Espinosa pour avoir fait partie de mon jury de thèse. Notre échange lors de la soutenance de thèse fut très instructif et a contribué à enrichir ce travail.

Benoît, Claude, je tiens à vous exprimer ma profonde gratitude pour votre accompagnement et votre soutien tout au long de ma thèse. Merci pour votre disponibilité, pour vos conseils et pour votre bienveillance. Vous avez partagé avec moi votre passion et votre expertise, j'ai beaucoup appris grâce à vous. Je vous remercie de m'avoir fait confiance pour ce projet. Plus que des encadrants, vous avez été pour moi de véritables mentors.

Je suis également reconnaissante envers tous mes collègues du laboratoire CRM2 pour m'avoir accueillie parmi eux. Merci à Dominik, directeur du laboratoire, à Rémi, Frédérique et Christian de l'équipe BioMIMIC, aux autres doctorants Asma, Amira, Julien, Vedran et Vishnu, et à Pierrick, Emmanuel, Elodie, Manu, El Eulmi, Abdel, Bruno, Anne, Valérie, et à tous les autres. Grâce à vous, le laboratoire est baigné d'une ambiance chaleureuse dans laquelle je me suis tout de suite sentie bien entourée.

Je remercie tout particulièrement Morgane, ma voisine de bureau qui est devenue mon amie. Grâce à toi, les journées de travail étaient toujours ponctuées de papotages, de fous rires et de desserts de la cafétéria (surtout de choco-cakes!). Merci d'avoir été là jusqu'au bout et pour avoir relu ma thèse jusqu'à la dernière ligne.

Je tiens également à exprimer ma gratitude envers Dr. Jean-François Wax et Dr. Alessandro Genoni pour avoir accepté de faire partie de mon comité de suivi. Vous avez parfaitement tenu votre rôle et j'ai beaucoup apprécié nos échanges lors de ces rencontres. Je remercie notamment Jean-François pour m'avoir accompagnée depuis le début de mes études supérieures et de m'avoir fait découvrir le monde de la recherche. Merci également de m'avoir appris à rédiger, vos conseils m'ont suivi jusqu'à la fin de la rédaction de ce manuscrit.

J'ai eu la chance d'enseigner à l'UFR SciFA pendant mes trois années de doctorat et pour cela je remercie Dr. Stéphane Dalmasso et toute l'équipe pédagogique de la licence Physique-Chimie de Metz. Cette expérience a été très enrichissante et m'a permis de beaucoup évoluer.

Je salue également mes collègues doctorants, savoir que d'autres partagent les mêmes difficultés que celles que j'ai pu rencontrer fut réconfortant. Merci notamment à Héroïse pour les déjeuners entre la FST et l'IJL, ces moments de détente ont été précieux.

Je remercie chaleureusement mes parents qui m'ont toujours soutenue dans mes études et pour tout le reste. J'ai toujours pu compter sur vous.

Je souhaite rendre hommage ici à mes grands-parents qui nous ont quittés ces trois dernières années. J'espère vous rendre fiers.

Enfin, à toi mon amour, depuis plus de 10 ans tu as toujours été là pour moi. Tout particulièrement pendant les dernières semaines qui ont été difficiles, merci de m'avoir soutenue et réconfortée (et de m'avoir supportée). Je t'aime de tout mon cœur.

Table des figures

1.1	Comparaison des densités électroniques obtenues par le modèle IAM, le formalisme kappa et le modèle multipolaire.	9
1.2	Illustration des trois contributions du modèle multipolaire de la densité électronique : électrons de cœur à symétrie sphérique, électrons de valence à symétrie sphérique et déformations multipolaires des électrons de valence.	10
1.3	Descripteurs issus de la topologie de la densité électronique moléculaire dans le complexe N-méthylacétamide - méthanol.	17
1.4	Représentation avec MoProViewer de l'indice d'interaction non-covalente NCI dans le complexe N-méthylacétamide - méthanol.	20
1.5	Représentation du potentiel électrostatique calculé à partir de la densité électronique moléculaire (a) transférée et (b) polarisée sur les surfaces moléculaires du complexe N-méthylacétamide - méthanol.	23
1.6	Lignes de champ électrique dans l'espace intermoléculaire du complexe N-méthylacétamide - méthanol.	25
1.7	Profil du potentiel d'interaction de Lennard-Jones.	35
2.1	Les points critiques du potentiel électrostatique dans le complexe N-méthylacétamide - méthanol.	47
2.2	Potentiel électrostatique $V(\mathbf{r})$ autour de l'atome d'oxygène de la molécule d'eau.	50
2.3	Etapas de détermination des points critiques d'un champ scalaire moléculaire.	51
2.4	Faisceaux primaires fermés dans l'espace intermoléculaire du complexe N-méthylacétamide - méthanol.	53
2.5	Etapas de détermination de la surface entourant un faisceau primaire fermé.	55
2.6	Définition de la zone d'influence d'un site électrophile et d'un site nucléophile.	57
2.7	Partition de l'espace moléculaire du complexe N-méthylacétamide - méthanol en zones d'influence électrophile et en zones d'influence nucléophile.	59
2.8	Lignes de champ électrique dans la poche de fixation de la FABP.	63
2.9	Zones d'influence électrophile et nucléophile dans la poche de fixation de la FABP.	65
2.10	Interface utilisateur de l'outil de calcul des points critiques du potentiel électrostatique dans MoProViewer.	68
2.11	Interface utilisateur de l'outil de calcul des zones d'influence et des faisceaux primaires dans MoProViewer.	69

3.1	Comparaison des résultats obtenus par les modèles ELMAM d'énergie électrostatique permanente et d'énergie d'induction dipolaire par rapport aux valeurs de référence SAPT.	80
3.2	Influence de la composition des dimères sur l'accord entre le modèle ELMAM et la référence SAPT.	82
3.3	Résultats obtenus par le modèle ELMAM de l'énergie de dispersion à partir des polarisabilités isotropes moyennes et des tenseurs de polarisabilités anisotropiques comparés aux valeurs de référence SAPT.	84
3.4	Résultats obtenus par le modèle ELMAM de l'énergie de dispersion avec les paramètres empiriques comparés aux valeurs de référence SAPT.	87
3.5	Influence de la composition des dimères sur l'accord entre le modèle de dispersion ELMAM et la référence SAPT.	89
3.6	Schéma des différentes méthodes d'intégration du produit des densités électroniques pour le calcul du potentiel d'échange-répulsion ELMAM.	92
3.7	Résultats obtenus par le modèle ELMAM de l'énergie d'échange-répulsion pour différentes méthodes d'intégration du produit des densités, comparés aux valeurs de référence SAPT.	93
3.8	Dépendance de l'accord ELMAM-SAPT pour le potentiel d'échange-répulsion (R^2 en ordonnées) en fonction de l'intervalle d'intégration choisi ($Z_{\text{cut-off}}$ en abscisses).	95
3.9	Résultats obtenus par le modèle ELMAM de l'énergie d'échange-répulsion comparés aux valeurs de référence SAPT pour plusieurs facteurs dépendant de la distance interatomique $f(R_{ij})$	97
3.10	Résultats obtenus par le modèle ELMAM de l'énergie d'échange-répulsion comparés aux valeurs de référence SAPT pour d'autres méthodes d'intégration et avec l'introduction de facteurs dépendant de la distance interatomique.	99
3.11	Résultats obtenus par le modèle ELMAM de l'énergie d'échange-répulsion avec les paramètres empiriques comparés aux valeurs de référence SAPT.	101
3.12	Influence de la composition des dimères sur l'accord entre le modèle d'échange-répulsion ELMAM et la référence SAPT.	103
3.13	Résultats obtenus par le modèle ELMAM de l'énergie de van der Waals sans et avec les paramètres empiriques comparés aux valeurs de référence SAPT.	104
3.14	Résultats obtenus par le modèle ELMAM de l'énergie de van der Waals par combinaison linéaire des termes de dispersion et de répulsion, sans et avec les paramètres empiriques, comparés aux valeurs de référence SAPT.	106
3.15	Influence de la composition des dimères sur l'accord entre le modèle de van der Waals ELMAM et la référence SAPT.	109
4.1	Zones d'influence nucléophile intersectant le site de fixation de NRP1.	140
4.2	Déformation multipolaire de la densité électronique du rétinol.	182
4.3	Zone d'influence nucléophile de l'atome O δ 1 du résidu Asn98 dans la barrière moléculaire stérique ouverte.	183
4.4	Zone d'influence nucléophile de l'atome O δ 1 du résidu Asn98 dans la barrière moléculaire stérique fermée.	184

4.5	Sous-système utilisé dans les calculs d'énergies QM pour modéliser le site de fixation CL352.	185
4.6	Zones d'influence électrophile dans la barrière moléculaire électrostatique.	188
4.7	Zones d'influence nucléophile dans la barrière moléculaire électrostatique.	189

Liste des tableaux

2.1	Vérification du théorème de Gauss dans les zones d'influence électrophile.	61
3.1	Coefficients statistiques pour caractériser l'accord entre le modèle ELMAM et la référence SAPT pour les énergies électrostatique et d'induction.	81
3.2	Coefficients statistiques pour caractériser l'accord entre le modèle ELMAM et la référence SAPT pour l'énergie de dispersion.	85
3.3	Paramètres empiriques dépendant de l'espèce chimique introduits dans le modèle ELMAM d'énergie de dispersion.	86
3.4	Coefficients statistiques pour caractériser l'accord entre le modèle ELMAM avec paramètres empiriques et la référence SAPT pour l'énergie de dispersion.	88
3.5	Coefficients statistiques pour caractériser l'accord entre le modèle ELMAM pour les différentes méthodes d'intégration et la référence SAPT pour l'énergie d'échange-répulsion.	96
3.6	Coefficients statistiques pour caractériser l'accord entre le modèle ELMAM pour les différents facteurs $f(R_{ij})$ dépendant de la distance interatomique R_{ij} et la référence SAPT pour l'énergie d'échange-répulsion.	98
3.7	Paramètres empiriques dépendant de l'espèce chimique introduits dans le modèle ELMAM d'énergie d'échange-répulsion.	100
3.8	Coefficients statistiques pour caractériser l'accord entre le modèle ELMAM avec paramètres empiriques et la référence SAPT pour l'énergie d'échange-répulsion. .	102
3.9	Coefficients statistiques pour caractériser l'accord entre le modèle ELMAM sans paramètre empirique et la référence SAPT pour l'énergie de van der Waals. . . .	105
3.10	Coefficients statistiques pour caractériser l'accord entre le modèle ELMAM avec paramètres empiriques et la référence SAPT pour l'énergie de van der Waals. . .	105
3.11	Coefficients statistiques pour caractériser l'accord entre le modèle ELMAM de combinaison linéaire des termes de dispersion et de répulsion, sans paramètre empirique, et la référence SAPT d'énergie de van der Waals.	107
3.12	Coefficients statistiques pour caractériser l'accord entre le modèle ELMAM de combinaison linéaire des termes de dispersion et de répulsion, avec paramètres empiriques, et la référence SAPT d'énergie de van der Waals.	108
4.1	Comparaison des énergies d'interaction dans le site de fixation CL352 obtenues par méthodes de chimie quantique avec les résultats des potentiels d'interaction ELMAM.	186

4.2	Energies d'interaction entre l'Asn98 et la Thr102 avec la Lys235 et le rétinol à partir des potentiels ELMAM.	187
-----	---	-----

Chapitre 1

Introduction

1.1 Introduction générale

Pourquoi le sel se dissout-il dans l'eau mais pas dans l'huile ? Pourquoi le carbone du diamant est-il plus dur que celui de la mine de crayon de papier ? Pourquoi l'aspirine possède-t-elle un pouvoir anti-inflammatoire ? La réponse à ces questions et bien d'autres se trouve dans l'arrangement spatial des atomes et les interactions qui les lient entre eux. A l'échelle du dixième de nanomètre, les atomes s'assemblent par liaisons covalentes pour former des molécules. Ces molécules sont alors capables d'interagir entre elles pour conférer à la matière ses diverses propriétés. Dans le domaine du vivant, la structure tridimensionnelle peut permettre de déterminer la fonction biologique d'une molécule. C'est pourquoi chimistes, physiciens et biologistes cherchent à comprendre et à modéliser les lois qui régissent les interactions interatomiques et intermoléculaires depuis des décennies.

Au début du XX^e siècle, l'arrivée des méthodes de cristallographie moderne ou radiocristallographie a permis de déterminer expérimentalement les structures atomiques tridimensionnelles de nombreux composés. La découverte de la diffraction des rayons X par des cristaux de M. von Laue [Friedrich *et al.*, 1912, von Laue, 1915], récompensée par le prix Nobel de physique de 1914, suivie des travaux de W. L. Bragg et W. H. Bragg [Bragg, 1913, Bragg et Bragg, 1913], qui ont obtenu le prix Nobel de physique en 1915, sont le point de départ de ces techniques de résolution structurale. Les expériences de diffraction de rayons X fournissent les intensités diffractées par le cristal qui donnent accès aux facteurs de structures permettant de reconstruire la densité électronique de sa maille élémentaire par transformée de Fourier. La disposition spatiale des atomes contenus dans la maille est extraite de ces données et, dans le cas de cristaux moléculaires, révèle la structure tridimensionnelle des molécules présentes dans le système¹. En un peu plus d'un siècle, de nombreuses approches de détermination de structures ont vu le jour. La diffraction des rayons X est toujours l'une des plus fructueuses notamment grâce aux développements de grandes infrastructures expérimentales, les synchrotrons, qui ont permis d'améliorer considérablement la qualité des données expérimentales récoltées, et d'outils informatiques automatisant le traitement de ces données. Le nombre de structures de petites molécules résolues par cette méthode dépasse aujourd'hui le million dans des bases de données telles que la Cambridge

1. Le livre de W. Massa [Massa, 2004] fait partie des ouvrages fréquemment recommandés pour se familiariser à la théorie et aux méthodes de résolution des structures atomiques par diffraction des rayons X.

Structural Database [Groom *et al.*, 2016].

Grâce à l’analyse des structures atomiques, diverses propriétés des petites molécules ont été dévoilées, permettant notamment de mieux comprendre les liaisons chimiques et les interactions non-covalentes. Par exemple, la résolution de la structure de l’hexaméthylbenzène par K. Lonsdale [Lonsdale, 1928] a confirmé expérimentalement la symétrie hexagonale du benzène et renforcé l’idée d’un phénomène de résonance dans les cycles aromatiques. Les propriétés chimiques des molécules sont liées à la distribution des électrons autour des noyaux, qui est accessible expérimentalement grâce à la diffraction des rayons X. En effet, les rayons X sont des ondes électromagnétiques qui interfèrent avec les nuages électroniques, ce qui permet lors d’expériences de diffraction de ces rayons X par un cristal d’effectuer une mesure indirecte de la densité électronique du système. A partir de cette mesure, la structure atomique, c’est-à-dire les positions des noyaux, et le modèle de la distribution des électrons dans la maille sont accessibles et la densité de charge moléculaire peut donc être modélisée. Or, le premier théorème de Hohenberg et Kohn dans le cadre de la théorie de la fonctionnelle de la densité [Hohenberg et Kohn, 1964] implique que les propriétés d’un système dans son état fondamental peuvent être déterminées par la connaissance de sa densité de charge. C’est pourquoi de nombreuses méthodes d’analyse et descripteurs des petites molécules ont été développés sur la base de la modélisation de la distribution de charge du système. Cette distribution de charge peut être modélisée directement à partir des données expérimentales de cristallographie mais aussi sur la base de modèles théoriques de mécanique moléculaire ou de méthodes de chimie quantique. Les descripteurs qui en sont issus sont utilisés pour améliorer la compréhension et permettre la prédiction de propriétés physico-chimiques des petites molécules, par exemple pour la conception de nouveaux matériaux.

Les macromolécules biologiques sont composées de plusieurs milliers d’atomes et la détermination de leurs structures atomiques et densités électroniques est beaucoup plus complexe et ne peut pas atteindre un niveau de précision comparable à ceux obtenus pour les petites molécules [Dauter *et al.*, 1997]. La structure en double hélice de l’ADN a pu être proposée en 1953 par James Watson et Francis Crick [Watson et Crick, 1953], récompensés par le prix Nobel de médecine en 1962, sur la base des clichés de diffraction dont les premiers ont été obtenus par Rosalind Franklin et Maurice Wilkins en 1951. La première structure atomique de protéine a quant à elle été obtenue en 1958 par John Kendrew par diffraction des rayons X sur un cristal de myoglobine [Kendrew *et al.*, 1958], ce qui lui valut le prix Nobel de chimie de 1962 aux côtés de Max Perutz. Aujourd’hui, la base de données des structures de macromolécules biologiques, la Protein Data Bank (PDB) [Berman *et al.*, 2000], contient plus de 200 000 structures déterminées expérimentalement². La majeure partie de ces structures ont été résolues par diffraction des rayons X (85,8%) mais d’autres techniques comme la microscopie électronique (7,1%) et la résonance magnétique nucléaire (6,9%) sont également utilisées. Notons par ailleurs que la PDB recense également maintenant plus d’un million de modèles théoriques de structures de protéines prédits essentiellement par AlphaFold [Jumper *et al.*, 2021], le logiciel de DeepMind basé sur des méthodes d’apprentissage automatique d’intelligence artificielle.

Les méthodes de cristallographie sont centrales dans le domaine de la biologie structurale

2. Les statistiques de la PDB sont disponibles sur le site : <https://www.rcsb.org/stats/summary> et ont été relevés pour ce paragraphe le 23/02/2023.

qui vise à analyser les structures et la dynamique des macromolécules biologiques à l'échelle atomique pour déterminer le lien avec leurs fonctions dans le milieu cellulaire. Outre les aspects fondamentaux de biologie moléculaire, la principale application des études de biologie structurale est la conception de médicaments [Anderson, 2003]. En effet, ces méthodes permettent à la fois d'identifier et analyser une molécule cible, typiquement une protéine, impliquée dans une pathologie donnée et également de rationaliser le développement de médicaments en prédisant les candidats les plus adaptés pour interagir avec la cible. Ces approches ont permis de réaliser d'importants progrès dans le développement de traitements de diverses maladies, comme par exemple dans le cas de pathologies liées au VIH, à certains cancers et à certaines infections bactériennes [Thomas *et al.*, 2017]. Pour cela, la biologie structurale s'appuie à la fois sur des données expérimentales structurales et de biochimie, mais aussi sur des méthodes computationnelles qui nécessitent la modélisation des interactions intermoléculaires, notamment entre une protéine et un ligand. Chez les protéines, trois catégories d'interactions non-covalentes sont communément distinguées [Karshikoff, 2006]³ : les interactions dites de van der Waals, les liaisons hydrogène et les interactions électrostatiques. La connaissance de la structure atomique seule ne suffit pas au développement de descripteurs pour l'analyse et la prédiction de ces interactions, la modélisation de la distribution de charge moléculaire est également nécessaire. Cependant, contrairement aux petites molécules, les données expérimentales de cristallographie des protéines ne permettent pas de construire une densité électronique faisant apparaître les détails nécessaires à la définition de descripteurs précis. Du côté des approches théoriques, les méthodes de mécanique moléculaire sont généralement basées sur des modèles de mécanique classique dépendant de paramètres empiriques qui sont optimisés pour chaque type de système, comme les charges atomiques partielles par exemple. Ces méthodes peuvent donc manquer certains détails de la distribution de charge et s'en retrouver limitées dans le champ d'application de leurs modèles. Quant aux méthodes de chimie quantique, ce sont des méthodes *ab initio* qui décrivent de façon très précise la distribution électronique moléculaire mais les ressources computationnelles qu'elles mobilisent dépendent très fortement du nombre d'atomes du système considéré. Ces méthodes requièrent donc des temps de calcul trop importants pour être appliquées aux macromolécules biologiques complètes. Aussi, pour développer des descripteurs applicables aux protéines, l'enjeu est de proposer un modèle décrivant la distribution de charge dans ces systèmes de façon très précise et avec des moyens computationnels raisonnables. Les méthodes hybrides entre la mécanique quantique et la mécanique moléculaire (QM/MM), qui seront présentées dans la partie 1.4, répondent avec succès à ces enjeux.

Dans ce projet doctoral, nous nous inspirons plutôt des méthodes développées dans le cadre de la cristallographie quantique. Ce domaine, alliant données expérimentales de cristallographie et méthodes théoriques de mécanique quantique, a notamment pour but d'améliorer la détermination et l'analyse des structures atomiques et des densités électroniques et/ou fonctions d'onde pour révéler les propriétés physico-chimiques des systèmes. En particulier, mes travaux sont

3. Les interactions dites hydrophobes jouent également un rôle très important en biochimie, par exemple pour l'existence d'un noyau hydrophobe dans de nombreux repliements de protéines. Néanmoins, étant donnée leur nature entropique, ces interactions sont rarement étudiées explicitement par les méthodes de chimie computationnelle classiques qui s'intéressent généralement aux effets enthalpiques. Par contre, ces effets peuvent être pris en compte de manière implicite, notamment par la modélisation de la surface d'accessibilité au solvant des molécules.

basés sur le modèle multipolaire de Hansen et Coppens [Hansen et Coppens, 1978] permettant de décrire finement les densités électroniques atomiques. Par transférabilité, la densité électronique d’une protéine est décrite à partir des modèles multipolaires de densités électroniques de petites molécules ou de peptides déterminés expérimentalement. Ces méthodes seront davantage détaillées dans la partie 1.2 de ce manuscrit. L’objectif de cette thèse est de développer des descripteurs électrostatiques qui sont issus de cette densité électronique transférée d’origine expérimentale et qui soient applicables aux protéines. Ces descripteurs ont pour ambition de devenir des outils supplémentaires en biologie structurale pour interpréter les structures atomiques des protéines.

Les descripteurs électrostatiques que nous développons reposent sur deux approches distinctes. D’une part, nous nous basons sur l’analyse de la topologie du potentiel électrostatique, directement issu de la densité électronique, pour appréhender spatialement les interactions entre les molécules biologiques. Le potentiel électrostatique permet notamment de localiser les contributeurs aux interactions de nature électrostatique, les sites électrophiles et les sites nucléophiles. Son analyse topologique donne en plus accès à la topographie des lignes de champ électrique qui sont directement reliées aux forces électrostatiques et montre l’étendue spatiale des interactions. D’autre part, nous développons un modèle d’énergie d’interaction intermoléculaire totale qui a l’originalité d’être extrait de la densité électronique transférée d’origine expérimentale. Ce potentiel d’interaction total a pour but d’identifier et de comparer les principaux groupements en quantifiant leurs contributions aux interactions intermoléculaires. En outre, nos deux types de descripteurs aspirent à participer à l’analyse des processus de reconnaissance protéine-ligand, comme la spécificité de substrat d’une enzyme, mais aussi des mécanismes ayant une direction spatiale privilégiée, comme par exemple le transport d’ions entre le milieu extra-cellulaire et le cytoplasme par des protéines spécialisées.

Pour rendre nos descripteurs accessibles et faciles à mettre en œuvre, nous avons implémenté des outils intégrés au logiciel de modélisation moléculaire MoProViewer [Guillot *et al.*, 2014] développé au laboratoire CRM2. Ce logiciel propose de nombreuses fonctionnalités dont la possibilité de travailler avec le modèle multipolaire transférable de la densité électronique et offre déjà plusieurs outils permettant d’appliquer d’autres descripteurs basés sur ce modèle.

Pour déterminer l’étendue des possibilités d’analyse offertes par nos descripteurs, nous les avons appliqués à plusieurs systèmes macromoléculaires largement étudiés. Nous avons analysé deux enzymes : la trypsine qui est une protéase à sérine [Hedstrom, 2002] et la glutathion transférase Chi [Hayes *et al.*, 2005] qui appartient à une superfamille d’enzymes très étudiée au laboratoire CRM2, ainsi que deux protéines membranaires : la neuropiline [Dumond *et al.*, 2020], une glycoprotéine et cible thérapeutique importante pour le traitement de certains cancers, et l’halorhodopsine [Engelhard *et al.*, 2018], appartenant à la famille des rhodopsines qui utilisent la lumière pour effectuer des changements conformationnels liés à leur fonction.

Ce manuscrit de thèse rend compte des travaux de développements méthodologiques et des applications réalisés. Dans un premier temps, cette introduction sera poursuivie par une présentation des méthodes issues du domaine de la cristallographie quantique en insistant sur celles qui permettent de définir le transfert de la densité électronique multipolaire. Puis, les approches

basées sur la topologie des champs scalaires moléculaires et sur la modélisation des énergies d'interaction seront discutées. Dans un deuxième temps, les développements méthodologiques que j'ai effectués seront exposés. D'abord, je présenterai les descripteurs issus de l'analyse topologique du potentiel électrostatique. Ensuite, je détaillerai le développement du potentiel d'interaction total issu de la densité électronique transférée d'origine expérimentale. Pour finir, l'application de ces descripteurs à divers systèmes biologiques sera analysée. Je commencerai par la trypsine, une protéase à sérine dont le mécanisme a été très largement étudié, pour montrer l'avantage d'employer les descripteurs de la topologie du potentiel pour la description des enzymes. L'illustration d'une utilisation de ces outils en biologie structurale sera ensuite proposée avec l'exemple de la neuropiline. Puis, je décrirai une autre enzyme, la glutathion transférase de classe chi, pour laquelle nous avons résolu la première structure atomique appartenant à la famille d'organisme des cyanobactéries, que nous avons caractérisée par le calcul des énergies d'interaction électrostatique [Mocchetti *et al.*, 2022]. Pour caractériser les aspects dynamiques des mécanismes moléculaires, je présenterai la perspective d'application des méthodologies développées pendant ma thèse au pompage d'ions chlorure photo-induit chez les archées halophiles par l'halorhodopsine. Finalement, les conclusions des réalisations effectuées lors de ce projet doctoral et les perspectives qu'elles permettent d'ouvrir seront discutées.

1.2 Cristallographie quantique et transferabilité des paramètres du modèle multipolaire de la densité électronique

La cristallographie quantique est le domaine de recherche dans lequel s'inscrit le développement des descripteurs électrostatiques qui font l'objet de cette thèse. Sa définition va à présent être discutée et les méthodes qui en sont issues seront présentées⁴. En particulier, le modèle multipolaire de Hansen et Coppens de la densité électronique atomique sera détaillé. La transférabilité des paramètres de ce modèle entre atomes de même type ainsi que la possibilité d'obtenir les propriétés électrostatiques des molécules par cette approche seront finalement décrites.

1.2.1 Cristallographie quantique

L'expression « cristallographie quantique » apparaît pour la première fois en 1995 dans l'article de L. Massa, L. Huang et J. Karle [Massa *et al.*, 1995]. Elle définit alors le domaine regroupant les méthodes pour, d'une part, affiner les calculs de mécanique quantique à l'aide de données cristallographiques et pour, d'autre part, améliorer les données de cristallographie et la détermination des modèles moléculaires par des approches de mécanique quantique. Ce terme formalise le lien historique entre la cristallographie et la mécanique quantique [Macchi, 2020]. En effet, alors que la physique quantique en était à ses débuts, des physiciens tels que P. Debye et A. Compton ont perçu l'arrivée des techniques de diffraction des rayons X par les cristaux comme une opportunité pour améliorer leur compréhension de la distribution des électrons dans les atomes [Debye, 1915, Compton, 1915]. Par exemple, l'expérience de diffraction des rayons

4. Les développements effectués dans le cadre de ce projet doctoral visant à être appliqués aux molécules biologiques, seule la dimension moléculaire des méthodes présentées sera discutée dans cette partie ainsi que dans l'ensemble de ce manuscrit, bien que la plupart de ces méthodes soient également applicables aux composés non-moléculaires de types ioniques, organo-métalliques ou autre.

X par le cristal de diamant [Bragg et Bragg, 1913] a permis d'observer la densité électronique autour d'un atome de carbone lié à quatre autres atomes et a ainsi montré le potentiel de la cristallographie pour comprendre les liaisons chimiques. Par ailleurs, les expériences de diffraction des électrons, une autre méthode de cristallographie, ont appuyé la théorie de la dualité onde-corpuscule proposée par Louis de Broglie [de Broglie, 1926], pour laquelle il a été récompensé par le prix Nobel de physique en 1929. Les avancées théoriques de la mécanique quantique ont aussi été à l'origine d'importants progrès dans l'interprétation des données issues des expériences de diffraction de rayons X. Le traitement quantique de la diffusion a permis de mieux comprendre la corrélation entre la densité électronique et les intensités des rayons X diffusés par les cortèges électroniques des atomes dans le cristal. Les facteurs de forme atomiques, qui sont nécessaires pour le calcul des facteurs de structure⁵, ont été obtenus par des méthodes de mécanique quantique sur la base de modèles théoriques d'atomes isolés.

De nos jours, la cristallographie quantique peut être définie comme le domaine scientifique qui a pour objectif d'étudier les propriétés et les phénomènes en lien avec l'état cristallin et qui ne peuvent être expliqués que par les lois de la mécanique quantique [Genoni *et al.*, 2018, Macchi, 2020]. De nombreuses approches ont vu le jour pour répondre aux problématiques qui en découlent. Dans la volonté de déterminer expérimentalement la fonction d'onde d'un système, les méthodes "X-ray Constrained Wavefunctions" (XCW) [Jayatilaka, 1998, Jayatilaka et Grimwood, 2001] ont été mises au point, conciliant des données de diffraction des rayons X et calculs quantiques. En effet, ces méthodes XCW permettent d'affiner la fonction d'onde de manière à reproduire les facteurs de structure expérimentaux tout en minimisant l'énergie du système.

D'autres approches ont pour objectif de décrire finement la densité électronique moléculaire à partir des données de cristallographie. En effet, les méthodes classiques d'affinement de structures atomiques font appel au modèle des atomes indépendants (IAM pour "Independent Atom Model") pour modéliser la distribution électronique. Dans ce modèle, les atomes sont neutres et indépendants et leurs densités électroniques ont une symétrie sphérique, ce qui permet de n'avoir que les coordonnées atomiques et les paramètres d'agitation thermique à affiner contre les intensités ou les facteurs de structure expérimentaux. Cependant, le modèle IAM ne permet pas de tenir compte des transferts de charge ni de représenter les électrons des liaisons chimiques ou des doublets non-liants. De plus, les longueurs des liaisons chimiques faisant intervenir un atome d'hydrogène obtenues par ce modèle sont trop courtes par rapport aux valeurs de références qui sont fournies par les expériences de diffraction des neutrons⁶. L'approche d'affinement HAR pour "Hirshfeld Atom Refinement" [Jayatilaka et Dittrich, 2008, Capelli *et al.*, 2014] va au-delà du modèle IAM en proposant pour le calcul des facteurs de structure de remplacer les facteurs de formes atomiques classiques, qui sont calculés à partir d'atomes isolés, par des facteurs de formes atomiques asphériques, obtenus à partir des atomes dans leur environnement moléculaire et calculés à chaque pas de la procédure itérative d'affinement. Ces facteurs de formes asphé-

5. Le facteur de structure $F(h, k, l)$ est défini à partir des facteurs de diffusion atomiques f_j et des positions atomiques $\mathbf{r}_j = (x_j, y_j, z_j)$ selon : $F(h, k, l) = \sum_j f_j e^{i2\pi(hx_j + ky_j + lz_j)}$, où l'indice j porte sur les atomes de la maille élémentaire.

6. L'ouvrage [Bacon, 1975] est souvent proposé comme référence pour une introduction aux méthodes cristallographiques reposant sur la diffraction des neutrons.

riques sont déterminés par méthodes *ab initio* de chimie quantique en utilisant la partition de Hirshfeld [Hirshfeld, 1977] pour subdiviser la molécule en "atomes de Hirshfeld" et décomposer la densité électronique moléculaire en densités atomiques asphériques. Les paramètres atomiques ainsi obtenus sont très précis mais le fait que les facteurs de forme soient calculés à chaque itération de l'affinement implique que le coût computationnel de cette approche augmente fortement avec la taille du système étudié, ce qui compromet son application directe aux macromolécules biologiques.

Une autre méthode permettant de dépasser les limitations du modèle IAM est le modèle multipolaire de la densité électronique atomique de Hansen et Coppens [Hansen et Coppens, 1978]. Ce modèle propose des fonctions analytiques paramétrées capables de rendre compte de l'asphéricité de la densité électronique atomique due aux liaisons chimiques et à la présence de paires d'électrons non-liantes notamment. C'est sur la base du modèle de Hansen et Coppens que sont construites les distributions électroniques utilisées pour développer les descripteurs électrostatiques développés dans cette thèse. Je vais donc à présent décrire ce modèle plus en détail.

1.2.2 Modèle multipolaire de Hansen et Coppens

Le modèle IAM

Pour décrire la densité électronique moléculaire, un des modèles les plus simples est le modèle IAM mentionné dans la section précédente. Dans ce modèle, des fonctions gaussiennes ou des fonctions de Slater $1s$ sont utilisées pour modéliser la distribution isotrope des électrons autour du noyau. Il est classiquement employé lors d'affinements structuraux pour lesquels les données de diffraction des rayons X disponibles n'atteignent pas une résolution subatomique car il permet de limiter les paramètres à affiner aux positions atomiques et aux paramètres d'agitation thermique. Néanmoins, cette limitation se fait au prix de la précision du modèle car la symétrie sphérique qu'il impose à la densité électronique atomique l'empêche de caractériser certains aspects tels que les électrons des liaisons chimiques et des doublets non-liants. Les nuages électroniques ne sont pas sphériques, car ils sont déformés par les interactions interatomiques notamment, et les atomes dans les molécules ne sont pas neutres.

Une variante du modèle IAM est le modèle des diffuseurs interatomiques IAS (pour "Interatomic Scatterers") dans lequel la densité électronique sphérique est modélisée autour des noyaux mais aussi autour de diffuseurs artificiels localisés sur des concentrations locales d'électrons tels que les liaisons covalentes [Hellner, 1977, Afonine *et al.*, 2007]. Plus récemment, le modèle des atomes sphériques réels et virtuels (VIR) [Nassour *et al.*, 2017] a repris cette idée en ajoutant des charges partielles aux centres de diffusion interatomiques qui sont ajustables contre les données de diffraction des rayons X.

Le formalisme kappa

Une autre évolution fut le formalisme kappa [Coppens *et al.*, 1979] qui introduit une séparation des descriptions des électrons de cœur et des électrons de valence. La densité électronique

atomique $\rho_{\text{atom}}(r)$ s'exprime alors sous la forme de deux termes :

$$\rho_{\text{atom}}(r) = \rho_{\text{core}}(r) + P_{\text{val}}\kappa^3\rho_{\text{val}}(\kappa r). \quad (1.1)$$

Deux nouveaux paramètres atomiques à affiner apparaissent dans ce modèle : la population de valence P_{val} et le coefficient d'extension-contraction κ . Les termes $\rho_{\text{core}}(r)$ et $\rho_{\text{val}}(r)$ sont les densités électroniques atomiques à symétrie sphérique de cœur et de valence obtenues par calculs théoriques de chimie quantique. Les fonctions orbitales calculées par C. Roetti et E. Clementi [Roetti et Clementi, 1974] sont souvent utilisées pour obtenir ces densités électroniques. Comme dans le modèle IAM, la densité électronique atomique décrite par le formalisme kappa est isotrope mais l'introduction des deux paramètres supplémentaires le rend plus réaliste car la charge totale de l'atome n'est plus nécessairement nulle et les transferts de charge entre atomes sont permis. La population de valence P_{val} permet en effet d'ajuster le nombre d'électrons dans la couche de valence de l'atome et est liée à la notion de charge partielle atomique. Le coefficient κ module la dépendance radiale de la densité de valence sphérique en permettant son extension pour $\kappa < 1$ et sa contraction pour $\kappa > 1$. Dans les figures 1.1a et 1.1b, la comparaison entre le modèle IAM et le formalisme kappa est illustrée dans la molécule N-méthylacétamide⁷. La principale différence entre ces deux modèles se trouve sur les atomes d'hydrogène des groupements méthyles pour lesquels une densité significative apparaît dans le formalisme kappa mais pas dans le modèle IAM.

Le modèle multipolaire de Hansen et Coppens

Le modèle multipolaire de Hansen et Coppens [Hansen et Coppens, 1978] ajoute au formalisme kappa la contribution non-sphérique des électrons de valence grâce à un troisième terme de forme multipolaire dans la description de la densité électronique atomique :

$$\rho_{\text{atom}}(r, \theta, \varphi) = \rho_{\text{core}}(r) + P_{\text{v}}\kappa^3\rho_{\text{val}}(\kappa r) + \sum_{l=0}^{l_{\text{max}}} \kappa'^3 R_l(\kappa' r) \sum_{m=-l}^{+l} P_{lm} d_{lm}(\theta, \varphi). \quad (1.2)$$

Les fonctions $R_l(r)$ sont les fonctions radiales de Slater et les fonctions $d_{lm}(\theta, \varphi)$ sont les fonctions harmoniques sphériques réelles. Les coordonnées (r, θ, φ) sont définies dans une base cartésienne orthogonale centrée sur le noyau et qui permet de tirer profit des symétries locales des liaisons chimiques pour réduire le nombre de paramètres du modèle. Les paramètres P_{lm} sont les populations de valence multipolaires et le paramètre κ' est le coefficient d'extension-contraction multipolaire. Le développement multipolaire est généralement tronqué aux quadrupôles ($l_{\text{max}} = 2$) pour les atomes d'hydrogène, aux octupôles ($l_{\text{max}} = 3$) pour les atomes de couche de valence $2s2p$ (carbone, oxygène, azote) et aux hexadécapôles ($l_{\text{max}} = 4$) pour les atomes plus lourds. Ces mul-

7. La géométrie de la molécule N-méthylacétamide est extraite du jeu de données NENCI2021 [Sparrow *et al.*, 2021]. J'ai choisi cette molécule pour illustrer la plupart des concepts abordés dans ce manuscrit car elle reproduit la liaison peptidique, la liaison covalente qui lie les acides aminés entre eux pour former la chaîne principale des protéines. Dans la suite du manuscrit, cette molécule apparaîtra en complexe avec la molécule de méthanol dans une géométrie d'équilibre également extraite du jeu de donnée NENCI2021. Les propriétés électrostatiques qui seront représentées dans ce complexe ont été calculées à partir de la densité électronique moléculaire reconstruite par transfert des paramètres multipolaires ELMAM2 (voir section 1.2.3 pour plus de détails) et à l'aide de la suite logiciel MoPro [Jelsch *et al.*, 2005].

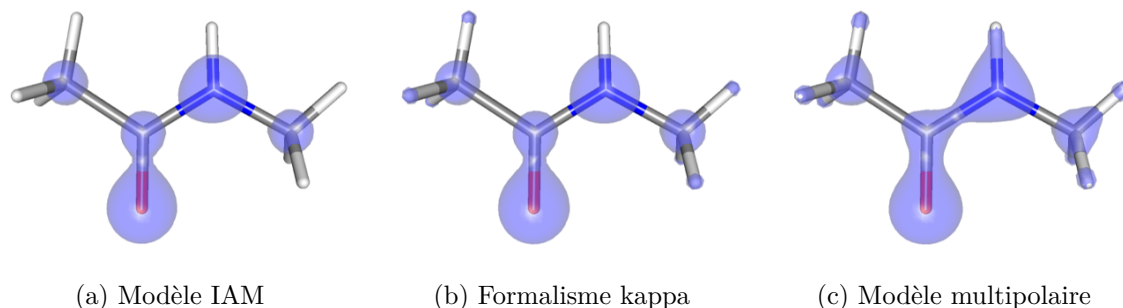


FIGURE 1.1 – Comparaison des densités électroniques obtenues par le modèle IAM, le formalisme kappa et le modèle multipolaire.

L'isosurface $\rho(\mathbf{r}) = 2,0 \text{ e.}\text{\AA}^{-3}$ de la densité électronique moléculaire du N-méthylacétamide est représentée en bleu dans le cadre (a) du modèle IAM (Independent Atom Model), (b) du formalisme kappa et (c) du modèle multipolaire. Le formalisme kappa permet l'extension des densités électroniques des atomes d'hydrogène (blancs) des groupements méthyle qui ne sont pas visibles dans le modèle IAM. Les densités électroniques des atomes d'oxygène (rouge) et d'azote (bleu) sont également légèrement plus étendues dans le formalisme kappa que dans le modèle IAM et portent alors des charges partielles reflétant leurs électronégativités. Le modèle multipolaire ajoute les déformations de la densité dues aux liaisons chimiques et aux paires d'électrons non-liantes de l'atome d'oxygène.

tipôles permettent de décrire les déformations anisotropes de la densité électronique atomique dans les directions des liaisons chimiques et des paires d'électrons non-liantes. La comparaison du modèle multipolaire avec les modèles à symétrie sphérique est illustrée par la figure 1.1c dans laquelle la contribution multipolaire est particulièrement visible sur la liaison peptidique tandis qu'elle n'apparaît pas dans le modèle IAM ni dans le formalisme kappa. La figure 1.2 représente les contributions des trois termes du modèle multipolaire à la densité électronique atomique totale. Les distributions sphériques des électrons de cœur (figure 1.2a) et de valence (figure 1.2b) sont centrés autour des noyaux tandis que les déformations multipolaires (figure 1.2c) mettent en évidence les concentrations d'électrons sur les liaisons covalentes et sur les doublets non-liants de l'atome d'oxygène ainsi que les déplétions d'électrons autour des atomes d'hydrogène, notamment.

La notion de pseudo-atome multipolaire introduit la description d'un atome dans son environnement moléculaire par la définition de la distribution asphérique de ses électrons selon le modèle multipolaire. La densité électronique multipolaire d'une molécule est alors définie comme l'union des densités électroniques des pseudo-atomes qui la composent.

Le modèle multipolaire introduit beaucoup plus de paramètres à affiner que le modèle IAM et nécessite donc de disposer de données expérimentales de très bonne qualité, mesurées à température cryogénique et de résolution subatomique. Bien qu'il existe quelques structures de macromolécules biologiques recensées dans la PDB qui ont été affinées à une résolution inférieure à $0,6\text{\AA}$ [Jelsch *et al.*, 2000, Brzezinski *et al.*, 2011, Schmidt *et al.*, 2011, Hirano *et al.*, 2016], ces systèmes conservent une flexibilité et une dynamique importantes dans l'état cristallin qui empêchent l'affinement de l'ensemble des paramètres du modèle multipolaire. En revanche, ces paramètres devraient être similaires entre pseudo-atomes de même type, ce qui permet de transférer ceux qui ont été obtenus à très haute résolution pour de petites molécules vers des

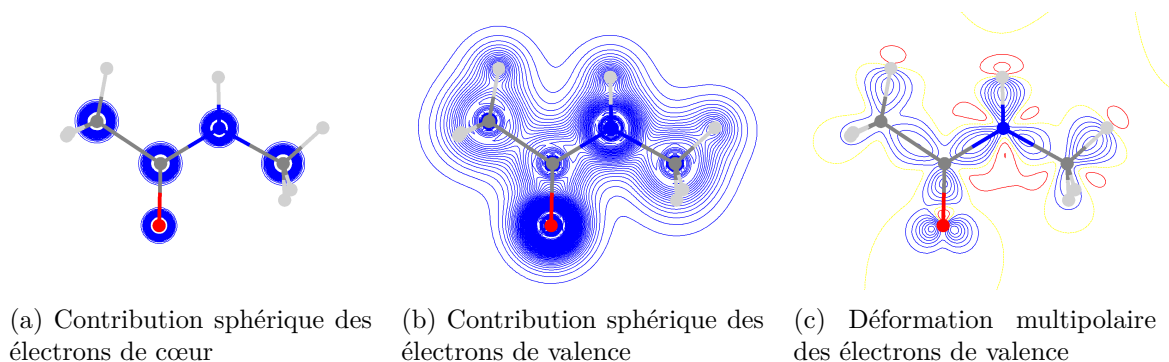


FIGURE 1.2 – Illustration des trois contributions du modèle multipolaire de la densité électronique : électrons de cœur à symétrie sphérique, électrons de valence à symétrie sphérique et déformations multipolaires des électrons de valence.

Les isocontours, à intervalle de densité $\delta\rho(\mathbf{r}) = 0,1 \text{ e.}\text{\AA}^{-3}$, correspondant aux trois contributions du modèle multipolaire à la densité électronique moléculaire du N-méthylacétamide sont représentés par des lignes continues dans le plan de la liaison peptidique. Les contours bleus représentent des isovaleurs positives de la densité, les contours rouges des isovaleurs négatives et les contours jaune l'isovaleur nulle. (a) Les distributions sphériques des électrons de cœur sont concentrées autour des noyaux. (b) Les contributions sphériques des électrons de valence prennent en compte la présence de charges partielles notamment sur les atomes d'oxygène et d'hydrogène dans ce plan. Notons que la plupart des atomes d'hydrogène ne sont pas dans ce plan de cette figure. (c) Les déformations multipolaires font apparaître des zones d'accumulation d'électrons, sur les liaisons chimiques et aux positions des doublets non-liants de l'atome d'oxygène, ainsi que des zones de déplétion d'électrons notamment à proximité des atomes d'hydrogène dans la direction opposée à celle de la liaison covalente.

structures de macromolécules biologiques. Cette approche est utilisée pour améliorer l'affinement des structures en prenant en compte la déformation asphérique de la densité électronique tout en gardant les paramètres du modèle multipolaire fixes lorsque la structure est obtenue à résolution proche de la résolution atomique. De plus, un modèle fin de la densité électronique moléculaire peut tout de même être reconstruit par transfert des paramètres multipolaires afin de dériver avec précision les propriétés électrostatiques du système. En pratique, ce principe de transférabilité est mis en œuvre grâce à l'existence de bases de données de paramètres du modèle multipolaire qui vont être présentées dans la section suivante.

1.2.3 Transférabilité des paramètres de densité électronique atomique du modèle multipolaire

Il a été montré que les paramètres atomiques P_{val} , κ , $P_{l,m}$ et κ' du modèle multipolaire de la densité électronique apparaissant dans l'équation 1.2 sont très similaires pour des atomes appartenant à un même type atomique [Pichon-Pesme *et al.*, 1995]. Un type atomique est défini par un ensemble de propriétés chimiques - incluant notamment la nature de l'élément chimique, son environnement covalent et sa géométrie dans la molécule - qui permet de regrouper les pseudo-atomes multipolaires comparables apparaissant dans des molécules différentes. Les paramètres du modèle multipolaire sont donc transférables entre pseudo-atomes d'un même type atomique. Cette approche a été introduite en 1991 avec l'étude de C. Pratt-Brock, J. D. Dunitz et F. L. Hirshfeld [Brock *et al.*, 1991] dans laquelle ils ont transféré les populations de valence des pseudo-atomes du pérylène aux pseudo-atomes équivalents du naphthalène et de l'anthracène dans

le but d'améliorer l'affinement et la détermination des paramètres de déplacement thermique. Quelques années plus tard, V. Pichon-Pesme, C. Lecomte et H. Lachekar [Pichon-Pesme *et al.*, 1995] ont proposé de regrouper dans une base de données les paramètres du modèle multipolaire de divers types atomiques affinés contre des facteurs de structure obtenus par diffraction des rayons X à très haute résolution. Ainsi, ces paramètres peuvent être réutilisés pour améliorer l'affinement d'autres composés. En effet, si les données cristallographiques ne sont pas de qualité et de résolution suffisantes, l'affinement de l'ensemble des paramètres du modèle multipolaire ne peut pas être réalisé. Une alternative est alors de transférer ces paramètres multipolaires pour affiner la structure en prenant tout de même en compte l'asphéricité de la densité électronique afin d'améliorer la qualité des autres paramètres du modèle, c'est-à-dire les positions atomiques et les paramètres thermiques. Le terme affinement TAAM (pour "Transferable Aspherical Atom Model") a été introduit par K. Woźniak et ses collègues [Bąk *et al.*, 2009] pour qualifier ce type d'affinement. L'autre application importante issue du transfert des paramètres du modèle multipolaire est le calcul des propriétés électrostatiques. La densité électronique d'une molécule peut être modélisée uniquement à partir des positions de ses atomes en la reconstruisant dans une approche "LEGO" à partir des densités transférées des pseudo-atomes. La densité moléculaire ainsi obtenue est de qualité comparable aux densités calculées par méthodes de chimie quantique [Meyer *et al.*, 2016b], elle est capable de décrire les détails fins de la structure électronique. De nombreuses propriétés peuvent donc être dérivées avec précision de cette densité transférée comme par exemple le potentiel électrostatique ou l'énergie d'interaction électrostatique.

Les bases de données de paramètres multipolaires

La première base de données de paramètres du modèle multipolaire fut la librairie ELMAM (Experimental Library of Multipolar Atom Model), développée par les équipes du laboratoire CRM2 et qui propose des paramètres d'origine expérimentale [Pichon-Pesme *et al.*, 1995]. Pour chaque type atomique, les paramètres correspondants ont été affinés contre des facteurs de structure issus de la diffraction des rayons X à très haute résolution sur plusieurs dizaines de cristaux moléculaires. Ces paramètres ont ensuite été moyennés pour gommer les influences des interactions intermoléculaires. Il en résulte un ensemble de paramètres transférables à de nombreux types de pseudo-atomes compatibles. A l'origine, la librairie ELMAM contenait essentiellement des paramètres obtenus à partir de cristaux de peptides [Jelsch *et al.*, 1998, Pichon-Pesme *et al.*, 2000] grâce auxquels des affinements TAAM de protéines ont pu être réalisés : [Jelsch *et al.*, 2000, Housset *et al.*, 2000, Guillot *et al.*, 2001, Guillot *et al.*, 2008, Schmidt *et al.*, 2011]. A présent, la librairie ELMAM2 [Domagała *et al.*, 2012] contient les paramètres de tous les types atomiques présents dans les 20 acides aminés protéinogènes pour l'affinement de protéines [Held et van Smaalen, 2014, Zarychta *et al.*, 2015, Howard *et al.*, 2016, Hirano *et al.*, 2016], ainsi que de nombreux types atomiques rencontrés couramment dans les composés organiques. La suite logiciel MoPro [Jelsch *et al.*, 2005, Guillot *et al.*, 2014], également développée au sein du laboratoire CRM2, fournit des outils faciles à utiliser pour réaliser les affinements TAAM et pour le calcul des propriétés électrostatiques à partir du transfert des paramètres multipolaires contenus dans la librairie ELMAM2.

Au début des années 2000, T. Koritsanszky, A. Volkov et P. Coppens [Koritsanszky *et al.*,

2002] ont proposé de créer une base de données de paramètres du modèle multipolaire d'origine théorique. Ils ont ainsi développé, avec P. M. Dominiak notamment, la base de données UBDB ("University at Buffalo pseudoatom DataBank") [Dominiak *et al.*, 2007, Jarzemska et Dominiak, 2012, Kumar *et al.*, 2019] qui est aujourd'hui incluse dans la base de données MATTS ("Multipolar Atom Types from Theory and Statistical clustering") [Jha *et al.*, 2022, Rybicka *et al.*, 2022]. Les paramètres UBDB/MATTS ont été affinés contre des facteurs de structure théoriques dérivés de densités électroniques calculées par des méthodes *ab initio* de chimie quantique, sur la base des géométries des molécules extraites de la Cambridge Structural Database [Groom *et al.*, 2016]. Comme ceux de la librairie ELMAM2, les paramètres UBDB/MATTS de chaque type atomique ont été moyennés sur un échantillon issu de plusieurs molécules contenant des pseudo-atomes correspondant à ce type atomique.

La base de données GID ("Generalized Invariom Database") [Dittrich *et al.*, 2004, Dittrich *et al.*, 2006, Dittrich *et al.*, 2013] propose également des paramètres du modèle multipolaire d'origine théorique. Les types atomiques de cette base de données sont appelés Invarioms ou atomes invariants et sont déterminés de façon unique, c'est-à-dire que leurs paramètres ne sont pas moyennés sur plusieurs molécules contrairement aux autres approches. En effet, pour chaque Invariom, une unique structure est modélisée en faisant apparaître explicitement ses plus proches voisins et en remplaçant ses voisins plus éloignés par des atomes d'hydrogène. Les facteurs de structure théoriques sont obtenus à partir de ces modèles moléculaires par calculs de chimie quantique et sont utilisés pour affiner les paramètres du modèle multipolaire GID.

Applications du transfert des pseudo-atomes multipolaires

La librairie ELMAM2 est la seule à être basée sur des données expérimentales mais l'avantage des bases de données de paramètres d'origine théorique MATTS et GID est de ne pas être limitées en termes de systèmes d'application, ce qui leur permet de fournir des paramètres pour la plupart des types atomiques apparaissant dans les petites molécules organiques, les protéines, l'ARN et l'ADN ainsi que les ions présents dans les systèmes biologiques. Le premier objectif commun à ces trois approches est d'améliorer la résolution des structures atomiques en permettant la modélisation de la densité électronique non-sphérique, sans ajouter de paramètres à affiner. Cet affinement TAAM permet de déconvoluer les effets d'agitation thermique des déformations de la densité électronique de valence et donc d'obtenir des paramètres de déplacements thermiques de meilleure qualité que ceux obtenus par affinement IAM [Bak *et al.*, 2011]. Les longueurs de liaison impliquant des atomes d'hydrogène sont plus proches des valeurs de référence de diffraction neutrons et les électrons des doublets non-liants et des liaisons chimiques sont décrits par la modélisation TAAM la densité électronique. Généralement, les affinements TAAM sont appliqués aux données issues des expériences de diffraction des rayons X mais la base de données MATTS permet de réaliser cet affinement aussi à partir de données de diffraction des électrons et de microscopie électronique [Gruza *et al.*, 2020, Jha *et al.*, 2021, Kulik *et al.*, 2022]. Au-delà de l'affinement de structures, le transfert des paramètres du modèle multipolaire permet de modéliser une densité électronique moléculaire précise à partir de laquelle certaines propriétés électrostatiques peuvent être extraites. En effet, l'analyse de la densité électronique transférée permet de caractériser les interactions dans les complexes enzymatiques [Liebschner

et al., 2011, Malińska *et al.*, 2014, Held et van Smaalen, 2014, Zarychta *et al.*, 2015]. De plus, l'énergie d'interaction électrostatique peut être estimée à partir du modèle multipolaire de la densité électronique [Li *et al.*, 2002, Volkov *et al.*, 2004b, Dominiak *et al.*, 2007, Dominiak *et al.*, 2009, Fournier *et al.*, 2009, Jarzemska et Dominiak, 2012, Kumar *et al.*, 2014b, Malinska *et al.*, 2015, Malinska et Dauter, 2016, Kumar et Dominiak, 2021, Vuković *et al.*, 2021]. Il permet aussi de calculer le potentiel électrostatique moléculaire [Muzet *et al.*, 2003, Malińska *et al.*, 2014, Malinska et Dauter, 2016, Dittrich et Luger, 2017, Weatherly *et al.*, 2021] et les lignes de champ électrique [Howard *et al.*, 2016]. Les trois bases de données ELMAM, UBDB et GID ont été comparées entre elles, à des méthodes expérimentales directes et à des méthodes purement théoriques, à la fois pour l'affinement structurale et pour le calcul des propriétés électrostatiques [Bak *et al.*, 2011]. Par ailleurs, il existe également des bases de données permettant de reconstruire la fonction d'onde moléculaire à partir de fragments tels que les ELMO ("Extremely Localized Molecular Orbitals") [Sironi *et al.*, 2007] qui sont des orbitales moléculaires strictement localisées sur de petits fragments moléculaires comme un atome seul, une liaison covalente ou groupement chimique. La transférabilité des ELMO a été établie et comparée avec celle des pseudo-atomes du modèle multipolaire [Meyer *et al.*, 2016a, Meyer *et al.*, 2016b, Meyer et Genoni, 2018], ce qui leur permet d'être utilisés pour les affinements HAR ("Hirshfeld Atom Refinement") [Malaspina *et al.*, 2019] mentionnés dans la section 1.2.1.

Densité électronique transférée et interactions intermoléculaires

Les paramètres multipolaires de ces bases de données ne tiennent pas compte des influences intermoléculaires de l'environnement local. La densité électronique reconstruite à partir du transfert de ces paramètres est dite non-perturbée par l'environnement non-covalent, c'est-à-dire que les effets de polarisation mutuelle entre les densités électroniques de deux molécules en interaction sont sous-estimés, voire négligés. Or, ces effets jouent un rôle important pour la compréhension de certains processus, notamment pour l'étude des interactions dans les complexes protéine-ligand. Pour retrouver ces aspects intermoléculaires, la base de données ELMAM2 fournit, en plus des paramètres atomiques du modèle multipolaire d'origine expérimentale, des polarisabilités atomiques anisotropes moyennes d'origine théorique et qui sont transférables entre pseudo-atomes similaires [Leduc *et al.*, 2019]. La polarisabilité est la grandeur qui traduit la capacité de la densité électronique à se déformer en réponse à l'application d'un champ électrique extérieur. Pour chaque type atomique i , le tenseur de polarisabilité α_i de la librairie ELMAM2 a été dérivé du moment dipolaire μ obtenu par l'intégration de la quantité $\mathbf{r} \times \rho(\mathbf{r})$ dans le bassin atomique⁸, où $\rho(\mathbf{r})$ est calculée par méthode de chimie quantique et soumise à un champ électrique externe $\mathbf{E}$. Les éléments du tenseur de polarisabilité α_i sont alors définis comme : $[\alpha_i]_{lm} = \frac{\partial \mu_l}{\partial E_m}$, où les indices $l, m = \{x, y, z\}$ portent sur les composantes des vecteurs μ et $\mathbf{E}$.

Grâce à la procédure développée par T. Leduc [Leduc *et al.*, 2019], le transfert de ces tenseurs de polarisabilité α_i est employé pour polariser la densité électronique d'un groupement d'atomes A sous l'influence d'un groupement d'atomes B et inversement, A et B pouvant être deux molécules distinctes, un site actif d'enzyme et un substrat ou encore deux domaines d'une même

8. Les bassins atomiques sont définis dans le cadre de la théorie Atoms in Molecule de Bader qui sera présentée dans la section 1.3.1.

protéine par exemple. Dans cette procédure auto-cohérente, les moments dipolaires atomiques $\mu_{i \in A}$ induits sur le groupement A sont calculés à chaque itération d’après la relation :

$$\mu_{i \in A} = \alpha_{i \in A} \cdot E_B(\mathbf{r}_{i \in A}), \quad (1.3)$$

où $E_B(\mathbf{r}_{i \in A})$ est le champ électrique généré par le groupement B et ressenti à la position de l’atome i du groupement A . Les moments dipolaires $\mu_{j \in B}$ induits sur les atomes du groupement B sont calculés de façon similaire. Ces dipôles induits sont ensuite injectés dans les populations dipolaires $P_{1,-1}$, $P_{1,0}$ et $P_{1,1}$ du modèle multipolaire de la densité électronique pour l’itération suivante. Ces étapes de calculs sont répétées jusqu’à ce que le critère de convergence soit atteint, c’est-à-dire jusqu’à ce que les normes des dipôles induits μ_i soient toutes inférieures à 10^{-4} e.Å. Cet algorithme de polarisation a été implémenté [Leduc *et al.*, 2019] dans le logiciel MoProViewer [Guillot *et al.*, 2014] de la suite MoPro.

En résumé,

les méthodes de cristallographie quantique sont issues de la combinaison des données expérimentales de cristallographie et des modèles de mécanique quantique. Parmi ces méthodes, certaines ont été développées pour modéliser la densité électronique ou la fonction d’onde d’un système afin d’en déterminer les propriétés. En particulier, le modèle multipolaire de la densité électronique permet de décrire les détails fins de la distribution des électrons comme les liaisons chimiques et les paires d’électrons non-liantes. Les paramètres de ce modèle peuvent être affinés contre les données issues des expériences de diffraction des rayons X quand celles-ci sont de qualité et de résolution suffisantes. Lorsque ce n’est pas le cas, ces paramètres peuvent être transférés depuis des bases de données d’origine expérimentale, c’est le cas de la librairie ELMAM2, ou d’origine théorique, pour les bases de données MATTS et GID. Le transfert des paramètres multipolaires permet de réaliser des affinements TAAM, c’est-à-dire des affinements où les déformations asphériques de la densité électronique sont modélisées sans ajouter de paramètres à affiner. Cette transférabilité permet également de reconstruire la densité électronique d’un système afin d’en dériver ses propriétés électrostatiques telles que le potentiel électrostatique, le champ électrique ou l’énergie d’interaction électrostatique. Ces approches ont déjà été mises en oeuvre pour l’étude de systèmes protéine-ligand. Le transfert de tenseurs de polarisabilité ajoute la possibilité de modéliser la polarisation mutuelle des densités électroniques de deux groupements d’atomes en interaction. Dans la suite de ce manuscrit, les descripteurs issus des champs scalaires moléculaires que sont la densité électronique et le potentiel électrostatique seront discutés. Puis, des méthodes de modélisation de l’énergie d’interaction seront également présentées.

1.3 Champs scalaires moléculaires et analyse topologique

La diffraction des rayons X par les cortèges électroniques des atomes dans les cristaux moléculaires est utilisée pour déterminer les structures atomiques des molécules mais ces expériences permettent également d’aller plus loin et de proposer un modèle de la densité électronique moléculaire. Comme exposé dans la section 1.2.2, le modèle multipolaire de Hansen et Cop-

pens [Hansen et Coppens, 1978] permet de décrire de façon très précise la densité électronique expérimentale. Ce modèle fin permet d'extraire les propriétés d'une molécule grâce à l'existence de nombreux descripteurs basés sur la densité électronique et sa topologie. Le potentiel électrostatique moléculaire, qui est un champ scalaire particulièrement riche en information sur les interactions intermoléculaires, est également défini à partir de la densité électronique moléculaire et de la distribution des charges positives correspondant aux noyaux. Le premier type de descripteur que nous avons développé dans le cadre de ce projet doctoral pour la caractérisation des interactions entre molécules biologiques est justement basé sur l'analyse topologique du potentiel électrostatique. Dans cette partie, les deux champs scalaires moléculaires, densité électronique et potentiel électrostatique, ainsi que leur analyse topologique vont être présentés.

1.3.1 Densité électronique moléculaire

Le lien entre densité électronique et propriétés moléculaires a été formalisé d'un point de vue théorique dans le cadre du premier théorème de Hohenberg et Kohn [Hohenberg et Kohn, 1964] et de la théorie de la fonctionnelle de la densité (DFT). Grâce aux méthodes de cristallographie par diffraction des rayons X, ce lien est également investigué du point de vue expérimental dans les études de densité de charge⁹ pour comprendre les liaisons chimiques covalentes et non-covalentes. Une des premières études de densité de charge expérimentale fut réalisée par P. Coppens en 1967 [Coppens, 1967] dans laquelle il a analysé les déformations asphériques de la densité électronique dans la molécule s-triazine dues aux liaisons chimiques, aux paires d'électrons non-liantes des atomes d'azote et au caractère aromatique de la molécule. Ces études nécessitent des données expérimentales de très haute résolution et de très bonne qualité pour pouvoir être appliquées. C'est pourquoi il a fallu attendre les années 2000 et les avancées techniques des infrastructures expérimentales et du traitement informatique des données de diffraction des rayons X pour que l'analyse de la densité de charge expérimentale puisse devenir une étape à part entière de l'étude cristallographique des petites molécules [Coppens, 2005]. Pour les macromolécules biologiques, la densité électronique moléculaire peut être reconstruite grâce au transfert des paramètres du modèle multipolaire d'origine expérimentale de la librairie ELMAM2 [Domagała *et al.*, 2012], comme expliqué dans la section 1.2.3.

La théorie quantique des atomes dans les molécules

Les études de densité de charge expérimentale (ou théorique) sont généralement basées sur l'application de la théorie quantique des atomes dans les molécules ou QTAIM ("Quantum Theory of Atoms In Molecules") développée par R. Bader [Bader, 1990]. La QTAIM repose sur l'analyse topologique de la densité électronique moléculaire $\rho(\mathbf{r})$ pour décrire la molécule en

9. Les méthodes de densité de charge sont les fondations du domaine de recherche aujourd'hui appelé "Cristallographie Quantique" (voir section 1.2.1). Ces méthodes cherchent à déterminer les propriétés d'un système à partir de sa distribution de charge d'origine expérimentale ou théorique. Les charges positives Z_i des noyaux sont considérées comme ponctuelles et localisées aux positions $\mathbf{R}_i$ des noyaux tandis que les distributions continues des électrons $\rho(\mathbf{r})$ sont modélisées soit à partir des données de diffraction des rayons X soit à partir de méthodes de chimie quantique. La distribution de charge totale $\rho_{\text{tot}}(\mathbf{r})$ est alors définie par : $\rho_{\text{tot}}(\mathbf{r}) = \sum_i Z_i \delta(\mathbf{r} - \mathbf{R}_i) + \rho(\mathbf{r})$. Pour aller plus loin sur les méthodes de densité de charge expérimentale et théorique, le lecteur peut se reporter aux références suivantes : [Tsirelson et Ozerov, 1996, Coppens, 1997, Spackman, 1998, Koritsanszky et Coppens, 2001, Stalke, 2011, Gatti et Macchi, 2012, Chopra, 2012, Dittrich et Matta, 2014, Matta, 2014].

termes de contributions atomiques appelés atomes topologiques. La sommation des propriétés additives des atomes topologiques permet alors d'obtenir les propriétés de la molécule. Une application importante de cette méthodologie est la caractérisation des liaisons covalentes et non-covalentes dans les matériaux et les macromolécules biologiques [Koritsanszky et Coppens, 2001, Gatti et Macchi, 2012].

La topologie d'un champ scalaire est basée sur l'identification de ses points critiques et sur la partition de l'espace réalisée par la distribution de ses lignes de gradient [Matta et Boyd, 2007]. Un point critique de coordonnées $\mathbf{r}_c$ est un point de l'espace où toutes les composantes du gradient du champ scalaire sont nulles. Pour la densité électronique $\rho(\mathbf{r})$ qui est définie en trois dimensions, un point critique est défini par $\nabla\rho(\mathbf{r}_c) = \mathbf{0}$. Il existe différents types de points critiques : maximum local, minimum local et points-selles. Ils sont classifiés à partir des valeurs propres de la matrice hessienne H de la densité électronique au point critique définie par :

$$H(\rho(\mathbf{r}_c)) = \begin{pmatrix} \frac{\partial^2 \rho(\mathbf{r}_c)}{\partial x^2} & \frac{\partial^2 \rho(\mathbf{r}_c)}{\partial x \partial y} & \frac{\partial^2 \rho(\mathbf{r}_c)}{\partial x \partial z} \\ \frac{\partial^2 \rho(\mathbf{r}_c)}{\partial y \partial x} & \frac{\partial^2 \rho(\mathbf{r}_c)}{\partial y^2} & \frac{\partial^2 \rho(\mathbf{r}_c)}{\partial y \partial z} \\ \frac{\partial^2 \rho(\mathbf{r}_c)}{\partial z \partial x} & \frac{\partial^2 \rho(\mathbf{r}_c)}{\partial z \partial y} & \frac{\partial^2 \rho(\mathbf{r}_c)}{\partial z^2} \end{pmatrix} = \begin{pmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_3 \end{pmatrix} \quad (1.4)$$

où λ_1, λ_2 et λ_3 sont les valeurs propres de $H(\rho(\mathbf{r}_c))$ obtenues après diagonalisation, qui sont également appelées courbures de la densité électronique, dans la base des vecteurs propres $(\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3)$. Une courbure λ_i positive implique un minimum de la densité dans la direction du vecteur propre $\mathbf{e}_i$ et une courbure négative, un maximum. Ces courbures sont utilisées pour caractériser les différents types de points critiques grâce aux valeurs (R, S) . Le rang R de la matrice hessienne H correspond au nombre de courbures λ_i non-nulles et sa signature S est la somme algébrique des signes des courbures. Puisque la densité est un champ scalaire $\mathbb{R}^3$, pour un système à l'équilibre : $R = 3$. La signature S peut prendre quatre valeurs différentes, il existe donc quatre types de points critiques de la densité électronique moléculaire :

- Point critique $(3, -3)$: les trois courbures sont négatives, la densité est maximale dans les trois directions de l'espace. Ces maxima locaux indiquent les positions des noyaux, ils sont appelés attracteurs des lignes du champ vectoriel $\nabla\rho(\mathbf{r})$ et peuvent être notés NCP pour "Nucleus Critical Point".
- Point critique $(3, -1)$: deux courbures sont négatives et la troisième est positive, la densité est minimale selon une direction et maximale dans les deux autres. Sur ces points-selles, la direction le long de laquelle se trouve le minimum de densité caractérise une liaison chimique. Le terme BCP pour "Bond Critical Point" est très fréquemment employé pour désigner ces points critiques.
- Point critique $(3, +1)$: deux courbures sont positives et la troisième est négative, la densité est minimale dans deux directions formant un plan et maximale selon l'autre direction. Ces points-selles sont associés à la présence d'un cycle dans la structure moléculaire et sont notés RCP pour "Ring Critical Point".
- Point critique $(3, +3)$: les trois courbures sont positives, la densité est minimale dans les trois directions de l'espace. Ces minima locaux correspondent à des déplétions locales de densité qui sont notamment rencontrées lorsque la molécule ou l'empilement cristallin

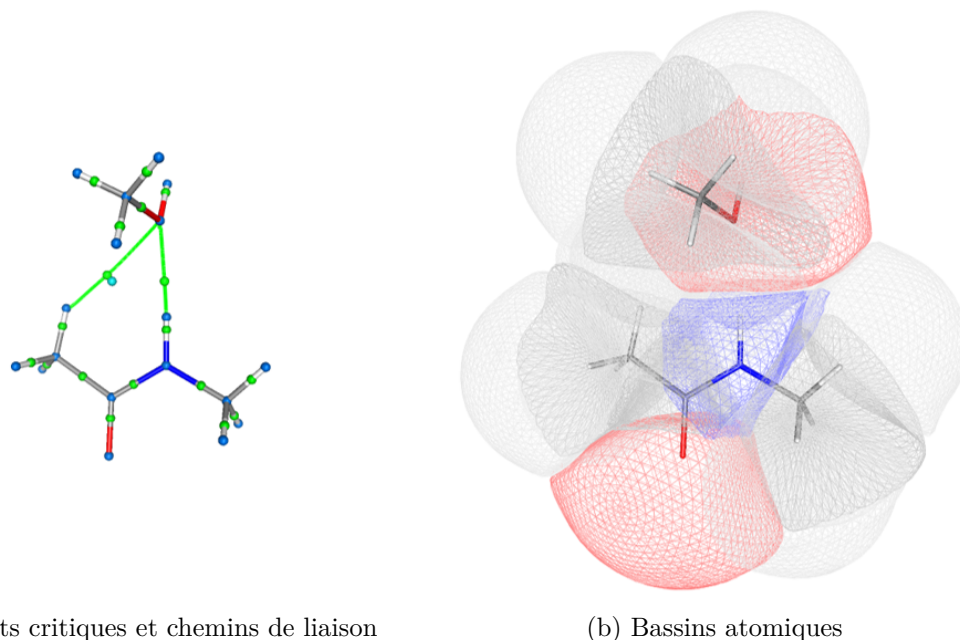


FIGURE 1.3 – Descripteurs issus de la topologie de la densité électronique moléculaire dans le complexe N-méthylacétamide - méthanol.

Les points critiques de densité électronique moléculaire, les chemins de liaison non-covalente et les bassins atomiques sont illustrés dans le complexe N-méthylacétamide - méthanol. (a) Les points critiques de densités sont représentés par des sphères bleues pour les NCP, vertes pour les BCP et cyan pour le RCP. Ce complexe ne présente pas de structure en cage, c'est pourquoi aucun CCP n'apparaît. Les NCP sont tous situés sur des noyaux. Les BCP sont localisés sur les liaisons covalentes mais aussi sur les deux liaisons hydrogène entre de l'atome d'oxygène du méthanol et deux atomes d'hydrogène du N-méthylacétamide. Ces deux liaisons hydrogène sont également caractérisées par deux chemins de liaison non-covalente représentés par des lignes vertes. Par soucis de clarté, les chemins de liaison covalente ne sont pas représentés ici. Un RCP apparaît entre les deux molécules bien qu'il n'y ait pas de cycle moléculaire complet. (b) Les surfaces entourant les bassins atomiques associés à chaque atome des deux molécules sont représentées par des maillages blancs pour les atomes d'hydrogène, gris pour les atomes de carbone, bleus pour les atomes d'azote et rouges pour les atomes d'oxygène. Dans cet exemple, les bassins ont été tronqués à une valeur de densité de 10^{-4} e.Å^{-3} .

forme une structure de cage, par exemple dans les molécules de type fullerène [Wagner *et al.*, 2002]. Ces points critiques peuvent être appelés CCP pour "Cage Critical Point".

Les différents types de points critiques de densité électronique apparaissant dans le complexe N-méthylacétamide - méthanol sont représentés dans la figure 1.3a. Les NCP (ou maxima locaux $(3, -3)$) sont bien positionnés sur chacun des 18 noyaux du complexe. Des BCP (ou points-selles $(3, -1)$) sont localisés sur chaque liaison covalente des deux molécules mais aussi sur deux liaisons non-covalentes, de type liaison hydrogène, entre l'oxygène du méthanol et les hydrogènes du N-méthylacétamide. Un unique RCP (ou points-selles $(3, +1)$) apparaît car, bien que les molécules ne présentent pas de cycle moléculaire complet, la structure du complexe forme un pseudo-cycle autour de la position de ce RCP. Le système ne comportant pas de structure en cage, il ne possède aucun CCP.

Les nombres de points critiques de chaque type dans un système pseudo-isolé obéissent à la relation de Poincaré-Hopf :

$$n_{\text{NCP}} - n_{\text{BCP}} + n_{\text{RCP}} - n_{\text{CCP}} = 1, \quad (1.5)$$

où n_{NCP} est le nombre de points critiques de type NCP, et de même pour les autres termes.

Pour chaque liaison covalente ou non-covalente caractérisée par un BCP, un chemin de liaison, souvent noté BP pour "Bond Path", est défini entre les deux atomes participant à la liaison. Par exemple, dans le complexe N-méthylacétamide - méthanol (voir figure 1.3a), les BP non-covalents (représentés par des lignes vertes) caractérisent les deux liaisons hydrogène entre les deux molécules, déjà identifiées par les BCP de liaisons non-covalentes. Les BP covalents ne sont pas représentés dans cette figure par souci de clarté car ils sont superposés aux cylindres représentant les liaisons covalentes.

En plus des points critiques, l'analyse topologique de la densité électronique repose également sur la topographie de son gradient $\nabla\rho(\mathbf{r}) = \left(\frac{\partial\rho(\mathbf{r})}{\partial x}, \frac{\partial\rho(\mathbf{r})}{\partial y}, \frac{\partial\rho(\mathbf{r})}{\partial z}\right)$. Les lignes de gradient de densité convergent toutes vers les maxima locaux, les attracteurs ou NCP, en se regroupant par faisceaux, chacun associé à un noyau. La surface entourant un faisceau de lignes de gradient de densité est une surface de flux électronique nul S_ρ , telle que $\nabla\rho(\mathbf{r}) \cdot \mathbf{n}(\mathbf{r}) = 0, \forall \mathbf{r} \in S_\rho$ et où $\mathbf{n}(\mathbf{r})$ est le vecteur normal à S_ρ . Dans la théorie QTAIM de Bader, les volumes contenus dans ces surfaces de flux nul sont appelés bassins atomiques. Chaque bassin atomique ne contenant qu'un seul noyau, il permet de définir le volume associé à un unique atome et ainsi réaliser une partition de l'espace. La frontière entre deux bassins atomiques présente un point critique de type BCP qui caractérise la présence d'une liaison covalente ou non-covalente. La densité électronique $\rho(\mathbf{r})$ ne présentant généralement pas de minimum local, ses lignes de gradient émanent de l'infini avant de converger sur les noyaux. Les bassins atomiques, qui sont dits ouverts, sont donc de volume infini pour une molécule pseudo-isolée. Néanmoins, pour pouvoir représenter graphiquement ces bassins, les extrémités des lignes de gradient sont tronquées à une valeur arbitraire faible de densité, généralement 0,001 u.a. qui permet (1) de recouvrir un volume très proche du volume correspondant aux rayons de van der Waals des atomes en phase gazeuse et (2) d'inclure plus de 99% de la densité électronique de la molécule [Bader, 1990]. Une illustration des bassins atomiques dans le complexe N-méthylacétamide - méthanol est proposée dans la figure 1.3b. Ces bassins atomiques définissent la notion d'atome topologique, le volume d'un bassin délimitant l'espace dans lequel les propriétés associées à l'atome topologique peuvent être intégrées.

Ces descripteurs topologiques - points critiques, chemins de liaison et bassins atomiques - sont utilisés pour déterminer les propriétés moléculaires [Matta et Boyd, 2007]. Par exemple, l'estimation de l'ordre d'une liaison ("Bond Order") à partir de la valeur de la densité $\rho(\mathbf{r})$ à la position du BCP permet de distinguer les liaisons covalentes et non-covalentes. Le Laplacien de la densité $\nabla^2\rho(\mathbf{r}) = \frac{\partial^2\rho(\mathbf{r})}{\partial x^2} + \frac{\partial^2\rho(\mathbf{r})}{\partial y^2} + \frac{\partial^2\rho(\mathbf{r})}{\partial z^2}$ permet de localiser les régions de concentration ($\nabla^2\rho(\mathbf{r}) < 0$) et de déplétion ($\nabla^2\rho(\mathbf{r}) > 0$) locales des électrons. Sa valeur sur le BCP est égale à la somme des trois courbures : $\nabla^2\rho(\mathbf{r}_{\text{BCP}}) = \lambda_1 + \lambda_2 + \lambda_3$, avec λ_1 et $\lambda_2 < 0$ et $\lambda_3 > 0$, et permet de caractériser les différents types de liaisons chimiques. Les densités d'énergie potentielle et cinétique locales peuvent également être évaluées à partir du Laplacien aux positions des BCP [Abramov, 1997]. La densité de charge aux points critiques de liaison permet également d'estimer les énergies de dissociation pour caractériser par exemple : les liaisons hydrogène intermoléculaires [Espinosa *et al.*, 2002], les interactions faibles dans l'ADN [Matta *et al.*, 2006], les empilements dominés par les interactions entre électrons π [Lyssenko *et al.*, 2007], les liaisons hydrogène entre zwitterions [Nelyubina *et al.*, 2009], ou encore le rôle des liaisons hydrogène

dans le transfert de charge [Lyssenko *et al.*, 2008]. Par ailleurs, les charges atomiques, les moments dipolaires et quadripolaires, les polarisabilités et certains aspects énergétiques du système peuvent être estimés par l'intégration des opérateurs correspondant dans les bassins atomiques.

Applications aux molécules biologiques

L'analyse topologique de la densité électronique moléculaire est très largement employée pour la compréhension des interactions intra- et intermoléculaires [Koch et Popelier, 1995, Munshi et Guru Row, 2005a, Munshi et Guru Row, 2005b, Dominiak *et al.*, 2006, Matta et Boyd, 2007] et de la réactivité chimique [Bader *et al.*, 1979] des petites molécules organiques ou organométalliques. Par exemple, la revue de B. Dittrich et C. F. Matta [Dittrich et Matta, 2014] discute des analyses de densité de charge expérimentale pour les molécules ayant un intérêt pharmaceutique. Grâce à la densité électronique reconstruite par transfert des pseudo-atomes asphériques depuis les bases de données de paramètres du modèle multipolaire, l'analyse topologique de la densité peut aussi être appliquée à l'étude des interactions impliquant des macromolécules biologiques [Muzet *et al.*, 2003, Lecomte *et al.*, 2005]. Par exemple, au-delà des considérations géométriques traditionnelles, l'analyse topologique de la densité électronique permet de caractériser les liaisons hydrogène participant à la fixation d'un ligand dans le site actif d'une protéine [Liebschner *et al.*, 2009] et entre les groupements carbonyles et amides de la chaîne principale stabilisant les hélices des protéines [Liebschner *et al.*, 2011]. L'identification des contacts d'un ligand dans le site actif d'une protéine ou avec les molécules d'eau environnantes a également été réalisée à partir de l'étude des points critiques de liaison non-covalente et des chemins de liaison [Howard *et al.*, 2016]. Les interactions particulières au sein de clusters fer-soufre des métalloprotéines peuvent également être investiguées [Hirano *et al.*, 2016].

Autres descripteurs issus de la densité électronique

Les points critiques, les chemins de liaison et les bassins atomiques ne sont pas les seuls descripteurs issus de la densité électronique. Parmi ces descripteurs, peuvent notamment être cités : la fonction de localisation des électrons ELF ("Electron Localization Function") [Becke et Edgecombe, 1990], la fonction source ("Source Function") [Bader, 1998, Gatti *et al.*, 2003, Farrugia et Macchi, 2009], les indices de délocalisation des électrons [Bader et Stephens, 1975] et l'indice d'interaction non-covalente NCI ("Non-Covalent Interaction") [Johnson *et al.*, 2010]. L'indice NCI est particulièrement intéressant pour l'étude des interactions intermoléculaires dans les macromolécules biologiques [Laplaza *et al.*, 2021, Mous *et al.*, 2022]. Il permet de visualiser dans l'espace moléculaire tridimensionnel la localisation et la nature des interactions non-covalentes qu'elles soient stabilisantes ou non, comme par exemple les liaisons hydrogène et les encombrements stériques [Contreras-García *et al.*, 2011, Saleh *et al.*, 2012]. Il est basé sur la notion de gradient de densité réduite $s(\mathbf{r})$ défini à partir de la densité électronique moléculaire $\rho(\mathbf{r})$ et de son gradient par :

$$s(\mathbf{r}) = \frac{|\nabla\rho(\mathbf{r})|}{2(3\pi^2)^{1/3}\rho(\mathbf{r})^{4/3}}. \quad (1.6)$$

En pratique, les isosurfaces de $s(\mathbf{r})$ sont représentées et colorées selon la force et la nature des interactions qui sont généralement estimées à partir du produit de la densité électronique

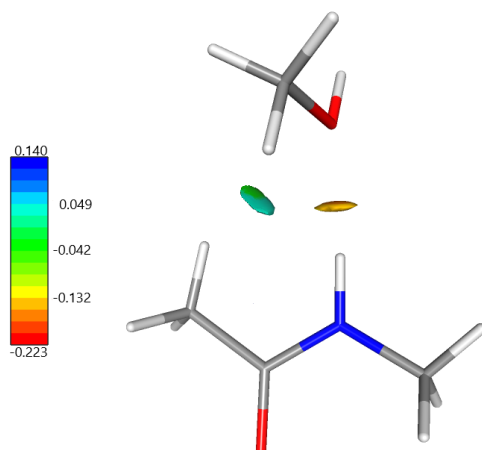


FIGURE 1.4 – Représentation de l'indice d'interaction non-covalente NCI dans le complexe N-méthylacétamide - méthanol.

L'isosurface $s(\mathbf{r}) = 0,45$ du gradient de densité réduite est représentée dans le complexe N-méthylacétamide - méthanol. Elle est colorée en fonction de la valeur du produit de la densité électronique $\rho(\mathbf{r})$ avec le signe de sa courbure λ_2 . Les deux régions qui se distinguent dans cette représentation de l'indice NCI caractérisent la double liaison hydrogène de l'oxygène hydroxyle du méthanol avec l'hydrogène du groupement amide et un des hydrogènes de groupement méthyle du N-méthylacétamide. Elles correspondent également aux deux points critiques de liaisons non-covalentes observés dans la figure 1.3a. La liaison hydrogène impliquant l'atome d'hydrogène amide est caractérisée par une valeur significativement négative de la quantité $\text{sign}(\lambda_2)\rho(\mathbf{r})$, c'est une liaison hydrogène forte, tandis que celle impliquant l'atome d'hydrogène méthyle présente une valeur de $\text{sign}(\lambda_2)\rho(\mathbf{r})$ proche zéro, c'est une liaison hydrogène faible.

$\rho(\mathbf{r})$ et du signe de la deuxième valeur propre λ_2 de sa matrice hessienne car il permet de traduire la contribution de $\rho(\mathbf{r})$ dans les directions interatomiques. Le calcul et la visualisation graphique de l'indice NCI sont implémentés dans le logiciel NCIPLOT [Laplaza *et al.*, 2021]. Le logiciel MoProViewer permet également de représenter l'indice NCI, par exemple dans le complexe N-méthylacétamide - méthanol (voir figure 1.4). Cette représentation met en évidence deux régions importantes dans l'espace intermoléculaire. La première région, caractérisée par une surface de couleur rouge/orangée, se situe entre l'oxygène du méthanol et l'hydrogène du groupement amide du N-méthylacétamide et présente une valeur significativement négative de l'indicateur $\text{sign}(\lambda_2)\rho(\mathbf{r})$: il s'agit d'une liaison hydrogène forte. La seconde région, caractérisée par une surface de couleur vert/bleu, est quant à elle située entre l'oxygène du méthanol et l'un des hydrogènes du groupement méthyle du N-méthylacétamide et sa valeur de $\text{sign}(\lambda_2)\rho(\mathbf{r})$ est très proche de zéro : il s'agit d'une liaison hydrogène faible. Ces deux liaisons hydrogène correspondent à celles identifiées par les BCP et les BP de liaison non-covalente dans la figure 1.3a.

La densité électronique moléculaire $\rho(\mathbf{r})$ permet également de calculer une propriété très importante pour l'étude des interactions intermoléculaires : le potentiel électrostatique moléculaire $V(\mathbf{r})$. Dans la section suivante, les descripteurs qui découlent de $V(\mathbf{r})$ vont être présentés ainsi que leur rôle dans la compréhension des mécanismes enzymatiques et des processus de reconnaissance et de fixation protéine-ligand.

1.3.2 Potentiel électrostatique et champ électrique moléculaires

Définition du potentiel électrostatique moléculaire

Le potentiel électrostatique moléculaire $V(\mathbf{r})$ est dérivé de la densité de charge moléculaire $\rho_{\text{tot}}(\mathbf{r})$ qui peut être soit calculée théoriquement par méthodes de chimie quantique, soit modélisée à partir des expériences de diffraction de rayons X grâce au modèle multipolaire notamment (voir partie 1.2). En effet, en considérant les noyaux comme des charges positives ponctuelles Z_i et les électrons comme une distribution continue $\rho(\mathbf{r})$, $V(\mathbf{r})$ est défini par :

$$V(\mathbf{r}) = \frac{1}{4\pi\epsilon_0} \sum_i \left(\frac{Z_i}{|\mathbf{R}_i - \mathbf{r}|} \right) - \frac{1}{4\pi\epsilon_0} \int \frac{\rho(\mathbf{r}')}{|\mathbf{r}' - \mathbf{r}|} d^3\mathbf{r}', \quad (1.7)$$

où l'indice i porte sur les noyaux dans les molécules du système. En mécanique moléculaire, le potentiel électrostatique est souvent calculé à partir de l'équation de Poisson :

$$\Delta V(\mathbf{r}) = -\frac{\rho_{\text{tot}}(\mathbf{r})}{\epsilon_d}. \quad (1.8)$$

Cette expression permet de tenir compte facilement de la permittivité diélectrique du solvant ϵ_d . Des serveurs en lignes comme APBS ("Adaptative Poisson-Boltzmann Solver") [Jurrus *et al.*, 2018] permettent d'obtenir les cartes de potentiel électrostatique calculées à partir des structures moléculaires en résolvant l'équation 1.8 pour des distributions de charge des ions du solvant de type statistique de Boltzmann.

Evaluation du potentiel électrostatique sur les surfaces moléculaires

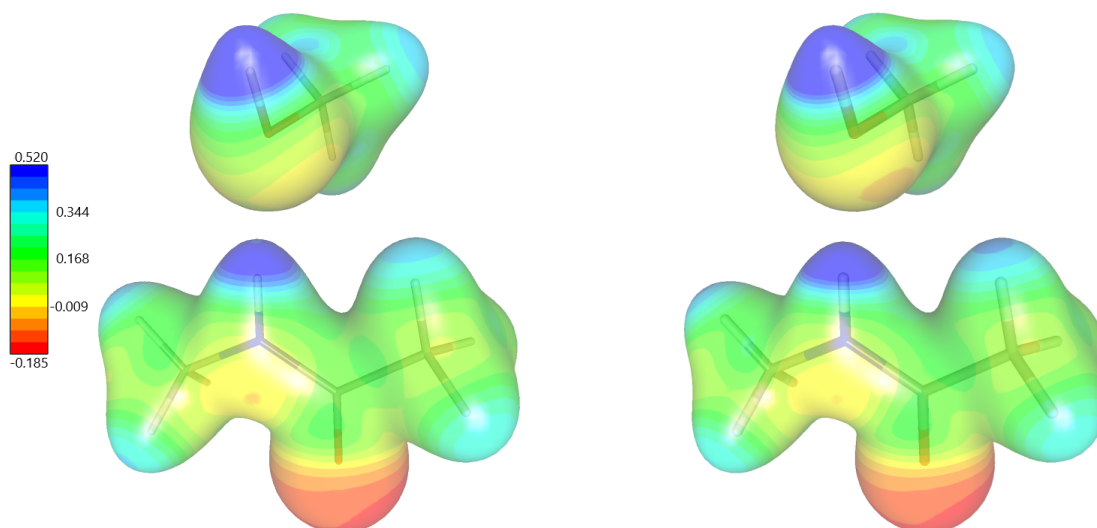
Dès 1970, R. Bonaccorsi, E. Scrocco et J. Tomasi ont montré l'importance du potentiel électrostatique moléculaire pour la compréhension des réactions chimiques en caractérisant les sites moléculaires susceptibles de subir une attaque électrophile dans une série de 6 petites molécules à partir de cartes d'isocontours de $V(\mathbf{r})$ [Bonaccorsi *et al.*, 1970]. Plus tard, C. Petrongolo et J. Tomasi ont suggéré que le potentiel électrostatique pourrait servir de base pour la prédiction et l'interprétation des interactions entre un pharmacophore et sa macromolécule cible en caractérisant les groupements fonctionnels les plus importants [Petrongolo et Tomasi, 1975]. Bien que le pouvoir de l'analyse du potentiel électrostatique $V(\mathbf{r})$ pour la compréhension des interactions intermoléculaires impliquant des macromolécules biologiques ait déjà été accepté à cette époque, il n'existait pas encore d'outil facile à utiliser pour décrire efficacement $V(\mathbf{r})$. C'est en partant de ce constat que P. K. Weiner et ses collègues ont proposé en 1982 de représenter graphiquement $V(\mathbf{r})$ en colorant une surface entourant la molécule selon la valeur du champ scalaire correspondant [Weiner *et al.*, 1982]. Pour cela, ils ont utilisé l'estimation du potentiel électrostatique à partir de distributions de charges ponctuelles sur une surface moléculaire, la surface d'accessibilité au solvant, afin de colorer par exemple en bleu les régions électropositives, telles que les groupements donneurs de protons, et en rouge les régions électronégatives, telles que les accepteurs de protons. Ils ont également montré que ce type de représentation permet de mettre en évidence les complémentarités électrostatique et stérique dans les complexes protéine-ligand et qu'il possède donc un pouvoir prédictif des interactions essentielles à la fixation du

ligand. Cette approche a notamment été appliquée en chimie pharmaceutique pour l'étude de la toxicité des molécules organiques [Politzer, 1988, Murray et Politzer, 2003] et la conception rationnelle de médicaments [Murray et Politzer, 2003, Kumar *et al.*, 2019]. La surface d'accessibilité au solvant est toujours couramment employée comme surface moléculaire pour projeter le potentiel électrostatique mais de plus en plus d'études utilisent préférentiellement les isosurfaces de la densité électronique moléculaire [Howard *et al.*, 2016, Malinska et Dauter, 2016, Dittrich et Luger, 2017], notamment de l'isovaleur $\rho(\mathbf{r}) = 0,002 \text{ u.a}^{10}$ qui a l'avantage d'entourer les atomes des molécules d'une surface proche de celle de van der Waals en phase condensée.

La représentation de $V(\mathbf{r})$ sur la surface moléculaire est aujourd'hui très répandue dans les études de biologie structurale pour décrire les interactions protéine-ligand. Cependant, elles reposent généralement sur un potentiel électrostatique déterminé à partir de distributions de charges ponctuelles issues des champs de force de mécanique moléculaire. La reconstruction de la densité électronique moléculaire fine à partir du transfert des paramètres du modèle multipolaire des bases de données expérimentale ELMAM/ELMAM2 et théorique UBDB/MATTS (voir section 1.2.3) permet quant à elle de déterminer le potentiel électrostatique généré par les macromolécules biologiques avec une précision proche des méthodes quantiques *ab initio* [Muzet *et al.*, 2003, Domagała *et al.*, 2012, Kumar *et al.*, 2019]. Les cartes de potentiel électrostatique ELMAM/ELMAM2 ont permis de caractériser la complémentarité électrostatique et les processus de reconnaissance moléculaire dans des systèmes variés [Muzet *et al.*, 2003, Fournier *et al.*, 2009, Liebschner *et al.*, 2009, Howard *et al.*, 2016]. Par exemple, grâce à la caractérisation de $V(\mathbf{r})$ dans le site de fixation d'une protéine de type phosphate-SBP ("phosphate Solute Binding Proteins"), la stabilisation préférentielle de la forme HPO_4^{2-} du phosphate par cette protéine a pu être établie [Liebschner *et al.*, 2009]. Le potentiel électrostatique UBDB/MATTS a également permis d'analyser des complexes impliquant des macromolécules biologiques - peptides, fragments d'ADN et protéines - grâce à son évaluation sur les surfaces moléculaires [Malinska et Dauter, 2016, Kumar *et al.*, 2019]. Dans la figure 1.5a, les surfaces moléculaires du complexe N-méthylacétamide - méthanol, colorées par l'estimation du potentiel électrostatique issu de la densité électronique transférée, sont représentées. La complémentarité électrostatique entre ces deux molécules dans leurs orientations relatives est particulièrement visible entre la région chargée négativement (en rouge) de l'oxygène du méthanol et la région chargée positivement (en bleu) de l'hydrogène du groupement amide du N-méthylacétamide.

Comme décrit dans la section 1.2.3, la librairie ELMAM2 fournit des polarisabilités anisotropes atomiques qui sont transférables sur les structures des molécules biologiques. Ces polarisabilités sont notamment utilisées pour polariser les densités électroniques en fonction des moments dipolaires mutuellement induits entre deux molécules en complexe [Leduc *et al.*, 2019]. Ces effets d'induction sont visibles dans la représentation du potentiel électrostatique calculé à partir de la densité électronique polarisée, comme le montre la figure 1.5b. En particulier, la comparaison entre la figure 1.5a, à partir de la densité électronique transférée, et la figure 1.5b, à partir de la densité électronique polarisée, met en évidence l'effet des dipôles induits sur les régions des doublets non-liants de l'oxygène du méthanol qui deviennent plus électronégatives

10. L'unité atomique (u. a.) pour la densité électronique est $e.a_0^{-3}$, où $e = 1,60.10^{-19}\text{C}$ est la charge élémentaire et $a_0 = 0,529.10^{-10}\text{m}$ est le rayon de Bohr. L'unité $e.\text{\AA}^{-3}$ est également fréquemment utilisée pour la densité électronique, avec $1 e.\text{\AA}^{-3} \simeq 0,148 e.a_0^{-3}$.



(a) A partir de la densité électronique transférée non-perturbée. (b) A partir de la densité électronique transférée puis polarisée.

FIGURE 1.5 – Représentation du potentiel électrostatique calculé à partir de la densité électronique moléculaire (a) transférée et (b) polarisée sur les surfaces moléculaires du complexe N-méthylacétamide - méthanol.

L'isosurface $\rho(\mathbf{r}) = 0,2 \text{ e.}\text{\AA}^{-3}$ de la densité électronique moléculaire est représentée dans le complexe N-méthylacétamide - méthanol. Elle est colorée selon les valeurs prises par le potentiel électrostatique moléculaire $V(\mathbf{r})$ sur cette surface, indiquées par l'échelle en $\text{e.}\text{\AA}^{-1}$ sur la gauche (dégradé du rouge pour les valeurs négatives au bleu pour les valeurs positives). Le potentiel $V(\mathbf{r})$ a été calculé (a) directement à partir de la densité électronique reconstruite par le transfert des paramètres multipolaires de la base de données ELMAM2 [Domagała *et al.*, 2012] et (b) après avoir appliqué à cette densité reconstruite la procédure de polarisation développée par T. Leduc [Leduc *et al.*, 2019] et décrite dans la section 1.2.3. Le potentiel issu de la densité transférée (a) montre la complémentarité électrostatique entre les deux molécules, et plus particulièrement entre l'atome d'oxygène du groupement hydroxyle du méthanol et l'atome d'hydrogène du groupement amide du N-méthylacétamide, qui est capable de diriger l'approche de ces deux molécules. Le potentiel issu de la densité polarisée (b) caractérise quant à lui les effets de polarisation mutuelle entre les deux molécules en complexe en exposant un oxygène du méthanol plus électronégatif et des hydrogènes méthyle du N-méthylacétamide plus électropositifs que dans le cas non polarisé.

(valeur du $V(\mathbf{r})$ plus basse) et des hydrogènes du N-méthylacétamide faisant face au méthanol qui deviennent plus électropositifs (valeur de $V(\mathbf{r})$ plus haute).

Le champ électrique moléculaire

Le potentiel électrostatique moléculaire $V(\mathbf{r})$ permet de définir une autre grandeur très importante pour la compréhension des interactions intermoléculaires : le champ électrique moléculaire $\mathbf{E}(\mathbf{r}) = -\nabla V(\mathbf{r})$. En effet, le champ électrique $\mathbf{E}(\mathbf{r})$ est directement proportionnel aux forces électrostatiques qui restent significatives à de longues distances dans l'espace intermoléculaire. De plus, s'agissant d'un champ vectoriel, il donne l'information sur la directionnalité de ces forces pour la compréhension du rôle de l'orientation mutuelle des molécules dans les mécanismes de reconnaissance lors de leur approche. Le terme de « pilotage électrostatique » a même été utilisé pour qualifier l'influence des forces électrostatiques sur l'approche d'un ligand dans la poche de fixation d'une protéine [Wade *et al.*, 1998]. D'un point de vue local, $\mathbf{E}(\mathbf{r})$ fournit également des indications sur les interactions stabilisantes dans un complexe. Par exemple,

la description de la topographie et de la densité des lignes de champ électrique entre les bases d'acides nucléiques a permis de caractériser la force des interactions intermoléculaires dans les paires de Watson-Crick en fonction de leur état de protonation [Alkorta *et al.*, 2019]. Pour illustrer l'apparence des lignes de champ électrique dans l'espace intermoléculaire, la figure 1.6 montre dans le complexe N-méthylacétamide - méthanol les lignes de champ électrique émanant des trois atomes d'hydrogène du N-méthylacétamide faisant face à l'oxygène du méthanol. Toutes ces lignes de champ se rejoignent sur les deux doublets non-liants de l'oxygène mais empruntent des chemins différents, étendant ainsi dans l'espace l'influence électrostatique à la fois des hydrogènes dont elles émergent mais aussi de l'atome d'oxygène qu'elles atteignent.

Le champ électrique moléculaire $\mathbf{E}(\mathbf{r})$ semble donc être une grandeur pertinente pour l'analyse de la contribution électrostatique aux mécanismes de reconnaissance moléculaire dans les complexes protéine-ligand. Dans l'étude de la poche de fixation de la protéine FABP ("Fatty Acid Binding Protein") contenant un acide oléique et un cluster de molécules d'eau [Howard *et al.*, 2016], le champ électrique dérivé du potentiel électrostatique moléculaire du complexe a été caractérisé. L'intensité du champ électrique calculée au centre de la cavité est de l'ordre de 10^9 V/m. Cet ordre de grandeur correspond aux valeurs obtenues habituellement par mesure directe de champ électrique dans une cavité de protéine qui peuvent être réalisées par méthodes de spectroscopie vibrationnelle exploitant l'effet Stark [Lehle *et al.*, 2005, Suydam *et al.*, 2006, Wang *et al.*, 2013]. Une représentation de la distribution des lignes de champ électrique dans ce complexe FABP - acide oléique - molécules d'eau a également été réalisée par B. Guillot et A. Podjarny¹¹. Cette représentation, bien qu'elle ait été fastidieuse à mettre en œuvre malgré les outils à disposition dans le logiciel MoProViewer [Guillot *et al.*, 2014], a fourni des résultats préliminaires prometteurs, justifiant la pertinence de la caractérisation des interactions électrostatiques dans les protéines par l'étude des lignes de champ électrique. Mon sujet de thèse est né du besoin de développements méthodologiques de descripteurs et d'outils permettant de mettre en place ces études. Plus précisément, mon objectif est de développer des descripteurs basés sur la topologie du potentiel électrostatique, et donc sur la topographie des lignes de champ électrique.

Analyse topologique du potentiel électrostatique

Le potentiel électrostatique $V(\mathbf{r})$ étant un champ scalaire moléculaire, son analyse topologique peut être réalisée de manière similaire à celle la densité électronique $\rho(\mathbf{r})$ décrite dans la section précédente. En effet, les points critiques $\mathbf{r}_c$ de $V(\mathbf{r})$, définis par $\nabla V(\mathbf{r}_c) = \mathbf{0}$, sont également distingués par la paire (R, S) en quatre types : $(3, -3)$, $(3, -1)$, $(3, +1)$ et $(3, +3)$. Néanmoins, l'interprétation des points critiques de $V(\mathbf{r})$ diffère de celle des points critiques de $\rho(\mathbf{r})$. Les maxima locaux $(3, -3)$ de $V(\mathbf{r})$ restent localisés sur les positions des noyaux, comme dans le cas de $\rho(\mathbf{r})$. Par contre, les minima locaux $(3, +3)$, qui correspondent à une concentration locale d'électrons, révèlent les localisations des paires d'électrons non-liantes [Kumar *et al.*, 2014a, Gadre et Kumar, 2016], qui n'apparaissent pas dans les points critiques de la densité. Pour les points-selles $(3, -1)$ et $(3, +1)$, certains sont équivalents aux points critiques de liaisons

11. Cet aspect n'a pas été publié dans l'article [Howard *et al.*, 2016] avec le reste de l'étude de la FABP mais a été introduite par ailleurs dans la revue de C. Matta [Matta, 2014] (voir notamment la figure 18 de cette revue).

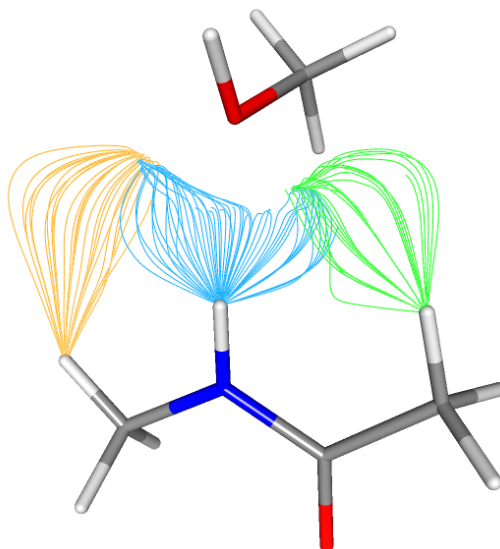


FIGURE 1.6 – Lignes de champ électrique dans l'espace intermoléculaire du complexe N-méthylacétamide - méthanol.

Les lignes de champ électrique dérivées du potentiel électrostatique moléculaire dans le complexe N-méthylacétamide - méthanol sont représentées par les lignes continues, de couleur bleue pour celles émergeant de l'atome d'hydrogène du groupement amide du N-méthylacétamide, et de couleurs orange et verte pour les deux hydrogène des groupements méthyle se trouvant de part et d'autre du N-méthylacétamide. Par soucis de clarté, seules les lignes de champ émanant de ces trois atomes et s'évanouissant sur les paires d'électrons non-liantes de l'atome d'oxygène du méthanol apparaissent ici. Cette représentation met en évidence de façon visuelle la répartition spatiale des forces électrostatiques stabilisant les liaisons hydrogène dans le complexe.

(BCP) et de cycles (RCP) de $\rho(\mathbf{r})$ mais les autres sont dû à la topologie plus complexe de $V(\mathbf{r})$ et ne possèdent pas une interprétation en termes de liaison ou d'interaction moléculaire [Gadre et Kumar, 2016, Gadre et Bendale, 1986, Leboeuf *et al.*, 1999, Balanarayan et Gadre, 2003, Mata *et al.*, 2007, Anjalikrishna *et al.*, 2019].

Contrairement au gradient de la densité, le gradient du potentiel électrostatique correspond à une grandeur physique mesurable : le champ électrique moléculaire $\mathbf{E}(\mathbf{r}) = -\nabla V(\mathbf{r})$. Les lignes de champ électrique se regroupent par faisceaux démarrant du même maximum local (correspondant à un site électrophile) et se terminant sur un même minimum local (associé à un site nucléophile). Ces faisceaux, appelés faisceaux primaires, sont entourés par des surfaces de flux nul du potentiel S_V telles que $\mathbf{E}(\mathbf{r}) \cdot \mathbf{n}(\mathbf{r}) = 0, \forall \mathbf{r} \in S_V$. Les bassins topologiques définis par ces surfaces de flux nul ne sont pas associés à un unique atome mais à une paire site électrophile - site nucléophile, à la différence des bassins atomiques de $\rho(\mathbf{r})$. Notons qu'un sujet de recherche actuel, mené par E. Espinosa et ses collègues, est la caractérisation de l'intersection entre les bassins topologiques de la densité électronique et ceux du potentiel électrostatique, permettant de définir la région d'attraction électrostatique EAR ("Electrostatic Attraction Region") [Espinosa, 2023]. Cette approche a permis d'expliquer les interactions paradoxalement attractives dans les agrégats anion-anion [Mata *et al.*, 2015] et cation-cation [Alkorta *et al.*, 2019].

Les faisceaux primaires peuvent être regroupés selon les sites électrophiles ou selon les sites nucléophiles auxquels ils sont associés. Deux partitions de l'espace moléculaire supplémentaires sont alors définies : la partition en zones d'influence électrophile et la partition en zones d'in-

fluence nucléophile. Grâce à ce découpage du volume entourant la molécule, la répartition spatiale des influences électrostatiques de chaque site est révélée. Par exemple, considérant une charge-test q à une position $\mathbf{r}_q$ dans l'espace moléculaire, si $q > 0$ alors elle sera attirée par les forces électrostatiques vers le site nucléophile associé à la zone d'influence nucléophile contenant sa position $\mathbf{r}_q$ et sera repoussée par le site électrophile associé à la zone d'influence électrophile contenant $\mathbf{r}_q$. Inversement, si $q < 0$, alors la charge-test sera attirée vers le site électrophile associé à la zone d'influence électrophile à laquelle sa position $\mathbf{r}_q$ appartient et sera repoussée par le site nucléophile associé à la zone d'influence nucléophile contenant $\mathbf{r}_q$. Aussi, ces descripteurs - les zones d'influence électrophile et nucléophile - ont un pouvoir prédictif de la direction de tractation de la charge-test lors de son approche et permettent d'identifier les contributeurs impliqués. Ils ont été décrits par les groupes de recherche de E. Espinosa et de S. R. Gadre pour l'étude des interactions intermoléculaires dans les petites molécules [Mata *et al.*, 2007, Gadre et Kumar, 2016, Kumar et Gadre, 2016, Alkorta *et al.*, 2019].

Les travaux de ma thèse ont pour objectif de développer ce type de descripteurs issus de l'analyse topologique du potentiel électrostatique $V(\mathbf{r})$ et de les appliquer aux macromolécules biologiques, notamment pour la compréhension des mécanismes de reconnaissance moléculaire protéine-ligand et de réactivité enzymatique. L'objet du chapitre 2 de ce manuscrit sera de détailler davantage et d'illustrer les notions et concepts associés à la topologie de $V(\mathbf{r})$ qui viennent d'être évoqués ainsi que de décrire nos développements méthodologiques. Dans le chapitre 4, les parties 4.1 et 4.2 proposeront quant à eux la preuve de concept de nos descripteurs pour l'étude d'enzymes, avec l'exemple de la protéase à sérine, et de protéines membranaires, telles que la neuropiline.

En résumé,

grâce au transfert des paramètres du modèle multipolaire depuis la librairie ELMAM2 par exemple, la densité électronique moléculaire $\rho(\mathbf{r})$ d'un système dont la structure atomique est connue peut être reconstruite de façon très précise. L'analyse de la densité de charge qui en découle permet de caractériser les propriétés électrostatiques des molécules. Les principaux descripteurs topologiques de la densité électronique définis dans le cadre de la théorie quantique des atomes dans les molécules (QTAIM) de Bader sont : les points critiques (notamment les BCP), les chemins de liaison et les bassins atomiques. Les propriétés aux points critiques de liaison et chemins de liaison sont utilisées pour identifier et déterminer la nature des liaisons covalentes et non-covalentes. Les bassins atomiques définissent les volumes topologiques associés à chaque atome, ce qui permet de déterminer les propriétés moléculaires à partir de celles des atomes topologiques.

La densité électronique transférée, associée à la distribution ponctuelle des charges positives des noyaux, permet de calculer le potentiel électrostatique moléculaire $V(\mathbf{r})$. Classiquement, le potentiel électrostatique est représenté sur des surfaces entourant les molécules afin de caractériser la complémentarité électrostatique entre molécules en interaction. Cette représentation est particulièrement utile pour comprendre les mécanismes de reconnaissance moléculaire. Le champ électrique moléculaire $\mathbf{E}(\mathbf{r})$ dérivé du potentiel électrostatique est mesurable expérimentalement dans les cavités des protéines. La représentation des lignes de champ donne accès à

la visualisation graphique de la directionnalité des forces électrostatiques intermoléculaires qui peuvent diriger l'approche d'un ligand et stabiliser un complexe. Le potentiel électrostatique peut également être étudié d'un point de vue topologique de façon similaire à la densité électronique par la définition des points critiques et des surfaces de flux nul. Sa topologie permet de réaliser deux partitions de l'espace différentes associées aux sites électrophiles et nucléophiles des molécules : la partition en zones d'influence électrophile et la partition en zones d'influence nucléophile. Ces partitions possèdent un fort potentiel pour la compréhension des mécanismes d'approche d'un ligand vers le site actif d'une protéine, en lien avec les notions d'affinité et de spécificité des protéines pour leur substrat. C'est pourquoi nous avons développé des méthodes et des outils permettant de mettre en œuvre ces descripteurs dans les études des macromolécules biologiques, ce que je présenterai plus en détail dans le chapitre 2.

Le second type de descripteur que nous avons développé repose quant à lui sur la définition d'un modèle d'énergie d'interaction intermoléculaire totale à partir de la densité électronique transférée. La modélisation de l'énergie d'interaction est un sujet de recherche investigué depuis des décennies et de très nombreuses approches ont vu le jour. La partie suivante de ce manuscrit présente les principales méthodes de calcul de cette énergie d'interaction sans avoir la prétention d'être exhaustive.

1.4 Modélisation de l'énergie d'interaction intermoléculaire totale

A partir de leurs structures atomiques et de leurs modèles de distributions de charge associés, l'énergie d'interaction entre deux molécules peut être évaluée. Cette énergie permet de prédire les orientations relatives des deux molécules et la stabilité de leur complexe. Cette méthode est notamment utilisée pour la conception rationalisée de médicaments [Merz Jr *et al.*, 2010]. En effet, la caractérisation biochimique du mécanisme pathologique permet d'identifier une cible thérapeutique (souvent une protéine) dont la fonction peut être modifiée par la fixation d'un ligand spécifique. La prédiction *in silico* des pharmacophores ayant le meilleur potentiel pour interagir avec la cible et se fixer dans son site actif sert à réduire les coûts expérimentaux et accélérer la recherche en limitant significativement le nombre de molécules à tester. Pour cela, des fonctions de scoring permettant de hiérarchiser les candidats, développées dans les programmes de docking moléculaire tels que GOLD [Jones *et al.*, 1997] ou Glide [Friesner *et al.*, 2004], reposent sur la quantification des différentes contributions à l'interaction protéine-ligand. De nombreux effets doivent être pris en compte par ces fonctions de scoring : vibrations et rotations intramoléculaires, interactions intermoléculaires, solvation, contacts hydrophobes et hydrophiles, entropie, variabilité conformationnelle, etc. L'énergie d'interaction intermoléculaire totale E_{inter} peut être déterminée à partir de la structure des molécules en interaction. Elle est généralement découpée en cinq contributions [Stone, 2013] :

$$E_{\text{inter}} = E_{\text{elst}} + E_{\text{pol}} + E_{\text{CT}} + E_{\text{disp}} + E_{\text{rep}}, \quad (1.9)$$

où E_{elst} est l'énergie d'interaction électrostatique, E_{pol} est la contribution de la polarisation, E_{CT} est l'énergie due aux transferts de charge, E_{disp} est l'énergie de dispersion et E_{rep} , l'énergie

de répulsion. Pour calculer ces différentes contributions, de nombreuses méthodes ont vu le jour, à partir des distributions de charge expérimentales, sur la base de modèles empiriques ou encore par application de méthodes de chimie quantique.

1.4.1 A partir des densités de charge expérimentales et transférées

Parmi les différentes contributions énergétiques de l'équation 1.9, les interactions de nature électrostatique font partie des plus importantes. L'énergie d'interaction électrostatique E_{elst} est due à l'interaction entre deux distributions de charge moléculaires non-perturbées $\rho_{\text{tot}}^A(\mathbf{r})$ et $\rho_{\text{tot}}^B(\mathbf{r})$, exprimée par la loi de Coulomb sous sa forme intégrale :

$$E_{\text{elst}}^{AB} = \frac{1}{4\pi\epsilon_0} \int_A \int_B \frac{\rho_{\text{tot}}^A(\mathbf{r}_1)\rho_{\text{tot}}^B(\mathbf{r}_2)}{|\mathbf{r}_1 - \mathbf{r}_2|} d^3\mathbf{r}_1 d^3\mathbf{r}_2. \quad (1.10)$$

Grâce aux données de diffraction des rayons X à résolution subatomique, la distribution de charge expérimentale peut être modélisée de façon très précise, à l'aide du modèle multipolaire notamment (voir section 1.2.2). Néanmoins, dans le cristal, les densités de charge sont perturbées, elles sont notamment déformées par la polarisation mutuelle et par la répulsion de Pauli. Aussi, l'énergie E_{elst} calculée par application directe de la loi de Coulomb sur ces densités de charge serait donc une combinaison de plusieurs contributions, voire même une approximation de l'énergie d'interaction totale [Ma et Politzer, 2004], sans que cette relation ne soit clairement définie, comme discuté par P. Dominiak, E. Espinosa et J. Ángyán dans le chapitre 11 du livre *Modern Charge-Density Analysis* publié par C. Gatti and P. Macchi [Gatti et Macchi, 2012].

La distribution de charge moléculaire peut également être reconstruite à partir du transfert des paramètres du modèle multipolaire de la densité électronique atomique des bases de données expérimentale ELMAM/ELMAM2 et théorique UBDB/MATTS présentées dans la section 1.2.3. Grâce au principe de moyenne de ces paramètres dans diverses structures moléculaires, les densités électroniques transférées sont considérées comme non-perturbées par les interactions non-covalentes. Ces densités permettent donc de calculer la contribution électrostatique E_{elst} à l'énergie d'interaction totale. Cependant, l'application directe de la loi de Coulomb n'est pas suffisante car l'énergie qui en découle ne tient pas compte des effets de pénétration des densités électroniques stabilisant l'interaction grâce à un écrantage incomplet des charges positives des noyaux par les densités électroniques. Plusieurs méthodes permettant de modéliser ces effets de pénétration ont été développées [Gavezzotti, 2002, Volkov *et al.*, 2004a, Bojarowski *et al.*, 2016]. Par ailleurs, pour simplifier les calculs, la densité électronique moléculaire est généralement partitionnée en contributions atomiques, certaines méthodes utilisent la partition en atomes de Hirshfeld [Hirshfeld, 1977] et d'autres celles en atomes topologiques de Bader [Bader, 1990] définis dans le cadre de la QTAIM discutée dans la section 1.3.1. Une approche numérique très répandue pour calculer E_{elst} à partir des distributions multipolaires est le modèle de potentiel exact et modèle multipolaire EP/MM ("Exact Potential and Multipole Model") [Volkov *et al.*, 2004a, Volkov *et al.*, 2006, Spackman, 2007, Volkov et Coppens, 2007] initialement développée par A. Volkov, T. Koritsanszky et P. Coppens. Cette méthode nécessite la définition d'une distance critique (ou cut-off), généralement choisie entre 4 et 5 Å, séparant le traitement exact EP et le traitement approximatif MM, ce qui permet un gain de temps de calcul considérable par rapport

à une évaluation exacte sur tout le système. Dans la partie EP, les interactions électrostatiques à courte portée entre paires de pseudo-atomes sont quantifiées par l'évaluation exacte de l'énergie électrostatique grâce à une quadrature numérique de l'intégrale tridimensionnelle de la loi de Coulomb (équation 1.10) réécrite en termes de potentiels électrostatiques $V(\mathbf{r})$:

$$E_{elst}^{AB} = \int_A \rho_{tot}^A(\mathbf{r}) V^B(\mathbf{r}) d^3\mathbf{r}, \quad (1.11)$$

où $V^B(\mathbf{r})$ est le potentiel électrostatique généré par la distribution de charge totale $\rho_{tot}^B(\mathbf{r})$ de la molécule B . Cette écriture a l'avantage de ne présenter qu'une seule intégrale 3D contrairement à l'équation 1.20 qui en présente deux. Les potentiels électrostatiques sont estimés à partir des pseudo-atomes par sommation dans l'espace direct. Pour la partie MM, les interactions à plus longue portée, qui sont aussi les plus nombreuses, sont évaluées par approximation multipolaire de Buckingham. Cette approximation est basée sur le développement de Taylor du terme $|\mathbf{r}_1 - \mathbf{r}_2|^{-1}$ de l'équation 1.10 [Volkov *et al.*, 2004b] :

$$\begin{aligned} E_{elst}^{AB} \simeq & \frac{1}{4\pi\epsilon_0} \left[T q^A q^B + T_\alpha (q^A \boldsymbol{\mu}_\alpha^B - q^B \boldsymbol{\mu}_\alpha^A) \right. \\ & + T_{\alpha\beta} \left(\frac{1}{3} q^A \boldsymbol{\Theta}_{\alpha\beta}^B + \frac{1}{3} q^B \boldsymbol{\Theta}_{\alpha\beta}^A - \boldsymbol{\mu}_\alpha^A \boldsymbol{\mu}_\alpha^B \right) \\ & + T_{\alpha\beta\gamma} \left(\frac{1}{15} q^A \boldsymbol{\Omega}_{\alpha\beta\gamma}^B - \frac{1}{15} q^B \boldsymbol{\Omega}_{\alpha\beta\gamma}^A - \frac{1}{3} \boldsymbol{\mu}_\alpha^A \boldsymbol{\Theta}_{\beta\gamma}^B + \frac{1}{3} \boldsymbol{\mu}_\alpha^B \boldsymbol{\Theta}_{\beta\gamma}^A \right) \Big] \\ & + \dots \end{aligned} \quad (1.12)$$

où q , $\boldsymbol{\mu}$, $\boldsymbol{\Omega}$ et $\boldsymbol{\Theta}$ sont respectivement les moments monopolaires (charges), dipolaires, quadrupolaires et octopolaires permanents des distributions de charge non-perturbées. Les tenseurs T sont les tenseurs d'interaction symétriques : $T_{\alpha\beta\gamma\dots} = \nabla_\alpha \nabla_\beta \nabla_\gamma \dots |\mathbf{r}_1 - \mathbf{r}_2|^{-1}$, où les indices α , β , γ portent sur les coordonnées x , y , z . Cette méthode EP/MM a été implémentée dans la suite logiciel MoPro [Fournier, 2010]. Elle a été appliquée au calcul de l'énergie d'interaction électrostatique E_{elst} entre ligand et protéine à partir des densités transférées de la base de données ELMAM/ELMAM2 [Fournier *et al.*, 2009, Howard *et al.*, 2016] et à partir de la base de données UBDB/MATTS [Dominiak *et al.*, 2007, Dominiak *et al.*, 2009, Kumar *et al.*, 2014b, Malinska et Dauter, 2016, Kumar et Dominiak, 2021]. Cependant, cette approche numérique est coûteuse en temps de calcul car elle utilise une intégration en trois dimensions pour obtenir la partie courte distance d'interaction. Une méthode similaire mais de nature analytique a été proposée par D. Nguyen, Z. Kisiel et A. Volkov [Nguyen *et al.*, 2018], elle est nommée méthode analytique du potentiel exact et modèle multipolaire (aEP/MM) par opposition à la méthode numérique (nEP/MM). La méthode aEP/MM est plus rapide jusqu'à deux ordres de grandeur en temps de calcul que la méthode nEP/MM tout en restant très précise [Vuković *et al.*, 2021]. Elle a été implémentée par V. Vuković dans la librairie *Charger* [Vuković *et al.*, 2021] qui est disponible dans le logiciel MoProViewer. Une application de cette librairie sur un complexe protéine-ligand a également été réalisée [Vuković *et al.*, 2021] pour déterminer les contributions individuelles de chaque résidu du site actif de la glutathion transférase à son énergie d'interaction électrostatique avec un ligand benzophénone. Cette étude a permis de mettre en évidence les résidus stabilisant l'interaction mais aussi ceux qui présentent une contribution électrostatique défavorable

à la fixation du ligand. Aussi, cette approche permet de suggérer des mutations permettant d'éliminer les interactions déstabilisantes.

Les effets d'induction dipolaire participant à l'énergie de polarisation E_{pol} sont quant à eux accessibles à partir des polarisabilités fournies par la librairie ELMAM2. Comme expliqué dans la section 1.2.3, la librairie ELMAM2 permet de transférer en plus des paramètres de densité électronique des pseudo-atomes, des tenseurs de polarisabilités atomiques anisotropes [Leduc *et al.*, 2019]. Grâce à ces tenseurs de polarisabilité, les moments dipolaires induits par la polarisation mutuelle de deux molécules en interaction sont calculés et intégrés aux populations dipolaires du modèle multipolaire de la densité électronique de chacune des deux molécules. La densité de charge moléculaire ainsi obtenue n'est perturbée que par ces effets d'induction dipolaire. L'énergie d'interaction électrostatique calculée à partir de cette densité de charge polarisée contient donc une contribution supplémentaire de polarisation par rapport à celle calculée sur la base de la densité de charge simplement transférée. Aussi, une estimation de l'énergie de polarisation E_{pol} qui en découle est définie comme la différence entre l'énergie d'interaction électrostatique basée sur la densité polarisée $E_{\text{elst}}^{\text{POL}}$ et l'énergie d'interaction électrostatique directement dérivée de la densité transférée non-perturbée $E_{\text{elst}}^{\text{TRF}}$ [Leduc *et al.*, 2019] :

$$E_{\text{pol}} = E_{\text{elst}}^{\text{POL}} - E_{\text{elst}}^{\text{TRF}} \quad (1.13)$$

L'énergie de polarisation E_{pol} ainsi obtenue ne rend compte que des effets d'induction dipolaire. Les contributions quadripolaires et d'ordres supérieurs nécessiteraient une modélisation plus avancée.

L'évaluation des termes de polarisation E_{pol} et de transfert de charge E_{CT} sont difficiles à extraire de la densité de charge expérimentale, ils sont négligés dans la plupart des modèles empiriques mais peuvent être estimés par méthodes *ab initio* [Abramov *et al.*, 2000b]. De même, il n'existe pas de méthode pour extraire les contributions de dispersion et de répulsion directement de la densité de charge expérimentale. Des méthodes semi-empiriques ont été développées, comme par exemple les méthodes de sommes de densités de Gavezzotti [Gavezzotti, 2002, Gavezzotti, 2003, Gavezzotti, 2005] pour la prédiction des empilements cristallins ou encore de potentiels atome-atome isotropes [Spackman, 1986a, Coppens *et al.*, 1999, Abramov *et al.*, 2000a, Grabowsky *et al.*, 2008] utilisant les paramétrisations de D. Williams et S. Cox [Cox *et al.*, 1981, Williams et Cox, 1984] ou du modèle promoléculaire de M. Spackman [Spackman, 1986b, Spackman et Maslen, 1986, Spackman, 1986a, Spackman, 1987]. Néanmoins, ces potentiels reposent sur des paramètres qui ne sont disponibles que pour un nombre limité d'atomes et dont l'interprétation physique n'est pas clairement définie, et requièrent un traitement spécial des atomes d'hydrogène impliqués dans une liaison hydrogène. A. Volkov et ses collègues ont également proposé de compléter l'énergie d'interaction électrostatique calculée par méthode EP/MM à partir des densités transférées UBDB en déterminant les termes manquants à l'aide d'un modèle paramétré pour reproduire les énergies obtenues par la théorie quantique SAPT [Li *et al.*, 2006] qui sera évoquée plus tard dans cette partie. Par conséquent, la détermination de l'énergie d'interaction intermoléculaire totale ne peut pas être réalisée uniquement à partir de la distribution de densité de charge et nécessite l'évaluation de certaines contributions par des modèles empiriques (ou semi-empiriques) et/ou par des méthodes de chimie quantique.

Ainsi, le second objectif de ce projet doctoral est de développer un potentiel¹² d'interaction intermoléculaire total basé sur les paramètres du modèle multipolaire et les polarisabilités atomiques fournis par la librairie ELMAM2. Avant de détailler ces développements dans le chapitre 3, une vue d'ensemble de principaux modèles empiriques et quantiques existant est proposée dans la suite de cette partie.

1.4.2 Méthodes empiriques

Bien avant de pouvoir prouver la nature atomistique de la matière, les scientifiques des XVIII^e et XIX^e siècles tentaient déjà de comprendre les interactions entre les molécules, en développant la théorie de la cinétique des gaz notamment. En effet, pouvoir décrire la nature exacte des atomes n'est pas nécessaire pour se convaincre de l'existence de forces au sein de la matière. En introduction de son livre [Stone, 2013], A. Stone donne cet exemple : des forces attractives doivent exister entre les molécules en phase condensée sinon rien ne retiendrait les molécules d'eau confinées à l'intérieur d'un verre. De même, il doit également exister des forces répulsives à courte distance entre les molécules car l'eau possède une densité définie et il n'est pas aisé de la compresser dans un volume inférieur. L'énergie d'interaction entre deux molécules est donc une fonction de la distance R qui les sépare, avec une région attractive à longue distance et une région répulsive à très courte distance. J. D. van der Waals fut l'un des premiers à formuler la prise en compte de ces effets en suggérant que pour la mesure du volume V d'un gaz, il faut considérer qu'un mole de molécules occupe un volume incompressible b , tandis que le volume restant $V - nb$ permet aux molécules de se mouvoir librement, où n est le nombre de mole. Les forces attractives entre les molécules ont pour effet de réduire la pression exercée par un gaz sur son contenant, la pression mesurée P est donc différente de la pression définie dans le cadre de la loi des gaz parfaits. L'expression proposée pour cette pression est : $P + an^2/V^2$, la loi des gaz parfaits $PV = nRT$ devenant alors : $(P + an^2/V^2)(V - nb) = nRT$. Malgré sa simplicité, ce modèle permet de retrouver avec une précision acceptable certaines propriétés des gaz réels. Les travaux de J. D. van der Waals furent la première pierre apportée à l'édifice de la recherche de la modélisation des interactions intermoléculaires, c'est pourquoi encore aujourd'hui certaines forces d'attraction et de répulsion sont appelées « forces de van der Waals ».

Potentiels interatomiques en dynamique moléculaire

Le développement des potentiels d'interaction est particulièrement utile pour la définition des champs de force en dynamique moléculaire classique. Dans les méthodes de simulation numérique par dynamique moléculaire classique, l'évolution temporelle du système considéré est prédite par intégration des équations classiques du mouvement sur la base des forces conservatives $\mathbf{F} = -\nabla U$ calculées à partir des potentiels interatomiques U [Allen et Tildesley, 1987]. Parmi ces potentiels interatomiques, certains sont utilisés pour décrire les effets intramoléculaires

12. Le terme « potentiel », souvent utilisé pour désigner « énergie potentielle » par abus de langage, sera employé dans ce manuscrit pour qualifier un modèle d'énergie potentielle, comme pratiqué en mécanique moléculaire. Il sera noté U contrairement à l'énergie réelle qui est notée E .

(ou covalents), tels que les vibrations d'élongation des liaisons covalentes, les rotations autour d'une liaison et les torsions des angles dièdres. Les autres potentiels interatomiques permettent de prendre en compte les interactions intermoléculaires (ou non-covalentes). Il existe des potentiels d'interaction non-covalente spécifiques à chaque type d'interaction particulière mais deux d'entre eux se retrouvent dans tous les systèmes : les interactions électrostatiques coulombiennes et les forces de van der Waals attractives et répulsives. Ces potentiels interatomiques (covalents et non-covalents) sont empiriques, ils ont été construits à partir d'approximations reposant sur la connaissance *a priori* du système et dépendent de paramètres optimisés pour reproduire des résultats expérimentaux ou de chimie quantique, comme les enthalpies de vaporisation ou des paramètres spectroscopiques. Le terme « champ de force » désigne l'ensemble des potentiels interatomiques et de leurs paramètres pour un type de système donné. De très nombreux champs de force ont été développés pour la simulation numérique et certains sont spécifiques aux macromolécules biologiques tels que : AMBER ("Assisted Model Building and Energy Refinement") [Case *et al.*, 2005], CHARMM ("Chemistry at HARvard Molecular Mechanics") [Brooks *et al.*, 2009], GROMACS ("GRONingen MACHine for Chemical Simulations") [Berendsen *et al.*, 1995] et OPLS ("Optimized Potentials for Liquid Simulations") [Jorgensen *et al.*, 1996]. Dans ces champs de force, les potentiels d'interaction non-covalente U_{inter} modélisent les forces de van der Waals attractives de dispersion U_{disp} et répulsives U_{rep} et les effets électrostatiques U_{elst} . Les effets de polarisation U_{pol} sont rarement pris en compte, comme dans le champ de force AMOEBA par exemple, tandis que les phénomènes de transfert de charge U_{CT} , qui sont de nature quantique, sont négligés.

Forces de van der Waals

Le terme "force de van der Waals" peut porter à confusion car il est utilisé pour qualifier à la fois les forces issues des interactions entre multipôles, dont les forces de Keesom, Debye et London, et les forces attractives et répulsives qui sont traitées de façon simultanée par les potentiels de type Lennard-Jones. Les forces de van der Waals de type dipolaire sont des forces émergeant des corrélations entre les dipôles permanents, induits ou instantanés des molécules en interaction. Trois types de forces sont alors distingués :

- Les forces de Keesom dipôle permanent - dipôle permanent : elles n'apparaissent que lorsque le système contient des molécules polaires, portant des moments dipolaires permanents μ . Cette contribution est directement prise en compte dans le calcul de l'énergie électrostatique à partir de distributions électroniques multipolaires. Elle peut également être estimée entre deux molécules A et B grâce au potentiel de Keesom :

$$U_{\text{Keesom}}^{AB} = -\frac{2\mu_A^2\mu_B^2}{3k_B T (4\pi\epsilon_0)^2 R_{AB}^6}, \quad (1.14)$$

où μ_A et μ_B sont les normes des moments dipolaires des molécules A et B , R_{AB} est la distance entre les centres de masse de ces deux molécules, k_B est la constante de Boltzmann, T est la température du système et ϵ_0 est la constante diélectrique du vide. Il est intéressant de noter que le potentiel de Keesom est issu de la moyenne statistique de l'interaction entre deux dipôles isolés : $U_{\text{dipôle 1 - dipôle 2}} = -\frac{\mu_1 \cdot \mu_2}{4\pi\epsilon_0 R_{12}^3}$.

- Les forces de Debye dipôle permanent - dipôle induit : elles se manifestent lorsque le dipôle permanent μ_A d'une molécule A polaire induit une déformation du nuage électronique d'une molécule B qui peut être polaire ou non. Cette interaction provoque l'apparition d'un dipôle induit sur la molécule B proportionnel à sa polarisabilité α_B ¹³. Ces effets d'induction sont généralement pris en compte dans la contribution de polarisation. L'expression du potentiel de Debye est la suivante :

$$U_{\text{Debye}}^{AB} = -\frac{\mu_A^2 \alpha_B + \mu_B^2 \alpha_A}{4\pi\epsilon_0 R_{AB}^6}, \quad (1.15)$$

où l'influence du dipôle de la molécule B sur la molécule A est également prise en compte lorsque B est aussi une molécule polaire ($\mu_B \neq 0$).

- Les forces de London dipôle instantané - dipôle instantané : elles interviennent lorsque les fluctuations dans la densité électronique d'une molécule A font apparaître sur celle-ci un dipôle instantané qui peut interagir avec la densité électronique d'une molécule B et ainsi provoquer l'apparition d'un dipôle induit sur cette dernière. Ces forces existent dans tous les types de système (polaires ou non) et sont appelées forces de dispersion de London ou simplement dispersion. Le potentiel de dispersion de London, qui a été développé pour la modélisation des cristaux de gaz rares, est un potentiel atome-atome reposant sur l'hypothèse que cette interaction est additive. Il s'exprime donc entre deux molécules comme la somme des termes entre paires d'atomes :

$$U_{\text{London}}^{AB} = -\frac{3}{2} \sum_{i \in A} \sum_{j \in B} \frac{I_i I_j}{I_i + I_j} \frac{\alpha_i \alpha_j}{R_{ij}^6}, \quad (1.16)$$

où α_i, I_i et α_j, I_j sont respectivement les polarisabilités isotropes atomiques et les énergies de première ionisation des atomes i appartenant à la molécule A et j appartenant à la molécule B , et R_{ij} est la distance séparant ces deux atomes.

Ces forces d'interaction entre distributions électroniques moléculaires sont toutes attractives et sont proportionnelles à l'inverse de la distance intermoléculaire R à la puissance 6.

Dans les champs de force classiques, le terme "force de van der Waals" désigne simultanément les interactions non-covalentes attractives de dispersion et répulsives. Le potentiel de Lennard-Jones est souvent utilisé pour modéliser ces interactions entre paires d'atomes. Il comprend une partie attractive proportionnelle à R_{ij}^{-6} et une partie répulsive proportionnelle à $R_{ij}^{-\gamma}$, où $\gamma = 12$ ou 14 et R_{ij} est la distance entre deux atomes i et j . Pour $\gamma = 12$:

$$U_{\text{LJ}}^{ij}(R_{ij}) = 4\epsilon_{ij} \left[\left(\frac{\sigma_{ij}}{R_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{R_{ij}} \right)^6 \right] \quad (1.17)$$

Le paramètre ϵ_{ij} est la profondeur du puit de potentiel, c'est-à-dire le minimum de la fonction $U_{\text{LJ}}^{ij}(R_{ij})$ tandis que σ_{ij} est la distance interatomique pour laquelle $U_{\text{LJ}}^{ij}(R_{ij}) = 0$. La figure 1.7 illustre la forme de la fonction $U_{\text{LJ}}^{ij}(R_{ij})$. A courte distance, le potentiel est fortement répulsif

13. La polarisabilité discutée ici et dans le reste de ce manuscrit est homogène à un volume et est exprimée en unités du système international (SI). Il est cependant intéressant de noter que l'expression de la polarisabilité en système d'unités CGS est encore souvent utilisée, avec $\alpha_{\text{CGS}} = 4\pi\epsilon_0\alpha_{\text{SI}}$.

pour modéliser le caractère défavorable du recouvrement des distributions électroniques tandis qu'à plus longue distance il devient attractif grâce aux forces de dispersion de type London, avant de tendre vers zéro. L'hypothèse d'additivité de ces effets est également utilisée pour exprimer ce potentiel d'interaction entre deux molécules A et B : $U_{LJ}^{AB} = \sum_{i \in A} \sum_{j \in B} U_{LJ}^{ij}$. Une version légèrement différente de ce potentiel est utilisée dans les champs de force tels que AMBER, CHARMM, GROMACS et OPLS :

$$U_{VDW}^{AB} = \sum_{i \in A} \sum_{j \in B} \left(\frac{C_{ij}}{R_{ij}^{12}} - \frac{D_{ij}}{R_{ij}^6} \right). \quad (1.18)$$

Les paramètres empiriques C_{ij} et D_{ij} font partie du champ de force et sont optimisés pour chaque type d'atome dans le système considéré. La puissance 6 de la partie attractive a un sens physique car elle provient de l'interaction entre les dipôles instantanés comme dans le potentiel de London (équation 1.16) alors que l'exposant 12 ou 14 de la partie répulsive a été choisi comme compromis entre la modélisation d'une très forte pente et l'optimisation des moyens computationnels. L'amélioration des moyens techniques de calcul a permis d'appliquer d'autres formes de potentiel comme le potentiel de Buckingham :

$$U_{Buck}^{AB}(R) = \sum_{i \in A} \sum_{j \in B} \left(b_{ij} e^{-a_{ij} R_{ij}} - \frac{c_{ij}}{R_{ij}^6} - \frac{d_{ij}}{R_{ij}^8} \right). \quad (1.19)$$

Cette variante couramment utilisée dans les champs de force fait intervenir une fonction exponentielle décroissante pour la partie répulsive. Pour les parties attractives, le terme en R_{ij}^{-6} correspond aux effets dipôle instantané - dipôle induit et le terme en R_{ij}^{-8} aux effets dipôle instantané - quadripôle instantané. Les paramètres a_{ij} , b_{ij} , c_{ij} et d_{ij} sont également des paramètres empiriques optimisés pour chaque type d'atome. Un autre modèle qui est notamment utilisé dans le champ de force AMOEBA est le potentiel amorti (ou "buffered") 14-7 de Halgren [Halgren, 1992] qui emploie l'exposant 7 pour la partie attractive et l'exposant 14 pour la partie répulsion, ainsi que des constantes d'amortissement δ et γ qui sont optimisées pour reproduire les propriétés des gaz rares.

Forces électrostatiques

En général, les distributions de charge moléculaire sont modélisées dans les champs de force par des charges partielles ponctuelles localisées sur les noyaux des atomes. Le potentiel d'interaction électrostatique qui en découle est décrit par la loi de Coulomb pour des charges ponctuelles et peut être soit attractif soit répulsif. Ce potentiel étant additif, l'interaction entre deux molécules A et B est la somme des interactions atome-atome entre les deux molécules :

$$U_{elst}^{AB} = \frac{1}{4\pi\epsilon_0} \sum_{i \in A} \sum_{j \in B} \frac{q_i q_j}{R_{ij}}, \quad (1.20)$$

où q_i et q_j sont les charges partielles portées par l'atome i de la molécule A et par l'atome j de la molécule B respectivement, et R_{ij} est la distance séparant ces deux atomes. Les interactions électrostatiques étant persistantes à longue distance, ce potentiel ne peut pas être simplement

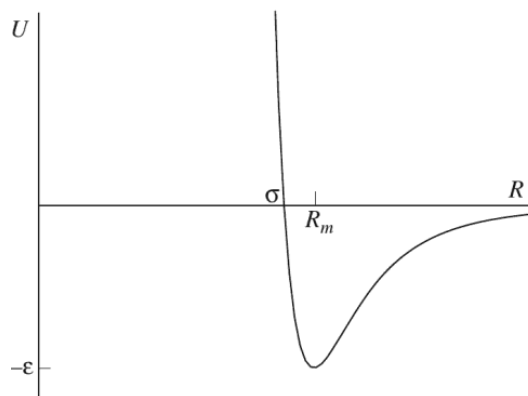


FIGURE 1.7 – Profil du potentiel d'interaction de Lennard-Jones.

Ce graphique issu du livre de A. Stone [Stone, 2013] représente la variation du potentiel interatomique $U (U_{LJ}^{ij}(R_{ij}))$ dans l'équation ??) en fonction de la distance interatomique R . Il existe une distance d'équilibre $R = R_m$ où le potentiel est minimal, c'est le puit de potentiel ϵ . Il s'agit de la configuration la plus stable d'un point de vue énergétique. A une distance plus courte $R = \sigma$, le potentiel passe par zéro. Il devient brusquement répulsif pour $R < \sigma$ tandis que, à plus longue distance $R > \sigma$, il devient attractif avant de tendre lentement vers zéro.

tronqué au-delà d'une certaine distance. Il existe des méthodes spécifiques pour traiter ce problème, dont la plus répandue est la sommation d'Ewald [Toukmaji et Board Jr, 1996], permettant de réduire autant que possible l'erreur de troncature dans le cas d'un système périodique. Par ailleurs, la nature ponctuelle des charges empêche de prendre en compte les effets de pénétration dus à l'écrantage incomplet des noyaux par les densités électroniques discutés précédemment, et sont généralement négligés.

Pour tenir compte de la polarisation, certains champs de force introduisent une constante diélectrique macroscopique ϵ à la place de la constante diélectrique du vide ϵ_0 dans l'équation 1.20 de la loi de Coulomb. Cependant, l'utilisation d'une seule valeur de ϵ pour l'ensemble du système est une approximation qui peut s'avérer trop sévère dans certains environnements hétérogènes comme l'intérieur des cavités des protéines ou encore dans les membranes cellulaires. Une autre possibilité est d'utiliser un champ de force dit polarisable pour traiter la contribution de polarisation de manière explicite. Par exemple, le champ de force AMOEBA ("Atomic Multipole Optimized Energetics for Biomolecular simulation") [Shi *et al.*, 2013] propose des moments dipolaires et quadripolaires atomiques ainsi que des polarisabilités atomiques calculés par méthode quantique en phase gazeuse pour la plupart des types atomiques rencontrés dans les molécules biologiques.

Modélisation des protons et des molécules d'eau

L'application des potentiels d'interaction aux macromolécules biologiques nécessite la modélisation des protons et des molécules d'eau de la couche de solvation mais aussi internes aux cavités des protéines notamment. En effet, les atomes d'hydrogène sont les principaux sites électrophiles participant aux interactions électrostatiques de ces systèmes et les transferts de protons sont courants dans les processus biologiques comme les réactions enzymatiques, par exemple. La prise en compte des molécules d'eau est également essentielle pour comprendre

beaucoup de mécanismes enzymatiques [Ball, 2008] et pour la reconnaissance entre les molécules biologiques [Hufner-Wulsdorf et Klebe, 2020]. Cependant, les méthodes de détermination des structures par diffraction des rayons X ne permettent généralement pas de modéliser les atomes d’hydrogène qui ont un faible pouvoir de diffusion, et seules les positions des atomes d’oxygène des molécules d’eau les plus stables peuvent être décelées, les orientations des hydrogènes restant inconnues. Lorsque les conditions expérimentales le permettent, il est possible de réaliser des affinements joints rayons X / neutrons pour modéliser les atomes d’hydrogène des molécules biologiques et des molécules d’eau [Howard *et al.*, 2016, Schiebel *et al.*, 2018]. Lorsque ce n’est pas possible, la mécanique moléculaire et certains champs de force offrent la possibilité de protoner *a posteriori* les structures et de modéliser explicitement les molécules de solvant. Pour les acides aminés et les acides nucléiques, la plupart des logiciels de modélisation moléculaire propose d’ajouter les protons dans une configuration géométriquement satisfaisante. Il est possible ensuite d’optimiser ces positions en fonction de leur environnement local, par exemple en estimant le pKa du groupement où ils sont attachés, grâce à des serveurs en lignes tels que MolProbity [Chen *et al.*, 2010], H++ [Anandakrishnan *et al.*, 2012] ou Yasara [Krieger *et al.*, 2009]. En dynamique moléculaire classique, les atomes d’hydrogène et les molécules d’eau du solvant sont explicitement modélisés et leurs orientations moyennes peuvent être estimées à partir des configurations les plus stables obtenues dans la simulation numérique. Néanmoins, ces configurations sont issues de moyennes statistiques reposant sur la validité du principe d’ergodicité, ce qui peut limiter la pertinence des structures statiques qui en sont extraites.

Les potentiels d’interaction empiriques permettent de décrire les principales caractéristiques des interactions électrostatiques, de dispersion et de répulsion. Néanmoins, ils reposent sur de nombreuses approximations parfois très sévères comme la réduction des distributions de charge à des charges partielles ponctuelles. De plus, la transférabilité des paramètres empiriques est limitée car ceux-ci sont optimisés pour un type de système donné. Par ailleurs, la polarisation électronique est sous-estimée dans la plupart des modèles alors que cette contribution peut s’avérer significative voire déterminante dans certains processus de reconnaissance moléculaire. Les méthodes de chimie quantique, qui vont être présentées dans la section suivante, sont dites *ab initio*, c’est-à-dire sans *a priori* sur le système considéré, et permettent de considérer l’ensemble des contributions à l’énergie d’interaction totale au niveau de précision le plus élevé atteignable aujourd’hui.

1.4.3 Méthodes de chimie quantique

En mécanique quantique, les positions des charges (noyaux et électrons) ne sont pas déterminées, elles sont probabilistes. En effet, le principe d’incertitude de Heisenberg stipule que la position et la vitesse d’une particule ne peuvent être mesurées simultanément. En revanche, la densité de probabilité $P(r, t)$ de présence d’une particule à la position r à un temps t donné est calculable à partir de la fonction d’onde $\psi(r, t)$ du système par : $P(r, t) = |\psi(r, t)|^2$. La fonction d’onde $\psi(r, t)$ du système est obtenue par résolution de l’équation de Schrödinger dépendante

du temps [Schrödinger, 1926] :

$$\hat{H}(r, t)\psi(r, t) = i\frac{\partial\psi(r, t)}{\partial t}, \quad (1.21)$$

où $\hat{H}(r, t)$ est l'opérateur hamiltonien du système dépendant du temps. En régime stationnaire, l'équation de Schrödinger devient :

$$\hat{H}\psi(r) = E\psi(r), \quad (1.22)$$

où E est l'énergie totale du système dans son état stationnaire. L'opérateur hamiltonien $\hat{H}$ est défini comme la somme des opérateurs d'énergie cinétique $\hat{T}$ et d'énergie potentielle $\hat{V}$, telle que : $\hat{H} = \hat{T} + \hat{V}$. L'opérateur énergie cinétique $\hat{T}$ comporte un premier terme se rapportant à l'énergie cinétique des M noyaux de masse m_k , et un second terme pour l'énergie cinétique des N électrons du système :

$$\hat{T} = \sum_{k=1}^M \frac{\hbar^2}{2m_k} \nabla_k^2 - \sum_{i=1}^N \frac{\hbar^2}{2m_e} \nabla_i^2, \quad (1.23)$$

avec $\hbar = h/2\pi$, la constante de Planck et m_e , la masse de l'électron. L'opérateur d'énergie potentielle $\hat{V}$ se compose de plusieurs contributions :

$$V = \frac{1}{4\pi\epsilon_0} \left[- \sum_{i=1}^n \sum_{k=1}^M \frac{eZ_k}{R_{i,k}} + \sum_{i=1}^n \sum_{j<i} \frac{e^2}{r_{ij}} + \sum_{k=1}^M \sum_{l<k} \frac{Z_k Z_l}{R_{k,l}} \right], \quad (1.24)$$

représentant respectivement l'attraction coulombienne électron-noyau et les répulsions coulombiennes électron-électron et noyau-noyau, avec Z_k , la charge du noyau k et e , la charge élémentaire.

Résolution de l'équation de Schrödinger

L'approximation de Born-Oppenheimer permet de négliger le couplage des mouvements des électrons et des noyaux. Le terme d'énergie cinétique des noyaux dans l'équation 1.23 peut donc être négligé et le terme de répulsion noyau-noyau de l'équation 1.24 est constant. Excepté pour l'atome d'hydrogène et les ions hydrogénoïdes, la résolution exacte de l'équation de Schrödinger (équation 1.22) est impossible à cause du terme électron-électron dont l'expression est inconnue. Les méthodes de chimie quantique proposent des approximations permettant d'obtenir des solutions approchées, les plus connues étant [Veszprémi et Fehér, 1999] : les méthodes Hartree-Fock (HF) et post-HF, et la théorie de la fonctionnelle de la densité (DFT). Dans les méthodes HF, les N électrons sont considérés comme indépendants les uns des autres mais soumis à un champ moyen généré par les $N - 1$ autres électrons. Cette approximation peut s'avérer trop abrupte notamment car elle a pour conséquence une surestimation de la probabilité que deux électrons soient très proches dans l'espace, appelée erreur de corrélation. Les méthodes post-HF permettent quant à elles de tenir compte de la corrélation statique des électrons, due aux configurations multiples dans la fonction d'onde, et de la corrélation dynamique des électrons, liée à la probabilité non-nulle d'avoir deux électrons à la même position. Parmi ces méthodes, les plus populaires sont la méthode de cluster couplé CC ("coupled cluster") [Coester et Kümmel,

1960, Bartlett et Musiał, 2007] et la théorie de la perturbation de Møller-Plesset MP [Møller et Plesset, 1934, Cremer, 2011]. Elles sont d'une grande précision mais sont particulièrement coûteuses en temps de calcul.

En revanche, la théorie de la fonctionnelle de la densité (DFT) a un coût computationnel moins élevé. Elle a été développée dans le cadre théorique des théorèmes de Hohenberg et Kohn [Hohenberg et Kohn, 1964] et des équations de Kohn-Sham [Kohn et Sham, 1965], et repose sur le principe selon lequel l'énergie électronique est entièrement déterminée par la densité électronique ρ . L'énergie E peut donc être exprimée comme une fonctionnelle de la densité ρ :

$$E[\rho] = T[\rho] + V_{ee}[\rho] + V_{Ne}[\rho], \quad (1.25)$$

où $T[\rho]$ est l'énergie cinétique des électrons, $V_{ee}[\rho]$ est l'énergie de répulsion coulombienne électron-électron et $V_{Ne}[\rho]$ est l'énergie d'attraction coulombienne électron-noyau. Les expressions exactes de $T[\rho]$ et $V_{ee}[\rho]$ sont inconnues mais peuvent être approximées par la fonctionnelle $F[\rho]$ proposée par Kohn et Sham [Kohn et Sham, 1965] qui remplace le système réel constitué de particules en interaction par un gaz de particules sans interaction :

$$F[\rho] = T_s[\rho] + J[\rho] + E_{xc}[\rho], \quad (1.26)$$

où $T_s[\rho]$ est l'énergie cinétique du gaz de particules sans interaction, $J[\rho]$ est le terme coulombien classique, et $E_{xc}[\rho]$ est l'énergie d'échange-corrélation dont l'expression analytique exacte est inconnue. Ce dernier terme peut toutefois être estimé par les approximations de la densité locale (LDA) ou du gradient généralisé (GGA) [Dobson *et al.*, 2013].

Ces méthodes de chimie quantique permettent de déterminer l'énergie totale d'un système par résolution de l'équation de Schrödinger (équation 1.22). Dans les méthodes dites supermoléculaires [Chalaśiński et Szcześniak, 2000], l'énergie d'interaction E_{inter}^{AB} entre deux molécules A et B est définie comme la différence entre l'énergie totale du complexe AB , E_{tot}^{AB} , et les énergies totales des molécules isolées, $E_{\text{tot}}^A + E_{\text{tot}}^B$. Cependant, l'énergie d'interaction étant inférieure de plusieurs ordres de grandeur aux énergies totales qui sont soustraites, il peut arriver que les erreurs sur les énergies totales dues aux approximations des méthodes utilisées soient du même ordre que l'énergie d'interaction elle-même [Patkowski, 2020]. De plus, les approches supermoléculaires ne donnent aucune information sur la nature et l'interprétation physique des interactions.

Méthodes de décomposition de l'énergie

Les méthodes d'analyse de décomposition de l'énergie ou EDA ("Energy Decomposition Analysis") [von Hopffgarten et Frenking, 2012] permettent de distinguer les multiples contributions aux interactions intermoléculaires dans un système. Les approches EDA variationnelles introduites par J. Morokuma et ses collègues [Morokuma, 1971] décrivent les différents types d'interactions intervenant dans le système comme des variations de l'énergie : $\Delta E_{\text{int}} = \Delta E_{\text{elect}} + \Delta E_{\text{Pauli}} + \Delta E_{\text{orb}}$. Le premier terme est l'interaction électrostatique classique coulombienne, le second la répulsion de Pauli et le dernier est l'énergie d'interaction orbitaire

stabilisante. Il est possible d'ajouter un terme énergétique de dispersion grâce à la correction de Grimme [Grimme *et al.*, 2010]. La décomposition SAPT ("Symmetry-Adapted Perturbation Theory") [Szalewicz et Jeziorski, 1979, Szalewicz, 2012, Patkowski, 2020] est une méthode EDA perturbative. Elle fournit diverses contributions à l'énergie d'interaction totale sur la base de la théorie de la perturbation. Dans sa forme la plus simple, les termes électrostatique E_{elst} , d'induction E_{elst} , de dispersion E_{elst} et d'échange-répulsion E_{elst} sont calculés :

$$E_{\text{int}}^{\text{SAPT}} = E_{\text{elst}} + E_{\text{ind}} + E_{\text{disp}} + E_{\text{exch-rep}}. \quad (1.27)$$

L'énergie électrostatique E_{elst} est l'interaction coulombienne entre les distributions de charge des molécules isolées (ou non-perturbées) et prend en compte les effets de pénétration des densités. Cette contribution peut être soit attractive soit répulsive. L'énergie d'induction E_{ind} , aussi appelée énergie de polarisation, est toujours attractive et provient de l'interaction entre les multipôles mutuellement induits dans la distribution de charge d'une molécule par le champ électrique statique généré par la distribution de charge non-perturbée de l'autre molécule. La dispersion E_{disp} émerge des corrélations entre les multipôles instantanés causés par les fluctuations dans les distributions électroniques. La contribution d'échange-répulsion $E_{\text{exch-rep}}$ est la conséquence du recouvrement des densités électroniques moléculaires et résulte de la combinaison de deux effets, le premier étant attractif et l'autre répulsif. La partie attractive provient des électrons devenus libres de se déplacer d'une molécule à l'autre, ce qui a pour effet d'augmenter l'incertitude sur leurs positions et de diminuer leur énergie. La partie répulsive est une conséquence du principe de Pauli de l'antisymétrisation de la fonction d'onde pour que les électrons de même spin ne puissent pas se trouver à la même position, ce qui coûte de l'énergie. Le second effet étant dominant, l'effet global de l'échange-répulsion est répulsif. Les méthodes SAPT de niveaux de théorie plus élevés font également intervenir des termes croisés tels que l'échange-induction ou l'échange-dispersion, ou encore la contribution des transferts de charge qui est généralement incluse dans le terme d'induction E_{ind} . Les méthodes EDA variationnelles et perturbatives ont été utilisées pour caractériser les interactions entre molécules biologiques et ont montré un fort potentiel d'application à la conception de médicaments [Phipps *et al.*, 2015].

Méthodes de fragmentation des molécules

Pour diminuer les temps de calculs élevés de ces méthodes, des approches de fragmentation ont été développées pour partitionner le système en fragments plus petits et traitables d'un point de vue computationnel. Les méthodes diviser et conquérir "Divide and Conquer" [Yang, 1991, Dixon et Merz Jr, 1997, He et Merz Jr, 2010] et d'adaptation moléculaire "Molecular Tailoring" [Gadre *et al.*, 1994, Sahu et Gadre, 2014] réalisent les calculs d'énergie sur des sous-unités puis les recombinaient pour construire la fonction d'onde ou la densité électronique totale du système. Dans les méthodes de fragmentation de l'interaction, les énergies d'interaction d'un système sont obtenues par sommation des propriétés calculées dans des sous-unités et des corrections d'interactions entre ces sous-unités. Des exemples de ces méthodes sont : la fragmentation moléculaire avec capsules conjuguées MFCC ("Molecular Fractionation with Conjugated Caps") [Zhang et Zhang, 2003, Li *et al.*, 2005], les orbitales moléculaires fragmentées FMO ("Fragment Molecular Orbital") [Kitaura *et al.*, 1999, Nakano *et al.*, 2000, Fedorov et Kitaura,

2006, Fedorov, 2017] et la méthode d'énergie de cœur KEM ("Kernel Energy Method") [Huang *et al.*, 2005].

Méthodes hybrides multi-échelles

Malgré ces méthodes de fragmentation, le coût computationnel des calculs d'énergie *ab initio* reste trop élevé pour appliquer ces méthodes aux systèmes de grande taille tels que les protéines. Les méthodes hybrides mécanique quantique / mécanique moléculaire ou QM/MM ("Quantum Mechanics / Molecular Mechanics") [Warshel et Levitt, 1976, Field *et al.*, 1990, Gao, 1996, Senn et Thiel, 2009] sont des approches dites d'intégration multi-échelles ("multiscale embedding approaches") dans lesquelles les systèmes larges sont traités par des méthodes de chimie quantique de haut niveau théorique dans leur région chimiquement active, comme le site actif d'une protéine, et par des méthodes classiques ailleurs. M. Karplus, M. Levitt et A. Warshel ont obtenu le prix Nobel de chimie en 2013 pour le développement de ces méthodes multi-échelles [Karplus, 2014, Levitt, 2014, Warshel, 2014]. Dans ces méthodes, le système est scindé en deux régions, une traitée en QM et l'autre en MM, l'énergie totale du système étant alors la somme des énergies de ces deux régions et d'un terme d'interaction entre celles-ci :

$$E_{\text{tot}} = E^{\text{QM}} + E^{\text{MM}} + E^{\text{QM/MM}}. \quad (1.28)$$

La partie soumise à un traitement quantique E^{QM} est obtenue par méthodes de calculs *ab initio* de chimie quantique de haut niveau théorique. La partie mécanique moléculaire E^{MM} est calculée par des potentiels interatomiques de champs de force comme définis dans la section 1.4.2. Plusieurs méthodes ont été développées pour décrire la frontière entre ces deux régions et obtenir l'énergie d'interaction $E^{\text{QM/MM}}$. Une des méthodes QM/MM les plus répandues est l'approche ONIOM développée par Morokuma et ses collègues [Svensson *et al.*, 1996, Chung *et al.*, 2015]. Cette méthode traite le problème de la frontière entre les deux régions en remplaçant les atomes se trouvant dans une des deux régions et proches de la frontière par des atomes d'hydrogène dans l'autre région. L'approche QM/ELMO [Macetti et Genoni, 2019, Macetti *et al.*, 2021] est un autre type d'approche multi-échelles dans laquelle la région la plus importante du système est traitée à un niveau de théorie complètement quantique tandis que le reste est décrit à l'aide du transfert des ELMO mentionné dans la section 1.2.3. Les approches multi-échelles sont très utilisées dans les simulations numériques par dynamique moléculaire [Engkvist *et al.*, 2000, Senn et Thiel, 2009] et ont permis d'obtenir des résultats très prometteurs pour la modélisation des réactions enzymatiques [Bennie *et al.*, 2016, Ranaghan *et al.*, 2019].

En résumé,

les potentiels d'interaction classiques des champs de force permettent d'obtenir à faible coût l'estimation des contributions à l'énergie d'interaction mais leur caractère empirique et les approximations sur lesquelles ils reposent les rendent insuffisants pour décrire avec précision les nombreuses interactions entre molécules. A l'inverse, les méthodes de chimie quantique décrivent les interactions de façon très précise mais ont un coût computationnel élevé qui les empêche d'être appliquées aux macromolécules biologiques. Les méthodes hybrides multi-échelles de QM/MM

tire le meilleur de ces deux types de méthodologie en proposant une description précise des régions de petite taille importante pour la réactivité des macromolécules et une description classique peu coûteuse en temps de calcul et de précision suffisante pour le reste du système. Néanmoins, ces méthodes restent difficiles à mettre en œuvre dans l'étude des macromolécules [Tzeliou *et al.*, 2022, Vennelakanti *et al.*, 2022, Clemente *et al.*, 2023]. En effet, d'un point de vue méthodologique, elles nécessitent une expertise particulière pour par exemple [Clemente *et al.*, 2023] : la définition de la frontière entre les traitements QM et MM, la préparation et l'équilibration du système ou encore le choix du niveau de théorie de la partie QM. De plus, d'un point de vue technique, malgré d'importants progrès ces dernières années, les méthodes QM/MM nécessitent de disposer de moyens de calculs très sophistiqués, de type cluster de calcul GPU ("Graphical Processing Units").

Dans mon projet doctoral, j'ai développé un modèle d'énergie d'interaction qui se veut simple d'utilisation, ou "user-friendly". Il repose sur des potentiels d'interaction semi-classiques, qui sont peu coûteux en temps de calcul, et sur la description expérimentale fine de la densité électronique, qui décrit le système avec précision et qui est rapidement obtenue grâce aux transferts des paramètres multipolaires de la librairie ELMAM2. Le potentiel d'interaction total U_{int} est exprimé suivant une décomposition de type SAPT de l'énergie (équation 1.27) en quatre contributions : électrostatique U_{elst} , d'induction U_{ind} , de dispersion U_{disp} et d'échange-répulsion $U_{\text{exch-rep}}$. Le calcul des contributions U_{elst} et U_{ind} a déjà été proposé à partir des paramètres du modèle multipolaire de la densité électronique d'origine expérimentale et les tenseurs de polarisabilités anisotropiques transférables d'origine théorique fournis par la librairie ELMAM2 [Fournier, 2010, Leduc *et al.*, 2019, Vuković *et al.*, 2021]. Des modèles reposant sur ces quantités transférables des contributions de dispersion U_{disp} et de répulsion U_{rep} sont proposés dans cette thèse. Les détails de ces modèles ainsi que la discussion de leur validité seront décrits dans le chapitre 3. Des calculs d'énergies d'interaction dans les protéines glutathion transférase et halorhodopsine seront présentés dans le chapitre 4.

1.5 Résumé de l'introduction et objectifs de la thèse

Pour conclure ce premier chapitre d'introduction, ma thèse de doctorat s'inscrit dans la volonté d'étendre les approches de cristallographie quantique à l'étude des macromolécules biologiques. Plus particulièrement, elle vise à développer des descripteurs électrostatiques basés sur la densité électronique moléculaire transférée d'origine expérimentale. La description fine de la densité électronique moléculaire permise par le modèle multipolaire de Hansen et Coppens peut être appliquée aux macromolécules biologiques grâce à la transférabilité des paramètres de ce modèle entre pseudo-atomes de même type. En effet, les bases de données de ces paramètres ELMAM/ELMAM2 et UBDB/MATTS ont été développées pour reconstruire les densités électroniques de protéines. En particulier, la librairie ELMAM2 propose des paramètres multipolaires d'origine expérimentale, obtenus par affinement multipolaire contre des données de diffraction des rayons X à résolution subatomique dans des cristaux de peptides et de petites molécules organiques, et également des polarisabilités atomiques anisotropes déterminées théoriquement et transférables aux macromolécules pour polariser les densités moléculaires en interactions.

Grâce à la précision des distributions électroniques transférées $\rho(\mathbf{r})$, les méthodes d'analyse

de densité de charge reposant sur l'étude de la topologie de $\rho(\mathbf{r})$ sont appliquées aux macromolécules biologiques. La densité de charge obtenue à partir de $\rho(\mathbf{r})$ transférée et des charges ponctuelles des noyaux permet également de calculer le potentiel électrostatique. La projection bidimensionnelle du potentiel électrostatique sur une surface moléculaire est traditionnellement utilisée pour caractériser la complémentarité électrostatique entre deux molécules en interaction comme un ligand et le site actif d'une protéine notamment. L'analyse des lignes de champ électrique permet quant à elle de visualiser la directionnalité des forces électrostatiques dans l'espace moléculaire et de déterminer les influences à longues portées pouvant aider à comprendre les mécanismes de reconnaissance moléculaire par exemple. L'analyse topologique du potentiel électrostatique est similaire à celle de la densité électronique mais donne accès à des informations supplémentaires. En particulier, elle fournit une partition de l'espace en bassins topologiques associés aux sites électrophiles et nucléophiles. Ces approches ont été appliquées uniquement sur les petites molécules pour l'instant. Le premier objectif de cette thèse est de développer des descripteurs issus de la topologie du potentiel électrostatique ainsi que les outils permettant de les appliquer aux complexes protéine-ligand afin de révéler visuellement les contributeurs électrostatiques aux interactions entre le ligand et le site actif de la protéine ainsi que l'étendue spatiale de leurs influences.

La distribution de charge modélisée à partir des données expérimentales en utilisant le modèle multipolaire permet d'évaluer avec précision l'énergie d'interaction électrostatique entre deux molécules. Les densités de charge reconstruites à partir des paramètres multipolaires et des polarisabilités ELMAM2 permettent également de déterminer cette énergie mais aussi la contribution des dipôles induits à l'énergie de polarisation entre deux molécules et notamment entre un ligand et les résidus de la poche de fixation de la protéine. Néanmoins, pour estimer l'énergie d'interaction intermoléculaire totale, il manque l'évaluation des contributions de dispersion et de répulsion. Les potentiels d'interaction des champs de force proposent des modèles paramétrés pour estimer ces contributions à faible coût computationnel mais leur caractère empirique ainsi que les approximations utilisées pour leur définition limitent la précision de leurs résultats pour étudier la réactivité enzymatique par exemple. De plus, ils sont paramétrés pour être consistants avec les autres paramètres définissant le champ de force, limitant leur transférabilité à d'autres applications. Les méthodes *ab initio* de chimie quantique fournissent quant à elles des résultats d'une grande précision mais sont trop coûteuses en temps de calcul pour être appliquées sur l'ensemble d'une protéine. Les approches hybrides de QM/MM permettent d'étudier les macromolécules biologiques en tirant avantage de la précision des méthodes quantiques pour décrire la région chimiquement active et des méthodes classiques pour traiter le reste du système à coût computationnel moindre. Ainsi, le second objectif de ma thèse est de développer un potentiel d'interaction total qui quant à lui repose sur la description complète du système sur la base des données cristallographiques expérimentales ELMAM2, tout en introduisant un minimum de paramètres n'ayant pas d'interprétation physique, et en minimisant les temps de calcul. Pour cela, j'ai complété les évaluations des contributions électrostatiques et de polarisation déjà établies à partir des données de la librairie ELMAM2 par l'estimation des effets de dispersion et de répulsion grâce à des modèles reposant sur les paramètres multipolaires et les polarisabilités ELMAM2.

Dans la suite de ce manuscrit, les développements méthodologiques réalisés dans le but d'atteindre ces deux objectifs seront détaillés. Le chapitre 2 traitera des descripteurs issus de la topologie du potentiel électrostatique tandis que le chapitre 3 décrira la construction du potentiel d'interaction total ELMAM2. Puis, plusieurs applications à des complexes de protéines seront présentées dans le chapitre 4. En particulier, dans la partie 4.1, l'apport de nos descripteurs électrostatiques comme nouveau point de vue complétant les analyses classiques des structures atomiques de macromolécules sera exposé par l'application de ceux-ci au complexe largement étudié de la trypsine avec un inhibiteur canonique. Une application à la neuropiline dans la partie 4.2 permettra d'illustrer l'utilisation de ces descripteurs en biologie structurale. La partie 4.3 proposera un exemple de calculs d'énergie d'interaction électrostatique pour l'étude du complexe enzyme-ligand de la glutathion transférase avec le glutathion. Dans la partie 4.4, la perspective d'analyse de la dynamique d'un mécanisme, le pompage d'ion chlorure par l'halorhodopsine, par le calcul des énergies d'interaction ELMAM et du point de vue de nos descripteurs issus de la topologie du potentiel électrostatique sera présentée. Enfin, je finirai par une conclusion globale des développements et applications réalisés et les perspectives ouvertes par ces travaux seront discutées.

Chapitre 2

Descripteurs issus de la topologie du potentiel électrostatique

Traditionnellement, le potentiel électrostatique moléculaire $V(\mathbf{r})$ est utilisé en biologie structurale pour caractériser les surfaces d'accessibilité au solvant entourant les protéines et leurs ligands (voir section 1.3.2). En colorant cette surface par les valeurs de $V(\mathbf{r})$, cette approche permet de localiser les régions chargées positivement ou négativement d'une protéine et d'estimer la complémentarité électrostatique entre un ligand et la poche de fixation. Néanmoins, elle est limitée à une évaluation en surface de $V(\mathbf{r})$ et repose souvent sur des distributions ponctuelles des charges partielles atomiques. Les contributeurs aux interactions électrostatiques ne peuvent donc pas être localisés avec précision par ces méthodes. Par ailleurs, pour étudier la réactivité chimique des petites molécules, un autre type d'approche basé sur le potentiel électrostatique a été développé : l'analyse topologique de $V(\mathbf{r})$. La topologie du potentiel électrostatique repose sur la détermination de ses points critiques et de la distribution de ses lignes de gradient, c'est-à-dire les lignes de champ électrique $\mathbf{E}(\mathbf{r}) = -\nabla V(\mathbf{r})$, ainsi que sur la partition de l'espace réalisée par les bassins topologiques qui découlent de la topographie du champ électrique. Lorsque le potentiel $V(\mathbf{r})$ est issu d'un modèle fin de la densité électronique, comme par transfert des pseudo-atomes multipolaires (voir section 1.2.3), les positions des sites électrophiles et nucléophiles ainsi que l'étendue et la direction dans l'espace moléculaire de leurs influences électrostatiques sont révélées par ces descripteurs. L'application de ces méthodes aux complexes protéine-ligand est donc prometteuse pour la description des mécanismes de reconnaissance moléculaire, pour comprendre l'approche ainsi que la fixation du ligand dans la cavité de la protéine. Dans le cadre de ma thèse, j'ai développé des descripteurs électrostatiques basés sur la topologie de $V(\mathbf{r})$ et qui ont pour ambition d'être appliqués aux études de complexes protéine-ligand en biologie structurale. Dans ce chapitre, je vais détailler les différents développements méthodologiques que j'ai réalisés pour définir ces descripteurs, en commençant par les points critiques du potentiel, la topographie des lignes de champ électrique, puis la partition de l'espace en zones d'influence électrophile et nucléophile. L'implémentation des outils pour mettre en pratique ces descripteurs dans le logiciel MoProViewer ainsi que quelques perspectives de développements seront également présentées.

2.1 Points critiques

2.1.1 Définition des points critiques du potentiel électrostatique

Définitions

La définition des points critiques du potentiel électrostatique moléculaire $V(\mathbf{r})$ est très similaire à celle des points critiques de la densité électronique $\rho(\mathbf{r})$ présentée dans la section 1.3.1. Les points critiques de $V(\mathbf{r})$ sont les points de coordonnées $\mathbf{r}_c$ tels que : $\nabla V(\mathbf{r} = \mathbf{r}_c) = \mathbf{0}$. Ils sont identifiés par le couple (R, S) . Le rang R de la matrice hessienne $H(V(\mathbf{r}))$ du potentiel est le nombre de valeurs propres non-nulles de H en $\mathbf{r} = \mathbf{r}_c$. Sa signature S est la somme algébrique des signes de ses valeurs propres, qui sont aussi appelées courbures de $V(\mathbf{r})$. Le potentiel électrostatique étant défini dans $\mathbb{R}^3$, en général $R = 3$ dans un système à l'équilibre¹ et la signature S peut prendre les valeurs $-3, -1, +1$ ou $+3$. Il existe donc quatre types de points critiques de $V(\mathbf{r})$: les maxima locaux $(3, -3)$, les minima locaux $(3, +3)$ et les points-selles $(3, -1)$ et $(3, +1)$.

Interprétations

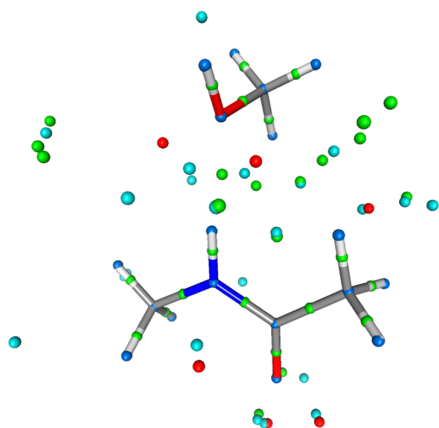
L'interprétation des points critiques de la densité électronique $\rho(\mathbf{r})$, en termes de structure et de liaisons moléculaires, est clairement définie dans le cadre de la théorie QTAIM de R. Bader des atomes dans les molécules (voir section 1.3.1). En revanche, le potentiel électrostatique $V(\mathbf{r})$ ne bénéficie pas d'un tel cadre théorique tandis que sa topologie est plus riche et complexe que celle de la densité électronique. Certains des points critiques de $V(\mathbf{r})$ peuvent être associés à des interprétations similaires à ceux de $\rho(\mathbf{r})$ mais ce n'est pas le cas de beaucoup d'autres qui sont moins discutés dans la littérature [Gadre et Bendale, 1986, Leboeuf *et al.*, 1999, Balanarayan et Gadre, 2003, Mata *et al.*, 2007, Gadre et Kumar, 2016, Anjalikrishna *et al.*, 2019]. Pour détailler les différents types de points critiques de $V(\mathbf{r})$, je vais m'appuyer dans ce qui suit sur l'exemple du complexe N-méthylacétamide - méthanol représenté dans la figure 2.1.

Comme le montre la figure 2.1b, les maxima locaux $(3, -3)$ de $V(\mathbf{r})$ sont localisés sur les positions des noyaux², comme ceux de $\rho(\mathbf{r})$. Les minima locaux $(3, +3)$ de $V(\mathbf{r})$ correspondent quant à eux à des concentrations locales d'électrons. Ces concentrations d'électrons révèlent les positions des paires d'électrons non-liantes [Kumar *et al.*, 2014a, Gadre et Kumar, 2016]³. Notons que celles-ci ne sont pas caractérisées par les points critiques de $\rho(\mathbf{r})$ mais plutôt par ceux du Laplacien $\nabla^2 \rho(\mathbf{r})$. Dans la figure 2.1c, les minima locaux Cp50 et Cp52 correspondent aux deux doublets non-liants de l'atome d'oxygène du méthanol, les Cp51 et Cp76 aux doublets non-liants de l'oxygène du N-méthylacétamide, et le Cp53 au doublet non-liant de l'azote du

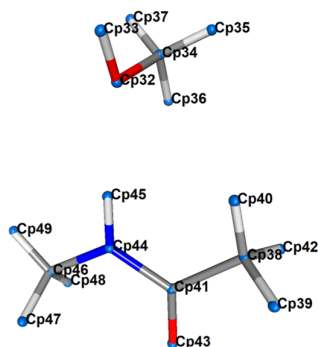
1. Il est possible d'avoir $R < 3$ même si le système est dans son état fondamental si celui-ci présente des symétries particulières [Gadre *et al.*, 1992].

2. Il est intéressant de noter que lorsque les noyaux sont traités comme des distributions ponctuelles, modélisées par des fonctions de Dirac $\delta(\mathbf{R}_i)$ aux positions $\mathbf{R}_i$, le potentiel électrostatique $V(\mathbf{r})$ tend vers l'infini en $\mathbf{R}_i$ donc ses courbures ne sont pas définies en ses positions. Les maxima locaux $(3, -3)$ émergent donc du comportement asymptotique de $V(\mathbf{r})$ autour des positions des noyaux [Mata *et al.*, 2007].

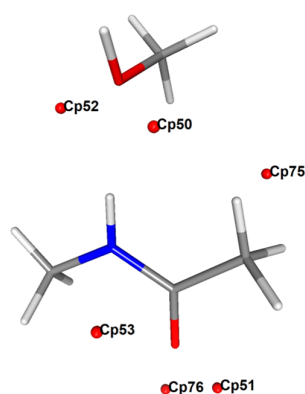
3. Le terme de paire d'électrons non-liante ou de doublet non-liant (ou "lone pair") est utilisé dans ce manuscrit au sens de la définition donnée par A. Kumar et S. R. Gadre [Kumar *et al.*, 2014a], c'est-à-dire pour qualifier une concentration localisée d'électrons de valence non-impliqués dans une liaison covalente. Il est important de noter ici que tous les minima locaux du potentiel électrostatique ne correspondent pas nécessairement à un doublet non-liant.



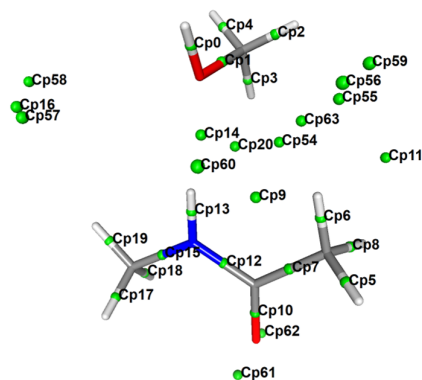
(a) Ensemble des points critiques.



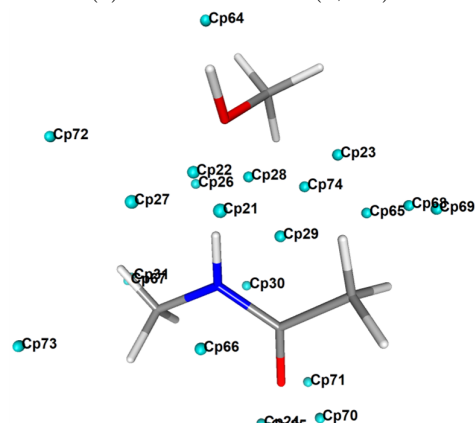
(b) Maxima locaux (3, -3).



(c) Minima locaux (3, +3).



(d) Points-selles (3, -1).



(e) Points-selles (3, +1).

FIGURE 2.1 – Les points critiques du potentiel électrostatique dans le complexe N-méthylacétamide - méthanol.

(a) Tous les points critiques du potentiel électrostatique $V(\mathbf{r})$ sont représentés dans le complexe N-méthylacétamide - méthanol, avec des sphères bleues pour les maxima locaux (3, -3), rouges pour les minima locaux (3, +3), vertes pour les points-selles (3, -1) et cyan pour les points-selles (3, +1). (b) Les maxima locaux (3, -3) sont localisés sur les positions des noyaux. (c) Les minima locaux (3, +3) correspondent à des concentrations locales d'électrons telles que les doublets non-liants. (d) Les points-selles (3, -1) sont positionnés sur les liaisons covalentes pour certains et dans l'espace intermoléculaire pour d'autres. (e) Les points-selles (3, +1) peuvent apparaître dans les structures de cycle et ailleurs dans l'espace intermoléculaire.

N-méthylacétamide. Le minimum local Cp75 n'est quant à lui associé à aucun doublet non-liant, il apparaît en raison de la présence des déplétions locales d'électrons sur les atomes d'hydrogène des groupements méthyles qui l'entourent. En effet, le potentiel électrostatique à cette position est positif, avec une valeur de $V(\mathbf{r}_{\text{Cp75}}) = 0,016 \text{ e.}\text{\AA}^{-1}$ tandis qu'aux positions des noyaux des hydrogènes (Cp36 et Cp40 de la figure 2.1b) $V(\mathbf{r}_{\text{Cp36}}) = V(\mathbf{r}_{\text{Cp40}}) = 17,6 \text{ e.}\text{\AA}^{-1}$. Il s'agit donc bien d'un minimum local du potentiel mais, puisque $V(\mathbf{r}_{\text{Cp75}})$ est proche de zéro et positif, il ne correspond pas à une concentration d'électrons. Les paires d'électrons non-liantes de l'oxygène du N-méthylacétamide (Cp51 et Cp76 de la figure 2.1c) portent quant à eux des valeurs négatives du potentiel : $V(\mathbf{r}_{\text{Cp51}}) = -0,206 \text{ e.}\text{\AA}^{-1}$ et $V(\mathbf{r}_{\text{Cp76}}) = -0,218 \text{ e.}\text{\AA}^{-1}$, propres à une concentration locale de charge négative. En effet, les doublets non-liants sont nettement définis dans ce système, ce qui leur permet d'être caractérisés par un minimum local de potentiel. Ce n'est pas toujours le cas, il peut arriver que ces paires d'électrons soient diffuses dans une région de l'espace [Ahmed *et al.*, 2013] et ne puissent pas être localisées par un point critique. Ce cas de figure sera rencontré dans les applications présentées dans le chapitre 4.

Parmi les points-selles $(3, -1)$ et $(3, +1)$ de $V(\mathbf{r})$, certains correspondent aux points-selles des liaisons (BCP) et des cycles (RCP) de $\rho(\mathbf{r})$. Par exemple, dans la figure 2.1d, des points-selles $(3, -1)$ apparaissent sur les liaisons covalentes et non-covalentes. En particulier, les points Cp14 et Cp54 correspondent aux deux BCP de $\rho(\mathbf{r})$ caractérisant les liaisons hydrogène observées dans la figure 1.3a. Pour les points-selles $(3, +1)$ sur la figure 2.1e, le point Cp29 est similaire au RCP de la densité apparaissant dans la figure 1.3a. Les autres points-selles sont plus difficiles à associer à des aspects structuraux ou d'interactions interatomiques. E. Espinosa et ses collègues [Mata *et al.*, 2007] ont proposé une interprétation en termes de délimitations des bassins topologiques de $V(\mathbf{r})$ et de points d'entrée des zones d'influence électrophile pour les $(3, -1)$ et nucléophile pour les $(3, +1)$ lors d'attaques nucléophiles ou électrophiles, respectivement. Ce dernier aspect sera davantage discuté dans la partie 2.3.

2.1.2 Algorithme de recherche des points critiques

Minimisation de Newton-Raphson

D'une manière générale, les points critiques $\mathbf{r}_c$ d'un champ scalaire $f(\mathbf{r})$ sont définis par : $\nabla f(\mathbf{r} = \mathbf{r}_c) = \mathbf{0}$. Leurs positions $\mathbf{r}_c$ peuvent donc être déterminées en utilisant une méthode numérique de recherche des racines (ou zéros) de la fonction $\nabla f(\mathbf{r})$ dans $\mathbb{R}^3$. Pour cela, une des méthodes les plus connues et simple à implémenter est la minimisation de Newton-Raphson. La minimisation de Newton-Raphson est une méthode itérative permettant de suivre la direction dans laquelle se trouve le point critique à partir d'une position initiale $\mathbf{r}_0$. A chaque itération, la nouvelle position $\mathbf{r}_{k+1}$ est calculée à partir de la position actuelle $\mathbf{r}_k$ et du pas de déplacement $\mathbf{h}$, telle que : $\mathbf{r}_{k+1} = \mathbf{r}_k + \mathbf{h}$.

L'expression du pas de déplacement $\mathbf{h}$ provient du développement limité à l'ordre 2 de la fonction $f(\mathbf{r}_{k+1})$. Par exemple, pour une fonction $f(x_{k+1})$ définie dans $\mathbb{R}$, ce développement limité s'écrit :

$$f(x_{k+1}) = f(x_k + h) \simeq f(x_k) + hf'(x_k) + \frac{1}{2}h^2 f''(x_k), \quad (2.1)$$

où f' et f'' sont respectivement les dérivées première et seconde de f . La minimisation de

l'équation 2.1 donne :

$$\begin{aligned}
 0 &= \frac{d}{dh} f(x_k + h) \\
 &= \frac{d}{dh} \left(f(x_k) + hf'(x_k) + \frac{1}{2}h^2 f''(x_k) \right) \\
 &= f'(x_k) + hf''(x_k)
 \end{aligned} \tag{2.2}$$

d'où :

$$h = -\frac{f'(x_k)}{f''(x_k)}. \tag{2.3}$$

Pour une fonction $f(\mathbf{r})$ définie dans $\mathbb{R}^3$, le pas de déplacement est un vecteur $\mathbf{h}$, la dérivée première f' devient le gradient $\nabla f(\mathbf{r})$ et la dérivée seconde f'' correspond à la matrice hessienne $H_f(\mathbf{r})$. La direction du point critique $\mathbf{r}_c$ est donc donnée par :

$$\mathbf{r}_{k+1} = \mathbf{r}_k - [H_f(\mathbf{r}_k)]^{-1} \nabla f(\mathbf{r}_k). \tag{2.4}$$

L'algorithme est stoppé lorsque la norme du gradient $\nabla f(\mathbf{r}_k)$ devient très proche de zéro. Cependant, la convergence n'est possible qu'à condition que le point de départ $\mathbf{r}_0$ soit déjà suffisamment proche d'un point critique réel.

Recherche des points critiques d'un champ scalaire moléculaire

Dans le cas d'un champ scalaire moléculaire, tel que la densité électronique $\rho(\mathbf{r})$ ou le potentiel électrostatique $V(\mathbf{r})$, plusieurs méthodes pour définir les points de départ $\mathbf{r}_0$ à partir de considérations moléculaires ont déjà été proposées [Leboeuf *et al.*, 1999, Balanarayan et Gadre, 2003, Shirsat *et al.*, 1992, Malcolm et Popelier, 2003]. Dans le cadre des développements réalisés pour ma thèse, j'ai choisi de reprendre la méthode à la fois élégante et efficace de P. Balanarayan et S. R. Gadre [Balanarayan et Gadre, 2003] pour la détermination des points critiques de $V(\mathbf{r})$. Cette méthode repose sur le principe suivant [Balanarayan et Gadre, 2003] : « la présence d'un point critique doit être ressentie lors de l'évaluation du champ scalaire moléculaire sur une surface fermée entourant les atomes de la molécule ». En pratique, cela signifie que si un point critique de $V(\mathbf{r})$ se situe à proximité d'un atome, alors l'effet de sa présence doit se manifester dans les valeurs prises par $V(\mathbf{r})$ autour de cet atome. Par exemple, dans le cas de la molécule d'eau (figure 2.2), l'existence des maxima locaux $(3, -3)$ localisés sur les noyaux d'hydrogène et des minima locaux $(3, +3)$ correspondant aux doublets non-liants de l'oxygène est détectable en sondant les valeurs de $V(\mathbf{r})$ sur la surface entourant l'atome d'oxygène.

Pour illustrer l'algorithme de recherche des points critiques, la figure 2.3 présente graphiquement les différentes étapes qui sont suivies. La première étape consiste à définir autour de chaque atome une surface sphérique centrée sur le noyau et dont le rayon est égal au rayon de covalence ("covalent radius") de l'espèce chimique. Sur chacune de ces sphères, une grille 2D en coordonnées sphériques (θ, φ) est construite, correspondant aux points verts formant des sphères autour des noyaux dans la figure 2.3a. Les extrema (maxima et minima) locaux en 2D du champ scalaire sont recherchés sur ces grilles sphériques en analysant les premiers voisins de chaque point (θ, φ) , un maximum 2D local portant une valeur du champ scalaire plus haute

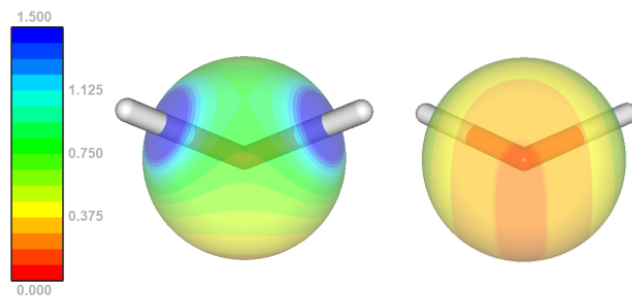


FIGURE 2.2 – Potentiel électrostatique $V(\mathbf{r})$ autour de l'atome d'oxygène de la molécule d'eau.

La sphère centrée sur le noyau de l'atome d'oxygène et de rayon égal au rayon de covalence de l'oxygène ($0,68\text{\AA}$) est colorée par la valeur du potentiel électrostatique $V(\mathbf{r})$ (donnée par l'échelle à gauche en $\text{e.\AA}^{-1}$) généré par la molécule d'eau. Du premier point de vue de la molécule d'eau où l'atome d'oxygène est derrière les atomes d'hydrogène (à gauche), la présence des maxima locaux $(3, -3)$ associés aux noyaux des atomes d'hydrogène à proximité est signalée par des concentrations de valeurs hautes de $V(\mathbf{r})$ sur la surface (de couleur bleue). De même, du second point de vue où l'oxygène est devant les hydrogènes (à droite), la présence des minima locaux $(3, +3)$ correspondant aux doublets non-liants de l'oxygène à proximité est caractérisée par des concentrations de valeurs faibles de $V(\mathbf{r})$ sur la surface (de couleur rouge).

que tous ses premiers voisins, et un minimum 2D local, une valeur plus faible. Dans la figure 2.3b, ces maxima et minima locaux 2D sont représentés par des points bleus et des points rouges respectivement.

Ensuite, pour chaque sphère atomique, des rayons partant du centre et passant par chacun des extrema locaux 2D sont construits. Leur longueur est égale à quatre fois le rayon de covalence pour assurer une couverture de l'espace moléculaire suffisante. Ces rayons sont dessinés de couleur bleue pour ceux passant par les maxima locaux 2D et de couleur rouge pour ceux passant par les minima locaux 2D dans la figure 2.3c. Une recherche des maxima et minima en 1D du champ scalaire sur ces directions radiales est ensuite effectuée, à l'intérieur et à l'extérieur des sphères, qui sont représentés par des points respectivement bleus et rouges dans la figure 2.3d. Les extrema radiaux ainsi obtenus indiquent la présence à proximité de points critiques 3D. Ils sont donc utilisés pour définir les points de départ $\mathbf{r}_0$ de l'algorithme de minimisation de Newton-Raphson présenté précédemment.

Pour finir, afin de s'assurer que tous les points critiques ont été répertoriés, la recherche de Newton-Raphson est également appliquée aux points médians entre chaque paire de points critiques déjà trouvés, jusqu'à ce que plus aucun nouveau point critique ne soit découvert.

Grâce à l'implémentation de ces algorithmes dans le logiciel MoProViewer, qui sera présentée dans la section 2.4.2, les positions des points critiques du potentiel électrostatique peuvent être déterminées avec précision dans les complexes protéine-ligand. En particulier, la localisation des sites électrophiles (maxima locaux $(3, -3)$) et des sites nucléophiles (minima locaux $(3, +3)$) permet d'identifier les contributeurs aux interactions électrostatiques qui se manifestent par la présence de lignes de champ électrique les reliant. Les points-selles apparaissent quant à eux à l'interface entre les surfaces de flux nul de $V(\mathbf{r})$.

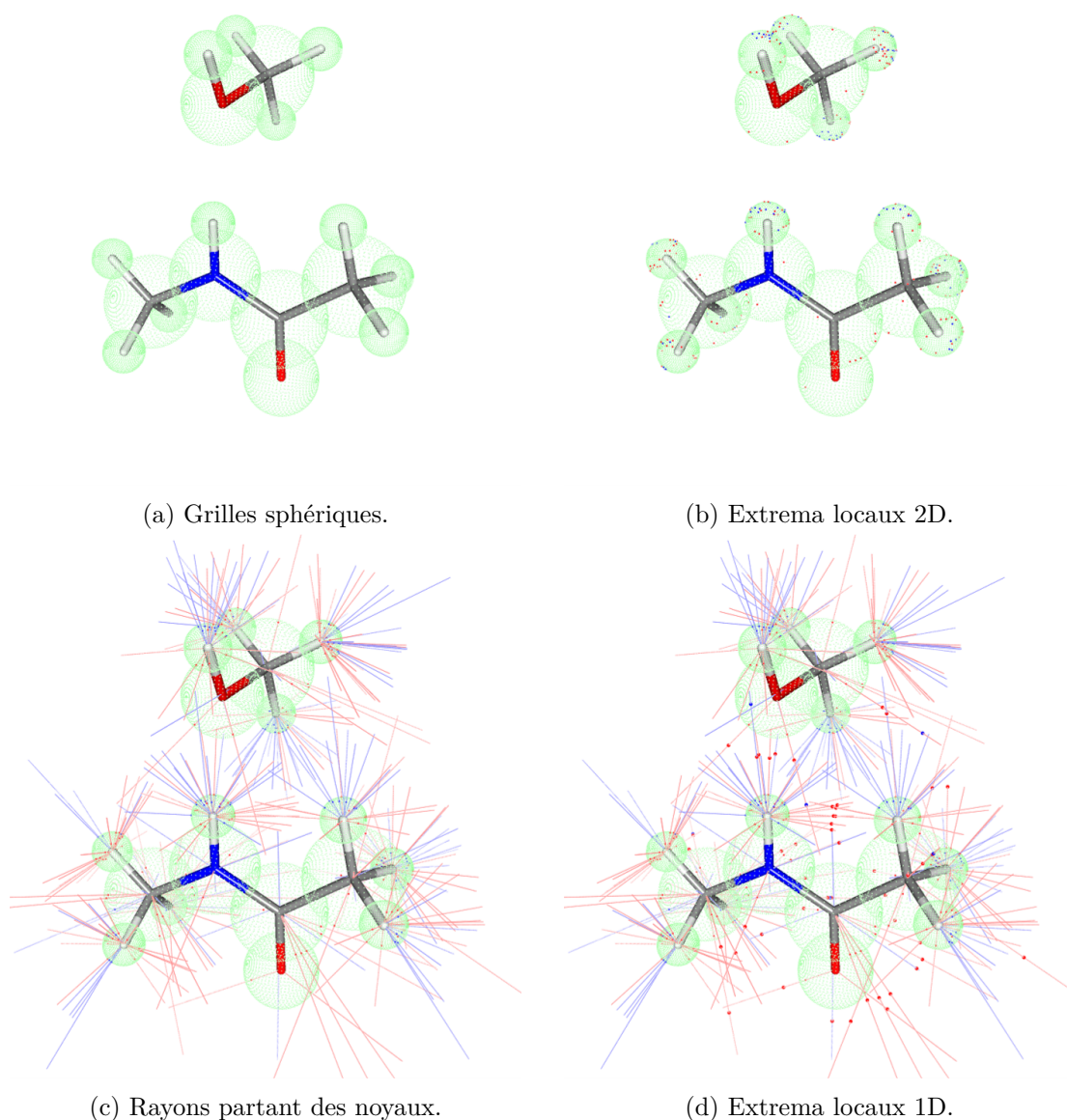


FIGURE 2.3 – Etapes de détermination des points critiques d'un champ scalaire moléculaire.

La méthode de détermination des points critiques du potentiel électrostatique $V(\mathbf{r})$ se déroule en 5 étapes. La première étape (a) consiste à construire des grilles sphériques 2D, dont les points (points verts) sont définis en coordonnées sphériques (θ, φ) autour de chaque atome, centrés sur les noyaux et de rayons égaux à leurs rayons de covalence. Ensuite, (b) les maxima (points bleus) et minima (points rouges) locaux 2D sont recherchés sur ces surfaces. Pour chaque atome, (c) des rayons partant du noyau et passant par chaque maximum local 2D (rayons bleus) et par chaque minimum local 2D (rayons rouges) sont construits. Puis, (d) les maxima (points bleus) et minima (points rouges) locaux 1D sont recherchés sur chaque direction radiale partant des noyaux. Enfin, les points critiques de $V(\mathbf{r})$, montrés sur la figure 2.1, sont obtenus en appliquant l'algorithme de minimisation de Newton-Raphson en utilisant ces extrema locaux 1D comme positions initiale $\mathbf{r}_0$.

2.2 Topographie des lignes de champ électrique et faisceaux primaires

2.2.1 Lignes de champ électrique

En plus de la détermination des points critiques, l'analyse topologique du potentiel électrostatique moléculaire $V(\mathbf{r})$ repose également sur la caractérisation de son gradient $\nabla V(\mathbf{r})$. Contrairement au gradient de $\rho(\mathbf{r})$, le gradient de $V(\mathbf{r})$ correspond à une observable physique : le champ électrique moléculaire $\mathbf{E}(\mathbf{r}) = -\nabla V(\mathbf{r})$. De plus, le champ électrique est directement relié aux forces électrostatiques, avec par exemple $\mathbf{F}(\mathbf{r}) = q\mathbf{E}(\mathbf{r})$, la force électrostatique ressentie par une charge-test q dans le champ électrique $\mathbf{E}(\mathbf{r})$ généré par une molécule. Enfin, le champ électrique est responsable des effets de polarisation via la relation $\boldsymbol{\mu}(\mathbf{r}) = \boldsymbol{\alpha}(\mathbf{r}) \cdot \mathbf{E}(\mathbf{r})$, ou $\boldsymbol{\alpha}(\mathbf{r})$ et $\boldsymbol{\mu}(\mathbf{r})$ sont les polarisabilités et moments dipolaires induits en $\mathbf{r}$ par $\mathbf{E}(\mathbf{r})$.

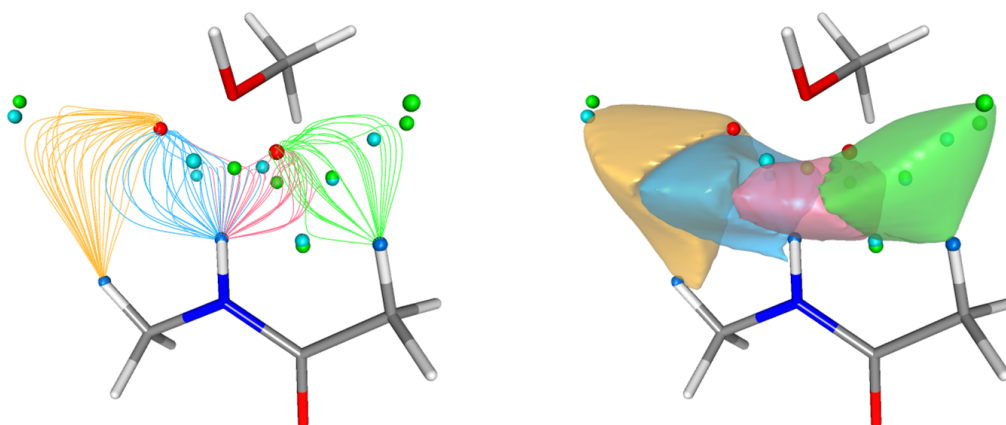
Aussi, l'étude des lignes de champ électrique est pertinente pour étudier les interactions dans les complexes moléculaires. Si le système est neutre, toutes les lignes de champ commencent sur les sites électrophiles (points critiques $(3, -3)$ de $V(\mathbf{r})$) et se terminent sur les sites nucléophiles (points critiques $(3, +3)$ de $V(\mathbf{r})$). Si le système est pseudo-isolé et porte une charge globale non-nulle, certaines lignes de champ possèdent une extrémité partant à l'infini. Les influences électrostatiques des contributeurs générant ce type de lignes de champ s'étendent donc à travers l'espace intermoléculaire et peuvent conduire l'approche d'un autre groupement chargé [Pathak et Gadre, 1990, Gadre et Shrivastava, 1991, Bouhmaida *et al.*, 2002, Kumar et Gadre, 2016].

2.2.2 Faisceaux primaires et surfaces de flux nul

Définition

Les lignes de champ électrique se regroupent en faisceaux et réalisent une partition de l'espace en surfaces de flux nul du potentiel électrostatique⁴ S_V telles que $\mathbf{E}(\mathbf{r}) \cdot \mathbf{n}(\mathbf{r}) = 0, \forall \mathbf{r} \in S_V$ et où $\mathbf{n}$ est le vecteur normal à la surface S_V . Un faisceau contenant les lignes de champ émergeant d'un même site électrophile et se terminant sur un même site nucléophile est enfermé dans un volume fini et est appelé faisceau primaire fermé (ou "closed primary bundle"). Par exemple, quelques faisceaux primaires fermés dans l'espace intermoléculaire du complexe N-méthylacétamide - méthanol sont représentés sur la figure 2.4. Ces faisceaux primaires commencent sur les sites électrophiles correspondant à quatre protons du N-méthylacétamide et se terminent sur les deux doublets non-liants de l'oxygène du méthanol. Dans les systèmes pseudo-isolés, les lignes de champ qui partent à l'infini peuvent également se regrouper sous la forme d'un faisceau si leur extrémité finie se trouve sur un même point critique. Ces faisceaux sont contenus dans un volume infini, on parle alors de faisceau primaire ouvert (ou "open primary bundle") [Mata *et al.*, 2007]. La partition de l'espace en bassins topologiques du potentiel, qui sont les volumes contenant les faisceaux primaires de lignes de champ électrique, diffère de celle obtenue à partir des bassins atomiques de la topologie de la densité $\rho(\mathbf{r})$. En effet, cette partition ne permet pas d'associer

4. Le terme surface de flux nul du potentiel électrostatique fait référence à l'expression "zero flux surface of the electrostatic potential" qui est couramment employée dans la littérature. Il désigne par abus de langage la surface au travers de laquelle le flux du gradient du potentiel électrostatique (ou du champ électrique) est nul.



(a) Faisceaux primaires de lignes de champ. (b) Surfaces entourant les faisceaux primaires.

FIGURE 2.4 – Faisceaux primaires fermés dans l’espace intermoléculaire du complexe N-méthylacétamide - méthanol.

Les faisceaux primaires peuvent être représentés soit par (a) les lignes de champ électrique qu’ils contiennent, soit par (b) les surfaces qui entourent ces faisceaux primaires. Le faisceau primaire entre l’hydrogène du groupement méthyle gauche du N-méthylacétamide et le doublet non-liant gauche de l’oxygène du méthanol est coloré en orange, celui entre l’hydrogène amide et le doublet non-liant gauche en bleu, celui avec le doublet non-liant droit en rose, et celui entre l’hydrogène du groupement méthyle droit et le doublet non-liant droit en vert. Les maxima locaux (sphères bleues), correspondant aux sites électrophiles sur les protons, sont les points de départ des lignes de champ tandis que les minima locaux (sphères rouges), associés aux sites nucléophiles sur les doublets non-liants, sont leurs points d’arrivée. Les points-selles, représentés par des sphères vertes pour les types $(3, -1)$ et cyan pour les $(3, +1)$, sont présents sur les interfaces entre les surfaces de flux nul de potentiel. La représentation des faisceaux primaires par les lignes de champ électrique qu’ils contiennent a l’avantage d’être intuitive car elle est liée à la notion de forces électrostatiques mais les limites spatiales exactes des faisceaux primaires ne sont pas visibles. Grâce à la représentation en surfaces contenant les faisceaux primaires, les limitations de chaque bassin topologique de $V(\mathbf{r})$, ainsi que la partition de l’espace qu’ils réalisent, apparaissent clairement.

les volumes topologiques de $V(\mathbf{r})$ aux atomes, comme c’est le cas pour $\rho(\mathbf{r})$, mais à des paires « site électrophile - site nucléophile ».

Représentation des faisceaux primaires

Les faisceaux primaires sont généralement représentés par les lignes de champ qu’ils contiennent [Mata *et al.*, 2007, Kumar et Gadre, 2016], comme dans la figure 2.4a. Le champ électrique étant directement associé aux forces électrostatiques, cette représentation en lignes de champ permet de visualiser les chemins que pourraient emprunter une charge-test soumise à ces forces. Cependant, chaque faisceau contenant une infinité de lignes de champ, elles ne peuvent pas toutes être représentées et les délimitations spatiales exactes des bassins topologiques de $V(\mathbf{r})$ ne sont pas visibles dans cette représentation. C’est pourquoi, j’ai mis au point une nouvelle méthodologie permettant de représenter les surfaces entourant les faisceaux primaires, comme dans la figure 2.4b. Cette représentation fait ressortir la partition spatiale en bassins topologiques du potentiel, de façon à ce que chaque point de l’espace moléculaire puisse être associé à un unique faisceau primaire.

2.2.3 Algorithme de détermination des surfaces entourant les faisceaux primaires

Dans le but de mettre en évidence les délimitations de leurs volumes, j'ai développé une nouvelle méthodologie pour obtenir les surfaces entourant les faisceaux primaires. Pour les faisceaux primaires fermés, je suis partie de la définition suivante : un faisceau primaire fermé est constitué des lignes de champ dont une extrémité est localisée sur un unique maximum local et l'autre extrémité est localisée sur un unique minimum local. En sondant l'espace moléculaire, les points appartenant au volume du faisceau primaire peuvent être discriminés en testant ces conditions aux extrémités des lignes de champ. Les différentes étapes de cette méthode sont illustrées dans le cas du faisceau primaire entre le site électrophile du proton amide du N-méthylacétamide et du site nucléophile de l'un des doublets non-liants de l'oxygène du méthanol sur la figure 2.5.

Pour générer le faisceau primaire entre un maximum local et un minimum local, la première étape est d'estimer le volume de la boîte rectangulaire contenant ce faisceau primaire. Pour cela, un ensemble de lignes de champ est généré à partir du maximum local, et celles rejoignant le minimum local considéré sont retenues (voir figure 2.5a). Les coordonnées de tous les points par lesquels ces lignes de champ passent sont analysées pour définir les coordonnées des coins de la boîte $(x_{\min}, y_{\min}, z_{\min})$ et $(x_{\max}, y_{\max}, z_{\max})$. Comme montré dans la figure 2.5b, une grille 3D orthogonale et régulière est définie à l'intérieur de cette boîte.

Ensuite, pour chaque point (x, y, z) de la grille, un nouveau champ scalaire binaire $\eta(x, y, z)$ est calculé. Cette fonction implicite prend la valeur 1 si le point (x, y, z) est à l'intérieur du volume $\mathcal{V}$ du faisceau primaire et la valeur 0 sinon :

$$\eta(x, y, z) = \begin{cases} 1 & \text{si } (x, y, z) \in \mathcal{V} \\ 0 & \text{si } (x, y, z) \notin \mathcal{V} \end{cases} \quad (2.5)$$

Pour tester si le point (x, y, z) se trouve à l'intérieur de $\mathcal{V}$ ou non, les positions des extrémités de la ligne de champ électrique passant par ce point sont déterminées en suivant la direction du vecteur champ électrique $\mathbf{E}(\mathbf{r})$, puis la direction inverse, c'est-à-dire celle du vecteur gradient du potentiel $\nabla V(\mathbf{r})$. Si les extrémités remplissent les conditions définissant le faisceau primaire, c'est-à-dire si elles correspondent aux positions des points critiques considérés, alors $\eta(x, y, z) = 1$, et $\eta(x, y, z) = 0$ sinon. La figure 2.5c montre les points pour lesquels $\eta(x, y, z) = 1$ dans la grille rectangulaire.

Le champ scalaire $\eta(x, y, z)$ valant 1 à l'intérieur du faisceau primaire et 0 à l'extérieur, l'isosurface $\eta(x, y, z) = 0,99$ doit être très proche de la surface de flux nul de potentiel entourant exactement le volume du faisceau primaire. J'ai utilisé l'algorithme des "Marching Cubes" [Lorenson et Cline, 1987] pour reconstruire le maillage triangulaire (voir figure 2.5d) de cette surface et la librairie OpenMesh [Kobbelt *et al.*, 2002] pour afficher la surface 3D autour du faisceau primaire (voir figures 2.5e et 2.5f). L'implémentation de ces méthodes dans le logiciel MoProViewer sera décrite dans la section 2.4.3.

La surface entourant un faisceau primaire ouvert peut être déterminée de façon équivalente en changeant simplement les conditions que les extrémités des lignes de champ calculées doivent respecter. En effet, pour les faisceaux primaires ouverts, seule une extrémité doit être localisée sur un unique extremum local tandis que l'autre extrémité part à l'infini.

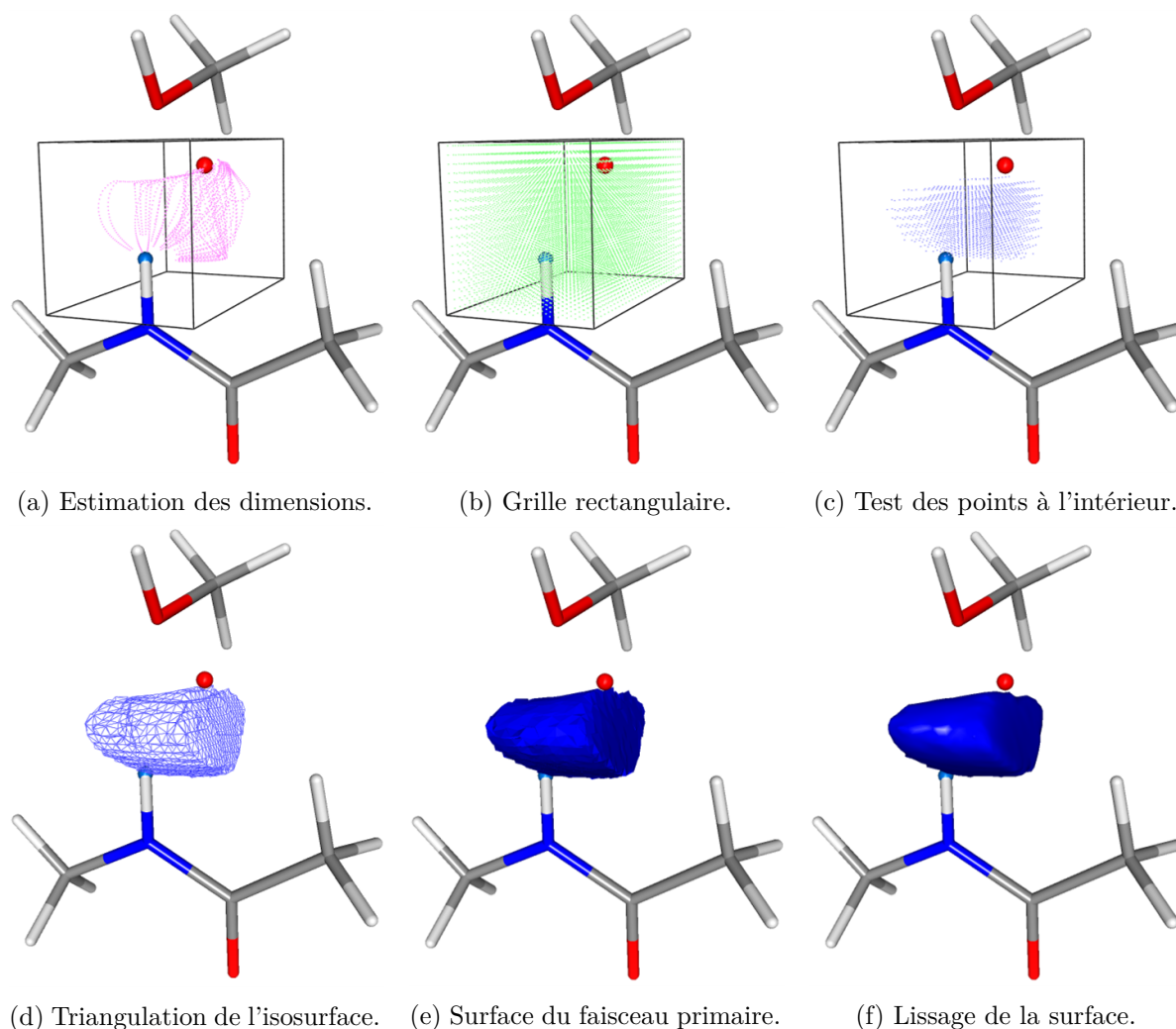


FIGURE 2.5 – Etapes de détermination de la surface entourant un faisceau primaire fermé.

Les différentes étapes de l'algorithme de détermination de la surface entourant un faisceau primaire fermé sont imagées dans le complexe N-méthylacétamide - méthanol, entre le site électrophile (sphère bleue) du proton amide du N-méthylacétamide et le site nucléophile (sphère rouge) d'un des deux doublets non-liants de l'oxygène du méthanol. (a) Un premier faisceau de quelques lignes de champ appartenant au faisceau primaire (lignes roses) est généré pour estimer la taille de la boîte rectangulaire (en noir) l'enfermant. (b) Une grille orthogonale et régulière de points (x, y, z) (points verts) est définie dans cette boîte. (c) Chaque point de la grille est testé pour évaluer le champ scalaire $\eta(x, y, z)$, qui prend la valeur 1 pour les points à l'intérieur du faisceau primaire (points bleus) et la valeur 0 sinon. (d) La triangulation de l'isosurface $\eta(x, y, z) = 0,99$ (triangles bleus) est réalisée à l'aide de l'algorithme des Marching Cubes [Lorensen et Cline, 1987]. (e) La surface entourant le faisceau primaire (surface bleue) est reconstruite à partir du maillage triangulaire grâce à la librairie OpenMesh [Kobbelt *et al.*, 2002]. (f) Pour finir, cette surface est lissée pour améliorer son aspect visuel, également grâce à la librairie OpenMesh [Kobbelt *et al.*, 2002].

Les faisceaux primaires étant associés aux paires de sites électrophiles et nucléophiles, ils peuvent être regroupés soit par sites électrophiles soit par sites nucléophiles. Les unions de faisceaux primaires qui en découlent permettent de définir les zones d'influence électrophile et les zones d'influence nucléophile.

2.3 Partition de l'espace en zones d'influence électrophile et en zones d'influence nucléophile

2.3.1 Définition des zones d'influence électrophile et nucléophile

Comme énoncé précédemment, un faisceau primaire est associé à une paire site électrophile - site nucléophile. Aussi, les faisceaux primaires peuvent être regroupés autour d'un site électrophile commun ou d'un site nucléophile commun, définissant ainsi les zones d'influence électrostatique comme introduites par I. Mata, E. Molins et E. Espinosa [Mata *et al.*, 2007].

Zone d'influence électrophile

L'union des faisceaux primaires contenant les lignes de champ électrique émergeant d'un même site électrophile (point critique $(3, -3)$ du potentiel électrostatique) constitue la zone d'influence de ce site électrophile. La zone d'influence électrophile (ZIE) correspond à la région de l'espace soumise à l'influence électrophile d'un unique maximum $(3, -3)$ de $V(\mathbf{r})$, c'est-à-dire d'un noyau. Par conséquent, une ZIE peut être associée à un unique atome. Par exemple, la figure 2.6a montre les trois faisceaux primaires émergeant du site électrophile du proton du groupement amide du N-méthylacétamide en complexe avec le méthanol, et la figure 2.6b représente la ZIE formée par leur union. A la surface d'une ZIE, le point où le potentiel électrostatique $V(\mathbf{r})$ est localement maximal est caractérisé par un point-selle $(3, -1)$, c'est donc en ce point que la barrière de potentiel pour une attaque nucléophile est la plus basse [Mata *et al.*, 2007].

Zone d'influence nucléophile

De même, l'union des faisceaux primaires contenant les lignes de champ électrique convergeant vers un même site nucléophile (point critique $(3, +3)$ du potentiel électrostatique) forme la zone d'influence de ce site nucléophile. Comme expliqué dans la partie 2.1, un site nucléophile correspond souvent à une paire d'électrons non-liantes. Certains atomes, comme l'atome d'oxygène qui est très présent dans les molécules biologiques, possèdent plusieurs doublets non-liants. J'ai choisi de définir dans mes travaux la zone d'influence nucléophile (ZIN) comme la région de l'espace soumise à l'influence du ou des minimum(minima) $(3, +3)$ de $V(\mathbf{r})$ correspondant au(x) doublet(s) non-liant(s) appartenant à un unique atome. Cette définition à l'avantage de relier les ZIN aux atomes, comme c'est le cas pour les ZIE. D'ailleurs, puisqu'il existe un faisceau primaire de lignes de champ électrique entre les doublets non-liants d'un atome et son propre noyau, ce dernier est également contenu dans le volume de la ZIN. Par exemple, les faisceaux primaires reliés aux deux doublets non-liants de l'atome d'oxygène du groupement hydroxyle du méthanol dans le complexe avec le N-méthylacétamide, montrés dans la figure 2.6c, forment la ZIN de cet atome d'oxygène, représentée par la surface rouge sur la figure 2.6d. Les deux lobes de la ZIN, chacun associé à un doublet non-liant, sont d'ailleurs bien visibles. Le point où le potentiel électrostatique $V(\mathbf{r})$ est localement minimal à la surface d'une ZIN est marqué par la présence d'un point-selle $(3, +1)$, et constitue le point d'entrée le plus favorable pour une attaque électrophile [Mata *et al.*, 2007].

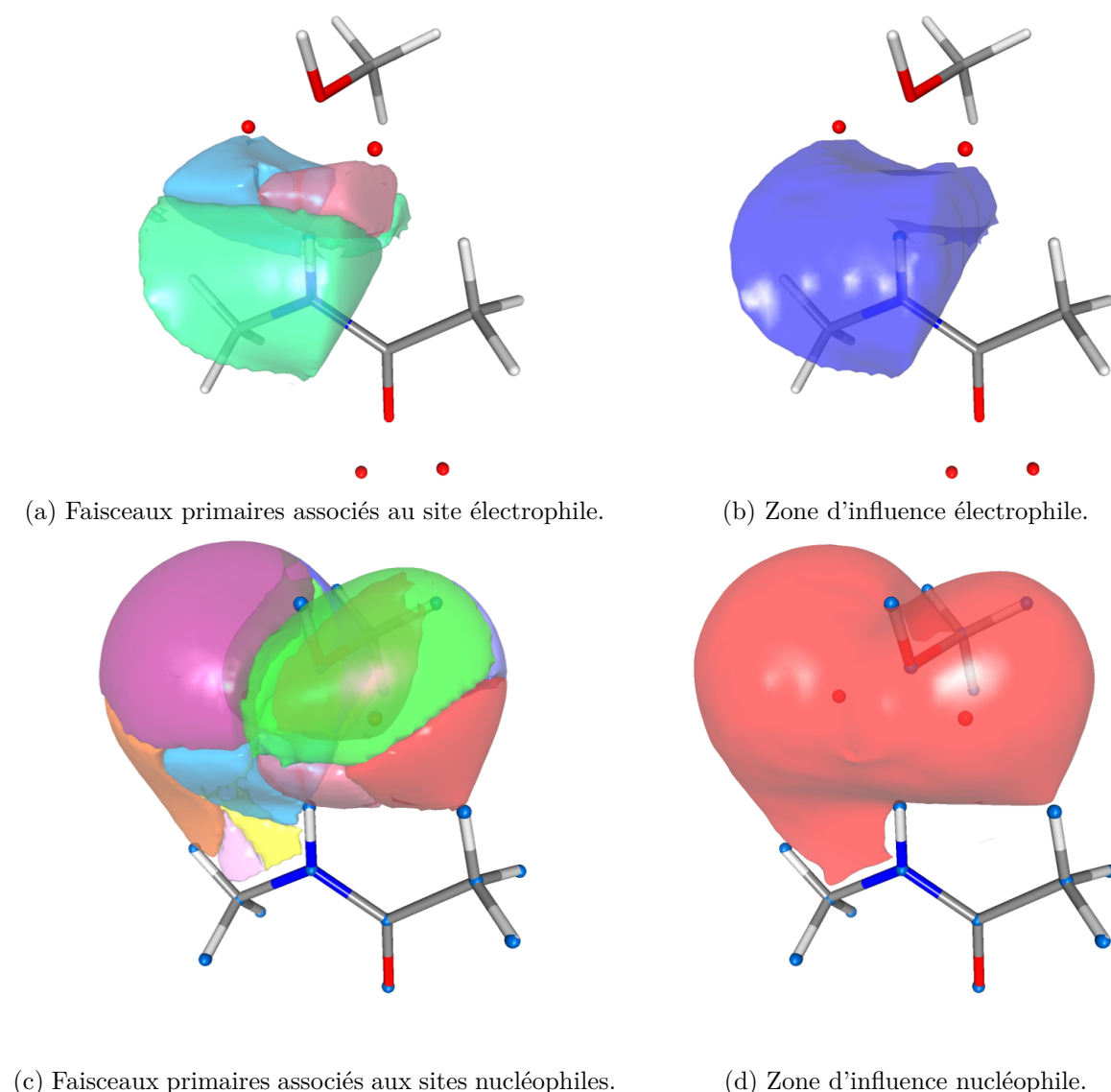


FIGURE 2.6 – Définition de la zone d'influence d'un site électrophile et d'un site nucléophile.

Dans le complexe N-méthylacétamide - méthanol, (a) les faisceaux primaires et (b) la zone d'influence qui sont associés au site électrophile localisé sur le proton du groupement amide du N-méthylacétamide, ainsi que (c) les faisceaux primaires et (d) la zone d'influence qui sont associés aux deux sites nucléophiles correspondant aux doublets non-liants de l'atome d'oxygène du méthanol, sont représentés. La zone d'influence électrophile (surface bleu foncé dans (b)), associée à l'atome d'hydrogène amide, est la région de l'espace soumise à l'influence électrophile de ce proton uniquement. De même, la zone d'influence nucléophile (surface rouge dans (d)), associée aux deux paires d'électrons non-liantes de l'atome d'oxygène hydroxyle, est la région de l'espace soumise à l'influence nucléophile de cet atome d'oxygène uniquement.

Algorithme de détermination des surfaces entourant les zones d'influence

D'un point de vue algorithmique, il aurait été envisageable de construire les zones d'influence en partant de la détermination des faisceaux primaires puis de les fusionner. Cependant, comme le montre la figure 2.6c, il peut arriver qu'une zone d'influence soit composée d'un grand nombre de faisceaux primaires. Cette approche serait donc trop coûteuse en temps de calcul. La méthode

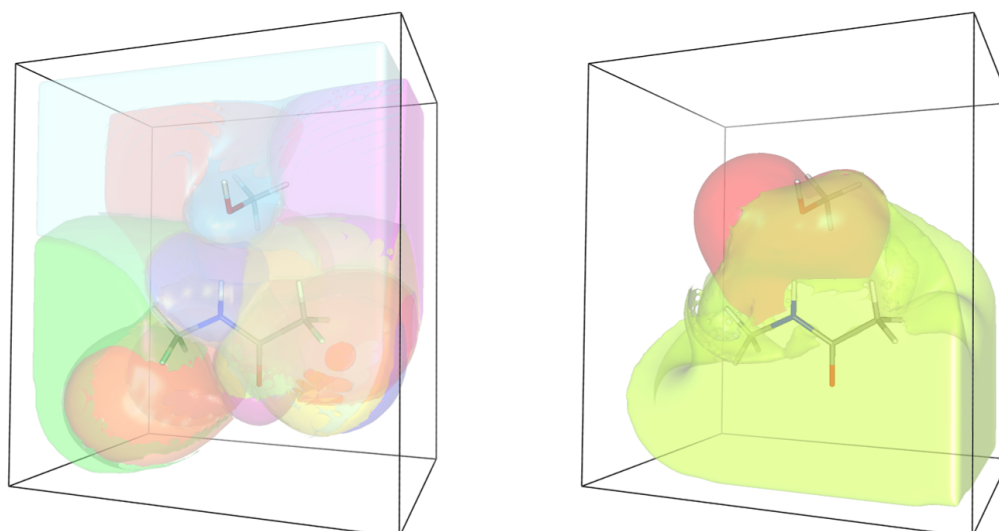
que j’ai proposée se base sur une détermination directe de la zone d’influence. En effet, comme pour les faisceaux primaires, la définition des ZIE et des ZIN repose sur les conditions remplies par les extrémités des lignes de champ contenues dans leurs volumes. Par conséquent, l’algorithme défini dans la section 2.2.3 pour les faisceaux primaires peut également être appliqué aux zones d’influence. La seule étape qui doit être adaptée est celle du calcul de la fonction implicite $\eta(x, y, z)$. Pour les faisceaux primaires, cette fonction $\eta(x, y, z)$ valait 1 si la ligne de champ passant par le point (x, y, z) avait bien pour extrémités la paire site électrophile - site nucléophile considérée, et valait 0 sinon. Pour une ZIE, il suffit que la ligne de champ passant par (x, y, z) débute sur le site électrophile associé au noyau de l’atome considéré pour avoir $\eta(x, y, z) = 1$, et $\eta(x, y, z) = 0$ sinon. De même, pour une ZIN, si la ligne de champ passant par (x, y, z) se termine sur un site nucléophile associé au(x) doublet(s) non-liant(s) de l’atome considéré alors $\eta(x, y, z) = 1$, et $\eta(x, y, z) = 0$ sinon. Cette méthode a également été implémentée dans le logiciel MoProViewer, comme le détaillera la partie 2.4.3.

Par ailleurs, il est intéressant de noter qu’un faisceau primaire fermé, joignant une paire site électrophile - site nucléophile, appartient à la fois à la ZIE associée au site électrophile et à la ZIN associée au site nucléophile. De même, un faisceau primaire ouvert émergeant d’un site électrophile appartient à la ZIE associée même si les lignes de champ partent vers l’infini, et un faisceau primaire ouvert convergeant vers un site nucléophile appartient à la ZIN correspondante, même si les lignes de champ proviennent de l’infini. De plus, les bords de ces zones d’influence étant communs avec les bords des faisceaux primaires, leurs volumes sont donc également enfermés dans des surfaces de flux nul de potentiel S_V .

2.3.2 Partitions de l’espace moléculaire

La définition des zones d’influence permet de définir deux partitions de l’espace distinctes et concomitantes : la partition en ZIE et la partition en ZIN. Les zones d’influence étant associées aux atomes, les partitions de l’espace qu’elles réalisent peuvent s’avérer plus pertinentes que la partition en faisceaux primaires, qui eux sont associés aux paires site électrophile - site nucléophiles. La figure 2.7a illustre la partition de l’espace en ZIE dans le complexe N-méthylacétamide - méthanol. Une ZIE est associée à chacun des maxima de $V(\mathbf{r})$, représentés sur la figure 2.1b, c’est-à-dire à chaque noyau. Grâce à cette partition, l’atome portant l’influence électrophile à laquelle une charge-test q à la position $\mathbf{r}_q$ est soumise peut être identifié facilement. En effet, cette charge-test sera attirée vers (si $q < 0$) ou repoussée par (si $q > 0$) l’atome dont la ZIE contient la position $\mathbf{r}_q$. Dans cette figure, toutes les ZIE ont volontairement été représentées pour mettre en évidence la complétude de la partition de l’espace obtenue, en dépit de la clarté de l’image. En pratique, la représentation de seulement quelques ZIE suffit pour étudier une région particulière de l’espace, comme par exemple un site de fixation dans la cavité d’une protéine.

La figure 2.7b illustre quant à elle la partition en ZIN dans le complexe N-méthylacétamide - méthanol. Dans ce complexe, seuls six sites nucléophiles ont été identifiés (voir figure 2.1c). Parmi ceux-ci, deux sont associés aux doublets non-liants de l’oxygène du méthanol (Cp50 et Cp52 de la figure 2.7b) et forment la ZIN représentée par une surface rouge dans la figure 2.7b, et deux autres sont associés aux doublets non-liants de l’oxygène du N-méthylacétamide (Cp51 et Cp76 de la figure 2.7b) et sont associés à la ZIN représentée par une surface jaune. En revanche, le



(a) Partition de l'espace en zones d'influence électrophile. (b) Partition de l'espace en zones d'influence nucléophile.

FIGURE 2.7 – Partition de l'espace moléculaire du complexe N-méthylacétamide - méthanol en zones d'influence électrophile et en zones d'influence nucléophile.

L'espace moléculaire peut être partitionné soit (a) en zones d'influence électrophile (ZIE), soit (b) en zones d'influence nucléophile (ZIN). Pour la partition en ZIE (a), les surfaces des zones d'influence des sites électrophiles associés à chacun des 18 noyaux du complexe sont représentées, avec une couleur différente pour chaque site électrophile. Pour la partition en ZIN (b), les surfaces des zones d'influence des sites nucléophiles correspondant aux doublets non-liants des atomes d'oxygène du groupement carbonyle du N-méthylacétamide et du groupement hydroxyle du méthanol sont représentées respectivement en jaune et en rouge. En pratique, le potentiel électrostatique est calculé dans une boîte de taille finie (boîte noire), c'est pourquoi les zones d'influence sont tronquées aux bords de cette boîte. De plus, le potentiel électrostatique n'étant jamais exactement nul à longue distance pour un système pseudo-isolé, j'ai introduit une valeur de coupure du champ électrique pour ramener sa norme à 0 lorsqu'elle est en-dessous d'une faible valeur $\varepsilon = 0,001 \text{ e.Å}^{-2}$. C'est pourquoi la boîte n'est pas entièrement remplie bien que les ZIE et les ZIN réalisent des partitions complètes de l'espace.

Cp53 supposé associé au doublet non-liant de l'atome d'azote du N-méthylacétamide, ne permet pas de former une ZIN car il n'est pas suffisamment électronégatif ($V(\mathbf{r}_{\text{Cp53}}) = -0,073 \text{ e.Å}^{-1}$) pour faire converger vers lui les lignes de champ, celles-ci allant plutôt converger vers le doublet non-liant Cp76 de l'oxygène se trouvant à proximité ($V(\mathbf{r}_{\text{Cp76}}) = -0,206 \text{ e.Å}^{-1}$). Quant au Cp75, il n'est pas associé à un atome et ne rentre donc pas dans la définition de ZIN donnée précédemment. Cependant, ces deux minima peuvent tout de même être étudiés du point de vue des faisceaux primaires. D'une manière générale, la partition de l'espace en ZIN permet d'identifier l'atome portant les doublets non-liants ayant une influence nucléophile à laquelle sera soumise une charge-test q à la position $\mathbf{r}_q$. Cette charge-test sera attirée vers (si $q > 0$) ou repoussée par (si $q < 0$) l'atome associé à la ZIN contenant la position $\mathbf{r}_q$. Il est intéressant de noter que puisque les ZIN ne peuvent être associées qu'à des atomes portant des doublets non-liants ou groupements chargés négativement, celles-ci sont moins nombreuses que les ZIE et plus volumineuses pour pouvoir remplir tout l'espace.

2.3.3 Surface de flux nul du potentiel et vérification du théorème de Gauss

Théorème de Gauss appliqué aux surfaces de flux nul de potentiel électrostatique

D'après le théorème de Gauss, le flux du champ électrique $\mathbf{E}$ à travers une surface fermée S est proportionnel à la charge électrique totale Q_{int} contenue à l'intérieur du volume $\mathcal{V}$ délimité par cette surface :

$$\oint_S \mathbf{E} \cdot \mathbf{n} = \frac{Q_{\text{int}}}{\varepsilon_0}, \quad (2.6)$$

où ε_0 est la permittivité diélectrique du vide, et $\mathbf{n}$ est le vecteur normal à la surface S . Une surface de flux nul du potentiel électrostatique S_V est définie telle que : $\mathbf{E}(\mathbf{r}) \cdot \mathbf{n}_V(\mathbf{r}) = 0$, $\forall \mathbf{r} \in S_V$ et où $\mathbf{n}_V$ est le vecteur normal à la surface S_V . L'application du théorème de Gauss à S_V donne :

$$\oint_{S_V} \mathbf{E} \cdot \mathbf{n}_V = \frac{Q_{\text{int}}}{\varepsilon_0} = 0. \quad (2.7)$$

Par conséquent, la charge totale Q_{int} à l'intérieur d'une surface de flux nul de $V(\mathbf{r})$ est nulle. Les faisceaux primaires ainsi que les zones d'influence étant entourés par des surfaces de flux nul, la charge totale à l'intérieur de leurs volumes est également nulle. La charge totale Q_{int} est la somme des charges nucléaires et des charges des électrons, ces dernières étant obtenues par intégration volumique de la densité électronique. Les faisceaux primaires étant associés aux paires site électrophile - site nucléophile, ils ne contiennent pas un nombre de noyaux connu a priori. Les ZIN recouvrent quant à elles de larges volumes et peuvent englober plusieurs noyaux. Par contre, les ZIE, par définition associées à un unique site électrophile chacune, ne contiennent quant à elle que la charge nucléaire du noyau de l'atome i correspondant. L'application du théorème de Gauss sur une ZIE donne donc : $Q_{\text{int}} = Z_i + Q_{\text{elec}} = 0$, avec Z_i , la charge nucléaire de l'atome i et Q_{elec} , la charge obtenue par intégration de la densité électronique $\rho(\mathbf{r})$ dans la ZIE :

$$Q_{\text{elec}} = -e \int_{\text{ZIE}} \rho(\mathbf{r}) d^3\mathbf{r}, \quad (2.8)$$

où e est la charge élémentaire. En pratique, pour calculer numériquement cette intégrale, j'ai utilisé la grille cartésienne orthogonale et régulière définie pour la construction de la ZIE, de la même manière que dans l'algorithme de détermination des faisceaux primaires (voir figures 2.5b et 2.5c), et j'ai sommé la valeur de la densité électronique $\rho(x, y, z)$ sur chaque point de la grille (x, y, z) à l'intérieur de la ZIE ($\eta(x, y, z) = 1$), multipliée par l'élément de volume $d^3\mathbf{r}$ qui est simplement le cube de la valeur du pas de la grille.

Vérification dans un système-test

Pour tester la validité des volumes définis dans mes travaux, j'ai vérifié le théorème de Gauss sur les ZIE dans le complexe N-méthylacétamide - méthanol. Le tableau 2.1 rassemble les valeurs de Q_{elec} et de Q_{int} calculées dans les ZIE associées à chaque atome du complexe. Globalement, les charges totales résiduelles Q_{int} obtenues sont de l'ordre de $\pm 0,10e$. Les valeurs de Q_{int} les plus élevées en valeur absolue sont obtenues pour l'atome d'oxygène O1 du méthanol avec $Q_{\text{int}} = +0,33e$ et pour les atomes de carbone C11 et C15 du N-méthylacétamide avec respectivement $Q_{\text{int}} = -0,21e$ et $Q_{\text{int}} = +0,17e$. Pour l'ensemble de la molécule, la somme des

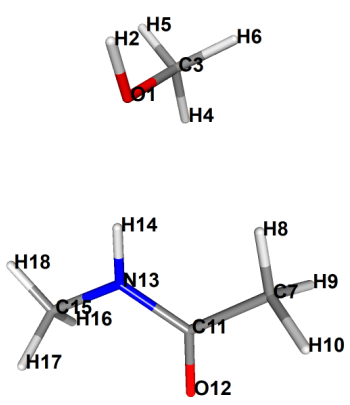
Labels des atomes	Atome	Q_{elec}	Q_{int}	ER
	O1	-7,67	+0,33	4,1%
	H2	-1,10	-0,10	10,0%
	C3	-5,90	+0,10	1,7%
	H4	-1,07	-0,07	7,0%
	H5	-1,04	-0,04	4,0%
	H6	-1,04	-0,04	4,0%
	C7	-5,92	+0,08	1,3%
	H8	-1,08	-0,08	8,0%
	H9	-1,07	-0,07	7,0%
	H10	-1,07	-0,07	7,0%
	C11	-6,21	-0,21	3,5%
	O12	-7,92	+0,08	1,0%
	N13	-6,88	-0,12	1,7%
	H14	-1,09	-0,09	9,0%
	C15	-5,83	+0,17	2,8%
	H16	-1,07	-0,07	7,0%
	H17	-1,08	-0,08	8,0%
	H18	-1,08	-0,08	8,0%
	Total :	-58,12	-0,12	0,2%

TABLEAU 2.1 – Vérification du théorème de Gauss dans les zones d'influence électrophile.

La densité électronique a été intégrée dans les ZIE associées à chaque atome du complexe N-méthylacétamide - méthanol. La première colonne « Atome » liste les labels, la première lettre étant le symbole de l'espèce chimique, correspondant aux différents atomes comme indiqué sur la figure de gauche. La deuxième colonne indique les valeurs de la densité intégrée dans la ZIE, Q_{elec} , donnée en unité atomique $e = 1,60 \cdot 10^{-19} \text{C}$. La troisième colonne fournit la charge totale à l'intérieur de la ZIE : $Q_{\text{int}} = Z_i - Q_{\text{elec}}$, avec Z_i , la charge du noyau de l'espèce chimique i ($Z_H = 1e$, $Z_C = 6e$, $Z_N = 7e$ et $Z_O = 8e$), également donnée en unité atomique e . Les sommes de Q_{elec} et Q_{int} sur l'ensemble du complexe sont indiquées dans la dernière ligne du tableau. La dernière colonne présente le calcul de l'erreur relative $\text{ER} = |(Q_{\text{elec}} - Q_{\text{elec, th}})/Q_{\text{elec, th}}|$, où $Q_{\text{elec, th}}$ est la valeur théorique de la charge électronique, correspondant à la charge nucléaire au signe près. Ces valeurs ont été obtenues en utilisant les paramètres par défaut de la génération des zones d'influence (voir section 2.4.3) et en utilisant un grand volume de boîte pour le calcul du potentiel électrostatique $V(\mathbf{r})$.

charges totales est seulement de $Q_{\text{int}} = -0,12e$ grâce aux compensations entre charges résiduelles de signes différents. L'erreur absolue commise sur Q_{int} est identique à l'erreur absolue sur Q_{elec} , pour laquelle il est également possible de calculer l'erreur relative (ER) par rapport à la valeur théoriquement attendue qui correspond à l'opposé de la charge nucléaire. Les plus fortes erreurs relatives sur Q_{elec} sont obtenues pour les atomes d'hydrogène, pour lesquels la valeur attendue est $Q_{\text{elec}} = 1e$, allant de 10,0% pour l'hydrogène H2 du méthanol à seulement 4,0% pour les hydrogènes H5 et H6. Pour les atomes plus lourds, les erreurs relatives les plus élevées sont obtenues pour l'atome d'oxygène O1 avec $Q_{\text{elec}} = -7,67e$ au lieu de $-8,00e$ attendu, soit $\text{ER} = 4,1\%$, et pour les atomes de carbone C11 et C15 avec respectivement $Q_{\text{elec}} = -6,21e$ et $Q_{\text{elec}} = -5,83e$ au lieu de $-6,00e$ attendu, soit $\text{ER} = 3,5\%$ et $2,8\%$. Pour l'ensemble de la molécule, la somme des charges électroniques calculées par intégration de $\rho(\mathbf{r})$ dans le volume de la ZIE est de $-58,12e$, contre $-58,00e$ attendu. L'erreur relative sur cette grandeur n'est que de 0,2% grâce aux compensations entre charges résiduelles de signes différents.

Pour une intégration de la densité électronique dans les bassins topologiques de $\rho(\mathbf{r})$, c'est-

à-dire dans les bassins atomiques définis dans la théorie QTAIM de R. Bader [Bader, 1990], la déviation standard sur la charge atomique est de l'ordre de 0,02e [Fournier *et al.*, 2018]. Dans l'exemple de la molécule d'acide (E)-5-phenylpent-1-enylboronique, les erreurs relatives sur les charges électroniques atomiques Q_{elec} ⁵ vont de 0,2% pour les atomes d'oxygène à 2,0% pour les atomes d'hydrogène [Fournier *et al.*, 2018]. Les erreurs obtenues par l'intégration de la densité électronique dans les ZIE sont légèrement plus élevées mais d'ordre de grandeur similaire, ce qui permet de valider la méthode de définition des ZIE. Ces erreurs ont trois origines possibles : la valeur de $\rho(\mathbf{r})$ elle-même, la méthode d'intégration (notamment l'échantillonnage du volume) et la définition des bords du volume dans lequel la densité est intégrée. Cette dernière cause est sans doute la plus importante, notamment dans des régions où les faisceaux primaires et les zones d'influence forment des volumes dont les bords sont très étroits et dont les volumes sont difficiles à modéliser avec l'algorithme des Marching Cubes.

2.3.4 Interprétation des descripteurs électrostatiques de partitions en zones d'influence électrophile et nucléophile

Applications possibles pour l'étude des complexes protéine-ligand

Dans le paradigme des zones d'influence électrophile (ZIE) et nucléophile (ZIN), bien que le potentiel électrostatique $V(\mathbf{r})$ employé soit celui généré par l'ensemble du système considéré, les partitions de l'espace qui en découlent permettent d'associer tout point de l'espace moléculaire à un unique contributeur électrophile et à un unique contributeur nucléophile, pouvant être un point critique, un atome ou un groupement d'atomes. Les analyses traditionnelles, uniquement à partir de la connaissance de la structure atomique, se limitent le plus souvent à la caractérisation des interactions interatomiques comme les liaisons hydrogène et les ponts salins, sans permettre de déduire les interactions plus distantes. En revanche, les descripteurs que sont les ZIE et les ZIN révèlent quant à eux l'étendue spatiale de l'influence électrostatique d'un atome (ou groupe d'atomes). Aussi, un contributeur en apparence éloigné mais capable d'étendre son influence sur de longues distance via les lignes de champ électrique pourra être caractérisé par cette approche. Par exemple, un résidu enfoui au fond de la poche de fixation d'une protéine et dont l'influence électrostatique parvient jusqu'à l'entrée de la cavité pour guider l'approche d'un ligand sera identifié grâce à la détermination des zones d'influence des atomes qui le composent.

De plus, contrairement aux représentations classiques de $V(\mathbf{r})$ sur les surfaces d'accessibilité au solvant où les contributeurs électrostatiques sont déterminés de façon individuelle, les ZIE et ZIN mettent en évidence la concomitance entre influences électrophiles et nucléophiles. Cette dualité souligne l'importance de la présence à la fois d'un contributeur électrophile et d'un contributeur nucléophile pour que les forces électrostatiques et les effets de polarisation puissent s'établir selon une direction particulière, les deux étant donc nécessaires pour expliquer les mécanismes dominés par ces effets. Par exemple, l'approche d'un ligand de nature nucléophile est souvent guidée par l'influence d'un résidu électrophile dans le site de fixation, mais il peut

5. Dans l'article [Fournier *et al.*, 2018], ce ne sont pas les valeurs de charges électroniques Q_{elec} qui sont données mais de charge totale Q_{int} (notée Q_{topo} dans la publication). Les erreurs relatives sur les Q_{elec} atomiques peuvent être estimées par le calcul : $ER = SSD / Z$, où SSD ("Sample Standard Deviation") est la déviation standard du nombre d'électrons et Z est la charge nucléaire, qui est égale à l'opposé de la charge électronique théorique.

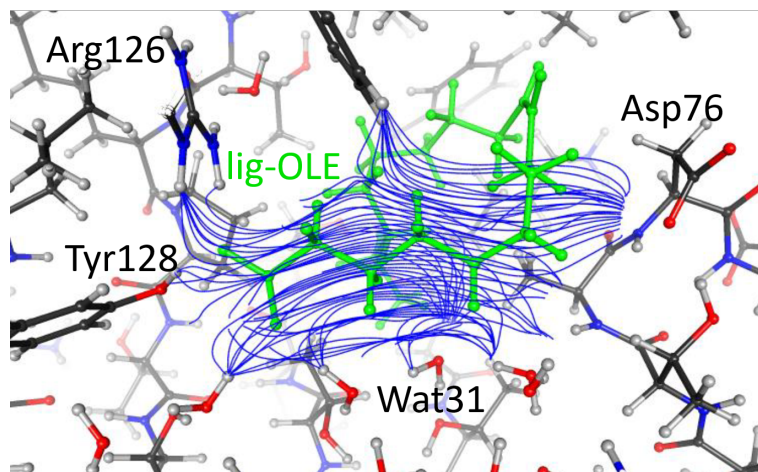


FIGURE 2.8 – Lignes de champ électrique dans la poche de fixation de la FABP.

Cette figure a été réalisée par B. Guillot (laboratoire CRM²) pour sa présentation "The role of water in ligand binding process : charge density study of a fatty acid binding protein", dans le cadre de la conférence Gordon "Electron Distribution and Chemical Bonding" de 2013. Les lignes de champ électrique dans la poche de fixation de la protéine FABP sont représentées par des lignes bleues. Elles ont été calculées à partir du potentiel électrostatique $V(\mathbf{r})$ généré par la protéine apo (sans l'acide oléique) et par les molécules d'eau à l'intérieur de la cavité. Ces lignes de champ électrique sont donc celles qui sont ressenties par l'acide oléique (lig-OLE), qui est représenté en vert.

l'être aussi par la répulsion due à la présence d'un résidu nucléophile à l'entrée de la poche. De même, la direction des lignes de champ électrique, et donc des forces ressenties par un groupement chargé, dépend simultanément des positions respectives des groupes électrophiles et nucléophiles impliqués. C'est ce que je vais illustrer avec l'exemple de la protéine de fixation des acides gras FABP ("Fatty Acid Binding Protein").

Retour sur la FABP

Les FABP sont des protéines de transport des acides gras qui sont connues pour présenter une large cavité remplie par des molécules d'eau, qui sont présumées nécessaires pour établir un environnement électrostatique favorable à la fixation du ligand [Hanhoff *et al.*, 2002]. L'étude de la FABP réalisée par E. I. Howard et ses collègues [Howard *et al.*, 2016], déjà mentionnée dans la section 1.3.2, avait effectivement permis de discuter l'importance du cluster de molécules d'eau conservé à l'intérieur de la cavité pour la fixation de l'acide oléique à partir des énergies d'interaction électrostatique et de la topologie de la densité électronique. Les lignes de champ électrique avaient également été envisagées pour caractériser les influences électrostatiques auxquelles l'acide oléique est soumis. C'est ce que montre la figure 2.8 qui a été réalisée par B. Guillot. Ces lignes de champ sont générées par les contributeurs des résidus de la protéine et des molécules d'eau uniquement (sans ceux de l'acide oléique) à l'intérieur de la cavité. Les conformations des résidus à l'intérieur de la poche étant conservées par rapport à la structure cristallographique de la forme apo de la protéine (code PDB : 3RZY [González et Fisher, 2015]), les lignes de champ ainsi obtenues caractérisent les forces électrostatiques ressenties par l'acide gras lors de son approche. Plusieurs contributeurs dominants se distinguent : les résidus Arg126, Tyr128 et Asp76 et quelques molécules d'eau conservées à l'intérieur de la cavité. Il est inté-

ressant de noter que, dans le complexe FABP - acide oléique, la tête carboxylate de cet acide gras est fixée par la charge positive du résidu Arg128 et par le groupement polaire hydroxyle du résidu Tyr128. De plus, le résidu Asp76 se trouve sur une boucle située à l'entrée de la cavité et fait partie du « portail » de la poche, pouvant passer d'une conformation ouverte (plus stable dans la forme apo [Matsuoka *et al.*, 2015]) à une conformation fermée (plus stable dans la forme holo [Matsuoka *et al.*, 2015]). Finalement, cette approche n'a pas été publiée dans l'article [Howard *et al.*, 2016], bien qu'elle ait été mentionnée dans la revue de C. F. Matta [Matta, 2014], car une description rigoureuse à partir de ces lignes de champ était difficile à mettre en œuvre à l'époque.

Grâce à l'implémentation dans MoProViewer des descripteurs issus de l'analyse topologique de $V(\mathbf{r})$, c'est-à-dire points critiques, faisceaux primaires et zones d'influence, ceux-ci peuvent être appliqués aux complexes protéine-ligand pour ce type d'étude. Les figures 2.9a et 2.9b montrent respectivement les ZIE et les ZIN correspondant aux lignes de champ électrique dans la poche de fixation de la FABP. Pour les ZIE, les deux contributeurs électrophiles principaux (sphères bleues), situés au fond de la poche, sont identifiés avec précision : le proton $H\epsilon$ de l'Arg126 (ZIE représentée par la surface bleu foncé) et le proton $H\eta$ de la Tyr128 (ZIE représentée par la surface cyan). Ce sont bien ces sites électrophiles qui interagissent directement avec le groupement carboxylate de l'acide oléique dans le complexe. De plus, l'extension de leur influence à travers la cavité peut permettre de guider l'approche de ce ligand. Ces ZIE sont capables de s'étendre dans une large partie de la poche grâce à la présence des contributeurs nucléophiles (sphères rouges) localisés à l'entrée de la cavité : l'un des doublets non-liants de l'atome d'oxygène O δ 1 de l'Asp76 et l'un des doublets non-liants de l'atome d'oxygène de la molécule d'eau Wat31. Notons que la Wat31 fait partie des molécules d'eau conservées dans les autres structures de FABP de la PDB, même dans les complexes formés avec un acide gras différent. La représentation complémentaire des ZIN, surface rouge pour l'atome Asp76-O δ 1 et surface orange pour Wat31-O dans la figure 2.9b, confirme l'étendue des influences électrostatiques ressenties dans la cavité. Notamment, la présence du résidu chargé négativement Asp76, qui est ici dans la conformation dite fermée de la poche, permet aux forces électrostatiques ressenties par l'acide oléique de s'établir. De même que les résidus Arg126 et Tyr128 sont capables d'attirer la tête nucléophile de ce ligand vers eux, ce résidu Asp76 pourrait guider sa fixation en le repoussant. Par conséquent, l'approche en zones d'influence électrophile et nucléophile offre un nouveau point de vue pour l'étude des structures de protéines, permettant d'obtenir des informations supplémentaires comme l'identification de contributeurs lointains à la fixation du ligand qui ne seraient pas caractérisés par des approches classiques.

2.4 Implémentation dans MoProViewer

2.4.1 La suite logiciel MoPro et le logiciel MoProViewer

La suite de logiciels MoProSuite [Jelsch *et al.*, 2005]⁶, développée par l'équipe BioMIMIC du CRM², est dédiée d'une part à l'affinement des structures atomiques des composés cristallins et

6. La suite logiciel MoProSuite est disponible sur le site du laboratoire CRM² : <https://crm2.univ-lorraine.fr/logiciels/mopro/>

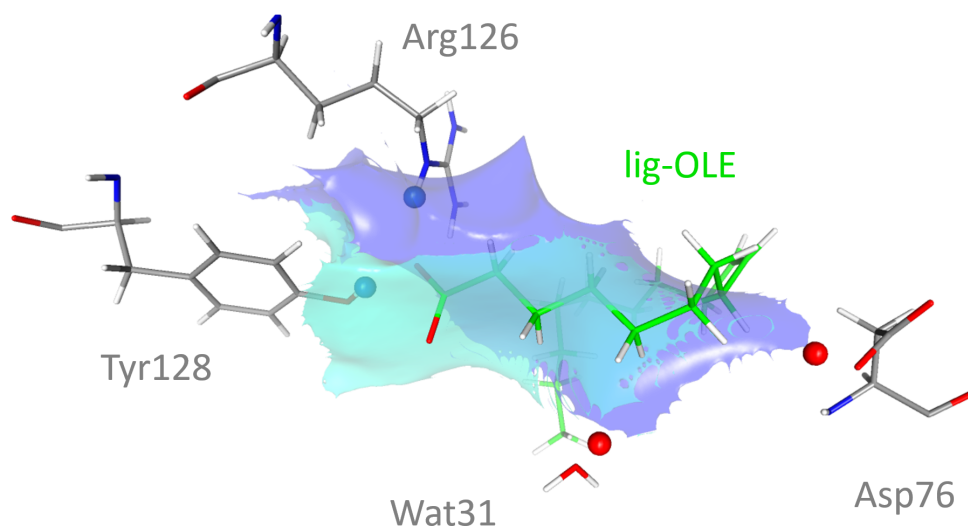
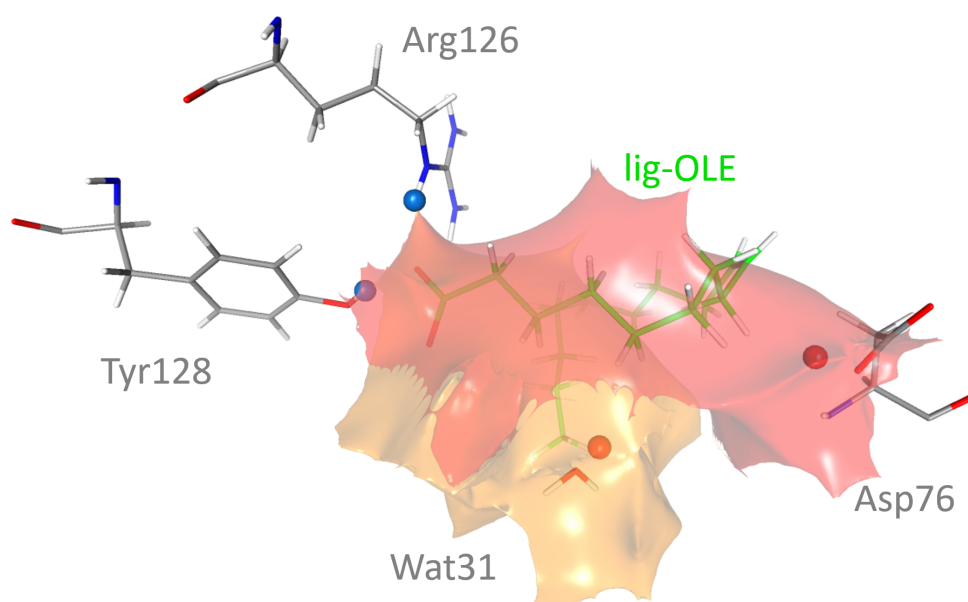
(a) Zones d'influence électrophile de Arg126-H ϵ et Tyr128-H η .(b) Zones d'influence nucléophile de Asp76-O $\delta 1$ et Wat31-O.

FIGURE 2.9 – Zones d'influence électrophile et nucléophile dans la poche de fixation de la FABP.

Les zones d'influence (a) électrophile et (b) nucléophile ont été déterminées dans la poche de fixation de la protéine FABP. Ces zones d'influence ont été calculées à partir du potentiel électrostatique $V(\mathbf{r})$ dans la protéine apo (sans l'acide oléique) en tenant compte des molécules d'eau à l'intérieur de sa cavité (dont Wat31 qui est affichée ici). Les acides aminés de la protéine (Asp76, Arg126 et Tyr128) sont représentés avec des atomes de carbone gris. Le ligand acide oléique (lig-OLE) est quant à lui représenté avec des atomes de carbone verts, et permet de comparer les régions recouvertes par les zones d'influence à la position qu'il prend dans la structure du complexe, bien qu'il n'ait pas été pris en compte dans le calcul de $V(\mathbf{r})$. Les sphères bleues correspondent aux points critiques (3, -3) caractérisant les sites électrophiles des protons H ϵ de l'arginine Arg126 et H η de la tyrosine Tyr128. La zone d'influence électrophile de Arg126-H ϵ est colorée en bleu foncé et celle de Tyr128-H η en couleur cyan. Les sphères rouges correspondent aux points critiques (3, +3) caractérisant les sites nucléophiles de doublets non-liants des atomes d'oxygène O $\delta 1$ de l'aspartate Asp76 et O de la molécule d'eau Wat31. La zone d'influence nucléophile de Asp76-O $\delta 1$ est colorée en rouge et celle de Wat31-O en orange.

d'autre part au calcul et à l'analyse des propriétés moléculaires. Pour les affinements cristallographiques, le logiciel MoPro et son interface utilisateur MoProGUI permet d'employer le modèle multipolaire de densité électronique de Hansen et Coppens [Hansen et Coppens, 1978] présenté dans la section 1.2.2. Les paramètres du modèle multipolaire peuvent soit être affinés contre des données de diffraction obtenues à résolution subatomique, soit transférés depuis les bases de données ELMAM2 ou MATTS introduites dans la section 1.2.3. Par ailleurs, le programme VMoPro réalise le calcul de cartes en 2D et en 3D de nombreuses propriétés moléculaires comme notamment la densité électronique $\rho(\mathbf{r})$, son gradient $\nabla\rho(\mathbf{r})$, son Laplacien $\nabla^2\rho(\mathbf{r})$, le potentiel électrostatique $V(\mathbf{r})$ et le champ électrique $\mathbf{E}(\mathbf{r})$. Les cartes de ces champs scalaires comportent les valeurs de la quantité calculée sur un maillage uniforme 3D d'un espace qui peut recouvrir une partie ou la totalité d'une molécule ou d'un ensemble de molécules. Le calcul des énergies d'interaction électrostatique selon la méthode nEP/MM (voir section 1.4.1) est également proposé par VMoPro, en lignes de commande.

Le logiciel MoProViewer [Guillot *et al.*, 2014] est quant à lui doté d'une interface graphique permettant la visualisation 3D des structures cristallines et moléculaires. Il propose la visualisation des propriétés géométriques, de déplacement thermique et électrostatiques dans les structures atomiques. En particulier, les cartes de $\rho(\mathbf{r})$ et de $V(\mathbf{r})$ calculées par VMoPro peuvent être affichées autour des molécules sous la forme d'isosurfaces 3D ou d'isocontours dans des plans 2D. Pour l'étude des protéines, MoProViewer propose des outils faciles à utiliser pour transférer les paramètres multipolaires de la librairie ELMAM2 sur ces structures et construire un modèle de densité électronique moléculaire précis pour ces systèmes. Grâce à ces modèles, les outils d'analyse basés sur la densité, sur le potentiel et sur l'énergie d'interaction électrostatique (méthode nEP/MM ou aEP/MM) peuvent également être appliqués à ces macromolécules. De plus, le calcul et l'affichage des points critiques de la densité électronique $\rho(\mathbf{r})$ et des bassins atomiques sont disponibles dans MoProViewer. Les lignes de champ électrique intersectant un plan de l'espace peuvent également être représentées. Aussi, MoProViewer propose déjà beaucoup d'outils sur lesquels j'ai pu me reposer pour le développement de mes descripteurs de la topologie du potentiel électrostatique. Notamment, les fonctions permettant de calculer la valeur du potentiel électrostatique $V(\mathbf{r})$ ou du champ électrique $\mathbf{E}(\mathbf{r})$ en des points quelconques de l'espace étaient déjà mises en place à partir des champs scalaires calculés par VMoPro et par des méthodes d'interpolation numérique de type polynômes de Lagrange. Le logiciel MoProViewer est programmé en langage C++ et utilise la librairie Qt pour son interface utilisateur.

2.4.2 Implémentation de la détermination des points critiques du potentiel électrostatique

L'algorithme de recherche des points critiques du potentiel électrostatique $V(\mathbf{r})$ détaillé dans la partie 2.1 a été implémenté dans le logiciel MoProViewer. Comme déjà mentionné, ce logiciel proposait déjà une méthode de détermination des points critiques de la densité électronique $\rho(\mathbf{r})$. Cependant, cette méthode s'appuyant sur le fait que $\rho(\mathbf{r})$ est toujours positive mais ce n'est pas le cas pour le potentiel électrostatique $V(\mathbf{r})$. Les outils déjà en place dans MoProViewer m'ont tout de même permis de mettre en place facilement la méthode de calcul des points critiques de $V(\mathbf{r})$.

Fenêtre de dialogue et paramètres utilisateur

J'ai implémenté cette méthode comme une alternative ajoutée à l'interface utilisateur de l'outil de points critiques ("Critical Point") déjà existant dans MoProViewer, cette fois pour le potentiel électrostatique. Une capture de la fenêtre de dialogue associée à l'outil est proposée sur la figure 2.10. Cette interface propose cinq paramètres ajustables par l'utilisateur. Le premier est le nombre maximum d'itérations autorisé pour l'algorithme de Newton-Raphson explicité dans la partie 2.1, c'est-à-dire le nombre de positions successives $\mathbf{r}_{k+1}$ testées avant de considérer que l'algorithme n'a pas convergé. La valeur de α est un coefficient qu'il est courant d'ajouter devant le pas de déplacement $\mathbf{h}$ dans cet algorithme : $\mathbf{r}_{k+1} = \mathbf{r}_k + \alpha\mathbf{h}$. Pour $\alpha < 1$, la minimisation de Newton-Raphson est dite amortie ("damped") et la recherche des points critiques s'effectue par pas plus petits afin d'améliorer la convergence au prix d'un temps de calcul plus élevé. La valeur minimale ε de la norme du gradient (ou du champ électrique $\mathbf{E}(\mathbf{r})$) est la valeur de $|\mathbf{E}(\mathbf{r})|$ en-dessous de laquelle l'algorithme est considéré convergé et qu'un point critique a été trouvé. Le pas angulaire $d\theta = d\varphi$ est le pas (en radians) utilisé entre les points des grilles sphériques définies autour des noyaux (voir figure 2.3a). Le pas radial dr est quant à lui le pas (en Å) utilisé pour rechercher les extrema locaux 1D sur les rayons partant des noyaux (voir figure 2.3d). Si aucun atome n'est sélectionné dans la structure alors tous les atomes sont pris en compte dans la recherche des points critiques et donc pour la définition des grilles sphériques, sinon il est possible de ne sélectionner que ceux appartenant à une région d'intérêt. Les points critiques trouvés apparaissent dans la fenêtre de visualisation 3D de la molécule avec des couleurs différentes pour chaque type point critique (voir figure 2.1), comme déjà implémenté pour la méthode de recherche des points critiques de la densité électronique $\rho(\mathbf{r})$.

Temps d'exécution

Le temps d'exécution de cet algorithme de recherche des points critiques de $V(\mathbf{r})$ dépend fortement du nombre d'atomes sélectionnés et des différents paramètres, notamment la valeur de α et des pas angulaires et radiaux. Par exemple, en gardant les paramètres recommandés (figure 2.10), il faut seulement 3,2 secondes⁷ pour obtenir 59 points critiques à partir des 18 atomes du complexe N-méthylacétamide - méthanol. Pour retrouver les 76 points critiques apparaissant dans la figure 2.1a, il faut choisir $\alpha = 0,1$, le temps de calcul étant alors de 3,9s. Par contre, pour une macromolécule, ces temps d'exécution peuvent devenir très longs. Par exemple, dans la trypsine en complexe avec un inhibiteur peptidique, en sélectionnant 1110 atomes entourant le site actif, il faut attendre 2407s (soit 40 minutes) pour obtenir les 42 068 points critiques autour de ces atomes. Il est donc fortement recommandé pour ce type de système de profiter des fonctions de MoProViewer pour calculer les points critiques dans une petite région ciblée de l'espace moléculaire, comme par exemple l'interface protéine-ligand, en ne sélectionnant que les atomes à l'intérieur de cette région d'intérêt avant de démarrer le calcul.

7. L'ordinateur utilisé pour ces estimations possède un processeur Intel(R) Core(TM) i7-10700 de 2,90 GHz et 16,0 Go de RAM.

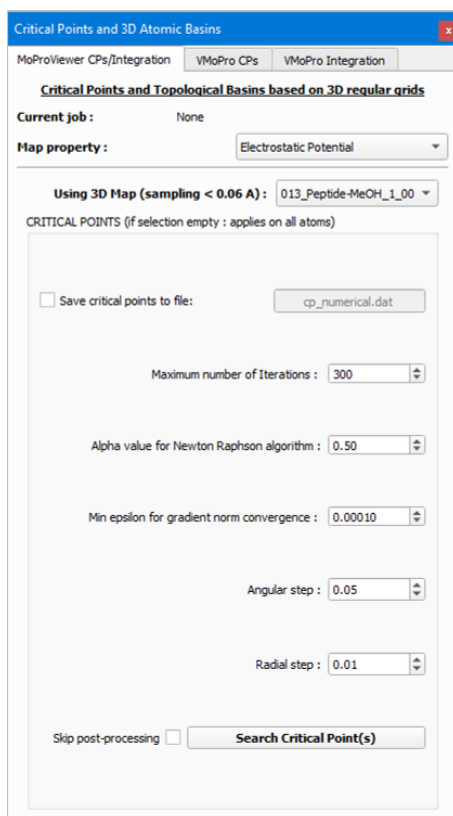


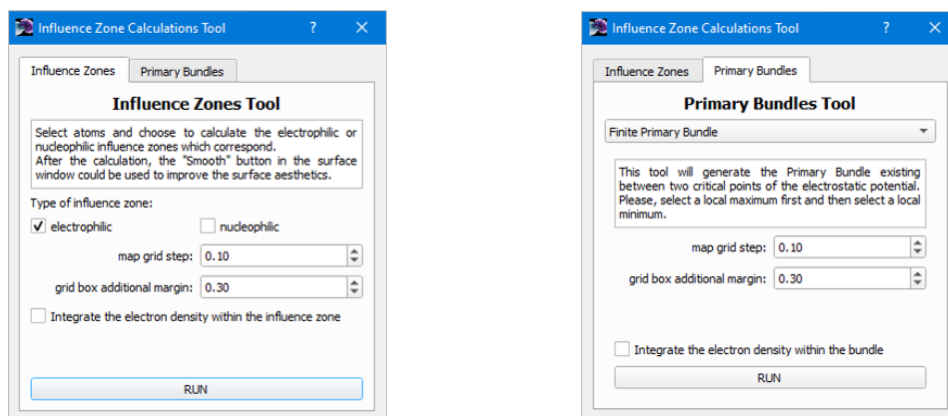
FIGURE 2.10 – Interface utilisateur de l'outil de calcul des points critiques du potentiel électrostatique dans MoProViewer.

La fenêtre de calcul des points critiques de $\rho(\mathbf{r})$ et de $V(\mathbf{r})$ dans MoProViewer est intitulée "Critical Points and 3D Atomic Basins". Elle comporte trois onglets mais seul le premier onglet "MoProViewer CPs/Integration" permet d'accéder au calcul des points critiques de $V(\mathbf{r})$. Sur cet onglet, la première option est un menu déroulant permettant de choisir le champ scalaire considéré (densité électronique ou potentiel électrostatique). Le choix de la carte VMOPro du champ scalaire est également proposé. Pour le potentiel électrostatique, cinq paramètres du calcul peuvent être ajustés par l'utilisateur : le nombre maximum d'itérations, la valeur de α pour l'algorithme de Newton-Raphson, la valeur minimale ε de la norme du gradient (ou du champ électrique) pour la convergence, le pas angulaire et le pas radial. Ces différents paramètres sont détaillés dans le texte. Le bouton "Search Critical Point(s)" permet de lancer le calcul.

2.4.3 Implémentation de la détermination des faisceaux primaires et des zones d'influence

La méthode de détermination des faisceaux primaires et des zones d'influence, présentée dans les parties 2.2 et 2.3, a également été implémentée dans MoProViewer. Le logiciel comportait déjà un outil de calcul des bassins topologiques de la densité électronique $\rho(\mathbf{r})$, les bassins atomiques, mais celui-ci étant basé sur la convexité systématique de ces bassins, il n'est pas applicable aux bassins topologiques du potentiel électrostatique $V(\mathbf{r})$ qui peuvent être convexes ou concaves et prendre des formes très complexes. Néanmoins, les outils de représentation des surfaces 3D, basés sur la librairie OpenMesh, utilisés pour les bassins atomiques ont été également employés pour la représentation des faisceaux primaires et des zones d'influence.

J'ai intégré une nouvelle fenêtre graphique dans MoProViewer, intitulée "Influence Zone Calculations Tool", afin de pouvoir utiliser l'algorithme de détermination de ces surfaces, détaillé



(a) Interface de calcul des zones d'influence. (b) Interface de calcul des faisceaux primaires.

FIGURE 2.11 – Interface utilisateur de l'outil de calcul des zones d'influence et des faisceaux primaires dans MoProViewer.

La fenêtre de calcul des faisceaux primaires et des zones d'influence dans MoProViewer est intitulée "Influence Zone Calculations Tool". Elle comporte deux onglets : (a) pour les zones d'influence et (b) pour les faisceaux primaires. (a) Le premier onglet "Influence Zone Tool" propose deux options : "electrophilic" pour calculer une ZIE et "nucleophilic" pour calculer une ZIN. Deux paramètres sont ajustables par l'utilisateur : le pas de la grille de calcul du champ scalaire binaire $\eta(x, y, z)$ et la marge additionnelle pouvant être ajoutée pour agrandir la boîte dans laquelle la détermination de la zone d'influence est effectuée. Une autre option est ensuite proposée afin de calculer l'intégrale de la densité électronique à l'intérieur de la zone d'influence. Enfin, le bouton "RUN" permet de lancer le calcul. (b) Le second onglet "Primary Bundles Tool" est très similaire. La seule différence est qu'à la place de choisir entre "electrophilic" et "nucleophilic", un menu déroulant propose de choisir entre faisceau primaire fermé ("Finite Primary Bundle") et un faisceau primaire ouvert ("Open Primary Bundle"). Pour l'instant, seule la représentation des faisceaux primaires fermés a été implémentée.

dans la partie 2.2. Cet outil utilise comme champ scalaire de potentiel électrostatique la carte active dans MoProViewer. Deux captures de l'interface utilisateur associée à l'outil sont présentées dans la figure 2.11, l'une pour l'onglet de calcul des zones d'influence (figure 2.11a) et l'autre pour l'onglet de calcul des faisceaux primaires (figure 2.11b).

Outil de zones d'influence

Pour le calcul des zones d'influence, après avoir sélectionné les atomes d'intérêts, l'utilisateur a le choix de calculer soit les ZIE en cochant "electrophilic" ou les ZIN en cochant "nucleophilic". Tous les atomes peuvent être associés à une ZIE mais pour les ZIN, si un des atomes sélectionnés n'est associé à aucun site nucléophile alors il ne possède pas de ZIN et est éliminé de la liste. Deux paramètres de la grille orthogonale et régulière (voir figure 2.5b) sont ensuite ajustables : son pas, c'est-à-dire la distance en Å entre deux points (x, y, z) voisins, et la marge additionnelle m pour agrandir la boîte dans laquelle la grille est définie : les coordonnées des coins de la boîte $(x_{\min}, y_{\min}, z_{\min})$ et $(x_{\max}, y_{\max}, z_{\max})$ deviennent $(x_{\min} - m, y_{\min} - m, z_{\min} - m)$ et $(x_{\max} + m, y_{\max} + m, z_{\max} + m)$. Les valeurs par défaut de ces paramètres sont de 0,1Å pour le pas et 0,3Å pour la marge. Diminuer la valeur du pas permet d'améliorer la précision des bords des zones d'influence mais rend le calcul plus long. La valeur de 0,1Å a été choisie comme meilleur compromis entre la qualité de la délimitation et le coût computationnel. Pour la marge, une valeur faible telle que 0,3Å est suffisante pour les ZIE tandis qu'une valeur plus haute (de 1,0

à 3,0Å) peut s'avérer nécessaire pour les ZIN qui ont généralement un volume plus large. Le volume de la zone d'influence est estimé par sommation des volumes élémentaires de la grille de définition du champ scalaire $\eta(x, y, z)$ $dx dy dz$, soit le cube du pas de grille, qui sont à l'intérieur de la surface ($\eta(x, y, z) = 1$). Le résultat apparaît à la fin du calcul dans la console de sortie du logiciel. Une autre option cochable est également disponible pour calculer l'intégrale de la densité électronique dans le volume de la zone d'influence, également donnée dans la console de sortie. Cette option a notamment été utilisée pour l'application du théorème de Gauss dans la partie 2.3.3.

Le temps de calcul dépend très fortement du volume de la zone d'influence et peut durer de quelques secondes à quelques minutes. Par exemple, pour la ZIE de l'oxygène carbonyle du N-méthylacétamide (O12), qui est très petite (volume $\mathcal{V} = 6\text{\AA}^3$) car elle s'arrête sur les doublets non-liants du même atome, il ne faut que 1,21s pour le calcul et l'affichage. Par contre, il faut compter 1min et 36s pour sa ZIN, dont la surface représentée en jaune dans la figure 2.7b recouvre une large partie de l'espace (volume $\mathcal{V} = 355\text{\AA}^3$ sur une carte de potentiel électrostatique contenue dans une boîte de $858\text{\AA}^3$). L'étape la plus coûteuse est de loin l'évaluation du champ scalaire binaire $\eta(x, y, z)$ sur tous les points de la grille de calcul. Néanmoins, puisque le calcul en un point (x, y, z) ne dépend pas des autres points, cet algorithme est fondamentalement parallélisable, ce qui pourrait réduire son temps d'exécution. Une fois le maillage triangulaire de la surface englobant la zone d'influence obtenu (voir figure 2.5d), cette surface est affichée dans la fenêtre de visualisation 3D de la molécule en utilisant la librairie OpenMesh. La surface directement obtenue par l'algorithme des Marching Cubes (voir figure 2.5e) a un aspect cabossé, faisant apparaître des « marches » dues au caractère binaire de la fonction implicite $\eta(x, y, z)$. La librairie OpenMesh fournit également des fonctions permettant de lisser ces surfaces afin d'améliorer leur apparence (voir figure 2.5f).

Outil de faisceaux primaires

Pour les faisceaux primaires, une option a été prévue pour que l'utilisateur puisse choisir entre faisceau primaire fermé et faisceau primaire ouvert, mais pour l'instant seule l'implémentation des faisceaux primaires fermés a été réalisée. En effet, la définition des faisceaux primaires ouverts, qui repose sur des lignes de champ électrique ayant une extrémité partant à l'infini, est complexe à mettre en pratique. Le potentiel électrostatique étant calculé dans une boîte finie, les lignes de champ électrique s'arrêtent aussi sur les bords de cette boîte et on ne peut donc pas vérifier si ces lignes de champ vont finalement converger vers un point critique extérieur à la boîte ou si elles partent effectivement à l'infini. Pour le calcul d'un faisceau primaire fermé, l'utilisateur commence par sélectionner la paire de points critiques correspondant aux extrémités des lignes de champ. Les paramètres ajustables et méthodes d'affichage sont les mêmes que pour les zones d'influence et leurs temps de calcul sont similaires. L'option de calcul de l'intégrale de la densité électronique est également disponible à l'intérieur du volume du faisceau primaire.

2.4.4 Perspectives de développements

Les outils de détermination des surfaces entourant les faisceaux primaires et les zones d'influence ont vocation à être employés de façon routinière dans les études de complexes protéine-

ligand. Ils doivent être simples et pratiques à mettre en œuvre pour des utilisateurs qui ne sont pas experts de la topologie du potentiel électrostatique. Pour cela, la méthode de calcul doit devenir plus performante et les représentations plus faciles à interpréter. Comme expliqué précédemment, la programmation du test des points à l'intérieur ou à l'extérieur de ces surfaces peut être optimisée de façon à ce que plusieurs points soient testés en même temps, de manière parallèle, dans le but de rendre le calcul de la fonction implicite plus rapide. De plus, l'esthétique des surfaces obtenues peut également être améliorée. Une méthode de lissage des surfaces plus élaborée que celle de la librairie OpenMesh déjà proposée pourrait être implémentée. Une autre possibilité serait de traiter directement le champ scalaire binaire pour combler les trous et gommer les points isolés avant la génération de l'isosurface.

Pour rendre l'interprétation des faisceaux primaires et des zones d'influence plus aisée, les natures électrophile ou nucléophile des extrémités peuvent être davantage mises en évidence. Par exemple, les surfaces pourraient être colorées en fonction de la valeur du potentiel électrostatique, avec un gradient de couleur allant du rouge pour les valeurs négatives (sites nucléophiles) vers le bleu pour les valeurs positives (sites électrophiles), de façon similaire aux représentations sur les surfaces d'accessibilité au solvant. L'affichage dans les volumes des faisceaux primaires et des zones d'influence de quelques lignes de champ électrique ou d'une ligne de champ moyenne, marquées par des flèches indiquant leurs directions, pourrait également permettre de mieux visualiser la directionnalité des influences électrostatiques.

Par ailleurs, d'autres outils peuvent être ajoutés pour compléter ceux développés pendant ma thèse. Notamment, une méthodologie pour la représentation des faisceaux primaires ouverts pourra être définie. De plus, un outil « sonde » pourra être implémenté de façon à générer le faisceau primaire, la ZIE ou la ZIN intersectant sa position. En effet, pour l'instant, la méthode de détermination de ces surfaces part de la sélection des contributeurs générant les lignes de champ électrique (atomes ou points critiques). Avec un outil de type sonde, il serait possible de prendre comme point de départ un point quelconque de l'espace. Un outil dynamique peut même être envisagé, dans lequel la surface affichée serait mise à jour en fonction de la position courante de la sonde, permettant ainsi de caractériser les influences électrostatiques tout en se déplaçant dans l'espace moléculaire.

Pour l'instant, les surfaces des zones d'influence sont calculées dans les structures statiques de cristallographie. L'ajout de la lecture des fichiers de trajectoires de dynamique moléculaire dans MoProViewer permettrait d'analyser l'évolution temporelle de ces zones d'influence. Les surfaces inchangées le long d'une trajectoire pourraient notamment permettre de caractériser finement un environnement électrostatique conservé, et donc pertinent pour expliquer des aspects de relation structure-fonction. En outre, la représentation des zones d'influence dans une trajectoire de dynamique moléculaire simulant par exemple l'amarrage d'un ligand dans la poche de fixation d'une protéine ou le mécanisme réactionnel d'une enzyme contribuerait à l'identification des influences électrostatiques déterminantes pour ces processus.

2.5 Conclusion partielle de chapitre

L'analyse topologique du potentiel électrostatique $V(\mathbf{r})$ repose sur l'étude de ses points critiques et de ses bassins topologiques. Pour la détermination des points critiques, j'ai utilisé

l'algorithme de minimisation de Newton-Raphson pour trouver les points où le gradient du potentiel, c'est-à-dire au signe près le champ électrique, s'annule. Pour trouver les positions de départ sur lesquels démarrer cet algorithme, j'ai suivi la méthode proposée par P. Balanarayan et S. R. Gadre partant de l'hypothèse que la présence d'un point critique de champ scalaire moléculaire doit être ressentie en sondant une surface fermée autour de la molécule. Parmi les points critiques ainsi déterminés, les maxima locaux $(3, -3)$ et les minima locaux $(3, +3)$ localisent respectivement les sites électrophiles, c'est-à-dire les noyaux, et les sites nucléophiles, qui sont souvent associés à des paires d'électrons non-liantes. Quant aux points-selles, certains de type $(3, -1)$ sont situés sur les liaisons covalentes et non-covalentes et certains de type $(3, +1)$ caractérisent des structures cycliques, mais la plupart de ces points critiques ne possèdent pas d'interprétation moléculaire clairement définie.

Les bassins topologiques de $V(\mathbf{r})$ sont les faisceaux primaires de lignes de champ électrique. Contrairement à un bassin atomique de la densité $\rho(\mathbf{r})$ qui est associé à un atome, un faisceau primaire est associé à une paire site électrophile - site nucléophile. Néanmoins, par union des faisceaux primaires associés au même site électrophile, la ZIE (zone d'influence électrophile) obtenue est associée à l'atome dont le noyau correspond à ce site électrophile. De façon similaire, l'union des faisceaux primaires associés au(x) site(s) nucléophile(s) correspondant aux doublets non-liants portés par un atome définit la ZIN (zone d'influence nucléophile) de cet atome. La méthode que j'ai proposée pour déterminer les surfaces délimitant les faisceaux primaires et les zones d'influence repose sur la définition des conditions aux extrémités des lignes de champ électrique à l'intérieur de ces surfaces. Les faisceaux primaires et les zones d'influence étant entourés par des surfaces du flux nul de potentiel, elles doivent vérifier le théorème de Gauss. J'ai testé la validité de ce théorème dans les ZIE du complexe N-méthylacétamide - méthanol et obtenu des erreurs absolues de l'ordre de $\pm 0,10e$ et des erreurs relatives allant de 1% à 10% selon les atomes considérés, et globalement une erreur de $0,12e$ sur tout le système, soit 0,02% de la charge électronique totale. Ces résultats permettent de conforter la capacité de ma méthode à déterminer les bords des faisceaux primaires et des zones d'influence. Grâce à la partition de l'espace en ZIE, chaque point de l'espace est associé à l'influence électrophile d'un unique atome. De même, la partition de l'espace en ZIN permet d'identifier l'atome responsable de l'influence nucléophile à laquelle un point de l'espace est soumise.

Grâce à l'implémentation dans le logiciel MoProViewer des outils permettant de mettre en œuvre ces différents descripteurs, ces derniers peuvent également être appliqués à l'étude des systèmes protéine-ligand. En effet, ces larges systèmes sont habituellement décrits en termes de descripteurs électrostatiques par des représentations se limitant à l'évaluation de $V(\mathbf{r})$ sur des surfaces moléculaire [Weiner *et al.*, 1982]. Les descripteurs basés sur la topologie de $V(\mathbf{r})$ que j'ai développés, notamment la partition en ZIE et ZIN, ont l'ambition d'apporter un nouveau point de vue pour analyser les interactions de nature électrostatique dans les complexes protéine-ligand grâce à : (i) la localisation précise des contributeurs électrostatiques même situés à longue distance, (ii) la visualisation de l'étendue et de la directionnalité spatiale des forces électriques et (iii) l'association de chaque région de l'espace à un influenceur électrophile et un influenceur nucléophile. Comme illustré avec l'exemple de la protéine FABP, ces différents aspects ont un fort potentiel pour aider à la compréhension des mécanismes de reconnaissance moléculaire guidant

l'approche d'un ligand dans la poche d'une protéine et stabilisant leur interaction. Il est supposé qu'ils aient également un pouvoir prédictif des chemins réactionnels lorsque ceux-ci sont dominés par des effets de nature électrostatique [Mata *et al.*, 2007, Mata *et al.*, 2015, Alkorta *et al.*, 2019], pouvant être appliqués aux mécanismes enzymatiques.

Dans le chapitre 4, plusieurs applications aux protéines de ces descripteurs issus de l'analyse topologique du potentiel électrostatique seront proposées. Tout d'abord, dans la partie 4.1, l'intérêt des méthodologies que j'ai développées pour l'étude des macromolécules sera discuté dans le cadre de l'analyse d'un système enzymatique : le complexe d'une protéase à sérine avec un inhibiteur canonique, mimant le complexe de Michaelis [Wahlgren *et al.*, 2011]. Ensuite, dans la partie 4.2, l'application routinière des zones d'influence aux études de biologie structurale sera illustrée dans le cas de la neuropiline, une glycoprotéine membranaire qui est une cible thérapeutique importante dans le traitement de différents cancers [Dumond *et al.*, 2020]. Puis, dans la partie 4.4, la caractérisation d'une protéine de transport transmembranaire, l'halorhodopsine qui est impliquée dans le pompage d'ions chlorure [Mous *et al.*, 2022], sera introduite. Avant cela, le chapitre 3 détaille le second développement méthodologique réalisé pendant ma thèse : la construction d'un potentiel d'interaction intermoléculaire reposant sur les paramètres de densité électronique et de polarisabilité de la librairie ELMAM2.

Chapitre 3

Potentiel d'interaction basé sur la densité électronique expérimentale transférée ELMAM

L'énergie d'interaction est une quantité centrale pour l'étude des effets intermoléculaires. Comme expliqué dans la partie 1.4, de nombreuses méthodes ont vu le jour pour calculer cette énergie d'interaction, à partir de potentiels empiriques de mécanique moléculaire (MM), de méthodes *ab initio* de mécanique quantique (QM) ou d'approches hybrides de QM/MM. Durant ma thèse, j'ai travaillé sur un potentiel d'interaction total qui est quant à lui basé sur la densité électronique d'origine expérimentale. En effet, d'après le théorème de Hohenberg et Kohn [Hohenberg et Kohn, 1964], l'énergie d'un système dans son état fondamental est déterminée de façon unique par sa distribution de charge. La densité de charge moléculaire peut être reconstruite de façon très précise grâce au transfert des paramètres de densité électronique du modèle multipolaire, depuis une base de données telle que la librairie ELMAM2 introduite dans la section 1.2.3. La librairie ELMAM2 fournit des paramètres multipolaires affinés à partir de données de diffraction des rayons X à résolution subatomique, ainsi que des polarisabilités atomiques anisotropes déterminées à l'aide de méthodes quantiques de haut niveau. Le modèle d'énergie d'interaction présenté ici repose sur ces quantités ELMAM et se décompose en quatre contributions : électrostatique, d'induction, de dispersion et d'échange-répulsion. Les méthodes de calcul des énergies électrostatique et d'induction avaient déjà été validées [Leduc *et al.*, 2019, Vuković *et al.*, 2021] mais les contributions d'interaction de type van der Waals (dispersion et répulsion) n'étaient pas encore disponibles. C'est dans cet objectif que j'ai développé des potentiels d'interaction de dispersion et de répulsion reposant sur les quantités transférables ELMAM.

Dans ce chapitre, je vais commencer par détailler la méthodologie employée pour construire le potentiel d'interaction total $U_{\text{int}}^{\text{ELMAM}}$ ainsi que pour le tester grâce à un jeu de données de benchmarking de petites molécules organiques. Les méthodes de calcul des contributions électrostatique et d'induction seront vérifiées par application sur ce jeu de données. Ensuite, les modèles de potentiel de dispersion et de répulsion que j'ai développés seront présentés et testés. Puis, la validité du potentiel de van der Waals ELMAM obtenu à partir de ces deux contributions sera discutée.

3.1 Modèle basé sur les données de la librairie ELMAM2

3.1.1 Décomposition du potentiel d'interaction total ELMAM

Décomposition de type SAPT

Le potentiel d'interaction intermoléculaire total ELMAM, $U_{\text{inter}}^{\text{ELMAM}}$, suit une décomposition de type SAPT comme présenté dans la section 1.4.3, qui comporte quatre termes :

$$U_{\text{inter}}^{\text{ELMAM}} = U_{\text{elst}}^{\text{ELMAM}} + U_{\text{ind-dip}}^{\text{ELMAM}} + U_{\text{disp}}^{\text{ELMAM}} + U_{\text{exch-rep}}^{\text{ELMAM}}. \quad (3.1)$$

Le premier terme $U_{\text{elst}}^{\text{ELMAM}}$ est la contribution permanente à l'énergie électrostatique due à l'interaction coulombienne entre deux distributions de charge non-perturbées, qui sont construites à partir des paramètres multipolaires transférés depuis la librairie ELMAM2. Le second terme $U_{\text{ind-dip}}^{\text{ELMAM}}$ est un terme d'induction dipolaire, il provient de la polarisation mutuelle entre les distributions de charges en interaction, et est obtenu à partir du calcul des moments dipolaires induits sur les atomes proportionnels aux polarisabilités atomiques anisotropes ELMAM transférées. Les méthodes de calcul de ces deux termes, $U_{\text{elst}}^{\text{ELMAM}}$ et $U_{\text{ind-dip}}^{\text{ELMAM}}$, ont déjà été publiées [Leduc *et al.*, 2019, Vuković *et al.*, 2021] au contraire des deux autres contributions dont la modélisation fait partie de mes travaux de thèse. Le troisième terme $U_{\text{disp}}^{\text{ELMAM}}$ est la contribution de dispersion pour laquelle un modèle dérivé de l'approximation de London basé sur les polarisabilités transférées ELMAM a été utilisé. Les détails du développement de ce modèle de dispersion seront présentés dans la partie 3.2. Le dernier terme $U_{\text{exch-rep}}^{\text{ELMAM}}$ est le terme d'échange-répulsion qui repose sur le modèle du recouvrement des densités électroniques moléculaires, en utilisant à nouveau les densités électroniques transférées ELMAM. Le développement de ce modèle sera quant à lui détaillé dans la partie 3.3. Ces deux contributions ont été optimisées séparément pour ensuite obtenir le potentiel d'interaction de van der Waals par sommation.

Il est intéressant de noter qu'A. Gavezzotti a proposé un modèle similaire pour calculer l'énergie d'interaction dans les cristaux, appelé méthode des sommes de densités semi-classiques (SCDS pour "SemiClassic Density Sum") [Gavezzotti, 2002, Gavezzotti, 2003, Gavezzotti, 2005, Gavezzotti, 2011]. Ce modèle a notamment été utilisé dans le cadre de la prédiction d'empilements cristallins [Johnston *et al.*, 2011]. La contribution électrostatique est calculée par intégration numérique directe des densités électroniques de valence calculées théoriquement dans les géométries cristallines des molécules, selon la méthode « Pixel » d'A. Gavezzotti [Gavezzotti, 2002]. Les effets de polarisation dipolaire sont pris en compte par un modèle de polarisabilités distribuées sur les Pixels, calculées à partir de polarisabilités moléculaires isotropes d'origine expérimentale [Gavezzotti, 2003]. La dispersion est évaluée à partir de ces polarisabilités Pixels dans l'approximation London atténuées à courte distance et la répulsion est estimée à partir du recouvrement des densités électroniques de valence calculées théoriquement les géométries en phase gazeuse des molécules [Gavezzotti, 2003]. Ce modèle introduit quatre paramètres empiriques dans les potentiels de polarisation, de dispersion et de répulsion qui ont été optimisés séparément pour reproduire des données thermodynamiques expérimentales et des résultats de calculs de chimie quantique de haut niveau théorique [Gavezzotti, 2003]. Ce modèle s'appuie donc à la fois sur des densités électroniques calculées théoriquement, des polarisabilités molé-

culaires d'origine expérimentale et des paramètres déterminés de manière empirique. Le modèle ELMAM est quant à lui construit dans une approche auto-cohérente, de façon ne reposer que sur les données de la librairie ELMAM.

Calcul des énergies électrostatiques

Pour calculer l'énergie électrostatique à partir d'un modèle multipolaire de la densité électronique, le méthode EP/MM analytique [Nguyen *et al.*, 2018] est à la fois très précise et peu coûteuse en temps de calcul [Vuković *et al.*, 2021]. Cette méthode a été implémentée par V. Vuković dans la librairie Charger qui est utilisée par le logiciel MoProViewer. La densité électronique reconstruite par transfert des paramètres du modèle multipolaire d'origine expérimentale fournis par la librairie ELMAM2 est dite non-perturbée, c'est-à-dire qu'elle n'est pas déformée par les interactions intermoléculaires (voir section 1.4.1). La contribution électrostatique permanente $U_{\text{elst}}^{\text{ELMAM}}$ est donc directement calculée à partir de ces densités transférées.

Par ailleurs, les polarisabilités anisotropes atomiques transférables proposées par la librairie ELMAM2 permettent d'obtenir la densité électronique polarisée en calculant les moments dipolaires mutuellement induits entre deux distributions de charge en interaction, en utilisant la procédure établie par T. Leduc [Leduc *et al.*, 2019]. L'énergie électrostatique $U_{\text{elst, POL}}^{\text{ELMAM}}$ calculée par application de la méthode aEP/MM sur cette densité polarisée contient, en plus de la contribution électrostatique permanente $U_{\text{elst}}^{\text{ELMAM}}$, la contribution d'induction dipolaire $U_{\text{ind-dip}}^{\text{ELMAM}}$. Cette dernière peut donc être déduite par soustraction des énergies électrostatiques polarisée et permanente :

$$U_{\text{ind-dip}}^{\text{ELMAM}} = U_{\text{elst, POL}}^{\text{ELMAM}} - U_{\text{elst}}^{\text{ELMAM}}. \quad (3.2)$$

Les deux termes $U_{\text{elst}}^{\text{ELMAM}}$ et $U_{\text{ind-dip}}^{\text{ELMAM}}$ sont exprimés séparément plutôt que par leur somme $U_{\text{elst, POL}}^{\text{ELMAM}}$ dans le but de conserver une décomposition de type SAPT et de pouvoir comparer les résultats du potentiel d'interaction ELMAM à d'autres méthodes. Par ailleurs, il est important de noter que le terme $U_{\text{ind-dip}}^{\text{ELMAM}}$ ne correspond pas tout à fait au terme d'induction SAPT car il ne tient compte que des effets de polarisation dipolaire tandis que le terme SAPT prend aussi en compte les effets de polarisation quadripolaire (et d'ordres supérieurs), les transferts de charge et les termes croisés d'échange-induction.

3.1.2 Jeu de données d'entraînement et de validation du modèle

Le jeu de données de benchmarking NENCI-2021

Pour tester la validité des potentiels de dispersion et de répulsion que j'ai développés, le jeu de données de benchmarking NENCI-2021 ("Non-Equilibrium Non-Covalent Interaction") [Sparrow *et al.*, 2021] a été choisi. Il contient au total 7763 systèmes de 141 complexes dimériques différents dans des géométries variées. Certains de ces systèmes contenant des types atomiques qui ne sont pas pris en charge par la librairie ELMAM2, le transfert des paramètres multipolaires de la densité électronique a été effectué sur 77 dimères de ce jeu de données dans 45 géométries différentes chacun, soit un total de 3465 systèmes. Ces 45 géométries correspondent à des variations de la distance entre les centres de masse des molécules et des variations d'orientations relatives des monomères. Les distances intermoléculaires les plus courtes, de 0,7 et 0,8 fois

la distance d'équilibre, ont été écartées car les paramètres de densité transférés étant issus de structures à l'équilibre, ils ne sont plus pertinents à ces distances. Les dimères utilisés incluent les espèces chimiques les plus couramment rencontrées dans les protéines : hydrogène, carbone, azote, oxygène et soufre. De plus, ils sont caractérisés par des interactions de natures différentes, classés selon les auteurs en quatre catégories à dominantes : électrostatique (ELEC), dispersive (DISP), d'induction (IND) et mixte (MIXD). Parmi les 3465 systèmes retenus pour le transfert, 1417 (41%) sont de type DISP, 1240 (36%) de type ELEC, 808 (23%) de type MIXD et aucun n'est de type IND. En effet, dans le jeu de données NENCI-2021, les systèmes de type IND sont essentiellement des dimères contenant des ions lithium Li^+ ou sodium Na^+ , qui n'ont pas été transférés pour cette étude. Par ailleurs, les valeurs calculées par méthode SAPT des contributions à l'énergie d'interaction $E_{\text{elst}}^{\text{SAPT}}$, $E_{\text{ind}}^{\text{SAPT}}$, $E_{\text{disp}}^{\text{SAPT}}$ et $E_{\text{exch-rep}}^{\text{SAPT}}$ sont fournies pour chaque système, avec une erreur absolue moyenne de 0,3 kcal/mol par rapport à des calculs quantiques de plus haut niveau [Sparrow *et al.*, 2021]. Ces valeurs SAPT sont utilisées comme valeurs de référence pour tester individuellement les quatre contributions du potentiel d'interaction total ELMAM.

Entraînement et validation des potentiels d'interaction

De manière générale, la comparaison des résultats obtenus à partir du potentiel d'interaction ELMAM aux valeurs de référence SAPT permet de tester la validité du modèle. Pour cela, j'ai utilisé quatre indicateurs statistiques : le coefficient de détermination R^2 , le coefficient de corrélation linéaire R , l'erreur absolue moyenne MAE ("Mean Absolute Error") et la déviation quadratique moyenne, plus communément appelée RMSD (pour "Root Mean Square Deviation"). Le coefficient de détermination R^2 est couramment utilisé pour mesurer la linéarité d'une relation $y = f(x)$, plus sa valeur est proche de 1, plus la relation entre x et y est linéaire. Pour tester les résultats des potentiels ELMAM $y = U^{\text{ELMAM}}$ par comparaison aux valeurs de référence SAPT $x = E^{\text{SAPT}}$, l'expression de R^2 est :

$$R^2 = 1 - \frac{\sum_i (U_i^{\text{ELMAM}} - E_i^{\text{SAPT}})^2}{\sum_i (U_i^{\text{ELMAM}} - \overline{U^{\text{ELMAM}}})^2}, \quad (3.3)$$

où l'indice i porte sur les systèmes transférés et $\overline{U^{\text{ELMAM}}}$ est la moyenne des valeurs du potentiel ELMAM sur l'ensemble de ces systèmes. Pour une régression linéaire, le coefficient de détermination R^2 est le carré du coefficient de corrélation R . L'erreur absolue moyenne MAE est, comme son nom l'indique, une mesure de l'erreur moyenne commise sur l'énergie pour l'ensemble du jeu de données :

$$\text{MAE} = \frac{1}{N} \sum_i |U_i^{\text{ELMAM}} - E_i^{\text{SAPT}}|. \quad (3.4)$$

Le RMSD, qui est utilisé par exemple en bioinformatique pour mesurer l'alignement de deux structures différentes, est plus généralement appliqué en statistiques pour mesurer l'écart moyen entre deux séries de valeurs. Ici, il est utilisé de la façon suivante :

$$\text{RMSD} = \sqrt{\frac{1}{N} \sum_i (U_i^{\text{ELMAM}} - E_i^{\text{SAPT}})^2}. \quad (3.5)$$

Plus les valeurs de MAE et de RMSD sont faibles, plus la qualité du potentiel d'interaction est bonne. Pour comparer ces coefficients statistiques, j'ai employé l'erreur chimique typique ("typical chemical accuracy") de 1 kcal/mol qui est très souvent utilisée, comme par exemple dans l'article de J. W. Ponder [Rackers et Ponder, 2019].

Par ailleurs, pour les potentiels de dispersion et d'échange-répulsion, j'ai introduit quelques paramètres empiriques dans le but d'améliorer la qualité des modèles qui seront présentés dans les parties 3.2 et 3.3. Ces paramètres ont été entraînés à l'aide de 80% des systèmes transférés du jeu de données par la méthode d'affinement des moindres carrés¹. Les 20% des systèmes restant ont été utilisés pour la validation croisée des résultats obtenus par les modèles paramétrés.

3.1.3 Validation des contributions électrostatique et d'induction

Contribution électrostatique permanente

La validité du terme électrostatique permanent $U_{\text{elst}}^{\text{ELMAM}}$ calculé à partir des densités électroniques non-perturbées construites par transfert des paramètres multipolaires de la librairie ELMAM2 a déjà été testé sur les jeux de données de benchmarking S66 et S66x8 [Rezác *et al.*, 2011b, Rezác *et al.*, 2011a] par T. Leduc et ses collègues [Leduc *et al.*, 2019]. Ces jeux de données fournissent les différentes contributions énergétiques SAPT de 66 complexes moléculaires différents avec 8 variations de la distance intermoléculaire chacun, soit 528 systèmes. D'après la figure 4 de l'article [Leduc *et al.*, 2019], l'accord entre les résultats ELMAM, $U_{\text{elst}}^{\text{ELMAM}}$, obtenus à partir des densités électroniques transférées d'origine expérimentale et les valeurs SAPT, $E_{\text{elst}}^{\text{SAPT}}$, qui sont déterminées théoriquement, est modélisé par la régression linéaire suivante :

$$U_{\text{elst}}^{\text{ELMAM}} = 1,053 E_{\text{elst}}^{\text{SAPT}}, \quad (3.6)$$

avec le coefficient de détermination $R^2 = 0,970$ et RMSD = 1,4 kcal/mol. Ces indicateurs statistiques montrent un excellent accord entre les valeurs ELMAM et SAPT. Sans l'introduction de paramètre empirique, les ordres de grandeur obtenus par ces deux approches sont les mêmes, ce qui est indiqué par la pente de la régression linéaire proche de 1.

Le jeu de données NENCI-2021 étant plus large que les S66 et S66x8, il est intéressant de vérifier si les résultats obtenus pour $U_{\text{elst}}^{\text{ELMAM}}$ conservent leur très bon accord avec les valeurs de référence SAPT dans ce nouveau jeu de données. Le graphique 3.1a montre les résultats du potentiel d'interaction électrostatique ELMAM (axe des ordonnées) en fonction des valeurs SAPT (axe des abscisses). La régression linéaire sur ces points donne :

$$U_{\text{elst}}^{\text{ELMAM}} = 0,981 E_{\text{elst}}^{\text{SAPT}}, \quad (3.7)$$

avec le facteur détermination $R^2 = 0,941$. La pente est légèrement moins proche de l'unité que pour les jeux de données S66 et S66x8 mais reste tout de même très correcte compte tenu de l'origine expérimentale du modèle et de l'absence de tout paramètre empirique, par rapport à la

1. Les moindres carrés sont une méthode de régression très répandue, détaillée dans la plupart des ouvrages d'outils numériques tels que [Cornillon et Matzner-Lober, 2006]. Il existe des implémentations de l'algorithme de minimisation par moindres carrés dans divers langages informatiques mais j'ai choisi de réaliser mon propre programme en langage Python pour adapter ces calculs à mes besoins.

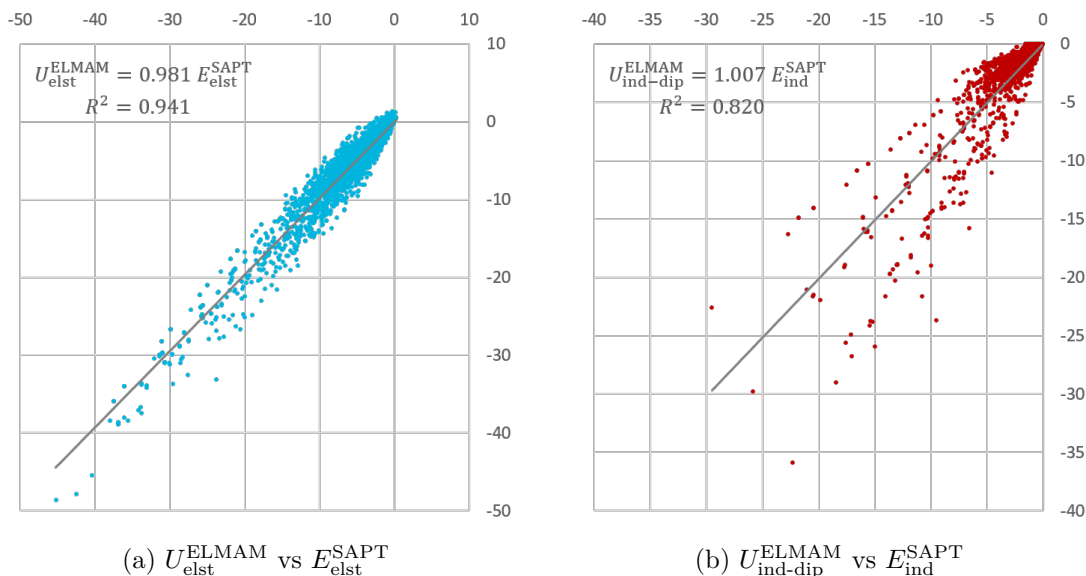


FIGURE 3.1 – Comparaison des résultats obtenus par les modèles ELMAM d'énergie électrostatique permanente et d'énergie d'induction dipolaire par rapport aux valeurs de référence SAPT.

Ces graphiques représentent les valeurs du modèle ELMAM en ordonnées et les valeurs SAPT en abscisses (a) des énergies d'interaction électrostatique permanente et (b) des énergies d'induction dans les dimères du jeu de données NENCI-2021 [Sparrow *et al.*, 2021] pour lesquels le transfert des paramètres multipolaire a été réalisé. Les valeurs d'énergie sont données en kcal/mol. Les régressions linéaires sur les deux graphiques sont représentées par les droites grises. Les équations de ces deux régressions linéaires ainsi que les coefficients de détermination R^2 correspondant sont affichés en haut à gauche de chaque graphique. La corrélation ELMAM-SAPT est très forte pour l'énergie électrostatique tandis qu'elle est moins bonne pour l'énergie d'induction.

nature quantique des valeurs SAPT. Le coefficient de détermination est également un peu moins bon, certainement à cause du plus grand nombre de systèmes considérés, mais le RSMD reste très similaire avec une valeur de 1,42 kcal/mol. L'erreur absolue moyenne atteint quant à elle la précision chimique typiquement attendue avec $\text{MAE} = 1,02$ kcal/mol. Les différents indicateurs statistiques sont récapitulés dans le tableau 3.1.

Pour analyser l'influence de la composition du dimère sur la corrélation entre les valeurs ELMAM et SAPT, le coefficient de détermination R^2 a été calculé pour chacun des 77 dimères transférés de NENCI-2021 en utilisant les 45 géométries différentes disponibles pour chacun. Les résultats obtenus sont représentés sous la forme d'un histogramme sur la figure 3.2a. La plupart des dimères (44/77 soit 57,1%) présentent un excellent accord ELMAM-SAPT avec une valeur de R^2 comprise entre 0,90 et 0,95. Pour 18 complexes sur 77 (soit 23,4%), R^2 devient même supérieur à 0,95, les trois meilleures valeurs étant obtenues pour les complexes suivants : néopentane-pentane ($R^2 = 0,995$), pyridine-pyridine ($R^2 = 0,987$) et benzène-néopentane ($R^2 = 0,977$). Néanmoins, 4 dimères (soit 5,2%) ont une valeur de R^2 inférieure à 0,85. Il s'agit des complexes suivants : benzène-méthylamine ($R^2 = 0,847$), méthanethiol-méthanethiol ($R^2 = 0,841$), benzène-benzène ($R^2 = 0,827$) et benzène-acétamide ($R^2 = 0,801$). Pour l'homodimère du méthanethiol, l'erreur peut s'expliquer par la sous-représentation des types atomiques du soufre dans la librairie ELMAM2. Les autres complexes présentant tous une molécule de benzène, il est possible que la validité des paramètres multipolaires des types atomiques correspondant soient à améliorer.

Coefficients statistiques	U_{elst}	U_{ind}
R^2	0,941	0,820
R	0,970	0,905
MAE (kcal/mol)	1,02	0,65
RMSD (kcal/mol)	1,42	1,30

TABLEAU 3.1 – Coefficients statistiques pour caractériser l'accord entre le modèle ELMAM et la référence SAPT pour les énergies électrostatique et d'induction.

Les coefficients statistiques pour comparer les modèles ELMAM et SAPT de l'énergie d'interaction électrostatique (colonne U_{elst}) et de l'énergie d'induction (colonne U_{ind}) sont rassemblés ici. Ces indicateurs, dont les formulations sont données dans la section 3.1.2, sont : le coefficient de détermination R^2 , le coefficient de corrélation R , l'erreur absolue moyenne MAE et la déviation quadratique moyenne RMSD. Les coefficients R et R^2 confirment l'excellente corrélation entre les modèles ELMAM et SAPT de l'énergie électrostatique, et la moins bonne corrélation pour l'énergie d'induction. Les valeurs de MAE et RMSD sont en revanche plus faibles pour la contribution d'induction mais cela ne signifie pas nécessairement que les erreurs relatives sur ce terme sont plus faibles que sur le terme électrostatique car elles sont à contraster avec les ordres de grandeur plus faibles en valeurs absolues des énergies d'induction.

Ainsi, l'application au jeu de données NENCI-2021 confirme que le potentiel d'interaction électrostatique permanente, qui est issu des densités électroniques transférées non-perturbées d'origine expérimentale et sans introduction de paramètre empirique, produit des résultats présentant une forte corrélation avec les valeurs de référence SAPT, qui sont quant à elles calculées par méthodes purement théoriques.

Contribution d'induction dipolaire

Dans l'étude [Leduc *et al.*, 2019], le terme d'induction dipolaire ELMAM, $U_{\text{ind-dip}}^{\text{ELMAM}}$, défini comme la différence entre les énergies électrostatiques obtenues à partir de la densité électronique polarisée et à partir de la densité électronique non-perturbée transférée, a également été comparé à la contribution d'induction SAPT, $E_{\text{ind}}^{\text{SAPT}}$, dans les jeux de données S66 et S66x8. En particulier, la figure 6 de cet article montre la régression linéaire suivante :

$$U_{\text{ind-dip}}^{\text{ELMAM}} = 0,957 E_{\text{ind}}^{\text{SAPT}}, \quad (3.8)$$

avec le coefficient de détermination $R^2 = 0,880$ et RMSD = 1,6 kcal/mol. L'accord ELMAM-SAPT n'est clairement pas aussi bon que pour le terme électrostatique permanent. Les auteurs [Leduc *et al.*, 2019] expliquent ceci par le fait que le terme d'induction ELMAM ne tient compte que de la polarisation due aux moments dipolaires tandis que le terme SAPT couvre plus de contributions comme les transferts de charge et les effets de polarisation des moments électrostatiques d'ordres supérieurs. De plus, les polarisabilités utilisées sont des moyennes et la procédure d'induction de T. Leduc [Leduc *et al.*, 2019] ne tient compte que des effets de polarisation sur un atome dus au champ électrique généré par les moments dipolaires des atomes de l'autre molécule uniquement, plutôt que ceux de l'ensemble des atomes du système. Néanmoins la pente proche de l'unité montre tout de même que les ordres de grandeur obtenus sont satisfaisants étant donné les natures très différentes des méthodes de calcul SAPT, purement quantique, et le potentiel ELMAM pour lequel aucun paramètre empirique n'a été introduit.

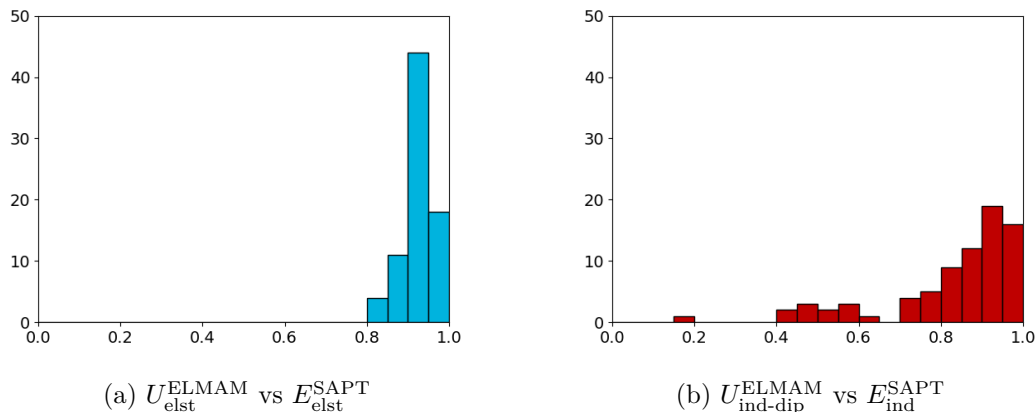


FIGURE 3.2 – Influence de la composition des dimères sur l'accord entre le modèle ELMAM et la référence SAPT.

Le coefficient de détermination R^2 entre les valeurs ELMAM et SAPT (a) de l'énergie électrostatique et (b) de l'énergie d'induction a été calculé séparément pour chacun des 77 dimères transférés du jeu de données NENCI-2021 en utilisant les 45 géométries par systèmes disponibles. Les résultats sont présentés sous la forme d'histogrammes montrant le nombre de dimères qui appartiennent à chaque intervalle de valeurs de R^2 . La hauteur d'une barre de graphique représente donc le nombre de complexes pour lesquels R^2 est compris entre deux valeurs séparées par pas de 0,05. Pour la contribution électrostatique, tous les dimères présentent une valeur de R^2 supérieure à 0,80, et même supérieure à 0,90 pour la plupart. En revanche, pour la contribution d'induction, bien que des valeurs de R^2 associées à de fortes corrélations soient obtenues pour une grande partie des complexes, certains présentent des valeurs correspondant à des corrélations faibles. Le modèle d'induction dipolaire n'est peut-être donc pas suffisant pour décrire les effets de polarisation de ces systèmes.

Pour le jeu de données NENCI-2021, les résultats du modèle d'énergie d'induction ELMAM (axe des ordonnées) en fonction des valeurs SAPT (axe des abscisses) sont présentés dans la figure 3.1b. La régression linéaire sur ces points a pour équation :

$$U_{\text{ind-dip}}^{\text{ELMAM}} = 1,007 E_{\text{ind}}^{\text{SAPT}}, \quad (3.9)$$

avec le coefficient de détermination $R^2 = 0,820$. Le graphique en nuages de points fait apparaître plusieurs branches qui correspondent à des dimères différents pour lesquels les effets de polarisation sont soit surestimés soit sous-estimés par le modèle ELMAM. La moins bonne corrélation ELMAM-SAPT pour le terme d'induction que pour le terme électrostatique permanent apparaît donc aussi dans le jeu de données NENCI-2021. Les valeurs de MAE = 0,65 kcal/mol et de RMSD = 1,30 kcal/mol sont plus faibles pour le terme d'induction que pour le terme électrostatique mais il faut contraster ces valeurs avec l'ordre de grandeur de cette contribution qui est plus faible en valeur absolue. Les coefficients statistiques sont regroupés dans le tableau 3.1.

Pour étudier l'influence de la composition des dimères sur la corrélation entre les termes d'induction ELMAM et SAPT, le coefficient de détermination R^2 a été calculé pour chacun des 77 dimères transférés et représenté sous la forme d'un histogramme sur la figure 3.2b. La majorité des complexes, 35 sur les 77 soit 45,5% d'entre eux, correspondent à une corrélation forte, c'est-à-dire à une valeur de R^2 supérieure à 0,90. Ces dimères présentent quasiment tous des interactions à dominante électrostatique (type ELEC défini par [Sparrow *et al.*, 2021]). Par exemple, les trois meilleures valeurs de R^2 sont obtenues pour les complexes suivants : pyridine-

éthyne ($R^2 = 0,988$), N-méthylacétamide-eau ($R^2 = 0,985$) et imidazole-eau ($R^2 = 0,979$). Une autre partie des dimères correspond à une corrélation ELMAM-SAPT moins forte, avec R^2 compris entre 0,70 et 0,90 pour 30 des 77 dimères (soit 39,0%). En revanche, une corrélation faible, avec R^2 compris entre 0,40 et 0,65, est obtenue pour 11 dimères (14,3%) ainsi qu’une corrélation négligeable avec $R^2 = 0,187$ pour le dimère pentane-pentane. Ces complexes associés à une corrélation faible sont quasiment tous dominés par des effets de dispersion (type DISP défini par [Sparrow *et al.*, 2021]). Aussi, l’induction dipolaire serait suffisante pour retrouver la majeure partie des effets de polarisation dans les systèmes de type ELEC mais insuffisante pour les systèmes de type DISP, pour lesquels des contributions supplémentaires, notamment les effets de polarisation quadripolaires et d’ordres supérieurs, seraient non-négligeables.

Pour compléter ces contributions électrostatique et d’induction dans la perspective de définir un potentiel d’interaction total ELMAM, j’ai développé les modèles d’énergie de dispersion et d’échange-répulsion reposant sur les densités électroniques transférées et sur les polarisabilités anisotropes atomiques fournies par la librairie ELMAM2. Ces modèles ont été entraînés et testés sur une sous-partie du jeu de données NENCI-2021. La partie suivante 3.2 présente les travaux réalisés pour obtenir le potentiel de dispersion tandis que la partie 3.3 détaillera les développements effectués pour le potentiel d’échange-répulsion. La combinaison de ces contributions, définissant le potentiel de van der Waals ELMAM, sera discutée dans la partie 3.4.

3.2 Potentiel de dispersion

3.2.1 Approximation de London de la dispersion

Les effets de dispersion apparaissent en raison des fluctuations dans les distributions électroniques atomiques qui font émerger des moments électrostatiques instantanés. Ces moments sont aléatoires mais leurs corrélations entre atomes voisins sont globalement stabilisantes pour le système. Ces interactions énergétiquement favorables existent dans tous les types de système et permettent d’expliquer les phases condensées des gaz rares. Pour définir la contribution de dispersion $U_{\text{disp}}^{\text{ELMAM}}$ du potentiel d’interaction ELMAM, je me suis basée sur l’approximation de London introduite dans la section 1.4.2. A l’origine, ce modèle a été développé par F. London [London, 1937] pour expliquer les interactions attractives entre atomes de gaz rares à partir de considérations de mécanique quantique. En utilisant l’hypothèse d’additivité de ces contributions [Yoffe et Maggiora, 1980], l’énergie de dispersion de London entre deux molécules A et B est donnée par :

$$U_{\text{disp, AB}}^{\text{London}} = -\frac{3}{2} \sum_{i \in A} \sum_{j \in B} \frac{I_i I_j}{I_i + I_j} \frac{\alpha_i \alpha_j}{R_{ij}^6}, \quad (3.10)$$

où I_i , α_i et I_j , α_j sont respectivement les énergies de première ionisation et les polarisabilités atomiques des atomes i appartenant à la molécule A et j appartenant à la molécule B . La distance R_{ij} est la distance entre les noyaux de ces deux atomes. Pour le potentiel ELMAM, $U_{\text{disp}}^{\text{ELMAM}}$, j’ai employé ce modèle en utilisant les énergies de première ionisation atomiques I disponibles

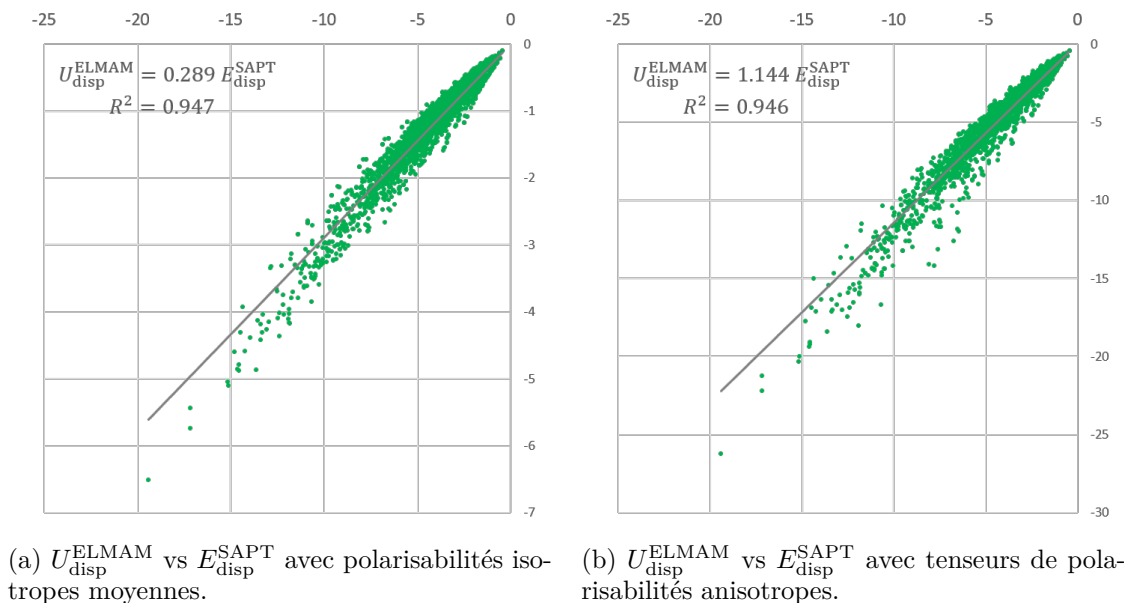


FIGURE 3.3 – Résultats obtenus par le modèle ELMAM de l'énergie de dispersion à partir des polarisabilités isotropes moyennes et des tenseurs de polarisabilités anisotropiques comparés aux valeurs de référence SAPT.

Ces graphiques représentent les valeurs du modèle ELMAM en ordonnées et les valeurs SAPT en abscisses de l'énergie d'interaction de dispersion sur le jeu de données NENCI-2021 [Sparrow *et al.*, 2021]. Le modèle ELMAM a été testé pour (a) des polarisabilités isotropes moyennes et pour (b) des tenseurs de polarisabilités anisotropes. Les valeurs d'énergie sont données en kcal/mol. Les régressions linéaires sur les deux graphiques sont représentées par les droites grises. Les équations de ces deux régressions linéaires ainsi que le coefficient de détermination correspondant sont affichés en haut à gauche de chaque graphique. La corrélation ELMAM-SAPT est très bonne pour les deux modèles bien que, dans le cas des polarisabilités isotropes moyennes, les énergies de dispersion sont sous-estimées, ce qui est indiqué par la faible valeur de la pente. Avec les tenseurs de polarisabilités anisotropes, les énergies obtenues sont plus proches des valeurs SAPT bien que cette fois-ci elles soient légèrement surestimées.

dans les tables² et les polarisabilités atomiques transférables α de la librairie ELMAM2.

3.2.2 Polarizabilités atomiques

Polarisabilités isotropes moyennes

Les polarisabilités atomiques anisotropes fournies par la librairie ELMAM2 sont des quantités tensorielles. Pour obtenir l'énergie $U_{\text{disp}}^{\text{ELMAM}}$, il faut que le produit $\alpha_i \alpha_j$ de l'équation 3.10 soit une grandeur scalaire. La première possibilité est de transformer les tenseurs de polarisabilité α en polarisabilités isotropes moyennes $\bar{\alpha}$ en prenant la moyenne des valeurs propres α_n de la matrice : $\bar{\alpha} = (\alpha_1 + \alpha_2 + \alpha_3)/3$. Le modèle de dispersion qui en découle est le suivant :

$$U_{\text{disp, AB}}^{\text{ELMAM}} = -\frac{3}{2} \sum_{i \in A} \sum_{j \in B} \frac{I_i I_j}{I_i + I_j} \frac{\bar{\alpha}_i \bar{\alpha}_j}{R_{ij}^6}, \quad (3.11)$$

où les indices i et j portent sur les atomes appartenant aux molécules A et B respectivement. J'ai appliqué ce potentiel sur l'ensemble des systèmes transférés du jeu de données NENCI-2021.

2. L'ouvrage [Lide, 2009] propose les valeurs d'énergies d'ionisations de tous les éléments chimiques dans les tableaux des pages 10-178 à 10-181.

Coefficients statistiques	Polarisabilités isotropes	Polarisabilités anisotropes
R^2	0,947	0,946
R	0,973	0,973
MAE (kcal/mol)	0,46	0,44
RMSD (kcal/mol)	0,59	0,59

TABLEAU 3.2 – Coefficients statistiques pour caractériser l'accord entre le modèle ELMAM et la référence SAPT pour les énergies de dispersion.

Les coefficients statistiques pour comparer les modèles ELMAM et SAPT de l'énergie de dispersion, dans le cas des polarisabilités isotropes moyennes (colonne « Polarisabilités isotropes ») et des tenseurs de polarisabilités anisotropes (colonne « Polarisabilités anisotropes ») sont rassemblés dans ce tableau. Ces indicateurs, dont les formulations sont données dans la section 3.1.2, sont : le coefficient de détermination R^2 , le coefficient de corrélation R , l'erreur absolue moyenne MAE et la déviation quadratique moyenne RMSD. Ces coefficients montrent un excellent accord ELMAM-SAPT, qui est quasiment identique pour les deux calculs du produit des polarisabilités.

La comparaison entre les résultats de ce potentiel et les valeurs de référence SAPT est donné sur la figure 3.3a. La régression linéaire obtenue est la suivante :

$$U_{\text{disp}}^{\text{ELMAM}} = 0,289 E_{\text{disp}}^{\text{SAPT}}, \quad (3.12)$$

avec le coefficient de détermination $R^2 = 0,947$. Cet accord est excellent mais la pente indique que les valeurs de dispersion sont largement sous-estimées par le modèle ELMAM. En effet, ce modèle est une approximation de la dispersion, la corrélation avec les valeurs SAPT est attendue mais pas l'accord quantitatif pour l'instant. J'ai donc introduit un facteur d'échelle K_{disp} sans dimension que j'ai affiné par minimisation des moindres carrés en utilisant l'ensemble des systèmes transférés du jeu de données NENCI-2021. Avec le facteur d'échelle obtenu $K_{\text{disp}} = 3,41$, la régression linéaire devient :

$$U_{\text{disp}}^{\text{ELMAM}} = 0,985 E_{\text{disp}}^{\text{SAPT}}, \quad (3.13)$$

et le coefficient de détermination R^2 étant inchangé. L'erreur absolue moyenne sur ces contributions est MAE = 0,46 kcal/mol et le RMSD = 0,59 kcal/mol, en-dessous de la précision chimique attendue. Ces coefficients statistiques sont regroupés dans le tableau 3.2.

Tenseurs de polarisabilité anisotrope

Une autre possibilité est de conserver les polarisabilités sous forme tensorielle et de prendre la trace de leur produit matriciel :

$$U_{\text{disp, AB}}^{\text{ELMAM}} = -\frac{3}{2} \sum_{i \in A} \sum_{j \in B} \frac{I_i I_j}{I_i + I_j} \frac{\text{Tr}(\boldsymbol{\alpha}_i^t \cdot \boldsymbol{\alpha}_j)}{R_{ij}^6}, \quad (3.14)$$

où l'exposant t signifie la transposition de la matrice $\boldsymbol{\alpha}_i$. Les résultats de ce modèle sont comparés aux valeurs SAPT dans la figure 3.3b. La régression linéaire sur ces points est :

$$U_{\text{disp}}^{\text{ELMAM}} = 1,144 E_{\text{disp}}^{\text{SAPT}}, \quad (3.15)$$

C_H	C_C	C_N	C_O	C_S
1,06	1,01	0,86	1,00	1,05

TABLEAU 3.3 – Paramètres empiriques dépendant de l'espèce chimique introduits dans le modèle ELMAM d'énergie de dispersion.

Les paramètres empiriques C_H , C_C , C_N , C_O et C_S ont été introduits comme coefficients devant les polarisabilités pour chaque espèce chimique (hydrogène, carbone, azote, oxygène et soufre) dans le modèle de London (équation 3.17). Ils ont été affinés par minimisation des moindres carrés contre les valeurs SAPT pour 80% des systèmes transférés du jeu de données NENCI-2021. Ils restent tous proches de l'unité, excepté C_N qui suggère que les polarisabilités ELMAM pour les atomes d'azote sont légèrement surestimées.

avec le coefficient de détermination $R^2 = 0,946$. Ce coefficient de détermination est quasiment le même qu'en utilisant les polarisabilités isotropes moyennes mais la pente est quant à elle plus proche de l'unité. Les valeurs de dispersion sont tout de même légèrement surestimées dans ce modèle, c'est pourquoi j'ai introduit le facteur d'échelle $K_{\text{disp}} = 0,86$, obtenu par affinement des moindres carrés sur l'ensemble des systèmes transférés. La nouvelle régression linéaire a pour équation :

$$U_{\text{disp}}^{\text{ELMAM}} = 0,983 E_{\text{disp}}^{\text{SAPT}}. \quad (3.16)$$

Les valeurs de MAE et de RMSD, qui sont regroupées dans le tableau 3.2, restent inchangées par rapport au modèle avec les polarisabilités isotropes. Pour la suite du développement du potentiel de dispersion ELMAM, j'ai utilisé le modèle qui conserve les polarisabilités sous forme tensorielle (équation 3.14).

3.2.3 Introduction de paramètres atomiques empiriques

Les résultats du modèle de dispersion sont déjà très satisfaisants mais pour tenter de les améliorer davantage, j'ai introduit des paramètres empiriques C_i dépendant de l'espèce chimique de l'atome i (hydrogène, carbone, azote, oxygène ou soufre). Les polarisabilités atomiques α utilisées étant des polarisabilités moyennées et transférées, ces paramètres empiriques ont été employés comme coefficients de mise à l'échelle ou "scaling" des α_i :

$$U_{\text{disp, AB}}^{\text{ELMAM}} = -\frac{3}{2} K_{\text{disp}} \sum_{i \in A} \sum_{j \in B} \frac{I_i I_j}{I_i + I_j} \frac{\text{Tr}(C_i \alpha_i^t \cdot C_j \alpha_j)}{R_{ij}^6}. \quad (3.17)$$

J'ai affiné ces paramètres C_i par minimisation des moindres carrés à partir de 80% des systèmes transférés du jeu de données NENCI-2021. Les 20% de systèmes restant ont été utilisés pour la validation croisée du modèle avec les paramètres empiriques obtenus. Ces paramètres sont regroupés dans le tableau 3.3. A part pour l'azote ($C_N = 0,86$), ces coefficients restent très proches de 1. Ces résultats confirment la qualité des polarisabilités de la librairie ELMAM2, sauf peut-être pour les types atomiques de l'azote pour lesquels elles seraient légèrement surestimées. Le graphique 3.4a représente les résultats obtenus par ce modèle en fonction des valeurs SAPT pour les 20% des systèmes transférés qui n'ont pas été utilisés pour entraîner ces paramètres. La régression linéaire donne :

$$U_{\text{disp}}^{\text{ELMAM}} = 0,989 E_{\text{disp}}^{\text{SAPT}}, \quad (3.18)$$

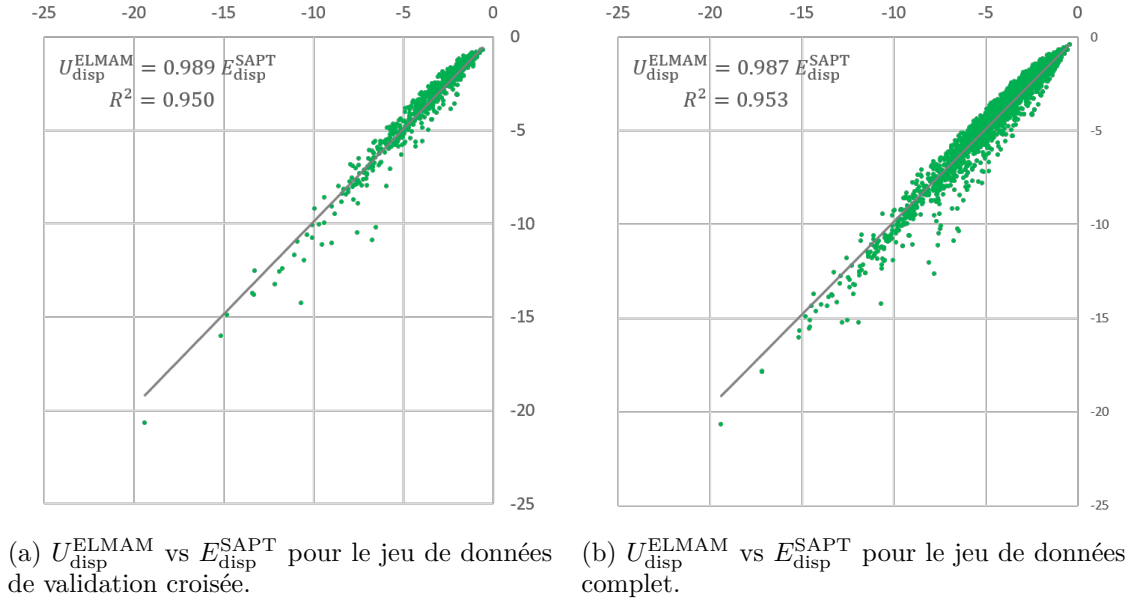


FIGURE 3.4 – Résultats obtenus par le modèle ELMAM de l'énergie de dispersion avec les paramètres empiriques comparés aux valeurs de référence SAPT.

Ces graphiques représentent les valeurs du modèle ELMAM comportant les paramètres empiriques en ordonnées et les valeurs SAPT en abscisses de l'énergie d'interaction, pour (a) les 20% des systèmes transférés du jeu de données NENCI-2021 [Sparrow *et al.*, 2021] qui n'ont pas été utilisés pour entraîner ces paramètres et (b) sur l'ensemble des systèmes transférés. Les valeurs d'énergie sont données en kcal/mol. Les régressions linéaires sur les deux graphiques sont représentées par les droites grises. Les équations de ces deux régressions linéaires ainsi que le coefficient de détermination correspondant sont affichés en haut à gauche de chaque graphique. La corrélation ELMAM-SAPT est légèrement améliorée par l'introduction des paramètres empiriques, ce qui est confirmé par la validation croisée sur les systèmes non-utilisés pour leur entraînement.

avec $R^2 = 0,950$. Les valeurs de $\text{MAE} = 0,39$ kcal/mol et de $\text{RMSD} = 0,55$ kcal/mol sont légèrement meilleures que sans les paramètres empiriques, comme attendu. Pour le jeu de données complet des systèmes transférés, le graphique 3.4b présente la comparaison entre le modèle paramétré et les valeurs SAPT. La régression linéaire associée est :

$$U_{\text{disp}}^{\text{ELMAM}} = 0,987 E_{\text{disp}}^{\text{SAPT}}, \quad (3.19)$$

avec $R^2 = 0,953$. Les valeurs de MAE et de RMSD, données dans le tableau 3.4, sont les mêmes que pour le jeu de données de validation croisée. Finalement, l'introduction de ces paramètres empiriques montre certes une amélioration mais peu significative par rapport au modèle non-paramétré et ne sont peut-être pas nécessaires pour la définition du modèle de dispersion. L'erreur persistante par rapport aux valeurs SAPT doit être analysée plus en profondeur.

3.2.4 Influence de la composition du dimère

Comme pour les termes électrostatique et d'induction, j'ai calculé pour la contribution de dispersion le coefficient de détermination R^2 entre les résultats ELMAM et les valeurs SAPT pour chacun des 77 dimères transférés, en utilisant leurs 45 géométries différentes. L'histogramme présentant ces résultats est donné sur la figure 3.5. Une large majorité des dimères, 68 sur les

Coefficients statistiques	Validation croisée	Jeu de données complet
R^2	0,950	0,953
R	0,975	0,976
MAE (kcal/mol)	0,39	0,39
RMSD (kcal/mol)	0,55	0,55

TABLEAU 3.4 – Coefficients statistiques pour caractériser l'accord entre le modèle ELMAM avec paramètres empiriques et la référence SAPT pour les énergies de dispersion.

Les coefficients statistiques pour comparer les modèles ELMAM avec paramètres empiriques et SAPT de l'énergie de dispersion, pour les systèmes non-utilisés pour l'entraînement des paramètres (colonne « Validation croisée ») et pour l'ensemble des systèmes transférés du jeu de données NENCI-2021 (colonne « Jeu de données complet ») sont rassemblés dans ce tableau. Ces indicateurs, dont les formulations sont données dans la section 3.1.2, sont : le coefficient de détermination R^2 , le coefficient de corrélation R , l'erreur absolue moyenne MAE et la déviation quadratique moyenne RMSD. Ces coefficients confirment la validation croisée car l'accord reste aussi bon pour les systèmes qui n'ont pas servi à l'entraînement des paramètres empiriques que pour l'ensemble des systèmes.

77 (soit 88,3%), sont associés à une excellente valeur de R^2 supérieure à 0,95, ce qui démontre la pertinence des choix de modélisation. Globalement, les dimères de type DISP présentent de meilleures valeurs de R^2 pour la contribution de dispersion que ceux de type ELEC. En effet, sur les 32 dimères à dominante DISP, 29 (90,6%) présentent une valeur de R^2 supérieure à 0,99 et la moins bonne corrélation est obtenue pour le complexe méthane-thiol-méthane-thiol avec $R^2 = 0,941$. Les trois meilleures valeurs de R^2 sont obtenues pour les complexes suivant : benzène-éthène ($R^2 = 0,999$), benzène-éthyne ($R^2 = 0,999$) et pyridine-uracile ($R^2 = 0,998$). Pour les 36 dimères à dominante ELEC, la plupart (30 complexes soit 83,3%) ont une valeur de R^2 inférieure à 0,99, la meilleure corrélation étant obtenue pour le complexe imidazole-eau avec $R^2 = 0,996$. Deux complexes (soit 2,6% de tous les dimères transférés) ont un R^2 inférieur à 0,90 : acide acétique-acide acétique ($R^2 = 0,896$) et acide acétique-uracile ($R^2 = 0,886$). Ces deux systèmes comportent la molécule d'acide acétique qui est la seule parmi toutes les molécules des systèmes transférés à posséder le groupement acide carboxylique (COOH). Cette molécule est également présente dans plusieurs autres hétérodimères mais ces deux complexes sont les seuls dans lesquels ce groupement COOH est impliqué dans une double liaison hydrogène. Il est possible que les polarisabilités transférées depuis la librairie ELMAM2 pour les types atomiques de ce groupement ne soient pas adaptées pour ce type d'interaction (ou « synthon ») bien particulier.

3.2.5 Autres modèles d'énergie de dispersion

Le potentiel de dispersion que j'ai développé à partir de l'approximation de London fournit déjà d'excellents résultats malgré sa relative simplicité. D'autres modèles plus élaborés pourront également être testés dans de futurs travaux. Par exemple, le modèle de Slater-Kirkwood [Slater et Kirkwood, 1931, Aquilanti *et al.*, 1996] est assez semblable au modèle de London mais fait intervenir, en plus des polarisabilités atomiques α_i , le « nombre d'électrons effectif » N_i de l'atome i :

$$U_{\text{disp, AB}}^{\text{S-K}} = -\frac{3}{2} \sum_{i \in A} \sum_{j \in B} \left(\sqrt{\frac{\alpha_i}{N_i}} + \sqrt{\frac{\alpha_j}{N_j}} \right)^{-1} \frac{\alpha_i \alpha_j}{d_{ij}^6}. \quad (3.20)$$

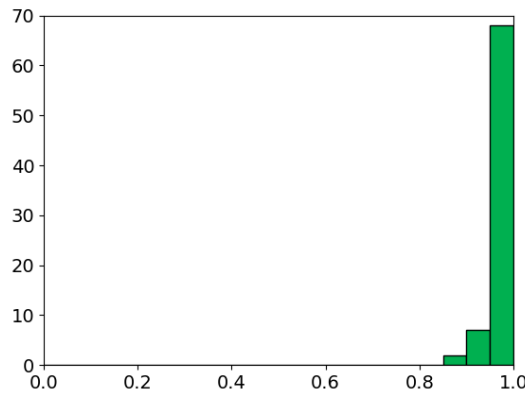


FIGURE 3.5 – Influence de la composition des dimères sur l'accord entre le modèle de dispersion ELMAM et la référence SAPT.

Le coefficient de détermination R^2 entre les valeurs ELMAM et SAPT de l'énergie de dispersion a été calculé séparément pour chacun des 77 dimères transférés du jeu de données NENCI-2021 en utilisant les 45 géométries par système disponibles. Les résultats sont présentés sous la forme d'un histogramme montrant le nombre de dimères qui appartiennent à chaque intervalle de valeurs de R^2 . La hauteur d'une barre de graphique représente donc le nombre de complexes pour lesquels R^2 est compris entre deux valeurs séparées par pas de 0,05. La valeur de R^2 est supérieure à 0,95 pour une large majorité des dimères, confirmant l'excellent accord avec la référence SAPT.

Ce nombre d'électrons effectif N_i pourrait être évalué par l'intégration de la densité électronique dans le bassin atomique de l'atome i , défini dans la section 1.3.1 ou sur la base des paramètres P_{val} du modèle multipolaire.

Le modèle "Exchange-hole" de A. Becke et E. Johnson [Becke et Johnson, 2005, Becke et Johnson, 2007, Ángyán, 2007], qui est plutôt utilisé pour les calculs de DFT, pourrait également servir d'inspiration pour développer un nouveau modèle de dispersion à partir des quantités transférables de la librairie ELMAM2.

3.3 Potentiel d'échange-répulsion

3.3.1 Modèle de recouvrement des densités électroniques

Recouvrement des orbitales et recouvrement des densités électroniques

La répulsion de Pauli ou échange-répulsion, parfois simplement appelée répulsion, prend son origine dans la nature quantique des électrons. En effet, d'après le principe d'exclusion de Pauli, puisque les électrons sont des particules fermioniques, deux électrons ne peuvent pas occuper le même état et lorsque deux atomes sont suffisamment proches pour interagir ensemble, la fonction d'onde du système doit rester antisymétrique. Comme très clairement expliqué dans l'article de J. A. Rackers et J. W. Ponder de 2019 [Rackers et Ponder, 2019], cette condition induit une variation $\delta\rho(\mathbf{r})$ dans la densité électronique par rapport à $\rho(\mathbf{r})$ non-perturbée qui a deux conséquences : une déplétion d'électrons dans la région de recouvrement entre les deux noyaux et une accumulation de $\rho(\mathbf{r})$ à proximité des noyaux. La déplétion dans la région internucléaire provoque une diminution de l'écrantage électronique des noyaux et donc une augmentation

de la répulsion noyau-noyau. L'accumulation de densité électronique proche des noyaux est quant à elle favorable du point de vue de l'énergie électrostatique. Néanmoins, l'énergie de répulsion due à la déplétion est plus importante que la baisse induite par l'accumulation [Salem, 1961, Rackers et Ponder, 2019]. L'effet du principe d'exclusion se traduit donc globalement par une augmentation de l'énergie du système. L'échange-répulsion peut donc être comprise comme la conséquence de l'application de la loi de Coulomb sur la variation de densité électronique due à l'antisymétrisation de la fonction d'onde.

Dans les champs de force classiques, cette interaction est généralement modélisée par des potentiels paramétrés, tels que les potentiels de Lennard-Jones, de Buckingham et de Halgren mentionnés dans la section 1.4.2, qui ne rendent pas compte de cette interprétation physique de la répulsion. En 1961, L. Salem a introduit le modèle dit de recouvrement des orbitales ("orbital overlap") de la répulsion de Pauli [Salem, 1961]. Il a montré que l'énergie d'échange-répulsion peut être modélisée de façon très précise par la quantité S^2/R , où S est l'intégrale du produit des orbitales atomiques et R est la distance entre les atomes. Néanmoins, ces orbitales étant rarement définies dans les champs de force, ce modèle n'a pas rencontré un succès massif. Par exemple, il est employé dans les champs de force SIBFA ("Sum of Interacting Fragment *ab initio*") [Piquemal *et al.*, 2006, Piquemal *et al.*, 2007] et EFP ("Effective Fragment Potential") [Gordon, 1996, Jensen et Gordon, 1998] où le recouvrement des orbitales est explicitement calculé. Ces modèles sont très coûteux en temps de calcul et ne sont donc pas applicables aux larges systèmes tels que les protéines. Les équipes de S. Kita [Kita *et al.*, 1976] et de Y. S. Kim [Kim *et al.*, 1981] ont proposé un autre modèle semi-classique de l'échange-répulsion qui est quant à lui basé sur le recouvrement des densités électroniques Ω ("density overlap") :

$$U_{\text{exch-rep}} = K_{\text{exch-rep}} \Omega = K_{\text{exch-rep}} \int \rho_1(\mathbf{r})\rho_2(\mathbf{r})d^3\mathbf{r}, \quad (3.21)$$

où $K_{\text{exch-rep}}$ est une constante qui permet de tenir compte de la différence de dimensions entre l'énergie $U_{\text{exch-rep}}$ et l'intégrale du produit des densités électroniques Ω . Pour deux molécules A et B , le recouvrement des densités électroniques moléculaires peut être exprimé par la sommation des recouvrements atome-atome :

$$U_{\text{exch-rep},AB} = K_{\text{exch-rep}} \sum_{i \in A} \sum_{j \in B} \Omega_{ij} = K_{\text{exch-rep}} \sum_{i \in A} \sum_{j \in B} \int \rho_i(\mathbf{r})\rho_j(\mathbf{r})d^3\mathbf{r}, \quad (3.22)$$

où $\rho_i(\mathbf{r})$ et $\rho_j(\mathbf{r})$ sont respectivement les densités électroniques atomiques de l'atome i appartenant à la molécule A et de l'atome j appartenant à la molécule B . Ce modèle est utilisé dans les champs de force GEM* ("Gaussian Electrostatic Model") [Duke *et al.*, 2014] et MASTIFF ("Multipolar Anisotropic Slater-Type Intermolecular Force Field") [Van Vleet *et al.*, 2018] qui utilisent les distributions de charges ponctuelles et de dipôles AMOEBA pour l'un et un modèle de densité électronique paramétré à symétrie sphérique pour l'autre. Le modèle de recouvrement des densités fournit des résultats relativement similaires au modèle du recouvrement des orbitales [Söderhjelm *et al.*, 2006, Rackers et Ponder, 2019]. Il peut être optimisé en introduisant un facteur $f(R_{ij})$ dépendant de la distance interatomique R_{ij} [Söderhjelm *et al.*, 2006] ou un exposant γ tel que $U_{\text{exch-rep}} = K_{\text{exch-rep}}\Omega^\gamma$ [Misquitta et Stone, 2016].

Unité de la constante de mise à l'échelle $K_{\text{exch-rep}}$

Le potentiel d'échange-répulsion $U_{\text{exch-rep}}$ est homogène à une énergie qui est généralement exprimée par les chimistes et biologistes en kcal/mol. Le recouvrement des densités électroniques Ω n'est pas homogène à une énergie mais au produit des densités multiplié par l'élément d'intégration. Les densités électroniques sont des charges par unité de volume, souvent exprimées en $\text{e} \cdot \text{\AA}^{-3}$, et pour une intégrale en trois dimensions, l'élément d'intégration est homogène à un volume qui peut être donné en $\text{\AA}^3$. Le recouvrement Ω serait alors exprimé en $(\text{e} \cdot \text{\AA}^{-3})^2 \times \text{\AA}^3$, soit $\text{e}^2 \cdot \text{\AA}^{-3}$. Une unité arbitraire (UA) pour $K_{\text{exch-rep}}$ pourrait donc être : $\text{kcal} \cdot \text{mol}^{-1} \cdot \text{e}^2 \cdot \text{\AA}^{-3}$. Sachant que $1 \text{ kcal} = 4,187 \text{ J}$, $1 \text{ e} = 1,602 \cdot 10^{-19} \text{ C}$ et $1 \text{ \AA} = 10^{-10} \text{ m}$, $K_{\text{exch-rep}}$ devient en unités du système international :

$$K_{\text{exch-rep}} = K_{\text{exch-rep}}^{\text{UA}} \times 1,632 \cdot 10^{11} \text{ J} \cdot \text{mol}^{-1} \cdot \text{m}^3 \cdot \text{C}^{-2}. \quad (3.23)$$

Or, cette unité n'est pas sans rappeler celle de la constante de Coulomb :

$$\frac{1}{4\pi\epsilon_0} = 8,988 \cdot 10^9 \text{ J} \cdot \text{m} \cdot \text{C}^{-2}. \quad (3.24)$$

Par conséquent, il est possible de choisir un facteur d'échelle d'échange-répulsion réduit $k_{\text{exch-rep}}$ tel que :

$$k_{\text{exch-rep}} = 4\pi\epsilon_0 K_{\text{exch-rep}} = K_{\text{exch-rep}}^{\text{UA}} \times 1,816 \text{ m}^2 \cdot \text{mol}^{-1}. \quad (3.25)$$

La dimension de ce facteur est difficile à rattacher à une grandeur physique interprétable mais permet d'utiliser le modèle de recouvrement des densités qui lui possède un sens physique. Notons par ailleurs que la dimension de $k_{\text{exch-rep}}$ dépend de celle du recouvrement Ω qui peut être amenée à changer selon sa méthode de calcul.

Potentiel d'échange-répulsion ELMAM

Pour le terme d'échange-répulsion $U_{\text{exch-rep}}^{\text{ELMAM}}$ du potentiel d'interaction ELMAM, j'ai utilisé le modèle de recouvrement basé sur les densités électroniques atomiques $\rho(\mathbf{r})$ reconstruites par transfert des paramètres multipolaires de la librairie ELMAM2. En utilisant le facteur d'échelle réduit $k_{\text{exch-rep}}$, l'expression du potentiel ELMAM est la suivante :

$$U_{\text{exch-rep},AB}^{\text{ELMAM}} = \frac{k_{\text{exch-rep}}}{4\pi\epsilon_0} \sum_{i \in A} \sum_{j \in B} \int \rho_i(\mathbf{r}) \rho_j(\mathbf{r}) d^3\mathbf{r}. \quad (3.26)$$

Pour obtenir les meilleurs résultats possibles, j'ai essayé plusieurs méthodes d'intégration, plusieurs facteurs dépendant de la distance interatomique R_{ij} et aussi l'introduction de paramètres empiriques $\{C_i, C_j\}$ qui seront détaillés dans les sections suivantes.

3.3.2 Intégration du produit des densités électroniques

Le recouvrement des densités électroniques atomiques Ω_{ij} est défini comme l'intégrale du produit des distributions des électrons des atomes i et j . Or, le modèle multipolaire, qui est utilisé pour exprimer les densités électroniques atomiques reconstruites par le transfert des paramètres

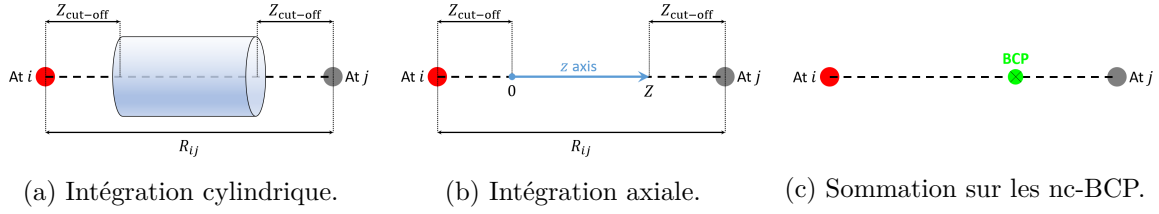


FIGURE 3.6 – Schéma des différentes méthodes d'intégration du produit des densités électroniques pour le calcul du potentiel d'échange-répulsion ELMAM.

Les trois différentes méthodes d'intégration du produit des densités électroniques qui ont été testées pour définir le potentiel d'échange-répulsion ELMAM sont : (a) l'intégration dans un volume cylindrique, (b) l'intégration le long de l'axe internucléaire, et (c) la sommation sur les points critiques de liaison non-covalente (nc-BCP) de la densité électronique totale. (a) Le volume cylindrique défini entre les atomes $At\ i$ et $At\ j$ est de hauteur $R_{ij} - 2Z_{\text{cut-off}}$ et sa base est rayon égal au plus grand des rayons de van der Waals des deux atomes. (b) L'intégration axiale est réalisée dans la direction de l'axe z qui joint les noyaux des atomes $At\ i$ et $At\ j$ et sur une longueur $Z = R_{ij} - 2Z_{\text{cut-off}}$. (c) Le produit des densités électroniques atomiques est évalué sur le point critique nc-BCP de la densité électronique moléculaire totale, défini dans la section 1.3.1.

ELMAM2, repose sur des fonctions de type Slater et harmoniques sphériques qui ne sont jamais exactement nulles même à longue distance des noyaux. Aussi, le produit $\rho_i(\mathbf{r})\rho_j(\mathbf{r})$ n'est pas exactement nul même en dehors de la région de recouvrement des densités. Une intégration dans un large volume entourant les molécules engendrerait donc une erreur due à la sommation des densités résiduelles. C'est pourquoi j'ai testé différentes méthodes d'intégration, permettant également de réduire le temps de calcul, en se concentrant sur la région internucléaire où se produit le phénomène de recouvrement.

Intégration dans un volume cylindrique

Pour se focaliser sur la région internucléaire, la première approche que j'ai employée consiste à intégrer dans un volume cylindrique. Pour chaque paire atome i - atome j , un cylindre est défini autour de l'axe joignant les deux noyaux, comme illustré sur la figure 3.6a. Sa hauteur z est égale à la distance interatomique R_{ij} tronquée par une longueur de coupure $Z_{\text{cut-off}}$ permettant de ne pas tenir compte des électrons de cœur : $Z = R_{ij} - 2Z_{\text{cut-off}}$, et son rayon $\mathcal{R}$ est égal au plus grand des rayons de van der Waals des deux atomes. L'intégrale du recouvrement des densités électroniques $\Omega_{\text{cyl},ij}$ est exprimée en coordonnées cylindrique (r, θ, z) :

$$\Omega_{\text{cyl},ij} = \int_0^{\mathcal{R}} \int_0^{2\pi} \int_0^Z \rho_i(r, \theta, z) \rho_j(r, \theta, z) r dr d\theta dz. \quad (3.27)$$

Le potentiel d'interaction d'échange-répulsion ELMAM a pour expression :

$$U_{\text{exch-rep},AB}^{\text{ELMAM}} = \frac{k_{\text{exch-rep}}}{4\pi\epsilon_0} \Omega_{\text{cyl}} = \frac{k_{\text{exch-rep}}}{4\pi\epsilon_0} \sum_{i \in A} \sum_{j \in B} \Omega_{\text{cyl},ij}. \quad (3.28)$$

Pour déterminer la valeur de $k_{\text{exch-rep}}$, j'ai d'abord comparé directement le recouvrement Ω_{cyl} aux valeurs de référence SAPT, $E_{\text{exch-rep}}^{\text{SAPT}}$, sur l'ensemble des systèmes transférés du jeu de données

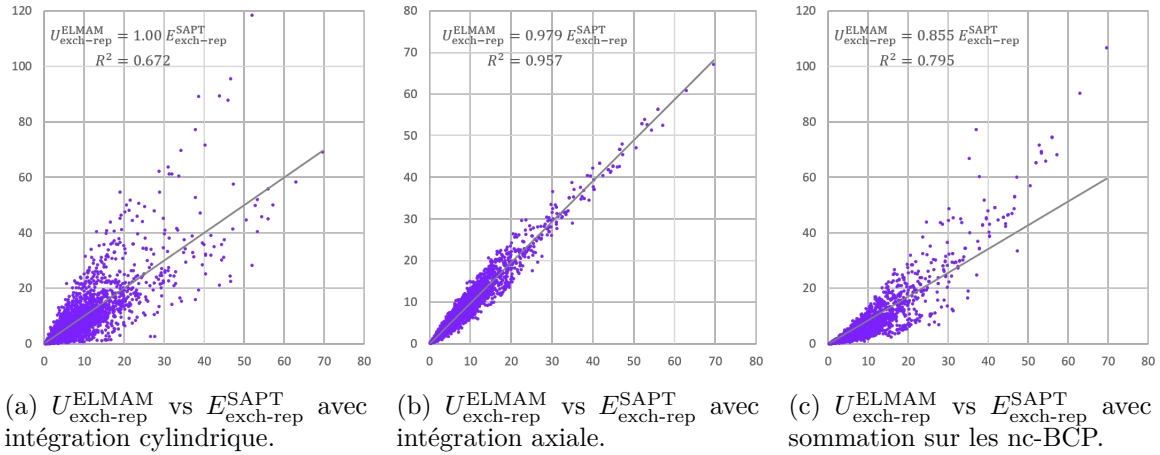


FIGURE 3.7 – Résultats obtenus par le modèle ELMAM de l'énergie d'échange-répulsion pour différentes méthodes d'intégration du produit des densités, comparés aux valeurs de référence SAPT.

Ces graphiques représentent les valeurs du modèle ELMAM en ordonnées et les valeurs SAPT en abscisses de l'énergie d'interaction d'échange-répulsion sur le jeu de données NENCI-2021 [Sparrow *et al.*, 2021]. Le modèle ELMAM a été testé pour le recouvrement des densités électroniques atomiques obtenu (a) par intégration dans un volume cylindrique, (b) par intégration le long de l'axe internucléaire, et (c) par sommation sur les points critiques de liaison non-covalente (nc-BCP) de la densité électronique moléculaire totale. Les valeurs d'énergie sont données en kcal/mol. Les régressions linéaires sur les trois graphiques sont représentées par les droites grises. Les équations de ces trois régressions linéaires ainsi que les coefficients de détermination correspondant sont affichés en haut à gauche de chaque graphique. La corrélation ELMAM-SAPT est excellente pour la méthode d'intégration axiale (b) mais nettement moins bonne pour les deux autres méthodes. L'intégration cylindrique (a) contient le bruit dans les régions où les densités ne recouvrent pas et la sommation sur les nc-BCP (c) manque des informations.

NENCI2021. La régression linéaire obtenue a pour équation :

$$\Omega_{\text{cyl}} = 6,32 \cdot 10^{-4} E_{\text{exch-rep}}^{\text{SAPT}}, \quad (3.29)$$

avec le coefficient de détermination $R^2 = 0,672$. La linéarité de la relation entre Ω_{cyl} et $E_{\text{exch-rep}}^{\text{SAPT}}$ n'étant pas très bonne, la valeur du facteur d'échelle $k_{\text{exch-rep}}$ n'a pas pu être affinée par la méthode des moindres carrés. J'ai donc choisi de prendre l'inverse de la pente de la régression linéaire : $k_{\text{exch-rep}} = 2,87 \cdot 10^3 \text{ m}^2 \cdot \text{mol}^{-1}$. Avec ce facteur d'échelle, les résultats du modèle $U_{\text{exch-rep},AB}^{\text{ELMAM}}$ ont été confrontés aux valeurs de référence $E_{\text{exch-rep}}^{\text{SAPT}}$ dans la figure 3.7a. La régression linéaire de ces points a pour équation :

$$U_{\text{exch-rep}}^{\text{ELMAM}} = 1,00 E_{\text{exch-rep}}^{\text{SAPT}}, \quad (3.30)$$

avec le coefficient de détermination $R^2 = 0,672$ caractérisant une faible corrélation ELMAM-SAPT. Les valeurs de MAE = 2,89 kcal/mol et de RMSD = 5,27 kcal/mol confirment l'erreur importante commise en utilisant la méthode d'intégration cylindrique. Ces coefficients statistiques, qui sont regroupés dans le tableau 3.5, ont été obtenus pour $Z_{\text{cut-off}} = 0,4 \text{ \AA}$, la valeur R^2 étant encore plus faible pour une coupure plus petite et similaire pour une coupure plus grande. Comme cette mauvaise corrélation pouvait provenir de l'intégration de densités résiduelles dans le volume, un modèle encore plus ciblé sur la région internucléaire a donc été tenté.

Intégration le long de la direction internucléaire

Afin de focaliser davantage sur la région internucléaire, j'ai utilisé une intégration unidimensionnelle, le long de l'axe z reliant les deux noyaux des atomes i et j , illustrée par la figure 3.6b. Comme précédemment, la longueur de l'intervalle d'intégration $\mathcal{Z}$ est définie par la distance interatomique R_{ij} tronquée par une longueur de coupure $Z_{\text{cut-off}}$ à chaque extrémité : $\mathcal{Z} = R_{ij} - 2Z_{\text{cut-off}}$. Le recouvrement des densités est alors exprimé par l'intégration axiale suivante :

$$\Omega_{\text{axe},ij} = \int_0^{\mathcal{Z}} \rho_i(\mathbf{r})\rho_j(\mathbf{r})dz, \quad (3.31)$$

où la coordonnée z est ici définie par rapport à l'axe d'intégration et ne correspond pas à une composante du vecteur position $\mathbf{r}$ qui lui est défini par rapport au système de coordonnées global du système. Le potentiel d'interaction $U_{\text{exch-rep}}^{\text{ELMAM}}$ a pour expression :

$$U_{\text{exch-rep},AB}^{\text{ELMAM}} = \frac{k_{\text{exch-rep}}}{4\pi\epsilon_0} \Omega_{\text{axe}} = \frac{k_{\text{exch-rep}}}{4\pi\epsilon_0} \sum_{i \in A} \sum_{j \in B} \Omega_{\text{axe},ij}. \quad (3.32)$$

La comparaison directe entre le recouvrement Ω_{axe} et les valeurs de référence $E_{\text{exch-rep}}^{\text{SAPT}}$ a permis d'affiner par minimisation des moindres carrés le facteur d'échelle³ $k_{\text{exch-rep}} = 2,00 \cdot 10^{-16} \text{ m}^4 \cdot \text{mol}^{-1}$. Les résultats du modèle $U_{\text{exch-rep}}^{\text{ELMAM}}$ en fonction des valeurs SAPT sont présentés sur le graphique 3.7b. La régression linéaire de ces points est :

$$U_{\text{exch-rep}}^{\text{ELMAM}} = 0,979 E_{\text{exch-rep}}^{\text{SAPT}}, \quad (3.33)$$

avec un coefficient de détermination très satisfaisant $R^2 = 0,957$. Les valeurs de MAE = 1,10 kcal/mol et de RMSD = 1,55 kcal/mol se rapprochent des résultats obtenus pour le modèle d'énergie électrostatique $U_{\text{elst}}^{\text{ELMAM}}$, présentés dans la section 3.1.3, dont les ordres de grandeur d'énergies sont similaires à ce terme de répulsion.

Le bon accord ELMAM-SAPT a ici été obtenu grâce à une valeur de coupure de l'intervalle d'intégration : $Z_{\text{cut-off}} = 0,6 \text{ \AA}$. Comme le montre le graphique 3.8, cet accord est nettement moins bon lorsque l'intervalle d'intégration n'est pas suffisamment tronqué ($Z_{\text{cut-off}} < 0,1 \text{ \AA}$) car les électrons de cœur sont pris en compte. De même, lorsque cet intervalle est trop court ($Z_{\text{cut-off}} > 0,7 \text{ \AA}$), une partie du recouvrement des densités électroniques atomiques est omis, ce qui fait baisser la qualité du modèle. Entre $Z_{\text{cut-off}} = 0,1 \text{ \AA}$ et $Z_{\text{cut-off}} = 0,7 \text{ \AA}$, un plateau est observé et l'accord ELMAM-SAPT est maximal autour de $Z_{\text{cut-off}} = 0,6 \text{ \AA}$. C'est pourquoi cette valeur de coupure a été choisie dans mon modèle pour la présentation des résultats. Notons ici que le ciblage de la région où se produit le recouvrement des densités est déterminant pour la qualité du modèle.

3. Avec cette méthode d'intégration, le recouvrement Ω n'est plus exprimé en $\text{e}^2 \cdot \text{\AA}^{-3}$ mais en $\text{e}^2 \cdot \text{\AA}^{-5}$, ce qui a pour conséquence de changer également l'unité de $k_{\text{exch-rep}}$.

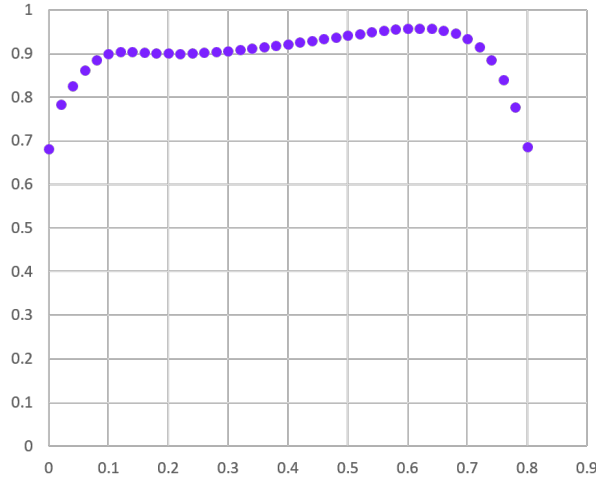


FIGURE 3.8 – Dépendance de l'accord ELMAM-SAPT pour le potentiel d'échange-répulsion (R^2 en ordonnées) en fonction de l'intervalle d'intégration choisi ($Z_{\text{cut-off}}$ en abscisses).

Le coefficient de détermination R^2 entre les valeurs ELMAM et SAPT de l'énergie d'échange-répulsion a été calculé sur le jeu de données NENCI-2021 pour plusieurs valeurs de la coupure $Z_{\text{cut-off}}$, données en Å. Plus la valeur de $Z_{\text{cut-off}}$ est élevée, plus la longueur de l'intervalle d'intégration $\mathcal{Z} = R_{ij} - 2Z_{\text{cut-off}}$ est courte. Pour $Z_{\text{cut-off}} < 0,1\text{Å}$, la corrélation est faible car l'intégration prend aussi en compte les électrons de cœur qui ne participent pas au recouvrement des densités électroniques. Entre $Z_{\text{cut-off}} = 0,1\text{Å}$ et $Z_{\text{cut-off}} = 0,7\text{Å}$, la valeur de R^2 atteint un plateau où le maximum $R^2 = 0,957$ est obtenu pour $Z_{\text{cut-off}} = 0,60\text{Å}$. Pour $Z_{\text{cut-off}} > 0,7\text{Å}$, la corrélation diminue car une partie de l'information est perdue. Notons que la distance entre atomes non-liés R_{ij} la plus courte de tous les systèmes transférés est de $1,55\text{Å}$, obtenue pour les liaisons hydrogène $\text{C}=\text{O} \cdots \text{H}-\text{O}$ dans l'homodimère de l'acide acétique à 90% de sa distance d'équilibre et dans son orientation relative d'équilibre. La longueur d'intervalle d'intégration $\mathcal{Z} = R_{ij} - 2Z_{\text{cut-off}}$ reste donc bien positive tant que $Z_{\text{cut-off}} < 0,775\text{Å}$.

Sommation sur les points critiques de liaison non-covalente de la densité électronique moléculaire totale

La réduction de dimensionnalité de l'intégration volumique (3D) à l'intégration le long d'un axe (1D) a considérablement amélioré le modèle. Une réduction supplémentaire, par une évaluation ponctuelle (0D) du produit des densités, a donc été envisagée pour permettre de diminuer également le temps de calcul. Or, il est bien connu en analyse de la densité de charge que l'évaluation des propriétés locales sur les points critiques (3, -1) de liaison non-covalente (nc-BCP pour "non-covalent Bond Critical Point") de la densité électronique moléculaire permet de renseigner sur la nature des interactions [Espinosa *et al.*, 1998]. J'ai donc proposé de remplacer l'intégration du produit $\rho_i(\mathbf{r})\rho_j(\mathbf{r})$ par une sommation de sa valeur sur les positions des nc-BCP de la densité moléculaire totale $\rho_{\text{mol}}(\mathbf{r})$. Puisque $\rho_{\text{mol}}(\mathbf{r})$ est minimale sur un point critique nc-BCP entre deux atomes dans la direction joignant les noyaux, ce type de point peut donc indiquer la localisation du recouvrement des densités. Ce dernier est alors exprimé de la façon suivante :

$$\Omega_{\text{nc-BCP}} = \sum_{i \in A} \sum_{j \in B} \rho_i(\mathbf{r}_{\text{nc-BCP},ij}) \rho_j(\mathbf{r}_{\text{nc-BCP},ij}), \quad (3.34)$$

où nc-BCP, ij est le nc-BCP situé entre les atomes i et j , s'il existe. Sinon, la paire d'atomes ij n'est pas prise en compte dans la sommation. La confrontation directe des valeurs de $\Omega_{\text{nc-BCP}}$ avec les valeurs de référence SAPT est suffisamment linéaire pour affiner par moindres carrés

Coefficients statistiques	Cylindrique	Axial	Point critique nc-BCP
R^2	0,672	0,957	0,795
R	0,820	0,978	0,892
MAE (kcal/mol)	2,89	1,10	2,80
RMSD (kcal/mol)	5,27	1,55	3,98

TABLEAU 3.5 – Coefficients statistiques pour caractériser l'accord entre le modèle ELMAM pour les différentes méthodes d'intégration et la référence SAPT pour l'énergie d'échange-répulsion.

Les coefficients statistiques pour comparer les modèles ELMAM et SAPT de l'énergie d'échange-répulsion sont rassemblés dans ce tableau pour les différentes méthodes d'intégration : dans le volume cylindrique pour la colonne « Cylindrique », le long de l'axe interatomique pour la colonne « Axial », et sur les positions des points critiques de liaison non-covalente (nc-BCP) pour la colonne « Point critique nc-BCP ». Ces indicateurs, dont les formulations sont données dans la section 3.1.2, sont : le coefficient de détermination R^2 , le coefficient de corrélation R , l'erreur absolue moyenne MAE et la déviation quadratique moyenne RMSD. Ces coefficients confirment l'excellent accord pour l'intégration axiale, et qui est nettement moins bon pour les deux autres méthodes.

le facteur d'échelle : $K_{\text{exch-rep}} = 8,26 \cdot 10^{-27} \text{ m}^5 \cdot \text{mol}^{-1}$. Les résultats du modèle $U_{\text{exch-rep}}^{\text{ELMAM}} = k_{\text{exch-rep}}/4\pi\epsilon_0 \Omega_{\text{nc-BCP}}$ ont été confrontés aux valeurs de référence $E_{\text{exch-rep}}^{\text{SAPT}}$ dans le graphique 3.7c. La régression linéaire de ces points a pour équation :

$$U_{\text{exch-rep}}^{\text{ELMAM}} = 0,855 E_{\text{exch-rep}}^{\text{SAPT}}, \quad (3.35)$$

avec le coefficient de détermination $R^2 = 0,795$ caractérisant un accord ELMAM-SAPT meilleur que pour l'intégration cylindrique mais loin d'être aussi bon que pour l'intégration axiale. Les valeurs de MAE = 2,80 kcal/mol et de RMSD = 3,98 kcal/mol, qui sont regroupés dans le tableau 3.5, confirment cette observation. Ce modèle de sommation sur les points critiques doit donc manquer une partie de l'information, de façon similaire à l'intégration axiale pour une trop grande coupure de l'intervalle ($Z_{\text{cut-off}} > 0,7 \text{ \AA}$).

3.3.3 Modèle dépendant de la distance interatomique

Les résultats présentés précédemment avec l'intégration axiale sont excellents mais dans mes premiers développements, ce niveau de précision n'avait pas été atteint. C'est pourquoi j'ai cherché à améliorer mon modèle en m'inspirant de précédentes études [Söderhjelm *et al.*, 2006, Misquitta et Stone, 2016, Rackers et Ponder, 2019] où l'ajout de facteurs $f(R_{ij})$ dépendant de la distance interatomique R_{ij} a été proposé. Pour le potentiel d'interaction d'échange-répulsion ELMAM, $U_{\text{exch-rep}}^{\text{ELMAM}}$, j'ai testé l'introduction de différentes fonctions $f(R_{ij})$ telles que :

$$U_{\text{exch-rep},AB}^{\text{ELMAM}} = \frac{k_{\text{exch-rep}}}{4\pi\epsilon_0} \sum_{i \in A} \sum_{j \in B} f(R_{ij}) \Omega_{ij}, \quad (3.36)$$

où Ω_{ij} est le recouvrement des densités obtenus à partir de l'intégration le long de la direction internucléaire et avec une coupure $Z_{\text{cut-off}} = 0,6 \text{ \AA}$. Dans une approche purement empirique, j'ai d'abord essayé trois fonctions simples : $f(R_{ij}) = R_{ij}$, $f(R_{ij}) = R_{ij}^2$ et $f(R_{ij}) = R_{ij}^{-1}$. Puis, j'ai également employé la fonction proposée par P. Söderhjelm, G. Karlström et U. Ryde [Söderhjelm *et al.*, 2006] : $f(R_{ij}) = 1 + Ce^{-R_{ij}/a_0}$, avec C , un paramètre empirique sans dimension à affiner

et $a_0 = 0,529\text{\AA}$, le rayon de Bohr.

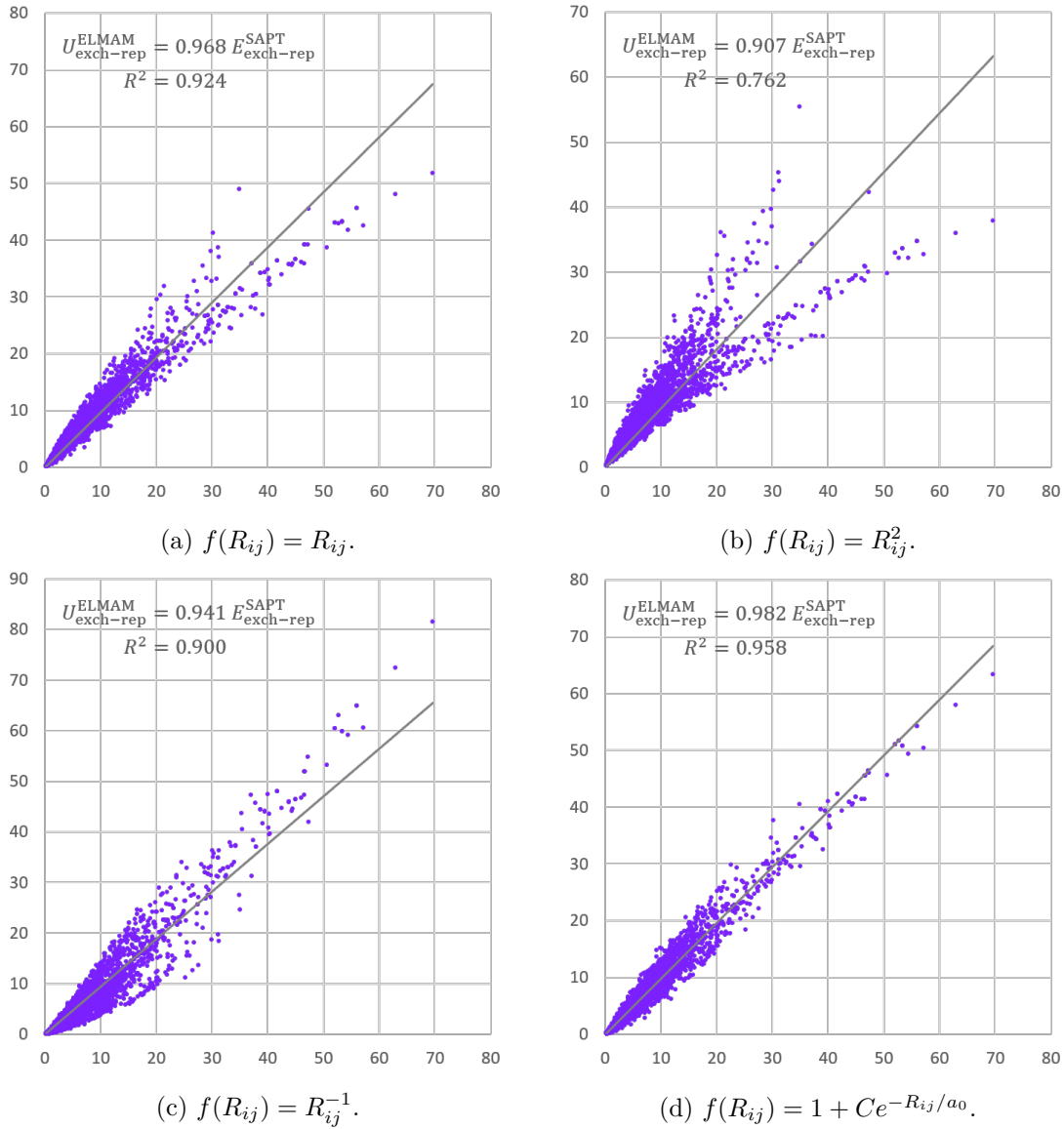


FIGURE 3.9 – Résultats obtenus par le modèle ELMAM de l'énergie d'échange-répulsion comparés aux valeurs de référence SAPT pour plusieurs facteurs dépendant de la distance interatomique $f(R_{ij})$.

Ces graphiques représentent les valeurs du modèle ELMAM du potentiel d'échange-répulsion en ordonnées confrontées aux valeurs SAPT en abscisses, en utilisant différents facteurs dépendant de la distance interatomique R_{ij} : (a) $f(R_{ij}) = R_{ij}$, (b) $f(R_{ij}) = R_{ij}^2$, (c) $f(R_{ij}) = R_{ij}^{-1}$ et (d) $f(R_{ij}) = 1 + C e^{-R_{ij}/a_0}$. Les valeurs d'énergie sont données en kcal/mol. Les régressions linéaires sur les quatre graphiques sont représentées par les droites grises. Les équations de ces quatre régressions linéaires ainsi que les coefficients de détermination correspondant sont affichés en haut à gauche de chaque graphique. La corrélation ELMAM-SAPT est moins bonne pour les trois premières fonctions (a, b et c) que pour le modèle sans facteur dépendant de R_{ij} . Pour la dernière fonction (d), qui a été proposée par [Söderhjelm *et al.*, 2006], cet accord est faiblement amélioré.

Pour la première fonction $f(R_{ij}) = R_{ij}$, le facteur d'échelle calculé par moindres carrés est $k_{\text{exch-rep}} = 8,39.10^{-8} \text{ m}^3.\text{mol}^{-1}$. La régression linéaire entre les résultats du modèle correspondant $U_{\text{exch-rep}}^{\text{ELMAM}}$ et la référence SAPT $E_{\text{exch-rep}}^{\text{SAPT}}$, présentée sur le graphique 3.9a, montre un moins bon accord ELMAM-SAPT ($R^2 = 0,924$) que dans le cas sans facteur dépendant de la

Coefficients statistiques	$f(R_{ij}) = R_{ij}$	$f(R_{ij}) = R_{ij}^2$	$f(R_{ij}) = R_{ij}^{-1}$	$f(R_{ij}) = 1 + Ce^{-\frac{R_{ij}}{a_0}}$
R^2	0,924	0,762	0,900	0,958
R	0,961	0,873	0,949	0,979
MAE (kcal/mol)	1,17	1,89	1,86	1,06
RMSD (kcal/mol)	1,90	3,18	2,58	1,51

TABLEAU 3.6 – Coefficients statistiques pour caractériser l'accord entre le modèle ELMAM pour les différents facteurs $f(R_{ij})$ dépendant de la distance interatomique R_{ij} et la référence SAPT pour l'énergie d'échange-répulsion.

Les coefficients statistiques pour comparer les modèles ELMAM et SAPT de l'énergie d'échange-répulsion sont rassemblés dans ce tableau pour les différentes fonctions de la distance interatomique R_{ij} : $f(R_{ij}) = R_{ij}$ (colonne n°2), $f(R_{ij}) = R_{ij}^2$ (colonne n°3), $f(R_{ij}) = R_{ij}^{-1}$ (colonne n°4) et $f(R_{ij}) = 1 + Ce^{-R_{ij}/a_0}$ (colonne n°5). Ces indicateurs, dont les formulations sont données dans la section 3.1.2, sont : le coefficient de détermination R^2 , le coefficient de corrélation R , l'erreur absolue moyenne MAE et la déviation quadratique moyenne RMSD. Ces coefficients montrent que l'introduction de ces facteurs $f(R_{ij})$ n'apporte pas d'amélioration significative par rapport au modèle initial, donc les coefficients statistiques sont présentés dans le tableau 3.5.

distance interatomique. Il en est de même pour les fonctions $f(R_{ij}) = R_{ij}^2$ et $f(R_{ij}) = R_{ij}^{-1}$, dont les résultats sont représentés respectivement avec $k_{\text{exch-rep}} = 3,11.10^2 \text{ m}^2.\text{mol}^{-1}$ dans le graphique 3.9b, et avec $k_{\text{exch-rep}} = 4,21.10^{-26} \text{ m}^5.\text{mol}^{-1}$ dans le graphique 3.9c. Pour la fonction $f(R_{ij}) = 1 + Ce^{-R_{ij}/a_0}$, j'ai affiné de manière simultanée par moindres carrés le facteur d'échelle $k_{\text{exch-rep}} = 2,11.10^{-16} \text{ m}^4.\text{mol}^{-1}$ et le paramètre empirique $C = -2,57$ à l'aide de 80% des systèmes transférés du jeu de données NENCI-2021, en gardant les 20% restant pour la validation croisée. Les résultats du modèle $U_{\text{exch-rep}}^{\text{ELMAM}}$ qui en découle sont confrontés aux valeurs de référence $E_{\text{exch-rep}}^{\text{SAPT}}$ dans le graphique 3.9d. La régression linéaire sur ces points a pour équation :

$$U_{\text{exch-rep}}^{\text{ELMAM}} = 0,982 E_{\text{exch-rep}}^{\text{SAPT}}, \quad (3.37)$$

avec le coefficient de détermination $R^2 = 0,958$ qui reste quasiment identique au modèle sans facteur dépendant de R_{ij} . Les valeurs de MAE = 1,06 kcal/mol et de RMSD = 1,51 kcal/mol ne montrent pas d'amélioration significative non plus. Les différents coefficients statistiques obtenus avec ces quatre fonctions de la distance interatomique sont rassemblés dans le tableau 3.6. Finalement, les différentes fonctions $f(R_{ij})$ introduites ici n'ont pas permis d'améliorer significativement les résultats déjà très bons issus de l'intégration axiale du produit des densités avec une coupure $Z_{\text{cut-off}} = 0,6\text{\AA}$.

Néanmoins, pour d'autres méthodes de calcul de l'intégrale Ω , l'introduction d'un facteur $f(R_{ij})$ dépendant de la distance interatomique R_{ij} permet de réduire significativement les écarts ELMAM-SAPT. Pour une intégration axiale mais avec une coupure plus courte $Z_{\text{cut-off}} = 0,4\text{\AA}$, le coefficient de détermination valait $R^2 = 0,918$ (voir graphique 3.8), contre $R^2 = 0,957$ pour $Z_{\text{cut-off}} = 0,6\text{\AA}$. En introduisant le facteur $f(R_{ij}) = R_{ij}$ dans le modèle avec $Z_{\text{cut-off}} = 0,4\text{\AA}$, et le facteur d'échelle $k_{\text{exch-rep}} = 5,77.10^{-8} \text{ m}^3.\text{mol}^{-1}$ affiné par moindres carrés, j'ai obtenu les résultats de modèle $U_{\text{exch-rep}}^{\text{ELMAM}}$ qui sont confrontés aux valeurs SAPT dans le graphique 3.10a. La régression linéaire de ces points a pour équation :

$$U_{\text{exch-rep}}^{\text{ELMAM}} = 0,983 E_{\text{exch-rep}}^{\text{SAPT}}, \quad (3.38)$$

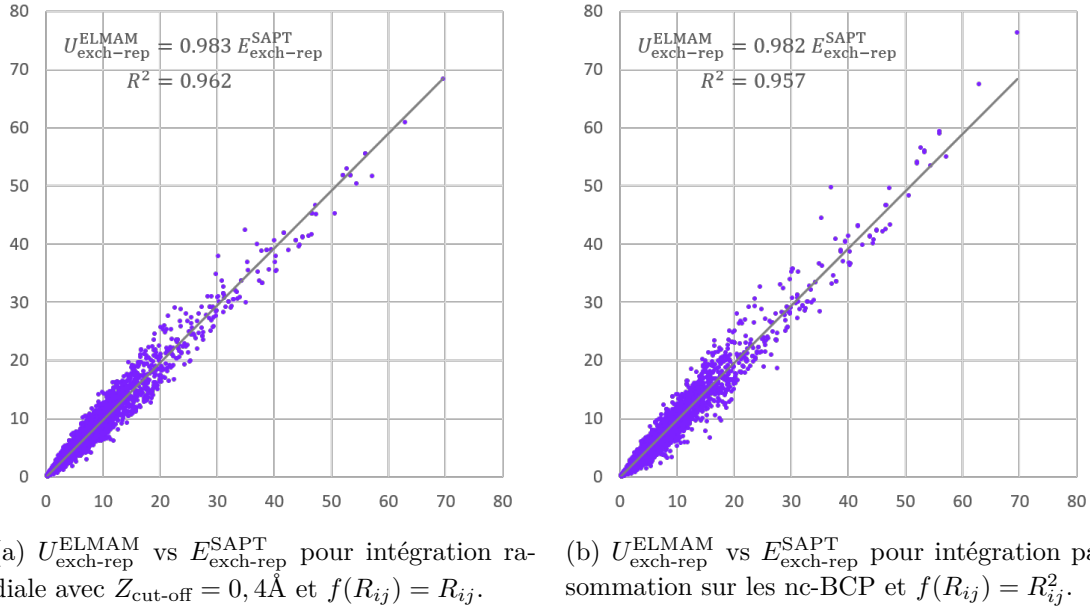


FIGURE 3.10 – Résultats obtenus par le modèle ELMAM de l'énergie d'échange-répulsion comparés aux valeurs de référence SAPT pour d'autres méthodes d'intégration et avec l'introduction de facteurs dépendant de la distance interatomique.

Ces graphiques représentent les valeurs du modèle ELMAM du potentiel d'échange-répulsion en ordonnées confrontées aux valeurs SAPT en abscisses, en utilisant : (a) la méthode d'intégration axiale avec $Z_{\text{cut-off}} = 0,4 \text{ \AA}$ et le facteur $f(R_{ij}) = R_{ij}$ et (b) la méthode d'intégration par sommation sur les nc-BCP et le facteur $f(R_{ij}) = R_{ij}^2$. Les valeurs d'énergie sont données en kcal/mol. Les régressions linéaires sur les deux graphiques sont représentées par les droites grises. Les équations de ces deux régressions linéaires ainsi que les coefficients de détermination correspondant sont affichés en haut à gauche de chaque graphique. Les excellentes corrélations ELMAM-SAPT obtenues montre que l'introduction des facteurs $f(R_{ij})$ dans les autres méthodes d'intégration du recouvrement des densités permet de retrouver des résultats de qualité équivalente à ceux de l'intégration axiale avec $Z_{\text{cut-off}} = 0,6 \text{ \AA}$.

avec $R^2 = 0,962$ (et $R = 0,981$) qui est très légèrement meilleur que celui obtenu avec le modèle à $Z_{\text{cut-off}} = 0,6 \text{ \AA}$. Les valeurs de $\text{MAE} = 1,00 \text{ kcal/mol}$ et $\text{RMSD} = 1,42 \text{ kcal/mol}$ confirment une légère amélioration. De même, pour la méthode d'intégration par sommation sur les points critiques de liaison non-covalente (nc-BCP), le coefficient de détermination était égal à seulement $R^2 = 0,795$. J'ai introduit dans ce modèle la fonction $f(R_{ij}) = R_{ij}^2$, et le facteur d'échelle $k_{\text{exch-rep}} = 2,11 \cdot 10^{-8} \text{ m}^3 \cdot \text{mol}^{-1}$ affiné par moindres carrés, dont les résultats obtenus sont comparés aux valeurs $E_{\text{exch-rep}}^{\text{SAPT}}$ dans le graphique 3.10b. La régression linéaire a pour expression :

$$U_{\text{exch-rep}}^{\text{ELMAM}} = 0,982 E_{\text{exch-rep}}^{\text{SAPT}}, \quad (3.39)$$

avec $R^2 = 0,957$ (et $R = 0,978$) qui est égal à celui obtenu pour la méthode d'intégration axiale avec $Z_{\text{cut-off}} = 0,6 \text{ \AA}$. Les valeurs de $\text{MAE} = 1,06 \text{ kcal/mol}$ et $\text{RMSD} = 1,56 \text{ kcal/mol}$ sont également très proches. Par conséquent, l'introduction des facteurs $f(R_{ij})$ permet de retrouver des résultats de qualité équivalente à ceux obtenus par la méthode d'intégration optimale à partir d'autres méthodes d'intégration. Ceci est notamment intéressant pour le cas de la sommation sur les nc-BCP qui est moins couteuse en temps de calcul, même s'il faut déterminer les points critiques en amont.

C_H	C_C	C_N	C_O	C_S
0,90	1,12	1,19	1,00	1,36

TABLEAU 3.7 – Paramètres empiriques dépendant de l'espèce chimique introduits dans le modèle ELMAM d'énergie d'échange-répulsion.

Les paramètres empiriques C_H , C_C , C_N , C_O et C_S ont été introduits comme coefficients devant les densités électroniques atomiques dans l'intégrale de recouvrement Ω (équation 3.40) pour chaque espèce chimique (hydrogène, carbone, azote, oxygène et soufre). Ils ont été affinés par minimisation des moindres carrés contre les valeurs SAPT pour 80% des systèmes transférés du jeu de données NENCI-2021. Ils restent tous relativement proches de l'unité, en particulier pour l'oxygène, et suggèrent une légère surestimation de la densité pour les atomes d'hydrogène et une légère sous-estimation pour les atomes de carbone, azote et soufre.

3.3.4 Introduction de paramètres atomiques empiriques

En partant de l'hypothèse selon laquelle la qualité des valeurs des densités électroniques atomiques $\rho(\mathbf{r})$, qui sont reconstruites à partir des paramètres multipolaires ELMAM2, puisse expliquer les erreurs observées entre le modèle ELMAM et la référence SAPT, des paramètres empiriques C_i dépendant de l'espèce chimique de l'atome i ont été affinés. Ces paramètres ont été introduits comme coefficients devant les densités électroniques atomiques des atomes i et j dans le calcul de l'intégrale le long de l'axe reliant les noyaux :

$$U_{\text{exch-rep},AB}^{\text{ELMAM}} = \frac{k_{\text{exch-rep}}}{4\pi\epsilon_0} \sum_{i \in A} \sum_{j \in B} \int_z C_i \rho_i(\mathbf{r}) C_j \rho_j(\mathbf{r}) dz, \quad (3.40)$$

en conservant $k_{\text{exch-rep}} = 2,00 \cdot 10^{-16} \text{ m}^4 \cdot \text{mol}^{-1}$. Pour affiner les paramètres C_i , C_j par la méthode des moindres carrés, j'ai utilisé 80% des systèmes transférés du jeu de données NENCI-2021. Les valeurs obtenues sont présentées dans le tableau 3.7. Ces paramètres sont proches de l'unité, notamment pour l'oxygène où C_O est exactement égal à 1,00, ce qui permet de valider l'utilisation des paramètres multipolaires de la librairie ELMAM2 pour le calcul du recouvrement des densités électroniques. Les valeurs obtenues suggèrent néanmoins que les recouvrements impliquant des atomes d'hydrogène sont légèrement surestimés tandis que ceux impliquant des atomes de carbone, d'azote et de soufre sont légèrement sous-estimés.

Les 20% des systèmes transférés du jeu de données NENCI-2021 ont été utilisés pour la validation croisée de ces paramètres. Les résultats du modèle paramétré sur ces systèmes sont confrontés aux valeurs de référence SAPT dans la figure 3.11a. La régression linéaire de ces points a pour équation :

$$U_{\text{exch-rep}}^{\text{ELMAM}} = 0,985 E_{\text{exch-rep}}^{\text{SAPT}}, \quad (3.41)$$

avec le coefficient de détermination $R^2 = 0,971$. Pour le jeu de données complet, la confrontation du modèle ELMAM paramétré aux valeurs SAPT est donnée sur le graphique 3.11b. La régression linéaire devient :

$$U_{\text{exch-rep}}^{\text{ELMAM}} = 0,984 E_{\text{exch-rep}}^{\text{SAPT}}, \quad (3.42)$$

avec le coefficient de détermination $R^2 = 0,969$. Les résultats obtenus sur les 20% des systèmes transférés non-utilisés pour l'entraînement des paramètres empiriques sont donc bien équivalents à ceux obtenus sur l'ensemble du jeu de données, confirmant ainsi la pertinence du modèle grâce

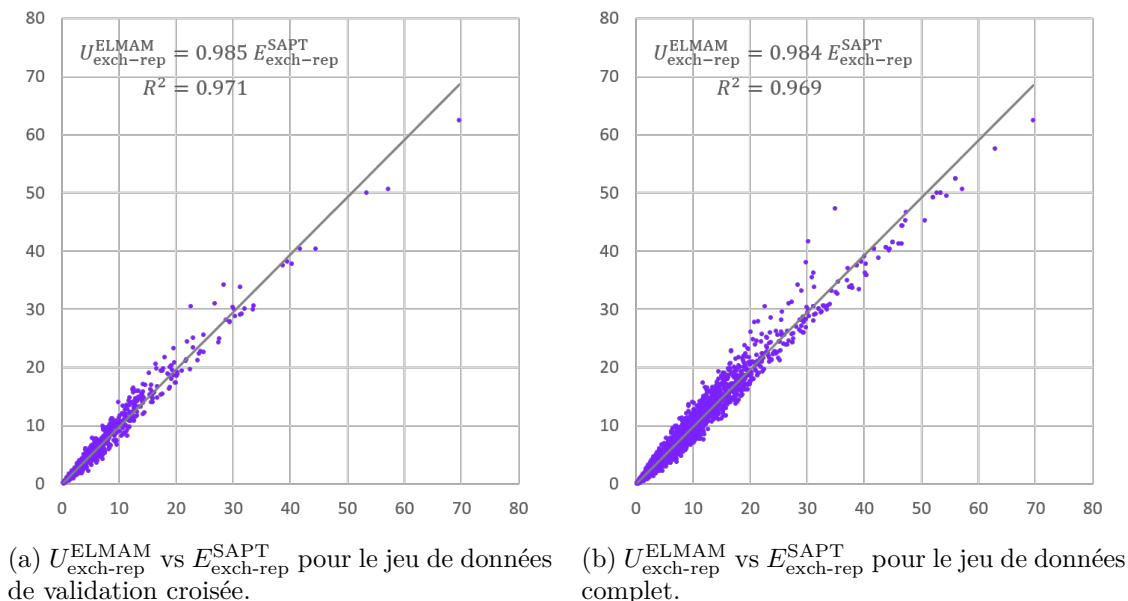


FIGURE 3.11 – Résultats obtenus par le modèle ELMAM de l'énergie d'échange-répulsion avec les paramètres empiriques comparés aux valeurs de référence SAPT.

Ces graphiques représentent les valeurs du modèle ELMAM comportant les paramètres empiriques en ordonnées et les valeurs SAPT en abscisses de l'énergie d'échange-répulsion, pour (a) les 20% des systèmes transférés du jeu de données NENCI-2021 [Sparrow *et al.*, 2021] qui n'ont pas été utilisés pour entraîner ces paramètres et (b) sur l'ensemble des systèmes transférés. Les valeurs d'énergie sont données en kcal/mol. Les régressions linéaires sur les deux graphiques sont représentées par les droites grises. Les équations de ces deux régressions linéaires ainsi que le coefficient de détermination correspondant sont affichés en haut à gauche de chaque graphique. La corrélation ELMAM-SAPT est légèrement améliorée par l'introduction des paramètres empiriques, ce qui est confirmé par la validation croisée sur les systèmes non-utilisés pour leur entraînement.

à l'approche de validation croisée.

Les différents coefficients statistiques caractérisant les résultats du modèle $U_{\text{exch-rep}}^{\text{ELMAM}}$ avec les paramètres empiriques, pour les systèmes utilisés pour la validation croisée uniquement et pour l'ensemble du jeu de données, sont regroupés dans le tableau 3.8. Le coefficient de détermination étant plus élevé que sans ces paramètres empiriques ($R^2 = 0,957$), la corrélation ELMAM-SAPT est légèrement améliorée par leur introduction. De même, les valeurs de MAE = 0,89 kcal/mol et de RMSD = 1,31 kcal/mol sont significativement réduites (MAE = 1,10 kcal/mol et de RMSD = 1,55 kcal/mol auparavant) grâce à ces paramètres empiriques. L'erreur absolue moyenne du potentiel d'échange-répulsion ELMAM est d'ailleurs à présent inférieure à la précision chimique attendue de 1 kcal/mol.

3.3.5 Influence de la composition du dimère

Comme pour les autres contributions énergétiques, l'influence de la composition des dimères du jeu de données NENCI-2021 sur l'accord entre le potentiel d'échange-répulsion ELMAM et les valeurs de référence SAPT a été analysée en calculant le coefficient de détermination R^2 pour chacun des 77 dimères transférés en utilisant les 45 géométries différentes associées. Ces résultats sont présentés sous la forme d'un histogramme sur la figure 3.12. Comme le montre la vue globale sur l'histogramme 3.12a, les 77 dimères présentent tous une valeur de R^2 supérieure à 0,90, et

Coefficients statistiques	Validation croisée	Jeu de données complet
R^2	0,971	0,969
R	0,985	0,984
MAE (kcal/mol)	0,90	0,89
RMSD (kcal/mol)	1,29	1,31

TABLEAU 3.8 – Coefficients statistiques pour caractériser l'accord entre le modèle ELMAM avec paramètres empiriques et la référence SAPT pour les énergies d'échange-répulsion.

Les coefficients statistiques pour comparer les modèles ELMAM avec paramètres empiriques et SAPT de l'énergie d'échange-répulsion, pour les systèmes non-utilisés pour l'entraînement des paramètres (colonne « Validation croisée ») et pour l'ensemble des systèmes transférés du jeu de données NENCI-2021 (colonne « Jeu de données complet ») sont rassemblés dans ce tableau. Ces indicateurs, dont les formulations sont données dans la section 3.1.2, sont : le coefficient de détermination R^2 , le coefficient de corrélation R , l'erreur absolue moyenne MAE et la déviation quadratique moyenne RMSD. Ces coefficients confirment la validation croisée car l'accord reste aussi bon pour les systèmes qui n'ont pas servi à l'entraînement des paramètres empiriques que pour l'ensemble des systèmes.

64 d'entre eux (83,1%), une valeur de R^2 supérieure à 0,95. Pour commenter plus précisément ces résultats, les valeurs prises par R^2 entre 0,90 et 1,0 sont détaillées dans l'histogramme 3.12b. Pour la majorité des dimères, soit 21 sur les 77 (27,3%), R^2 est compris entre 0,97 et 0,98, caractérisant une très forte corrélation ELMAM-SAPT. Pour 13 dimères (soit 16,9%), la valeur de R^2 est même supérieure à 0,99. Les trois meilleurs accords sont obtenus pour les complexes suivants : néopentane-néopentane ($R^2 = 0,997$ et de type DISP), acétamide-acétamide ($R^2 = 0,997$ et de type ELEC) et uracile-uracile ($R^2 = 0,996$ et de type ELEC). Un unique dimère (1,3%) possède une valeur de R^2 inférieure à 0,92 : pyridine-éthyne ($R^2 = 0,911$ et de type ELEC), et deux (2,6%) une valeur comprise entre 0,92 et 0,93 : N-méthylacétamide-méthylamine ($R^2 = 0,927$ et de type ELEC) et indole-eau ($R^2 = 0,929$ et de type ELEC). Les systèmes de type ELEC apparaissent à la fois pour les moins bonnes valeurs de R^2 et pour les meilleures car ils représentent la majorité des complexes (36/77, 32 étant de type DISP et 9 de type MIXD). La qualité de l'accord ELMAM-SAPT ne semble pas dépendre du type d'interaction dominante et aucun type atomique particulier ne ressort parmi les moins bons résultats. Notons que cette conclusion n'est pas surprenante car les effets de répulsion sont présents dans toutes les interactions, quelles que soient les natures des molécules considérées.

3.3.6 Autres modèles d'énergie de répulsion

Finalement, le modèle d'énergie d'échange-répulsion ELMAM basé sur le recouvrement des densités électroniques fournit des résultats présentant un accord très satisfaisant avec les valeurs de référence SAPT. Cet accord semble principalement dépendre de la méthode d'intégration et de la détermination de la région dans laquelle se produit le recouvrement. Une méthode davantage précise pour la définition de cette région pourrait peut-être améliorer les résultats.

Par ailleurs, l'accord avec les valeurs de référence SAPT pourrait être renforcé en partant de modèles de type potentiel de Lennard-Jones, proportionnel à l'inverse de la distance interatomique avec un exposant 12 ou 14, ou Buckingham, avec une dépendance en exponentielle décroissante. Cependant, ces potentiels fortement empiriques ne peuvent pas être raccrochés à l'interprétation physique du phénomène de répulsion en termes de perte d'écrantage noyau-

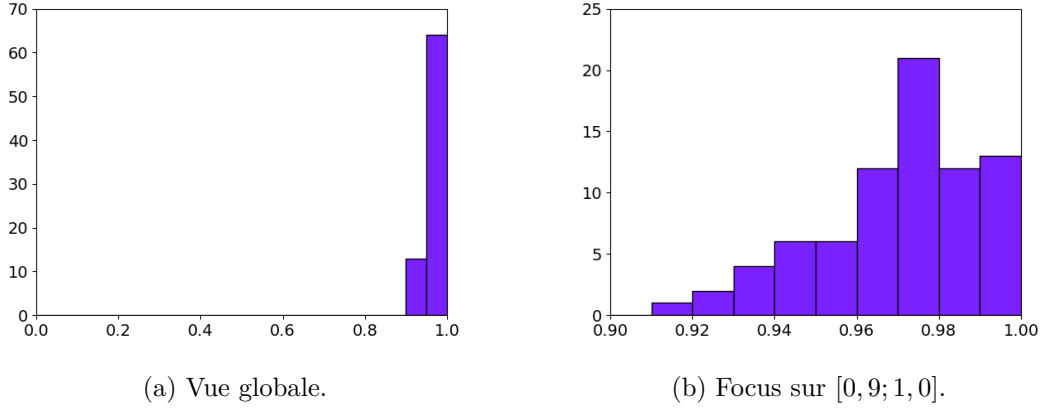


FIGURE 3.12 – Influence de la composition des dimères sur l'accord entre le modèle d'échange-répulsion ELMAM et la référence SAPT.

Le coefficient de détermination R^2 entre les valeurs ELMAM et SAPT de l'énergie d'échange-répulsion a été calculé séparément pour chacun des 77 dimères transférés du jeu de données NENCI-2021 en utilisant les 45 géométries par système disponibles. Les résultats sont présentés sous la forme d'histogrammes montrant le nombre de dimères qui appartiennent à chaque intervalle de valeurs de R^2 . La hauteur d'une barre de graphique représente donc le nombre de complexes pour lesquels R^2 est compris entre deux valeurs. Dans la vue globale (a), les valeurs sont séparées par pas de 0,05, ce qui permet de comparer ces résultats aux autres contributions énergétiques. Dans le focus sur la région de R^2 compris entre 0,9 et 1,0 (b) les valeurs sont séparées par pas de 0,01 afin d'approfondir l'analyse. La valeur de R^2 est supérieure à 0,90 pour tous les dimères, confirmant l'excellent accord avec la référence SAPT, et est compris entre 0,97 et 0,98 pour la plupart d'entre eux.

noyau, comme c'est le cas pour le modèle du recouvrement des densités.

3.4 Potentiel d'interaction van der Waals ELMAM

3.4.1 Sommation des contributions ELMAM de dispersion et d'échange-répulsion

Pour définir le potentiel d'interaction total ELMAM, les contributions électrostatique permanente et d'induction dipolaire avaient déjà été établies dans des travaux antérieurs [Leduc *et al.*, 2019, Vuković *et al.*, 2021]. Dans le cadre de ma thèse, j'ai développé les potentiels de dispersion (voir partie 3.2) et d'échange-répulsion (voir partie 3.3) pour compléter ce potentiel d'interaction total. Ces deux contributions définissent le potentiel de van der Waals ELMAM, $U_{\text{vdw}}^{\text{ELMAM}}$, qui a pour expression :

$$U_{\text{vdw}}^{\text{ELMAM}} = U_{\text{disp}}^{\text{ELMAM}} + U_{\text{exch-rep}}^{\text{ELMAM}}. \quad (3.43)$$

Sans paramètre empirique

Dans les développements des potentiels de dispersion $U_{\text{disp}}^{\text{ELMAM}}$ et d'échange-répulsion $U_{\text{exch-rep}}^{\text{ELMAM}}$, de très bons résultats ont été obtenus uniquement à partir des données de la librairie ELMAM2, sans l'introduction des paramètres empiriques excepté les facteurs d'échelle K_{disp} et $k_{\text{exch-rep}}$. Pour définir le potentiel de van der Waals ELMAM, j'ai donc choisi de sommer la contribution de dispersion, obtenue par l'approximation de London avec les tenseurs de polarisabilités

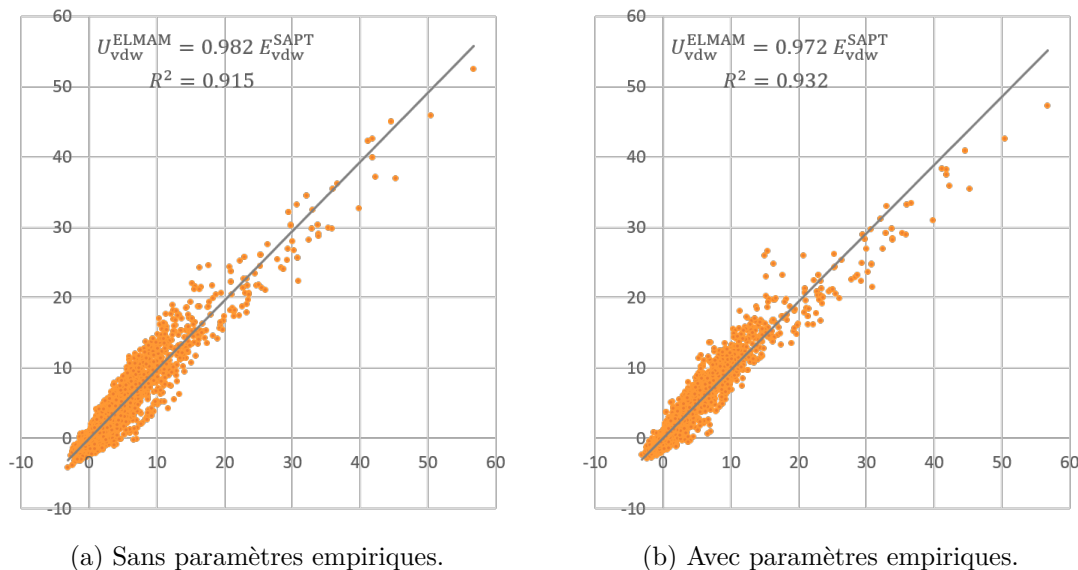


FIGURE 3.13 – Résultats obtenus par le modèle ELMAM de l'énergie de van der Waals sans et avec les paramètres empiriques comparés aux valeurs de référence SAPT.

Ces graphiques représentent les valeurs du modèle ELMAM en ordonnées et les valeurs SAPT en abscisses de l'énergie de van der Waals, (a) sans aucun paramètre empirique dans le modèle ELMAM et (b) avec les paramètres empiriques de dispersion et d'échange-répulsion. Les valeurs d'énergie sont données en kcal/mol. Les régressions linéaires sur les deux graphiques sont représentées par les droites grises. Les équations de ces deux régressions linéaires ainsi que le coefficient de détermination correspondant sont affichés en haut à gauche de chaque graphique. La corrélation ELMAM-SAPT est légèrement améliorée par l'introduction des paramètres empiriques.

anisotropes ELMAM2 (équation 3.14), et la contribution d'échange-répulsion, obtenue par l'intégration axiale du recouvrement des densités électroniques reconstruites à partir des paramètres multipolaires ELMAM2 (équation 3.32). Le potentiel de van der Waals ainsi obtenu a été comparé aux valeurs de référence SAPT, $E_{\text{vdw}}^{\text{SAPT}} = E_{\text{disp}}^{\text{SAPT}} + E_{\text{exch-rep}}^{\text{SAPT}}$, dans le graphique 3.13a. La régression linéaire de ces points a pour équation :

$$U_{\text{vdw}}^{\text{ELMAM}} = 0,982 E_{\text{vdw}}^{\text{SAPT}}, \quad (3.44)$$

avec le coefficient de détermination $R^2 = 0,915$. La corrélation ELMAM-SAPT pour le potentiel de van der Waals est donc moins bonne que pour les potentiels de dispersion et d'échange-répulsion individuellement, dont les coefficients statistiques sont rappelés dans le tableau 3.9, mais reste forte. De même, les valeurs de MAE = 1,21 kcal/mol et de RMSD = 1,73 kcal/mol sont plus élevées pour $U_{\text{vdw}}^{\text{ELMAM}}$ mais restent très correctes, notamment l'erreur absolue moyenne qui reste proche de l'erreur chimique typique de 1 kcal/mol.

Avec les paramètres empiriques dépendant de l'espèce chimique

Puisque l'introduction de paramètres empiriques, dans le potentiel de dispersion (équation 3.17) et dans le potentiel d'échange-répulsion (équation 3.40), avait permis d'améliorer légèrement l'accord avec les valeurs de référence SAPT, il est intéressant d'analyser les résultats du potentiel de van der Waals basé sur ces modèles. Le graphique 3.13b présente la confrontation

Coefficients statistiques	Dispersion	Echange-répulsion	van der Waals
R^2	0,946	0,957	0,915
R	0,973	0,978	0,957
MAE (kcal/mol)	0,44	1,10	1,21
RMSD (kcal/mol)	0,59	1,55	1,73

TABLEAU 3.9 – Coefficients statistiques pour caractériser l'accord entre le modèle ELMAM sans paramètre empirique et la référence SAPT pour l'énergie de van der Waals.

Les coefficients statistiques pour comparer les modèles ELMAM sans paramètre empirique et SAPT des énergies de dispersion (colonne n°2), d'échange-répulsion (colonne n°3) et de van der Waals (colonne n°4) sont rassemblés dans ce tableau. Ces indicateurs, dont les formulations sont données dans la section 3.1.2, sont : le coefficient de détermination R^2 , le coefficient de corrélation R , l'erreur absolue moyenne MAE et la déviation quadratique moyenne RMSD. Ces coefficients montrent que la somme des contributions de dispersion et d'échange-répulsion, c'est-à-dire le potentiel de van der Waals, présente un accord ELMAM-SAPT moins bon et des erreurs plus élevées que ces contributions prises individuellement.

Coefficients statistiques	Dispersion	Echange-répulsion	van der Waals
R^2	0,953	0,969	0,932
R	0,976	0,984	0,965
MAE (kcal/mol)	0,39	0,89	1,00
RMSD (kcal/mol)	0,55	1,31	1,50

TABLEAU 3.10 – Coefficients statistiques pour caractériser l'accord entre le modèle ELMAM avec paramètres empiriques et la référence SAPT pour l'énergie de van der Waals.

Les coefficients statistiques pour comparer les modèles ELMAM, avec paramètres empiriques dépendant de l'espèce chimique, et SAPT des énergies de dispersion (colonne n°2), d'échange-répulsion (colonne n°3) et de van der Waals (colonne n°4) sont rassemblés dans ce tableau. Ces indicateurs, dont les formulations sont données dans la section 3.1.2, sont : le coefficient de détermination R^2 , le coefficient de corrélation R , l'erreur absolue moyenne MAE et la déviation quadratique moyenne RMSD. Le potentiel de van der Waals présente toujours un accord ELMAM-SAPT un peu moins bon et des erreurs légèrement plus élevées que les contributions individuelles de dispersion et de répulsion mais l'introduction des paramètres empiriques permet d'améliorer ces résultats.

de ces résultats aux valeurs SAPT dont la régression linéaire a pour expression :

$$U_{\text{vdw}}^{\text{ELMAM}} = 0,972 E_{\text{vdw}}^{\text{SAPT}}, \quad (3.45)$$

avec le coefficient de détermination $R^2 = 0,932$. Cet accord, ainsi que les valeurs de MAE = 1,00 kcal/mol et de RMSD = 1,50 kcal/mol, sont moins bons que ceux des contributions de dispersion et d'échange-répulsion (rappelées dans le tableau 3.10) prises séparément mais sont significativement meilleurs que dans le modèle sans paramètre empirique. L'erreur absolue moyenne sur $U_{\text{vdw}}^{\text{ELMAM}}$ atteint maintenant la précision chimique attendue.

3.4.2 Combinaison linéaire

Avec ou sans paramètres empiriques, l'accord ELMAM-SAPT du potentiel de van der Waals est significativement moindre que celui associé aux contributions individuelles de dispersion et d'échange-répulsion. Une possibilité d'amélioration est de remplacer la simple sommation de ces deux termes par une combinaison linéaire, avec des pondération κ sans dimension affinés par la

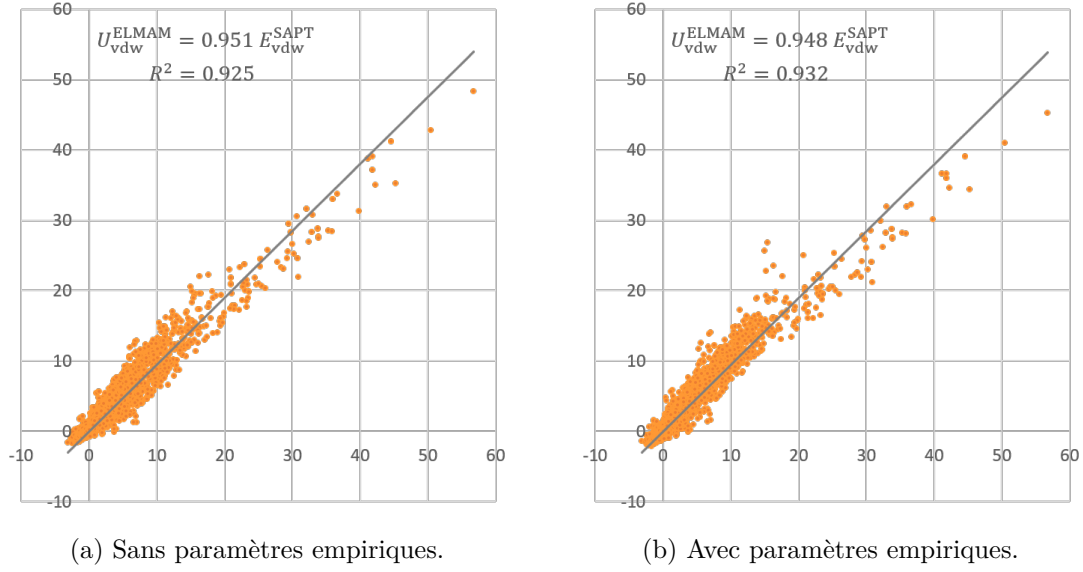


FIGURE 3.14 – Résultats obtenus par le modèle ELMAM de l'énergie de van der Waals par combinaison linéaire des termes de dispersion et de répulsion, sans et avec les paramètres empiriques, comparés aux valeurs de référence SAPT.

Ces graphiques représentent les valeurs du modèle ELMAM en ordonnées et les valeurs SAPT en abscisses de l'énergie de van der Waals obtenue par combinaison linéaire des contributions de dispersion et de répulsion, (a) sans aucun paramètre empirique dans le modèle ELMAM et (b) avec les paramètres empiriques de dispersion et d'échange-répulsion. Les valeurs d'énergie sont données en kcal/mol. Les régressions linéaires sur les deux graphiques sont représentées par les droites grises. Les équations de ces deux régressions linéaires ainsi que le coefficient de détermination correspondant sont affichés en haut à gauche de chaque graphique. La corrélation ELMAM-SAPT est légèrement améliorée par l'introduction des paramètres empiriques. Les coefficients de la combinaison linéaire améliorent les résultats pour le cas sans paramètres empiriques mais les laissent inchangés pour le cas avec paramètres empiriques.

méthode des moindres carrés, telle que :

$$U_{\text{vdw}}^{\text{ELMAM}} = \kappa_{\text{disp}} U_{\text{disp}}^{\text{ELMAM}} + \kappa_{\text{exch-rep}} U_{\text{exch-rep}}^{\text{ELMAM}}. \quad (3.46)$$

Sans paramètre empirique

Les pondérations κ de la combinaison linéaire ont été affinées par moindres carrés sur 80% des systèmes transférés du jeu de données NENCI-2021, les 20% restant ont été conservés pour la validation croisée du modèle. Pour les modèles de dispersion et d'échange-répulsion sans paramètre empirique, les pondérations suivants ont été obtenues : $\kappa_{\text{disp}} = 0,635$ et $\kappa_{\text{exch-rep}} = 0,859$. Le modèle de van der Waals qui en découle a pour expression :

$$U_{\text{vdw}}^{\text{ELMAM}} = 0,635 U_{\text{disp}}^{\text{ELMAM}} + 0,859 U_{\text{exch-rep}}^{\text{ELMAM}}. \quad (3.47)$$

Ce modèle est confronté aux valeurs SAPT sur le graphique 3.14a. La régression linéaire de ces points donne :

$$U_{\text{vdw}}^{\text{ELMAM}} = 0,951 E_{\text{vdw}}^{\text{SAPT}}, \quad (3.48)$$

Coefficients statistiques	Validation croisée	Jeu de données complet
R^2	0,925	0,925
R	0,962	0,962
MAE (kcal/mol)	1,09	1,10
RMSD (kcal/mol)	1,50	1,51

TABLEAU 3.11 – Coefficients statistiques pour caractériser l'accord entre le modèle ELMAM de combinaison linéaire des termes de dispersion et de répulsion, sans paramètre empirique, et la référence SAPT d'énergie de van der Waals.

Les coefficients statistiques pour comparer les modèles ELMAM sans paramètres empiriques et SAPT de l'énergie de van der Waals, pour les systèmes non-utilisés pour l'entraînement des coefficients de combinaison linéaire (colonne « Validation croisée ») et pour l'ensemble des systèmes transférés du jeu de données NENCI-2021 (colonne « Jeu de données complet ») sont rassemblés dans ce tableau. Ces indicateurs, dont les formulations sont données dans la section 3.1.2, sont : le coefficient de détermination R^2 , le coefficient de corrélation R , l'erreur absolue moyenne MAE et la déviation quadratique moyenne RMSD. Ces indicateurs statistiques confirment la validation croisée car l'accord reste identique pour les systèmes qui n'ont pas servi à l'entraînement des coefficients de combinaison linéaire que pour l'ensemble des systèmes.

avec le coefficient de détermination $R^2 = 0,925$. L'accord ELMAM-SAPT est donc légèrement amélioré par cette combinaison linéaire, ce qui est confirmé par une diminution des valeurs de MAE = 1,10 kcal/mol et de RMSD = 1,51 kcal/mol. Ces coefficients statistiques sont regroupés dans le tableau 3.11, pour les systèmes utilisés pour la validation croisée et pour l'ensemble du jeu de données. Ces coefficients sont identiques pour les systèmes qui n'ont pas été utilisés pour l'affinement par moindres carrés que pour le jeu de données complet, ce qui confirme la validation croisée du modèle.

Avec les paramètres empiriques dépendant de l'espèce chimique

Pour les modèles de dispersion et d'échange-répulsion avec les paramètres empiriques dépendant de l'espèce chimique, les pondérations de la combinaison linéaire obtenues par affinement des moindres carrés sur 80% des systèmes transférés sont : $\kappa_{\text{disp}} = 0,827$ et $\kappa_{\text{exch-rep}} = 0,927$. Le modèle de van der Waals est alors exprimé par :

$$U_{\text{vdw}}^{\text{ELMAM}} = 0,827 U_{\text{disp}}^{\text{ELMAM}} + 0,927 U_{\text{exch-rep}}^{\text{ELMAM}}. \quad (3.49)$$

Ce modèle est confronté aux valeurs SAPT sur le graphique 3.14b. La régression linéaire de ces points a pour équation :

$$U_{\text{vdw}}^{\text{ELMAM}} = 0,948 E_{\text{vdw}}^{\text{SAPT}}, \quad (3.50)$$

avec le coefficient de détermination $R^2 = 0,932$, qui est égal à celui obtenu par simple sommation des contributions de dispersion et d'échange-répulsion. Les valeurs de MAE = 0,95 kcal/mol et de RMSD = 1,46 kcal/mol sont légèrement plus faibles pour le jeu de données complet. Cependant, ce n'est pas le cas en prenant uniquement les systèmes qui n'ont pas été utilisés pour l'affinement des coefficients de combinaison linéaires, pour lesquels ces coefficients (regroupés dans le tableau 3.12) sont identiques à ceux de la simple sommation. Finalement, la combinaison linéaire ne permet pas d'améliorer les résultats de potentiel de van der Waals à partir des modèles de dispersion et d'échange-répulsion avec paramètres empiriques.

Coefficients statistiques	Validation croisée	Jeu de données complet
R^2	0,933	0,932
R	0,966	0,965
MAE (kcal/mol)	1,00	0,95
RMSD (kcal/mol)	1,52	1,46

TABLEAU 3.12 – Coefficients statistiques pour caractériser l'accord entre le modèle ELMAM de combinaison linéaire des termes de dispersion et de répulsion, avec paramètres empiriques, et la référence SAPT d'énergie de van der Waals.

Les coefficients statistiques pour comparer les modèles ELMAM avec paramètres empiriques et SAPT de l'énergie de van der Waals, pour les systèmes non-utilisés pour l'entraînement des coefficients de combinaison linéaire (colonne « Validation croisée ») et pour l'ensemble des systèmes transférés du jeu de données NENCI-2021 (colonne « Jeu de données complet ») sont rassemblés dans ce tableau. Ces indicateurs, dont les formulations sont données dans la section 3.1.2, sont : le coefficient de détermination R^2 , le coefficient de corrélation R , l'erreur absolue moyenne MAE et la déviation quadratique moyenne RMSD. Ces indicateurs statistiques confirment la validation croisée car l'accord reste identique pour les systèmes qui n'ont pas servi à l'entraînement des coefficients de combinaison linéaire que pour l'ensemble des systèmes.

3.4.3 Influence de la composition du dimère

Pour chercher à interpréter les erreurs du potentiel de van der Waals ELMAM, $U_{\text{vdw}}^{\text{ELMAM}}$, l'influence de la composition des dimères du jeu de données NENCI-2021 sur l'accord ELMAM-SAPT a été analysée en calculant le coefficient de détermination R^2 pour chacun des 77 dimères transférés en utilisant les 45 géométries différentes associées. L'histogramme regroupant ces résultats est présenté sur la figure 3.15.

Pour la majorité des dimères (54/77 soit 70,1% des complexes), l'accord ELMAM-SAPT est excellent, avec des valeurs de R^2 supérieures à 0,90. Parmi ces systèmes, la plupart (35/54) sont dominés par des interactions de type électrostatique (type ELEC défini par [Sparrow *et al.*, 2021]). Les meilleures valeurs de R^2 sont obtenues pour les complexes suivants : acétamide - acétamide ($R^2 = 0,995$), uracile - uracile ($R^2 = 0,990$) et acétamide - uracile ($R^2 = 0,989$). Ces trois dimères font partie de ceux ayant les moins bons résultats pour la contribution de dispersion ($R^2 = 0,913$ pour acétamide - acétamide, $R^2 = 0,927$ pour uracile - uracile et $R^2 = 0,911$ pour acétamide - uracile) mais des meilleurs pour la contribution d'échange-répulsion ($R^2 = 0,997$ pour acétamide - acétamide, $R^2 = 0,996$ pour uracile - uracile et $R^2 = 0,995$ pour acétamide - uracile). Par ailleurs, ces dimères présentent des fortes valeurs d'énergie d'échange-répulsion, $E_{\text{exch-rep}}^{\text{SAPT}}$ autour de +30 kcal/mol dans la géométrie d'équilibre, par rapport à leurs valeurs d'énergie de dispersion, avec $E_{\text{disp}}^{\text{SAPT}}$ autour de -8 kcal/mol dans la géométrie d'équilibre. Aussi, l'énergie de van der Waals de ces dimères présente un excellent accord avec SAPT car elle est dominée par la contribution d'échange-répulsion pour laquelle mon modèle permet une reproduction fiable.

Pour 21 dimères (27,3%), l'accord est moins bon (R^2 compris entre 0,75 et 0,90) mais reste acceptable. En revanche, la corrélation ELMAM-SAPT est très faible pour le dimère méthanethiol - méthanethiol ($R^2 = 0,359$) et nulle pour le dimère néopentane - néopentane ($R^2 = 0,000$). Pour le complexe méthanethiol - méthanethiol, les accords obtenus pour les contributions de dispersion ($R^2 = 0,943$) et d'échange-répulsion ($R^2 = 0,935$) font partie des moins bonnes, ce qui peut expliquer en partie les mauvais résultats pour le potentiel de van der Waals. Par contre,

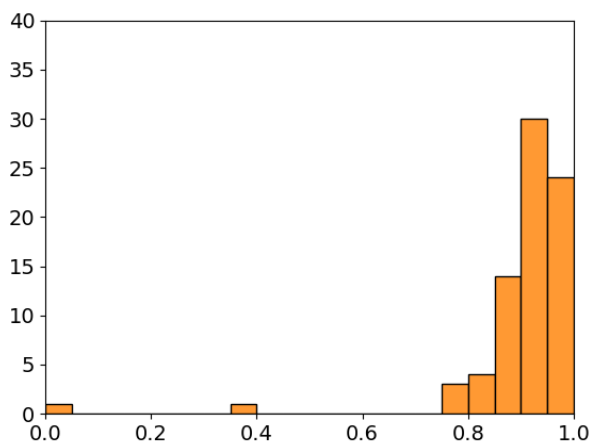


FIGURE 3.15 – Influence de la composition des dimères sur l'accord entre le modèle de van der Waals ELMAM et la référence SAPT.

Le coefficient de détermination R^2 entre les valeurs ELMAM et SAPT de l'énergie de van der Waals a été calculé séparément pour chacun des 77 dimères transférés du jeu de données NENCI-2021 en utilisant les 45 géométries par système disponibles. Les résultats sont présentés sous la forme d'un histogramme montrant le nombre de dimères qui appartiennent à chaque intervalle de valeurs de R^2 . La hauteur d'une barre de graphique représente donc le nombre de complexes pour lesquels R^2 est compris entre deux valeurs séparées par pas de 0,05. La valeur de R^2 est supérieure à 0,90 pour la plupart des dimères mais la corrélation avec la référence SAPT est très faible pour un complexe ($R^2 = 0,359$ pour méthaneethiol - méthaneethiol) et nulle pour un autre ($R^2 = 0,000$ pour néopentane - néopentane).

pour le complexe néopentane - néopentane, les résultats pour le terme de dispersion ($R^2 = 0,996$) sont excellents et ceux pour le terme d'échange-répulsion ($R^2 = 0,997$) sont les meilleurs parmi les 77 dimères. La non-corrélation du potentiel de van der Waals avec les valeurs SAPT pour ce dimère ne peut donc pas être expliquée par l'erreur sur les contributions individuelles. Par ailleurs, pour ce complexe, les valeurs d'énergie de dispersion ($E_{\text{disp}}^{\text{SAPT}} = -3,48$ kcal/mol dans la configuration d'équilibre) et d'échange répulsion ($E_{\text{exch-rep}}^{\text{SAPT}} = +2,68$ kcal/mol dans la configuration d'équilibre) sont très proches en valeurs absolues et puisqu'elles sont de signes opposés, leur somme est très proche de zéro ($E_{\text{vdw}}^{\text{SAPT}} = -0,80$ kcal/mol) et inférieure à l'erreur moyenne absolue $\text{MAE} = 1,00$ kcal/mol. De même, pour le complexe méthaneethiol - méthaneethiol dans la géométrie d'équilibre : $E_{\text{disp}}^{\text{SAPT}} = -3,79$ kcal/mol, $E_{\text{exch-rep}}^{\text{SAPT}} = +5,50$ kcal/mol et $E_{\text{vdw}}^{\text{SAPT}} = +1,71$ kcal/mol. Aussi, lorsque les contributions de dispersion et d'échange-répulsion sont proches en valeurs absolues (ce qui est souvent le cas pour les systèmes dominés par des interactions dispersives, de type DISP), puisque leurs signes sont opposés alors leur somme prend une valeur très faible, de l'ordre de grandeur de l'erreur moyenne, ce qui a pour conséquence la perte de la corrélation entre le modèle ELMAM et la référence SAPT.

3.5 Conclusion partielle de chapitre

Dans ce chapitre, les développements réalisés pour obtenir le potentiel d'interaction total ELMAM, basé sur les paramètres multipolaires et sur les tenseurs de polarisabilité de la librairie ELMAM2, ont été présentés. Ce potentiel d'interaction est composé de quatre contributions :

électrostatique, d'induction, de dispersion et d'échange-répulsion. Les termes électrostatique et d'induction avait déjà été étudiés dans de précédentes publications [Leduc *et al.*, 2019, Vuković *et al.*, 2021] contrairement aux modèles de dispersion et d'échange-répulsion sur lesquels j'ai travaillé durant ma thèse. J'ai optimisé ces deux contributions de manière individuelle et j'ai testé leur validité grâce au jeu de données de benchmarking de petites molécules organiques NENCI-2021 [Sparrow *et al.*, 2021].

Pour le potentiel de dispersion ELMAM, je suis partie de l'approximation de London en introduisant les polarisabilités anisotropes de la librairie ELMAM2. En utilisant la trace du produit des tenseurs de polarisabilités, j'ai obtenu un bon facteur de détermination $R^2 = 0,946$ avec les valeurs SAPT du jeu de données NENCI-2021 et une erreur absolue moyenne de 0,44 kcal/mol, largement en-dessous de l'erreur chimique typique. Ces résultats ont été obtenus sans introduire de paramètre empirique, excepté un facteur d'échelle, et ne dépendent que des valeurs d'énergie de première ionisation (qui peuvent être considérées comme des constantes puisqu'elles sont définies de manière univoque dans la littérature), des distances interatomiques (qui sont imposées par les structures moléculaires du jeu de données) et des polarisabilités atomiques qui sont transférées de la librairie ELMAM2. Dans le but de tester la pertinence des polarisabilités ELMAM2 pour le calcul du potentiel de dispersion, j'ai ajouté des paramètres empiriques dépendant de l'espèce chimique comme coefficients devant ces quantités transférables. Finalement, ces paramètres restent très proches de l'unité, ce qui valide l'utilisation de ces polarisabilités, et le modèle n'est pas significativement amélioré avec $R^2 = 0,953$ et une erreur absolue moyenne de 0,39 kcal/mol. L'analyse de l'influence de la composition des dimères étudiés sur les résultats du potentiel de dispersion a révélé que les systèmes dominés par des interactions de type dispersif présentent de meilleures valeurs de R^2 que ceux dominés par des interactions électrostatiques coulombiennes. En particulier, les moins bons résultats sont obtenus pour les dimères présentant des groupements acides carboxyliques impliqués dans une liaison hydrogène. Une perspective d'amélioration du potentiel de dispersion ELMAM est l'utilisation d'un modèle plus élaboré tel que le modèle de Slater-Kirkwood qui fait intervenir la notion de nombre d'électrons effectif d'un atome, qu'il serait possible d'estimer à partir du paramètre de population de valence du modèle multipolaire.

Pour le potentiel d'échange-répulsion ELMAM, j'ai basé mes développements sur le modèle de recouvrement des densités électroniques qui a l'avantage d'être lié à l'interprétation physique du phénomène de répulsion en termes de perte d'écrantage noyau-noyau due à la déplétion de densité électronique dans la région de recouvrement. Pour calculer ce recouvrement, j'ai testé plusieurs méthodes d'intégration du produit des densités électroniques, ces dernières étant reconstruites par transfert des paramètres multipolaires de la librairie ELMAM2. L'intégration dans un volume cylindrique semble bruitée car elle tient compte de densités qui sont à l'évidence moins représentatives du recouvrement. Par contre, l'intégration le long de l'axe internucléaire fournit de très bons résultats, $R^2 = 0,957$ et une erreur absolue de 1,10 kcal/mol, à condition de choisir la bonne longueur de coupure de l'intervalle d'intégration. Ce modèle ne repose sur aucun paramètre empirique, excepté un facteur d'échelle nécessaire pour prendre en compte la différence de dimension entre l'énergie et le recouvrement des densités. La sommation du produit des densités sur les points critiques de liaisons non-covalentes a également été étudiée

mais semble manquer une partie de l'information. Par ailleurs, l'introduction dans le modèle d'un facteur dépendant de la distance interatomique, qui avait été bénéfique dans de précédentes études [Söderhjelm *et al.*, 2006, Misquitta et Stone, 2016, Rackers et Ponder, 2019], n'a pas permis d'améliorer davantage les excellents résultats de la méthode d'intégration axiale. Dans le but de tester la validité des densités électroniques reconstruites à partir des paramètres ELMAM2, des paramètres empiriques dépendant de l'espèce chimique ont été ajoutés comme coefficients devant les densités dans l'expression du recouvrement. Ces paramètres restent très proche de l'unité comme attendu au vu des résultats de la partie électrostatique, et améliorent significativement les résultats avec $R^2 = 0,969$ et une erreur absolue moyenne de 0,89 kcal/mol, qui est maintenant inférieure à la précision chimique attendue. L'analyse de l'influence de la composition des dimères étudiés sur les résultats du modèle n'a pas permis de révéler de dépendance particulière. La qualité du potentiel d'échange-répulsion semble essentiellement dépendre de la méthode d'intégration choisie pour obtenir le recouvrement des densités. Une perspective d'amélioration possible est donc une optimisation de la paramétration de cette intégrale.

Le potentiel de van der Waals ELMAM est défini comme la somme des potentiels de dispersion et d'échange-répulsion ELMAM. En prenant les modèles sans paramètre empirique de ces deux termes, les résultats obtenus sont corrects, $R^2 = 0,915$ et une erreur absolue moyenne de 1,21 kcal/mol, mais pas aussi bons que pour les contributions individuelles. En considérant les potentiels de dispersion et d'échange-répulsion avec les paramètres empiriques dépendant de l'espèce chimique, les résultats sont améliorés avec $R^2 = 0,932$ et une erreur absolue moyenne de 1,00 kcal/mol atteignant la précision chimique typique. Ces résultats restant inférieurs à ceux des contributions prises séparément, des coefficients de combinaison linéaire de ces deux termes ont été affinés. Ils permettent d'améliorer les résultats du modèle sans paramètre, avec $R^2 = 0,925$ et une erreur absolue moyenne de 1,10 kcal/mol, mais pas ceux avec paramètres empiriques. L'étude de l'influence de la composition des dimères sur le potentiel de van der Waals a permis de mettre en évidence d'excellents résultats pour les systèmes dominés par des interactions électrostatiques coulombiennes pour lesquels, la contribution d'échange-répulsion étant majoritaire, l'erreur sur l'énergie de van der Waals est similaire à celle de cette contribution. En revanche, pour les systèmes dominés par des effets dispersifs, les contributions de dispersion et d'échange-répulsion sont semblables en valeurs absolues mais, puisque leurs signes sont opposés, leur somme est du même ordre de grandeur que l'erreur absolue moyenne, causant une décorrélation avec les valeurs d'énergie de référence. Une piste pour éviter cet effet en réduisant l'erreur serait d'entraîner le modèle de van der Waals directement contre les valeurs $E_{\text{vdw}}^{\text{SAPT}}$, plutôt que les contributions individuelles.

Pour le potentiel d'interaction total ELMAM, c'est-à-dire la somme des quatre contribution électrostatique, d'induction, de dispersion et d'échange-répulsion, un effet similaire est observé. De même que pour les systèmes dominés par des interactions dispersives, les systèmes dominés par des effets coulombiens présentent des énergies d'interaction électrostatique et d'induction dont la somme est du même ordre que la contribution d'échange-répulsion en valeurs absolues. Par conséquent, le potentiel total prend de faibles valeurs, souvent proches des erreurs absolues moyennes sur les différentes contributions, ce qui réduit considérablement la corrélation entre les résultats du modèle ELMAM et les valeurs de référence. A. Gavezzotti, qui avait proposé

un modèle semblable, a déjà remarqué que « *la partition de l'énergie en composantes révèle que les énergies totales sont des sommes de larges contributions de signes opposés* », et que même si ces modèles ne sont pas assez précis pour que l'énergie totale soit directement employée, ils permettent d'estimer les contributions individuelles et de les comparer entre elles de façon à pouvoir discuter la nature des interactions dominantes dans un système [Gavezzotti, 2003]. C'est dans cet esprit que sont employés les calculs d'énergies d'interaction protéine-ligand qui seront présentés dans le chapitre suivant. Notons qu'il reste cependant des pistes pour améliorer le potentiel ELMAM, notamment en complétant le terme d'induction qui pour l'instant ne rend compte que des effets dipolaires, et en suivant les perspectives d'amélioration des potentiels de dispersion et d'échange-répulsion proposées ci-dessus.

Chapitre 4

Applications aux complexes protéine-ligand

Dans le cadre de ma thèse, j’ai développé des méthodologies qui ont pour vocation d’être employées dans l’étude des macromolécules biologiques. Le premier type de développement est celui des descripteurs issus de l’analyse topologique du potentiel électrostatique : les points critiques de $V(\mathbf{r})$, les faisceaux primaires et les zones d’influence électrophile et nucléophile. Ces descripteurs étant basés sur le gradient de $V(\mathbf{r})$, ils révèlent par une représentation intuitive de la topographie des lignes de champ électrique les contributeurs principaux aux forces électrostatiques mais aussi l’étendue spatiale et la direction de ces forces. L’interprétation de ces descripteurs dans les complexes protéine-ligand possède un potentiel prometteur pour la compréhension des mécanismes de reconnaissance moléculaire de nature électrostatique, lors de l’approche d’un ligand vers la poche de fixation d’une protéine par exemple [Wade *et al.*, 1998]. Le second travail méthodologique réalisé pendant ma thèse est le développement d’un modèle d’énergie d’interaction de van der Waals (dispersion et répulsion) reposant sur les quantités transférables de la librairie ELMAM2. Ce modèle complète les potentiels d’interaction ELMAM électrostatique et d’induction déjà établis [Leduc *et al.*, 2019, Vuković *et al.*, 2021]. Les différentes contributions aux interactions protéine-ligand peuvent notamment être quantifiées grâce à ces modèles d’énergie ELMAM.

Dans ce chapitre, ces méthodologies sont appliquées à divers systèmes protéiques choisis. Tout d’abord, différentes utilisations possibles des descripteurs issus de la topologie du potentiel électrostatique pour la caractérisation d’un complexe enzyme-inhibiteur sont montrées par l’analyse d’un système qui a été largement étudié, la trypsine bovine en complexe avec un inhibiteur canonique [Wahlgren *et al.*, 2011]. Ensuite, un exemple de l’apport de ces descripteurs dans des travaux de biologie structurale est présenté dans le cadre de l’étude de la neuropiline en complexe avec un inhibiteur peptidique [Goudiaby *et al.*, 2023]. Puis, l’emploi des calculs d’énergies électrostatiques ELMAM dans les complexes protéine-ligand est illustré dans le cas d’une glutathion transférase de cyanobactérie, dont nous avons résolu la structure en complexe avec le ligand glutathion [Mocchetti *et al.*, 2022]. Pour finir, une perspective d’application prometteuse à un système dynamique des zones d’influence et des calculs d’énergies d’interaction ELMAM est décrite sur la base du mécanisme de transport membranaire de l’ion chlorure par l’halorhodopsine [Mous *et al.*, 2022].

4.1 Intérêts de l'étude des zones d'influence : complexe Trypsine Bovine - Inhibiteur SGPI

4.1.1 Contexte de l'étude

Dans l'article [Mocchetti *et al.*, 2023] (non publié au dépôt de ce manuscrit), nous présentons les méthodologies développées pendant ma thèse pour la définition des descripteurs issus de la topologie du potentiel électrostatique. Pour illustrer différents intérêts de l'emploi de nos nouveaux descripteurs dans les études des complexes protéine-ligand, nous avons analysé du point de vue des zones d'influence électrophile (ZIE) et nucléophile (ZIN) un complexe de Michaelis de la trypsine de bovin qui est une protéase à sérine dont le mécanisme catalytique est très connu [Hedstrom, 2002]. En particulier, nous avons choisi le complexe de la trypsine avec l'inhibiteur SGPI-1-PO-2 qui est un inhibiteur peptidique canonique [Wahlgren *et al.*, 2011].

Les protéases sont des enzymes impliquées dans le clivage de liaisons peptidiques [Hedstrom, 2002]. Les protéases à sérine se distinguent par la présence dans leur site actif d'une triade catalytique : sérine (Ser195) – histidine (His57) – aspartate (Asp102)¹. Le réseau de liaisons hydrogène entre ces trois résidus serait à l'origine de l'activation de la sérine catalytique Ser195 attaquant la liaison peptidique du substrat à cliver. Cette attaque nucléophile est suivie de deux réactions, l'acylation et la dé-acylation, faisant intervenir plusieurs intermédiaires tétraédriques [Liu *et al.*, 2006, Radisky *et al.*, 2006]. Ces intermédiaires portent une charge négative stabilisée dans le trou oxyanion formé par les influences électrophiles des protons amides des résidus Gly193 et Ser195 [Ménard et Storer, 1992]. Les trypsines sont des protéases à sérine produites sous forme inactive par le pancréas et impliquées dans les processus de digestion [Hirota *et al.*, 2006]. Elles clivent spécifiquement les acides aminés portant une charge positive (lysine ou arginine) dont la spécificité est attribuée à la charge négative du résidu Asp189 situé au fond de la poche de fixation [Graf *et al.*, 1987]. Les complexes de la trypsine avec des inhibiteurs canoniques permettent de mimer la structure du complexe de Michaelis [Ascenzi *et al.*, 2003, Krowarsch *et al.*, 2003].

Pour appliquer nos descripteurs, nous avons choisi la structure de la trypsine bovine en complexe avec un inhibiteur canonique de [Wahlgren *et al.*, 2011] (code PDB 2XTT), résolue par diffraction des rayons X à résolution atomique (0,93Å), car les états de protonation des résidus catalytiques sont discutés. En effet, dans le modèle de [Wahlgren *et al.*, 2011], le résidu catalytique Asp102 est neutre (groupement COOH), contrairement à d'autres travaux où Asp102 apparaît avec une charge négative (forme COO⁻) [Schmidt *et al.*, 2003, Kawamura *et al.*, 2011, Liebschner *et al.*, 2013, Schiebel *et al.*, 2018]. L'inhibiteur est un peptide mutant de 35 acides aminés appelé SGPI-1-PO-2 qui a été obtenu par évolution dirigée de l'inhibiteur sauvage SGTI (S. Gregaria Trypsin Inhibitor).

4.1.2 Caractérisation de la trypsine du point de vue des zones d'influence

Une première analyse des énergies d'interaction électrostatique entre les résidus de l'enzyme et ceux de l'inhibiteur a permis d'identifier les principales interactions stabilisant le complexe et

1. La numérotation des résidus utilisée ici suit la numérotation basée sur la séquence de la chymotrypsine qui est communément employée pour toutes les protéases à sérine [Greer, 1981].

d'orienter le choix des zones d'influence étudiées par la suite. La description des zones d'influence associées aux atomes (ou groupements d'atomes) de la trypsine a mis en évidence plusieurs utilisations possibles de ces descripteurs pour l'étude des complexes protéiques. En effet, les résidus impliqués dans la fixation du substrat sont révélés par la directionnalité de leurs zones d'influence, même lorsque ceux-ci ne sont pas caractérisés par des analyses classiques de la structure atomique. Par exemple, la partition de l'espace de la poche de fixation du substrat dans la trypsine en ZIN a mis en évidence le rôle majeur de l'Asp189 vers lequel les lignes de champ électrique convergent à l'intérieur de la cavité. Ces lignes de champ émanent d'une multitude de sites électrophiles localisés sur les atomes d'hydrogène de résidus répartis autour du site actif et participent donc également aux forces fixant le substrat bien que ceux-ci n'aient pas été mentionnés dans des travaux précédents.

De même, les contributeurs aux forces électrostatiques participant au guidage de l'approche du substrat à travers le solvant peuvent être identifiés aisément car ceux-ci présentent des zones d'influence qui s'étendent au-delà de la surface de la protéine. Dans le cas de la trypsine, nous avons ainsi mis en évidence les rôles d'attracteur de groupements nucléophiles du proton amide du résidu Gly193 stabilisant le trou oxyanion et du groupement amine de la Lys224 fixant le résidu Asp22I de l'inhibiteur.

De plus, l'identification de ces résidus peut être utilisée pour prédire des interactions enzyme-substrat et suggérer des mutations pouvant améliorer l'affinité de leur complexe. Notamment, pour la trypsine en complexe avec l'inhibiteur muté de [Wahlgren *et al.*, 2011], la ZIE du groupement amine de la Lys224 montre que les lignes de champ, qui s'étendaient vers le solvant dans la protéine apo, sont concentrées vers le résidu 22 de l'inhibiteur ayant subi la mutation T22D. Cette mutation du résidu de l'inhibiteur pour lui attribuer une charge négative aurait donc pu être suggérée à partir de la visualisation de cette ZIE dans la structure apo. En effet, son extension au-delà de la surface de la protéine traduit la capacité de la Lys224 à attirer vers elle un groupement nucléophile et à le fixer.

Par ailleurs, la représentation des zones d'influence peut être utilisée pour discuter les mécanismes enzymatiques lorsque ceux-ci sont dominés par des effets électrostatiques. Les ZIE des protons amide de Gly193 et Ser195 ont permis de proposer que la stabilisation du trou oxyanion serait assurée essentiellement par Gly193. L'influence de l'état de protonation du résidu catalytique Asp102 sur son interaction avec His57 a également pu être discutée par comparaison de la ZIN du groupement carboxyle dans le cas où celui-ci est sous forme non-protonée (COO^-) par rapport au cas protoné (COOH). En effet, ces ZIN montrent que, dans la forme COO^- , Asp102 est l'unique contributeur électrostatique à interagir avec l'atome H δ 1 de l'histidine catalytique, jouant donc un rôle important dans le mécanisme d'activation de Ser195, ce qui est cohérent avec les résultats de mutagenèse dirigée qui ont montré le caractère essentiel de Asp102 pour l'activité enzymatique [Sprang *et al.*, 1987]. En revanche, dans la forme COOH , l'influence de Asp102 semble écrantée et His57-H δ 1 interagit avec plusieurs partenaires, dont Ser214, Val213 et Tyr94, bien que par exemple la Ser214 ne soit pas nécessaire pour la catalyse [McGrath *et al.*, 1992, Epstein et Abeles, 1992].

4.1.3 Draft de l'article décrivant les descripteurs issus de la topologie du potentiel électrostatique et leur application à la trypsine

Graphical descriptors of protein electrostatics based on the topology of the molecular electrostatic potential.

Eva Mocchetti, Claude Didierjean and Benoit Guillot.

Abstract

A new paradigm for representing and interpreting electrostatic properties in protein structures is proposed, based on the topology of the molecular electrostatic potential. This approach allows for the partitioning of molecular space into Electrophilic or Nucleophilic Influence Zones, which are volumes encompassing all electric field lines associated with electrophilic or nucleophilic sites. Visualizing these influence zones offers intuitive graphical descriptors that illustrate the spatial extent of electrostatic forces and pinpoint, at an atomic level, the contributions of nucleophilic and electrophilic sites to the electrostatic environment found in and around protein structures.

This article presents the development of these novel electrostatic descriptors and their integration into the MoProViewer software. An application example is provided, featuring a bovine Trypsin complexed with a modified peptide inhibitor. The electrostatic influence zone perspective is utilized to discuss molecular recognition, electrostatic steering, and enzymatic mechanisms.

Introduction

Visual examination by a researcher of the structure of a biological macromolecule using the graphical representation of its 3D model is essential in structural biology [1]. Even in the era of powerful artificial intelligences, the “*Human brain aided*” interpretation of protein crystal structures provides valuable clues regarding structure-function relationships [2].

The representation of protein structures using computer graphics can be augmented by visual descriptors of physical properties, which are relevant to the understanding of important phenomena like ligand binding or enzymatic mechanisms. For instance, when studying a protein-ligand complex, interactions between a drug and its target can be visualized at the atomic level through graphical descriptors indicating the nature of the interactions. At much lower resolution, graphical representations using ribbons and arrows denoting α -helices and β -strands respectively, enable to visually apprehend the relative orientations of secondary structure elements. Voids within protein structures are better visualized as 3D volumes, offering direct insights into their shape, their extent and therefore their possible role in the macromolecule function [3]. As exemplified here, graphical descriptors can take many forms to convey information. From them, a specialist can gain intuitive understanding and appreciation of molecular properties, allowing to solve problems, to formulate hypothesis or simply to spark ideas.

It is well known that electrostatic properties play crucial roles in most biomolecular processes [4]. Among electrostatic properties, the most subjected to graphical representation and to visual interpretation is, by far, the molecular electrostatic potential $V(\mathbf{r})$. In structural biology (and related fields), the molecular electrostatic potential is usually represented as projected on a molecular surface, either solvent accessible/exclusion surface or electron density iso-surface. Adequate colour-mapping techniques allows to highlight the location of electropositive and electronegative patches on the molecular surface. $V(\mathbf{r})$ can be computed for instance from point charges distributions as defined in

molecular mechanics force fields [5], using Poisson Boltzmann equation solvers [6] (Figure 1a), or using electron density [7,8] (Figure 1b) or localized orbitals parameters [9,10] transferred from available databases. This surface representation is usually interpreted from an electrostatic complementarity perspective. For instance, the inner surface of a protein binding pocket can be colour-mapped by values of the protein electrostatic potential, allowing to visually appreciate its propensity to accommodate negatively or positively charged groups of a putative ligand. Therefore most, if not all, molecule viewer software used in the structural or molecular biology field feature the possibility to compute a protein surface, then to colour it by values of the protein electrostatic potential (for instance PyMol [11], VMD [12] or Chimera [13]). However, even if indeed popular and intuitive, this representation of electrostatic potential limits it extends to a molecular surface, a bi-dimensional object even though $V(\mathbf{r})$ is by essence a three-dimensional property. A lot of insights that $V(\mathbf{r})$ could provide are therefore hidden by limiting its representation to a qualitative surface projection.

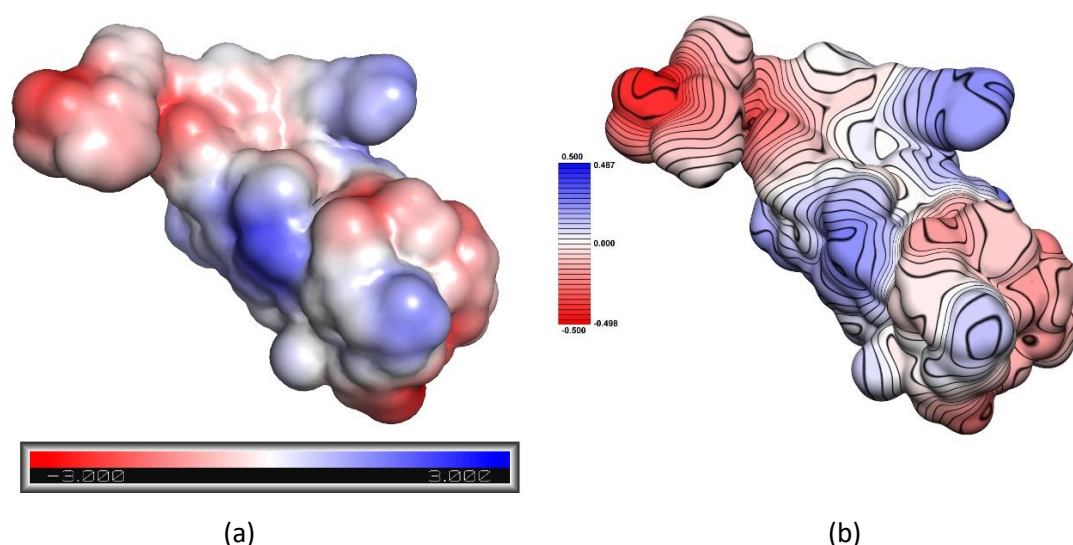


Figure 1: Examples of molecular surfaces coloured by values of molecular electrostatic potentials for the SGPI-1-PO-2 modified trypsin inhibitor as extracted from the 2XTT PDB structure.

(a) Solvent accessible surface colored by $V(\mathbf{r})$ values calculated by solving the Poisson-Boltzmann equation using the APBS plugin [16] of the PyMol software [11]. Computation was performed at 310K, using 150 mM solvent concentration for positive (+1) and negative (-1) ions. Dielectric constant values of 2.0 and 78.0 were used for protein and solvent regions, respectively. Color ranges vary from -3.0 (red) to +3.0 (blue) units of kT/e. This image was rendered in PyMol, using default settings.

(b) Electron density iso-surface ($10^{-3} \text{ e.}\text{\AA}^{-3}$) coloured by $V(\mathbf{r})$ computed *in vacuo* with the Charger module [14] of the MoProViewer software [15], using a multipolar electron density model transferred from the ELMAMII database [7]. Color ranges vary from -0.5 (red) to +0.5 (blue) $\text{e.}\text{\AA}^{-1}$. Image was produced with MoProViewer without specular reflections.

The molecular electrostatic potential is indeed a 3D scalar field displaying a rich topology [17]. This topology should, using adequate graphical descriptors, convey more information than the usual surface projection. It is known that the topology of a scalar field provides abstract descriptive features which allows to simplify its representation and to make more intuitive its interpretation [18,19]. For instance, in Quantum Crystallography (a scientific field at the fringe between ultrahigh-resolution crystallography and quantum chemistry [20,21]), molecular properties and intermolecular interactions are often studied through the topology of another 3D scalar field: the molecular electron density $\rho(\mathbf{r})$. The topological analysis of $\rho(\mathbf{r})$ benefits of a mature theoretical frame and is rigorously defined in the Quantum Theory of Atoms In Molecules (QTAIM) [22]. Briefly, it relies on the partition of the molecular space into atomic volumes, called atomic basins, limited by surfaces of zero-flux of the electron density gradient $\nabla\rho(\mathbf{r})$. This approach is similar to the determination of Morse-Smale complexes in differential

topology [23], used to provide compact and simplified representation of large scientific datasets [24]. In the case of $\rho(\mathbf{r})$, this leads to a complete partition of the molecular electron density, allowing to define molecular properties as additive atomic contributions. Electron density topology also relies on the location of Critical Points (CPs), points in the molecular space where $\nabla\rho(\mathbf{r}) = \mathbf{0}$. At critical point positions, $\rho(\mathbf{r})$ local curvatures (characterized by eigenvalues of the $\rho(\mathbf{r})$ Hessian matrix) present features allowing to define four types of critical points in $\mathbb{R}^3$. They can be found on local $\rho(\mathbf{r})$ maxima (nuclei positions), on local minima (in molecular cages), in atomic rings and, most importantly, between interacting atoms, whether they are covalently bonded, or not. These latter so-called bond critical points, characterized by a positive curvature in the inter-nuclei directions, and two perpendicular negatives ones, are usually represented with their bond paths *i.e.*, ridges of maximum electron density values passing through saddle critical points and joining atomic nuclei (Fig. 2a). The presence of such CP at $\mathbf{r} = \mathbf{r}_{cp}$ indicates an interatomic interaction and the values of $\rho(\mathbf{r}_{cp})$ and $\nabla^2\rho(\mathbf{r}_{cp})$ (the Laplacian of $\rho(\mathbf{r})$) inform on the nature and the strength of the interaction, notably in case of hydrogen bonds [25]. Non-Covalent Interactions (NCI) analysis is another method based on $\rho(\mathbf{r})$ and its topology, which is nowadays commonly used in Quantum Crystallography [26]. It relies on the computation of the Reduced Density Gradient $s(\mathbf{r})$, a functional of $\rho(\mathbf{r})$ and $\nabla^2\rho(\mathbf{r})$. Isosurfaces of $s(\mathbf{r})$, once colored by a quantity related to the local curvature of $\rho(\mathbf{r})$, enable real-space visualization of both attractive and repulsive interactions [27] (Fig. 2b). Topological features of $\rho(\mathbf{r})$ can therefore provide graphical descriptors of interatomic interactions in protein structures.

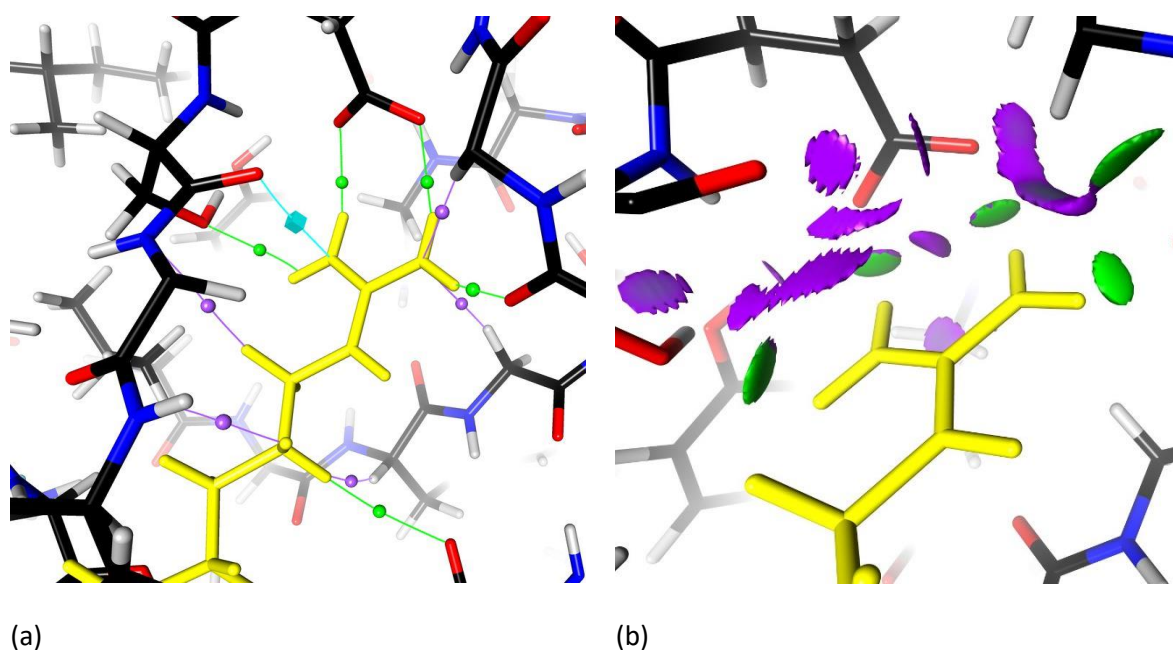


Figure 2: Graphical descriptors based on $\rho(\mathbf{r})$ topological features highlighting interatomic interactions between Arg29 of SGPI-1-PO-2 modified trypsin inhibitor (displayed in yellow) and the bovine trypsin (PDB code 2XTT).

(a) Interatomic interactions are highlighted by their corresponding electron density critical points and bond paths. Critical points indicating hydrogen bonds are represented as green spheres and van der Waals H...N interactions as purple spheres. The van der Waals contact between Ser190 O γ and a nitrogen atom of Arg29 guanidinium group is shown as blue cube.

(b) Isosurfaces of the Reduced Density Gradient ($s = 0.5$) coloured by the value of the total molecular electron density as preconized in the NCI analysis approach [27]. Hydrogen bonds and strong interactions appear in green, while regions of favourable van der Waals contacts are shown in purple.

These descriptors were computed and displayed with the MoProViewer software, using a multipolar electron density model transferred from the ELMAMIII database [7].

In the case of $V(\mathbf{r})$, there is no such elaborated theory as the QTAIM. Yet, several authors defined and discussed key concepts of its topology [17,28,29]. As for $\rho(\mathbf{r})$, the topology of $V(\mathbf{r})$ relies on the determination of critical points positions and of zero-flux basins of its gradient vectors $\nabla V(\mathbf{r})$. However, unlike in $\rho(\mathbf{r})$ topology, $V(\mathbf{r})$ gradient has an immediate physical meaning as it is related to the electric field $\mathbf{E}(\mathbf{r})$ through the relationship $\mathbf{E}(\mathbf{r}) = -\nabla V(\mathbf{r})$. In other words, the topology of $V(\mathbf{r})$ is a mean to rationalize the topography of electric field lines emerging from a molecular charge distribution. While the topology of the molecular electrostatic potential has already been used to study properties of small molecules [30–33], it is not yet the case for macromolecular systems. Surprisingly, even though $\mathbf{E}(\mathbf{r})$ is a fundamental, measurable quantity [34,35], the use of electric field lines as graphical descriptors of molecular electrostatic properties appears rarely in the structural biology literature. Few recent examples can be found [36–38] where $\mathbf{E}(\mathbf{r})$ is represented as field lines and at low resolution i.e., around whole protein structures. They are in these studies interpreted on the basis of their local density, qualitatively indicating the strength of the electric field.

However, analysis of $\mathbf{E}(\mathbf{r})$ distribution through the topology of $V(\mathbf{r})$ can convey more information. In fact, according to fundamental electrostatics, electric field lines either connect electronegative and electropositive sites or emanate from or converge towards infinity. Therefore, $\mathbf{E}(\mathbf{r})$ manifests as bundles of field lines either connecting electrophilic (positively charged) and nucleophilic (negatively charged) critical points, or extending towards the solvent region. As a result, in the case of a protein binding pocket for instance, the relative positions of both nucleophilic and electrophilic sites play a role in establishing the particular electric field lines distribution representing directions and magnitudes of electrostatic forces and polarization effects a bound ligand would experience.

Given the abundance of charged sites within macromolecular structures, the distribution of electric field lines therein becomes inherently complex. In order to simplify $\mathbf{E}(\mathbf{r})$ representation and to allow its intuitive interpretation, we propose in this article new graphical descriptors issued from the topology of $V(\mathbf{r})$, the property from which $\mathbf{E}(\mathbf{r})$ derives. These descriptors are based on the notion of Electrophilic and Nucleophilic Influence Zones [29], which are regions of space that contain all electric fields lines converging, or emanating from a given electrophilic or nucleophilic site, respectively. The representation of the corresponding volumes, within a protein structure, allows to display the spatial extend of the electrostatic influence of a given charged group, but also indicates which complementary sites (in terms of electrophilic or nucleophilic character) are required to properly shape these volumes. Therefore, this descriptor allows visual insights related to the structure-property relationship, where the structure is represented by the 3D distribution of electrophilic and nucleophilic sites, and the property refers to the electrostatic environment this distribution creates in terms of electrostatic forces.

This approach has been implemented in the MoProViewer software [15], the molecule and molecular properties viewer included in the MoProSuite program package [39,40]. In MoProViewer, a user-friendly tool allows the user to select a chemical group (for instance within a residue side chain) and to compute either its electrophilic or nucleophilic influence zones, displayed as 3D volumes superimposed to the structure of the protein in exam.

Next in this article, we will provide a brief description of some key aspects of $V(\mathbf{r})$ topology from which we will outline in more details the concept of Nucleophilic and Electrophilic Influence Zones, their corresponding graphical descriptors and their practical implementation in the MoProViewer software. In the final section, we will demonstrate a practical application to a model system, the bovine trypsin structure bound to a modified high affinity peptide inhibitor [41]. We will discuss protein-ligand recognition and some aspects of the enzymatic mechanism, focusing on the insights provided by electrostatic potential topology descriptors.

Descriptors from the molecular electrostatic potential topology

Electrostatic potential critical points

The critical points of the electrostatic potential $V(\mathbf{r})$, located at $\mathbf{r} = \mathbf{r}_{cp}$, are defined as $\nabla V(\mathbf{r}_{cp}) = -\mathbf{E}(\mathbf{r}_{cp}) = \mathbf{0}$. As in $\rho(\mathbf{r})$ topology, four types of critical points are defined and classified according to their rank ω (the number of non-zero eigenvalues in the $V(\mathbf{r})$ Hessian matrix) and signature σ (algebraic sum of signs of $V(\mathbf{r})$ Hessian matrix eigenvalues). Critical points are usually denoted using (ω, σ) symbols, where $\omega = 3$ in $\mathbb{R}^3$ and σ depends on the signs of the local curvatures. In molecular electrostatic potential can be found the $(3, -3)$ local maxima, the $(3, +3)$ local minima and the $(3, -1)$ and $(3, +1)$ saddle critical points.

The $(3, -3)$ critical points correspond to local $V(\mathbf{r})$ maxima and are, in a molecular charge density, necessarily localized at nuclei positions¹. The $(3, +3)$ CPs are local minima of $V(\mathbf{r})$. Their nature depends on the charge distribution model used to compute $V(\mathbf{r})$. Using point charges, $(3, +3)$ critical points will be found on atomic nuclei bearing negative partial charges. In a more physically grounded charge distribution, e.g., obtained from *ab-initio* computations or reconstructed using transferable electron density parameters, $(3, +3)$ CPs will arise from local concentrations of electronic charge such as on electron lone pairs [43,44]. Therefore, the local maxima and minima, highlighted by their associated $(3, -3)$ and $(3, +3)$ critical points, correspond to the electrophilic and nucleophilic sites in $V(\mathbf{r})$, respectively.

The $(3, +1)$ and $(3, -1)$ saddle critical points are located in $\rho(\mathbf{r})$ topology in atomic rings, and on interatomic bonds, respectively. Similar CPs appear in $V(\mathbf{r})$, but others saddle points can be found at the border of topological basins, owing to the complexity of $V(\mathbf{r})$ topology [28,45,46]. Their significance and interpretation have been discussed. For instance, $V(\mathbf{r})$ curvatures on $(3, +1)$ critical points found in hydrocarbon molecules have been related to their degree of aromaticity [33]. It was also proposed that the saddle critical points found at the border of electrostatic potential basins could indicate preferred reaction paths in nucleophilic or electrophilic attack for $(3, -1)$ and $(3, +1)$ CPs respectively [29], as they correspond to $V(\mathbf{r})$ local extrema on basins surfaces.

Electric field lines and primary bundles

The gradient of the electrostatic potential $V(\mathbf{r})$ corresponds to the electric field $\mathbf{E}(\mathbf{r}) = -\nabla V(\mathbf{r})$. In a neutral system, all electric field lines start on a $(3, -3)$ local maximum and end on a $(3, +3)$ local minimum. Electric field lines with a common pair of local maximum and minimum form a bundle surrounded by S_V , a surface of $\nabla V(\mathbf{r})$ zero-flux, defined by $\mathbf{E}(\mathbf{r}) \cdot \mathbf{n}(\mathbf{r}) = 0, \forall \mathbf{r} \in S_V$ with $\mathbf{n}(\mathbf{r})$, the normal vector to S_V . This bundle has a finite volume and is named a *closed primary bundle*. Interestingly, given this zero-flux condition, the Gauss theorem in classical electrostatics implies that the total charge included in volumes delimited by S_V should be null. Moreover, using the relationship $E_{self} = \frac{1}{2} \epsilon_0 \int |\mathbf{E}|^2 dV$ where the integral is taken over the volume bounded by S_V , one can determine the contribution of the corresponding primary bundle to the total self-electrostatic energy E_{self} of the charge distribution from which $V(\mathbf{r})$ was derived. If the system bears a non-zero global charge, some

¹ As pointed out by Gadre and Pathak [42], the existence of $(3, -3)$ critical points is actually not possible in a molecular electrostatic potential map. Local maxima are explicitly prohibited since they would result in negative values of $\rho(\mathbf{r})$, according to the Poisson equation. Maxima associated to atomic nuclei (or positively charged point charges) correspond to points in space where $V(\mathbf{r})$ diverges and are therefore undefined. Nevertheless, $(3, -3)$ CPs can still be found *numerically* in a discretized $V(\mathbf{r})$ scalar field, which is the context referred to in this study.

electric field lines will flee to infinity and extend their electrostatic influence through the intermolecular space so that they could drive the approach of a charged group [31]. A bundle formed by this type of electric field lines extends in infinite volumes and is named *open primary bundle*. The primary bundles are the topological basins of the electrostatic potential and, unlike the atomic basins in $\rho(\mathbf{r})$ which are associated to a unique atom, they are related to a unique *pair* of electrophilic and nucleophilic sites.

Electrophilic and nucleophilic influence zones

The union of primary bundles associated to a unique electrophilic site or to a unique nucleophilic site defines the electrostatic influence zone of this electrophilic or nucleophilic site [29]. Therefore, an *Electrophilic Influence Zone* (EIZ) contains all the electric field lines emerging from its associated electrophilic site. A probe charge q located inside an EIZ will be attracted toward the electrophilic site if $q < 0$ or repelled away from it if $q > 0$. Similarly, a *Nucleophilic Influence Zone* (NIZ) contains all the electric field lines converging toward its associated nucleophilic site. A probe charge q inside a NIZ will be attracted toward (if $q > 0$) or repelled away from (if $q < 0$) the nucleophilic site. In case of a neutral charge distribution, any point in the molecular space where reign a given electric field (i.e., $\mathbf{E}(\mathbf{r})$ vector at this point location) can be seen as simultaneously under the influence of the nucleophilic and electrophilic sites which are connected by the corresponding electric field line. In non-neutral charge distributions, points in the molecular space can be located in open primary bundles *i.e.*, under the influence of a single nucleophilic, or electrophilic site. In any case, the molecular space can therefore be concomitantly partitioned into EIZ and into NIZ.

Here, we propose the simplification of EIZ and NIZ concepts through the determination and the representation of the bounding surface encompassing the union of their contributing closed and open (if any) primary bundles. This representation has the advantage of overcoming the complex distribution of electric field lines (as found notably in protein structures), revealing the extend and the shape of the region of space globally under the electrostatic influence of the chosen electrophilic or nucleophilic site. Moreover, considering all sites belonging to a given chemical group, the union of their NIZ (or EIZ) would define the corresponding influence zone of the complete chemical group. This is particularly relevant in the studies of macromolecular structures as this allows the representation of electrostatic influence zones associated to, for instance, residues side chains.

Implementation in MoProViewer

In MoProViewer, surfaces delimiting electrostatic influence zones are computed numerically, on the basis of $V(\mathbf{r})$ critical points located in a preloaded 3D scalar field of molecular electrostatic potential. More details of these developments are given in Supplementary Materials. As both electrophilic and nucleophilic sites can be associated to atoms (either to atom nuclei or to electron lone pairs), in MoProViewer it is sufficient to select atom(s) of interest and prompt for the computation of their NIZ or EIZ. Once completed, their corresponding graphical descriptors are represented as triangulated surfaces delimiting their volumes. Their appearance can be configured in terms of colors, transparency level and representation mode as lines, points or filled surfaces. Subdivision, simplification and smoothing algorithms of the resulting triangle meshes are also proposed in MoProViewer, allowing to improve the surfaces aesthetics. If needed, the triangle meshes representing EIZ or NIZ can be exported in the .OFF file format, which can be read by any computer graphics rendering software.

In the following section, to demonstrate the interest of these electrostatic descriptors, we will present an application to the study of a well-known system: a bovine trypsin structure in complex with an inhibitor. We have chosen to use an electrostatic potential computed from a transferred multipolar molecular electron density [7], known to allow the calculation of molecular electrostatic potential of *ab initio* quality [47].

Application to the crystal structure of a Bovine Trypsin Michaelis complex.

Introduction to serine proteases

Serine proteases are ubiquitous enzymes with a widely recognized catalytic mechanism that is documented in all textbooks of structural enzymology, despite the fact that the mechanism has not been completely elucidated to date. They serve as an excellent model for showcasing the potential of our electrostatic descriptors in interpreting both catalytic mechanisms and enzyme-substrate interactions. Serine proteases hold a catalytic triad in their active site: serine (Ser195) – histidine (His57) – aspartate (Asp102). His57, in interaction with Asp102, activates Ser195, enabling it to attack the peptide cleavage site. The nucleophilic attack is followed by an acylation step and a deacylation step, allowing the enzyme to return to its initial state. These reactions involve tetrahedral intermediates [48] that feature an oxyanion stabilized by the amide protons of Gly193 and Ser195 [49].

To illustrate the use of our electrostatic potential topology descriptors, we have chosen an extensively studied serine protease: bovine trypsin. Among the numerous structures of this enzyme in the PDB, we have selected an atomic resolution crystal structure and a mimic of an intermediate state, i.e., the transient complex formed when the enzyme binds the substrate before the reaction (Michaelis complex). The complex of bovine trypsin with SGTI-1-PO-2 inhibitor (Figure 3) proved to be a suitable choice (0.93 Å resolution, PDB entry 2XTT) [41]. The SGTI-1-PO-2 inhibitor was obtained through directed evolution from the wild-type SGTI peptide inhibitor (*S. gregaria* trypsin inhibitor 1) [41]. The five mutations (T5E, N18R, T20G, P21S and T22D) enabled to significantly improve the inhibitory constant, from the micromolar to the picomolar range. Based on the high resolution of the electron density, the authors proposed a non-charged state for the residues of the catalytic triad in the Michaelis complex. In this model, carbonyl O δ 2 of Asp102 residue is hydrogen bonded to the proton of His57-N δ 1 and His57-N δ 2 is hydrogen bonded to the proton of Ser195-O γ . We have revisited the electrostatic influence of the Asp102 side chain on His57 as well as the influence of the amide protons of Gly193 and Ser195 on the stabilization of the oxyanion hole. Additionally, we investigated the interaction between the inhibitor and the enzyme using our original descriptors. The trypsin - SGTI-1-PO-2 inhibitor model has been prepared from the 2XTT PDB structure following the procedure described in Supplementary Materials.



Figure 3: Structure of the Bovine Trypsin – SGPI-2-PO-1 inhibitor complex.

The trypsin protein and the SGPI-2-PO-1 inhibitor structures are represented as purple and lime cartoons, respectively. Side chains of mutated residues in SGPI-2-PO-1 are shown in magenta. Arg29I SGPI side chain, which is deeply buried in the enzyme structure, is represented in green. This image was generated in PyMol [11]

Enzyme-inhibitor electrostatic interaction energies

In the structure of Wahlgren *et al.* [41], the trypsin is complexed with a canonical peptide inhibitor obtained by directed evolution on the wild type SGTI inhibitor. Through the five mutations (T5E, N18R, T20G, P21S and T22D), two negative charges and one positive charge have been added to the inhibitor. These charges are able to form strong hydrogen bonds with the trypsin and thus improve the inhibitory constant. To quickly analyse the enzyme-peptide interactions, an efficient method is to compute the electrostatic interaction energies between the enzyme residues and the peptide residues to identify the most favourable (and unfavourable) contacts.

Electrostatic interaction energies were computed in the trypsin – inhibitor complex by accounting for each residue-residue contribution, on the basis of the transferred multipolar $\rho(r)$ model. The most significant interactions are summarized Figure 4 as an interaction matrix. Two main favourable contacts appear clearly. First, the interaction between trypsin Asp189 residue and inhibitor Arg29I residue is characterized by an electrostatic energy of -31.2 kcal/mol. This contact corresponds to the binding of the Arg29I residue, which precedes the scissile peptide bond, in the cleavage site. Secondly, the trypsin Lys224 residue and inhibitor Asp22I residue present an electrostatic interaction energy of -27.0 kcal/mol. This interaction is particularly interesting to study because the inhibitor 22nd residue acquired a negative charge through the T22D mutation. From an electrostatic interaction energy perspective, other residues which were mutated to produce the high affinity SGPI-2-PO-1 inhibitor do not seem to contribute significantly to the stability of complex (Figure 4), which appear to largely depends on the two main anchor points that are Asp189 –Arg29I and Lys224 –Asp22I contacts.

Together, they represent about 50% of the total electrostatic interaction energy (~105 kcal/mol) between the protein and the inhibitor.

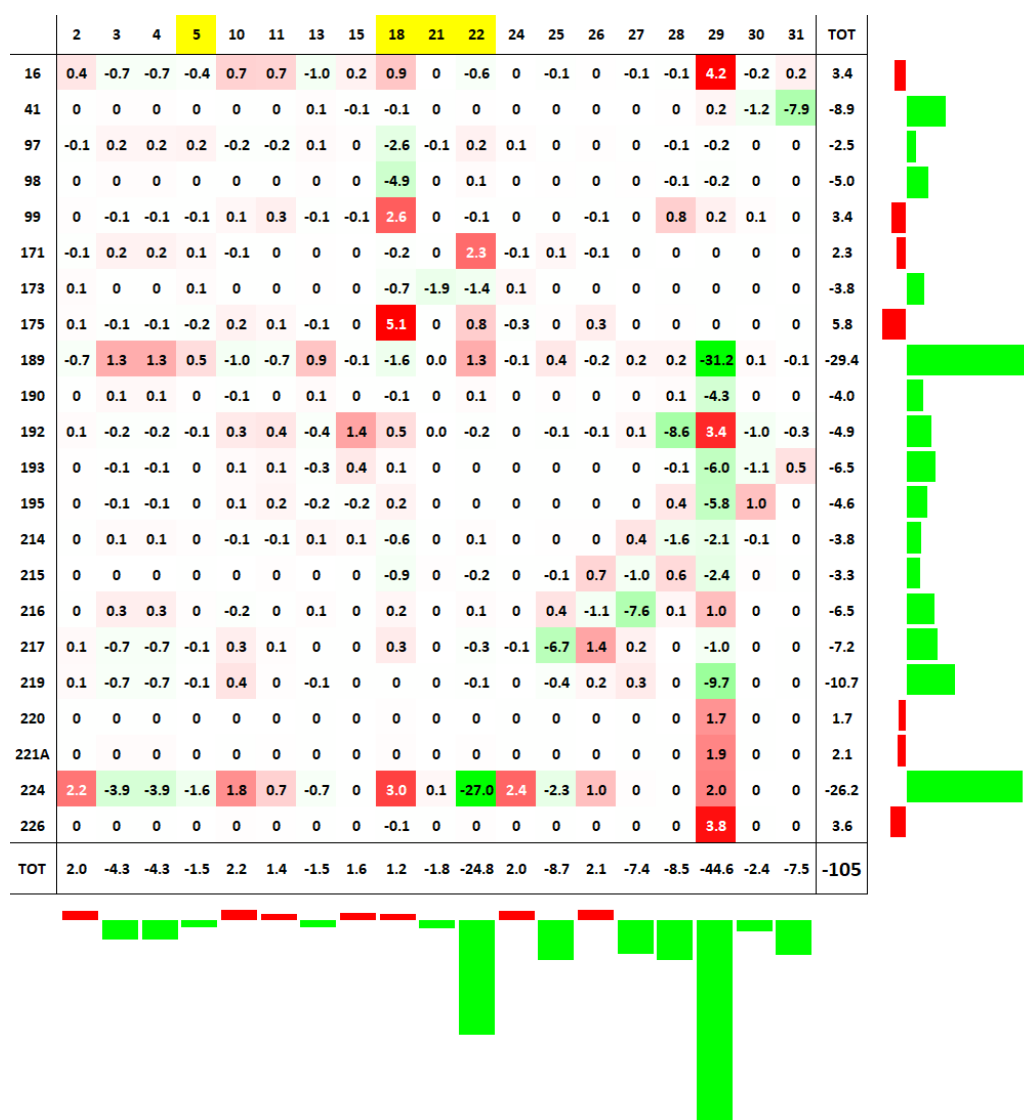


Figure 4: electrostatic interaction map between protein and inhibitor residues.

Electrostatic interaction energies (E_{elec} , kcal/mol) between BT and BTI residues are represented as an interaction matrix. Only residues whose total contribution satisfy $|E_{elec}| > 1$ kcal/mol are kept. Remaining residues in the protein and the inhibitor sequences are shown vertically and horizontally, respectively. Mutated residues in SGPI-2-PO-1 are highlighted in yellow. Vertical and horizontal bar graphs represent total E_{elec} values for a given residue in the protein and the inhibitor sequences, respectively. Overall, favorable ($E_{elec} < 0$) electrostatic energy contributions are displayed in shades of green, while unfavorable ($E_{elec} > 0$) are shown in red.

Molecular recognition from the influence zone descriptor perspective

Specificity of the trypsin binding pocket

Trypsin is specific for the hydrolysis of peptide bonds where the carbonyl group belongs to either a lysine or an arginine residue. The substrate-binding pocket encompasses segments 189-192, 214-220, and 224-228. The substrate specificity is attributed to the presence of a nucleophilic pocket formed by the negative charge of Asp189, along with the carbonyl groups of Gly216 and Gly226 [50–52]. From an energetic standpoint, Asp189 is the primary contributor to the inhibitor binding, while Gly216 has a significant impact, and Gly226 is negligible (Figure 4). We have computed the nucleophilic influence zones (NIZs) associated to the carbonyl oxygen atoms of Gly216 and Gly226, and O δ 1 and O δ 2 atoms of Asp189. A NIZ contains the electric field lines generated by the whole protein but specifically converging onto the nucleophilic site belonging to the associated atom. Thus, an electrophilic group within this volume would experience attractive electrostatic forces directed towards this atom.

The NIZ of the Asp189-O δ 2 atom, represented by the red surface in Figure 5, encompasses the position that the side chain of the substrate arginine residue preceding the scissile peptide bond is supposed to take. The electrostatic influence of this specific atom thus appears as the main nucleophilic contribution stabilizing the positive charge of Arg291 guanidinium group. The NIZ of the Gly216-O atom fills a volume of 88 Å³ outside the protein surface, including the positions of the inhibitor backbone for Cys271 and Thr281 residues so that the electric field lines contained in this NIZ may be important for substrate binding but not for the enzymatic specificity. The Gly226-O and Asp189-O δ 1 NIZs cover volumes of 40 and 132 Å³ respectively, which are buried inside the protein and do not include the substrate position.

Furthermore, the Asp189-O δ 2 NIZ extends over a large volume of 121 Å³ in the binding pocket and points to many electropositive atoms of the active sites. These electrophilic sites are marked with blue spheres in the Figure 5 and are mainly protons of the segments of the active site (189-192, 214-220, and 224-228). Asp189 has therefore a strong influence on the whole active site, and it is not surprising that its mutation to alanine or asparagine leads to a loss of activity [53]. Our analysis also agrees with the recent findings on the catalytic mechanism of thrombin, a trypsin-like enzyme [54]. This study indicates that the charges of Asp102 and Asp189 are crucial for the efficiency of the active site. Trypsin-like enzymes are believed to exist in both closed and open forms at equilibrium. When the charges of Asp102 and Asp189 are removed, the equilibrium shifts towards the closed form, resulting in a displacement of the 215-217 segment and impeding substrate entry. These new insights were considered as “*new paradigm for the control of the equilibrium of the open and closed forms in the trypsin fold*” [54].

The last but not least is the volume of the Asp189-O δ 2 NIZ that extends to the entry of the substrate binding-pocket. This indicate that the approach of the positive charge of the substrate into the binding pocket can be driven by the electrostatic forces oriented toward Asp189-O δ 2. The electrophilic sites surrounding the active site and the Asp189-O δ 2 nucleophilic site could maintain the positive charge of the substrate in the bottom of the binding pocket.

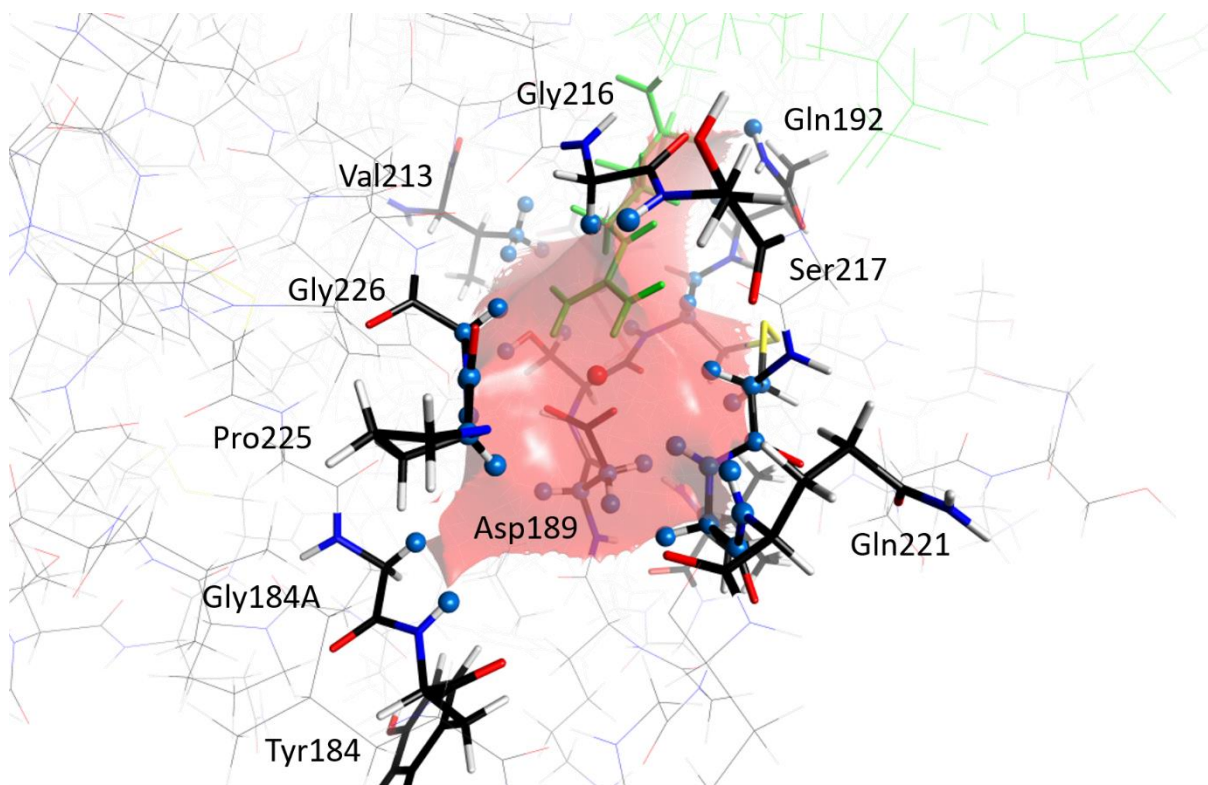


Figure 5: Nucleophilic influence zone associated to the O δ 2 atom of the trypsin Asp189 residue.

The nucleophilic influence zone (red surface) associated to the nucleophilic site (red sphere) corresponding to one electron lone pair of the Asp189-O δ 2 atom has been computed from the electrostatic potential generated by the contributors of the whole apo protein (excluding the contribution of the inhibitor, represented in green). This NIZ extends over a large volume (121 Å³), thanks to the presence of numerous electrophilic sites (blue spheres) surrounding the binding pocket. These electrophilic sites are located on the main chains of Gly184A, Tyr184, Cys191, Gly216, Ser217, Cys220, Ala221A, Gln221, Pro225 and Gly226 residues, and the side chains of Val17, Asp189, Ser190, Gln192, Asp194, and Val213 residues.

Trypsin – inhibitor electrostatic complementary

To study the Lys224 – Asp221 interaction, which presents the second strongest electrostatic interaction energy, from the electrostatic influence zone perspective, we have computed the EIZ associated to the electrophilic sites of Lys224 amine group (see Figure 6). In the case where only the contributors of the apo protein have been accounted for the electrostatic potential calculation (Figure 6.a), this EIZ contains electric field lines emerging from the Lys224 amine group which extend outside the protein surface to the solvent. The visualization of this electric field lines enables to predict that a nucleophilic group (such as a negatively charged residue) can be driven from the solvent to this amine group by the corresponding attractive electrostatic forces, then by this residue. Interestingly, when the contributors of the inhibitor are included in the computing of the electric field lines (Figure 6.b), the EIZ of the Lys224 amine group no longer extend outside the protein surface. All electric field lines emanating from Lys224 NH₃⁺ converge on the nucleophilic sites of the inhibitor Asp221 carboxylate group, showing the electrostatic forces now stabilizing this interaction. The hydrogen bond is moreover characterized as expected by a (3, -1) critical point of the electrostatic potential, displayed by a green sphere in the Figure 6b.

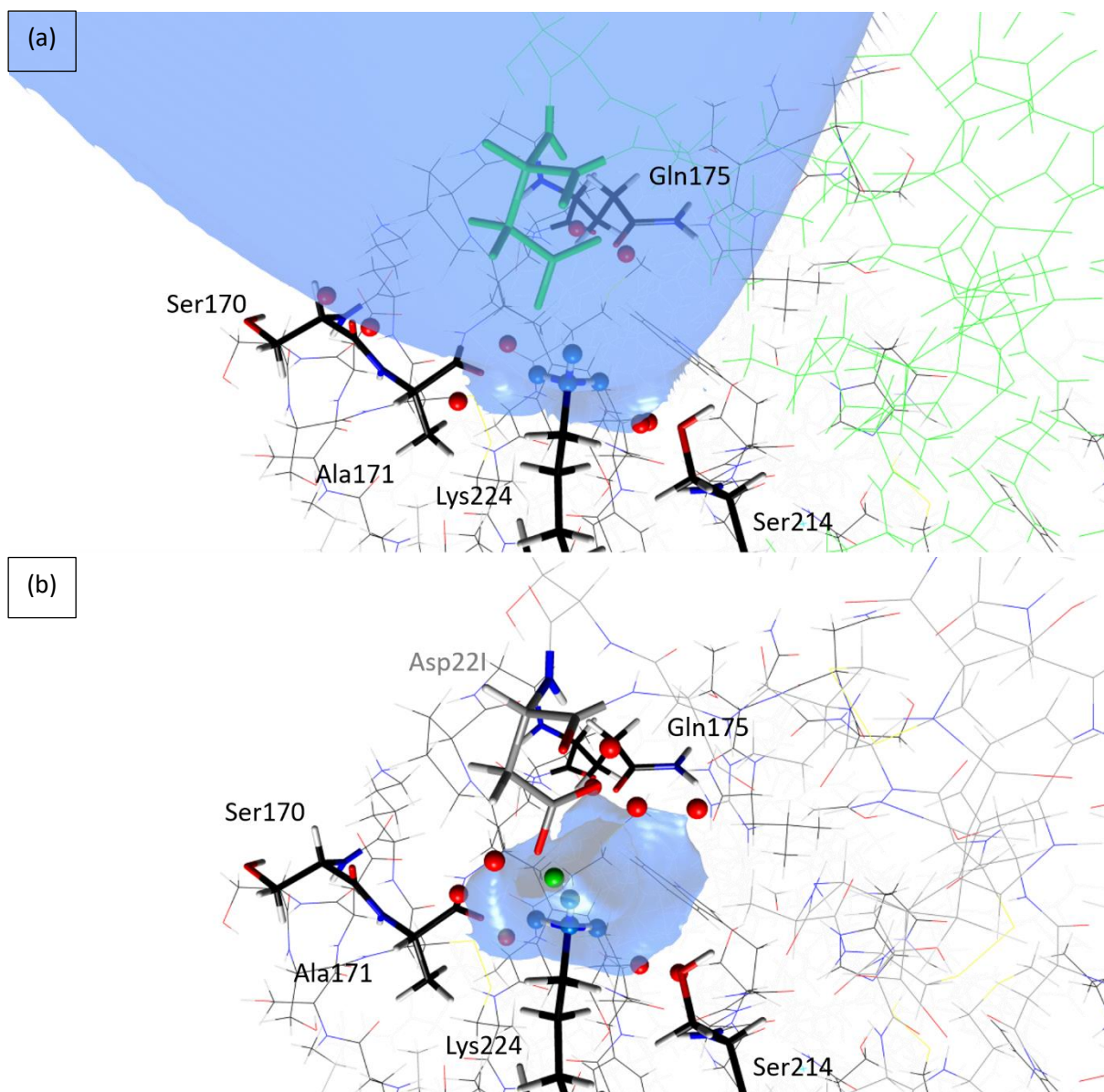


Figure 6: Electrophilic Influence Zone of Lys224 NH3 group

The electrophilic influence zone (EIZ) of the $N\zeta$, $H\zeta1$, $H\zeta2$ and $H\zeta3$ atoms of the Lys224 residue, represented by blue surfaces, was computed on the basis of (a) the whole apo trypsin charge distribution, excluding the inhibitor one, and (b) the whole trypsin and inhibitor charge distributions. Their union represents the volume occupied by the electric field lines emerging from the electrophilic sites (blue spheres) localized on the Lys224 amine group nuclei. Some of these field lines converge onto the nucleophilic sites (red spheres) corresponding to the electron lone pairs of the following oxygen atoms: Ser170-O, Ala171-O, Gln175-O ϵ 1 and Ser214-O γ from the protein Asp22-O δ 1 and O δ 2 from the inhibitor. Since the protein net charge is positive, other field lines are allowed to go outside of the protein surface in (a). The green sphere in (b) represent (3, -1) critical point of the electrostatic potential characterizing the hydrogen bond between enzyme Lys224-H ζ 3 and inhibitor Asp22-O δ 2.

Discussion of catalytic mechanisms using the influence zone descriptors

Stabilization of the oxyanion hole

Serine proteases contain in their active site an oxyanion hole, which is most often formed by two hydrogen bond donors from the main chain, namely those of the amide groups of Gly193 and Ser195 [49]. These hydrogen atoms induce a positively charged pocket that first interacts with the carbonyl group of the scissile peptide bond and then stabilizes the negatively charged oxyanion of both tetrahedral intermediates [51]. We have computed the electrophilic influence zones (EIZs) associated to the amide H atoms of Gly193 and Ser195, from the electrostatic potential generated by the apo protein (see Figure 7).

The position of the oxyanion hole (displayed as a green sphere in Figure 7) is exclusively included within the Gly193-H EIZ. The Ser195-H EIZ extends toward the binding pocket of the positive charge of the substrate and converges on nucleophilic sites of Ser214 and Trp215. These observations agree with the analysis of the trypsin/inhibitor complex. The total electrostatic interaction energy between the inhibitor and Gly193 (-6.5 kcal/mol) is more favourable than that between the inhibitor and Ser195 (-4.6 kcal/mol). This may seem surprising as Gly193 forms only one hydrogen bond with the inhibitor, while Ser195 forms two. On the other hand, Gly193 forms the shortest hydrogen bond distance (2.7 Å while 2.9 Å and 3.0 Å for Ser195).

Furthermore, the Gly193-H EIZ occupies a large volume that extends well outside the protein surface. While some electric field lines emerging from Gly193-H atom converge toward nucleophilic sites (red spheres) corresponding to the electron lone pairs of oxygen atoms in His40-O, Ser96-O, Asn97-O δ 1, and Ser195-O γ , others extend far outside the protein structure, due to its net positive total charge. This “infinite” EIZ suggests that the influence of the Gly193-H atom can extend beyond the protein toward the solvent region. Therefore, this proton may influence the electrostatic steering of the scissile peptide bond carbonyl group to bring it to the cleavage site.

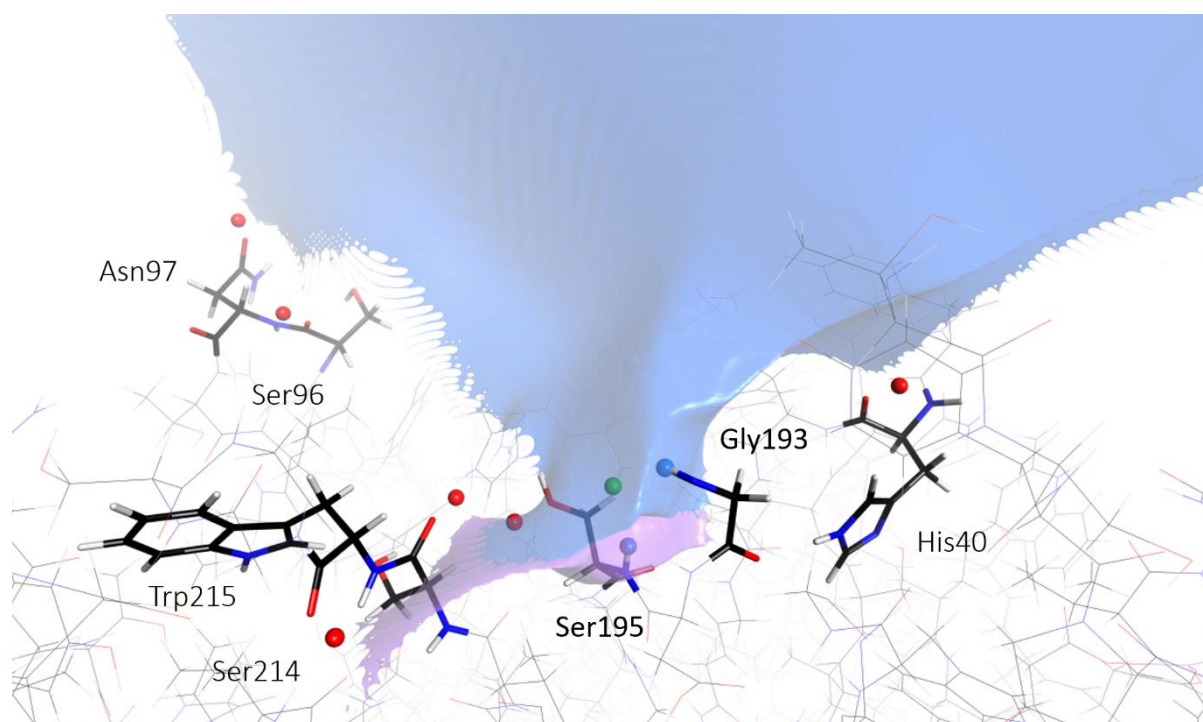


Figure 7: Electrophilic influence zone associated to the amide H atom of the trypsin Gly193 and Ser195 residues.

The electrophilic influence zone (EIZ) of the Gly193-H and Ser195-H atoms, represented by blue and purple surfaces, respectively, were computed on the basis of the whole apo trypsin charge distribution, excluding the contribution from the inhibitor. They represent volumes occupied by electric field lines emerging from the electrophilic sites (blue sphere) localized on the Gly193-H and Ser195-H nuclei. Some of these field lines converge onto the nucleophilic sites (red spheres) corresponding to the electron lone pairs of His40-O, Ser196-O, Asn97-O δ 1, Ser195-O for Gly193-H, and Ser214-O γ , Trp215-O atoms for Ser195-H EIZ, respectively. Since the protein net charge is positive, other field lines are allowed to go outside of the protein surface, to the infinity. The green sphere highlights the position of the oxyanion which is known to be stabilized by these two atoms.

Protonation state of the catalytic aspartate residue

The mechanism of the catalytic serine activation was first explained by Blow *et al.*, [55] with the "charge relay" model. In this model, the negatively charged side chain of the Asp102 residue forms a hydrogen bond with the proton of the His57-N δ 1 atom, making the other His57 nitrogen atom (N ϵ 2) more electronegative. This His57-N ϵ 2 atom is then able to accept the hydroxyl proton of the Ser195 residue, while the His57-N δ 1 proton is transferred to one of the oxygen atoms of Asp102 carboxyl group. The proton transfer from Ser195 to His57 has been supported by experimental evidence [56,57], but not the proton transfer from His57 to Asp102. NMR experiments have supported a mechanism involving a single proton transfer where the His57-N ϵ 2 atom receive the hydrogen atom from the hydroxyl group of S195. The positively charged His57 residue is then stabilized by an electrostatic interaction with the negatively charged Asp102 residue [58]. A more advanced two-proton mechanism has supposed a "low barrier hydrogen bond", in which the short (< 2.6 Å) and strong (> 30 kcal/mol) hydrogen bonds are formed [51]. These unusual properties of hydrogen bonds arise from the close distance between donor and acceptor, with equivalent pKa values, lowering the energy barrier for proton transfer, so that the proton can be shared between the two attractors in a single potential well. In conclusion, no unique explanation of proton transfer mechanism following the Ser195 activation is commonly accepted.

When the Ser195 is in activated form (*i.e.*, deprotonated), the His57 is positively charged and the Asp102 is negatively charged, which is supported by the X-ray and neutron joint refinement of the trypsin – BPTI (bovine pancreas trypsin inhibitor) structure [59]. Before the Ser195 activation, the neutral form of the His57 residue (proton on N δ 1 atom) in protein-inhibitor complexes is well accepted [60] whereas the protonation state of the Asp102 residue is still debated. Joint X-ray and neutron refinement of the apo trypsin structure [61] shows the Asp102 residue in its carboxylate form. Similarly, high-resolution structure refinements of trypsin in complex with small molecule inhibitors also show the negatively charged Asp102 residue [62,63]. Additionally, the pKa of Asp102 was estimated at 1.5 in a serine protease, suggesting the COO⁻ form of Asp102 in a wide range of pH [64]. However, Wahlgren *et al.* refined the structure of trypsin in complex with a canonical peptide inhibitor from high-resolution X-ray diffraction data and modelled a catalytic aspartic acid, with a proton bound to the Asp102-O δ 2 atom [41]. They based their model on the asymmetric character of the experimental electron density around the Asp102 side chain, which was found similar to theoretically computed density. They proposed that this neutral form of Asp102 residue maintains a sufficiently low pKa of His57 residue to remain in neutral form. Once histidine receives the proton after serine activation, aspartate would then become negatively charged, stabilizing the protonated histidine and the tetrahedral intermediate.

451 Here, we have computed the nucleophilic influence zones (NIZ) associated to the Asp102-O δ 1 and
452 Asp102-O δ 2 atoms in the case of a protonated O δ 2 atom (carboxyl state of Asp102), as in the model
453 of Wahlgren *et al*, and in carboxylate state of Asp102. In the carboxylate case (Figure 8.a), the NIZs of
454 Asp102 side chain oxygen atoms cover numerous His57 electrophilic sites, including the proton of the
455 His57-N δ 1 atom. In this model where Asp102 is charged, Asp102 electrostatic influence on His57
456 appears significant. In the model where Asp102 is protonated, the picture is totally different (see
457 Figure 8.b). The volume of both Asp102 oxygen atoms NIZ are dramatically reduced: from 69 Å³ to 2
458 Å³ for Asp102-O δ 1 atom and from 56 Å³ to 0.8 Å³ for Asp102-O δ 2 atom. These atoms are no longer
459 the main contributors interacting with the His57-H δ 1 proton. Now, the electric field lines are shared
460 on many electronegative nucleophilic sites including Ser214-O γ , Val213-O, and Tyr94-O η atoms. The
461 proton of His57-N δ 1 appears mainly under the electrostatic influence of these oxygen atoms while the
462 Asp102 contribution seems shielded by the influence of these nucleophilic sites. Indeed, electric field
463 magnitude experienced by His57-H δ 1 (excluding the contribution of His57 atoms) drops from 0.15 to
464 0.07 eÅ⁻² when Asp102 is protonated on O δ 2.

465 However, in previously proposed mechanisms, His57-H δ 1 is supposed to be either shared with or
466 stabilized by Asp102 side chain. For non-protonated Asp102, the His57-H δ 1 atom is uniquely
467 influenced by the carboxylate group of Asp102, which is consistent with the fact that trypsin D102N
468 mutant shows an essentially identical atomic structure to the wild type trypsin but with an activity
469 decreased by four orders of magnitude at pH = 7 [65], meaning that the role of the Asp102 residue
470 seems determinant for the enzymatic mechanism.

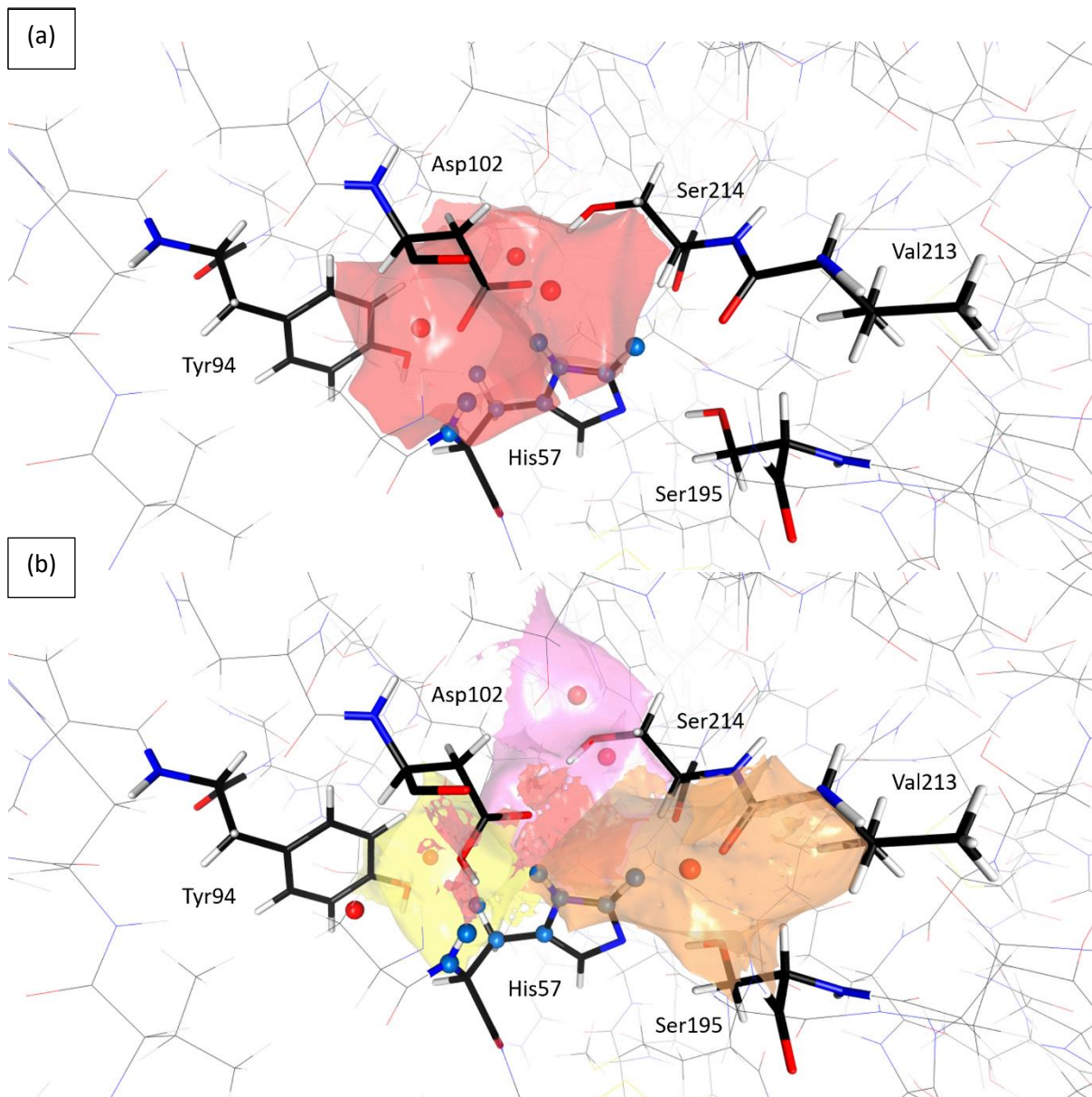


Figure 7: Nucleophilic influence zones interacting with the electrophilic sites on the His57 residue.

The nucleophilic influence zones (NIZ) which interact with electrophilic sites (blue spheres) on His57 residues has been computed from the electrostatic potential generated by the charge distributions of the whole trypsin-inhibitor complex. (a) In the case of non-protonated Asp102 carboxyl group, the NIZ associated to the nucleophilic sites (red spheres) corresponding to the electron lone pairs of the Asp102-O δ 1 and Asp102-O δ 2 atoms are represented by red surfaces. The electrostatic influences of this oxygen atoms on the electrophilic sites of the His57 residue (blue spheres) are dominant. (b) In the case of protonated Asp102-O δ 2 atom, the NIZs associated to the nucleophilic sites (red spheres) corresponding to the electron lone pairs of the Asp102-O δ 1 and Asp102-O δ 2 are represented by the red surfaces, the one of Tyr94-O η by the yellow surface, of the Ser214-O γ by the pink surface, and of the Val213-O by the orange surface. The influence of the Asp102 residue on the His57 electrophilic sites (blue spheres) is considerably reduced and replaced by the nucleophilic influences of the other oxygen atoms.

Conclusion and final remarks.

In this article we shown how the topology of the molecular electrostatic potential allows to retrieve $V(\mathbf{r})$ critical points and topological basins. These basins are the union of electric field lines associated with either a pair of nucleophilic/electrophilic sites or a single site in the case of a non-neutral total charge distribution. Based on this, we propose utilizing the surfaces of Nucleophilic or Electrophilic Influence Zones as new graphical descriptors of protein electrostatic properties. This approach offers a partition of the molecular space, such as a protein binding pocket, into influence zones that encompass all electric field lines converging to, or emanating from, a single actor. Using the MoProViewer software, which implements this method, the actor can be a single critical point, an atom, or even a group of atoms such as an amino acid side chain.

This approach is illustrated here with a test-case example based on the structure of a bovine trypsin in complex with a peptide inhibitor, in which the catalytic Asp102 residue was modelled as neutral [41]. Examining this system from the perspective of electrostatic influence zones resulted in several findings.

At first, the protein cleft in which Arg291 is buried to form a salt bridge with Asp189 is totally under the electrostatic influence of one lone electron pair of Asp189-O δ 2 (i.e. the corresponding (3, +3) $V(\mathbf{r})$ critical point). This NIZ extends to the surface of the protein, suggesting a possible role of Asp189 in the electrostatic steering of the electrophilic Arg291 residue upon ligand binding, despite Asp189 being buried in the protein structure (Fig. 5). This spatial extent is made possible by the presence of numerous electrophilic sites distributed in Arg291 binding site, notably by main chain hydrogen atoms of Gly216 and Ser217. Interestingly, it was recently shown that mutating Asp189 into a neutral residue leads to destabilization of the 215-217 segment [54], which could therefore be precisely explained by the loss of the electrostatic influence exerted by Asp189-O δ 2.

When it comes to molecular recognition, the EIZ of Lys224 NH 3^+ group in the apo structure plays a significant role. It extends largely into the solvent region (Fig. 6) and could have been utilized to rationally predict the necessity of a strong nucleophilic group at the appropriate position in the inhibitor sequence. Indeed, T22D is one of the mutations that greatly improved the binding affinity of this inhibitor, which is also supported by the strongly favorable electrostatic interaction energy between Asp221 and Lys224 (Fig. 4).

Furthermore, interpreting electrostatic influence zones provides insights into the catalytic mechanism. At first, it was shown that Gly193 and Ser195 amide H atoms, presumed to contribute to the stabilization of the oxyanion hole, present significantly different EIZ, both in term of volumes and anisotropy (Fig. 7). The oxyanion hole position is under the influence of Gly193 main-chain hydrogen atom which present an EIZ extending up to the catalytic site entrance, unlike the one of Ser195.

Finally, by comparing the partition of the molecular space into NIZ when Asp102 is charged or neutral, we concluded that the influence of neutral Asp102 on the catalytic His57 residue is significantly shielded by those of other nearby nucleophilic sites. The protonation state of Asp102 has a significant impact on both the magnitude and the direction of electric fields applied to His57, consequently affecting the electrostatic forces and the polarization effects experienced by this catalytic residue.

In this study we used a molecular electrostatic potential computed from an accurate multipolar charge distribution, transferred on a trypsin-inhibitor complex structure solved at atomic resolution. However, our method can be applied in MoProViewer using $V(\mathbf{r})$ derived from any source, as soon as

it can be provided as a fine-sampled three-dimensional scalar field, for instance in the Gaussian CUBE format.

Overall, in comparison to the traditional graphical representation of $V(\mathbf{r})$ mapped on a molecular surface, the approach proposed in this work presents several advantages:

- Intuitive interpretation: it allows for the intuitive interpretation of the molecular electrostatic potential as a three-dimensional property, and at atomic resolution (up to the position of electron lone pairs if an adequate model of molecular electron density is used to compute $V(\mathbf{r})$).
- Partitioning of molecular space: it allows a partition the molecular space into NIZ and EIZ, providing a mean to visually apprehend not only the positions of nucleophilic and electrophilic sites, but also the spatial extend of their respective influence in terms of electrostatic forces. Moreover, besides values of $V(\mathbf{r})$ and of $E(\mathbf{r})$ magnitudes, the space partition into NIZ or EIZ gives access to an estimation of their corresponding volumes, which are in turn related to their share in the total self-electrostatic energy of the corresponding charge distribution
- Communication between charged sites: by highlighting the communication between nucleophilic and electrophilic sites through electric field lines, this approach allows to identify the role of charged groups, even if they are spatially distant, in shaping specific electrostatic environments related to protein properties. This would not be possible, or at least much less intuitively, by examining only an atomic model or a surface-mapped representation of $V(\mathbf{r})$.
- Comparative study: it facilitates the comparative study of protonation states or of point mutations, offering a novel and easy way to visually interpret their consequences in terms of electrostatic environments.
- Electrostatic steering: locating NIZ or EIZ which are significantly extending toward the solvent region can highlights possible roles of related nucleophilic or electrophilic residues in the electrostatic driving of an approaching substrate.

This last point brings us to a noteworthy observation. In the aforementioned study of the catalytic mechanism of thrombin, Pozzi *et al.* presented a figure (figure 1 [54]) of the protein's electrostatic potential drawn using iso-surfaces of negative and positive $V(\mathbf{r})$ values. This representation, chosen instead of a surface-mapped one, was presumably aimed to highlight the extent of $V(\mathbf{r})$ in the extra molecular space surrounding the binding pocket. The figure was accompanied by the caption: "*Nine Asp and Glu-residues (magenta sticks) surround the active site and generate a negative electrostatic gradient that steers positively charged substrates into the active site pocket*". Considering their reference to $E(\mathbf{r})$, the approach we propose could have allowed these authors to delve deeper into their interpretation. Indeed, a partition of this region into NIZ would have precisely indicated the respective contribution of each of these nucleophilic residues, highlighting the actual contributors to the attraction exerted on a positively charged substrate as it approaches the binding site.

References

- [1] J.S. Richardson, D.C. Richardson, D.S. Goodsell, Seeing the PDB, J Biol Chem. 296 (2021) 100742. <https://doi.org/10.1016/j.jbc.2021.100742>.
- [2] X. Martinez, M. Chavent, M. Baaden, Visualizing protein structures - tools and trends, Biochem Soc Trans. 48 (2020) 499–506. <https://doi.org/10.1042/BST20190621>.

- [3] W. Tian, C. Chen, X. Lei, J. Zhao, J. Liang, CASTp 3.0: computed atlas of surface topography of proteins, *Nucleic Acids Res.* 46 (2018) W363–W367. <https://doi.org/10.1093/nar/gky473>.
- [4] F. Vascon, M. Gasparotto, M. Giacomello, L. Cendron, E. Bergantino, F. Filippini, I. Righetto, Protein electrostatics: From computational and structural analysis to discovery of functional fingerprints and biotechnological design, *Computational and Structural Biotechnology Journal*. 18 (2020) 1774–1789. <https://doi.org/10.1016/j.csbj.2020.06.029>.
- [5] A. Hospital, J.R. Goñi, M. Orozco, J.L. Gelpí, Molecular dynamics simulations: advances and applications, *Advances and Applications in Bioinformatics and Chemistry*. 8 (2015) 37–47. <https://doi.org/10.2147/AABC.S70333>.
- [6] E. Jurrus, D. Engel, K. Star, K. Monson, J. Brandi, L.E. Felberg, D.H. Brookes, L. Wilson, J. Chen, K. Liles, M. Chun, P. Li, D.W. Gohara, T. Dolinsky, R. Konecny, D.R. Koes, J.E. Nielsen, T. Head-Gordon, W. Geng, R. Krasny, G.-W. Wei, M.J. Holst, J.A. McCammon, N.A. Baker, Improvements to the APBS biomolecular solvation software suite, *Protein Sci.* 27 (2018) 112–128. <https://doi.org/10.1002/pro.3280>.
- [7] S. Domagała, B. Fournier, D. Liebschner, B. Guillot, C. Jelsch, An improved experimental databank of transferable multipolar atom models – ELMAM2. Construction details and applications, *Acta Cryst A*. 68 (2012) 337–351. <https://doi.org/10.1107/S0108767312008197>.
- [8] P.M. Rybicka, M. Kulik, M.L. Chodkiewicz, P.M. Dominiak, Multipolar Atom Types from Theory and Statistical Clustering (MATTS) Data Bank: Impact of Surrounding Atoms on Electron Density from Cluster Analysis, *J. Chem. Inf. Model.* 62 (2022) 3766–3783. <https://doi.org/10.1021/acs.jcim.2c00145>.
- [9] B. Meyer, B. Guillot, M.F. Ruiz-Lopez, A. Genoni, Libraries of Extremely Localized Molecular Orbitals. 1. Model Molecules Approximation and Molecular Orbitals Transferability, *J. Chem. Theory Comput.* 12 (2016) 1052–1067. <https://doi.org/10.1021/acs.jctc.5b01007>.
- [10] B. Meyer, B. Guillot, M.F. Ruiz-Lopez, C. Jelsch, A. Genoni, Libraries of Extremely Localized Molecular Orbitals. 2. Comparison with the Pseudoatoms Transferability, *J. Chem. Theory Comput.* 12 (2016) 1068–1081. <https://doi.org/10.1021/acs.jctc.5b01008>.
- [11] L. Schrödinger, W.L. DeLano, Pymol, (2020). <http://www.pymol.org/pymol>.
- [12] W. Humphrey, A. Dalke, K. Schulten, VMD: Visual molecular dynamics, *Journal of Molecular Graphics*. 14 (1996) 33–38. [https://doi.org/10.1016/0263-7855\(96\)00018-5](https://doi.org/10.1016/0263-7855(96)00018-5).
- [13] E.F. Pettersen, T.D. Goddard, C.C. Huang, G.S. Couch, D.M. Greenblatt, E.C. Meng, T.E. Ferrin, UCSF Chimera--a visualization system for exploratory research and analysis, *J Comput Chem*. 25 (2004) 1605–1612. <https://doi.org/10.1002/jcc.20084>.
- [14] V. Vuković, T. Leduc, Z. Jelić-Matošević, C. Didierjean, F. Favier, B. Guillot, C. Jelsch, A rush to explore protein–ligand electrostatic interaction energy with Charger, *Acta Cryst D*. 77 (2021) 1292–1304. <https://doi.org/10.1107/S2059798321008433>.
- [15] B. Guillot, E. Enrique, L. Huder, C. Jelsch, IUCr, MoProViewer: a tool to study proteins from a charge density science perspective, *Acta Crystallographica Section A: Foundations and Advances*. 70 (2014) C279–C279. <https://doi.org/10.1107/S2053273314097204>.
- [16] N.A. Baker, D. Sept, S. Joseph, M.J. Holst, J.A. McCammon, Electrostatics of nanosystems: Application to microtubules and the ribosome, *Proceedings of the National Academy of Sciences*. 98 (2001) 10037–10041. <https://doi.org/10.1073/pnas.181342398>.
- [17] S.R. Gadre, S.A. Kulkarni, I.H. Shrivastava, Molecular electrostatic potentials: A topographical study, *The Journal of Chemical Physics*. 96 (1992) 5253–5260. <https://doi.org/10.1063/1.462710>.
- [18] V. Pascucci, Topology Diagrams of Scalar Fields in Scientific Visualisation, in: *Topological Data Structures for Surfaces*, John Wiley & Sons, Ltd, 2004: pp. 121–129. <https://doi.org/10.1002/0470020288.ch8>.
- [19] C. Heine, H. Leitte, M. Hlawitschka, F. Iuricich, L. De Floriani, G. Scheuermann, H. Hagen, C. Garth, A Survey of Topology-based Methods in Visualization, *Computer Graphics Forum*. 35 (2016) 643–667. <https://doi.org/10.1111/cgf.12933>.
- [20] S. Grabowsky, A. Genoni, H.-B. Bürgi, Quantum crystallography, *Chem. Sci.* 8 (2017) 4159–4176. <https://doi.org/10.1039/C6SC05504D>.

- [21] A. Genoni, L. Bučinský, N. Claiser, J. Contreras-García, B. Dittrich, P.M. Dominiak, E. Espinosa, C. Gatti, P. Giannozzi, J.-M. Gillet, D. Jayatilaka, P. Macchi, A.Ø. Madsen, L. Massa, C.F. Matta, K.M. Merz Jr., P.N.H. Nakashima, H. Ott, U. Ryde, K. Schwarz, M. Sierka, S. Grabowsky, *Quantum Crystallography: Current Developments and Future Perspectives*, *Chemistry – A European Journal*. 24 (2018) 10881–10905. <https://doi.org/10.1002/chem.201705952>.
- [22] R.F.W. Bader, *Atoms in Molecules: A Quantum Theory*, Oxford University Press, Oxford, New York, 1994.
- [23] M.M. Mesmoudi, L. De Floriani, P. Magillo, Morphology analysis of 3D scalar fields based on morse theory and discrete distortion, in: *Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, Association for Computing Machinery, New York, NY, USA, 2009: pp. 187–196. <https://doi.org/10.1145/1653771.1653799>.
- [24] L. Yan, T.B. Masood, R. Sridharamurthy, F. Rasheed, V. Natarajan, I. Hotz, B. Wang, Scalar Field Comparison with Topological Descriptors: Properties and Applications for Scientific Visualization, *Computer Graphics Forum*. 40 (2021) 599–633. <https://doi.org/10.1111/cgf.14331>.
- [25] E. Espinosa, E. Molins, C. Lecomte, Hydrogen bond strengths revealed by topological analyses of experimentally observed electron densities, *Chemical Physics Letters*. 285 (1998) 170–173. [https://doi.org/10.1016/S0009-2614\(98\)00036-0](https://doi.org/10.1016/S0009-2614(98)00036-0).
- [26] J. Contreras-García, W. Yang, E.R. Johnson, Analysis of Hydrogen-Bond Interaction Potentials from the Electron Density: Integration of Noncovalent Interaction Regions, *J. Phys. Chem. A*. 115 (2011) 12983–12990. <https://doi.org/10.1021/jp204278k>.
- [27] R. Laplaza, F. Peccati, R. A. Boto, C. Quan, A. Carbone, J.-P. Piquemal, Y. Maday, J. Contreras-García, NCIPLOT and the analysis of noncovalent interactions using the reduced density gradient, *WIREs Computational Molecular Science*. 11 (2021) e1497. <https://doi.org/10.1002/wcms.1497>.
- [28] M. Leboeuf, A.M. Köster, K. Jug, D.R. Salahub, Topological analysis of the molecular electrostatic potential, *J. Chem. Phys.* 111 (1999) 4893–4905. <https://doi.org/10.1063/1.479749>.
- [29] I. Mata, E. Molins, E. Espinosa, Zero-Flux Surfaces of the Electrostatic Potential: The Border of Influence Zones of Nucleophilic and Electrophilic Sites in Crystalline Environment, *J. Phys. Chem. A*. 111 (2007) 9859–9870. <https://doi.org/10.1021/jp074032l>.
- [30] N. Mohan, C.H. Suresh, A. Kumar, S.R. Gadre, Molecular electrostatics for probing lone pair– π interactions, *Phys. Chem. Chem. Phys.* 15 (2013) 18401–18409. <https://doi.org/10.1039/C3CP53379D>.
- [31] I. Alkorta, I. Mata, E. Molins, E. Espinosa, Energetic, Topological and Electric Field Analyses of Cation-Cation Nucleic Acid Interactions in Watson-Crick Disposition, *ChemPhysChem*. 20 (2019) 148–158. <https://doi.org/10.1002/cphc.201800878>.
- [32] S.R. Gadre, I.H. Shrivastava, Shapes and sizes of molecular anions via topographical analysis of electrostatic potential, *J. Chem. Phys.* 94 (1991) 4384–4390. <https://doi.org/10.1063/1.460625>.
- [33] P.K. Anjalikrishna, C.H. Suresh, S.R. Gadre, Electrostatic Topographical Viewpoint of π -Conjugation and Aromaticity of Hydrocarbons, *J Phys Chem A*. 123 (2019) 10139–10151. <https://doi.org/10.1021/acs.jpca.9b09056>.
- [34] Z. Ji, J. Kozuch, I.I. Mathews, C.S. Diercks, Y. Shamsudin, M.A. Schulz, S.G. Boxer, Protein Electric Fields Enable Faster and Longer-Lasting Covalent Inhibition of β -Lactamases, *J. Am. Chem. Soc.* 144 (2022) 20947–20954. <https://doi.org/10.1021/jacs.2c09876>.
- [35] R. Schweitzer-Stenner, Internal Electric Field in Cytochrome C Explored by Visible Electronic Circular Dichroism Spectroscopy., *J. Phys. Chem. B*. 112 (2008) 10358–10366. <https://doi.org/10.1021/jp802495q>.
- [36] T. Kimmett, N. Smith, S. Witham, M. Petukh, S. Sarkar, E. Alexov, ProBLM Web Server: Protein and Membrane Placement and Orientation Package, *Computational and Mathematical Methods in Medicine*. 2014 (2014) e838259. <https://doi.org/10.1155/2014/838259>.
- [37] L. Breindel, J. Yu, D.S. Burz, A. Shekhtman, Intact ribosomes drive the formation of protein quinary structure, *PLOS ONE*. 15 (2020) e0232015. <https://doi.org/10.1371/journal.pone.0232015>.

- [38] W. Guo, Y. Xie, A.E. Lopez-Hernandez, S. Sun, L. Li, W. Guo, Y. Xie, A.E. Lopez-Hernandez, S. Sun, L. Li, Electrostatic features for nucleocapsid proteins of SARS-CoV and SARS-CoV-2, *MBE*. 18 (2021) 2372–2383. <https://doi.org/10.3934/mbe.2021120>.
- [39] B. Guillot, L. Viry, R. Guillot, C. Lecomte, C. Jelsch, Refinement of proteins at subatomic resolution with MOPRO, *Journal of Applied Crystallography*. 34 (2001) 214–223. <https://doi.org/10.1107/S0021889801001753>.
- [40] C. Jelsch, B. Guillot, A. Lagoutte, C. Lecomte, Advances in protein and small-molecule charge-density refinement methods using MoPro, *J Appl Cryst*. 38 (2005) 38–54. <https://doi.org/10.1107/S0021889804025518>.
- [41] W.Y. Wahlgren, G. Pál, J. Kardos, P. Porrogi, B. Szenthe, A. Patthy, L. Gráf, G. Katona, The Catalytic Aspartate Is Protonated in the Michaelis Complex Formed between Trypsin and an in Vitro Evolved Substrate-like Inhibitor, *J Biol Chem*. 286 (2011) 3587–3596. <https://doi.org/10.1074/jbc.M110.161604>.
- [42] R.K. Pathak, S.R. Gadre, Maximal and minimal characteristics of molecular electrostatic potentials, *J. Chem. Phys.* 93 (1990) 1770–1773. <https://doi.org/10.1063/1.459703>.
- [43] A. Kumar, S.R. Gadre, N. Mohan, C.H. Suresh, Lone Pairs: An Electrostatic Viewpoint, *J. Phys. Chem. A*. 118 (2014) 526–532. <https://doi.org/10.1021/jp4117003>.
- [44] A. Kumar, S.R. Gadre, Exploring the Gradient Paths and Zero Flux Surfaces of Molecular Electrostatic Potential, *J. Chem. Theory Comput*. 12 (2016) 1705–1713. <https://doi.org/10.1021/acs.jctc.6b00073>.
- [45] S.R. Gadre, R.D. Bendale, On the similarity between molecular electron densities, electrostatic potentials and bare nuclear potentials, *Chemical Physics Letters*. 130 (1986) 515–521. [https://doi.org/10.1016/0009-2614\(86\)80249-4](https://doi.org/10.1016/0009-2614(86)80249-4).
- [46] P. Balanarayan, S.R. Gadre, Topography of molecular scalar fields. I. Algorithm and Poincaré–Hopf relation, *J. Chem. Phys.* 119 (2003) 5037–5043. <https://doi.org/10.1063/1.1597652>.
- [47] S. Domagała, P. Munshi, M. Ahmed, B. Guillot, C. Jelsch, Structural analysis and multipole modelling of quercetin monohydrate—a quantitative and comparative study, *Acta Crystallogr B*. 67 (2011) 63–78. <https://doi.org/10.1107/S0108768110041996>.
- [48] E.S. Radisky, J.M. Lee, C.-J.K. Lu, D.E. Koshland, Insights into the serine protease mechanism from atomic resolution structures of trypsin reaction intermediates, *Proc Natl Acad Sci U S A*. 103 (2006) 6835–6840. <https://doi.org/10.1073/pnas.0601910103>.
- [49] R. Ménard, A.C. Storer, Oxyanion hole interactions in serine and cysteine proteases, *Biol Chem Hoppe Seyler*. 373 (1992) 393–400. <https://doi.org/10.1515/bchm3.1992.373.2.393>.
- [50] T.A. Steitz, R. Hendekson, D.M. Blow, Structure of crystalline α -chymotrypsin: III. Crystallographic studies of substrates and inhibitors bound to the active site of α -chymotrypsin, *Journal of Molecular Biology*. 46 (1969) 337–348. [https://doi.org/10.1016/0022-2836\(69\)90426-4](https://doi.org/10.1016/0022-2836(69)90426-4).
- [51] L. Hedstrom, Serine Protease Mechanism and Specificity, *Chem. Rev*. 102 (2002) 4501–4524. <https://doi.org/10.1021/cr000033x>.
- [52] S. Prasad, A.M. Cantwell, L.A. Bush, P. Shih, H. Xu, E. Di Cera, Residue Asp-189 controls both substrate binding and the monovalent cation specificity of thrombin, *J Biol Chem*. 279 (2004) 10103–10108. <https://doi.org/10.1074/jbc.M312614200>.
- [53] L.B. Evnin, J.R. Vásquez, C.S. Craik, Substrate specificity of trypsin investigated by using a genetic selection., *Proceedings of the National Academy of Sciences*. 87 (1990) 6659–6663. <https://doi.org/10.1073/pnas.87.17.6659>.
- [54] N. Pozzi, M. Zerbetto, L. Acquasaliente, S. Tescari, D. Frezzato, A. Polimeno, D.W. Gohara, E. Di Cera, V. De Filippis, Loop Electrostatics Asymmetry Modulates the Preexisting Conformational Equilibrium in Thrombin, *Biochemistry*. 55 (2016) 3984–3994. <https://doi.org/10.1021/acs.biochem.6b00385>.
- [55] D.M. Blow, J.J. Birktoft, B.S. Hartley, Role of a buried acid group in the mechanism of action of chymotrypsin, *Nature*. 221 (1969) 337–340. <https://doi.org/10.1038/221337a0>.

- [56] J.L. Markley, I.B. Ibañez, Zymogen activation in serine proteinases. Proton magnetic resonance pH titration studies of the two histidines of bovine chymotrypsinogen A and chymotrypsin A α , *Biochemistry*. 17 (1978) 4627–4640. <https://doi.org/10.1021/bi00615a008>.
- [57] W.W. Bachovchin, J.D. Roberts, Nitrogen-15 nuclear magnetic resonance spectroscopy. The state of histidine in the catalytic triad of α -lytic protease. Implications for the charge-relay mechanism of peptide-bond cleavage by serine proteases, *J. Am. Chem. Soc.* 100 (1978) 8041–8047. <https://doi.org/10.1021/ja00494a001>.
- [58] G. Robillard, R.G. Shulman, High resolution nuclear magnetic resonance study of the histidine—Aspartate hydrogen bond in chymotrypsin and chymotrypsinogen, *Journal of Molecular Biology*. 71 (1972) 507–511. [https://doi.org/10.1016/0022-2836\(72\)90366-X](https://doi.org/10.1016/0022-2836(72)90366-X).
- [59] K. Kawamura, T. Yamada, K. Kurihara, T. Tamada, R. Kuroki, I. Tanaka, H. Takahashi, N. Niimura, X-ray and neutron protein crystallographic analysis of the trypsin-BPTI complex, *Acta Crystallogr D Biol Crystallogr*. 67 (2011) 140–148. <https://doi.org/10.1107/S0907444910053382>.
- [60] J.L. Markley, M.A. Porubcan, The charge-relay system of serine proteinases: Proton magnetic resonance titration studies of the four histidines of porcine trypsin, *Journal of Molecular Biology*. 102 (1976) 487–509. [https://doi.org/10.1016/0022-2836\(76\)90330-2](https://doi.org/10.1016/0022-2836(76)90330-2).
- [61] J. Schiebel, R. Gaspari, T. Wulsdorf, K. Ngo, C. Sohn, T.E. Schrader, A. Cavalli, A. Ostermann, A. Heine, G. Klebe, Intriguing role of water in protein-ligand binding studied by neutron crystallography on trypsin complexes, *Nat Commun*. 9 (2018) 3559. <https://doi.org/10.1038/s41467-018-05769-2>.
- [62] A. Schmidt, C. Jelsch, P. Østergaard, W. Rypniewski, V.S. Lamzin, Trypsin Revisited: CRYSTALLOGRAPHY AT (SUB) ATOMIC RESOLUTION AND QUANTUM CHEMISTRY REVEALING DETAILS OF CATALYSIS*, *Journal of Biological Chemistry*. 278 (2003) 43357–43362. <https://doi.org/10.1074/jbc.M306944200>.
- [63] D. Liebschner, M. Dauter, A. Brzuszkiewicz, Z. Dauter, On the reproducibility of protein crystal structures: five atomic resolution structures of trypsin, *Acta Crystallogr D Biol Crystallogr*. 69 (2013) 1447–1462. <https://doi.org/10.1107/S0907444913009050>.
- [64] P. Everill, J.L. Sudmeier, W.W. Bachovchin, Direct NMR Observation and pKa Determination of the Asp102 Side Chain in a Serine Protease, *J. Am. Chem. Soc.* 134 (2012) 2348–2354. <https://doi.org/10.1021/ja210091q>.
- [65] S. Sprang, T. Standing, R.J. Fletterick, R.M. Stroud, J. Finer-Moore, N.H. Xuong, R. Hamlin, W.J. Rutter, C.S. Craik, The three-dimensional structure of Asn102 mutant of trypsin: role of Asp102 in serine protease catalysis, *Science*. 237 (1987) 905–909. <https://doi.org/10.1126/science.3112942>.

4.2 Application des zones d'influence en biologie structurale : complexe Neuropiline 1 - peptide KDKPPR

4.2.1 Contexte de l'étude

Présentation de la neuropiline 1

La neuropiline 1 (NRP1) est une glycoprotéine membranaire exprimée chez tous les vertébrés et impliquée dans de nombreux processus physiologiques [Pellet-Many *et al.*, 2008]. Elle est une cible thérapeutique importante notamment en raison de son rôle dans le développement et la progression de plusieurs types de cancer [Chaudhary *et al.*, 2014, Liu *et al.*, 2021]. Cette protéine a également été étudiée pour son implication dans le mécanisme d'infection par le coronavirus SARS-CoV-2 [Daly *et al.*, 2020]. Dans l'étude présentée ici [Goudiaby *et al.*, 2023], six acides aminés chargés situés à la surface du fragment b1 de NRP1 ont été mutés pour modifier son empilement en phase cristalline et ainsi rendre accessible la poche de fixation du VEGF ("vascular endothelial growth factor" ou facteur de croissance de l'endothélium vasculaire). La structure de ce variant du fragment NRP1-b1 a été déterminée par diffraction des rayons X en complexe avec le peptide KDKPPR, qui a été développé dans le cadre d'une thérapie photodynamique [Bechet *et al.*, 2014, Kamarulzaman *et al.*, 2015, Kamarulzaman *et al.*, 2017]. Dans cette étude [Goudiaby *et al.*, 2023], nous avons étudié les interactions entre le peptide KDKPPR et le fragment NRP1-b1 à partir de la structure cristallographique, par simulation de dynamique moléculaire et par deux outils basés sur le modèle de densité électronique multipolaire [Hansen et Coppens, 1978] : les facteurs d'enrichissement de contact développés par C. Jelsch [Jelsch *et al.*, 2014] et les zones d'influence développées dans ma thèse.

Résumé de l'analyse de la structure cristallographique

Les trois premiers résidus (KDK) du peptide KDKPPR n'étaient pas visibles à partir des données cristallographiques et n'ont donc pas été modélisés. En revanche, le fragment PPR du peptide est apparu lié dans la poche de fixation du VEGF. Le résidu arginine est le seul résidu dont la densité électronique est parfaitement définie. La densité partielle pour les deux prolines a permis d'obtenir un modèle avec confiance. L'analyse de la structure atomique a révélé que le groupement guanidinium du résidu arginine du peptide est impliqué dans une double liaison hydrogène avec le résidu Asp320 et la partie aliphatique de sa chaîne latérale interagit avec les cycles aromatiques des résidus Tyr297 et Tyr353. Les chaînes principales de l'arginine et de la dernière proline forment des liaisons hydrogène avec les chaînes latérales des résidus Tyr297, Ser346, Thr349 et Tyr353. Les simulations de dynamique moléculaire suggèrent que la fixation du peptide n'est pas modifiée par les six mutations du fragment NRP1-b1 et que le résidu arginine du peptide est le principal contributeur à son ancrage dans la poche.

4.2.2 Caractérisation de la fixation de l'inhibiteur du point de vue des zones d'influence nucléophile

Puisque le fragment PPR du peptide est globalement électrophile en raison de la charge positive de l'arginine et des nombreux sites électrophiles sur les protons des prolines, une étude

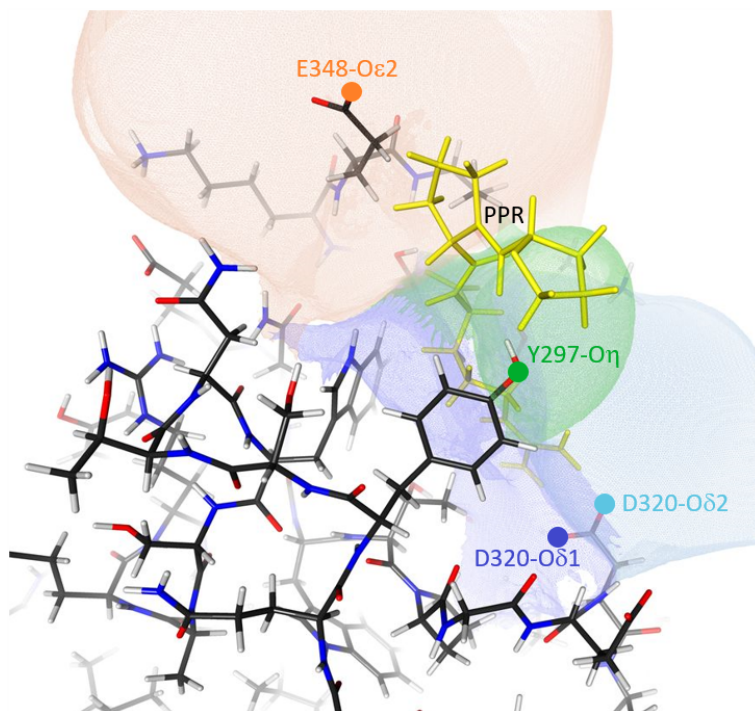


FIGURE 4.1 – Zones d'influence nucléophile intersectant le site de fixation de NRP1.

Les zones d'influence nucléophile (ZIN) associées aux atomes Tyr297-O η (surface verte), Asp320-O δ 1 (surface bleu foncé), Asp320-O δ 2 (surface bleu clair) et Glu348-O ϵ 2 (surface orange) ont été calculées à partir du potentiel électrostatique généré par les distributions de charge de l'ensemble des atomes de la NRP1 (en excluant ceux de l'inhibiteur). Ces ZIN se situent en surface de la protéine et recouvrent les positions que les atomes de la moitié PPR du ligand (en jaune) occupent dans le complexe. Cette figure est contenue dans l'article [Goudiaby *et al.*, 2023].

des influences nucléophiles de la protéine sur le site de fixation du VEGF est pertinente. Pour cela, nous avons transféré les paramètres multipolaires ELMAM2 [Domagała *et al.*, 2012] sur la structure de NRP1-b1 et calculé la carte du potentiel électrostatique $V(\mathbf{r})$ généré par l'ensemble des atomes de cette protéine (en excluant ceux du ligand) via le module Charger [Vuković *et al.*, 2021] de MoProViewer [Guillot *et al.*, 2014]. Nous avons calculé les zones d'influence nucléophile (NIZ) dans la poche de fixation, représentées sur la figure 4.1, associées aux sites nucléophiles correspondant aux doublets non-liants des atomes d'oxygène suivant : Tyr297-O η (surface verte), Asp320-O δ 1 (surface bleu foncé), Asp320-O δ 2 (surface bleu clair) et Glu348-O ϵ 2 (surface orange). Le volume défini par une NIZ contient toutes les lignes de champ électrique $\mathbf{E}(\mathbf{r}) = -\nabla V(\mathbf{r})$ convergeant vers le site nucléophile associé. Un ligand électrophile dans ce volume est donc soumis aux forces électrostatiques attractives générées par l'ensemble de la protéine et dirigées vers le site nucléophile.

La NIZ de l'atome Asp320-O δ 1 (surface bleu foncé) recouvre la position prise par l'un des deux groupements amines NH₂ de l'arginine du peptide tandis que celle de l'atome Asp320-O δ 2 (surface bleu clair) recouvre la position prise par l'autre groupement NH₂. Ces influences nucléophiles ressenties par le peptide correspondent bien à la double liaison hydrogène déjà observée par l'analyse structurale. La NIZ de l'atome Tyr297-O η (surface verte) englobe la

majorité de la première proline du peptide. Le groupement hydroxyle du résidu Tyr297 a déjà été identifié par l'analyse de la structure car il forme une liaison hydrogène avec la chaîne principale de la proline du peptide. La NIZ révèle qu'il existe également des forces électrostatiques attractives ressenties par les sites électrophiles de la chaîne latérale de cette proline qui sont dirigées vers Tyr297-O η . La NIZ de l'atome Glu348-O ϵ 2 (surface orange) recouvre quant à elle la position d'une partie de la chaîne latérale du second résidu proline du peptide. Cette influence à longue distance du résidu Glu348 participant à la fixation du peptide n'est pas visible par une analyse structurale traditionnelle. De plus, la représentation en NIZ suggère que l'une des prolines est soumise à des forces électrostatiques attractives dirigées vers Tyr297-O η tandis que l'autre est soumise à des forces électrostatiques attractives dirigées vers Glu348-O ϵ 2, ce qui peut se traduire par une stabilisation globale du peptide dans une position d'équilibre dans la poche de fixation. Par ailleurs, puisque les influences électrostatiques participent au guidage du ligand à travers le solvant vers la poche de fixation d'une protéine [Wade *et al.*, 1998], les NIZ fournissent également des informations sur les forces électrostatiques de la protéine orientant le ligand lors de son approche. Ici, les NIZ des atomes Glu348-O ϵ 2 et Asp320-O δ 2 s'étendent au-delà de la surface de la protéine, suggérant leur potentielle contribution au pilotage électrostatique facilitant la fixation du ligand électrophile dans le site de fixation VEGF de NRP1-b1.

4.2.3 Article publié décrivant la structure de NRP1

Article

New Crystal Form of Human Neuropilin-1 b1 Fragment with Six Electrostatic Mutations Complexed with KDKPPR Peptide Ligand

Ibrahima Goudiaby^{1,2}, Thérèse E. Malliavin³ , Eva Mocchetti¹, Sandrine Mathiot¹, Samir Acherar⁴ , Céline Frochot⁵ , Muriel Barberi-Heyob⁶, Benoît Guillot¹, Frédérique Favier¹ , Claude Didierjean^{1,*}  and Christian Jelsch^{1,*} 

¹ Université de Lorraine, CNRS, CRM2, F-54000 Nancy, France; i.goudiaby1592@zig.univ.sn (I.G.); eva.mocchetti@univ-lorraine.fr (E.M.); benoit.guillot@univ-lorraine.fr (B.G.)

² Université Assane Seck de Ziguinchor, Laboratoire de Chimie et de Physique des Matériaux (LCPM), 523 Ziguinchor, Senegal

³ Université de Lorraine, CNRS, LPCT, F-54000 Nancy, France; therese.malliavin@univ-lorraine.fr

⁴ Université de Lorraine, CNRS, LCPM, F-54000 Nancy, France

⁵ Université de Lorraine, CNRS, LRGP, F-54000 Nancy, France

⁶ Université de Lorraine, CNRS, CRAN, F-54000 Nancy, France; muriel.barberi@univ-lorraine.fr

* Correspondence: claudedidierjean@univ-lorraine.fr (C.D.); christian.jelsch@univ-lorraine.fr (C.J.)

Abstract: Neuropilin 1 (NRP1), a cell-surface co-receptor of a number of growth factors and other signaling molecules, has long been the focus of attention due to its association with the development and the progression of several types of cancer. For example, the KDKPPR peptide has recently been combined with a photosensitizer and a contrast agent to bind NRP1 for the detection and treatment by photodynamic therapy of glioblastoma, an aggressive brain cancer. The main therapeutic target is a pocket of the fragment b1 of NRP1 (NRP1-b1), in which vascular endothelial growth factors (VEGFs) bind. In the crystal packing of native human NRP1-b1, the VEGF-binding site is obstructed by a crystallographic symmetry neighbor protein, which prevents the binding of ligands. Six charged amino acids located at the protein surface were mutated to allow the protein to form a new crystal packing. The structure of the mutated fragment b1 complexed with the KDKPPR peptide was determined by X-ray crystallography. The variant crystallized in a new crystal form with the VEGF-binding cleft exposed to the solvent and, as expected, filled by the C-terminal moiety of the peptide. The atomic interactions were analyzed using new approaches based on a multipolar electron density model. Among other things, these methods indicated the role played by Asp320 and Glu348 in the electrostatic steering of the ligand in its binding site. Molecular dynamics simulations were carried out to further analyze the peptide binding and motion of the wild-type and mutant proteins. The simulations revealed that specific loops interacting with the peptide exhibited mobility in both the unbound and bound forms.

Keywords: Neuropilin 1; variant; ligand; X-ray crystallography; molecular dynamics simulation; Hirshfeld interface; electrostatic influence



Citation: Goudiaby, I.; Malliavin, T.E.; Mocchetti, E.; Mathiot, S.; Acherar, S.; Frochot, C.; Barberi-Heyob, M.; Guillot, B.; Favier, F.; Didierjean, C.; et al. New Crystal Form of Human Neuropilin-1 b1 Fragment with Six Electrostatic Mutations Complexed with KDKPPR Peptide Ligand.

Molecules **2023**, *28*, 5603. <https://doi.org/10.3390/molecules28145603>

Academic Editor: Chojiro Kojima

Received: 23 June 2023

Revised: 15 July 2023

Accepted: 20 July 2023

Published: 24 July 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Neuropilins (NRP1 and NRP2) are type I single-pass transmembrane glycoproteins expressed in all vertebrates and have many physiological roles. They act as co-receptors of a range of growth factors and other signaling molecules. A recent cryo-electron microscopy structure highlights the role of NRP1 in a ternary complex with a semaphorin protein and a plexin receptor, which altogether mediates signaling in neuronal axon guidance and other processes [1]. NRP1 also forms a ternary complex with vascular endothelial growth factor A 165 (VEGF-A165) and the receptor VEGFR2 [2]. This complexation is associated

with intracellular signaling, mitogenesis, cell migration and angiogenesis [3]. Research has shown that NRP1 plays a significant role in the development and progression of various cancer types [4] and also more recently in the infectivity of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) [5]. In particular, this transmembrane receptor has been suggested as a molecular therapeutic target for glioblastoma, with overexpression mainly due to endothelial cells of angiogenic phenotype and associated pro-tumor macrophages, both of which are linked to an unfavorable prognosis [6]. A recent review outlines the various functions of NRP1 in the context of cancer treatments [7].

The membrane protein NRP1 (as NRP2) contains a large N-terminal extracellular region (~850 amino acids (AA)), a single transmembrane domain (~25 AA) and a short C-terminal cytoplasmic domain (~45 AA) [8]. The ectodomain consists of five independent domains, where the first four (a1, a2, b1 and b2 domains) are involved in ligand binding, while the role of the last one (c domain) is still under debate (oligomerization, NRP1 homodimerization, etc.) [1]. The interdomain linkers are also important in the heterocomplex formation as spacers [1]. The determination of the crystal structure of the b1 domain provided the first structural insight at the atomic-level into NRP1 [9]. The interaction of NRP1 with VEGF has been extensively studied. Briefly, the NRP1-b1 domain folds into a distorted jelly roll barrel motif that is composed of two beta-sheets [9] (Figure 1a). The strands are connected by loops of varying length. The bottom of the beta-barrel core exhibits a triangular shape that contains an intramolecular disulfide bridge. At the top of the beta-barrel core, the loops are divided in six loop regions (L1–L6) to delimit the pocket of the positively charged tail of VEGF. The C-terminal arginine residue of VEGF-A165 is buried in this pocket and is a key feature of the binding. Numerous atomic experimental structures have been determined to characterize and/or inhibit the interaction of NRP1-b1 and VEGF using peptides or arginine derivatives as well as a fusion protein [10–13].

We have developed peptides combined with a photosensitizer to target NRP1 in the context of photodynamic therapy (PDT) to detect and treat glioblastoma [14–16]. Recently, a nanoparticle was designed that combines KDKPPR motif as a targeting peptide, porphyrin as photosensitizer and gadolinium chelate as contrast agent. This nanoparticle, called AGuIX@PS@KDKPPR, enables the detection of tumor tissue by magnetic resonance imaging and treatment by PDT [6,17,18]. The affinity of the nanoparticle for human NRP1 was validated, and it was found to be ten times lower than that of the free peptide (K_D of 4.7 μM for AGuIX@PS@KDKPPR and K_D of 0.5 μM for KDKPPR) [17].

In this study, the KDKPPR peptide was synthesized, and its molecular interactions with NRP1-b1 fragment were investigated by X-ray crystallography and molecular dynamics (MD) simulations. Previously, we have attempted to co-crystallize NRP1-b1 with a carbohydrate-based peptidomimetic [19]. However, we constantly obtained tetragonal crystals that were isomorphous to the crystals of the unbound protein. These crystals are unsuitable for obtaining structures of complexes because the site that binds the C-terminal tail of VEGF is obstructed by symmetry-related protein molecules [9]. We have used site-directed mutagenesis to modify the repartition of charges on the surface of NRP1-b1 to induce changes in the crystal packing. This approach is an efficient tool for crystallizing a protein in a new form and facilitating, for example, the formation of a protein–ligand complex in the crystal through co-crystallization or soaking techniques [20]. In this study, we successfully co-crystallized the KDKPPR peptide with a hexavariant of NRP1-b1 (Glu277Lys, Glu285Lys, Asp289Lys, Glu367Lys, Lys373Glu, Lys397Glu). An original crystal form was obtained where the VEGF-binding pocket is filled by the KDKPPR peptide and is located in large spaces connected by solvent channels in the crystal packing. The structure of the NRP1-b1/KDKPPR complex was analyzed by MD simulations and innovative tools based on a multipolar electron density model [21].

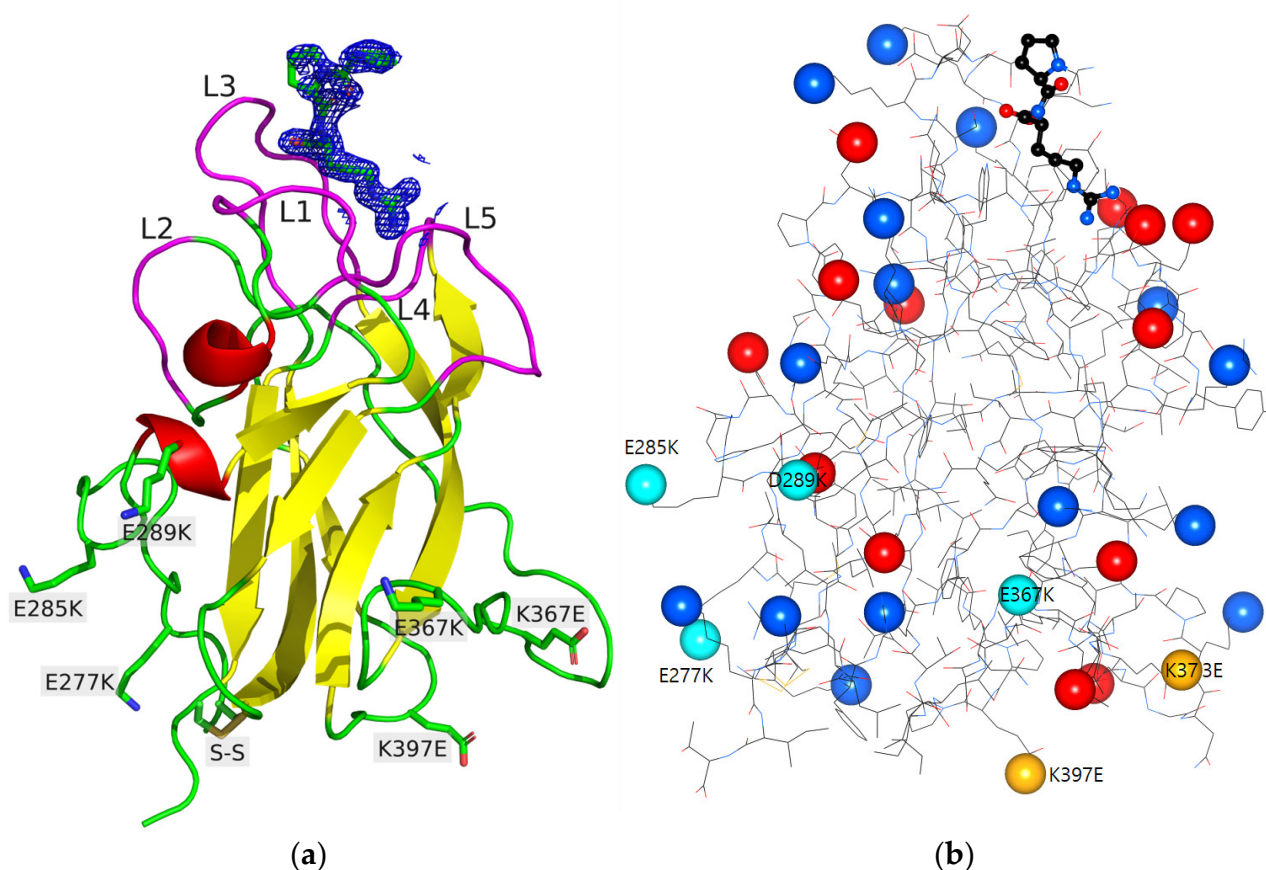


Figure 1. (a) Ribbon view of the crystal structure of NRP1-b1 hexavariant. The six mutations and the disulfide bridge Cys275-Cys424 are shown as sticks and labelled. The loops L1–L5 that line the VEGF-binding pocket are highlighted in magenta and labelled. The PPR moiety of KDKPPR peptide is shown as sticks with refined $2mFo-DFc$ electron density contoured at 1.0σ . (b) Repartition of the charged residues in NRP1-b1 hexavariant. The positive (blue) and negative (red) charges are shown as spheres on the protein structure. The mutated residues with change in charge are labeled and are shown in light blue and light yellow. The PR residues of KDKPPR peptide are shown on the top of the figure.

2. Results and Discussion

2.1. Design of the NRP1-b1 Hexavariant

The charge distribution on the surface of NRP1-b1 was significantly altered in order to promote the formation of a new crystal form, based on a visual inspection of the model. The point mutations were chosen to be far away from the VEGF-binding pocket to minimize disruption of the peptide-binding site (Figure 1b). Specifically, six charged residues on the protein surface were mutated to residues of opposite charge: Glu277Lys, Glu285Lys, Asp289Lys, Glu367Lys, Lys373Glu and Lys397Glu. Mutation of these residues did not lead to isoelectric conservation. In fact, the estimated isoelectric point of the hexavariant was 1 unit higher than that of the native protein, with a value of 9.2 for the hexavariant and 8.0 for the wild type. These choices resulted in the creation of a highly electropositive region around the Glu285Lys, Asp289Lys and Glu277Lys mutations, consisting of seven positively charged residues (Lys274, Lys277, Lys285, Lys289, Arg334, Arg402 and Lys425). Three of them (Lys277, Lys285 and Lys425) form hydrogen bonds with symmetry-related molecules in the crystal of the hexavariant (see below, Figure 2 and Table S1).

Form	PDB entry	β1'	β1	L1	α	L2	L5	β2	β3	β4'	L3
		5	15	25	35	45	55	65	75		
		277	287	297	307	317	327	337	347		
I	5C7G-A	----FKCMEA	LGME ES GEI HS	DQITASSQ S	TNWSAERS L	NYPENGWTPG	EDSYREWIVQ	DLGLLR F VTA	VG TQ GAIS KE		
II	5JGI-A	---MF K CMEA	LGME ES GEI HS	DQITASSQ S	TNWSAERS L	NYPENGWTPG	EDSYREWIVQ	DLGLLR F VTA	VG TQ GAIS KE		
III	5JGI-B	---MF K CMEA	LGME ES GEI HS	DQITASSQ S	TNWSAERS L	NYPENGWTPG	EDSYREWIVQ	DLGLLR F VTA	VG TQ GAIS KE		
IV	6FMC-A	---G H MF K CMEA	LGME ES GEI HS	DQITASSQ S	TNWSAERS L	NYPENGWTPG	EDSYREWIVQ	DLGLLR F VTA	VG TQ GAIS KE		
V	6TKK-A	---G H MF K CMEA	LGME ES GEI HS	DQITASSQ S	TNWSAERS L	NYPENGWTPG	EDSYREWIVQ	DLGLLR F VTA	VG TQ GAIS KE		
	7JJC-B	GG H MF K CMEA	LGME ES GEI HS	DQITASSQ S	TNWSAERS L	NYPENGWTPG	EDSYREWIVQ	DLGLLR F VTA	VG TQ GAIS KE		
	7JJC-C	GG H MF K CMEA	LGME ES GEI HS	DQITASSQ S	TNWSAERS L	NYPENGWTPG	EDSYREWIVQ	DLGLLR F VTA	VG TQ GAIS KE		
	7JJC-D	GG H MF K CMEA	LGME ES GEI HS	DQITASSQ S	TNWSAERS L	NYPENGWTPG	EDSYREWIVQ	DLGLLR F VTA	VG TQ GAIS KE		
VI	8PFE-A	GG H MF K CMEA	LGME ES CK HS	KQITASSQ S	TNWSAERS L	NYPENGWTPG	EDSYREWIVQ	DLGLLR F VTA	VG TQ GAIS KE		
	8PFE-C	GG H MF K CMEA	LGME ES CK HS	KQITASSQ S	TNWSAERS L	NYPENGWTPG	EDSYREWIVQ	DLGLLR F VTA	VG TQ GAIS KE		

Form	PDB entry	β4	β5'	β5''	β5	β6	β7	L4	β8
		85	95	105	115	125	135	145	155
		357	367	377	387	397	407	417	427
I	5C7G-A	TKKKY Y VK T	KIDVSS NG ED	WITIK E GN K P	VLFQGN T N P T	DVVVA V FP K P	LITR F VRI K P	ATWETG I SMR	FEVY G CK I T
II	5JGI-A	TKKKY Y VK T	KIDVSS NG ED	WITIK E GN K P	VLFQGN T N P T	DVVVA V FP K P	LITR F VRI K P	ATWETG I SMR	FEVY G CK I T
III	5JGI-B	TKKKY Y VK T	KIDVSS NG ED	WITIK E GN K P	VLFQGN T N P T	DVVVA V FP K P	LITR F VRI K P	ATWETG I SMR	FEVY G CK I T
IV	6FMC-A	TKKKY Y VK T	KIDVSS NG ED	WITIK E GN K P	VLFQGN T N P T	DVVVA V FP K P	LITR F VRI K P	ATWETG I SMR	FEVY G CK I T
V	6TKK-A	TKKKY Y VK T	KIDVSS NG ED	WITIK E GN K P	VLFQGN T N P T	DVVVA V FP K P	LITR F VRI K P	ATWETG I SMR	FEVY G CK I T
	7JJC-A	TKKKY Y VK T	KIDVSS NG ED	WITIK E GN K P	VLFQGN T N P T	DVVVA V FP K P	LITR F VRI K P	ATWETG I SMR	FEVY G CK I T
	7JJC-B	TKKKY Y VK T	KIDVSS NG ED	WITIK E GN K P	VLFQGN T N P T	DVVVA V FP K P	LITR F VRI K P	ATWETG I SMR	FEVY G CK I T
	7JJC-C	TKKKY Y VK T	KIDVSS NG ED	WITIK E GN K P	VLFQGN T N P T	DVVVA V FP K P	LITR F VRI K P	ATWETG I SMR	FEVY G CK I T
	7JJC-D	TKKKY Y VK T	KIDVSS NG ED	WITIK E GN K P	VLFQGN T N P T	DVVVA V FP K P	LITR F VRI K P	ATWETG I SMR	FEVY G CK I T
VI	8PFE-A	TKKKY Y VK T	KIDVSS NG ED	WITIE G GN K P	VLFQGN T N P T	DVVVA V FP E P	LITR F VRI K P	ATWETG I SMR	FEVY G CK I T
	8PFE-C	TKKKY Y VK T	KIDVSS NG ED	WITIE G GN K P	VLFQGN T N P T	DVVVA V FP E P	LITR F VRI K P	ATWETG I SMR	FEVY G CK I T

Figure 2. Highlights (in green) of the NRP1-b1 residues involved in contacts with a neighboring monomer in the crystal forms I to VI. Form VI corresponds to the NRP1-b1 hexavariant. Contacts are defined as residues with a proximity of less than 4 Å. Residues in bold characters forms intermolecular hydrogen bonds. The positions of the mutations have been highlighted with triangles above the two sets of residue numbering.

2.2. Crystal Structure of the NRP1-b1 Hexavariant

2.2.1. Description of the Structure

The hexavariant (Glu277Lys, Glu285Lys, Asp289Lys, Glu367Lys, Lys373Glu, Lys397Glu) of the NRP1-b1 domain crystallized in the $P3_221$ space group with a novel packing arrangement. The asymmetric unit contained two chains (A and C), each with a KDKPPR peptide (chains B and D) in its VEGF-binding site, plus 377 water molecules and an acetate ion. Only the last three residues of the peptide (PPR) in both monomers were included in the refined structure (Figure 1a). Positive residual peaks in the difference electron density maps persisted around the first proline residue of the KDKPPR peptide. These peaks were slightly stronger in monomer D. We made several attempts to improve the final $2mFo-DFc$ electron density map, such as modeling an additional residue in the peptide or modeling alternative conformations for the first proline residue. However, no satisfactory model has emerged from any of these efforts. The two monomers were nearly identical, with an overall coordinate root mean square deviation (RMSD, Å) of 0.30 Å on the 154 Cα atoms common to both chains. Slight conformational differences were observed at the very first and last residues of the two monomers, due to their different involvement in packing contacts. This probably also explained the differences observed at the neighboring disulfide bridge Cys275-Cys424 of the two monomers. Indeed, in monomer A, it showed two alternative conformations, one of which was rather ill-defined, whereas only one conformation was observed in monomer C. The same argument concerning packing effects probably applied for the slight differences observed in some side chain conformations (Phe335) and sometimes in some main chain regions (Glu374-Pro378, Pro396-Pro398). The mutations did not affect the protein fold, since the current hexavariant model showed an overall

RMSD of 0.51 Å with the wild-type structure (PDB entry 1KEX, [9]), compared to a 0.30 Å RMSD between the hexavariant independent monomers (A and C). Four of the mutations (Glu277Lys, Glu285Lys, Asp289Lys and Lys397Glu) resulted in no apparent change in the main chain fold. On the contrary, Glu367Lys and Lys373Glu were located in regions with larger observed displacements: the C α atom of residue 367 underwent a 1.75 Å shift due to an overall movement of region Ser363-Trp369, while the ψ angle of Gly375 rotated 180° in the rearrangement of the Glu373-Pro378 loop. Both of these observations are most likely the consequence of the change in crystal packing and protein–protein contacts (Figure 2), rather than the direct influence of the mutations on the polypeptide conformation.

2.2.2. Crystal Packing and Intermolecular Contacts

The mutations induced a new crystal packing in which a few introduced residues (Lys277 and Lys285 in chain A) were involved in modified intermolecular interactions. The two independent hexavariant models (chains A and C) have similar molecular environments. Indeed, more than half of the contacts are identical in both chains (Table S1). The KDKPPR peptide is not involved in the crystal cohesion. The hexavariant crystal contains large solvent spaces (volume of approximatively 30,000 Å³, Figure S1) into which it seems possible for small molecules to diffuse because they are connected by solvent channels. The N-terminal parts of the two independent KDKPPR peptides (KDK moiety, chains B and D) are located in the bulk solvent, while their C-terminal parts are tightly bound in the pocket that would welcome the C-terminal tail of VEGF, in chains A and C, respectively (see below). The electron density for the N-terminal regions of chains B and D was too weak to build a model, probably because these regions exhibited dynamic disorder.

We compared the hexavariant crystal form with those of the NRP1-b1 domain available in the Protein Data Bank [22]. Five structures of NRP1-b1 with distinct unit cell parameters were found (Figure 2, Table S2). The asymmetric units contain one to four independent chains. The crystal structure of the tetragonal form I represents the unbound state of NRP1-b1, in which the VEGF-binding site is obstructed by symmetry-related molecules. All other crystal forms were obtained by co-crystallization of NRP1-b1 with arginine or close derivatives (crystal forms II to V). The intermolecular environment analysis revealed that crystal forms II and III have similar packings (*P*₂₁ space group with two independent chains and *P*₄₁ space group with one independent chain, respectively) (Figure S2). We also noticed that in all cases, at least one independent ligand was positioned at the interface between two adjacent NRP1-b1 monomers (Figure S3). Their presence was probably necessary for the cohesion of these crystal packings. That is why we worked with a variant, trying to find a new crystal packing where the ligand only contacts the protein to which it is bound.

2.3. Protein Ligand Interaction

2.3.1. Description of Ligand Binding

The KDKPPR peptide was bound in the VEGF-binding pocket formed by loops L1 to L5 of NRP1-b1, with its C-terminal arginine positioned in a very similar manner to what was described in detail by [11]. Briefly, its guanidine group forms a salt bridge involving two hydrogen bonds with Asp320 (L5), while the aliphatic part of its side chain is stacked between the phenyl rings of Tyr297 (L1) and Tyr353 (L3) (Figure 3a). The arginine residue of the ligand is also tightly bound at the main chain by residues of L3, with one of its carboxylate oxygen atoms hydrogen-bonded to the side chain hydroxyl groups of Thr349 and Tyr353 and the second one to the lateral chain of Ser346. The last strong anchoring of the KDKPPR peptide arises from the main chain carbonyl group of the first proline, which forms a hydrogen bond with the hydroxyl group of Tyr297 (L1). The first three residues of the peptide apparently found no preferred binding to NRP1-b1 that could have fixed them in a given conformation and made them visible in the electron density.

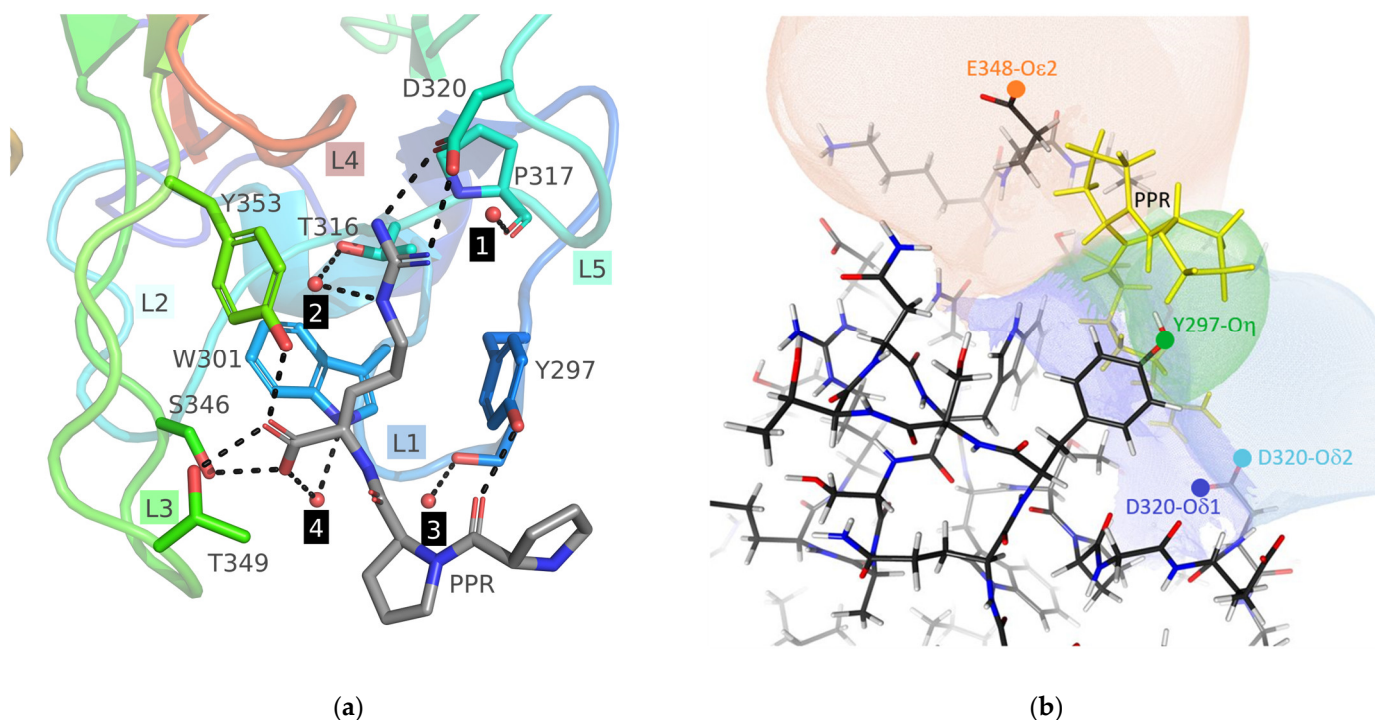


Figure 3. (a) Structure of the binding site of NRP1-b1 hexameric complex with the KDKPPR peptide. The pocket is mainly composed of five loops (L1 to L5), which are respectively colored blue, cyan, green, red and turquoise. The crystallographic model of the peptide includes only the PPR moiety, represented as sticks. The NRP1 residues in the close proximity of the peptide are also depicted as sticks. The four structural water molecules are highlighted as spheres, while hydrogen bonds are illustrated as dashed sticks. Various labels are provided to enhance clarity, indicating loops, peptide, residues and water molecules. (b) Nucleophilic Influence Zones associated with the oxygen atoms Asp320-O δ 1 (dark blue), Asp320-O ϵ 2 (light blue), Tyr297-O η (green) and Glu348-O ϵ 2 (major conformer, orange) in the vicinity of the ligand-binding site of monomer A. The corresponding atomic nucleophilic sites are indicated by colored circles, and the PPR moiety of the KDKPPR peptide is highlighted in yellow.

Ordered water molecules were checked and found identical to those discussed by [11]. Four structural water molecules in the peptide-binding site (HOH A690, A666, A717 and B101 for chain A and HOH C570, C549, C596 and D101 for chain B) were included in calculation of the contacts enrichment ratio (see below, Figure 3a).

2.3.2. Electrostatic Influence of the Protein on the Peptide

The VEGF-binding site of NRP1-b1 is occupied by the PPR moiety of the KDKPPR peptide, with its electrophilic C-terminal arginine residue forming a salt bridge with Asp320. For this reason, it is interesting to study the electrostatic influences from the whole NRP1 protein on the peptide from the point of view of Nucleophilic Influence Zones (NIZ) [23] (Figure 3b). A NIZ represents the volume containing all the electric field lines converging to a specific nucleophilic site, often an oxygen atom in proteins. Therefore, an electrophilic ligand within this space, like a positive charged entity, experiences attractive electrostatic forces directed toward the corresponding nucleophilic site. NIZs associated with the relevant oxygen atoms involved in the binding of the PPR moiety were calculated excluding ligand atoms and using CHARGER program [24].

As anticipated, the NH₂ groups of the arginine guanidinium of the KDKPPR peptide are influenced by their respective hydrogen-bonded oxygen atoms of Asp320 (dark blue surface for O δ 1 atom and light blue surface for O ϵ 2 atom) (Figure 3b). The NIZ of the Glu348-O ϵ 2 atom (orange surface) covers the position of one peptide proline residue, while

the NIZ of the Tyr297-O η atom (green surface) encompasses the majority of the other peptide proline residue. Hence, these NIZs illustrate the forces acting on both proline residues, specifically exerting an attraction on their hydrogen atoms, thus stabilizing the ligand conformation by pulling one proline residue towards Glu348 and the other towards Tyr297.

Given the close proximity of the solvent dielectric medium and the distances between the proline residues involved and the generators of the discussed NIZs (i.e., Glu348-O ϵ 2 and Tyr297-O η), it is likely that the electrostatic stabilization is relatively weak but cannot be disregarded. Electrostatic forces, being long-range interactions, play a role, and the presence of numerous charged hydrogen atoms in the pyrrolidine rings of prolines facilitates favorable interactions with the negatively charged oxygen atoms of the side chains of Glu348 and Tyr297, supporting our interpretation.

However, there are other factors contributing to the stabilization of the KDKPPR peptide that deserve attention. One such factor is the strong hydrogen bond between the hydroxyl group of Tyr297 and the carbonyl oxygen atom of the first proline residue in KDKPPR (Figure 3a). Furthermore, a stabilizing van der Waals contact can be inferred from the proximity of the Glu348-C γ hydrogen atoms and the C γ hydrogen atoms in the second proline residue of the KDKPPR peptide.

Furthermore, since electrostatic influences participate in the driving of ligand diffusion across the solvent toward a protein-binding site [25], the NIZs also provide insights on the electrostatic forces originating from the protein residues and directing the ligand during its approach. Here, the NIZs of Glu348-O ϵ 2 and Asp320-O δ 2 atoms extend beyond the protein surface, suggesting their potential contribution to the electrostatic steering effect that facilitates electropositive ligand fixation in the VEGF-binding site of NRP1-b1 (Figure 3b).

2.3.3. Hirshfeld Surface and Contacts Enrichment Ratio

The Hirshfeld surface [26] between the peptide and the protein was calculated with MoProViewer software [27]. The Hirshfeld surface allows the analysis and visualization of intermolecular interactions. The contact enrichment ratio E_{XY} between chemical species X and Y is obtained by comparing the actual C_{XY} contacts with those calculated as if all contact types had the same probability of forming [28]. The equiprobable proportions R_{XY} are derived by probability products from the chemical proportions on the Hirshfeld surface. An E_{XY} enrichment ratio greater than unity for a particular contact between chemical species X...Y indicates that these are over-represented. The chemical nature of the contacts and their enrichment in the complex of NRP1-b1 with KDKPPR peptide are shown in Table 1. The proportions of contact types are very similar in the two independent monomers of NRP1-b1 (correlation of C_{XY} contact type proportions of 99.9%).

The less polar Hc hydrogen atoms bonded to carbon were distinguished from the more electropositive Ho/n atoms bound to oxygen or nitrogen. Four structural water molecules in monomer A (and their equivalent in monomer B, see above) in the binding cleft were kept and attributed to the protein in the complex. Obviously, the O...Ho/n hydrogen bonds are strongly attractive from an electrostatic point of view and are overrepresented ($E = 2.58$, Table 1). Representing 17.4% of the interaction surface, they are recognized as the most favored contacts. This concerns interactions between C=O and COO $^-$ acceptors and N-H, NH $_2$ and O-H hydrogen bond donors. The O...Hc weak hydrogen bond contacts represent 15.3% of the interaction surface and can be considered as weakly favored contacts, with an enrichment ratio of $E = 1.20$. Some enrichment ratios close to zero concern the O...O and Hn/o...Hn/o contacts, which are absolutely avoided in the protein/ligand complex because they concern repulsive self-contacts between charged species.

Occupying the largest contact area (21.3%), non-polar Hc...C contacts are significantly over-represented ($E = 1.80$) and consist in particular of C-H... π interactions involving the aromatic rings of Tyr239, Tyr183 and Tyr187. The Hc...Hc contacts represent 10.3% of the surface and can be considered as weakly disfavored contacts, as they present an enrichment ratio lower than unity ($E = 0.81$). The Ho/n...W (oxygen atoms of water

molecules) contacts involving the four structural water molecules represent 6.9% of the interaction surface and can be considered as significantly favored contacts, with $E = 1.88$. The four water molecules interact essentially with Ho/n and secondarily with Hc atoms.

Table 1. Statistical analysis of intermolecular contacts on the Hirshfeld interface between the PPR moiety of the KDKPPR peptide and NRP1-b1.

Atom Type	C	Hc	N	Ho/n	O	W [#]
surface_peptide	8.8	42.7	5.5	22.9	20.1	0
surface_protein	21.4	30.0	1.2	15.6	15.9	16.0
C	1.4	21.3	1.3	3.4	0.4	1.0
Hc		10.3	3.5	7.2	15.3	4.8
N			0	0.5	0.2	1.3
Ho/n	C _{XY}	(%)		1.6	17.4	6.9
O					0.3	2.0
C	0.75	1.80	1.02	0.54	0.08	0.73
Hc		0.81	1.62	0.53	1.20	0.70
N			0	0.40	0.17	1.44
Ho/n	E _{XY}			0.44	2.58	1.88
O					0.09	0.64
		Hphob	Hphil	Hphob * Hphil		
surface %	peptide	57.1	43.0			
surface %	protein	52.6	47.4			
contacts	%	37.8	28.2	34.0		
enrichment		1.26	1.38	0.69		

* The Hirshfeld surface was limited to the regions where the electron density is larger than $0.0013 \text{ e}/\text{\AA}^3$ in order to omit the peptide surface exposed to the solvent. [#] Oxygen atom of water molecules. The second and third rows show the chemical content on the Hirshfeld surface. The next rows show the % C_{XY} of the contact types on the surface, followed by their enrichment ratios. The major surface components, the C_{XY} contacts and the significantly enriched contacts ($E > 1$) are highlighted in bold characters. In the lower part of the table, the atoms are grouped into hydrophobic (Hphob) and hydrophilic (Hphil) atoms.

The hydrophobic and hydrophilic atoms were regrouped in order to analyze the interactions between the two subgroups. The N contact surface occurs on sp² peptide and guanidinium nitrogen atoms (N without electron lone pair) and was considered hydrophobic together with the C and Hc atoms. The peptide and protein interaction surfaces are constituted by more hydrophobic (57.1 and 52.6%, respectively) than hydrophilic atoms. The protein/ligand complex shows an enrichment of contacts between hydrophilic atoms ($E = 1.26$) and between hydrophobic atoms ($E = 1.38$). On the other hand, despite the mild enrichment of the weak hydrogen bonds of O...Hc, the cross contacts Hphob × Hphil are strongly under-represented ($E = 0.69$), which indicates a good partitioning of hydrophilic and hydrophobic contacts.

In summary, the protein/ligand complex is mainly maintained by over-represented strong O...Ho/n interactions, which correspond notably to the salt bridge anchoring Asp206 and the arginine residue of the peptide. Secondarily, more moderately enriched interactions also play an important role, such as weak C-H...O hydrogen bonds and hydrophobic contacts, notably between Hc and C atoms. The enrichment values agree with trends found in studies of interactions in several families of oxygenated and nitrogenated hydrocarbon molecules [28,29]. The strong hydrogen bonds such as O...Ho/n are significantly enriched; in the case of small-molecule crystals, the over-representation reaches even larger values beyond 10.

Concerning the weak C-H...O hydrogen bonds, they tend to occur in a moderately under-represented way in crystal structures of small molecules containing both strong H-bond donors and acceptors (such as alcohols for example), due to the competition of strong H-bonds. On the contrary, in the present protein/peptide interface, they appear slightly enriched. This can be explained by the excess of strong H-bond acceptors on the protein ($S_O + S_W = 31.9\%$) compared to $S_{Ho/n} = 22.9\%$ of strong H-bond donors on the peptide.

2.4. Molecular Dynamics

Coordinate RMSD of NRP1-b1 was monitored along the molecular dynamics (MD) trajectories and stabilized after 50 ns in a range of 1–2 Å. Similar coordinate drift was observed for the bound and unbound proteins, as well as for the mutated and wild-type sequences. Only one copy of the mutated protein in complex with the peptide displays slightly larger coordinate drift (Figure 4).

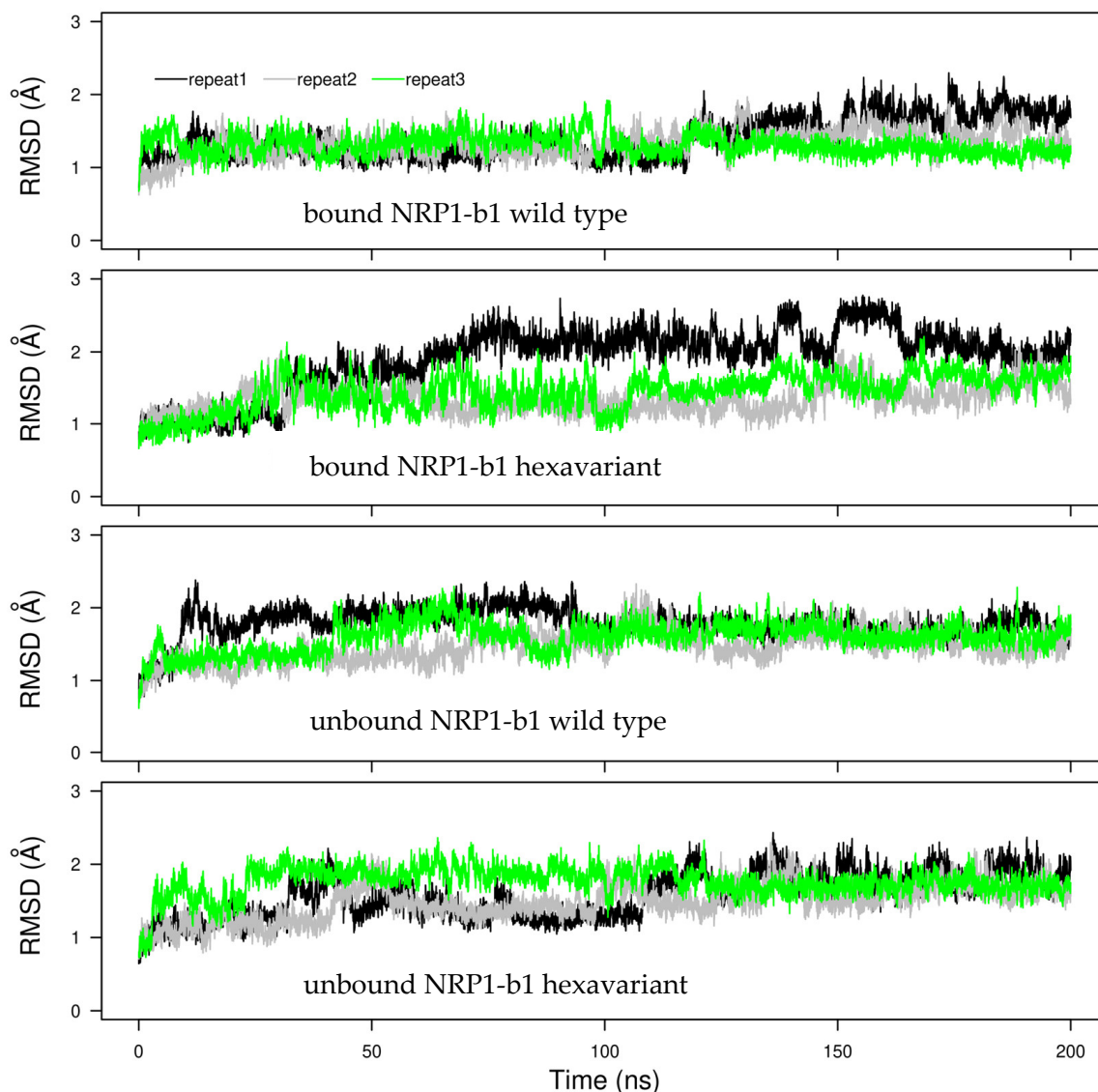


Figure 4. Coordinate RMSD (Å) calculated on the backbone heavy atoms of NRP1-b1 with respect to the initial X-ray crystallographic structure. The curves measured on the triplicated trajectories are colored in black, green and gray, respectively.

The atomic root mean square fluctuations (RMSFs, Å, Figure 5) of NRP1-b1 display quite superimposed profiles for all MD trajectories, with the unbound protein (dashed line) showing more mobile regions in particular for the loop of residues 317–322, which corresponds to the loop L5 of the VEGF-binding pocket. The very similar fluctuation profiles for the variant and WT forms show that the mutations introduced to alter crystal packing do not introduce a major bias in the dynamics of the protein. The N-terminal region of the KDKPPR peptide interacting with NRP1-b1 displays large fluctuations along the trajectories, in agreement with the invisible electronic density for this part of the peptide.

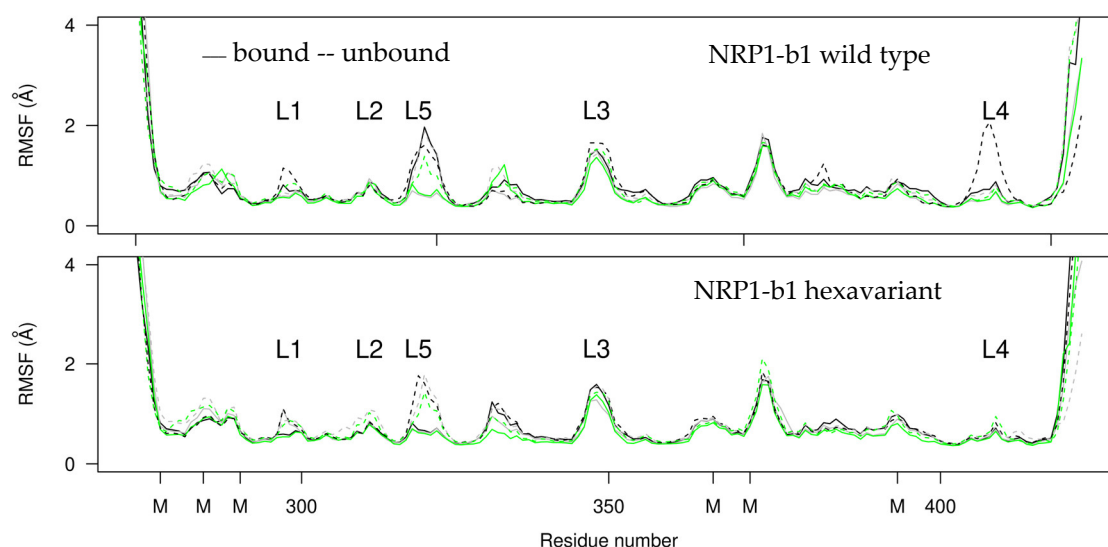


Figure 5. Atomic root mean square fluctuations (RMSFs, Å) calculated along the molecular dynamics (MD) trajectories by superimposing the heavy backbone atoms of NRP1-b1 on the corresponding atoms in the initial crystal structure. The upper panel corresponds to the trajectories recorded on the WT protein, while the lower panel corresponds to the trajectories recorded on the NRP1-b1 hexavariant used to determine the crystal structure. The curves measured on the triplicated trajectories are colored black, green and grey, respectively. They are plotted as solid and dashed lines for NRP1-b1 in complex with the peptide and for the unbound form, respectively. The positions in the sequence of the mutated residues are marked with the letter ‘M’. The NRP1 loops interacting with the peptide are labelled as defined in the text.

The root mean square thermal displacements (RMSTDs) of the C α atoms in the crystal structure were derived from the thermal parameters using the formula $RMSTD = \sqrt{B/8\pi^2}$. This estimation is meaningful because we obtained a high-resolution structure (1.35 Å, Table 2) [30]. The values were then compared to the average RMSF obtained from three MD simulations of NRP1-b1 hexavariant in complex with the peptide (Figure S4). RMSTD and RMSF show similar profiles, indicating a general agreement between both indicators. However, the values of RMSF are noticeably larger (up to 1.7 Å) in the most dynamic regions of the polypeptide chain, while RMSTDs consistently remain below 0.75 Å. One plausible explanation for this discrepancy is that proteins in the crystalline state typically exhibit reduced mobility compared to their counterparts in solution. Overall, RMSF and RMSTD show a correlation coefficient of 0.735, indicating a moderate correlation between the two parameters. The values of RMSF are noticeably larger (up to 1.7 Å) in the most dynamic region of the polypeptide chain, while RMSTDs consistently remain below 0.75 Å. Several MD simulations of the literature [31–34] have also shown larger fluctuations in solution compared to the crystal environment. Our observations agree with these references.

Looking more closely at the RMSF profiles (Figure 5), the loops interacting with the peptide all display fluctuation peaks whether the peptide is absent or present. The loop 347–350 (L3 loop) bearing Glu348 displays the same mobility. By contrast, a group of loops clustered on the other side of the binding side, the loop 296–300 (L1) bearing Tyr297 and Asn300, the loop 310–313 (L2) bearing Glu312 and the loop 318–322 (L5) bearing Glu319, are more mobile in the absence than in the presence of the ligand. Somehow, the two sides of the binding pocket behave differently with respect to the ligand. One moiety (L1, L2 and L5) stabilizes upon binding, while the other moiety (L3) retains the same mobility. A recent MD study by Alshawaf et al. [35] found similar RMSF profiles in NRP1-b1 when complexed with specialized metabolites 3-O-methylquercetin and esculetin. Specifically, the fluctuations in the esculetin/NRP1-b1 complex were similar to those observed in our study of the unbound form, while the 3-O-methylquercetin/NRP1-b1 complex was more

similar to our complex with the peptide. Notably, the study found that 3-O-methylquercetin had a more favorable energy of interaction with NRP1-b1 than esculetin [35]. Our “bound” RMSF profile may be indicative to a state of NRP1-b1 that allows stable interactions with a ligand.

Table 2. Statistics of X-ray diffraction data collection and model refinement.

Data Collection	
Diffraction source	ESRF FIP2-BM07
Wavelength (Å)	0.9795
Space group	$P3_22$
a, b, c (Å)	59.77, 59.77, 174.60
α, β, γ (°)	90, 90, 120
Resolution range (Å)	44.53–1.35 (1.37–1.35) ¹
Total number of measured intensities	696,189 (16,486) ¹
Number of unique reflections	80,230 (3895) ¹
Average redundancy	8.7 (4.2) ¹
Mean $I/\sigma(I)$	28.7 (1.9) ¹
Completeness (%)	99.7 (95.7) ¹
R_{merge} ² ; R_{meas} ³	0.031 (0.681) ¹ ; 0.033 (0.779) ¹
$CC_{1/2}$ ⁴	1.00 (0.69) ¹
Wilson B -factor (Å ²)	17.9 (Aimless)/21.14 (Buster)
Refinement and structure	
Resolution range (Å)	19.57–1.35 (1.36–1.35) ¹
Number of reflections	80,203 (1605) ¹
R_{work}/R_{free} ⁵	0.1942/0.2100 (0.2826/0.2760) ¹
Correlation $F_o - F_c/F_o - F_{c_{free}}$	0.966/0.963
Total number of atoms	2910
Average B factor (Å ²)	25.55
Model quality	
RMSZ bond lengths ⁶	1.28
RMSZ bond angles ⁶	1.14
Ramachandran favored (%)	97.5
Ramachandran allowed (%)	2.4
Rotamer outliers (%)	1.8
Clash-score ⁷	10

¹ Values in parentheses are for the highest resolution shell; ² $R_{merge} = \sum_h \sum_i |I_{hi} - \langle I_h \rangle| / \sum_h \sum_i \langle I_h \rangle$; ³ $R_{meas} =$

$\sum_h \sum_i \left(\frac{n_h}{n_h - 1} \right)^{1/2} |I_{hi} - \langle I_h \rangle| / \sum_h \sum_i \langle I_h \rangle$ (with I_{hi} being the intensity of an individual observation of the reflection

h and $\langle I_h \rangle$ being the average of all symmetry-related or replicate observations); ⁴ $CC_{1/2}$ is the correlation coefficient of the mean intensities between two random half-sets of data. ⁵ $R_{work} = \sum_h ||F_o| - |F_c|| / \sum_h |F_o|$, 95% of the reflections, R_{free} same formula (5% of the reflections) (F_o and F_c observed and calculated structure factors, respectively). ⁶ RMSZ: root mean square Z-score. ⁷ The MolProbity clash-score is the number of serious clashes per 1000 atoms.

Comparing the fluctuation profiles of the WT and of the mutated sequences of NRP1-b1 (Figure 5), only one major difference can be noticed: the loop 411–416 (L4), interacting with the peptide, displays a fluctuation peak for one of the unbound trajectories on the WT sequence, whereas this loop is quite rigid in all trajectories recorded for the modified sequence.

The loops 282–289 and 373–378, located in the bottom of the structure, display high mobility in all conditions. The loop 282–289, containing the charged and polar sequence ESGEIHSD, becomes more mobile in the absence of peptide. The sequence ESGEIHSD (residues 282–289) is located at the surface of b1 domain close to the surface of a2 domain in NRP1 structures, which contain a2, b1 and b2 domains (PDB entry 2QQM, [36]) or a1, a2, b1 and b2 domains (PDB entry 4GZ9, [37]). A similar configuration is also visible in the more recent cryo-EM structure of the Sema3A/PlexinA4/NRP1 tripartite complex [1]. As

the b1 and a2 interface do not form direct contact in any of these structures, it is difficult to speculate on the precise functional effect of the mobility of the region, but this mobility may have an influence on the propagation of conformational signals during the physiological processes.

3. Materials and Methods

3.1. Protein Production and Purification

The gene of NRP1-b1 hexavariant (residues Met272 to Thr427 of NRP1 with mutations Glu277Lys, Glu285Lys, Asp289Lys, Glu367Lys, Lys373Glu, Lys397Glu) was cloned into pET15b-NRP16mut-6His-3C (Novagen, Pretoria, South Africa) and expressed in *Escherichia coli* after induction at 18 °C. Cells were grown in Terrific-Broth at 18 °C to optical density (OD) = 0.36. After 8 h 25 min at 18 °C, they were induced with 0.2 mM Isopropyl β -D-1 thiogalactopyranoside. After growth at 18 °C for 10 h 25 min to OD = 0.7, cells were harvested by centrifugation, lysed and centrifuged. Proteins were purified on HIS-Select nickel affinity resin (Sigma-Aldrich, St. Louis, MO, USA) in 50 mM Tris (pH 8.0) and 300 mM NaCl with a 250 mM imidazole gradient. The protein was further purified by gel filtration using a Superdex75 HiLoad 16/60 column (GE Healthcare, Piscataway, NJ, USA) equilibrated and run in buffer A (50 mM Tris, pH 8, 300 mM NaCl, 20 mM imidazole). Analytical gel filtration experiments were performed using a Superdex 75 10/16 column (GE Healthcare) in buffer A. The protein was concentrated to 63 mg/mL by centrifugation on an Amicon ultrafiltration unit with a 10 kDa molecular weight cutoff, in a solution of Tris/HCl pH 8, NaCl 50 mM.

3.2. KDKPPR Synthesis on Solid Phase

The KDKPPR peptide was synthesized using the automated ResPepXL peptide synthesizer, with a Fmoc/tBu methodology. The side chains of arginine, lysine and aspartic acid were protected by Pbf, OtBu and Boc groups. A Fmoc-Arg(Pbf)-Wang resin swelled in DCM was used. The Fmoc group was removed by a piperidine solution (20% in DMF), and this step was performed two times (the first for 4 min and the second for 7 min). Then, the next AA was grafted by adding an excess of Fmoc-AA-OH (6 eq), HBTU (5 eq), NMP (3 eq) and NMM (10 eq) in DMF, and this step was repeated two times for 18 min. A last step of capping, using a solution of acetic anhydride (5% in DMF), was performed for 5 min to trap all amino functions that did not react. Deprotection, coupling and capping steps were repeated until the end of the synthesis of the peptide. After a last Fmoc deprotection, the resin was dried under vacuum and then cleaved (with full deprotection of lateral chains) using TFA/TIPS/water (92.5/2.5/5, v/v/v) for 2 h. The acidic resin was filtered and washed with DCM and EtOH. The filtrate was dried under vacuum, and the compound was precipitated in diethylether by centrifugation. TSK gel Amide-80 column was used for HILIC purification of KDKPPR peptide using acetonitrile/water (0.1% TFA; 95/5 (v/v) to 55/45 (v/v) gradient) for 15 min, followed by an isocratic elution (0.1% TFA; 55/45, v/v) for 10 min at a flow of 12 mL/min (R_t = 20.8 min). KDKPPR was isolated as a white powder with a yield of 64% and a purity of 95% (UV-vis detection at 214 nm).

3.3. Crystallization

Crystallization was conducted by sitting-drop vapor diffusion method. The reservoir solutions were prepared by mixing 0.3 μ L of commercial reservoir solutions (screens) and 0.3 μ L of protein solution. The crystallization trials of NRP1-b1 hexavariant were carried out in the presence of the KDKPPR hexapeptide at 20 °C. Crystals appeared with a JCSGplus solution composed of 0.2 M ammonium citrate, 0.1 M bis-tris pH 5.5 and 25% w/v PEG 3350.

3.4. X-ray Diffraction Data Collection and Crystal Structure Determination

Crystals of NRP1-b1 hexavariant that appeared suitable for X-ray diffraction data collection were quickly soaked in their mother liquor supplemented with 20% glycerol (v/v),

before flash freezing in a nitrogen stream at 100 K. Preliminary X-ray diffraction experiments were carried out in-house on an Agilent SuperNova diffractometer (Oxford Diffraction, Oxford, UK) equipped with a CCD detector, and high-resolution data were further collected by ESRF synchrotron on beamline BM07 (Grenoble, France). The data set was indexed and integrated with XDS [38], scaled, and merged with Aimless [39] from the CCP4 suite [40]. The atomic structure was solved by molecular replacement using MOLREP [41] with the coordinates of NRP1-b1 wild type (PDB code 5C7G, [19]) as the search model. The structure was manually adjusted with Coot [42] and refined with Buster [43]. Structure validation was performed with MolProbity [44] and the wwPDB validation server (<http://validate.wwpdb.org>, (accessed on 22 december 2022)). Diffraction data and refinement statistics are shown in Table 2. Figures of the protein structures were generated with Pymol (Schrödinger LLC, New York, NY, USA), MoProViewer [27] and Ligplot+ [45], and cleft volume calculations were performed with 3V [46]. Coordinates and structure factors were deposited in the Protein Data Bank (PDB ID: 8PFE, DOI:10.2210/pdb8pfe/pdb).

3.5. Nucleophilic Influence Zones

The Nucleophilic Influence Zones were calculated from the electrostatic potential using Charger module of MoProviewer [24,27]. The electrostatic potential was generated from an electron density model, based on transferred multipolar parameters of the ELMAM2 library [21].

3.6. Hirshfeld Analysis

MoProViewer software [27] was used to investigate the intermolecular interactions and the contacts enrichment on the Hirshfeld interface between the protein molecules and the ligand peptides. The intermolecular interactions were evaluated by computing the enrichment ratios (Table 1) in order to highlight which contacts are favored. The enrichment values are obtained as the ratio between the proportions of actual contacts C_{XY} and the equiprobable (random) contacts R_{XY} , the latter being obtained by probability products ($R_{XY} = S_X S_Y$).

Contacts $X \dots Y$, which are over-represented with respect to the share of X and Y chemical species on the Hirshfeld surface, have enrichments larger than unity. They are likely to represent interactions that are attractive from an electrostatic point of view and shall be the driving force in the complex formation [28]. Interactions between atoms that have electric charges of the same sign are repulsive and are generally under-represented ($E < 1$).

3.7. Molecular Dynamics

The protein and peptide chains C and D were selected from the X-ray crystallographic structure. The hydrogen atoms were added, and the flip of side chains was optimized using the Molprobity server [44]. The NRP1-b1 crystal structure with six mutations was unmodified, whereas for the wild-type (WT) system, the mutations Lys277Glu, Lys285Glu, Lys289Asp, Lys367Glu, Glu373Lys and Glu397Lys were introduced to return to the WT sequence of NRP1 for both protein sequences. The bound and unbound systems were simulated.

For each previously described system, the protein was embedded in a water box. Sodium and chloride counterions were added to obtain an ionic concentration of 0.15 M. The total number of atoms was about 31,000 in both cases. All MD simulations were performed using NAMD 2.14 [47], with the CHARMM36 force field [48] for protein and the TIP3P model for water [49]. A cutoff of 12 Å and a switching distance of 10 Å were used for non-bonded interactions, while long-range electrostatic interactions were calculated with the Particle Mesh Ewald (PME) method [50]. The RATTLE algorithm [51] was used to keep rigid all covalent bonds involving hydrogen atoms, enabling a time step of 2 fs. At the beginning of each trajectory, the system was minimized for 20,000 steps, and it was then heated up gradually from 0 K to 310 K in 31,000 integration steps. Finally, the system was

equilibrated for 1 ns in the NPT ensemble at 310 K. During the equilibration stage, the C α atoms were kept fixed. Simulations were then performed in the NPT ensemble ($P = 1$ bar, $T = 310$ K), with all atoms free to move. Atomic coordinates were saved every 10 ps. For each trajectory, 200 ns of production and the trajectories were triplicated for a cumulative trajectory duration of 3 μ s.

4. Conclusions

In this study, we have reported the crystal structure of a hexavariant of the domain b1 of human NRP1 in complex with the KDKPPR peptide. The mutant was designed to modify the monomer assembly observed in the crystal packing of the unbound form [9,19], in which the VEGF-binding pocket of NRP1-b1 is inaccessible. Molecular dynamic trajectories permitted investigating the differences in the structures of the wild type and variant. Both structures produced similar internal flexibility and protein/peptide interaction. This showed that the ability of the protein to bind small ligands was not affected by the designed mutations. As part of our future search for ligands, we need to check that the dissociation constant (K_D) of the molecules tested is the same for mutated and wild-type NRP1-b1.

The NRP1-b1 hexavariant crystallized in a new crystal form, in which the KDKPPR peptide was not involved in the cohesion of the solid state. In the crystal, the peptide-binding site was observed to communicate with a solvent cavity large enough to diffuse small molecules. Therefore, ligand soaking in the crystal of the unbound form of NRP1-b1 hexavariant could be considered as a strategy to prepare NRP1-b1 complexes with peptides that target the pocket where VEGF binds.

The structure of the NRP1-b1 hexavariant in complex with the KDKPPR peptide was analyzed with two original tools. First, the Nucleophilic Influence Zones (NIZ) of the ligand-binding cleft were analyzed. They revealed two additional residues (Tyr297 and Glu348) as probable attractors of the ligand electrophilic groups and two residues (Asp320 and Glu348) in the electrostatic steering of the ligand in its binding site. Secondly, the enrichment of contacts was calculated to analyze the interactions between the protein and the peptide. This metric has provided valuable insights into the diversity and specificity of the protein/ligand interaction. The complex was mainly stabilized by a notable presence of strong N-H...O and O-H...O hydrogen bonds, which were crucial due to the loop-rich nature of the VEGF-binding site. Indeed, these loops exhibited mobility both in the unbound and bound forms, as suggested by the MD simulation.

Supplementary Materials: The following supporting information can be downloaded at <https://www.mdpi.com/article/10.3390/molecules28145603/s1>, Figure S1. Large aqueous cavity in the crystal structure of NRP1-b1 hexavariant; Figure S2: Comparison of the packing of NRP1-b1 in crystal forms II (a) and III (b); Figure S3: Environment of the ligand in crystal forms II, III, IV and V; Figure S4: Comparison of atomic root mean square fluctuations (RMSFs) and root mean square thermal displacements (RMSTDs); Table S1: Intermolecular hydrogen bonds between NRP1-b1 chains in hexavariant crystal; Table S2: Crystal forms of NRP1-b1 fragment in the unbound form or in complex with small ligands.

Author Contributions: Conceptualization, C.D., C.F., C.J., F.F., S.A. and T.E.M.; methodology, C.D., C.J. and T.E.M.; formal analysis, I.G., S.A., B.G., E.M., F.F., C.D., C.J., S.M. and T.E.M.; investigation, I.G., S.A., B.G., E.M., F.F., C.D., C.J., S.M. and T.E.M.; writing—original draft preparation, I.G., T.E.M., F.F., C.D. and C.J.; writing—review and editing, C.F., I.G., S.A., B.G., E.M., F.F., C.D., C.J., S.M., M.B.-H. and T.E.M.; supervision, C.D. and C.J.; project administration, C.D. and C.J.; funding acquisition, C.D. and C.J. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: PDB data (8PFE) are made freely available by the wwPDB (<https://www.wwpdb.org/> (accessed on 21 July 2023)).

Acknowledgments: The authors appreciated the access to the “Plateforme de mesures de diffraction X” of the Université de Lorraine. We acknowledge ESRF (Grenoble, France) for providing synchrotron radiation facilities, and we thank the staffs of BM07 beamlines for assistance.

Conflicts of Interest: The authors declare no conflict of interest.

Sample Availability: Not applicable.

References

- Lu, D.; Shang, G.; He, X.; Bai, X.C.; Zhang, X. Architecture of the Sema3A/PlexinA4/Neuropilin tripartite complex. *Nat. Commun.* **2021**, *12*, 3172. [[CrossRef](#)] [[PubMed](#)]
- Soker, S.; Miao, H.-Q.; Nomi, M.; Takashima, S.; Klagsbrun, M. VEGF165 mediates formation of complexes containing VEGFR-2 and neuropilin-1 that enhance VEGF165-receptor binding. *J. Cell. Biochem.* **2002**, *85*, 357–368. [[CrossRef](#)] [[PubMed](#)]
- Soker, S.; Takashima, S.; Miao, H.Q.; Neufeld, G.; Klagsbrun, M. Neuropilin-1 Is Expressed by Endothelial and Tumor Cells as an Isoform-Specific Receptor for Vascular Endothelial Growth Factor. *Cell* **1998**, *92*, 735–745. [[CrossRef](#)]
- Chaudhary, B.; Khaled, Y.S.; Ammori, B.J.; Elkord, E. Neuropilin 1: Function and therapeutic potential in cancer. *Cancer Immunol. Immunother.* **2014**, *63*, 81–99. [[CrossRef](#)]
- Daly, J.L.; Simonetti, B.; Klein, K.; Chen, K.-E.; Williamson, M.K.; Antón-Plágaro, C.; Shoemark, D.K.; Simón-Gracia, L.; Bauer, M.; Hollandi, R.; et al. Neuropilin-1 is a host factor for SARS-CoV-2 infection. *Science* **2020**, *370*, 861–865. [[CrossRef](#)]
- Lerouge, L.; Gries, M.; Chateau, A.; Daouk, J.; Lux, F.; Rocchi, P.; Cedervall, J.; Olsson, A.K.; Tillement, O.; Frochot, C.; et al. Targeting Glioblastoma-Associated Macrophages for Photodynamic Therapy Using AGuIX((R))-Design Nanoparticles. *Pharmaceutics* **2023**, *15*, 997. [[CrossRef](#)] [[PubMed](#)]
- Liu, S.D.; Zhong, L.P.; He, J.; Zhao, Y.X. Targeting neuropilin-1 interactions is a promising anti-tumor strategy. *Chin. Med. J.* **2020**, *134*, 508–517. [[CrossRef](#)]
- Pellet-Many, C.; Frankel, P.; Jia, H.; Zachary, I. Neuropilins: Structure, function and role in disease. *Biochem. J.* **2008**, *411*, 211–226. [[CrossRef](#)]
- Lee, C.C.; Kreusch, A.; McMullan, D.; Ng, K.; Spraggon, G. Crystal Structure of the Human Neuropilin-1 b1 Domain. *Structure* **2003**, *11*, 99–108. [[CrossRef](#)]
- Powell, J.; Mota, F.; Steadman, D.; Soudy, C.; Miyauchi, J.T.; Crosby, S.; Jarvis, A.; Reisinger, T.; Winfield, N.; Evans, G.; et al. Small Molecule Neuropilin-1 Antagonists Combine Antiangiogenic and Antitumor Activity with Immune Modulation through Reduction of Transforming Growth Factor Beta (TGFbeta) Production in Regulatory T-Cells. *J. Med. Chem.* **2018**, *61*, 4135–4154. [[CrossRef](#)]
- Mota, F.; Fotinou, C.; Rana, R.R.; Chan, A.W.E.; Yelland, T.; Arooz, M.T.; O’Leary, A.P.; Hutton, J.; Frankel, P.; Zachary, I.; et al. Architecture and hydration of the arginine-binding site of neuropilin-1. *FEBS J.* **2018**, *285*, 1290–1304. [[CrossRef](#)] [[PubMed](#)]
- Parker, M.W.; Xu, P.; Li, X.; Vander Kooi, C.W. Structural basis for selective vascular endothelial growth factor-A (VEGF-A) binding to neuropilin-1. *J. Biol. Chem.* **2012**, *287*, 11082–11089. [[CrossRef](#)] [[PubMed](#)]
- Jarvis, A.; Allerston, C.K.; Jia, H.; Herzog, B.; Garza-Garcia, A.; Winfield, N.; Ellard, K.; Aqil, R.; Lynch, R.; Chapman, C.; et al. Small molecule inhibitors of the neuropilin-1 vascular endothelial growth factor A (VEGF-A) interaction. *J. Med. Chem.* **2010**, *53*, 2215–2226. [[CrossRef](#)]
- Kamarulzaman, E.E.; Vanderesse, R.; Gazzali, A.M.; Barberi-Heyob, M.; Boura, C.; Frochot, C.; Shawkataly, O.; Aubry, A.; Wahab, H.A. Molecular modelling, synthesis and biological evaluation of peptide inhibitors as anti-angiogenic agent targeting neuropilin-1 for anticancer application. *J. Biomol. Struct. Dyn.* **2017**, *35*, 26–45. [[CrossRef](#)]
- Kamarulzaman, E.E.; Gazzali, A.M.; Acherar, S.; Frochot, C.; Barberi-Heyob, M.; Boura, C.; Chaimbault, P.; Sibille, E.; Wahab, H.A.; Vanderesse, R. New Peptide-Conjugated Chlorin-Type Photosensitizer Targeting Neuropilin-1 for Anti-Vascular Targeted Photodynamic Therapy. *Int. J. Mol. Sci.* **2015**, *16*, 24059–24080. [[CrossRef](#)] [[PubMed](#)]
- Bechet, D.; Mordon, S.R.; Guillemin, F.; Barberi-Heyob, M.A. Photodynamic therapy of malignant brain tumours: A complementary approach to conventional therapies. *Cancer Treat. Rev.* **2014**, *40*, 229–241. [[CrossRef](#)]
- Gries, M.; Thomas, N.; Daouk, J.; Rocchi, P.; Choulier, L.; Jubreaux, J.; Pierson, J.; Reinhard, A.; Jouan-Hureaux, V.; Chateau, A.; et al. Multiscale Selectivity and in vivo Biodistribution of NRP-1-Targeted Theranostic AGuIX Nanoparticles for PDT of Glioblastoma. *Int. J. Nanomed.* **2020**, *15*, 8739–8758. [[CrossRef](#)]
- Thomas, E.; Colombeau, L.; Gries, M.; Peterlini, T.; Mathieu, C.; Thomas, N.; Boura, C.; Frochot, C.; Vanderesse, R.; Lux, F.; et al. Ultrasmall AGuIX theranostic nanoparticles for vascular-targeted interstitial photodynamic therapy of glioblastoma. *Int. J. Nanomed.* **2017**, *12*, 7075–7088. [[CrossRef](#)]
- Richard, M.; Chateau, A.; Jelsch, C.; Didierjean, C.; Manival, X.; Charron, C.; Maigret, B.; Barberi-Heyob, M.; Chapleur, Y.; Boura, C.; et al. Carbohydrate-based peptidomimetics targeting neuropilin-1: Synthesis, molecular docking study and in vitro biological activities. *Bioorgan. Med. Chem.* **2016**, *24*, 5315–5325. [[CrossRef](#)]
- Jelsch, C.; Longhi, S.; Cambillau, C. Packing forces in nine crystal forms of cutinase. *Proteins Struct. Funct. Genet.* **1998**, *31*, 320–333. [[CrossRef](#)]
- Domagala, S.; Fournier, B.; Liebschner, D.; Guillot, B.; Jelsch, C. An improved experimental databank of transferable multipolar atom models—ELMAM2. Construction details and applications. *Acta Crystallogr. A* **2012**, *68*, 337–351. [[CrossRef](#)] [[PubMed](#)]

22. wwPDB consortium. Protein Data Bank: The single global archive for 3D macromolecular structure data. *Nucleic Acids Res.* **2019**, *47*, D520–D528. [[CrossRef](#)] [[PubMed](#)]
23. 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. [[CrossRef](#)] [[PubMed](#)]
24. Vuković, V.; Leduc, T.; Jelić-Matošević, Z.; Didierjean, C.; Favier, F.; Guillot, B.; Jelsch, C. A rush to explore protein–ligand electrostatic interaction energy with Charger. *Acta Crystallogr. Sect. D Struct. Biol.* **2021**, *77*, 1292–1304. [[CrossRef](#)] [[PubMed](#)]
25. Wade, R.C.; Gabdoulhine, R.R.; Lüdemann, S.K.; Lounnas, V. Electrostatic steering and ionic tethering in enzyme–ligand binding: Insights from simulations. *Proc. Natl. Acad. Sci. USA* **1998**, *95*, 5942–5949. [[CrossRef](#)]
26. Hirshfeld, F.L. Bonded-atom fragments for describing molecular charge densities. *Theor. Chim. Acta* **1977**, *44*, 129–138. [[CrossRef](#)]
27. Guillot, B.; Enrique, E.; Huder, L.; Jelsch, C. MoProViewer: A tool to study proteins from a charge density science perspective. *Acta Crystallogr. Sect. A* **2014**, *70*, C279. [[CrossRef](#)]
28. Jelsch, C.; Ejsmont, K.; Huder, L. The enrichment ratio of atomic contacts in crystals, an indicator derived from the Hirshfeld surface analysis. *IUCr* **2014**, *1*, 119–128. [[CrossRef](#)]
29. Jelsch, C.; Bibila Mayaya Bisseyou, Y. Atom interaction propensities of oxygenated chemical functions in crystal packings. *IUCr* **2017**, *4*, 158–174. [[CrossRef](#)]
30. Sun, Z.; Liu, Q.; Qu, G.; Feng, Y.; Reetz, M.T. Utility of B-Factors in Protein Science: Interpreting Rigidity, Flexibility, and Internal Motion and Engineering Thermostability. *Chem. Rev.* **2019**, *119*, 1626–1665. [[CrossRef](#)]
31. Ahlstrom, L.S.; Miyashita, O. Packing interface energetics in different crystal forms of the λ Cro dimer. *Proteins Struct. Funct. Bioinform.* **2014**, *82*, 1128–1141. [[CrossRef](#)]
32. Ahlstrom, L.S.; Vorontsov, I.I.; Shi, J.; Miyashita, O. Effect of the Crystal Environment on Side-Chain Conformational Dynamics in Cyanovirin-N Investigated through Crystal and Solution Molecular Dynamics Simulations. *PLoS ONE* **2017**, *12*, e0170337. [[CrossRef](#)] [[PubMed](#)]
33. Janowski, P.A.; Liu, C.; Deckman, J.; Case, D.A. Molecular dynamics simulation of triclinic lysozyme in a crystal lattice. *Protein Sci.* **2016**, *25*, 87–102. [[CrossRef](#)] [[PubMed](#)]
34. Kuzmanic, A.; Pannu, N.S.; Zagrovic, B. X-ray refinement significantly underestimates the level of microscopic heterogeneity in biomolecular crystals. *Nat. Commun.* **2014**, *5*, 3220. [[CrossRef](#)] [[PubMed](#)]
35. Alshawaf, E.; Hammad, M.M.; Marafie, S.K.; Ali, H.; Al-Mulla, F.; Abubaker, J.; Mohammad, A. Discovery of natural products to block SARS-CoV-2 S-protein interaction with Neuropilin-1 receptor: A molecular dynamics simulation approach. *Microb. Pathog.* **2022**, *170*, 105701. [[CrossRef](#)]
36. Appleton, B.A.; Wu, P.; Maloney, J.; Yin, J.; Liang, W.-C.; Stawicki, S.; Mortara, K.; Bowman, K.K.; Elliott, J.M.; Desmarais, W.; et al. Structural studies of neuropilin/antibody complexes provide insights into semaphorin and VEGF binding. *EMBO J.* **2007**, *26*, 4902–4912. [[CrossRef](#)]
37. Janssen, B.J.C.; Malinauskas, T.; Weir, G.A.; Cader, M.Z.; Siebold, C.; Jones, E.Y. Neuropilins lock secreted semaphorins onto plexins in a ternary signaling complex. *Nat. Struct. Mol. Biol.* **2012**, *19*, 1293–1299. [[CrossRef](#)]
38. Kabsch, W. XDS. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **2010**, *66*, 125–132. [[CrossRef](#)]
39. Evans, P.R.; Murshudov, G.N. How good are my data and what is the resolution? *Acta Crystallogr. D Biol. Crystallogr.* **2013**, *69*, 1204–1214. [[CrossRef](#)]
40. Winn, M.D.; Ballard, C.C.; Cowtan, K.D.; Dodson, E.J.; Emsley, P.; Evans, P.R.; Keegan, R.M.; Krissinel, E.B.; Leslie, A.G.W.; McCoy, A.; et al. Overview of the CCP4 suite and current developments. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **2011**, *67*, 235–242. [[CrossRef](#)]
41. Vagin, A.; Teplyakov, A. Molecular replacement with MOLREP. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **2010**, *66*, 22–25. [[CrossRef](#)]
42. Emsley, P.; Lohkamp, B.; Scott, W.G.; Cowtan, K. Features and development of Coot. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **2010**, *66*, 486–501. [[CrossRef](#)]
43. Smart, O.S.; Womack, T.O.; Flensburg, C.; Keller, P.; Paciorek, W.; Sharff, A.; Vonrhein, C.; Bricogne, G. Exploiting structure similarity in refinement: Automated NCS and target-structure restraints in BUSTER. *Acta Crystallogr. Sect. D* **2012**, *68*, 368–380. [[CrossRef](#)] [[PubMed](#)]
44. Williams, C.J.; Headd, J.J.; Moriarty, N.W.; Prisant, M.G.; Videau, L.L.; Deis, L.N.; Verma, V.; Keedy, D.A.; Hintze, B.J.; Chen, V.B.; et al. MolProbity: More and better reference data for improved all-atom structure validation. *Protein Sci.* **2018**, *27*, 293–315. [[CrossRef](#)] [[PubMed](#)]
45. Laskowski, R.A.; Swindells, M.B. LigPlot+: Multiple Ligand–Protein Interaction Diagrams for Drug Discovery. *J. Chem. Inf. Model.* **2011**, *51*, 2778–2786. [[CrossRef](#)]
46. Voss, N.R.; Gerstein, M. 3V: Cavity, channel and cleft volume calculator and extractor. *Nucleic Acids Res.* **2010**, *38*, W555–W562. [[CrossRef](#)]
47. Phillips, J.C.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R.D.; Kalé, L.; Schulten, K. Scalable molecular dynamics with NAMD. *J. Comput. Chem.* **2005**, *26*, 1781–1802. [[CrossRef](#)]
48. Best, R.B.; Zhu, X.; Shim, J.; Lopes, P.E.M.; Mittal, J.; Feig, M.; MacKerell, A.D., Jr. Optimization of the Additive CHARMM All-Atom Protein Force Field Targeting Improved Sampling of the Backbone ϕ , ψ and Side-Chain χ_1 and χ_2 Dihedral Angles. *J. Chem. Theory Comput.* **2012**, *8*, 3257–3273. [[CrossRef](#)] [[PubMed](#)]

49. Jorgensen, W.L.; Chandrasekhar, J.; Madura, J.D.; Impey, R.W.; Klein, M.L. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **1983**, *79*, 926–935. [[CrossRef](#)]
50. Darden, T.; York, D.; Pedersen, L. Particle mesh Ewald: An $N \cdot \log(N)$ method for Ewald sums in large systems. *J. Chem. Phys.* **1993**, *98*, 10089–10092. [[CrossRef](#)]
51. Ryckaert, J.-P.; Ciccotti, G.; Berendsen, H.J.C. Numerical integration of the cartesian equations of motion of a system with constraints: Molecular dynamics of n-alkanes. *J. Comput. Phys.* **1977**, *23*, 327–341. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

4.3 Energies d'interaction électrostatique ELMAM pour l'étude des complexes protéine-ligand : complexe Glutathion Transférase - Glutathion

Au début de ma thèse, nous avons résolu la structure atomique de l'enzyme SynGSTC1, appartenant à la famille des glutathion transférases (GST) qui est beaucoup étudiée au laboratoire CRM², en complexe avec l'un de ses substrats, à savoir le pseudo-tripeptide γ Glu-Cys-Gly appelé glutathion [Mocchetti *et al.*, 2022]. En plus des études structurales, biochimiques et de dynamique moléculaire, nous avons caractérisé la fixation du glutathion dans la poche de SynGSTC1 par le calcul des énergies d'interaction électrostatique ELMAM. Cette analyse illustre une utilisation possible des potentiels d'interaction ELMAM décrits au chapitre 3 dans les études de biologie structurale.

4.3.1 Etude structurale de l'enzyme SynGSTC1

Les GST constituent une superfamille d'enzymes très répandues qui sont notamment impliquées dans des processus de détoxification [Hayes *et al.*, 2005]. Un des principaux rôles des GST est de fournir l'environnement électrostatique local nécessaire à l'activation du glutathion, permettant à ce dernier de se lier à des composés toxiques présentant un site électrophile. En particulier, dans les GST à sérine, le résidu sérine du site actif est supposé abaisser le pKa local du résidu Cys du glutathion pour favoriser sa forme active thiolate. L'enzyme SynGSTC1, de l'organisme *Synechocystis* sp. PCC 6803, appartient aux GST de classe Chi (GSTC) qui présentent un motif SRAS très conservé dans leur site actif. Cette enzyme participe notamment à la détoxification du méthylglyoxal [Kammerscheit *et al.*, 2020].

L'étude structurale de SynGSTC1 [Mocchetti *et al.*, 2022] a révélé que celle-ci adopte la conformation dimérique canonique des GST avec un site actif très ouvert en raison de sa séquence plus courte d'environ 30 résidus que la plupart des autres GST. De manière inattendue, le résidu Ser10 du site actif ne semble pas interagir avec le glutathion, contrairement à ce qui est généralement observé dans les GST à sérine, laissant le groupement thiol exposé au solvant. La mutation S10A n'a d'ailleurs que légèrement affecté l'activité enzymatique. Les simulations de dynamique moléculaire ont suggéré que la Ser10 avait plutôt un rôle dans la stabilisation de la structure atomique que dans la fixation du glutathion. Aucun autre résidu de SynGSTC1 ne semble interagir directement avec le groupement thiol du glutathion. En revanche, de nombreuses interactions entre l'enzyme et le reste du glutathion ont été identifiées (voir Figure 3 de l'article [Mocchetti *et al.*, 2022]).

4.3.2 Caractérisation de l'interaction protéine-ligand par les énergies électrostatiques ELMAM

Pour caractériser les principales interactions entre le glutathion et les résidus du site actif de SynGSTC1, nous avons calculé les énergies d'interaction électrostatique correspondantes sur la base de la densité électronique construite par transfert des paramètres multipolaires de la librairie ELMAM2 [Domagała *et al.*, 2012]. Comme décrit dans le chapitre 3, deux potentiels

d'interaction ELMAM de nature électrostatique sont définis : le terme électrostatique permanent $U_{\text{eslt}}^{\text{ELMAM}}$ (nommé $E_{\text{perm}}^{\text{elec}}$ dans [Mocchetti *et al.*, 2022]) et le terme d'induction dipolaire $U_{\text{ind-dip}}^{\text{ELMAM}}$ (nommé $E_{\text{pol}}^{\text{elec}}$ dans [Mocchetti *et al.*, 2022]). Ces contributions électrostatiques ont été évaluées entre le ligand et onze résidus du site actif dont huit appartenant au même monomère que le glutathion (Ser10, Arg11, Leu33, His38, Lys51, Val52, Glu64, Ser65 et Asn97) et trois à l'autre monomère (Ser98, Thr99 et Arg116). La contribution $U_{\text{elst}}^{\text{ELMAM}}$ a été interprétée comme un terme de reconnaissance moléculaire pouvant être favorable (négatif) ou défavorable (positif) tandis que la contribution $U_{\text{ind-dip}}^{\text{ELMAM}}$ est un terme d'adaptation qui est toujours favorable.

A partir des valeurs de ces contributions, nous avons identifié les trois interactions électrostatiques les plus fortes qui correspondent aux ponts salins entre les résidus chargés positivement du site actif (Arg11, Lys51 et Arg116) et les charges négatives des extrémités zwitterion du fragment γ -Glu et carboxylate du fragment Gly du glutathion. De plus, la contribution fortement favorable du résidu non-chargé Ser65 a été mise en avant par cette analyse, proche de celles des résidus Arg11 et Arg116 lorsque les effets de polarisation sont inclus. Par ailleurs, le résidu chargé négativement Glu64 présente une énergie $U_{\text{elst}}^{\text{ELMAM}}$ défavorable mais la forte contribution de polarisation $U_{\text{ind-dip}}^{\text{ELMAM}}$ induite par la présence du glutathion mène à une contribution électrostatique totale favorable. Grâce à l'analyse des énergies d'interaction, nous avons montré que, malgré la présence de la charge négative de ce résidu Glu64, le site de fixation du groupement zwitterion du fragment γ -Glu est globalement électrophile. Cette observation est cohérente avec la présence d'anions chlorure, acétate ou formate dans les structures cristallographiques de GST sans le glutathion [Sylvestre-Gonon *et al.*, 2022]. Le fragment Cys du glutathion est quant à lui fixé par deux liaisons hydrogène entre chaînes principales avec le résidu Val52 mais aucune interaction ne semble impliquer son groupement thiol. L'énergie d'interaction avec la Ser10 du site actif est quasiment nulle dans SynGSTC1 contrairement à ce que nous avons obtenu dans d'autres structures de GST à sérine et à tyrosine dans lesquelles le résidu du site actif forme une liaison hydrogène avec le thiol. Finalement, nous n'avons pas pu déterminer un unique résidu catalytique responsable de l'activation du groupement thiol, qui semble résulter d'effets électrostatiques complexes entre plusieurs résidus. La description en termes d'énergies d'interaction électrostatique a néanmoins permis de quantifier l'importance de chacun des résidus, dont notamment la Ser65 qui est fortement impliquée dans la formation du site de fixation électrophile du fragment γ -Glu.

Ce travail a été réalisé avant la mise au point des descripteurs issus de la topologie du potentiel électrostatique présentés dans le chapitre 2. Ces méthodologies pourraient apporter de nouvelles informations caractérisant les influences électrostatiques du site actif de l'enzyme sur le groupement thiol du glutathion.

4.3.3 Article publié décrivant la structure de SynGSTC1

Article

Biochemical and Structural Characterization of Chi-Class Glutathione Transferases: A Snapshot on the Glutathione Transferase Encoded by *sll0067* Gene in the Cyanobacterium *Synechocystis* sp. Strain PCC 6803

Eva Mocchetti ^{1,†}, Laura Morette ^{2,†}, Guillermo Mulliert ¹, Sandrine Mathiot ¹, Benoît Guillot ¹, François Dehez ³, Franck Chauvat ⁴ , Corinne Cassier-Chauvat ⁴ , Céline Brochier-Armanet ⁵, Claude Didierjean ^{1,*} , and Arnaud Hecker ^{2,*} 

¹ Université de Lorraine, CNRS, CRM2, F-54000 Nancy, France

² Université de Lorraine, INRAE, IAM, F-54000 Nancy, France

³ Université de Lorraine, CNRS, LPCT, F-54000 Nancy, France

⁴ Université Paris-Saclay, CEA, CNRS, Institute for Integrative Biology of the Cell (I2BC), F-91190 Gif-sur-Yvette, France

⁵ Université de Lyon 1, CNRS, LBBE, F-69622 Villeurbanne, France

* Correspondence: claudedidierjean@univ-lorraine.fr (C.D.); arnaud.hecker@univ-lorraine.fr (A.H.)

† These authors contributed equally to this work.



Citation: Mocchetti, E.; Morette, L.; Mulliert, G.; Mathiot, S.; Guillot, B.; Dehez, F.; Chauvat, F.; Cassier-Chauvat, C.; Brochier-Armanet, C.; Didierjean, C.; et al. Biochemical and Structural Characterization of Chi-Class Glutathione Transferases: A Snapshot on the Glutathione Transferase Encoded by *sll0067* Gene in the Cyanobacterium *Synechocystis* sp. Strain PCC 6803. *Biomolecules* **2022**, *12*, 1466. <https://doi.org/10.3390/biom12101466>

Academic Editor: Bengt Mannervik

Received: 4 August 2022

Accepted: 5 October 2022

Published: 13 October 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Abstract: Glutathione transferases (GSTs) constitute a widespread superfamily of enzymes notably involved in detoxification processes and/or in specialized metabolism. In the cyanobacterium *Synechocystis* sp. PCC 6803, SynGSTC1, a chi-class GST (GSTC), is thought to participate in the detoxification process of methylglyoxal, a toxic by-product of cellular metabolism. A comparative genomic analysis showed that GSTCs were present in all orders of cyanobacteria with the exception of the basal order Gloeobacterales. These enzymes were also detected in some marine and freshwater noncyanobacterial bacteria, probably as a result of horizontal gene transfer events. GSTCs were shorter of about 30 residues compared to most cytosolic GSTs and had a well-conserved SRAS motif in the active site (¹⁰SRAS¹³ in SynGSTC1). The crystal structure of SynGSTC1 in complex with glutathione adopted the canonical GST fold with a very open active site because the $\alpha 4$ and $\alpha 5$ helices were exceptionally short. A transferred multipolar electron-density analysis allowed a fine description of the solved structure. Unexpectedly, Ser10 did not have an electrostatic influence on glutathione as usually observed in serinyl-GSTs. The S10A variant was only slightly less efficient than the wild-type and molecular dynamics simulations suggested that S10 was a stabilizer of the protein backbone rather than an anchor site for glutathione.

Keywords: glutathione transferase; glutathione; cyanobacteria; *Synechocystis* sp. PCC 6803; crystallography; biochemistry; phylogeny

1. Introduction

Glutathione transferases (GSTs) constitute a widespread superfamily of enzymes playing crucial roles in the cell notably in detoxification processes and in specialized secondary metabolism by catalyzing three major kinds of reactions. These include catalytic reactions where glutathione (GSH) is consumed (GSH-conjugation), reactions where GSH is not consumed (isomerization and dehalogenation) and reactions where GSH is oxidized (thiol-transferase and reduction activities) [1]. At the structural level, canonical GSTs are mainly dimeric proteins, and each subunit adopts a conserved fold composed of an N-terminal thioredoxin (TRX) domain linked to an all α -helical C-terminal domain. The active site of the enzyme is located in a cleft at the interface between both domains and contains a GSH-binding site (G site) and a hydrophobic-substrate binding site (H site). Depending

on their primary sequence conservation, GSTs were classified into classes designated by a Greek letter. GSTs with a sequence identity greater than 40% belong to the same class, whereas proteins of different classes share less than 25% sequence identity [2]. GSTs were further distinguished into four catalytic types, tyrosine (TyrGSTs), serine (SerGSTs), cysteine (CysGSTs) and atypical (AtyGSTs), depending on an assumed important residue for catalysis [3]. The tyrosine, serine and cysteine residues have a conserved position in the structures. The tyrosine residue is located at the C-terminal end of the first strand ($\beta 1$). The serine and cysteine residues have the same position at the N-terminal end of the first helix ($\alpha 1$). AtyGSTs do not have a specific residue at a conserved position. In CysGSTs, the cysteine residue has a reactant role according to the enzyme mechanism ontology [4], since it forms a covalent bond with the substrate during the catalytic act. The important residue of the others (TyrGSTs, SerGSTs and AtyGSTs) has a spectator role in enhancing the nucleophilicity of the glutathione thiolate group, and the mutation of this residue does not abolish the activity. Through noncatalytic properties, so-called ligandins, hitherto underestimated compared to the other documented roles of GSTs, many GSTs also participate in the binding and transport of small heterocyclic ligands [5,6].

GSTs have been extensively investigated in animals and plants because of their great relevance to human health and agriculture [7–9]. In contrast, studies in bacteria remain scarce, especially in the cyanobacterial phylum that encompasses oxygenic photosynthetic prokaryotes considered to be the ancestors of chloroplasts. It has been speculated that cyanobacteria may be the first organisms to harbored GSTs [10]. Greek letters have been used for eight classes of GSTs from bacteria: beta, chi, eta, nu, rho, theta, xi and zeta [11,12]. Other classes exist that often have specific functions such as LigE, LigF and LigG involved in lignin degradation in soil bacteria [13,14].

The chi class is thought to be specific of cyanobacteria and three isoforms (TeGSTC1 from *Thermosynechococcus elongatus* BP-1, SeGSTC1 from *Synechococcus elongatus* PCC 6301 and SynGSTC1 from *Synechocystis* sp. PCC 6803) have been characterized biochemically [10,15,16]. A preliminary crystallographic study has been reported for TeGSTC1 and SeGSTC1 [17]. All the three isoforms (TeGSTC1, SeGSTC1 and SynGSTC1) exhibit similar activities. They efficiently catalyze the addition of GSH on various isothiocyanates and show moderate activities toward other classical substrates such as chlorodinitrobenzene [10,15,16]. Interestingly, we recently showed that SynGSTC1 is involved in the detoxification of methylglyoxal, a toxic by-product of the cellular metabolism [15]. Despite a shorter sequence length compared to other GSTs, homology modelling combined with secondary structure prediction suggested that chi GSTs (GSTCs) adopt the fold of canonical GSTs. Their amino acid sequences show two motifs usually found in GSTs. The motif I, which contains an invariant cis-proline residue as well as a $\beta\beta\alpha$ structure essential for the stabilization of the γ -glutamyl moiety of GSH, is the most conserved region in all of the GSTs [18]. Motif II, in turn, contains a very well conserved aspartic acid important for fold stability [19]. Recently, a conserved tyrosine residue located at the fifth position of the N-terminus of GSTCs has been proposed as the catalytic residue [20]. To better characterize the chi class of GSTs, it was necessary to obtain an experimental three-dimensional model. Therefore, we determined the first crystal structure of a chi-class GST (SynGSTC1), performed a robust phylogenetic study and completed the biochemical data by testing new substrates and modulating the active site residues by site-directed mutagenesis.

2. Materials and Methods

2.1. Cloning, Mutagenesis, Expression and Purification

SynGSTC1 (SI10067) encoding sequence was amplified by PCR from *Synechocystis* sp. PCC 6803 genomic DNA as template using specific forward and reverse primers containing *Nde*I and *Xho*I restriction sites, respectively (Table S1). The amplified sequence was subsequently digested and cloned in *E. coli* expression vector pET-26b between *Nde*I and *Xho*I restriction sites allowing the fusion of a His-tag at the C-terminal part of SynGSTC1 as previously described [15]. Various catalytic mutants (S10T, S10A, S10C and R11A) were generated

by site-directed mutagenesis using the QuikChange site-directed mutagenesis kit (Agilent Technologies) and specific mutagenic primers listed in Table S1. The sequences have been confirmed by DNA sequencing.

The expression of recombinant SynGSTC1 and variants were performed at 37 °C using *E. coli* Rosetta2 (DE3) pLysS expression strain (Novagen) transformed with appropriate plasmid in LB medium supplemented with kanamycin (50 µg/mL) and chloramphenicol (34 µg/mL). When the cell culture reached an OD_{600 nm} of 0.7–0.8, the expression of the SynGSTC1 (or S10T or S10A or S10C or R11A) recombinant protein was induced with 0.1 mM isopropyl β-D-1-thio-galactopyranoside (IPTG) for 4 h at 37 °C. Cells were then harvested by centrifugation, resuspended in a 30 mM Tris-HCl buffer (pH 8.0) supplemented with 200 mM NaCl (lysis buffer) and stored at −20 °C until use. After the lysis of the cells by sonication, the resulting cell extract was centrifuged at 40,000× g for 20 min at 4 °C to remove cellular debris and aggregated proteins. After the addition of 10 mM imidazole, SynGSTC1 was purified from the soluble extract by gravity-flow chromatography on a nickel nitrilotriacetate (Ni-NTA) agarose resin (Qiagen, Hilden Germany). After a washing step with lysis buffer containing 20 mM imidazole, recombinant SynGSTC1 was eluted using lysis buffer supplemented with 250 mM imidazole. The fractions of interest were pooled, concentrated by ultrafiltration, subjected to a size exclusion chromatography using a Superdex™200 16/600 column connected to an ÄKTA-Purifier™ device (Cytiva) and eluted with lysis buffer. The purified recombinant protein was concentrated and finally stored at −20 °C. The concentration of SynGSTC1 recombinant protein was determined at 280 nm using a theoretical molar absorption coefficient of 28,420 M^{−1}·cm^{−1}.

2.2. Crystallization, X-ray Data Collection, Processing and Refinement

A first screening of 288 crystallization conditions was carried out at the CRM2 crystallogenes platform (University of Lorraine) with an Oryx 8 crystallogenes robot (Douglas Instruments Ltd, Hungerford, UK). Crystals were optimized manually at 4 °C by the microbatch-under-oil method. Solutions of SynGSTC1 and the variants contained 30–40 mg·mL^{−1} protein in 30 mM Tris buffer (pH 8.0) supplemented with 200 mM NaCl, 1 mM Tris(2-carboxyethyl)phosphine (TCEP) and 10 mM glutathione. SynGSTC1 was crystallized by mixing 1 µL of protein with 1 µL of solution consisting of 16% (*w/v*) PEG 8000, 40 mM potassium phosphate monobasic and 20% (*w/v*) glycerol (condition no. 32, Wizard™ Classic Crystallization Screen III, Rigaku, Tokyo, Japan).

Preliminary X-ray diffraction experiments were carried out in-house on an Agilent SuperNova diffractometer (Rigaku Oxford Diffraction) equipped with a CCD detector. Data collections were carried out at the ESRF, on beamline FIP BM07 (ESRF, Grenoble, France) and (PX1 and PX2, SOLEIL, Gif-Sur-Yvette, France). Data sets were indexed and integrated with XDS [21], and scaled and merged with Aimless [22] from the CCP4 suite [23]. The structure of SynGSTC1 was solved by molecular replacement using *MoRDa* [24] with the coordinates of a GST from *Rhodobacter sphaeroides* (PDB entry 3LSZ) as the search model. Structures of SynGSTC1 and its variants were refined with *BUSTER* [25] and manually improved with *Coot* [26]. The validation of all structures was performed with the PDB validation service (<http://validate.wwpdb.org>, accessed on 30 September 2022). The coordinates and structure factors have been deposited in the Protein Data Bank (PDB entries 8AI8, 8AI9, 8AIB). Crystal data, diffraction and refinement statistics are shown in Table 1.

Table 1. Statistics of X-ray diffraction data collection and model refinement.

	Wild-Type	S10T	R11A
Data Collection			
Diffraction source	ESRF-BM07	ESRF-BM07	ESRF-BM07
Detector	Pilatus 6M	Pilatus 6M	Pilatus 6M
Wavelength (Å)	0.97951	0.97951	0.97951

Table 1. Cont.

	Wild-Type	S10T	R11A
Space Group	$P4_32_12$	$P4_32_12$	$P4_32_12$
Unit-cell $a; c$ (Å)	92.5; 193.6	92.9; 193.6	92.2; 193.1
Resolution Range (Å)	48.4 1.7 (1.73 1.70)	48.4 1.7 (1.73 1.70)	46.1 2.2 (2.27 2.20)
Tot. no. of meas. int.	1,200,217 (41158)	1,244,521 (62,785)	503,802 (22,268)
Unique reflections	92,986 (4529)	93,931 (4590)	39,298 (2006)
Average redundancy	13 (9)	13.2 (14)	13 (11)
Mean I/σ (I)	24.8 (1.8)	17.0 (2.0)	18.4 (2.4)
Completeness (%)	100.0 (99.6)	100.0 (100.0)	91.3 (55.5)
R_{merge}	0.056 (1.097)	0.084 (1.52)	0.097 (1.039)
R_{meas}	0.061 (1.168)	0.087 (1.59)	0.100 (1.142)
$CC_{1/2}$	1.00 (0.83)	1.00 (0.84)	1.00 (0.85)
Wilson B -factor ($\text{\AA}^2$)	29.6	28.6	41.5
Refinement			
Resolution Range (Å)	24.8 1.7	24.2 1.7	31.3 2.2
No. of reflections	92839	93783	39248
$R_{\text{work}}/R_{\text{free}}$	0.204/0.221	0.206/0.221	0.215/0.242
Corr F_o - F_c / F_o - $F_{c\text{free}}$	0.938/0.936	0.940/0.939	0.907/0.888
Total number of atoms	3469	3500	3253
Average B -factor ($\text{\AA}^2$)	34.0	32.5	44.0
Model quality			
RMSZ Bond lengths	0.41	0.42	0.42
RMSZ Bond angles	0.54	0.56	0.56
Ramachandran fav. (%)	98	98	98
Ramachandran all. (%)	2	2	2
Rotamer outliers (%)	0	0	1
Clashscore	1	1	1
PDB entry	8AI8	8AI9	8AIB

$R_{\text{merge}} = \sum_{hkl} \sum_i |I_i(hkl) - \bar{I}(hkl)| / \sum_{hkl} \sum_i I_i(hkl)$. $R_{\text{meas}} = \sum_{hkl} \{N(hkl) / [N(hkl) - 1]\}^{1/2} \sum_i |I_i(hkl) - \bar{I}(hkl)| / \sum_{hkl} \sum_i I_i(hkl)$. $CC_{1/2}$ is the correlation coefficient of the mean intensities between two random half-sets of data. $R_{\text{work}} = \sum_{hkl} ||F_{\text{obs}}| - |F_{\text{calc}}|| / \sum_{hkl} |F_{\text{obs}}|$. In total, 5% of reflections were selected for R_{free} calculation. RMSZ: root mean square Z-score. The MolProbity clashscore is the number of serious clashes per 1000 atoms. Values in parentheses are for highest resolution shell.

2.3. Structure Analysis Based on Electron Density Distribution

To calculate the electrostatic interaction energies between residues of SynGSTC1 active site and the glutathione ligand, the electron charge density of the complex based on the Hansen and Coppens multipolar model [27] was determined (method detailed in Supplementary Materials). The electron density parameters for the SynGSTC1-GSH complex were transferred from the ELMAM2 database2, which provides parameters averaged over experimental peptide electron densities from ultra-high resolution X-ray scattering data [28]. In addition, polarization effects due to the environment were estimated in the transferred electron density using the procedure described recently by Leduc et al. [29] and implemented in *MoProViewer* software (version 0.1.1302) [30]. The electrostatic interaction energy

(E_{tot}^{elec}) between the glutathione ligand and the SynGSTC1 active site residues was computed using *Charger*, which is a fast and analytical electrostatic energy calculation tool [31] also implemented in *MoProViewer*. E_{tot}^{elec} includes two terms, the electrostatic interaction permanent energy E_{perm}^{elec} and the polarization contribution E_{pol}^{elec} (hence $E_{tot}^{elec} = E_{perm}^{elec} + E_{pol}^{elec}$). The *MoProViewer* database transfer tool enables an automatic parameter transfer on the structure with appropriate formal charge assignment (+1e for arginine and lysine, −1e for aspartate and glutamate, 0 for others). The His38 and His61 of SynGSTC1 were protonated on the Nε atom and the formal charge of glutathione was set to −1e. The procedure is detailed in the Supplementary Materials.

2.4. Molecular Dynamics Simulation

The Molecular Dynamics simulations presented in this study were based on the crystallographic structure of SynGSTC1 in complex with GSH. This system was immersed in a cubic simulation cell of length equal to 73 Å filled by a solvent of 9881 water molecules with 150 mM NaCl. The simulations were performed using *NAMD 3.0* [32] with the CHARMM36 [33] force field for proteins and the TIP3P water model [34]. The parameters for the GSH ligand were generated by the CHARMM general force field (CGenFF) [35]. Long-range electrostatic forces were evaluated using the particle mesh Ewald algorithm with a grid spacing of 1.0 Å. A smoothed 12.0 Å spherical cutoff was applied to truncate the short-range van der Waals and electrostatic interactions. The temperature was maintained at 300 K thanks to the Langevin thermostat and the pressure at 1 atm thanks to the Langevin piston method. Covalent bonds involving hydrogen atoms were restrained to their equilibrium length by the Rattle algorithm [36] and the water molecules were constrained to their equilibrium geometry using the Settle algorithm [37]. In addition, a mass-repartitioning scheme was used to integrate the equations of motion with a time step of 4 fs, according to Hopkins et al. [38]. A smooth equilibration, along which the positions of the heavy atoms of the protein were restrained harmonically, was carried out during 8 ns before a non-restrained long equilibration of 100 ns. Then, the SynGSTC1-GSH complex was probed in production runs including a long simulation of 500 ns and five independent shorter simulations of 100 ns. These trajectories were visualized and analyzed using *VMD* [39]. These simulations were aimed at exploring the stability of the interactions between the glutathione and the active site as well as the flexibility of the protein interdomain linker.

2.5. Enzymatic Assays

The GSH-conjugation activity was assayed at 25 °C toward 1-chloro-2,4-dinitrobenzene (CDNB), benzyl-isothiocyanate (BITC), 2-phenethyl-isothiocyanate (PITC) or p-nitrophenyl butyrate (PNP-butyrate). The reactions were performed in 500 µL of 30 mM Tris-HCl (pH 8.0) and 1 mM EDTA for CDNB and PNP-butyrate and 100 mM phosphate buffer (pH 6.5) for ITC derivatives in the presence of various concentrations of CDNB (0–4000 µM), BITC (0–1000 µM), PITC (0–1000 µM) or PNP-butyrate (0–2000 µM) at a fixed saturating GSH concentration. Peroxidase and thiol-transferase activities were assayed at 25 °C toward cumene hydroperoxide (CuOOH) and 2-hydroxyethyl disulfide (HED) in a NADPH-coupled spectrophotometric method by following the absorbance at 340 nm. The reactions were carried out in 500 µL of 30 mM Tris-HCl (pH 8.0) containing 200 µM NADPH, 0.5 unit of yeast glutathione reductase and various concentrations of HED (0–500 µM) or CuOOH (0–3000 µM) at a fixed GSH concentration. The optimum pH of the wild-type enzyme and its variants was determined against PITC using 100 mM sodium citrate, phosphate, or borate buffers at pH ranging from 4.0 to 11.0. GSH-conjugation activity was determined as described above.

For all activity assays, the recombinant protein, used at a concentration (3 µM) within the linear response range of the enzyme, was added after 2 min of preincubation and the variation of absorbance monitored using a Cary 50 spectrophotometer. The activity recorded without enzymes was subtracted and three independent experiments were performed at each substrate concentration. The kinetic parameters, apparent K_m (Michaelis

constant) and $k_{\text{software_cat}}$ (turnover number) were determined by fitting the data to the nonlinear regression Michaelis–Menten model in *GraphPad Prism* (version 8, GraphPad Software, Inc., San Diego, CA, USA). The k_{cat} values were expressed as μmol of substrate oxidized per second per μmol of enzyme (i.e., the turnover number in s^{-1}) using specific molar absorption coefficients of $9600 \text{ M}^{-1}\cdot\text{cm}^{-1}$ at 340 nm for CDNB, $9250 \text{ M}^{-1}\cdot\text{cm}^{-1}$ at 274 nm for BITC, $8890 \text{ M}^{-1}\cdot\text{cm}^{-1}$ at 274 nm for PITC, $17700 \text{ M}^{-1}\cdot\text{cm}^{-1}$ at 412 nm for PNP-butyrate and $6220 \text{ M}^{-1}\cdot\text{cm}^{-1}$ at 340 nm for NADPH.

2.6. Phylogenetic Analysis

In total, 222 proteomes of the Cyanobacteria/Melainabacteria group were retrieved from the RefSeq database of the NCBI. These corresponded to 208 proteomes of Cyanobacteria labelled as RefSeq “reference proteomes” or from type strains, and 14 proteomes from noncyanobacterial lineages (i.e., Margulisbacteria, Melainabacteria, Gastranaerophilales) classified in the Cyanobacteria/Melainabacteria group (Table S2). The sequences of the 53 ribosomal protein families (rprots) were retrieved from the 222 proteomes using the riboDB database [40] (Table S3). The corresponding protein sequences were aligned using *MAFFT v7.453* with the accurate option L-INS-I [41]. The resulting multiple alignments were trimmed with *BMGE v1.2* using the BLOSUM30 substitution matrix [42]. The multiple alignments of the 52 rprots present in more than 30% of the 222 analyzed proteomes were combined to build a large supermatrix (222 sequences, 6430 amino acid positions) and used to build a phylogeny using the maximum likelihood method. The tree was inferred with *IQ-TREE* (multicore version 2.2.0 COVID-edition, June 2022) with the LG + C20 + F + R4 evolutionary model [43]. The branch robustness of the inferred tree was computed with the ultrafast bootstrap procedure implemented in *IQ-TREE* (1000 replicates). The resulting tree was rooted using the 14 noncyanobacterial sequences.

The 222 studied proteomes were queried with *BLASTP* using the GSTC1 sequence from the *Synechocystis* sp. PCC 6803 strain (RefSeq protein Id WP_010873500.1, locus tag SGL_RS13850) as seed. The 924 GST sequences displaying an E-value lower than 10^{-3} were retrieved and aligned using *MAFFT* with the auto option. A total of 54 partial sequences were discarded from the analysis. A survey of the nr database at the NCBI identified 11 sequences of GSTC in noncyanobacterial bacteria. These 11 sequences were added to the cyanobacterial GSTC sequences. The 881 GSTC sequences were realigned with *MAFFT* with the L-INS-I option and trimmed using *BMGE* with the BLOSUM30 substitution matrix. The 104 kept amino acid positions were used to infer a phylogeny using *FastTree v2* [44] with 20 rate categories of sites, the gamma optimization option, and the Le and Gascuel model [45]. The branch robustness of the inferred tree was estimated using the Shimodaira Hasegawa test (resampling the site likelihoods 1000 times). Finally, a phylogenetic analysis of the 147 cyanobacterial GSTC sequences was performed using *FastTree* and the same parameters (147 sequences, 110 amino acid positions).

The trees were drawn using *iToL v6.5.8* [46].

3. Results and Discussion

3.1. Crystal Structure of SynGSTC1

In this study, the crystal structure of the glutathione transferase chi1 from *Synechocystis* sp. PCC 6803 (SynGSTC1) in complex with GSH is presented. We also solved the structures of two variants (S10T and R11A variants in complex with GSH) which did not show significant differences from the wild-type. The protein samples were cocrystallized with an excess of GSH (10:1) in the presence of TCEP to avoid oxidation of the GSH thiol group into sulfenic acid. SynGSTC1 crystallized in space group $P4_32_12$ with two polypeptide chains in the asymmetric unit. They formed a two-fold dimer that had a globular shape with molecular dimensions of approximately $55 \text{ \AA} \times 55 \text{ \AA} \times 45 \text{ \AA}$ (Figure 1). The dimer buried $1710 \text{ \AA}^2$ of surface area for each monomer and was tightly stabilized by ten hydrogen bonds and six salt bridges (Table S4). At the core, a four-helix bundle consisting of the $\alpha 3$ and $\alpha 4$ helices of the two monomers buried aliphatic residues (L70 and L94 of chains A and B).

This interaction pattern was complemented by a lock-and-key motif where the F49 residue fitted into a low-polar cavity of the adjacent subunit (W92, F95, L117, L121) (Figure 1).

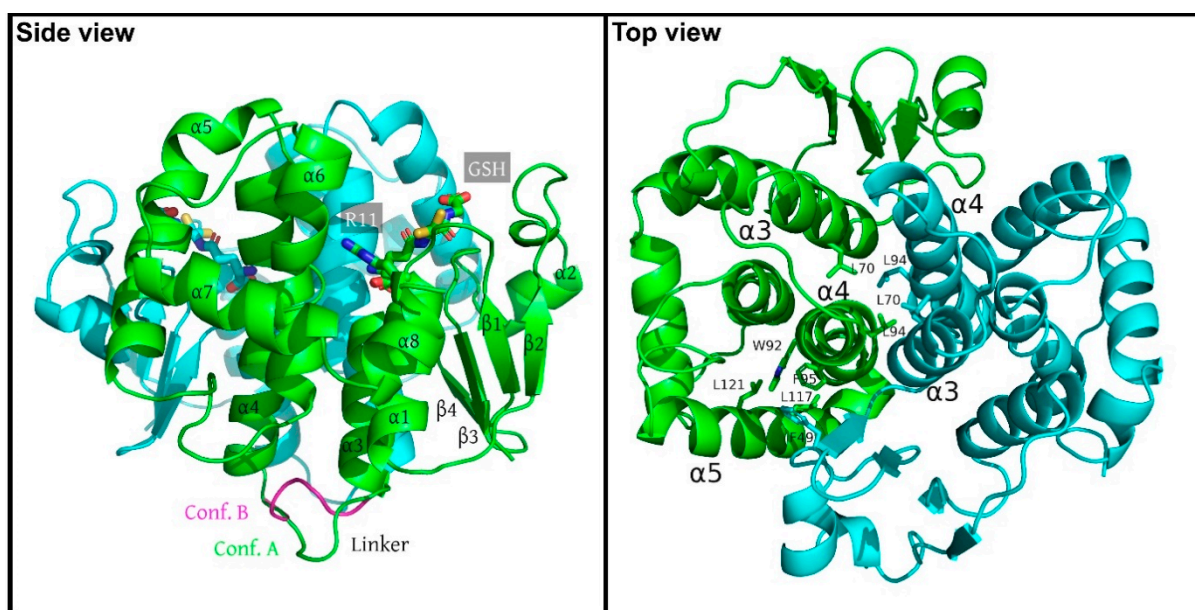


Figure 1. Crystal structure of the SynGSTC1 dimer, left, and rotated 90°, right. The monomers A and B are shown in ribbon mode and colored green and blue, respectively. (**Left**), side view. The secondary structures of the monomer A are labelled. In each monomer, the side chain of residue R11 and the glutathione molecules are labelled and highlighted as sticks. Both conformations of the linker are shown in monomer A. (**Right**), top view. The figure highlights the hydrophobic patches on SynGSTC1 dimer interface. L70 and L94 of both monomers are buried in the center of the dimer. This interaction pattern is complemented by a lock-and-key motif where the F49 residue (blue) fits into a low-polar cavity of the adjacent subunit (W92, F95, L117, L121) (green). The symmetry related lock-and-key motif is not shown for clarity.

Both subunits were very similar structures and could be superimposed within 0.33 Å root-mean-square deviation over 181 α -carbon atoms. The SynGSTC1 protomer adopted the conserved GST fold that was subdivided into two domains for clarity (N-terminal domain $\beta 1\alpha 1\beta 2\alpha 2\beta 3\beta 4\alpha 3$ and C-terminal domain $\alpha 4\alpha 5\alpha 6\alpha 7\alpha 8$). As mentioned in the introduction, the chain length of GSTCs (approximately 180 residues) was significantly shorter by at least 20 residues compared to most canonical GSTs [47]. The $\alpha 4$ – $\alpha 5$ hairpin pattern was significantly shortened (roughly 10 residues) and the angle between these two helices ($\sim 42^\circ$) was twice that usually observed (Figure 2). This “missing” region made the active site of SynGSTC1 very open, with no clear pocket for the hydrophobic substrate (H-site). Both motifs I (47–71) and II (129–147) played their expected structural roles. In motif I, the V52–P53 peptide bond was cis, and V52 formed the typical antiparallel β -sheet-like interaction with the cysteine moiety of GSH. Motif II contained the Ncap sequence $^{137}\text{SVVD}^{140}$ where the side chains of the serine and aspartic acid residues contributed to the stabilization of the $\alpha 6$ helix [19]. The linker ($^{76}\text{ASTIPAD}^{82}$) between the N- and C-terminal domains was peculiar because it had no aliphatic or aromatic residue wedged between these two domains as usually observed [48,49]. The consequence was an interdomain linker without a unique conformation. The quality of the electron density allowed the building of two major conformations (Figures 1 and S1). To investigate this property, we performed molecular dynamics simulations of the SynGSTC1-GSH complex in an aqueous environment. The simulation revealed a protein very stable with the linker as one of the most mobile regions. The time-evolution of the φ and ψ torsion angles of the linker residues revealed transitions between two main conformations during the trajectory

(Figure S2). Interestingly, these two conformations corresponded to those observed in the crystal structure.

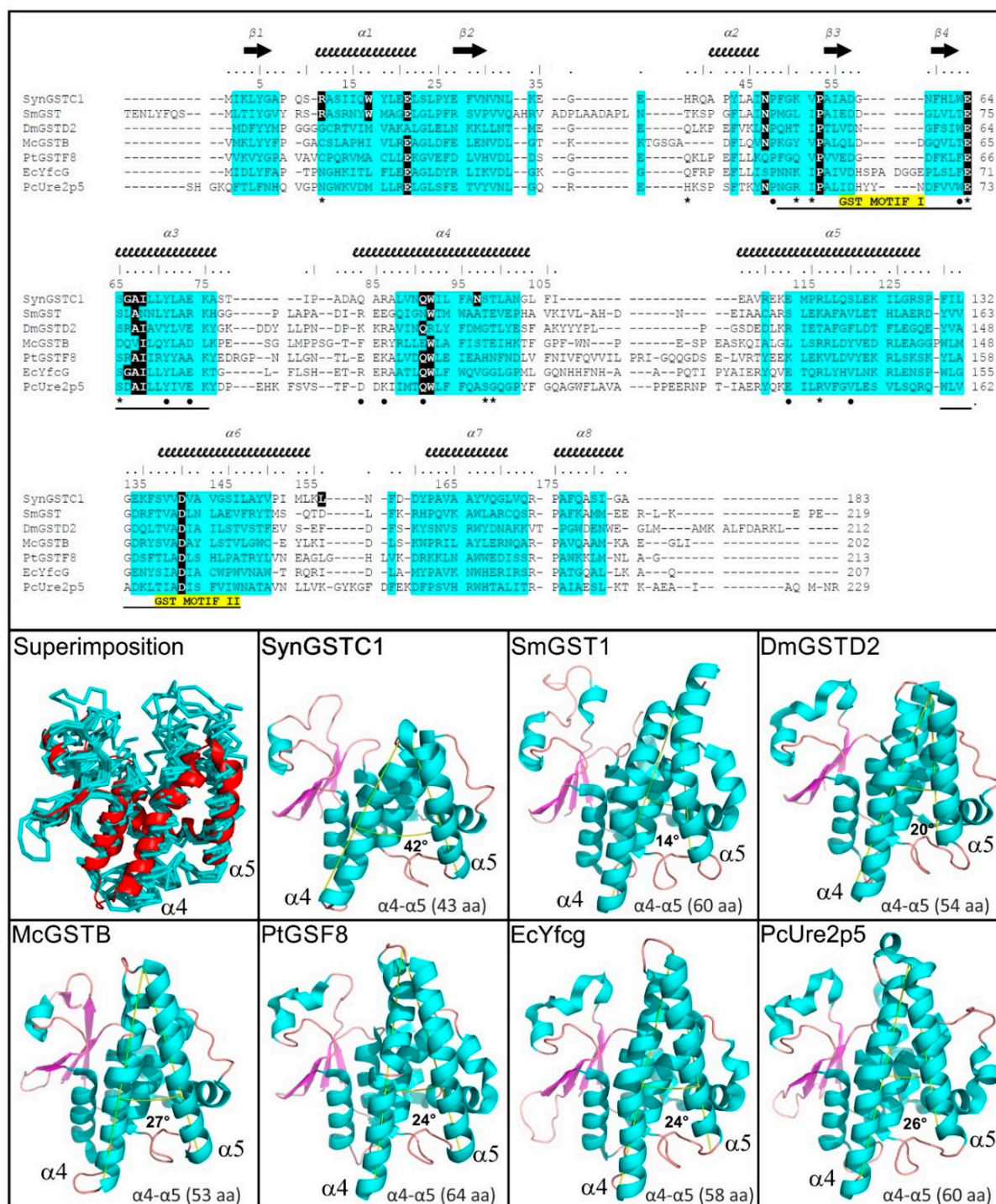


Figure 2. Comparison of SynGSTC1 with structural homologs. The top figure shows a structure-based sequence alignment, and the bottom figures highlight that SynGSTC1 has the shortest $\alpha 4$ – $\alpha 5$ hairpin and the highest angle between $\alpha 4$ and $\alpha 5$ helices. Crystal structure and sequences can be found at the Protein Data Bank (<http://www.rcsb.org>, accessed on 30 September 2022): SynGSTC1, this study, PDB ENTRY 8AI8; SmGST, GST from *Sinorhizobium meliloti* 2011, PDB entry 4NHW; DmGSTD2, GST delta 2

from *Drosophila melanogaster*, PDB entry 5F0G; McGSTB, GST beta from *Methylococcus capsulatus* str. Bath, PDB entry 3UAP; PtGSTF8, GST phi 8 from *Populus trichocarpa*, PDB entry 5F07; EcYfcG, GST nu from *Escherichia coli* K-12, PDB entry 5HFK; PcUre2p5, Ure2p 5 from *Phanerodontia chrysosporium*, PDB entry 4F0C. The characteristics of the top figure are as follows: secondary structures are labelled and shown using arrows (β -strands) and squiggles (helices); common regions, i.e., regions with no gaps and with pairwise residue distances less than 4 Å are highlighted in blue; the invariant residues in the GST chi class are in bold type, coloured white and highlighted in black; residues that participates in dimer stabilization of SynGSTC1 via strong polar interactions are marked with •; residues involved in binding glutathione (G-site) in SynGSTC1 are marked with *. The characteristics of the bottom figures are as follows: the models are shown in the cartoon or ribbon modes; the $\alpha 4$ and $\alpha 5$ helices are labelled; the first figure shows a superimposition of the seven structures where SynGSTC1 is colored red and the others cyan; in the other figures, the estimated angle between $\alpha 4$ and $\alpha 5$ helices is provided as well as the number of amino acids in the $\alpha 4$ – $\alpha 5$ hairpin; the angles were calculated using the *AngleBetweenHelices* script (<https://pymolwiki.org/index.php/AngleBetweenHelices>, accessed on 30 September 2022) implemented in *PyMol Molecular Graphics System* (Version 2.0 Schrödinger, LLC, New York, NY, USA).

3.2. Structural Comparison

A search for the structural homologs using the Dali server (<http://ekhidna2.biocenter.helsinki.fi/dali/>, accessed on 30 September 2022) ranked proteobacterial nu GSTs and fungal GSTs from the Ure2p class at the top of the list [50]. The other hits included proteobacterial beta GSTs, insect delta GSTs, an unclassified proteobacterial GST and plant phi GSTs. To better depict the proximities of these structures, an additional multiple structural alignment was performed using the mTM-align server (<https://yanglab.nankai.edu.cn/mTM-align/>, accessed on 30 September 2022) [51] (Figure 2). The resulting dendrogram based on the pairwise alignment scores (Figure S3) showed a distribution of the proteins into two clades, one of which containing SynGSTC1 and the unclassified proteobacterial GST (GST SMc00097 from *Sinorhizobium meliloti* 2011, PDB entry 4nhw). The latter had therefore the most similar structure to SynGSTC1. SMc00097 had one of the structural attributes of SynGSTC1, namely a SRAS motif at the beginning of the $\alpha 1$ helix in its active site (see below) (Figure 2). The first serine residue adopted the same orientation and did not participate in the stabilization of GSH while the arginine residue did (Figure S4). The closeness between SynGSTC1 and SMc00097 could be explained in a more comprehensive way by a domain-by-domain comparison. Indeed, the overall structures (i.e., both the N-ter and C-ter domains) of SynGSTC1 and SMc00097 overlapped well (Table S5). The proximity of SynGSTC1 with other hits (nu, beta, delta and phi GSTs) was rather due to the good overlap of N-terminal domains.

3.3. Active Site Structure and Its Analysis Using Transferred Multipolar Electron-Density

The active site contained GSH tightly bound to the G-site by numerous polar interactions (respectively, six, two and three for the γ -Glu, Cys and Gly moieties) (Figure 3). The GSH Cys moiety adopted two rotamers (**m**, $\chi_1 = -52^\circ$ and **t**, $\chi_1 = 172^\circ$) exposing the GSH thiol group towards the solvent (Figure 1). The three regular rotamers (**p**, **m**, and **t**) of the glutathione cysteine moiety were accessible during the molecular dynamics simulations with the frequencies of 0.35, 0.42 and 0.16, respectively (Figure S5). The crystal structure did not reveal a strong polar interaction between the sulfur atom of GSH and the enzyme. The smallest distance was 3.8 Å with the amide group of R11. The Y5 residue, recently proposed as a catalytic residue [20], was far too distant to stabilize the GSH-thiolate group during catalysis as the Y5 hydroxyl group was 17 Å away from the GSH sulfur-atom. Based on the sequence analysis of SynGSTC1, we could have thought that S10 played an important role in catalysis. Indeed, this serine residue belongs to the $^{10}\text{SRAS}^{13}$ motif, which is related to the CXXC active-site motif of thioredoxin [52]. The equivalent serine residue in Ser-GSTs (see introduction) is almost always found hydrogen-bonded to the GSH thiol group while this is not the case in SynGSTC1 [53,54]. Indeed, the OG atom invariably retained the same orientation in all structures (wild-type and variants), and was hydrogen bonded to the main chains of A7 and A12. This interaction network remained stable throughout most of

the molecular dynamics simulations showing that S10 was important for the stabilization of the protein backbone. Whatever its conformation, this serine residue never formed a strong interaction with the GSH thiol group during the simulation (Figure S6).

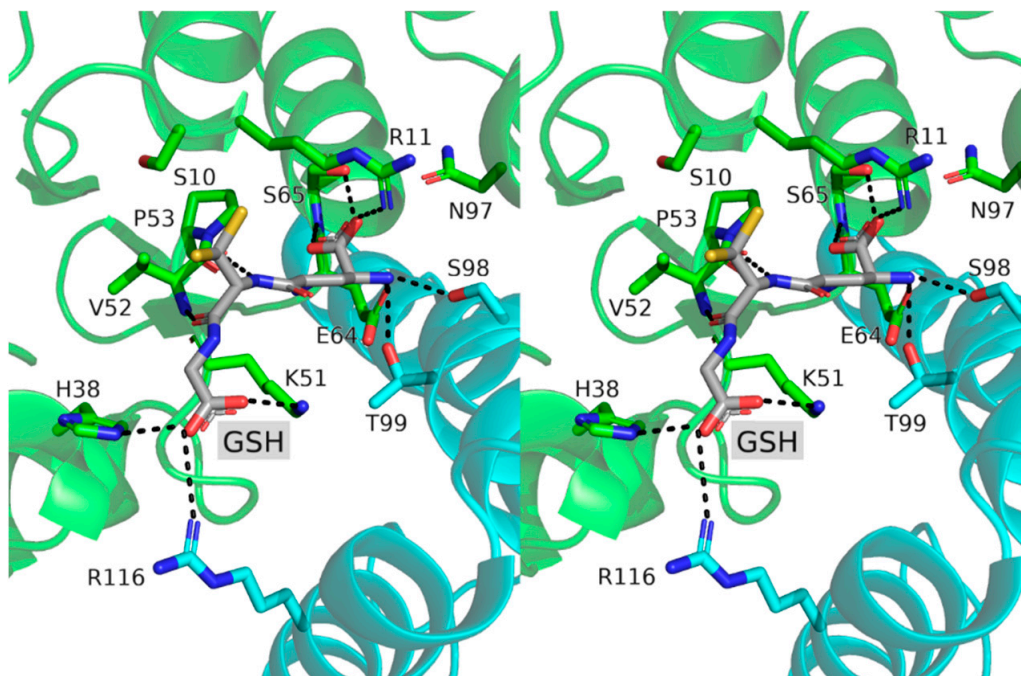


Figure 3. Stereoview of the glutathione binding site of SynGSTC1. The monomers A and B are shown in cartoon mode and colored green and blue, respectively. GSH and residues around it are shown as sticks and labelled. Numbering of residues is according to sequence of SynGSTC1. Strong intermolecular interactions are shown as dashed sticks.

The description of the interactions between a ligand and a protein is most often summarized by the list of residues involved, without quantifying the importance of each. We developed recently a fast and analytical procedure to estimate the electrostatic contribution of each residue to the ligand binding, based on a continuous distribution of electron density of experimental origin [31]. This method implemented in *MoProViewer* [30] was applied on SynGSTC1 in complex with GSH where the contributions of eleven residues were evaluated (distance cutoff of 3.5 Å away from GSH). This included eight residues from one chain (S10, R11, L33, H38, K51, V52, E64, S65 and N97) and three from the other (S98, T99 and R116). *MoproSuite* calculates electrostatic interaction energies $E_{\text{tot}}^{\text{elec}}$ that are divided into two contributions: a permanent electrostatic interaction term, $E_{\text{perm}}^{\text{elec}}$, and a polarization one, $E_{\text{pol}}^{\text{elec}}$, which can be interpreted as a molecular recognition term and an adaptation term, respectively (Figure 4, Table S6). By definition, the polarization term is negative and makes the total interaction energy more favorable for all the active site residues and especially for charged residues [29]. We performed the calculations for the two GSH thiol orientations observed in the crystal structure. The orientation of the thiol group did not affect notably the GSH binding, from an electrostatic and dipolar-induction point of view (Table S6). Thus, the following analysis did not depend on the GSH conformation.

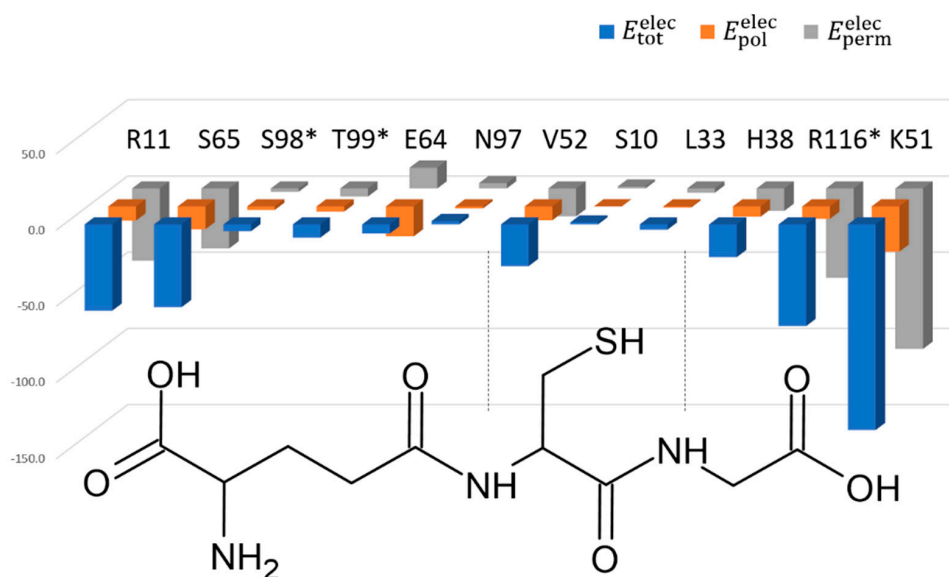


Figure 4. Permanent, polarization and total electrostatic interaction energies. The permanent $E_{\text{perm}}^{\text{elec}}$, polarization $E_{\text{pol}}^{\text{elec}}$ and total $E_{\text{tot}}^{\text{elec}}$ electrostatic interaction energies between the glutathione ligand and twelve residues of the SynGSTC1 active site are presented in kcal/mol. $E_{\text{perm}}^{\text{elec}}$ is computed using the electron density model transferred on the glutathione and the protein atoms, whereas $E_{\text{tot}}^{\text{elec}}$ is obtained after the electron density polarization procedure. Finally, $E_{\text{pol}}^{\text{elec}}$ is computed using $E_{\text{pol}}^{\text{elec}} = E_{\text{tot}}^{\text{elec}} - E_{\text{perm}}^{\text{elec}}$, and represents the polarization contribution to the total electrostatic interaction energy. The reported energy values have been averaged over the two conformations of the glutathione (A and B) observed in the crystal structure and over the two monomers. The GSH formula has been added to highlight the proximity of the residues to the GSH moieties. The residues marked with a star (*) in the figure are not from the same monomer as glutathione. The numerical values of these energies and the associated standard deviations are available in the Supplementary Materials.

The permanent interaction energy $E_{\text{perm}}^{\text{elec}}$ pictures the electrostatic complementarity between the GSH chemical groups and the residues lining the binding site. GSH was assumed to bear three charges: a zwitterionic γ -glutamic acid moiety and a terminal glycine carboxylate group. The SynGSTC1 residues with the largest contributions were R11, K51, R116, which formed salt bridges with the GSH negative charges (Figure 3, Table S6). As an example, the energy values $E_{\text{tot}}^{\text{elec}}$, $E_{\text{perm}}^{\text{elec}}$ and $E_{\text{pol}}^{\text{elec}}$ obtained for R11 were $-56.6 \text{ kcal}\cdot\text{mol}^{-1}$, $-47.3 \text{ kcal}\cdot\text{mol}^{-1}$ and $-9.3 \text{ kcal}\cdot\text{mol}^{-1}$, respectively. S65 showed the most favorable $E_{\text{perm}}^{\text{elec}}$ among the uncharged residues ($E_{\text{perm}}^{\text{elec}} = -39.3 \text{ kcal}\cdot\text{mol}^{-1}$), and its contribution was close to those of R11 and R116 when the dipolar induction is included ($E_{\text{tot}}^{\text{elec}} = -54.3 \text{ kcal}\cdot\text{mol}^{-1}$). This serine residue was double-hydrogen-bonded to the γ -Glu carboxylate group. This interaction pattern, well conserved in GSTs, is ensured either by a serine residue or a threonine residue [55]. The negatively charged E64 residue was an interesting case because it had an unfavorable $E_{\text{perm}}^{\text{elec}}$ ($13.6 \text{ kcal}\cdot\text{mol}^{-1}$) that underwent a significant dipolar induction ($E_{\text{pol}}^{\text{elec}} = -19.7 \text{ kcal}\cdot\text{mol}^{-1}$) to interact with the positively charged N-terminal amine group of GSH ($E_{\text{tot}}^{\text{elec}} = -6.0 \text{ kcal}\cdot\text{mol}^{-1}$, Figure 4, Table S6). The major contributors for the GSH γ -Glu moiety, $E_{\text{perm}}^{\text{elec}}$ speaking, were therefore R11 via its guanidium group and S65 via its amide and hydroxyl groups (Figure 3). This showed that the site where the zwitterionic fragment of GSH was located was an electrophilic site. This property is verified in the crystallographic structures of glutathione-free GSTs because they often contain a negative ion in this site such as chloride, acetate or formate ions [56]. In addition, this electrophilic site was found to be catalytically important as it hosts the γ -Glu carboxylate group which is presumed to decrease the pKa of the GSH thiol group [57]. The glycine part of GSH was surrounded by the two positively charged

K51 and R116 residues, and by the lateral chain of H38 residue. These residues tightly stabilized the GSH C-terminal carboxylate group (Figures 3 and 4). Finally, the GSH Cys part was strongly stabilized by a single residue (V52) via two main-chain–main-chain hydrogen bonds (Figure 3). This twofold contribution was significantly lower compared to that of S65 probably because the V52-GSH interaction did not involve charged groups. The S10 residue, which was assumed to be the catalytic residue interacting with the thiol group, presented an unfavorable electrostatic interaction energy and did not contribute to the GSH stabilization ($E_{\text{tot}}^{\text{elec}} = 1.3 \text{ kcal}\cdot\text{mol}^{-1}$, Table S6). It also showed an almost zero polarization energy so this residue was not affected by the binding of the glutathione. This correlated well with the fact that the crystal structure of SynGSTC1 revealed no intermolecular interaction between S10 and GSH. The side chain of the “main” tyrosine residue of TyrGSTs (Tyrosine type GSTs) was always observed interacting with the thiol group of GSH in the crystal structures. The “main” serine residue of SerGSTs plays the same role in most known structures. We evaluated the electrostatic contribution of residues to GSH binding in a TyrGST (and a SerGST) containing a putative hydrogen bond between the tyrosine (serine) residue and GSH (Table S6). In both cases, the important residue (tyrosine or serine) provided a stabilizing effect on the ligand ($E_{\text{tot}}^{\text{elec}} = -12 \text{ kcal}\cdot\text{mol}^{-1}$ and $E_{\text{tot}}^{\text{elec}} = -7.3 \text{ kcal}\cdot\text{mol}^{-1}$, respectively, Table S6). However, this contribution was never predominant. The main anchor points remained the positively charged residues that stabilized the terminal carboxylate groups of GSH.

3.4. Biochemical Characterization of SynGSTC1 and Variants

We recently detected an activity for SynGSTC1 toward methylglyoxal as substrate and also tested glutathione transferase reactions namely aromatic substitution, and addition using, respectively, 1-chloro-2,4-dinitrobenzene (CDNB) and isothiocyanates (ITCs) as substrates [15]. In addition to these activities, we tested here the ability of SynGSTC1 to conjugate GSH on 4-nitrophenyl butyrate (PNP-butyrate) by transacylation and to reduce hydroperoxide toward cumene hydroperoxide (CuOOH). The measured activities ($k_{\text{cat}}/K_{\text{m}}$), respectively of $49.0 \pm 1.7 \text{ M}^{-1}\cdot\text{s}^{-1}$ for PNP-butyrate and $604.6 \pm 37.7 \text{ M}^{-1}\cdot\text{s}^{-1}$ for CuOOH are similar to the one measured toward CDNB $112.5 \pm 14.2 \text{ M}^{-1}\cdot\text{s}^{-1}$. These activities remain significantly lower than those measured with ITCs ($6.7 \times 10^5 \pm 0.2 \times 10^5 \text{ M}^{-1}\cdot\text{s}^{-1}$ and $5.7 \times 10^5 \pm 0.2 \times 10^5 \text{ M}^{-1}\cdot\text{s}^{-1}$ for PITC and BITC, respectively) due to a higher affinity of the enzyme for PITC and BITC (31.4 ± 3.5 and $82.0 \pm 10.0 \text{ }\mu\text{M}$, respectively) associated to a higher turn-over number ($21.0 \pm 0.5 \text{ s}^{-1}$ and $45.0 \pm 1.6 \text{ s}^{-1}$, respectively) (Table S7).

We also investigated the structure–activity relationships of SynGSTC1 by targeting the first two residues of the $^{10}\text{SRAS}^{13}$ motif, S10 being suspected to activate glutathione as in Ser-GSTs and R11 because of its ubiquity in GSTCs (see below in Section 3.5). The kinetic constants and the effect of pH on activities toward PITC were determined for S10T, S10A, S10C and R11A variants (Table 2 and Figure S7). The optimal pH of SynGSTC1 WT (7.4 units) was in the same range as usually observed for GSTs [57]. The substitution of S10 by a threonine residue slightly decreased the optimal pH of the enzyme (6.9 vs. 7.4 for WT) and the catalytic efficiency ($k_{\text{cat}}/K_{\text{m}}$) of the protein ($2.1 \times 10^5 \text{ M}^{-1}\cdot\text{s}^{-1}$ for S10T vs. $3.7 \times 10^5 \text{ M}^{-1}\cdot\text{s}^{-1}$ for WT). This result was consistent with the crystal structure of S10T which was superimposable to the wild-type (Figure S8). The bulkier threonine side chain did not impair GSH binding. Indeed, the GSH apparent affinity (K_{m}) was not altered in the S10T variant (Table 2). Furthermore, a sequence analysis of GSTCs (see below in Section 3.5) showed either a serine or a threonine as the first residue of the active site motif ($^{10}\text{SRAS}^{13}$ in SynGSTC1). S10A remained effective even though the decrease was greater than for S10T, being divided by 10 and 1.2 in S10A and S10T, respectively, compared to WT. All the kinetic parameters were affected roughly similarly. The crystal structure of SynGSTC1 did not show interaction between S10 and GSH. Instead, S10 was rather involved in stabilizing the $\beta 1$ – $\alpha 1$ loop in the close vicinity of the G-site (see above in Section 3.3) suggesting that the S10A substitution most likely disrupted the integrity of the active site. This resulted in a degradation of the catalytic constants and a moderate increase of the

catalysis optimal pH (shift of 0.4 unit compared to WT). The R11A substitution also did not fully abolish the activity of the enzyme even though it decreased significantly (divided by a factor close to 250 as compared to WT). R11 formed a salt bridge with the N-terminal carboxylate group of GSH in the crystal structure (see above in Section 3.3). This interaction did not seem to be essential for the catalysis because the GSH apparent affinity in R11A was not much more degraded than in S10A (four and five times higher in S10A and R11A variants, respectively, compared to WT). The electrostatic influence of R11 on the catalytic process was, however, clear since the catalytic rate was 75 times lower in R11A compared to WT. This was accompanied by a significant one-unit increase in optimal pH suggesting a higher GSH-thiol pKa in the R11A variant than in the WT enzyme. These variations appeared small compared to those observed in eta GSTH1-1 from *Agrobacterium tumefaciens*, which harbored an arginine residue at the same position as in SynGSTC1. Indeed, the R34A mutation in AtuGSTH1-1 had a detrimental effect on the catalytic constant, which dropped by at least a factor of 5000 [58]. Finally, the substitution of S10 by a cysteine residue, had the same global effect as the S10A mutation ($2.9 \times 10^4 \text{ M}^{-1} \cdot \text{s}^{-1}$ for S10C vs. 3.6×10^4 for S10A). Unlike the WT protein and other variants, the S10C enzyme was also active (k_{cat}/K_m of $3.36 \times 10^3 \pm 0.08 \times 10^3 \text{ M}^{-1} \cdot \text{s}^{-1}$) with HED, a substrate commonly used to characterize Grxs and cysteinyl-GSTs, indicating that this variant acquired a significant thiol-transferase activity.

Table 2. Kinetic parameters of SynGSTC1 toward model substrates.

	PITC	GSH	HED
$k_{\text{cat}} (\text{s}^{-1})$			
WT	12.6 ± 0.2		ND
S10T	7.2 ± 0.1		ND
S10A	2.60 ± 0.05		ND
S10C	1.19 ± 0.02		0.0111 ± 0.0002
R11A	0.170 ± 0.003		ND
$K_m (\mu\text{M})$			
WT	33.8 ± 2.5	135.2 ± 7.9	ND
S10T	33.6 ± 2.6	142.8 ± 14.6	ND
S10A	89.7 ± 6.4	528.4 ± 33.0	ND
S10C	33.0 ± 3.0	2149 ± 123	3.3 ± 0.4
R11A	108.4 ± 6.0	719.2 ± 58.2	ND
$k_{\text{cat}}/K_m (\text{M}^{-1} \cdot \text{s}^{-1})$			
WT	$3.73 \times 10^5 \pm 0.06 \times 10^5$		ND
S10T	$2.14 \times 10^5 \pm 0.04 \times 10^5$		ND
S10A	$2.89 \times 10^4 \pm 0.06 \times 10^4$		ND
S10C	$3.60 \times 10^4 \pm 0.06 \times 10^4$		$3.36 \times 10^3 \pm 0.08 \times 10^3$
R11A	$1.53 \times 10^3 \pm 0.02 \times 10^3$		ND

The apparent K_m values of SynGSTC1 wild-type and variants (S10T, S10A, S10C and R11A) were determined by varying substrate concentrations at a fixed saturating GSH concentration. The apparent K_m and k_{cat} values were calculated with *Prism 8* software using the Michaelis–Menten equation as nonlinear regression model. Results are means $\pm$ S.D. ($n = 3$).

3.5. Comparative Genomic Analysis

A similarity-based survey of 222 reference proteomes of the Cyanobacteria/Melainabacteria group led to the identification of 870 full-length GSTC1 homologues (BLASTP E-value cutoff 10^{-3}). The phylogenetic analysis of these sequences led to a large tree (Figure S9). According to this tree, the glutathione transferase chi1 from *Synechocystis* sp. PCC 6803 (SynGSTC1) belonged to a large group of 147 sequences displaying a SRAS motif or related motifs (Figures S9 and S10 and Table S8). These 147 GSTC protein sequences displayed more than 35% of sequence identity and were largely distributed in Cyanobacteria, being present in 144 of the 208 analyzed cyanobacterial proteomes (Figures 5 and S11 for high-quality version). In contrast, they were absent in the noncyanobacterial members of the Cyanobacteria/Melainabacteria group. More precisely, they were present in all cyanobacterial orders excepted Gloeobacterales, the oldest branching extant group of cyanobacteria [59].

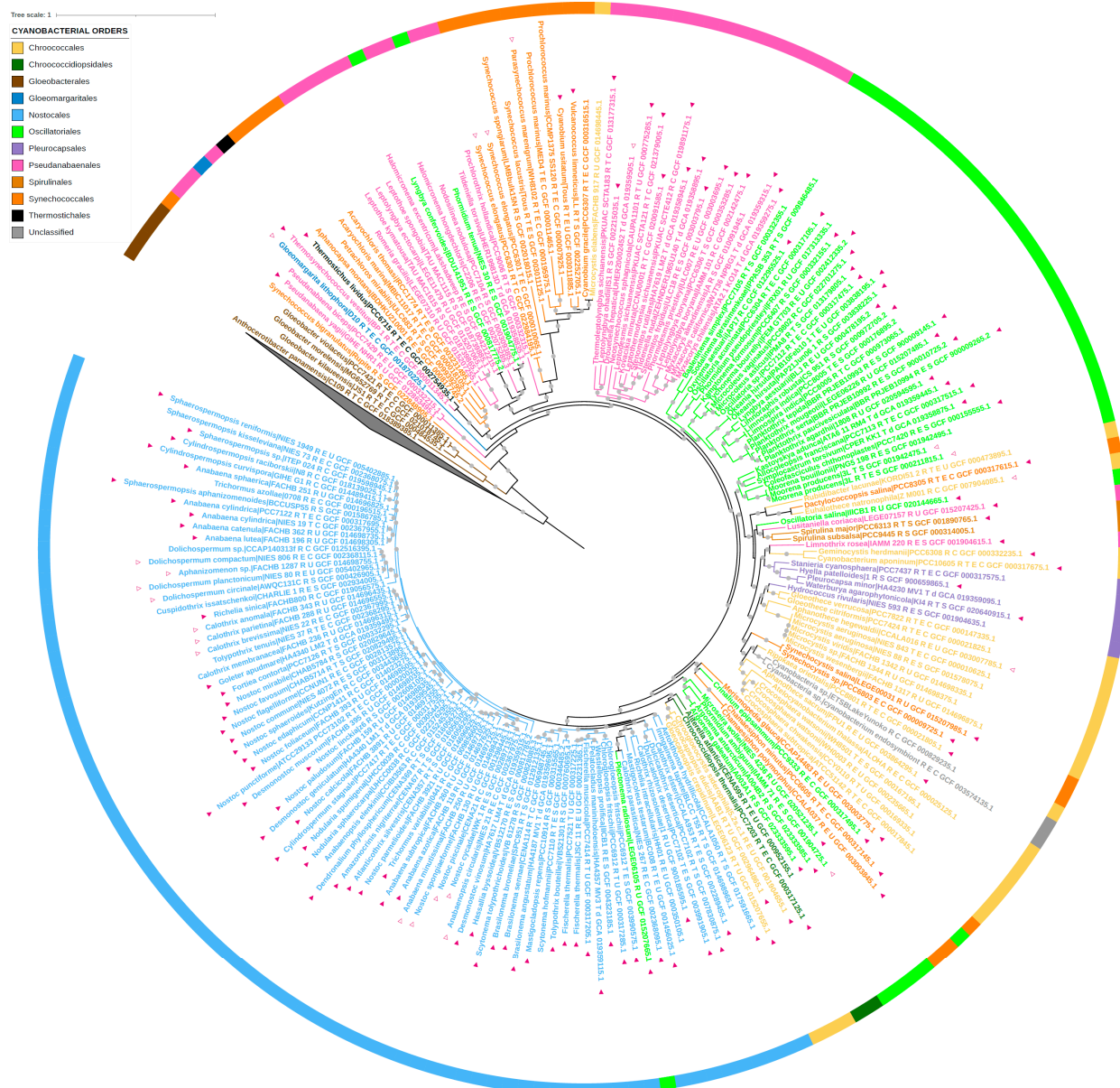


Figure 5. Phylogeny of the 222 proteomes of Cyanobacteria/Melainabacteria group considered in this study. The tree was inferred with *IQ-TREE* using the 52 rprots sequences present in more than 70% of the 222 proteomes (6430 amino acid sites, LG + C20 + F + R4 evolutionary model). The scale bar corresponds to the average number of substitutions per site. Gray circles correspond to ultrafast bootstrap values >90% (1000 replicates). The taxonomy of each proteome is indicated: Gloeobacterales (brown), Synechococcales (orange), Pseudanabaenales (pink), Gloeomargaritales (dark blue), Thermostichales (black), Oscillatoriales (light green), Chroococcales (yellow), Pleurocapsales (purple), Chroococcidiopsidales (dark green), Nostocales (light blue), and unclassified (gray). The 122 GSTC protein sequences harboring the SRAS motif are indicated with filled triangles, while the 25 GSTC sequences harboring variants of the SRAS motif are indicated with empty triangles. All the motifs are described in the Supplementary Table S8. The phylogeny of these 147 GSTC sequences is shown as Figure S10. A high-quality pdf version of the tree is provided as Figure S11 in the online Supplementary Materials.

To go further, we inferred the phylogeny of the 147 sequences displaying the SRAS motif (or related motifs) (Table S9). As expected, due to the restricted number of amino acid positions retained after the alignment trimming, branch supports were overall low

(Figure S10). Despite this global lack of support, the resulting tree showed clearly that sequences harboring the SRAS motif and sequences harboring related motifs were mixed on the tree, indicating that the canonical SRAS motif was lost several times independently during the diversification of GSTCs. Furthermore, the topology of the tree also showed some inconsistencies with the phylogeny of species (Figures 5 and S10). For instance, some Chroococcales sequences emerged within Nostocales (Figure S10), indicating that the evolutionary history of GSTCs harboring the SRAS (or related motifs) was impacted by horizontal gene transfers (HGTs) (Figure S10). Interestingly, these HGTs also contributed to spread GSTC1 outside of Cyanobacteria, since homologues were found in a few non-cyanobacterial bacteria (Figure S9). Most of them were marine and freshwater bacteria and some were recently closely related to cyanobacteria as Planctomycetaceae bacterium TMED241. Indeed, it was found that this bacterium contains a circadian clock *kaiABC* operon, which is typically found in cyanobacteria [60].

The majority of the 147 GSTC sequences had a length of less than 190 amino acids (Table S10). A dozen had longer sequences because they contained extensions at the N-terminus and/or between the secondary structures. All GSTCs had a reduced C-ter domain with a shortening of helices $\alpha 4$ and $\alpha 5$ as observed in the crystal structure of SynGSTC1. The SRAS motif ($^{10}\text{SRAS}^{13}$ in SynGSTC1) was well conserved. The arginine residue was invariant, the first position was replaced in a few cases by a threonine residue and the last two positions were a bit more variable. Surprisingly other residues involved in the structural attributes of SynGSTC1 were not conserved such as the patch of leucine residues (L70 and L94) in the core of the dimer, or the key residue of the lock and key motif (F49), or quaternary contributors to the stabilization of GSH (S98, T99, R116) (Figure 2). The sequence alignment revealed the conservation of 13 residues, most of which were located in the N-terminal domain (eight residues) and more precisely in the domain I (six residues) (Tables S8 and S9). The N-ter domain is generally better conserved than the C-ter domain because it contains an extended part of the active site [2]. In one subunit, the set of conserved residues was not centered on the active site but rather on the center of gravity of the monomer. The residues were distributed almost homogeneously around this center and most of them were located at a distance of less than 10 Å from it (Figure S12). This distribution was consistent with what is usually observed in proteins, namely that the most conserved positions tend to be situated in the core of the protein or on functional surfaces [61]. While the structural role of these conserved residues is obvious, it is difficult to identify those that form the signature of GSTs chi and most of them have been shown conserved in a class of GSTs. Only N97 seemed specific to the GST chi class; it most likely contributed electrostatically to the active site, as it was located near the γ -Glu moiety of GSH and close to the guanidinium group of the SRAS motif (Table S9).

4. Conclusions

This study increased the knowledge on the biochemical characteristic acquired on the chi class of GSTs (GSTCs) and detailed for the first time the structural attributes of this GST class, specific to cyanobacteria. These short-sequence GSTs (~180 aa) had a three-dimensional structure with a very open active site because the $\alpha 4$ and $\alpha 5$ helices were significantly shorter than those usually observed. The glutathione substrate was tightly bound to the enzyme with its reactive center exposed to the solvent. The transfer of multipolar density parameters from small peptides to SynGSTC1 permitted the gradation of residues involved in GSH stabilization. The two carboxylate groups of GSH were the two chemical groups that best adhered to the protein.

GSTCs contained a SRAS conserved motif at the N-terminus of the $\alpha 1$ helix indicating that they belonged to the SerGST group because the first residue of the motif was a serine residue. However, this serine residue was not directly involved in the catalytic act as assumed in SerGSTs [9]. The SRAS motif appeared to constrain the conformation of the serine side chain towards the interior of the protein and not towards the thiol group of GSH. S10 (in SynGSTC1) had a weak and unfavorable electrostatic influence on GSH and its

mutation did not drastically alter the catalytic properties of the enzyme. The denomination TyrGST, CysGST, SerGST and AtyGST (tyrosine type GST, . . . , Atypical GST) has the advantage of simplifying the confusing and cumbersome Greek letter classification. It is relevant in the case of TyrGSTs from a phylogenetic point of view [1]. It is also appropriate in the case of CysGSTs because the cysteine residue is covalently bound to the substrate in one step of the catalytic mechanism [62]. The disadvantage of the residue-based naming is its stigmatization on one residue that may not have a strong link to the activity of the enzyme as is the case for SynGSTC1.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/biom12101466/s1>; Table S1. List of the PCR and mutagenic primers used in this study; Table S2. List of the 222 studied proteomes; Table S3. List of the 53 ribosomal protein families present in bacteria according to riboDB database; Table S4. Strong intersubunit contacts in SynGSTC1; Table S5. Comparison of SynGSTC1 with its structural homologs; Table S6. Permanent, polarization and total electrostatic interaction energies in the active sites of SynGSTC1, of Epsilon 2 GST from *Anopheles Gambiae* (AgGSTE2) and of Alpha 1 GST from chicken (GgGSTA1); Table S7. Kinetic parameters of SynGSTC1 toward model substrates; Table S8. List of the 147 full-length protein sequences displaying a SRAS motif (or a related motif) identified in the 222 proteomes of Cyanobacteria/Melainabacteria group; Table S9. Invariant amino acid residues in the GST Chi Class; Table S10. Multiple sequence alignment of the 147 GSTCs (displaying a SRAS motif or a related motif) identified in the 222 proteomes of Cyanobacteria/Melainabacteria group; Figure S1. Stereoviews of the 2mFo-DFc map of the SynGSTC1 inter-domain linker; Figure S2. Φ and Ψ torsion angles for the inter-domain linker residues in SynGSTC1 during the simulation; Figure S3. Structure-based phylogenetic tree of SynGSTC1 with structural homologs; Figure S4. Stereoview of the comparison of the SRAS motif in SynGSTC1 and in GST SMc00097 from *Simorhizobium meliloti* 2011; Figure S5. N-C α -C β -S γ torsion angle of glutathione during molecular dynamics simulation of SynGSTC1; Figure S6. Interatomic distances (Å) between γ -oxygen atom of Ser10 and selected atoms during molecular dynamics simulation of SynGSTC1; Figure S7. Optimal reaction pH of SynGSTC1 and variants S10T, S10A, S10C and R11A; Figure S8. Structural comparison of the active site of SynGSTC1 WT with S10T and R11A variants; Figure S9. Phylogeny of the 870 GST sequences identified in the 222 studied proteomes of the Cyanobacteria / Melainabacteria group and the 11 GSTC-related sequences identified in noncyanobacterial bacteria; Figure S10. Phylogeny of the 147 cyanobacterial GST1 sequences harboring the SRAS motif or related motifs; Figure S11. Phylogeny of the 222 proteomes of Cyanobacteria/Melainabacteria group considered in this study (high quality version of the tree provided in Figure 5); Figure S12. Stereoview of the invariant amino acid residues in the GST Chi Class and WebLogos of aligned GSTCs from cyanobacteria. References [63–68] are cited in Supplementary Materials.

Author Contributions: Conceptualization, C.D. and A.H.; formal analysis, E.M., L.M., G.M., B.G., C.B.-A., C.D. and A.H.; funding acquisition, C.D. and A.H.; methodology, E.M., L.M., S.M., B.G. and A.H.; supervision, B.G., F.D., C.D. and A.H.; writing—original draft, E.M., C.D. and A.H.; writing—review and editing, E.M., G.M., B.G., F.C., C.C.-C., C.B.-A., C.D. and A.H. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by a grant from “Agence Nationale pour la Recherche” as part of the “Investissements d’Avenir” program (ANR-11-LABX-0002-01 and ANR-17-CE20-0008-01). This work was supported by the “Institut Jean Barriol”.

Institutional Review Board Statement: Not relevant for this study.

Informed Consent Statement: Not relevant for this study.

Data Availability Statement: PDB data (8AI8, 8AI9 and 8AIB) are made freely available by the wwPDB (<https://www.wwpdb.org/> (accessed on 30 September 2022)).

Acknowledgments: We thank LabEx ARBRE and Région Grand Est for funding LM’s PhD thesis. The authors appreciated the access to the “Plateforme de mesures de diffraction X” of the Université de Lorraine. We acknowledge SOLEIL (Gif Sur Yvette, France) and ESRF (Grenoble, France) for providing synchrotron radiation facilities, and we thank the staffs of PROXIMA-1, PROXIMA-2 and BM07 beamlines for assistance.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

Abbreviations

GST: glutathione transferase; GSTC, chi-class GST; SynGSTC1, GSTC1 from *Synechocystis* sp. PCC 6803; rmsd, root-mean-square deviation; ESRF, European Synchrotron Radiation Facility; PDB, Protein Data Bank; WT, wild-type.

References

- Mashiyama, S.T.; Malabanan, M.M.; Akiva, E.; Bhosle, R.; Branch, M.C.; Hillerich, B.; Jagessar, K.; Kim, J.; Patskovsky, Y.; Seidel, R.D.; et al. Large-scale determination of sequence, structure, and function relationships in cytosolic glutathione transferases across the biosphere. *PLoS Biol.* **2014**, *12*, e1001843. [\[CrossRef\]](#)
- Allocati, N.; Federici, L.; Masulli, M.; Di Ilio, C. Glutathione transferases in bacteria. *FEBS J.* **2009**, *276*, 58–75. [\[CrossRef\]](#) [\[PubMed\]](#)
- Ma, X.-X.; Jiang, Y.-L.; He, Y.-X.; Bao, R.; Chen, Y.; Zhou, C.-Z. Structures of yeast glutathione-S-transferase Gtt2 reveal a new catalytic type of GST family. *EMBO Rep.* **2009**, *10*, 1320–1326. [\[CrossRef\]](#) [\[PubMed\]](#)
- Ribeiro, A.J.M.; Tyzack, J.D.; Borkakoti, N.; Holliday, G.L.; Thornton, J.M. A global analysis of function and conservation of catalytic residues in enzymes. *J. Biol. Chem.* **2020**, *295*, 314–324. [\[CrossRef\]](#)
- Perrot, T.; Schwartz, M.; Saiag, F.; Salzet, G.; Dumarçay, S.; Favier, F.; Gérardin, P.; Girardet, J.-M.; Sormani, R.; Morel-Rouhier, M.; et al. Fungal Glutathione Transferases as Tools to Explore the Chemical Diversity of Amazonian Wood Extractives. *ACS Sustain. Chem. Eng.* **2018**, *6*, 13078–13085. [\[CrossRef\]](#)
- Schwartz, M.; Perrot, T.; Aubert, E.; Dumarçay, S.; Favier, F.; Gerardin, P.; Morel-Rouhier, M.; Mulliert, G.; Saiag, F.; Didierjean, C.; et al. Molecular recognition of wood polyphenols by phase II detoxification enzymes of the white rot *Trametes versicolor*. *Sci. Rep.* **2018**, *8*, 8472. [\[CrossRef\]](#)
- Allocati, N.; Masulli, M.; Di Ilio, C.; Federici, L. Glutathione transferases: Substrates, inhibitors and pro-drugs in cancer and neurodegenerative diseases. *Oncogenesis* **2018**, *7*, 8. [\[CrossRef\]](#) [\[PubMed\]](#)
- Nianiou-Obeidat, I.; Madesis, P.; Kissoudis, C.; Voulgari, G.; Chronopoulou, E.; Tsaftaris, A.; Labrou, N.E. Plant glutathione transferase-mediated stress tolerance: Functions and biotechnological applications. *Plant Cell Rep.* **2017**, *36*, 791–805. [\[CrossRef\]](#)
- Sylvestre-Gonon, E.; Law, S.R.; Schwartz, M.; Robe, K.; Keech, O.; Didierjean, C.; Dubos, C.; Rouhier, N.; Hecker, A. Functional, Structural and Biochemical Features of Plant Serinyl-Glutathione Transferases. *Front. Plant Sci.* **2019**, *10*, 608. [\[CrossRef\]](#)
- Wikteliu, E.; Stenberg, G. Novel class of glutathione transferases from cyanobacteria exhibit high catalytic activities towards naturally occurring isothiocyanates. *Biochem. J.* **2007**, *406*, 115–123. [\[CrossRef\]](#)
- Shehu, D.; Abdullahi, N.; Alias, Z. Cytosolic Glutathione S-transferase in Bacteria: A Review. *Pol. J. Environ. Stud.* **2019**, *28*, 515–528. [\[CrossRef\]](#)
- Meux, E.; Prosper, P.; Ngadin, A.; Didierjean, C.; Morel, M.; Dumarçay, S.; Lamant, T.; Jacquot, J.-P.; Favier, F.; Gelhaye, E. Glutathione Transferases of *Phanerochaete chrysosporium*: S-Glutathionyl-p-hydroquinone Reductase Belongs to a New Structural Class. *J. Biol. Chem.* **2011**, *286*, 9162–9173. [\[CrossRef\]](#) [\[PubMed\]](#)
- Masai, E.; Ichimura, A.; Sato, Y.; Miyauchi, K.; Katayama, Y.; Fukuda, M. Roles of the enantioselective glutathione S-transferases in cleavage of beta-aryl ether. *J. Bacteriol.* **2003**, *185*, 1768–1775. [\[CrossRef\]](#) [\[PubMed\]](#)
- Meux, E.; Prosper, P.; Masai, E.; Mulliert, G.; Dumarçay, S.; Morel, M.; Didierjean, C.; Gelhaye, E.; Favier, F. *Sphingobium* sp. SYK-6 LigG involved in lignin degradation is structurally and biochemically related to the glutathione transferase omega class. *FEBS Lett.* **2012**, *586*, 3944–3950. [\[CrossRef\]](#) [\[PubMed\]](#)
- Kammerscheit, X.; Hecker, A.; Rouhier, N.; Chauvat, F.; Cassier-Chauvat, C. Methylglyoxal Detoxification Revisited: Role of Glutathione Transferase in Model Cyanobacterium *Synechocystis* sp. Strain PCC 6803. *mBio* **2020**, *11*, e00882-20. [\[CrossRef\]](#)
- Pandey, T.; Singh, S.K.; Chhetri, G.; Tripathi, T.; Singh, A.K. Characterization of a Highly pH Stable Chi-Class Glutathione S-Transferase from *Synechocystis* PCC 6803. *PLoS ONE* **2015**, *10*, e0126811. [\[CrossRef\]](#)
- Feil, S.C.; Tang, J.; Hansen, G.; Gorman, M.A.; Wikteliu, E.; Stenberg, G.; Parker, M.W. Crystallization and preliminary X-ray analysis of glutathione transferases from cyanobacteria. *Acta Crystallogr. Sect. F Crystallogr. Commun.* **2009**, *65*, 475–477. [\[CrossRef\]](#)
- Allocati, N.; Casalone, E.; Masulli, M.; Ceccarelli, I.; Carletti, E.; Parker, M.W.; Di Ilio, C. Functional analysis of the evolutionarily conserved proline 53 residue in *Proteus mirabilis* glutathione transferase B1-1. *FEBS Lett.* **1999**, *445*, 347–350. [\[CrossRef\]](#)
- Dragani, B.; Stenberg, G.; Melino, S.; Petruzzelli, R.; Mannervik, B.; Aceto, A. The Conserved N-capping Box in the Hydrophobic Core of Glutathione S-Transferase P1-1 Is Essential for Refolding. Identification of A buried and Conserved Hydrogen Bond Important for Protein Stability. *J. Biol. Chem.* **1997**, *272*, 25518–25523. [\[CrossRef\]](#)
- ShylajaNaciyar, M.; Karthick, L.; Prakasam, P.A.; Deviram, G.; Uma, L.; Prabakaran, D.; Saha, S.K. Diversity of Glutathione S-Transferases (GSTs) in Cyanobacteria with Reference to Their Structures, Substrate Recognition and Catalytic Functions. *Microorganisms* **2020**, *8*, 712. [\[CrossRef\]](#)
- Kabsch, W. XDS. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **2010**, *66*, 125–132. [\[CrossRef\]](#) [\[PubMed\]](#)

22. Evans, P.R.; Murshudov, G.N. How good are my data and what is the resolution? *Acta Crystallogr. Sect. D Biol. Crystallogr.* **2013**, *69*, 1204–1214. [[CrossRef](#)] [[PubMed](#)]
23. Winn, M.D.; Ballard, C.C.; Cowtan, K.D.; Dodson, E.J.; Emsley, P.; Evans, P.R.; Keegan, R.M.; Krissinel, E.B.; Leslie, A.G.W.; McCoy, A.; et al. Overview of the CCP4 suite and current developments. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **2011**, *67*, 235–242. [[CrossRef](#)] [[PubMed](#)]
24. Vagin, A.; Lebedev, A. *MoRDa*, an automatic molecular replacement pipeline. *Acta Crystallogr. Sect. A Found. Adv.* **2015**, *71*, s19. [[CrossRef](#)]
25. Smart, O.S.; Womack, T.O.; Flensburg, C.; Keller, P.; Paciorek, W.; Sharff, A.; Vonrhein, C.; Bricogne, G. Exploiting structure similarity in refinement: Automated NCS and target-structure restraints in BUSTER. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **2012**, *68*, 368–380. [[CrossRef](#)]
26. Emsley, P.; Lohkamp, B.; Scott, W.G.; Cowtan, K. Features and development of Coot. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **2010**, *66*, 486–501. [[CrossRef](#)] [[PubMed](#)]
27. Hansen, N.K.; Coppens, P. Testing aspherical atom refinements on small-molecule data sets. *Acta Crystallogr. Sect. A Found. Adv.* **1978**, *34*, 909–921. [[CrossRef](#)]
28. Domagala, S.; Fournier, B.; Liebschner, D.; Guillot, B.; Jelsch, C. An improved experimental databank of transferable multipolar atom models—ELMAM2. Construction details and applications. *Acta Crystallogr. Sect. A Found. Adv.* **2012**, *68*, 337–351. [[CrossRef](#)]
29. Leduc, T.; Aubert, E.; Espinosa, E.; Jelsch, C.; Iordache, C.; Guillot, B. Polarization of Electron Density Databases of Transferable Multipolar Atoms. *J. Phys. Chem.* **2019**, *123*, 7156–7170. [[CrossRef](#)]
30. Guillot, B.; Enrique, E.; Huder, L.; Jelsch, C. MoProViewer: A tool to study proteins from a charge density science perspective. *Acta Crystallogr. Sect. A Found. Adv.* **2014**, *70*, C279. [[CrossRef](#)]
31. Vuković, V.; Leduc, T.; Jelić-Matošević, Z.; Didierjean, C.; Favier, F.; Guillot, B.; Jelsch, C. A rush to explore protein–ligand electrostatic interaction energy with Charger. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **2021**, *77*, 1292–1304. [[CrossRef](#)] [[PubMed](#)]
32. Phillips, J.C.; Hardy, D.J.; Maia, J.D.C.; Stone, J.E.; Ribeiro, J.V.; Bernardi, R.C.; Buch, R.; Fiorin, G.; Hénin, J.; Jiang, W.; et al. Scalable molecular dynamics on CPU and GPU architectures with NAMD. *J. Chem. Phys.* **2020**, *153*, 044130. [[CrossRef](#)] [[PubMed](#)]
33. Brooks, B.R.; Brooks, C.L., III; Mackerell, A.D., Jr.; Nilsson, L.; Petrella, R.J.; Roux, B.; Won, Y.; Archontis, G.; Bartels, C.; Boresch, S.; et al. CHARMM: The biomolecular simulation program. *J. Comput. Chem.* **2009**, *30*, 1545–1614. [[CrossRef](#)]
34. Jorgensen, W.L.; Chandrasekhar, J.; Madura, J.D.; Impey, R.W.; Klein, M.L. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **1983**, *79*, 926–935. [[CrossRef](#)]
35. Vanommeslaeghe, K.; Hatcher, E.; Acharya, C.; Kundu, S.; Zhong, S.; Shim, J.; Darian, E.; Guvench, O.; Lopes, P.; Vorobyov, I.; et al. CHARMM general force field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields. *J. Comput. Chem.* **2010**, *31*, 671–690. [[CrossRef](#)] [[PubMed](#)]
36. Andersen, H.C. Rattle: A “velocity” version of the shake algorithm for molecular dynamics calculations. *J. Comput. Phys.* **1983**, *52*, 24–34. [[CrossRef](#)]
37. Miyamoto, S.; Kollman, P.A. Settle: An analytical version of the SHAKE and RATTLE algorithm for rigid water models. *J. Comput. Chem.* **1992**, *13*, 952–962. [[CrossRef](#)]
38. Hopkins, C.W.; Le Grand, S.; Walker, R.C.; Roitberg, A.E. Long-Time-Step Molecular Dynamics through Hydrogen Mass Repartitioning. *J. Chem. Theory Comput.* **2015**, *11*, 1864–1874. [[CrossRef](#)]
39. Humphrey, W.; Dalke, A.; Schulten, K. VMD: Visual molecular dynamics. *J. Mol. Graph.* **1996**, *14*, 33–38. [[CrossRef](#)]
40. Jauffrit, F.; Penel, S.; Delmotte, S.; Rey, C.; de Vienne, D.M.; Gouy, M.; Charrier, J.P.; Flandrois, J.P.; Brochier-Armanet, C. RiboDB Database: A Comprehensive Resource for Prokaryotic Systematics. *Mol. Biol. Evol.* **2016**, *33*, 2170–2172. [[CrossRef](#)]
41. Katoh, K.; Standley, D.M. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.* **2013**, *30*, 772–780. [[CrossRef](#)] [[PubMed](#)]
42. Criscuolo, A.; Gribaldo, S. BMGE (Block Mapping and Gathering with Entropy): A new software for selection of phylogenetic informative regions from multiple sequence alignments. *BMC Evol. Biol.* **2010**, *10*, 210. [[CrossRef](#)] [[PubMed](#)]
43. Minh, B.Q.; Schmidt, H.A.; Chernomor, O.; Schrempf, D.; Woodhams, M.D.; von Haeseler, A.; Lanfear, R. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol. Biol. Evol.* **2020**, *37*, 1530–1534. [[CrossRef](#)] [[PubMed](#)]
44. Price, M.N.; Dehal, P.S.; Arkin, A.P. FastTree 2—approximately maximum-likelihood trees for large alignments. *PLoS ONE* **2010**, *5*, e9490. [[CrossRef](#)]
45. Le, S.Q.; Gascuel, O. An improved general amino acid replacement matrix. *Mol. Biol. Evol.* **2008**, *25*, 1307–1320. [[CrossRef](#)]
46. Letunic, I.; Bork, P. Interactive Tree Of Life (iTOL) v5: An online tool for phylogenetic tree display and annotation. *Nucleic Acids Res.* **2021**, *49*, W293–W296. [[CrossRef](#)]
47. Hayes, J.D.; Flanagan, J.U.; Jowsey, I.R. Glutathione transferases. *Annu. Rev. Pharmacol. Toxicol.* **2005**, *45*, 51–88. [[CrossRef](#)]
48. Polekhina, G.; Board, P.G.; Blackburn, A.C.; Parker, M.W. Crystal structure of maleylacetoacetate isomerase/glutathione transferase zeta reveals the molecular basis for its remarkable catalytic promiscuity. *Biochemistry* **2001**, *40*, 1567–1576. [[CrossRef](#)]
49. Wilce, M.C.J.; Parker, M.W. Structure and function of glutathione S-transferases. *Biochim. Biophys. Acta Prot. Struct. Mol. Enzymol.* **1994**, *1205*, 1–18. [[CrossRef](#)]
50. Holm, L.; Laakso, L.M. Dali server update. *Nucleic Acids Res.* **2016**, *44*, W351–W355. [[CrossRef](#)]
51. Dong, R.; Pan, S.; Peng, Z.; Zhang, Y.; Yang, J. mTM-align: A server for fast protein structure database search and multiple protein structure alignment. *Nucleic Acids Res.* **2018**, *46*, W380–W386. [[CrossRef](#)]

52. Jacquot, J.P.; Gelhaye, E.; Rouhier, N.; Corbier, C.; Didierjean, C.; Aubry, A. Thioredoxins and related proteins in photosynthetic organisms: Molecular basis for thiol dependent regulation. *Biochem. Pharmacol.* **2002**, *64*, 1065–1069. [[CrossRef](#)]
53. Dixon, D.P.; Edwards, R. Glutathione Transferases. *Arabidop. Book* **2010**, *8*, e0131. [[CrossRef](#)]
54. Pegeot, H.; Koh, C.S.; Petre, B.; Mathiot, S.; Duplessis, S.; Hecker, A.; Didierjean, C.; Rouhier, N. The poplar Phi class glutathione transferase: Expression, activity and structure of GSTF1. *Front. Plant. Sci.* **2014**, *5*, 712. [[CrossRef](#)]
55. Sheehan, D.; Meade, G.; Foley, V.M.; Dowd, C.A. Structure, function and evolution of glutathione transferases: Implications for classification of non-mammalian members of an ancient enzyme superfamily. *Biochem. J.* **2001**, *360*, 1–16. [[CrossRef](#)]
56. Sylvestre-Gonon, E.; Morette, L.; Vilorio, M.; Mathiot, S.; Boutilliat, A.; Favier, F.; Rouhier, N.; Didierjean, C.; Hecker, A. Biochemical and structural insights on the poplar tau glutathione transferase GSTU19 and 20 paralogs binding flavonoids. *Front. Mol. Biosci.* **2022**, *9*, 9585866. [[CrossRef](#)] [[PubMed](#)]
57. Dourado, D.F.; Fernandes, P.A.; Mannervik, B.; Ramos, M.J. Glutathione transferase: New model for glutathione activation. *Chemistry* **2008**, *14*, 9591–9598. [[CrossRef](#)]
58. Skopelitou, K.; Dhavala, P.; Papageorgiou, A.C.; Labrou, N.E. A glutathione transferase from *Agrobacterium tumefaciens* reveals a novel class of bacterial GST superfamily. *PLoS ONE* **2012**, *7*, e34263. [[CrossRef](#)]
59. Moreira, D.; Tavera, R.; Benzerara, K.; Skouri-Panet, F.; Couradeau, E.; Gerard, E.; Fonta, C.L.; Novelo, E.; Zivanovic, Y.; Lopez-Garcia, P. Description of *Gloeomargarita lithophora* gen. nov., sp. nov., a thylakoid-bearing, basal-branching cyanobacterium with intracellular carbonates, and proposal for *Gloeomargaritales* ord. nov. *Int. J. Syst. Evol. Microbiol.* **2017**, *67*, 653–658. [[CrossRef](#)]
60. Dvornyk, V.; Mei, Q. Evolution of *kaiA*, a key circadian gene of cyanobacteria. *Sci. Rep.* **2021**, *11*, 9995. [[CrossRef](#)]
61. Halabi, N.; Rivoire, O.; Leibler, S.; Ranganathan, R. Protein sectors: Evolutionary units of three-dimensional structure. *Cell* **2009**, *138*, 774–786. [[CrossRef](#)] [[PubMed](#)]
62. Board, P.G.; Coggan, M.; Chelvanayagam, G.; Eastal, S.; Jermini, L.S.; Schulte, G.K.; Danley, D.E.; Hoth, L.R.; Griffon, M.C.; Kamath, A.V.; et al. Identification, characterization, and crystal structure of the omega class glutathione transferases. *J. Biol. Chem.* **2000**, *275*, 24798–24806. [[CrossRef](#)]
63. Williams, C.J.; Headd, J.J.; Moriarty, N.W.; Prisant, M.G.; Videau, L.L.; Deis, L.N.; Verma, V.; Keedy, D.A.; Hintze, B.J.; Chen, V.B.; et al. MolProbity: More and better reference data for improved all-atom structure validation. *Protein Sci.* **2018**, *27*, 293–315. [[CrossRef](#)] [[PubMed](#)]
64. DeLano, W.L. Pymol: An open-source molecular graphics tool. CCP4 Newsl. *Protein Crystallogr.* **2002**, *40*, 82–92.
65. Roret, T.; Thuillier, A.; Favier, F.; Gelhaye, E.; Didierjean, C.; Morel-Rouhier, M. Evolutionary divergence of Ure2pA glutathione transferases in wood degrading fungi. *Fungal Genet. Biol.* **2015**, *83*, 103–112. [[CrossRef](#)] [[PubMed](#)]
66. Aceto, A.; Dragani, B.; Melino, S.; Allocati, N.; Masulli, M.; Ilio, C.D.; Petruzzelli, R. Identification of an N-capping box that affects the α 6-helix propensity in glutathione S-transferase superfamily proteins: A role for an invariant aspartic residue. *Biochem. J.* **1997**, *322*, 229–234. [[CrossRef](#)]
67. Dirr, H.; Reinemer, P.; Huber, R. X-ray crystal structures of cytosolic glutathione S-transferases: Implications for protein architecture, substrate recognition and catalytic function. *Eur. J. Biochem.* **1994**, *220*, 645–661. [[CrossRef](#)] [[PubMed](#)]
68. Crooks, G.E.; Hon, G.; Chandonia, J.M.; Brenner, S.E. WebLogo: A sequence logo generator. *Genome Res.* **2004**, *14*, 1188–1190. [[CrossRef](#)]

4.4 Perspective d'application à un système dynamique : pompage de l'ion chlorure par l'halorhodopsine

Pour montrer le pouvoir d'analyse des mécanismes moléculaires dans les protéines par les descripteurs électrostatiques et les potentiels d'interaction développés dans cette thèse, nous avons pour perspective de les appliquer à des systèmes dynamiques. Pour cela, la description d'une protéine de transport membranaire, l'halorhodopsine, a été entreprise. L'halorhodopsine est une pompe activée par la lumière et spécifique à l'anion chlorure. L'étude de cristallographie résolue en temps de cette protéine par [Mous *et al.*, 2022] leur a permis de discuter la dynamique des changements conformationnels permettant le passage de l'ion chlorure du milieu extracellulaire vers le cytoplasme. Les calculs de zones d'influence et d'énergies d'interaction qui sont présentés dans cette partie donnent un aperçu du potentiel de ces méthodologies pour revisiter le mécanisme de transport de cette protéine.

4.4.1 L'halorhodopsine NmHR

Cycle de transport

L'halorhodopsine est une protéine membranaire de la famille des rhodopsines et est présente chez les archées halophiles. Elle est impliquée dans le pompage d'ions chlorure du milieu extra-cellulaire vers le cytoplasme. Ce mécanisme est enclenché par l'absorption d'un photon par le cofacteur rétinale des rhodopsines [Engelhard *et al.*, 2018]. Les changements structuraux accompagnant le transport de l'anion seraient entraînés par des effets de nature principalement électrostatique [Mous *et al.*, 2022], ce qui rend ce système particulièrement intéressant à étudier du point de vue des descripteurs développés dans cette thèse.

En combinant cristallographie résolue en temps, spectroscopie et simulations multi-échelles, [Mous *et al.*, 2022] ont proposé des mécanismes structuraux détaillés pour expliquer la dynamique du transport de l'ion chlorure dans l'halorhodopsine NmHR de l'organisme *Nonlabens Marinus*. Dans l'état de repos ("dark state"), le chromophore, à savoir le rétinale, est lié de façon covalente au résidu Lys235 par l'intermédiaire d'une base de Schiff. Dans cet état le rétinale est en configuration *trans*. L'ion chlorure est localisé dans un site de fixation faisant face à la base de Schiff protonée du rétinale (RPSB pour "Retinal Protonated Schiff Base"). La photoisomérisation du rétinale en configuration *cis* initie le cycle de transport de l'anion. Ce dernier serait alors transféré de l'autre côté du rétinale par interaction avec les électrons π du chromophore polarisé. Une première barrière moléculaire, appelée « barrière moléculaire stérique », fermerait l'accès au site de fixation initial suite à la relaxation de l'hélice C de NmHR pour empêcher le reflux de l'ion. La libération du chlorure dans le cytoplasme serait ensuite entraînée par un phénomène de diffusion assisté par le moment dipolaire global de la protéine, qui présente des charges négatives sur sa surface extra-cellulaire et des charges positives sur sa surface cytoplasmique.

Le retour vers l'état de repos nécessite l'entrée d'un anion dans la protéine qui serait permise par la présence d'une petite région chargée positivement sur la surface extra-cellulaire, qui est chargée négativement par ailleurs. Après plusieurs étapes de transport assurées par les charges positives des deux résidus Arg223 et Arg95, l'anion rencontre une seconde barrière moléculaire, la « barrière moléculaire électrostatique ». Celle-ci servirait de goulot d'étranglement fermant le

passage vers le site de fixation du rétinale et s'ouvrirait par interaction avec la charge négative du chlorure. En passant cette barrière, l'anion retourne dans le site de fixation initial tandis que le rétinale retrouve sa conformation *trans* de l'état de repos.

Ici, les mécanismes des barrières moléculaires dites stérique et électrostatique vont être étudiés du point de vue des descripteurs issus de la topologie du potentiel électrostatique et par le calcul d'énergies d'interaction en utilisant les potentiels ELMAM.

Préparation des systèmes

La structure de l'halorhodopsine NmHR a été étudiée par cristallographie résolue en temps après l'exposition du cristal à un rayonnement lumineux au temps initial t_0 pour activer le mécanisme de pompage. Seize structures atomiques ont été affinées à différents intervalles de temps $\Delta t = t - t_0$ écoulé depuis la photoactivation [Mous *et al.*, 2022]. Pour caractériser la barrière moléculaire stérique, deux de ces structures ont été choisies. Elles correspondent aux conformations où la courbure de l'hélice C est maximale, à $\Delta t = 1 \mu\text{s}$ (code PDB : 7O8I), et minimale, à $\Delta t = 2,5 \text{ ms}$ (code PDB : 7O8M), suite à la relaxation de celles-ci. Pour la barrière moléculaire électrostatique, trois autres structures ont été sélectionnées : à $\Delta t = 12,5 \text{ ms}$ (code PDB : 7O8O) et à $\Delta t = 22,5 \text{ ms}$ (code PDB : 7O8Q) entre lesquelles se produit le changement de conformation interprété comme l'ouverture de la barrière, et à $\Delta t = 37,5 \text{ ms}$ (code PDB : 7O8T) à partir de laquelle la barrière apparaît à nouveau dans son état initial.

Pour chacune de ces structures, après avoir supprimé les doubles conformations des résidus de la protéine et les molécules du solvant, les atomes d'hydrogène ont été ajoutés dans leurs orientations optimisées à l'aide du serveur en ligne MolProbity [Chen *et al.*, 2010]. Les fichiers au format MoPro correspondant à ces structures ont été obtenus en utilisant le programme Import2MoPro [Jelsch *et al.*, 2005]. La densité électronique des résidus de la protéine a été directement reconstruite par transfert des paramètres multipolaires de la librairie ELMAM2 [Domagała *et al.*, 2012] dans MoProViewer [Guillot *et al.*, 2014]. Les charges formelles des acides aminés ont été ajustées à $+1,0e$ pour les lysines et les arginines, à $-1,0e$ pour les aspartates et les glutamates et $0,0e$ pour les autres. Le type atomique correspondant à l'ion chlorure n'étant pas défini dans la librairie ELMAM2, nous avons choisi de le modéliser par une densité électronique sphérique avec une population de valence $P_{\text{val}} = 8,0e$, de sorte à obtenir une charge formelle de $-1,0e$. Le chromophore rétinale et la base de Schiff protonée (PSB) le reliant au résidu Lys235 faisant apparaître des types atomiques particuliers, nous avons fait le choix d'affiner des paramètres multipolaires spécifiquement pour cette étude.

Nous avons extrait de la structure de l'état initial (code PDB : 7O8F) les coordonnées atomiques du rétinale en conformation *trans* et lié à la Lys235. Les atomes de chaîne latérale de Lys235 ont été conservés jusqu'au carbone C δ qui a été remplacé par un groupement méthyle. La densité électronique théorique de ce système a été calculée par la méthode de chimie quantique de haut niveau B3LYP_6-311. Les facteurs de structure obtenus à partir de la densité électronique théorique ont été utilisés pour affiner à l'aide du logiciel MoPro [Jelsch *et al.*, 2005] les paramètres multipolaires atomiques P_{val} , $P_{l,m}$, κ et κ' , en tirant profit des symétries de la molécule. Le facteur d'accord R entre les facteurs de structure d'origine théorique et ceux calculés à partir des paramètres multipolaires affinés est de 0,57%. Le facteur $w_R^2(I)$, basé sur les

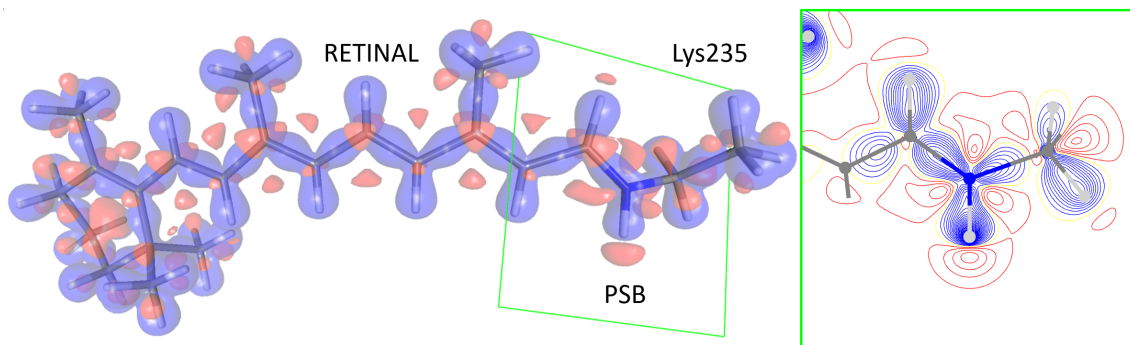


FIGURE 4.2 – Déformation multipolaire de la densité électronique du rétinale.

La déformation statique de la densité électronique multipolaire du rétinale liée à la Lys235 sont représentées par des isosurfaces, en bleu pour l'isovaleur $\delta\rho(\mathbf{r}) = +0,1e$ et en rouge pour l'isovaleur $\delta\rho(\mathbf{r}) = -0,1e$. Aussi, les surfaces bleues indiquent les régions d'accumulation d'électrons, sur les liaisons covalentes notamment, et les surfaces rouges les régions de déplétion d'électrons telles que celle entourant la base de Schiff protonnée (PSB). L'encadré de droite montre les isocontours de densité de déformation à intervalle $\delta\rho(\mathbf{r}) = 0,05e$ dans le plan du PSB.

intensités, est de 1,2%. La déformation de la densité électronique reconstruite à partir de ces paramètres multipolaires est illustrée sur la figure 4.2. Les déplétions électroniques autour du PSB ("Protonated Schiff Base") sont remarquables et sont cohérentes avec la charge positive attendue sur ce groupement. Les paramètres multipolaires ont ensuite été transférés aux structures de NmHR.

Les cartes de potentiel électrostatique utilisées pour calculer les zones d'influence présentées par la suite ont été obtenues à partir des distributions de charge multipolaires de la protéine et du rétinale, en excluant l'ion chlorure afin de visualiser les forces ressenties par celui-ci. Pour cela, nous avons employé le module Charger [Vuković *et al.*, 2021] de MoProViewer dans une boîte de calcul entourant la protéine avec une marge de 2,5 Å et un pas de grille de 0,2 Å.

4.4.2 Barrière moléculaire stérique

Mécanisme proposé par [Mous *et al.*, 2022]

Après avoir quitté le site de fixation initial, nommé CL351, l'ion chlorure est entraîné de l'autre côté du rétinale, dans le site de fixation CL352 stabilisé par le RPSB et par le résidu Thr102 (voir figure 4.3). Des calculs de chimie quantique ont révélé que cette stabilisation est dominée par la composante électrostatique entre le chlorure et le RPSB. Une relaxation de l'hélice C de la protéine est observée à partir de $\Delta t = 1\mu s$, faisant suite au déplacement de l'ion vers le site CL352 le rapprochant de la sortie cytoplasmique. La courbure de cette hélice est minimale à $\Delta t = 2,5ms$, lorsque la chaîne latérale du résidu Asn98 se positionne dans le site de fixation initial du chlorure CL351. Ce changement de conformation de l'Asn98 est interprété comme la fermeture stérique d'une barrière moléculaire empêchant un reflux de l'ion vers le milieu extra-cellulaire. La fermeture de cette barrière moléculaire stérique serait scellée par une liaison hydrogène entre les chaînes principales des résidus Asn98 et Thr102 (voir figures 4.3 et 4.4).

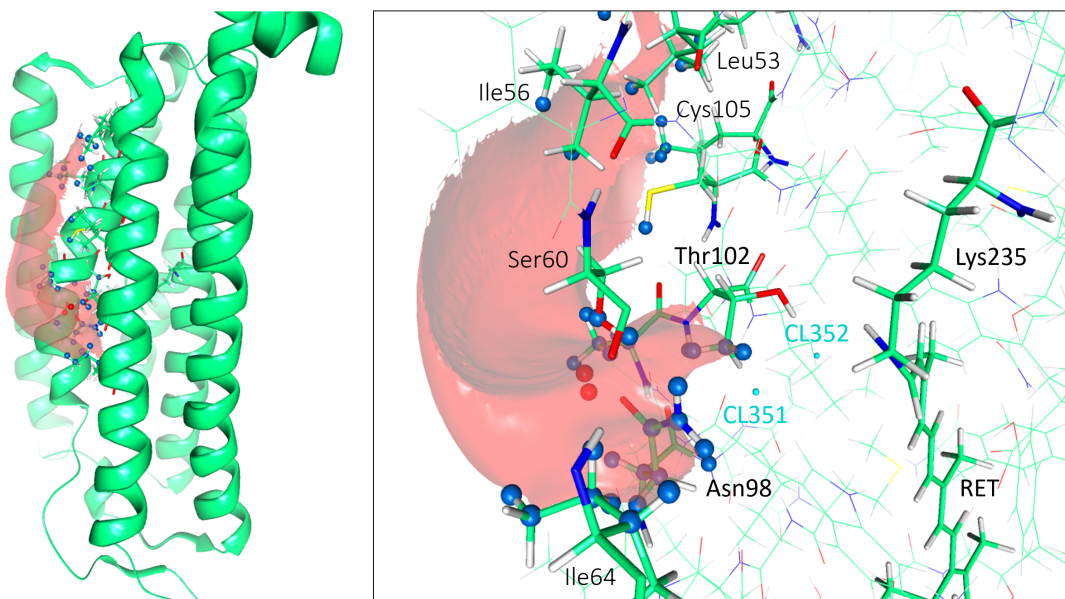


FIGURE 4.3 – Zone d'influence nucléophile de l'atome O δ 1 du résidu Asn98 dans la barrière moléculaire stérique ouverte.

La zone d'influence nucléophile (ZIN) associée aux sites nucléophiles (sphères rouges) correspondant aux doublets non-liants de l'atome Asn98-O δ 1 dans sa conformation à $\Delta t = 1\mu s$ est représentée par la surface rouge. À gauche, cette ZIN est montrée dans une vue globale de la protéine faisant apparaître sa structure secondaire, tandis que la vue de droite est centrée autour des sites de fixation CL351 et CL352 de l'ion chlorure (sphères cyan). Les lignes de champ électrique à l'intérieur de cette surface recouvrent un large volume de $210 \text{ \AA}^3$, s'étendant en dehors de la protéine. Ces lignes de champ rejoignent les sites électrophiles (sphères bleues) appartenant à plusieurs résidus : Ile49, Leu53, Ile56, Ser60, Ile64, Asn98, Ala101, Thr102, Cys105, Leu108 et Ile112, mais ne recouvrent pas le site de fixation CL351.

Caractérisation par les zones d'influence

L'Asn98 jouant un rôle essentiel dans le mécanisme de cette barrière moléculaire stérique, nous avons calculé la zone d'influence nucléophile (ZIN) de son oxygène O δ 1 dans les deux conformations à $\Delta t = 1\mu s$ et $\Delta t = 2,5ms$. Lorsque la barrière est en conformation ouverte, à $\Delta t = 1\mu s$, cette ZIN (représentée dans la figure 4.3) est extérieure à la protéine. En effet, les lignes de champ électrique à l'intérieur de ce large volume de $210 \text{ \AA}^3$, convergeant vers l'Asn98, rejoignent la surface cytoplasmique (haut de la figure 4.3) où se trouve le résidu Ile49. Ces lignes de champ émanent également de sites électrophiles localisés sur les atomes d'hydrogène et de carbone des résidus suivants : Leu53, Ile56, Ser60, Ile64, Asn98, Ala101, Thr102, Cys105, Leu108 et Ile112. Les lignes de champ allant du côté cytoplasmique vers le côté extra-cellulaire, l'orientation de cette ZIN est cohérente avec le moment dipolaire global de la protéine décrit par [Mous *et al.*, 2022], formé par les charges positives sur la surface cytoplasmique et les charges négatives sur la surface extra-cellulaire. Dans cette conformation, la ZIN de Asn98-O δ 1 ne couvre pas le site de fixation CL351 (ni le CL352 occupé par l'ion à $\Delta t = 1\mu s$) et ne montre pas d'interaction avec le RPSB qui participe à la stabilisation de ce site.

Dans la conformation fermée de la barrière moléculaire stérique, à $\Delta t = 2,5ms$, la ZIN de l'Asn98-O δ 1 (voir figure 4.4) n'est plus étendue vers l'extérieur de la protéine mais contenue à l'intérieur. Les lignes de champ électrique appartenant à cette ZIN s'étendent maintenant dans

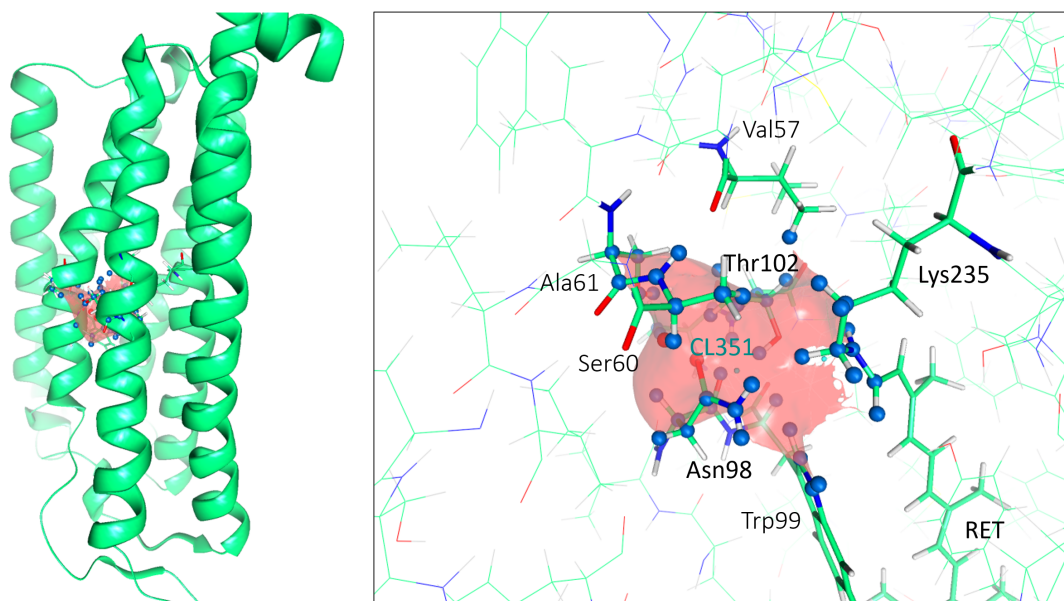


FIGURE 4.4 – Zone d'influence nucléophile de l'atome O δ 1 du résidu Asn98 dans la barrière moléculaire stérique fermée.

La zone d'influence nucléophile (ZIN) associée aux sites nucléophiles (sphères rouges) correspondant aux doublets non-liants de l'atome Asn98-O δ 1 dans sa conformation à $\Delta t = 2,5$ ms est représentée par la surface rouge. A gauche, cette ZIN est montrée dans une vue globale de la protéine faisant apparaître sa structure secondaire, tandis que la vue de droite est centrée sur la ZIN. Les lignes de champ électrique de cette surface recouvrent un volume de $96 \text{ \AA}^3$, plus petit qu'à $\Delta t = 1\mu\text{s}$ et contenu à l'intérieur de la protéine. Ces lignes de champ rejoignent les sites électrophiles (sphères bleues) appartenant à plusieurs résidus : Val57, Ser60, Ala61, Asn98, Trp99, Ala101, Thr102, Lys235 et le rétinale (RET).

un volume de $96 \text{ \AA}^3$ plus petit qu'à $\Delta t = 1\mu\text{s}$. Elles joignent toujours les sites électrophiles des résidus Ser60, Asn98, Ala101 et Thr102, ainsi que ceux appartenant à d'autres résidus : Val57, Ala61, Trp99, Lys235 et le rétinale (RET). Ce volume inclut notamment la position du site de fixation CL351 supposé fermé par la barrière moléculaire stérique, mais pas le site de fixation CL352 qui est vide à $\Delta t = 2,5$ ms. De plus, la ZIN montre maintenant une interaction entre l'Asn98 et le RPSB qui n'était pas présente dans la conformation précédente. Par conséquent, les changements de conformation des résidus de la protéine sont accompagnés par des modifications remarquables de la topographie des lignes de champ électrique dans le site CL351. L'Asn98, mais aussi du RPSB, interagissent avec des partenaires électrostatiques différents dans les deux structures. Ces observations suggèrent que la fermeture de cette barrière dite stérique [Mous *et al.*, 2022] pourrait aussi être assurée par des effets de nature électrostatique stabilisant la position de l'Asn98 dans ce site.

Caractérisation par calculs d'énergie d'interaction

Pour leur étude, [Mous *et al.*, 2022] ont réalisé des calculs d'énergie d'interaction par analyse de décomposition de l'énergie (EDA pour "Energy Decomposition Analysis") sur un sous-système modélisant le site de fixation CL352 à $\Delta t = 1\mu\text{s}$. Ce sous-système, représenté dans la figure 4.5, comprend le rétinale lié à la Lys235, l'ion chlorure et les résidus voisins Asn98, Thr102 et Asp231. Les résidus de la protéine ont été coupés pour ne garder que les chaînes latérales, entre

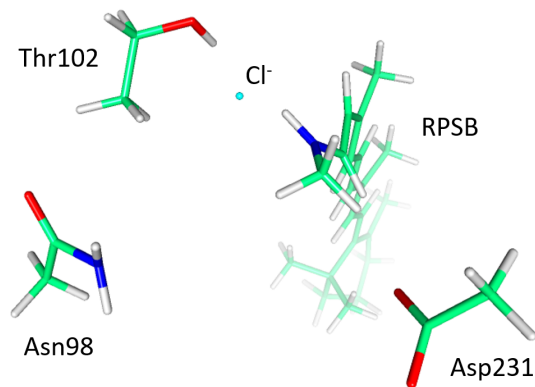


FIGURE 4.5 – Sous-système utilisé dans les calculs d'énergies QM pour modéliser le site de fixation CL352.

Le sous-système utilisé dans les calculs d'énergie d'interaction par méthode de chimie quantique [Mous *et al.*, 2022] a été extrait de la structure tridimensionnelle de l'halorhodopsine à $\Delta t = 1\mu s$. Il modélise le site de fixation CL352 et comprend l'ion chlorure Cl^- , La base de Schiff protonée (RPSB) liant le rétinale à la Lys235 et les chaînes latérales des résidus Asn98, Thr102 et Asp231.

les carbones $C\delta$ et $C\epsilon$ pour Lys235 et entre $C\alpha$ et $C\beta$ pour les autres. Les valeurs des énergies d'interaction électrostatique E_{elec} , de polarisation E_{pol} , de transfert de charge E_{CT} , de dispersion E_{disp} et de répulsion de Pauli E_{Pauli} de ce sous-système sont fournies en matériel supplémentaire de l'article [Mous *et al.*, 2022].

Ici, ces valeurs d'énergies issues de méthodes de chimie quantique de haut de niveau de théorie ont été utilisées pour évaluer la validité des potentiels d'interaction ELMAM dans un système protéique, ceux-ci étant en cours de développement comme décrit au chapitre 3. Pour cela, nous avons transféré les paramètres multipolaires utilisés dans la structure de l'halorhodopsine à $\Delta t = 1\mu s$ et les polarisabilités atomiques de la librairie ELMAM2 sur le sous-système modélisant le site de fixation CL352 (figure 4.5). Le terme électrostatique ELMAM, U_{elst} , qui est calculé par la méthode aEP/MM [Vuković *et al.*, 2021], peut être directement comparé aux valeurs théoriques E_{elec} . La contribution d'induction dipolaire $U_{ind-disp}$, définie comme la différence entre les énergies électrostatiques obtenues à partir des densités électroniques polarisée et non-polarisée [Leduc *et al.*, 2019], peut être utilisée pour estimer E_{pol} mais ce dernier prend également en compte les effets d'induction quadripolaire et d'ordres supérieurs. Le modèle de dispersion U_{disp} employé ici est basé sur les tenseurs de polarisabilité anisotropes, avec le facteur d'échelle $K_{disp} = 0,86$ et sans introduire de paramètres empiriques dépendant de l'espèce chimique, celui du chlore n'ayant pas pu être affiné (voir partie 3.2). Pour le terme d'échange-répulsion $U_{exch-rep}$, le modèle de recouvrement des densités électroniques a été utilisé avec la méthode d'intégration le long de l'axe internucléaire, la coupure de l'intervalle d'intégration $Z_{cut-off} = 0,6\text{\AA}$ et le facteur d'échelle $k_{exch-rep} = 2,00 \cdot 10^{-16} \text{ m}^4 \cdot \text{mol}^{-1}$, sans facteur dépendant de la distance interatomique et sans paramètres dépendant de l'espèce chimique (voir partie 3.3). Ces deux modèles développés dans mes travaux de thèse peuvent être comparés respectivement aux valeurs de E_{disp} et de E_{Pauli} . La contribution de transfert de charge E_{CT} n'est quant à elle pas estimée dans le modèle actuel du potentiel d'interaction ELMAM.

Energies QM	E_{elec}	E_{pol}	E_{CT}	E_{disp}	E_{Pauli}	E_{int}
Cl^- vs RPSB	-99,47	-17,92	-15,39	-5,75	52,33	-86,20
Cl^- vs RPSB, T102	-146,7	-22,11	-20,47	-10,67	107,0	-92,93
Cl^- vs RPSB, D231	-114,5	-30,03	-20,23	-7,63	64,00	-108,4
Cl^- vs RPSB, N98, T102, D231	-166,5	-34,88	-26,66	-14,26	125,8	-116,5
Energies ELMAM	U_{elst}	$U_{\text{ind-disp}}$	—	U_{disp}	$U_{\text{exch-rep}}$	U_{int}
Cl^- vs RPSB	-78,9	-32,7	—	-8,13	26,1	-93,7
Cl^- vs RPSB, T102	-108	-53,4	—	-13,8	58,1	-117
Cl^- vs RPSB, D231	-33,2	-36,5	—	-8,15	26,9	-51,0
Cl^- vs RPSB, N98, T102, D231	-59,2	-54,3	—	-14,9	58,1	-69,3

TABLEAU 4.1 – Comparaison des énergies d’interaction dans le site de fixation CL352 obtenues par méthodes de chimie quantique avec les résultats des potentiels d’interaction ELMAM.

Les valeurs théoriques (énergies QM) d’énergies d’interaction électrostatique E_{elec} , de polarisation E_{pol} , de transfert de charge E_{CT} , de dispersion E_{disp} , de répulsion de Pauli E_{Pauli} et totale E_{int} obtenues par [Mous *et al.*, 2022] sont présentées dans le haut de tableau (lignes 2 à 5). Les résultats des potentiels d’interaction ELMAM électrostatique U_{elst} , d’induction dipolaire $U_{\text{ind-disp}}$, de dispersion U_{disp} , d’échange-répulsion $U_{\text{exch-rep}}$ et total U_{int} sont donnés dans le bas du tableau (lignes 7 à 10). Plusieurs interactions sont évaluées : entre le chlorure (Cl^-) et la base de Schiff protonée du rétinol (RPSB) uniquement, entre Cl^- et RPSB avec la Thr102, entre Cl^- et le RPSB avec l’Asp231, et entre le Cl^- et le RPSB avec l’Asn98, la Thr102 et l’Asp231. Toutes les valeurs d’énergies sont données en kcal/mol.

Les valeurs de ces énergies d’interaction ont été calculées dans le sous-système modélisant la poche de fixation CL352, entre l’ion chlorure Cl^- et le RPSB uniquement, entre le Cl^- et le RPSB avec Thr102, entre le Cl^- et le RPSB avec Asp231 et entre le Cl^- et le RPSB avec Asn98, Thr102 et Asp231. L’ensemble des résultats est donné dans le tableau 4.1. En comparant les énergies ELMAM U_{elst} et $U_{\text{ind-disp}}$ aux références E_{elec} et E_{pol} , d’importants écarts sont observés. La contribution électrostatique est sous-estimée en valeur absolue par le modèle ELMAM, avec une erreur relative allant de 21% pour l’interaction entre le Cl^- et le RPSB, jusqu’à 71% pour l’interaction entre le Cl^- et le RPSB avec Asp231. Au contraire, la contribution d’induction est surestimée en valeur absolue, de 22% pour l’interaction entre le Cl^- et le RPSB avec Asp231, et de 142% pour l’interaction entre le Cl^- et le RPSB avec Thr102. La somme des termes $U_{\text{elst}} + U_{\text{ind-disp}}$ reproduit mieux la somme $E_{\text{elec}} + E_{\text{pol}}$ que les contributions individuelles, avec des écarts relatifs de seulement 5% pour les interactions entre le Cl^- et le RPSB et entre le Cl^- et le RPSB avec Thr102. Il est donc possible que les termes électrostatique et d’induction ne soit pas distingués de la même manière entre la méthode EDA et le potentiel ELMAM. Par exemple, l’interaction entre dipôles permanents qui est incluse dans le terme ELMAM U_{elst} appartient peut-être au terme quantique E_{pol} . En revanche, pour les interactions comprenant le résidu Asp231, l’erreur sur cette somme reste très élevée, de l’ordre de +80 kcal/mol. En effet, la répulsion électrostatique entre les charges négatives de Cl^- et de l’Asp231 est surestimée par le modèle ELMAM car celle-ci est calculée en utilisant la permittivité diélectrique du vide ϵ_0 tandis qu’en réalité le rétinol situé entre les deux charges écrante cette interaction déstabilisante.

Pour la contribution de dispersion, le modèle ELMAM reproduit les valeurs de référence avec des erreurs absolues allant de 0,3 kcal/mol pour l’interaction entre le Cl^- et le RPSB avec Asn98, Thr102 et Asp231, à 3,2 kcal/mol pour l’interaction entre Cl^- et le RPSB avec Thr102. En revanche, pour le terme d’échange-répulsion, l’accord entre les résultats ELMAM

Energies ELMAM	U_{elst}	$U_{\text{ind-dip}}$	U_{disp}	$U_{\text{exch-rep}}$	U_{int}
$\Delta t = 1\mu\text{s}$	-8,37	-6,16	-5,65	9,72	-10,5
$\Delta t = 2,5\text{ms}$	-27,9	-19,5	-13,4	32,0	-28,9

TABLEAU 4.2 – Energies d'interaction entre l'Asn98 et la Thr102 avec la Lys235 et le rétinale à partir des potentiels ELMAM.

Les résultats des potentiels d'interaction ELMAM électrostatique U_{elst} , d'induction dipolaire $U_{\text{ind-dip}}$, de dispersion U_{disp} , d'échange-répulsion U_{exch} et total U_{int} entre le résidu Asn98 et les résidus Thr102, Lys235 et RET (rétinal) sont donnés en kcal/mol. Ces valeurs ont été calculées dans la structure de l'halorhodopsine à $\Delta t = 1\mu\text{s}$ (conformation ouverte de la barrière moléculaire stérique) et à $\Delta t = 2,5\text{ms}$ (conformation fermée de la barrière moléculaire stérique). Les valeurs d'énergie de ce tableau sont données en kcal/mol.

et les énergies d'origine quantique est mauvais pour toutes les interactions, avec des erreurs relatives de l'ordre de 50%. Ces erreurs peuvent être expliquées par le choix de modélisation de l'ion chlorure avec une densité électronique à symétrie sphérique qui pourrait induire une mauvaise estimation du recouvrement des densités. Enfin, la contribution de transfert de charge s'avère significative dans ces interactions et pourrait faire l'objet d'une modélisation future dans le modèle ELMAM. Cette analyse a donc permis de mettre en lumière plusieurs perspectives d'amélioration du potentiel d'interaction ELMAM pour rendre les estimations plus proches des valeurs théoriques dans les protéines. Néanmoins, contrairement aux méthodes quantiques, le modèle ELMAM a l'avantage de ne pas être restreint à une application à un sous-système mais peut être appliqué directement dans la protéine en restant rapide en temps de calcul.

Nous avons également appliqué les modèles énergies d'interaction ELMAM dans la barrière moléculaire stérique, le rôle de l'Asn98 n'ayant pas été caractérisé par les calculs de chimie quantique. En particulier, les différentes contributions à l'énergie d'interaction entre le résidu Asn98 et les résidus Thr102, Lys235 et RET (rétinal) ont été calculées dans la conformation ouverte de la barrière à $\Delta t = 1\mu\text{s}$ et dans sa conformation fermée à $\Delta t = 2,5\text{ms}$. Les résultats obtenus sont regroupés dans le tableau 4.2. Toutes les contributions sont plus élevées (en valeurs absolues) dans la conformation fermée par rapport à la conformation ouverte, ce qui confirme une interaction plus forte de l'Asn98 avec ces résidus dans le site de fixation CL351. De plus, les composantes électrostatique et d'induction sont plus stabilisantes que la contribution de dispersion, appuyant une nature plutôt électrostatique de cette interaction. Les développements futurs des potentiels d'interaction ELMAM permettraient de confirmer (ou non) cette conclusion mais cet exemple de calcul permet déjà d'illustrer une utilisation possible de ces modèles pour l'étude des protéines.

4.4.3 Barrière moléculaire électrostatique

Mécanisme proposé par [Mous *et al.*, 2022]

Après son entrée dans la protéine, le chlorure diffuserait jusqu'à la barrière moléculaire électrostatique formée par le pont salin entre les résidus Arg95 et Asp231. L'interaction entre la charge négative de l'ion et les charges de ces deux résidus entraînerait le changement de conformation de la chaîne latérale de l'Asp231 observé entre $\Delta t = 12,5\text{ms}$ et $\Delta t = 22,5\text{ms}$. L'Asp231 interagissant alors avec son voisin His29 plutôt qu'avec Arg95 permettrait l'ouverture

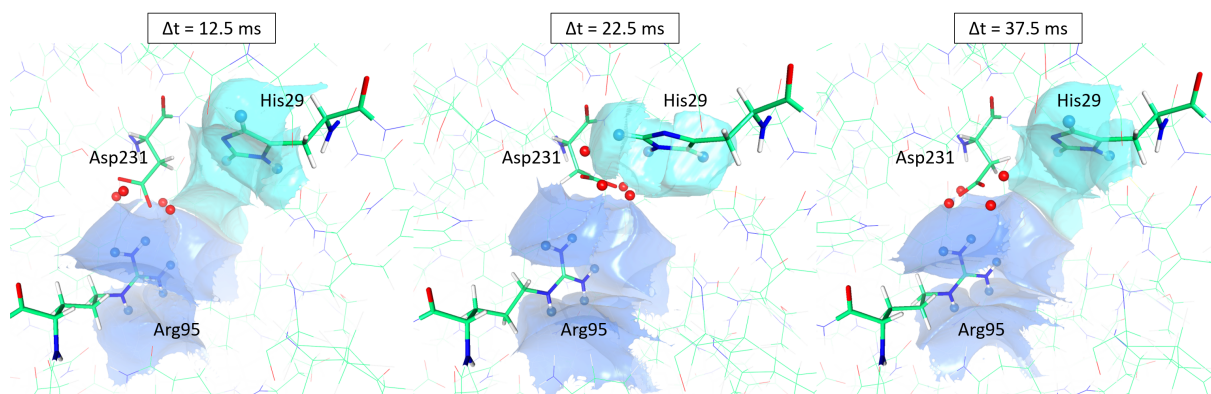


FIGURE 4.6 – Zones d'influence électrophile dans la barrière moléculaire électrostatique.

Les zones d'influence électrophile (ZIE) associées aux atomes $H\epsilon$, $H\eta11$, $H\eta12$, $H\eta21$ et $H\eta22$ du groupement guanidinium de l'Arg95 sont représentées par des surfaces bleu foncé et les ZIE associées aux atomes $H\delta2$, $H\epsilon1$ et $H\delta1$ (ou $H\epsilon2$ à $\Delta t = 22,5\text{ms}$) du groupement imidazole de l'His29 par des surfaces bleu clair. Elles ont été calculées dans les structures de l'halorhodopsine à $\Delta t = 12,5\text{ms}$, $22,5\text{ms}$ et $37,5\text{ms}$. Les lignes de champ électrique contenues dans ces ZIE émanent des sites électrophiles localisés par des sphères bleues et certaines convergent vers les sites nucléophiles (sphères rouges) correspondant aux doublets non-liants des atomes d'oxygène du groupement carboxyle de l'Asp231 dans les trois structures.

de la barrière moléculaire. L'anion serait alors capable de passer cette barrière avant de se diriger vers son site de fixation initial CL351. A $\Delta t = 37,5\text{ms}$, l'ion est effectivement localisé à proximité du site CL351 et le pont salin entre Arg95 et Asp231 est rétabli, refermant la barrière pour empêcher la fuite du chlorure vers le solvant (voir figure 4.6).

Caractérisation par les zones d'influence dans la barrière moléculaire stérique

Le mécanisme de la barrière électrostatique moléculaire repose sur les interactions entre les résidus Arg95, Asp231 et His29. Nous avons donc calculé les zones d'influence électrophile (ZIE) des protons des groupements guanidinium d'Arg95 et imidazole d'His29, ainsi que les zones d'influence nucléophile (ZIN) des atomes d'oxygène du groupement carboxylate d'Asp231 à $\Delta t = 12,5\text{ms}$, $22,5\text{ms}$ et $37,5\text{ms}$. Les ZIE de l'Arg95, représentées en bleu foncé sur la figure 4.6, occupent des régions quasiment identiques dans les trois structures, avec un volume variant de $130\text{\AA}^3$ à $\Delta t = 12,5\text{ms}$, $162\text{\AA}^3$ à $\Delta t = 22,5\text{ms}$ et $141\text{\AA}^3$ à $\Delta t = 37,5\text{ms}$. Les partenaires électrostatiques d'Arg95 vers lesquels convergent les lignes de champ électrique contenues dans ces volumes, dont notamment les oxygènes carboxylates de l'Asp231, ne changent pas dans la conformation ouverte de la barrière moléculaire à $\Delta t = 22,5\text{ms}$.

Pour les ZIE de l'His29, représentées en bleu clair dans la figure 4.6, l'ouverture de la barrière change la topographie des lignes de champ électrique suite au retournement du cycle imidazole et du passage du proton sur l'atome d'azote $N\delta1$ en conformation fermée vers l'atome d'azote $N\epsilon2$ en conformation ouverte. Lorsque la barrière est fermée, à $\Delta t = 12,5\text{ms}$ et $37,5\text{ms}$, seule la ZIE de l'hydrogène $H\epsilon1$ rejoint l'Asp231, et s'étend également vers l'atome d'oxygène du résidu Leu21. Les lignes de champ électrique émergeant de l'atome $H\delta2$ convergent quant à elles vers les oxygènes des résidus His29, Val57 et Met58, et celles du proton de l'azote $N\delta1$ vers l'atome Gly25-O. Ces ZIE occupent un volume total de $85\text{\AA}^3$ à $\Delta t = 12,5\text{ms}$ et de $69\text{\AA}^3$ à $37,5\text{ms}$. Lorsque la barrière moléculaire est ouverte, à $\Delta t = 22,5\text{ms}$, les trois ZIE se concentrent sur

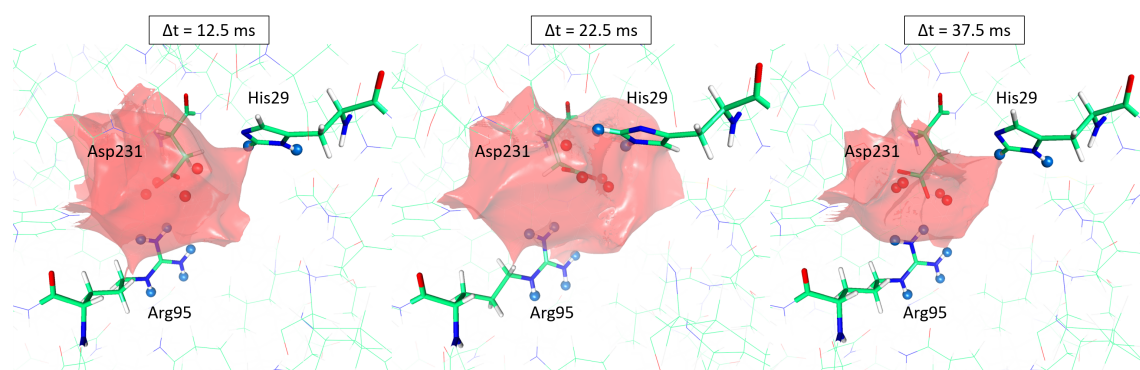


FIGURE 4.7 – Zones d'influence nucléophile dans la barrière moléculaire électrostatique.

Les zones d'influence nucléophile (ZIN) associées aux doublets non-liants des atomes O δ 1 et O δ 2 du résidu Asp231 sont représentées par des surfaces rouges. Ces ZIN ont été calculées dans les structures de l'halorhodopsine à $\Delta t = 12,5\text{ms}$, $22,5\text{ms}$ et $37,5\text{ms}$. Parmi les lignes de champ électrique à l'intérieur de ces surfaces, certaines émanent des sites électrophiles (sphères bleues) des résidus Arg95 et His29 avant de converger vers les sites nucléophiles (sphères rouges) de l'Asp231 dans les trois structures.

l'Asp231 et leur volume total est réduit à $30\text{\AA}^3$. Finalement, il existe dans les trois structures étudiées ici des lignes de champ électrique émanant à la fois de Arg95 et de His29 qui convergent vers Asp231, dans les conformations ouverte et fermée de la barrière moléculaire.

La représentation des ZIN des atomes O δ 1 et O δ 2 l'Asp231, par des surfaces rouges dans la figure 4.7, est complémentaire à celle des ZIE. Le volume total de ces ZIN est de $142\text{\AA}^3$ à $\Delta t = 12,5\text{ms}$, $265\text{\AA}^3$ à $\Delta t = 22,5\text{ms}$ et $300\text{\AA}^3$ à $\Delta t = 37,5\text{ms}$. L'augmentation de ce volume, qui n'apparaît pas de manière évidente depuis l'angle de vue choisi pour l'image 4.7, s'explique par une extension des ZIN vers des contributeurs plus éloignés, situés derrière le plan de cette image. Dans les trois structures, les ZIN confirment la présence d'interactions électrostatiques entre l'Asp231 et à la fois l'Arg95 et l'His29, pour les conformations ouverte et fermée de la barrière moléculaire. Contrairement au cas de la barrière moléculaire stérique, dans cette barrière moléculaire électrostatique les changements de conformations des résidus n'induisent pas d'importantes modifications dans la topographie des lignes de champ électrique. Cette observation suggère que le passage de l'ion chlorure par cette barrière est peut-être également assuré par d'autres phénomènes que les effets électrostatiques. Par exemple, l'écartement entre les chaînes latérales de l'Arg95 et de l'Asp231 à $\Delta t = 22,5\text{ms}$ pourrait réduire l'encombrement stérique de manière à laisser passer l'ion par diffusion.

En résumé,

l'application des calculs de zones d'influence et d'énergies d'interaction ELMAM aux protéines possède un pouvoir d'analyse de la dynamique des mécanismes structuraux prometteur, comme illustré ici dans le cas des deux barrières moléculaires de l'halorhodopsine. En effet, ces méthodologies ont déjà permis de décrire la conservation des contributeurs aux forces électrostatiques et les directions de celles-ci lors de changements conformationnels et de discuter la nature électrostatique ou non des phénomènes influant sur les ouvertures et fermetures de ces barrières. Dans la perspective de montrer l'utilisation de nos descripteurs pour décrire d'autres

effets dynamiques, l’analyse des autres étapes du transport de l’ion chlorure est en cours de réalisation. Par exemple, les résidus appartenant à la région chargée positivement de la surface extra-cellulaire de la protéine responsable de l’entrée de l’ion pourront être identifiés grâce à la visualisation des zones d’influence électrophile s’étendant vers le solvant. L’interaction entre la charge de l’anion et les électrons π du chromophore participant au mécanisme d’entraînement de l’ion du site de fixation CL351 vers le site CL352 pourra également être caractérisée par le calcul de la variation des énergies d’interaction chlorure – rétinale entre ces deux configurations.

4.5 Conclusion partielle de chapitre

Dans ce chapitre, les descripteurs issus de la topologie du potentiel électrostatique et les modèles d’énergie d’interaction ELMAM développés dans mes travaux thèse ont été appliqués à différents systèmes protéiques. Pour montrer différentes utilisations des descripteurs électrostatiques [Mocchetti *et al.*, 2023], notamment des zones d’influence électrophile et nucléophile, nous avons analysé la trypsine, une protéase à sérine dont le mécanisme est bien connu. En effet, grâce à la visualisation des surfaces correspondantes, les atomes ou groupements d’atomes participant à la fixation d’un substrat sont révélés, même lorsqu’ils sont situés à longue distance et ne sont donc pas caractérisés par des analyses classiques. Ces descripteurs permettent aussi de prédire les résidus pouvant diriger l’approche du substrat à travers le solvant avant de le fixer et donc de proposer des mutations potentielles pour améliorer l’affinité électrostatique du complexe. Ils peuvent par ailleurs être employés pour discuter la description d’un mécanisme enzymatique. Par conséquent, nos descripteurs électrostatiques proposent un nouveau point de vue pour caractériser les systèmes protéiques grâce auquel de nouvelles informations sont apportées dans les études de biologie structurale. C’est ce que nous avons montré dans l’analyse du complexe de la neuropiline 1 avec un ligand peptidique [Goudiaby *et al.*, 2023]. De même, les contributions de chaque résidu à la fixation du ligand peuvent être quantifiées et classifiées grâce aux modèles d’énergie d’interaction ELMAM et la nature de ces interactions peut être discutée. Nous avons notamment appliqué les calculs d’énergie d’interaction électrostatique à l’enzyme SynGSTC1 dans ce but [Mocchetti *et al.*, 2022].

Dans la perspective d’appliquer les descripteurs électrostatiques et les potentiels d’interaction ELMAM aux systèmes dynamiques, nous avons entrepris l’analyse du transport de l’ion chlorure par l’halorhodopsine NmHR. La description des deux barrières moléculaires identifiées par l’étude de [Mous *et al.*, 2022] a été réalisée du point de vue de nos méthodologies. L’étude des zones d’influence dans la barrière dite stérique a montré d’importantes modifications dans la topographie des lignes de champ électrique en réponse aux changements conformationnels, suggérant un possible rôle d’interactions de nature électrostatique dans la fermeture cette barrière. Cette idée est renforcée par la contribution majoritaire du terme électrostatique dans les énergies d’interaction calculées en utilisant les potentiels ELMAM. Au contraire, dans la barrière moléculaire dite électrostatique, les lignes de champ ne sont quasiment pas impactées par les changements de conformation des résidus Arg95, Asn231 et His29, suggérant la contribution d’autres phénomènes de type stérique ou diffusif par exemple pour assister le passage de l’ion. L’analyse des autres étapes de transport du chlorure permettrait de caractériser l’utilisation de nos approches pour d’autres aspects dynamiques, comme l’entrée de l’ion dans la protéine

depuis le milieu extra-cellulaire ou encore sa diffusion assistée par le moment dipolaire global de l'halorhodopsine.

En bref, les descripteurs issus du potentiel électrostatique et les potentiels d'interaction EL-MAM apportent des informations supplémentaires pour la caractérisation des protéines en complément des analyses structurales classiques. Leur implémentation dans le logiciel MoProViewer les rend pratiques à mettre en œuvre pour la communauté de biologie structurale, contrairement à d'autres outils théoriques tels que les méthodes QM/MM et de dynamique moléculaire qui nécessitent une expertise particulière.

Chapitre 5

Conclusion globale et perspectives

Dans l’objectif d’étendre les approches de cristallographie quantique aux études de biologie structurale, de nouveaux descripteurs électrostatiques applicables aux macromolécules biologiques ont été développés. En effet, grâce aux bases de données telles que la librairie ELMAM2 [Domagała *et al.*, 2012], le transfert des paramètres du modèle multipolaire de Hansen et Coppens [Hansen et Coppens, 1978] permet de reconstruire une densité électronique de protéine faisant apparaître des détails fins tels que les électrons des liaisons covalentes et des doublets non-liants. Les deux types de méthodologie développés dans cette thèse reposent sur cette description précise de la distribution électronique moléculaire. Le premier objectif était de définir des descripteurs issus de la topologie du potentiel électrostatique et de montrer leur intérêt pour l’analyse des systèmes protéiques. Le second consistait en la construction d’un potentiel d’interaction de van der Waals, comprenant les modèles d’énergies de dispersion et d’échange-répulsion, basé sur les quantités transférables de la librairie ELMAM2. Ce chapitre final propose un retour sur les développements et applications réalisés pour ces deux approches ainsi que les perspectives de ce travail.

Le potentiel électrostatique $V(\mathbf{r})$ est généralement employé en biologie structurale pour colorer la surface d’accessibilité au solvant entourant une protéine ou son ligand [Weiner *et al.*, 1982]. Ce type de représentation met en évidence les complémentarités électrostatiques dans un complexe mais restreint la description à une surface moléculaire. En réalité, le champ scalaire $V(\mathbf{r})$, défini dans $\mathbb{R}^3$, est riche en informations sur l’étendue et la directionnalité des interactions électrostatiques. L’analyse topologique de ce champ scalaire propose quant à elle une appréhension tridimensionnelle des forces électrostatiques, reposant sur la répartition spatiale des lignes de champ électrique. Des méthodologies basées sur la topologie du potentiel électrostatique avaient déjà été explorées pour la caractérisation de petites molécules [Mata *et al.*, 2007, Mohan *et al.*, 2013, Mata *et al.*, 2015, Kumar et Gadre, 2016, Alkorta *et al.*, 2019] mais pas dans les macromolécules.

Pour l’appliquer aux protéines, j’ai développé des descripteurs graphiques issus de cette approche reposant sur la définition des points critiques et des bassins topologiques du potentiel électrostatique. Les points critiques de $V(\mathbf{r})$ sont définis comme les points où le vecteur champ électrique $\mathbf{E}(\mathbf{r})$ est nul. Quatre types de points critiques sont distingués : les maxima locaux $(3, -3)$, les minima locaux $(3, +3)$ et les points-selles $(3, -1)$ et $(3, +1)$. Les maxima $(3, -3)$ sont

localisés sur les noyaux des atomes et les minima $(3, +3)$ correspondent à des concentrations locales d'électrons pouvant indiquer la présence d'un doublet non-liant. Ces maxima et minima locaux révèlent les positions des sites électrophiles et des sites nucléophiles respectivement. Les points-selles $(3, -1)$ caractérisent les interactions interatomiques covalentes ou non et les $(3, +1)$, les cycles moléculaires. Ces points-selles apparaissent également dans l'espace moléculaire sans interprétation clairement définie en termes de propriétés moléculaires. Pour déterminer les positions des points critiques, j'ai employé la méthode définie par [Balanarayan et Gadre, 2003] permettant de les estimer puis l'algorithme de minimisation de Newton-Raphson pour les affiner, comme décrit dans la partie 2.1.

Les lignes de champ électrique se regroupent sous la forme de faisceaux émergeant d'un site électrophile (point critique $(3, -3)$) et convergeant vers un site nucléophile (point critique $(3, +3)$). Un faisceau primaire de lignes de champ est associé à un unique site électrophile et un unique site nucléophile et constitue un bassin topologique du potentiel électrostatique. L'union des faisceaux primaires associés à un même site électrophile forme la zone d'influence électrophile (ZIE) de l'atome auquel appartient ce site. De même, l'union des faisceaux primaires associés à un même site nucléophile forme la zone d'influence nucléophile (ZIN) de l'atome auquel appartient ce site. Dans le but de déterminer les surfaces entourant les faisceaux primaires et les zones d'influence électrophile et nucléophile, j'ai développé une nouvelle méthode, détaillée dans la partie 2.2, reposant sur la vérification des conditions aux extrémités des lignes de champ électrique à l'intérieur de ces surfaces. Ces dernières étant par définition des surfaces de flux nul du potentiel électrostatique, l'application du théorème de Gauss permet de vérifier les charges contenues dans les volumes correspondant. Le test des ZIE du complexe N-méthylacétamide - méthanol dans la partie 2.3 a révélé des erreurs absolues sur la charge électronique de l'ordre de $\pm 0,10e$ et des erreurs relatives allant de 1% à 10% selon les atomes considérés, et globalement une erreur de $0,12e$ sur tout le système, soit 0,02% de la charge électronique totale. Ces résultats permettent de conforter la méthode développée ici pour déterminer les contours des faisceaux primaires et des zones d'influence.

L'implémentation de ces méthodes dans le logiciel MoProViewer, discutée dans la partie 2.4, permet de mettre en application facilement nos descripteurs graphiques dans les systèmes protéine-ligand. Grâce à ces outils, nous avons pu montrer différentes utilisations des approches basées sur la topologie du potentiel électrostatique pour l'étude des protéines. Pour cela, nous avons analysé les systèmes décrits au chapitre 4 : la trypsine, une protéase à sérine dont le mécanisme enzymatique est bien connu (partie 4.1), la neuropiline 1, une protéine membranaire impliquée dans de nombreux processus biologiques mais aussi dans le développement de certains cancers (partie 4.2), et l'halorhodopsine, une protéine de transport membranaire d'ions chlorure (partie 4.4). Ces applications ont révélé la capacité de nos descripteurs à mettre en évidence les atomes et groupements d'atomes de la protéine impliqués dans les forces électrostatiques stabilisant la fixation du ligand. Grâce à la visualisation de l'étendue et de la direction de ces forces, les contributeurs situés à longue distance sont également identifiés et les rôles de différents résidus peuvent être comparés pour déterminer le principal acteur d'une interaction. Les résidus capables de diriger l'approche d'un ligand sont prédits par la mise en évidence de leur influence s'étendant au-delà de la surface de la protéine. De plus, les mécanismes enzyma-

tiques peuvent être discutés sur la base de la répartition des lignes de champ électrique dans le site actif, celles-ci étant à l'origine des forces coulombiennes ressenties par les résidus impliqués dans ces mécanismes mais aussi des phénomènes de polarisation enzyme-substrat. Les mécanismes influençant les dynamiques structurales peuvent également être décrits en caractérisant les modifications de partenaires électrostatiques et de directions des forces faisant suite aux changements de conformation des résidus de la protéine.

Au-delà de la détermination des liaisons hydrogène par contacts locaux entre résidus et de la caractérisation des régions chargées en surface, les descripteurs issus de la topologie du potentiel électrostatique apportent des informations nouvelles dans les études de biologie structurale par rapport aux méthodes classiques. Ces approches originales, analysables et interprétables grâce à la topographie des lignes de champ électrique, proposent un nouveau paradigme considérant le potentiel électrostatique d'une protéine dans son ensemble et en trois dimensions. Contrairement aux surfaces colorées par le potentiel qui sont typiquement représentées à l'échelle des résidus, voire de la protéine dans son ensemble, les contributeurs aux interactions électrostatiques sont déterminés à l'échelle de l'atome par ces descripteurs. Ceux-ci sont pratiques à mettre en œuvre grâce à leur implémentation dans MoProViewer, dès lors qu'un potentiel électrostatique est disponible sur une grille 3D, et ne requièrent pas une expertise particulière pour les utiliser contrairement à d'autres outils avancés tels que les méthodes de QM/MM et de dynamique moléculaire.

Pour rendre les descripteurs graphiques encore plus facilement interprétables, plusieurs développements sont prévus dont l'affichage des directions moyennes des lignes de champ électrique dans les faisceaux primaires. La parallélisation de l'algorithme permettrait de réduire le temps d'exécution des calculs des surfaces entourant les faisceaux primaires et les zones d'influence. Des informations quantitatives pourraient être apportées par la coloration de ces surfaces en fonction de la valeur du potentiel électrostatique ou de la magnitude du champ électrique. L'évaluation de l'énergie électrostatique interne $E = 1/2\epsilon_0 \int |\mathbf{E}|^2 dV$ par intégration du champ électrique dans un faisceau primaire est également envisagée. D'autres outils pourraient rendre la description plus intuitive comme la programmation d'une sonde réactive qui afficherait la zone d'influence intersectant la position courante tout en se déplaçant dans l'espace moléculaire.

Une autre perspective importante est l'application de ces descripteurs à des systèmes dynamiques. L'analyse du mécanisme de pompage de l'ion chlorure par l'halorhodopsine que nous avons entreprise a déjà permis d'aborder l'influence des changements structuraux sur les phénomènes électrostatiques qui pourrait être discutée plus en détails avec la poursuite de cette étude sur l'ensemble du cycle de transport. Pour aller plus loin, la caractérisation de l'évolution temporelle des zones d'influence le long de trajectoires de dynamique moléculaire est envisagée. Notamment, le calcul de moyennes d'ensembles statistiques sur les volumes correspondant permettrait de discuter de leur conservation et donc du caractère essentiel d'interactions spécifiques pour un mécanisme. De même, il serait intéressant d'analyser le lien entre la conservation de résidus et la conservation de l'environnement électrostatique, par exemple dans le site actif d'enzymes d'une même famille. L'existence d'une signature électrostatique pourrait notamment être investiguée. Les effets de mutations ponctuelles de résidus sur la topographie des lignes de champ électrique pourraient également être caractérisés. La perspective d'application des zones

d’influence pour la description de mécanismes enzymatiques encore non-résolus est également considérée, en particulier dans le cas de l’activation du glutathion par l’enzyme SynGSTC1 mentionnée dans la partie 4.3. Par ailleurs, il serait intéressant de vérifier l’impact de l’utilisation de potentiels électrostatiques d’origines différentes, à partir de charges ponctuelles ou par résolution de l’équation de Poisson-Boltzmann par exemple, sur la définition des surfaces entourant les zones d’influence et les faisceaux primaires.

Le second objectif de mes travaux de thèse consistait à définir un modèle d’énergie d’interaction basé sur les paramètres de densité électronique et les tenseurs de polarisabilité atomique transférables de librairie ELMAM2 [Domagała *et al.*, 2012]. L’énergie d’interaction intermoléculaire est généralement décomposée en quatre contributions : énergie électrostatique coulombienne E_{elst} , énergie d’induction E_{ind} (ou de polarisation), énergie de dispersion E_{disp} et énergie d’échange-répulsion $E_{\text{exch-rep}}$ (ou simplement de répulsion). Les potentiels interatomiques classiques des champs de forces proposent des modèles paramétrés pour estimer ces contributions à faible coût computationnel mais leur caractère empirique ainsi que les approximations utilisées pour leur définition conditionnent la transférabilité de ces modèles d’un système à l’autre. Les méthodes *ab initio* de chimie quantique fournissent quant à elles des résultats d’une grande précision mais sont trop coûteuses en temps de calcul pour être appliquées sur une structure de macromolécule dans son ensemble. Les approches hybrides de QM/MM permettent d’étudier de larges systèmes en tirant avantage de la précision des méthodes quantiques pour décrire la région chimiquement active et des méthodes classiques pour traiter le reste du système à coût computationnel moindre. Néanmoins, ces approches restent difficiles à mettre en œuvre, nécessitent toujours des moyens computationnels importants et demandent une expertise particulière pour être employées.

Le potentiel d’interaction total ELMAM repose quant à lui sur la description précise de la distribution électronique moléculaire sur la base de données cristallographiques, en introduisant un minimum de paramètres empiriques, tout en restant simple à utiliser et performant en temps de calcul. Pour cela, les évaluations des contributions électrostatique $U_{\text{elst}}^{\text{ELMAM}}$ et d’induction dipolaire $U_{\text{ind-dip}}^{\text{ELMAM}}$ de ce potentiel avaient déjà été établies dans de précédents travaux [Leduc *et al.*, 2019, Vuković *et al.*, 2021]. Un modèle d’énergie de van der Waals, composé des termes de dispersion et d’échange-répulsion, a donc été développé pour compléter le potentiel d’interaction total ELMAM. Ces deux contributions ont été optimisées séparément et testées par comparaison aux valeurs de référence d’énergie d’interaction SAPT dans le jeu de données de benchmarking de petites molécules organiques NENCI-2021 [Sparrow *et al.*, 2021].

Le potentiel de dispersion $U_{\text{disp}}^{\text{ELMAM}}$, détaillé dans la partie 3.2, a été construit sur la base de l’approximation de London utilisant les tenseurs de polarisabilités anisotropes atomiques transférés. Le facteur de détermination des résultats de ce modèle par rapport aux valeurs SAPT était $R^2 = 0,946$ avec une erreur absolue moyenne de 0,44 kcal/mol. L’introduction de paramètres empiriques dépendant de l’espèce chimique a légèrement augmenté la qualité de ces résultats avec $R^2 = 0,953$ et une erreur absolue moyenne de 0,39 kcal/mol. L’analyse de l’influence de la composition du dimère considéré sur les résultats du modèle ELMAM a révélé que les systèmes dominés par des interactions de type dispersif présentent de meilleures valeurs

de R^2 que ceux dominés par des interactions électrostatiques coulombiennes.

Le potentiel d'échange-répulsion $U_{\text{exch-rep}}^{\text{ELMAM}}$, décrit dans la partie 3.3, a été défini à partir du modèle de recouvrement des densités électroniques, basé sur les paramètres multipolaires transférés de la librairie ELMAM2. Le recouvrement des densités correspondant à l'intégrale du produit de deux densités électroniques atomiques, plusieurs méthodes de calcul ont été testées. Par intégration le long de l'axe internucléaire avec une longueur d'intervalle optimisée, le coefficient de détermination $R^2 = 0,957$ et l'erreur absolue de 1,10 kcal/mol ont été obtenus. Ce modèle ne repose sur aucun paramètre empirique, excepté un facteur d'échelle nécessaire pour prendre en compte la différence de dimension entre l'énergie et le recouvrement des densités. Par ailleurs, l'ajout dans le modèle d'un facteur dépendant de la distance interatomique, qui avait été proposé dans de précédentes études [Söderhjelm *et al.*, 2006, Misquitta et Stone, 2016, Rackers et Ponder, 2019], n'a pas amélioré davantage ces résultats. L'introduction de paramètres empiriques dépendant de l'espèce chimique a par contre permis d'obtenir $R^2 = 0,969$ et une erreur absolue moyenne de 0,89 kcal/mol. L'analyse de l'influence de la composition des dimères étudiés sur les résultats du modèle n'a pas révélé de dépendance particulière.

Le potentiel de van der Waals ELMAM, discuté dans la partie 3.4, est défini comme la somme des potentiels de dispersion et d'échange-répulsion ELMAM. Sans paramètre empirique, les résultats obtenus sont corrects, avec $R^2 = 0,915$ et une erreur absolue moyenne de 1,21 kcal/mol, mais ne sont pas aussi bons que pour les contributions individuelles. En considérant les potentiels de dispersion et d'échange-répulsion avec les paramètres empiriques dépendant de l'espèce chimique, la corrélation avec les valeurs de référence avec $R^2 = 0,932$ et une erreur absolue moyenne de 1,00 kcal/mol. Ces résultats restant inférieurs à ceux des contributions prises séparément, des coefficients de combinaison linéaire de ces deux termes ont été affinés. Ils permettent d'améliorer les résultats du modèle sans paramètre, avec $R^2 = 0,925$ et une erreur absolue moyenne de 1,10 kcal/mol, mais pas ceux avec paramètres empiriques. L'étude de l'influence de la composition du dimère considéré sur le potentiel de van der Waals a permis de mettre en évidence de meilleurs facteurs R^2 pour les systèmes dominés par des interactions électrostatiques pour lesquels la contribution d'échange-répulsion est majoritaire. En revanche, pour les systèmes dominés par des effets dispersifs, les termes de dispersion et d'échange-répulsion sont du même ordre de grandeur mais, puisque leurs signes sont opposés, leur somme est proche de l'erreur absolue moyenne, causant une décorrélation avec les valeurs d'énergie de référence. Pour le potentiel d'interaction total ELMAM, c'est-à-dire la somme des quatre contributions électrostatique, d'induction dipolaire, de dispersion et d'échange-répulsion, un effet similaire est observé. De même que pour les systèmes dominés par des interactions dispersives, les systèmes dominés par des effets coulombiens présentent des énergies d'interaction électrostatique et d'induction dont la somme est du même ordre que la contribution d'échange-répulsion en valeurs absolues. Par conséquent, le potentiel d'interaction total prend de faibles valeurs, s'approchant des erreurs absolues moyennes sur les différentes contributions, ce qui réduit considérablement la corrélation entre les résultats du modèle ELMAM et les valeurs de référence. Néanmoins, les résultats individuels de ces modèles permettent d'estimer les différentes contributions à l'énergie et de les comparer entre elles de façon à pouvoir discuter la nature des interactions dominantes dans un système.

Dans les complexes protéine-ligand, les contributions de chaque résidu à la fixation du ligand peuvent être quantifiées et classifiées grâce aux modèles d'énergie d'interaction ELMAM. Nous avons notamment appliqué les calculs d'énergie d'interaction électrostatique à l'enzyme SynGSTC1, comme décrit dans la partie 4.3, afin de mettre en évidence les points d'ancrage principaux du ligand glutathion dans sa poche de fixation. Dans la perspective d'étude de la dynamique du mécanisme de transport de l'halorhodopsine présentée dans la partie 4.4, les quatre contributions du potentiel d'interaction ELMAM ont été calculées dans un site de fixation de l'ion chlorure et comparées aux valeurs obtenues par méthodes de chimie quantique. Les potentiels ELMAM fournissent des estimations correctes de ces contributions bien qu'ils soient conditionnés par la disponibilité de paramètres multipolaires pour tout le système, ce qui ne fut pas le cas pour l'ion chlorure. Néanmoins, notre modèle a l'avantage de pouvoir être appliqué sur le système réel, comprenant l'ensemble de la protéine, tandis que les méthodes quantiques sont restreintes à un sous-système de quelques dizaines d'atomes pour fournir des résultats dans un délais raisonnable. De plus, les estimations des potentiels ELMAM ont permis par comparaison des différentes contributions d'appuyer l'hypothèse de la nature électrostatique du mécanisme de la barrière moléculaire dite stérique de l'halorhodopsine.

Ces premières applications à des systèmes protéiques ont illustré le potentiel prometteur des modèles d'énergie d'interaction ELMAM pour décrire la force et la nature des interactions protéine-ligand, bien que ceux-ci soient encore en cours de développement pour améliorer leur pouvoir prédictif d'énergies de référence calculées théoriquement. Notamment, pour le potentiel de dispersion ELMAM, une variante plus avancée de l'approximation de London serait intéressante à considérer : le modèle de Slater-Kirkwood qui fait intervenir la notion de nombre d'électrons effectif d'un atome qu'il serait possible d'estimer à partir du paramètre de population de valence du modèle multipolaire. Pour le potentiel d'échange-répulsion qui semble essentiellement dépendre de la méthode d'intégration choisie pour obtenir le recouvrement des densités, la paramétrisation de cette intégrale peut être encore optimisée. Par ailleurs, pour réduire les effets de sommation des erreurs de termes de signes opposés dans le modèle de van der Waals et dans le potentiel d'interaction total, des paramètres empiriques affinés directement contre les valeurs de référence d'énergie totale plutôt que contre les termes individuels pourraient être introduits. Un modèle estimant la contribution de transfert de charge pourrait également être ajouté.

Pour aller plus loin, une fois que le potentiel d'interaction total ELMAM sera optimisé pour reproduire des résultats de méthodes quantiques, celui-ci pourrait être employé comme terme enthalpique d'une fonction de scoring de docking moléculaire. En effet, ces approches utilisées en conception de médicaments prédisent les ligands ayant le meilleur potentiel pour se fixer sur une protéine cible en calculant les différentes contributions de cette interaction tels que les termes enthalpiques, entropiques, ou de solvation. Pour cela, des modèles de calcul de ces interactions à la fois précis et rapides, tels que les potentiels ELMAM, sont nécessaires.

Pour finir, l'extension des approches de cristallographie quantique au monde de la biologie structurale ouvre un champ d'applications immense dont l'exploration ne fait que commencer. Partant d'une description fine de cette quantité fondamentale qu'est la densité électronique, les propriétés moléculaires deviennent accessibles. Dans le cas des protéines, les processus de recon-

naissance moléculaire et de guidage du ligand à travers le solvant, les mécanismes enzymatiques et les dynamiques structurales peuvent être mieux analysés sur la base de ces approches. Les descripteurs développés dans ce travail s'inscrivent dans cette démarche et ouvrent la voie à d'autres développements méthodologiques pour appréhender ces systèmes complexes.

Bibliographie

- [Abramov, 1997] ABRAMOV, Y. A. (1997). On the possibility of kinetic energy density evaluation from the experimental electron-density distribution. *Acta Crystallographica Section A*, 53(3): 264–272.
- [Abramov *et al.*, 2000a] ABRAMOV, Y. A., VOLKOV, A., WU, G. et COPPENS, P. (2000a). The experimental charge-density approach in the evaluation of intermolecular interactions. application of a new module of the xd programming package to several solids including a pentapeptide. *Acta Crystallographica Section A*, 56(6):585–591.
- [Abramov *et al.*, 2000b] ABRAMOV, Y. A., VOLKOV, A., WU, G. et COPPENS, P. (2000b). Use of x-ray charge densities in the calculation of intermolecular interactions and lattice energies : Application to glycylglycine, dl-histidine, and dl-proline and comparison with theory. *The Journal of Physical Chemistry B*, 104(9):2183–2188.
- [Afonine *et al.*, 2007] AFONINE, P. V., GROSSE-KUNSTLEVE, R. W., ADAMS, P. D., LUNIN, V. Y. et URZHUMTSEV, A. (2007). On macromolecular refinement at subatomic resolution with interatomic scatterers. *Acta Crystallographica Section D*, 63(11):1194–1197.
- [Ahmed *et al.*, 2013] AHMED, M., JELSCH, C., GUILLOT, B., LECOMTE, C. et DOMAGAŁA, S. (2013). Relationship between stereochemistry and charge density in hydrogen bonds with oxygen acceptors. *Crystal growth & design*, 13(1):315–325.
- [Alkorta *et al.*, 2019] ALKORTA, I., MATA, I., MOLINS, E. et ESPINOSA, E. (2019). Energetic, topological and electric field analyses of cation-cation nucleic acid interactions in watson-crick disposition. *Chem Phys Chem*, 20:148–158.
- [Allen et Tildesley, 1987] ALLEN, M. P. et TILDESLEY, D. J. (1987). *Computer simulation of liquids*. Oxford university press.
- [Anandakrishnan *et al.*, 2012] ANANDAKRISHNAN, R., AGUILAR, B. et ONUFRIEV, A. V. (2012). H++ 3.0 : automating p k prediction and the preparation of biomolecular structures for atomistic molecular modeling and simulations. *Nucleic acids research*, 40(W1):W537–W541.
- [Anderson, 2003] ANDERSON, A. C. (2003). The process of structure-based drug design. *Chemistry & biology*, 10(9):787–797.
- [Ángyán, 2007] ÁNGYÁN, J. G. (2007). On the exchange-hole model of london dispersion forces. *The Journal of chemical physics*, 127(2):024108.
- [Anjalikrishna *et al.*, 2019] ANJALIKRISHNA, P. K., SURESH, C. H. et GADRE, S. R. (2019). Electrostatic topographical viewpoint of π -conjugation and aromaticity of hydrocarbons. *The Journal of Physical Chemistry A*, 123(46):10139–10151.

- [Aquilanti *et al.*, 1996] AQUILANTI, V., CAPPELLETTI, D. et PIRANI, F. (1996). Range and strength of interatomic forces : dispersion and induction contributions to the bonds of dications and of ionic molecules. *Chemical Physics*, 209(2-3):299–311.
- [Ascenzi *et al.*, 2003] ASCENZI, P., BOCEDI, A., BOLOGNESI, M., SPALLAROSSA, A., COLETTA, M., CRISTOFARO, R. D. et MENEGATTI, E. (2003). The bovine basic pancreatic trypsin inhibitor (kunitz inhibitor) : a milestone protein. *Current Protein and Peptide Science*, 4(3): 231–251.
- [Bacon, 1975] BACON, G. E. (1975). *Neutron diffraction*. Oxford University Press.
- [Bader *et al.*, 1979] BADER, R. F., NGUYEN-DANG, T. T. et TAL, Y. (1979). Quantum topology of molecular charge distributions. ii. molecular structure and its change. *The Journal of Chemical Physics*, 70(9):4316–4329.
- [Bader, 1990] BADER, R. F. W. (1990). *Atoms In Molecules : A Quantum Theory*. Clarendon Press.
- [Bader, 1998] BADER, R. F. W. (1998). A bond path : a universal indicator of bonded interactions. *The Journal of Physical Chemistry A*, 102(37):7314–7323.
- [Bader et Stephens, 1975] BADER, R. F. W. et STEPHENS, M. E. (1975). Spatial localization of the electronic pair and number distributions in molecules. *Journal of the American Chemical Society*, 97(26):7391–7399.
- [Bał *et al.*, 2011] BAŁ, J. M., DOMAGAŁA, S., HÜBSCHLE, C., JELSCH, C., DITTRICH, B. et DOMINIAK, P. M. (2011). Verification of structural and electrostatic properties obtained by the use of different pseudoatom databases. *Acta Crystallographica Section A*, 67(2):141–153.
- [Bał *et al.*, 2009] BAŁ, J. M., DOMINIAK, P. M., WILSON, C. C. et WOŹNIAK, K. (2009). Experimental charge-density study of paracetamol–multipole refinement in the presence of a disordered methyl group. *Acta Crystallographica Section A : Foundations of Crystallography*, 65(6):490–500.
- [Balanarayan et Gadre, 2003] BALANARAYAN, P. et GADRE, S. R. (2003). Topography of molecular scalar fields. i. algorithm and poincaré–hopf relation. *The Journal of chemical physics*, 119(10):5037–5043.
- [Ball, 2008] BALL, P. (2008). Water as an active constituent in cell biology. *Chemical reviews*, 108(1):74–108.
- [Bartlett et Musiał, 2007] BARTLETT, R. J. et MUSIAŁ, M. (2007). Coupled-cluster theory in quantum chemistry. *Reviews of Modern Physics*, 79(1):291.
- [Bechet *et al.*, 2014] BECHET, D., MORDON, S. R., GUILLEMIN, F. et BARBERI-HEYOB, M. A. (2014). Photodynamic therapy of malignant brain tumours : A complementary approach to conventional therapies. *Cancer treatment reviews*, 40(2):229–241.
- [Becke et Edgecombe, 1990] BECKE, A. D. et EDGECOMBE, K. E. (1990). A simple measure of electron localization in atomic and molecular systems. *The Journal of chemical physics*, 92(9):5397–5403.
- [Becke et Johnson, 2005] BECKE, A. D. et JOHNSON, E. R. (2005). Exchange-hole dipole moment and the dispersion interaction. *The Journal of chemical physics*, 122(15):154104.

- [Becke et Johnson, 2007] BECKE, A. D. et JOHNSON, E. R. (2007). Exchange-hole dipole moment and the dispersion interaction revisited. *The Journal of chemical physics*, 127(15):154108.
- [Bennie et al., 2016] BENNIE, S. J., van der KAMP, M. W., PENNIFOLD, R. C. R., STELLA, M., MANBY, F. R. et MULHOLLAND, A. J. (2016). A projector-embedding approach for multiscale coupled-cluster calculations applied to citrate synthase. *Journal of chemical theory and computation*, 12(6):2689–2697.
- [Berendsen et al., 1995] BERENDSEN, H. J. C., van der SPOEL, D. et van DRUNEN, R. (1995). Gromacs : A message-passing parallel molecular dynamics implementation. *Computer physics communications*, 91(1-3):43–56.
- [Berman et al., 2000] BERMAN, H. M., WESTBROOK, J., FENG, Z., GILLILAND, G., BHAT, T. N., WEISSIG, H., SHINDYALOV, I. N. et BOURNE, P. E. (2000). The protein data bank. *Nucleic acids research*, 28(1):235–242.
- [Bojarowski et al., 2016] BOJAROWSKI, S. A., KUMAR, P. et DOMINIAK, P. M. (2016). A universal and straightforward approach to include penetration effects in electrostatic interaction energy estimation. *ChemPhysChem*, 17(16):2455–2460.
- [Bonaccorsi et al., 1970] BONACCORSI, R., SCROCCO, E. et TOMASI, J. (1970). Molecular scf calculations for the ground state of some three-membered ring molecules : (ch₂)₃, (ch₂)₂nh, (ch₂)₂nh₂⁺, (ch₂)₂o, (ch₂)₂s, (ch)₂ch₂, and n₂ch₂. *The Journal of Chemical Physics*, 52(10):5270–5284.
- [Bouhmaida et al., 2002] BOUHMAIDA, N., DUTHEIL, M., GHERMANI, N. E. et BECKER, P. (2002). Gradient vector field and properties of the experimental electrostatic potential : application to ibuprofen drug molecule. *The Journal of chemical physics*, 116(14):6196–6204.
- [Bragg et Bragg, 1913] BRAGG, W. H. et BRAGG, W. L. (1913). The structure of the diamond. *Proceedings of the Royal Society of London*, 89(610):277–291.
- [Bragg, 1913] BRAGG, W. L. (1913). The structure of some crystals as indicated by their diffraction of x-rays. *Proceedings of the Royal Society of London*, 89(610):248–277.
- [Brock et al., 1991] BROCK, C., DUNITZ, J. et HIRSHFELD, F. (1991). Transferability of deformation densities among related molecules : atomic multipole parameters from perylene for improved estimation of molecular vibrations in naphthalene and anthracene. *Acta Crystallographica Section B*, 47(5):789–797.
- [Brooks et al., 2009] BROOKS, B. R., BROOKS III, C. L., MACKERELL JR, A. D., NILSSON, L., PETRELLA, R. J., ROUX, B., WON, Y., ARCHONTIS, G., BARTELS, C., BORESCH, S., CAFLISCH, A., CAVES, L., CUI, Q., DINNER, A. R., FEIG, M., FISCHER, S., GAO, J., HODOSCEK, M., IM, W., KUCZERA, K., LAZARIDIS, T., MA, J., OVCHINNIKOV, V., PACI, E., PASTOR, R. W., POST, C. B., PU, J. Z., SCHAEFER, M., TIDOR, B., VENABLE, R. M., WOODCOCK, H. L., WU, X., YANG, W., YORK, D. M. et M, K. (2009). Charmm : the biomolecular simulation program. *Journal of computational chemistry*, 30(10):1545–1614.
- [Brzezinski et al., 2011] BRZEZINSKI, K., BRZUSZKIEWICZ, A., DAUTER, M., KUBICKI, M., JASKOLSKI, M. et DAUTER, Z. (2011). High regularity of z-dna revealed by ultra high-resolution crystal structure at 0.55 Å. *Nucleic acids research*, 39(14):6238–6248.

- [Capelli *et al.*, 2014] CAPELLI, S. C., BÜRGI, H.-B., DITTRICH, B., GRABOWSKY, S. et JAYATILAKA, D. (2014). Hirshfeld atom refinement. *IUCr Journals*, 1(5):361–379.
- [Case *et al.*, 2005] CASE, D. A., CHEATHAM, T. E., DARDEN, T., GOHLKE, H., LUO, R., MERZ, K. M., ONUFRIEV, A., SIMMERLING, C., WANG, B. et WOODS, R. J. (2005). The amber biomolecular simulation programs. *Journal of computational chemistry*, 26(16):1668–1688.
- [Chalaśiński et Szcześniak, 2000] CHAŁASIŃSKI, G. et SZCZEŚNIAK, M. M. (2000). State of the art and challenges of the ab initio theory of intermolecular interactions. *Chemical reviews*, 100(11):4227–4252.
- [Chaudhary *et al.*, 2014] CHAUDHARY, B., KHALED, Y. S., AMMORI, B. J. et ELKORD, E. (2014). Neuropilin 1 : function and therapeutic potential in cancer. *Cancer Immunology, Immunotherapy*, 63:81–99.
- [Chen *et al.*, 2010] CHEN, V. B., ARENDALL, W. B., HEADD, J. J., KEEDY, D. A., IMMORMINO, R. M., KAPRAL, G. J., MURRAY, L. W., RICHARDSON, J. S. et RICHARDSON, D. C. (2010). Molprobity : all-atom structure validation for macromolecular crystallography. *Acta Crystallographica Section D*, 66(1):12–21.
- [Chopra, 2012] CHOPRA, D. (2012). Advances in understanding of chemical bonding : inputs from experimental and theoretical charge density analysis. *The Journal of Physical Chemistry A*, 116(40):9791–9801.
- [Chung *et al.*, 2015] CHUNG, L. W., SAMEERA, W. M. C., RAMOZZI, R., PAGE, A. J., HATANAKA, M., PETROVA, G. P., HARRIS, T. V., LI, X., KE, Z., LIU, F., LI, H. B., DING, L. et MOROKUMA, K. (2015). The oniom method and its applications. *Chemical reviews*, 115(12):5678–5796.
- [Clemente *et al.*, 2023] CLEMENTE, C. M., CAPECE, L. et MARTÍ, M. A. (2023). Best practices on qm/mm simulations of biological systems. *Journal of Chemical Information and Modeling*, 63:2609–2627.
- [Coester et Kümmel, 1960] COESTER, F. et KÜMMEL, H. (1960). Short-range correlations in nuclear wave functions. *Nuclear Physics*, 17:477–485.
- [Compton, 1915] COMPTON, A. H. (1915). The distribution of the electrons in atoms. *Nature*, 95(2378):343–344.
- [Contreras-García *et al.*, 2011] CONTRERAS-GARCÍA, J., YANG, W. et JOHNSON, E. R. (2011). Analysis of hydrogen-bond interaction potentials from the electron density : integration of noncovalent interaction regions. *The Journal of Physical Chemistry A*, 115(45):12983–12990.
- [Coppens, 1967] COPPENS, P. (1967). Comparative x-ray and neutron diffraction study of bonding effects in s-triazine. *Science*, 158(3808):1577–1579.
- [Coppens, 1997] COPPENS, P. (1997). *X-Ray charge densities and chemical bonding*. Oxford university press.
- [Coppens, 2005] COPPENS, P. (2005). Charge densities come of age. *Angewandte Chemie International Edition*, 44(42):6810–6811.
- [Coppens *et al.*, 1999] COPPENS, P., ABRAMOV, Y., CARDUCCI, M., KORJOV, B., NOVOZHILOVA, I., ALHAMBRA, C. et PRESSPRICH, M. R. (1999). Experimental charge densities and

- intermolecular interactions : electrostatic and topological analysis of dl-histidine. *Journal of the American Chemical Society*, 121(11):2585–2593.
- [Coppens *et al.*, 1979] COPPENS, P., GURU ROW, T. N., LEUNG, P., STEVENS, E. D., BECKER, P. J. t. et YANG, Y. W. (1979). Net atomic charges and molecular dipole moments from spherical-atom x-ray refinements, and the relation between atomic charge and shape. *Acta Crystallographica Section A*, 35(1):63–72.
- [Cornillon et Matzner-Lober, 2006] CORNILLON, P.-A. et MATZNER-LOBER, E. (2006). *Régression : théorie et applications*. Springer Paris.
- [Cox *et al.*, 1981] COX, S. R., HSU, L.-Y. et WILLIAMS, D. E. (1981). Nonbonded potential function models for crystalline oxohydrocarbons. *Acta Crystallographica Section A*, 37(3):293–301.
- [Cremer, 2011] CREMER, D. (2011). Møller–plessset perturbation theory : from small molecule methods to methods for thousands of atoms. *Wiley Interdisciplinary Reviews : Computational Molecular Science*, 1(4):509–530.
- [Daly *et al.*, 2020] DALY, J. L., SIMONETTI, B., KLEIN, K., CHEN, K.-E., WILLIAMSON, M. K., ANTÓN-PLÁGARO, C., SHOEMARK, D. K., SIMÓN-GRACIA, L., BAUER, M., HOLLANDI, R. *et al.* (2020). Neuropilin-1 is a host factor for sars-cov-2 infection. *Science*, 370(6518):861–865.
- [Dauter *et al.*, 1997] DAUTER, Z., LAMZIN, V. S. et WILSON, K. S. (1997). The benefits of atomic resolution. *Current Opinion in Structural Biology*, 7(5):681–688.
- [de Broglie, 1926] de BROGLIE, L. (1926). Interference and corpuscular light. *Nature*, 118(2969):441–442.
- [Debye, 1915] DEBYE, P. (1915). Zerstreuung von röntgenstrahlen. *Annalen der Physik*, 351(6):809–823.
- [Dittrich *et al.*, 2006] DITTRICH, B., HÜBSCHLE, C. B., LUGER, P. et SPACKMAN, M. A. (2006). Introduction and validation of an invariom database for amino-acid, peptide and protein molecules. *Acta Crystallographica Section D*, 62(11):1325–1335.
- [Dittrich *et al.*, 2013] DITTRICH, B., HÜBSCHLE, C. B., PRÖPPER, K., DIETRICH, F., STOLPER, T. et HOLSTEIN, J. (2013). The generalized invariom database (gid). *Acta Crystallographica Section B*, 69(2):91–104.
- [Dittrich *et al.*, 2004] DITTRICH, B., KORITSÁNSZKY, T. et LUGER, P. (2004). A simple approach to nonspherical electron densities by using invarioms. *Angewandte Chemie International Edition*, 43(20):2718–2721.
- [Dittrich et Luger, 2017] DITTRICH, B. et LUGER, P. (2017). Invariom-based comparative electron density studies of iso-sildenafil and sildenafil. *Zeitschrift für Naturforschung B*, 72(1):1–10.
- [Dittrich et Matta, 2014] DITTRICH, B. et MATTA, C. F. (2014). Contributions of charge-density research to medicinal chemistry. *IUCr Journal*, 1(6):457–469.
- [Dixon et Merz Jr, 1997] DIXON, S. L. et MERZ JR, K. M. (1997). Fast, accurate semiempirical molecular orbital calculations for macromolecules. *The Journal of chemical physics*, 107(3):879–893.

- [Dobson *et al.*, 2013] DOBSON, J. F., VIGNALE, G. et DAS, M. P. (2013). *Electronic density functional theory : recent progress and new directions*. Springer Science & Business Media.
- [Domagała *et al.*, 2012] DOMAGAŁA, S., FOURNIER, B., LIEBSCHNER, D., GUILLOT, B. et JELSCH, C. (2012). An improved experimental databank of transferable multipolar atom models—elmam2. construction details and applications. *Acta Crystallographica Section A : Foundations of Crystallography*, 68(3):337–351.
- [Dominiak *et al.*, 2006] DOMINIAK, P. M., MAKAL, A., MALLINSON, P. R., TRZCINSKA, K., EILMES, J., GRECH, E., CHRUSZCZ, M., MINOR, W. et WOŹNIAK, K. (2006). Continua of interactions between pairs of atoms in molecular crystals. *Chemistry—A European Journal*, 12(7):1941–1949.
- [Dominiak *et al.*, 2009] DOMINIAK, P. M., VOLKOV, A., DOMINIAK, A. P., JARZEMBSKA, K. N. et COPPENS, P. (2009). 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 Crystallographica Section D : Biological Crystallography*, 65(5):485–499.
- [Dominiak *et al.*, 2007] DOMINIAK, P. M., VOLKOV, A., LI, X., MESSERSCHMIDT, M. et COPPENS, P. (2007). A theoretical databank of transferable aspherical atoms and its application to electrostatic interaction energy calculations of macromolecules. *Journal of Chemical Theory and Computation*, 3(1):232–247.
- [Duke *et al.*, 2014] DUKE, R. E., STAROVOYTOV, O. N., PIQUEMAL, J.-P. et CISNEROS, G. A. (2014). Gem* : A molecular electronic density-based force field for molecular dynamics simulations. *Journal of Chemical Theory and Computation*, 10(4):1361–1365.
- [Dumond *et al.*, 2020] DUMOND, A., DEMANGE, L. et PAGÈS, G. (2020). Les neuropilines-des cibles pertinentes pour améliorer le traitement des cancers. *médecine/sciences*, 36(5):487–496.
- [Engelhard *et al.*, 2018] ENGELHARD, C., CHIZHOV, I., SIEBERT, F. et ENGELHARD, M. (2018). Microbial halorhodopsins : light-driven chloride pumps. *Chemical reviews*, 118(21):10629–10645.
- [Engkvist *et al.*, 2000] ENGVIST, O., ÅSTRAND, P.-O. et KARLSTRÖM, G. (2000). Accurate intermolecular potentials obtained from molecular wave functions : Bridging the gap between quantum chemistry and molecular simulations. *Chemical Reviews*, 100(11):4087–4108.
- [Epstein et Abeles, 1992] EPSTEIN, D. M. et ABELES, R. H. (1992). Role of serine 214 and tyrosine 171, components of the s2 subsite of. alpha.-lytic protease, in catalysis. *Biochemistry*, 31(45):11216–11223.
- [Espinosa, 2023] ESPINOSA, E. (2023). "Topological analysis of the electrostatic potential in molecular systems" for the Distinguished Lectures on quantum crystallography and complementary fields, University of Warsaw.
- [Espinosa *et al.*, 2002] ESPINOSA, E., ALKORTA, I., ELGUERO, J. et MOLINS, E. (2002). From weak to strong interactions : A comprehensive analysis of the topological and energetic properties of the electron density distribution involving x–h... f–y systems. *The Journal of chemical physics*, 117(12):5529–5542.

- [Espinosa *et al.*, 1998] ESPINOSA, E., MOLINS, E. et LECOMTE, C. (1998). Hydrogen bond strengths revealed by topological analyses of experimentally observed electron densities. *Chemical physics letters*, 285(3-4):170–173.
- [Farrugia et Macchi, 2009] FARRUGIA, L. J. et MACCHI, P. (2009). On the interpretation of the source function. *The Journal of Physical Chemistry A*, 113(37):10058–10067.
- [Fedorov, 2017] FEDOROV, D. G. (2017). The fragment molecular orbital method : theoretical development, implementation in gamess, and applications. *Wiley Interdisciplinary Reviews : Computational Molecular Science*, 7(6):e1322.
- [Fedorov et Kitaura, 2006] FEDOROV, D. G. et KITAURA, K. (2006). Theoretical development of the fragment molecular orbital (fmo) method. In *Modern methods for theoretical physical chemistry of biopolymers*, pages 3–38. Elsevier.
- [Field *et al.*, 1990] FIELD, M. J., BASH, P. A. et KARPLUS, M. (1990). A combined quantum mechanical and molecular mechanical potential for molecular dynamics simulations. *Journal of computational chemistry*, 11(6):700–733.
- [Fournier, 2010] FOURNIER, B. (2010). *Modélisation des propriétés électrostatiques des complexes macromoléculaires à partir des données de diffraction des rayons X à très haute résolution*. Thèse de doctorat, Université Henri Poincaré, Nancy I.
- [Fournier *et al.*, 2009] FOURNIER, B., BENDEIF, E.-E., GUILLOT, B., PODJARNY, A., LECOMTE, C. et JELSCH, C. (2009). Charge density and electrostatic interactions of fidarestat, an inhibitor of human aldose reductase. *Journal of the American Chemical Society*, 131(31):10929–10941.
- [Fournier *et al.*, 2018] FOURNIER, B., GUILLOT, B., LECOMTE, C., ESCUDERO-ADÁN, E. C. et JELSCH, C. (2018). A method to estimate statistical errors of properties derived from charge-density modelling. *Acta Crystallographica Section A : Foundations and Advances*, 74(3):170–183.
- [Friedrich *et al.*, 1912] FRIEDRICH, W., KNIPPING, P. et von LAUE, M. (1912). Interferenzerscheinungen bei röntgenstrahlen. *Königlich-Bayerischen Akademie der Wissenschaften zu München*, pages 303–322.
- [Friesner *et al.*, 2004] FRIESNER, R. A., BANKS, J. L., MURPHY, R. B., HALGREN, T. A., KLICIC, J. J., MAINZ, D. T., REPASKY, M. P., KNOLL, E. H., SHELLEY, M., PERRY, J. K. *et al.* (2004). Glide : a new approach for rapid, accurate docking and scoring. 1. method and assessment of docking accuracy. *Journal of medicinal chemistry*, 47(7):1739–1749.
- [Gadre et Bendale, 1986] GADRE, S. R. et BENDALE, R. D. (1986). On the similarity between molecular electron densities, electrostatic potentials and bare nuclear potentials. *Chemical physics letters*, 130(6):515–521.
- [Gadre *et al.*, 1992] GADRE, S. R., KULKARNI, S. A. et SHRIVASTAVA, I. H. (1992). Molecular electrostatic potentials : A topographical study. *The Journal of chemical physics*, 96(7):5253–5260.
- [Gadre et Kumar, 2016] GADRE, S. R. et KUMAR, A. (2016). Bonding and reactivity patterns from electrostatic landscapes of molecules. *Journal of Chemical Sciences*, 128(10):1519–1526.

- [Gadre *et al.*, 1994] GADRE, S. R., SHIRSAT, R. N. et LIMAYE, A. C. (1994). Molecular tailoring approach for simulation of electrostatic properties. *The Journal of Physical Chemistry*, 98(37): 9165–9169.
- [Gadre et Shrivastava, 1991] GADRE, S. R. et SHRIVASTAVA, I. H. (1991). Shapes and sizes of molecular anions via topographical analysis of electrostatic potential. *The Journal of chemical physics*, 94(6):4384–4390.
- [Gao, 1996] GAO, J. (1996). Methods and applications of combined quantum mechanical and molecular mechanical potentials. *Reviews in computational chemistry*, 7:119–185.
- [Gatti *et al.*, 2003] GATTI, C., CARGNONI, F. et BERTINI, L. (2003). Chemical information from the source function. *Journal of computational chemistry*, 24(4):422–436.
- [Gatti et Macchi, 2012] GATTI, C. et MACCHI, P. (2012). *Modern Charge-Density Analysis*. Springer.
- [Gavezzotti, 2002] GAVEZZOTTI, A. (2002). Calculation of intermolecular interaction energies by direct numerical integration over electron densities. i. electrostatic and polarization energies in molecular crystals. *The Journal of Physical Chemistry B*, 106(16):4145–4154.
- [Gavezzotti, 2003] GAVEZZOTTI, A. (2003). 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. *The Journal of Physical Chemistry B*, 107(10):2344–2353.
- [Gavezzotti, 2005] GAVEZZOTTI, A. (2005). Quantitative ranking of crystal packing modes by systematic calculations on potential energies and vibrational amplitudes of molecular dimers. *Journal of Chemical Theory and Computation*, 1(5):834–840.
- [Gavezzotti, 2011] GAVEZZOTTI, A. (2011). Efficient computer modeling of organic materials. the atom–atom, coulomb–london–pauli (aa-clp) model for intermolecular electrostatic-polarization, dispersion and repulsion energies. *New Journal of Chemistry*, 35(7):1360–1368.
- [Genoni *et al.*, 2018] GENONI, A., BUČINSKÝ, L., CLAISER, N., CONTRERAS-GARCÍA, J., DITTRICH, B., DOMINIAK, P. M., ESPINOSA, E., GATTI, C., GIANNOZZI, P., GILLET, J. M., JAYATILAKA, D., MACCHI, P., MADSEN, A. O., MASSA, L., MATTA, C. F., MERZN, K. M. J., NAKASHIMA, P. N. H., OTT, H., RYDE, U., SCHWARZ, K., SIERKA, M. et GRABOWSKY, S. (2018). Quantum crystallography : current developments and future perspectives. *Chemistry–A European Journal*, 24(43):10881–10905.
- [González et Fisher, 2015] GONZÁLEZ, J. M. et FISHER, S. Z. (2015). Structural analysis of ibuprofen binding to human adipocyte fatty-acid binding protein (fabp4). *Acta Crystallographica Section F : Structural Biology Communications*, 71(2):163–170.
- [Gordon, 1996] GORDON, J. H. J. M. S. (1996). An approximate formula for the intermolecular pauli repulsion between closed shell molecules. *molecular physics*, 89(5):1313–1325.
- [Goudiaby *et al.*, 2023] GOUDIABY, I., MALLIAVIN, T. E., MOCCHETTI, E., MATHIOT, S., ACHE-RAR, S., FROCHOT, C., BARBERI-HEYOB, M., GUILLOT, B., FAVIER, F., DIDIERJEAN, C. et JELSCH, C. (2023). New crystal form of human neuropilin-1 b1 fragment with six electrostatic mutations complexed with KDKPPR peptide ligand. *Molecules*, 28:5603.

- [Grabowsky *et al.*, 2008] GRABOWSKY, S., PFEUFFER, T., MORGENROTH, W., PAULMANN, C., SCHIRMEISTER, T. et LUGER, P. (2008). A comparative study on the experimentally derived electron densities of three protease inhibitor model compounds. *Organic & biomolecular chemistry*, 6(13):2295–2307.
- [Graf *et al.*, 1987] GRAF, L., CRAIK, C. S., PATHY, A., ROCZNIK, S., FLETTERICK, R. J. et RUTTER, W. J. (1987). Selective alteration of substrate specificity by replacement of aspartic acid-189 with lysine in the binding pocket of trypsin. *Biochemistry*, 26(9):2616–2623.
- [Greer, 1981] GREER, J. (1981). Comparative model-building of the mammalian serine proteases. *Journal of Molecular Biology*, 153(4):1027–1042.
- [Grimme *et al.*, 2010] GRIMME, S., ANTONY, J., EHRLICH, S. et KRIEG, H. (2010). A consistent and accurate ab initio parametrization of density functional dispersion correction (dft-d) for the 94 elements h-pu. *The Journal of chemical physics*, 132(15):154104.
- [Groom *et al.*, 2016] GROOM, C. R., BRUNO, I. J., LIGHTFOOT, M. P. et WARD, S. C. (2016). The cambridge structural database. *Acta Crystallographica Section B*, 72(2):171–179.
- [Gruza *et al.*, 2020] GRUZA, B., CHODKIEWICZ, M. L., KRZESZCZAKOWSKA, J. et DOMINIAK, P. M. (2020). Refinement of organic crystal structures with multipolar electron scattering factors. *Acta Crystallographica Section A*, 76(1):92–109.
- [Guillot *et al.*, 2014] GUILLOT, B., ESPINOSA, E., HUDER, L. et JELSCH, C. (2014). Moprovier : a tool to study proteins from a charge density science perspective. volume A70, page C279.
- [Guillot *et al.*, 2008] GUILLOT, B., JELSCH, C., PODJARNY, A. et LECOMTE, C. (2008). Charge-density analysis of a protein structure at subatomic resolution : the human aldose reductase case. *Acta Crystallographica Section D*, 64(5):567–588.
- [Guillot *et al.*, 2001] GUILLOT, B., VIRY, L., GUILLOT, R., LECOMTE, C. et JELSCH, C. (2001). Refinement of proteins at subatomic resolution with mopro. *Journal of applied crystallography*, 34(2):214–223.
- [Halgren, 1992] HALGREN, T. A. (1992). The representation of van der waals (vdw) interactions in molecular mechanics force fields : potential form, combination rules, and vdw parameters. *Journal of the American Chemical Society*, 114(20):7827–7843.
- [Hanhoff *et al.*, 2002] HANHOFF, T., LÜCKE, C. et SPENER, F. (2002). Insights into binding of fatty acids by fatty acid binding proteins. In *Cellular Lipid Binding Proteins*, pages 45–54. Springer.
- [Hansen et Coppens, 1978] HANSEN, N. K. et COPPENS, P. (1978). Testing aspherical atom refinements on small-molecule data sets. *Acta Crystallographica Section A*, 34(6):909–921.
- [Hayes *et al.*, 2005] HAYES, J. D., FLANAGAN, J. U. et JOWSEY, I. R. (2005). Glutathione transferases. *Annual Reviews of Pharmacology and Toxicology*, 45:51–88.
- [He et Merz Jr, 2010] HE, X. et MERZ JR, K. M. (2010). Divide and conquer hartree-fock calculations on proteins. *Journal of chemical theory and computation*, 6(2):405–411.
- [Hedstrom, 2002] HEDSTROM, L. (2002). Serine protease mechanism and specificity. *Chemical Reviews*, 102(12):4501–4525.

- [Held et van Smaalen, 2014] HELD, J. et van SMAALEN, S. (2014). The active site of hen egg-white lysozyme : flexibility and chemical bonding. *Acta Crystallographica Section D*, 70(4): 1136–1146.
- [Hellner, 1977] HELLNER, E. (1977). A simple refinement of density distributions of bonding electrons. i. a description of the proposed method. *Acta Crystallographica Section B*, 33(12): 3813–3816.
- [Hirano *et al.*, 2016] HIRANO, Y., TAKEDA, K. et MIKI, K. (2016). Charge-density analysis of an iron–sulfur protein at an ultra-high resolution of 0.48 Å. *Nature*, 534(7606):281–284.
- [Hirota *et al.*, 2006] HIROTA, M., OHMURAYA, M. et BABA, H. (2006). The role of trypsin, trypsin inhibitor, and trypsin receptor in the onset and aggravation of pancreatitis. *Journal of gastroenterology*, 41(9).
- [Hirshfeld, 1977] HIRSHFELD, F. L. (1977). Xvii. spatial partitioning of charge density. *Israel Journal of Chemistry*, 16(2-3):198–201.
- [Hohenberg et Kohn, 1964] HOHENBERG, P. et KOHN, W. (1964). Inhomogeneous electron gas. *Physical review*, 136(3B):B864.
- [Housset *et al.*, 2000] HOUSSET, D., BENABICHA, F., PICHON-PESME, V., JELSCH, C., MAIERHOFFER, A., DAVID, S., FONTECILLA-CAMPS, J. C. et LECOMTE, C. (2000). Towards the charge-density study of proteins : a room-temperature scorpion-toxin structure at 0.96 Å resolution as a first test case. *Acta Crystallographica Section D*, 56(2):151–160.
- [Howard *et al.*, 2016] 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. et PODJARNY, A. (2016). 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*, 3(2):115–126.
- [Huang *et al.*, 2005] HUANG, L., MASSA, L. et KARLE, J. (2005). Kernel energy method illustrated with peptides. *International journal of quantum chemistry*, 103(6):808–817.
- [Hufner-Wulsdorf et Klebe, 2020] HUFNER-WULSDORF, T. et KLEBE, G. (2020). Role of water molecules in protein–ligand dissociation and selectivity discrimination : Analysis of the mechanisms and kinetics of biomolecular solvation using molecular dynamics. *Journal of Chemical Information and Modeling*, 60(3):1818–1832.
- [Jarzembska et Dominiak, 2012] JARZEMBSKA, K. N. et DOMINIAK, P. M. (2012). New version of the theoretical databank of transferable aspherical pseudoatoms, ubdb2011–towards nucleic acid modelling. *Acta Crystallographica Section A : Foundations of Crystallography*, 68(1):139–147.
- [Jayatilaka, 1998] JAYATILAKA, D. (1998). Wave function for beryllium from x-ray diffraction data. *Physical review letters*, 80(4):798.
- [Jayatilaka et Dittrich, 2008] JAYATILAKA, D. et DITTRICH, B. (2008). X-ray structure refinement using aspherical atomic density functions obtained from quantum-mechanical calculations. *Acta Crystallographica Section A : Foundations of Crystallography*, 64(3):383–393.

- [Jayatilaka et Grimwood, 2001] JAYATILAKA, D. et GRIMWOOD, D. J. (2001). Wavefunctions derived from experiment. i. motivation and theory. *Acta Crystallographica Section A : Foundations of Crystallography*, 57(1):76–86.
- [Jelsch et al., 2014] JELSCH, C., EJSMTONT, K. et HUDER, L. (2014). The enrichment ratio of atomic contacts in crystals, an indicator derived from the hirshfeld surface analysis. *IUCrJ*, 1(2):119–128.
- [Jelsch et al., 2005] JELSCH, C., GUILLOT, B., LAGOUTTE, A. et LECOMTE, C. (2005). Advances in protein and small-molecule charge-density refinement methods using mopro. *Journal of applied crystallography*, 38(1):38–54.
- [Jelsch et al., 1998] JELSCH, C., PICHON-PESME, V., LECOMTE, C. et AUBRY, A. (1998). Transferability of multipole charge-density parameters : application to very high resolution oligopeptide and protein structures. *Acta Crystallographica Section D*, 54(6):1306–1318.
- [Jelsch et al., 2000] JELSCH, C., TEETER, M. M., LAMZIN, V., PICHON-PESME, V., BLESSING, R. H. et LECOMTE, C. (2000). Accurate protein crystallography at ultra-high resolution : valence electron distribution in crambin. *Proceedings of the National Academy of Sciences*, 97(7):3171–3176.
- [Jensen et Gordon, 1998] JENSEN, J. H. et GORDON, M. S. (1998). An approximate formula for the intermolecular pauli repulsion between closed shell molecules. ii. application to the effective fragment potential method. *The Journal of chemical physics*, 108(12):4772–4782.
- [Jha et al., 2021] JHA, K. K., GRUZA, B., CHODKIEWICZ, M. L., JELSCH, C. et DOMINIAK, P. M. (2021). Refinements on electron diffraction data of β -glycine in mopro : a quest for an improved structure model. *Journal of Applied Crystallography*, 54(4):1234–1243.
- [Jha et al., 2022] JHA, K. K., GRUZA, B., SYPKO, A., KUMAR, P., CHODKIEWICZ, M. L. et DOMINIAK, P. M. (2022). Multipolar atom types from theory and statistical clustering (matts) data bank : Restructurization and extension of ubdb. *Journal of Chemical Information and Modeling*, 62(16):3752–3765.
- [Johnson et al., 2010] JOHNSON, E. R., KEINAN, S., MORI-SÁNCHEZ, P., CONTRERAS-GARCÍA, J., COHEN, A. J. et YANG, W. (2010). Revealing noncovalent interactions. *Journal of the American Chemical Society*, 132(18):6498–6506.
- [Johnston et al., 2011] JOHNSTON, A., BARDIN, J., JOHNSTON, B. F., FERNANDES, P., KENNEDY, A. R., PRICE, S. L. et FLORENCE, A. J. (2011). Experimental and predicted crystal energy landscapes of chlorothiazide. *Crystal growth & design*, 11(2):405–413.
- [Jones et al., 1997] JONES, G., WILLETT, P., GLEN, R. C., LEACH, A. R. et TAYLOR, R. (1997). Development and validation of a genetic algorithm for flexible docking. *Journal of molecular biology*, 267(3):727–748.
- [Jorgensen et al., 1996] JORGENSEN, W. L., MAXWELL, D. S. et TIRADO-RIVES, J. (1996). Development and testing of the opols all-atom force field on conformational energetics and properties of organic liquids. *Journal of the American Chemical Society*, 118(45):11225–11236.
- [Jumper et al., 2021] JUMPER, J., EVANS, R., PRITZEL, A., GREEN, T., FIGURNOV, M., RONEBERGER, O., TUNYASUVUNAKOOL, K., BATES, R., ŽÍDEK, A., POTAPENKO, A. et al. (2021). Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589.

- [Jurrus *et al.*, 2018] JURRUS, E., ENGEL, D., STAR, K., MONSON, K., BRANDI, J., FELBERG, L. E., BROOKES, D. H., WILSON, L., CHEN, J., LILES, K. *et al.* (2018). Improvements to the apbs biomolecular solvation software suite. *Protein Science*, 27(1):112–128.
- [Kamarulzaman *et al.*, 2015] KAMARULZAMAN, E., MOHD GAZZALI, A., ACHERAR, S., FROCHOT, C., BARBERI-HEYOB, M., BOURA, C., CHAIMBAULT, P., SIBILLE, E., WAHAB, H. A. et VANDERESSE, R. (2015). New peptide-conjugated chlorin-type photosensitizer targeting neuropilin-1 for anti-vascular targeted photodynamic therapy. *International journal of molecular sciences*, 16(10):24059–24080.
- [Kamarulzaman *et al.*, 2017] KAMARULZAMAN, E. E., VANDERESSE, R., GAZZALI, A. M., BARBERI-HEYOB, M., BOURA, C., FROCHOT, C., SHAWKATALY, O., AUBRY, A. et WAHAB, H. A. (2017). Molecular modelling, synthesis and biological evaluation of peptide inhibitors as anti-angiogenic agent targeting neuropilin-1 for anticancer application. *Journal of Biomolecular Structure and Dynamics*, 35(1):26–45.
- [Kammerscheit *et al.*, 2020] KAMMERSCHEIT, X., HECKER, A., ROUHIER, N., CHAUVAT, F. et CASSIER-CHAUVAT, C. (2020). Methylglyoxal detoxification revisited : role of glutathione transferase in model cyanobacterium *synechocystis* sp. strain pcc 6803. *Mbio*, 11(4):e00882–20.
- [Karplus, 2014] KARPLUS, M. (2014). Development of multiscale models for complex chemical systems : from h+ h2 to biomolecules (nobel lecture). *Angewandte Chemie International Edition*, 53(38):9992–10005.
- [Karshikoff, 2006] KARSHIKOFF, A. (2006). *Non-covalent interactions in proteins*. World Scientific.
- [Kawamura *et al.*, 2011] KAWAMURA, K., YAMADA, T., KURIHARA, K., TAMADA, T., KUROKI, R., TANAKA, I., TAKAHASHI, H. et NIIMURA, N. (2011). X-ray and neutron protein crystallographic analysis of the trypsin–bpti complex. *Acta Crystallographica Section D*, 67(2):140–148.
- [Kendrew *et al.*, 1958] KENDREW, J. C., BODO, G., DINTZIS, H. M., PARRISH, R., WYCKOFF, H. et PHILLIPS, D. C. (1958). A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. *Nature*, 181(4610):662–666.
- [Kim *et al.*, 1981] KIM, Y. S., KIM, S. K. et LEE, W. D. (1981). Dependence of the closed-shell repulsive interaction on the overlap of the electron densities. *Chemical Physics Letters*, 80(3):574–575.
- [Kita *et al.*, 1976] KITA, S., NODA, K. et INOUE, H. (1976). Repulsive potentials for cl—r and br—r (r= he, ne, and ar) derived from beam experiments. *The Journal of Chemical Physics*, 64(8):3446–3449.
- [Kitaura *et al.*, 1999] KITAURA, K., IKEO, E., ASADA, T., NAKANO, T. et UEBAYASI, M. (1999). Fragment molecular orbital method : an approximate computational method for large molecules. *Chemical Physics Letters*, 313(3-4):701–706.
- [Kobbelt *et al.*, 2002] KOBBELT, L., BISCHOFF, S., BOTSCH, M. et STEINBERG, S. (2002). Open-mesh : A generic and efficient polygon mesh data structure. Conference paper.

- [Koch et Popelier, 1995] KOCH, U. et POPELIER, P. L. A. (1995). Characterization of cho hydrogen bonds on the basis of the charge density. *The Journal of Physical Chemistry*, 99(24):9747–9754.
- [Kohn et Sham, 1965] KOHN, W. et SHAM, L. J. (1965). Self-consistent equations including exchange and correlation effects. *Physical review*, 140(4A):A1133–A1138.
- [Koritsanszky *et al.*, 2002] KORITSANSZKY, T., VOLKOV, A. et COPPENS, P. (2002). Aspherical-atom scattering factors from molecular wave functions. 1. transferability and conformation dependence of atomic electron densities of peptides within the multipole formalism. *Acta Crystallographica Section A*, 58(5):464–472.
- [Koritsanszky et Coppens, 2001] KORITSANSZKY, T. S. et COPPENS, P. (2001). Chemical applications of x-ray charge-density analysis. *Chemical reviews*, 101(6):1583–1628.
- [Krieger *et al.*, 2009] KRIEGER, E., JOO, K., LEE, J., LEE, J., RAMAN, S., THOMPSON, J., TYKA, M., BAKER, D. et KARPLUS, K. (2009). Improving physical realism, stereochemistry, and side-chain accuracy in homology modeling : Four approaches that performed well in casp8. *Proteins : Structure, Function, and Bioinformatics*, 77(S9):114–122.
- [Krowarsch *et al.*, 2003] KROWARSCH, D., CIERPICKI, T., JELEN, F. et OTLEWSKI, J. (2003). Canonical protein inhibitors of serine proteases. *Cellular and molecular life sciences CMLS*, 60:2427–2444.
- [Kulik *et al.*, 2022] KULIK, M., CHODKIEWICZ, M. L. et DOMINIAK, P. M. (2022). Theoretical 3d electron diffraction electrostatic potential maps of proteins modeled with a multipolar pseudoatom data bank. *Acta Crystallographica Section D*, 78(8).
- [Kumar et Gadre, 2016] KUMAR, A. et GADRE, S. R. (2016). Exploring the gradient paths and zero flux surfaces of molecular electrostatic potential. *Journal of chemical theory and computation*, 12(4):1705–1713.
- [Kumar *et al.*, 2014a] KUMAR, A., GADRE, S. R., MOHAN, N. et SURESH, C. H. (2014a). Lone pairs : an electrostatic viewpoint. *The Journal of Physical Chemistry A*, 118(2):526–532.
- [Kumar *et al.*, 2014b] KUMAR, P., BOJAROWSKI, S. A., JARZEMBSKA, K. N., DOMAGAŁA, S., VANOMMESLAEGHE, K., MACKERELL JR, A. D. et DOMINIAK, P. M. (2014b). A comparative study of transferable aspherical pseudoatom databank and classical force fields for predicting electrostatic interactions in molecular dimers. *Journal of Chemical Theory and Computation*, 10(4):1652–1664.
- [Kumar et Dominiak, 2021] KUMAR, P. et DOMINIAK, P. M. (2021). Combining molecular dynamic information and an aspherical-atom data bank in the evaluation of the electrostatic interaction energy in multimeric protein-ligand complex : A case study for hiv-1 protease. *Molecules*, 26(13).
- [Kumar *et al.*, 2019] KUMAR, P., GRUZA, B., BOJAROWSKI, S. A. et DOMINIAK, P. M. (2019). Extension of the transferable aspherical pseudoatom data bank for the comparison of molecular electrostatic potentials in structure–activity studies. *Acta Crystallographica Section A*, 75(2):398–408.

- [Laplaza *et al.*, 2021] LAPLAZA, R., PECCATI, F., BOTO, R., QUAN, C., CARBONE, A., PIQUEMAL, J., MADAY, Y. et CONTRERAS-GARCÍA, J. (2021). Nciplot and the analysis of non-covalent interactions using the reduced density gradient. *Wiley Interdisciplinary Reviews : Computational Molecular Science*, 11(2):e1497.
- [Leboeuf *et al.*, 1999] LEBOEUF, M., KÖSTER, A. M., JUG, K. et SALAHUB, D. R. (1999). Topological analysis of the molecular electrostatic potential. *The Journal of chemical physics*, 111(11):4893–4905.
- [Lecomte *et al.*, 2005] LECOMTE, C., GUILLOT, B., JELSCH, C. et PODJARNY, A. (2005). Frontier example in experimental charge density research : Experimental electrostatics of proteins. *International journal of quantum chemistry*, 101(5):624–634.
- [Leduc *et al.*, 2019] LEDUC, T., AUBERT, E., ESPINOSA, E., JELSCH, C., IORDACHE, C. et GUILLOT, B. (2019). Polarization of electron density databases of transferable multipolar atoms. *The Journal of Physical Chemistry A*, 123(32):7156–7170.
- [Lehle *et al.*, 2005] LEHLE, H., KRIEGL, J. M., NIENHAUS, K., DENG, P., FENGLER, S. et NIENHAUS, G. U. (2005). Probing electric fields in protein cavities by using the vibrational stark effect of carbon monoxide. *Biophysical journal*, 88(3):1978–1990.
- [Levitt, 2014] LEVITT, M. (2014). Birth and future of multiscale modeling for macromolecular systems (nobel lecture). *Angewandte Chemie International Edition*, 53(38):10006–10018.
- [Li *et al.*, 2005] LI, S., LI, W. et FANG, T. (2005). An efficient fragment-based approach for predicting the ground-state energies and structures of large molecules. *Journal of the American Chemical Society*, 127(19):7215–7226.
- [Li *et al.*, 2006] LI, X., VOLKOV, A. V., SZALEWICZ, K. et COPPENS, P. (2006). Interaction energies between glycopeptide antibiotics and substrates in complexes determined by x-ray crystallography : application of a theoretical databank of aspherical atoms and a symmetry-adapted perturbation theory-based set of interatomic potentials. *Acta Crystallographica Section D*, 62(6):639–647.
- [Li *et al.*, 2002] LI, X., WU, G., ABRAMOV, Y. A., VOLKOV, A. V. et COPPENS, P. (2002). Application of charge density methods to a protein model compound : calculation of coulombic intermolecular interaction energies from the experimental charge density. *Proceedings of the National Academy of Sciences*, 99(19):12132–12137.
- [Lide, 2009] LIDE, D. R. (2009). *CRC handbook of chemistry and physics*. CRC press, 90 édition.
- [Liebschner *et al.*, 2013] LIEBSCHNER, D., DAUTER, M., BRZUSZKIEWICZ, A. et DAUTER, Z. (2013). On the reproducibility of protein crystal structures : five atomic resolution structures of trypsin. *Acta Crystallographica Section D : Biological Crystallography*, 69(8):1447–1462.
- [Liebschner *et al.*, 2009] LIEBSCHNER, D., ELIAS, M., MONIOT, S., FOURNIER, B., SCOTT, K., JELSCH, C., GUILLOT, B., LECOMTE, C. et CHABRIERE, E. (2009). Elucidation of the phosphate binding mode of ding proteins revealed by subangstrom x-ray crystallography. *Journal of the American Chemical Society*, 131(22):7879–7886.
- [Liebschner *et al.*, 2011] LIEBSCHNER, D., JELSCH, C., ESPINOSA, E., LECOMTE, C., CHABRIERE, E. et GUILLOT, B. (2011). Topological analysis of hydrogen bonds and weak in-

- teractions in protein helices via transferred experimental charge density parameters. *The Journal of Physical Chemistry A*, 115(45):12895–12904.
- [Liu *et al.*, 2006] LIU, B., SCHOFIELD, C. J. et WILMOUTH, R. C. (2006). Structural analyses on intermediates in serine protease catalysis. *Journal of Biological Chemistry*, 281(33):24024–24035.
- [Liu *et al.*, 2021] LIU, S., ZHONG, L.-P., HE, J. et ZHAO, Y.-X. (2021). Targeting neuropilin-1 interactions is a promising anti-tumor strategy. *Chinese Medical Journal*, 134(05):508–517.
- [London, 1937] LONDON, F. (1937). The general theory of molecular forces. *Transactions of the Faraday Society*, 33:8b–26.
- [Lonsdale, 1928] LONSDALE, K. (1928). The structure of the benzene ring. *Nature*, 122(3082):810–810.
- [Lorensen et Cline, 1987] LORENSEN, W. E. et CLINE, H. E. (1987). Marching cubes : A high resolution 3D surface construction algorithm. *SIGGRAPH Computer Graphics*, 21(4):163–169.
- [Lyssenko *et al.*, 2008] LYSENKO, K. A., BARZILOVICH, P. Y., ALDOSHIN, S. M., ANTIPIN, M. Y. et DOBROVOLSKY, Y. A. (2008). The role of h-bonds in charge transfer in the crystal of 1, 5-naphthalenedisulfonic acid tetrahydrate. *Mendeleev Communications*, 6(18):312–314.
- [Lyssenko *et al.*, 2007] LYSENKO, K. A., BORISSOVA, A. O., BURLOV, A. S., VASILCHENKO, I. S., GARNOVSKII, A. D., MINKIN, V. I. et ANTIPINA, M. Y. (2007). Interplay of the intramolecular n–h · · · n bond and π -stacking interaction in 2-(2'-tosylaminophenyl) benzimidazoles. *Mendeleev Communications*, 17(3):164–166.
- [Ma et Politzer, 2004] MA, Y. et POLITZER, P. (2004). Determination of noncovalent interaction energies from electronic densities. *The Journal of chemical physics*, 120(19):8955–8959.
- [Macchi, 2020] MACCHI, P. (2020). The connubium between crystallography and quantum mechanics. *Crystallography Reviews*, 26(4):209–268.
- [Macetti et Genoni, 2019] MACETTI, G. et GENONI, A. (2019). Quantum mechanics/extremely localized molecular orbital method : A fully quantum mechanical embedding approach for macromolecules. *The Journal of physical chemistry A*, 123(43):9420–9428.
- [Macetti *et al.*, 2021] MACETTI, G., WIEDUWILT, E. K. et GENONI, A. (2021). Qm/elmo : A multi-purpose fully quantum mechanical embedding scheme based on extremely localized molecular orbitals. *The Journal of Physical Chemistry A*, 125(13):2709–2726.
- [Malaspina *et al.*, 2019] MALASPINA, L. A., WIEDUWILT, E. K., BERGMANN, J., KLEEMISS, F., MEYER, B., RUIZ-LÓPEZ, M. F., PAL, R., HUPF, E., BECKMANN, J., PILTZ, R. O., EDWARDS, A. J., GRABOWSKY, S. et GENONI, A. (2019). Fast and accurate quantum crystallography : from small to large, from light to heavy. *The journal of physical chemistry letters*, 10(22):6973–6982.
- [Malcolm et Popelier, 2003] MALCOLM, N. O. J. et POPELIER, P. L. A. (2003). An improved algorithm to locate critical points in a 3d scalar field as implemented in the program morphy. *Journal of computational chemistry*, 24(4):437–442.
- [Malinska et Dauter, 2016] MALINSKA, M. et DAUTER, Z. (2016). Transferable aspherical atom model refinement of protein and dna structures against ultrahigh-resolution x-ray data. *Acta Crystallographica Section D*, 72(6):770–779.

- [Malińska *et al.*, 2014] MALIŃSKA, M., JARZEMBSKA, K. N., GORAL, A. M., KUTNER, A., WOŹNIAK, K. et DOMINIAK, P. M. (2014). Sunitinib : from charge-density studies to interaction with proteins. *Acta Crystallographica Section D*, 70(5):1257–1270.
- [Malinska *et al.*, 2015] MALINSKA, M., KUTNER, A. et WOŹNIAK, K. (2015). Predicted structures of new vitamin d receptor agonists based on available x-ray structures. *Steroids*, 104:220–229.
- [Massa *et al.*, 1995] MASSA, L., HUANG, L. et KARLE, J. (1995). Quantum crystallography and the use of kernel projector matrices. *International Journal of Quantum Chemistry*, 56(S29):371–384.
- [Massa, 2004] MASSA, W. (2004). *Crystal Structure Determination*. Springer Berlin Heidelberg, 2nd edition. édition.
- [Mata *et al.*, 2015] MATA, I., MOLINS, E., ALKORTA, I. et ESPINOSA, E. (2015). The paradox of hydrogen-bonded anion–anion aggregates in oxoanions : A fundamental electrostatic problem explained in terms of electrophilic · · · nucleophilic interactions. *The Journal of Physical Chemistry A*, 119(1):183–194.
- [Mata *et al.*, 2007] MATA, I., MOLINS, E. et ESPINOSA, E. (2007). Zero-flux surfaces of the electrostatic potential : The border of influence zones of nucleophilic and electrophilic sites in crystalline environment. *The Journal of Physical Chemistry A*, 111(39):9859–9870.
- [Matsuoka *et al.*, 2015] MATSUOKA, D., SUGIYAMA, S., MURATA, M. et MATSUOKA, S. (2015). Molecular dynamics simulations of heart-type fatty acid binding protein in apo and holo forms, and hydration structure analyses in the binding cavity. *The Journal of Physical Chemistry B*, 119(1):114–127.
- [Matta, 2014] MATTA, C. F. (2014). Modeling biophysical and biological properties from the characteristics of the molecular electron density, electron localization and delocalization matrices, and the electrostatic potential. *Journal of Computational Chemistry*, 35(16):1165–1198.
- [Matta et Boyd, 2007] MATTA, C. F. et BOYD, R. (2007). *The Quantum Theory of Atoms in Molecules : From Solid State to DNA and Drug Design*. Wiley-VCH.
- [Matta *et al.*, 2006] MATTA, C. F., CASTILLO, N. et BOYD, R. J. (2006). Extended weak bonding interactions in dna : π -stacking (base- base), base- backbone, and backbone- backbone interactions. *The Journal of Physical Chemistry B*, 110(1):563–578.
- [McGrath *et al.*, 1992] MCGRATH, M. E., VASQUEZ, J. R., CRAIK, C. S., YANG, A., HONIG, B. et FLETTERICK, R. J. (1992). Perturbing the polar environment of asp102 in trypsin : consequences of replacing conserved ser214. *Biochemistry*, 31(12):3059–3064.
- [Ménard et Storer, 1992] MÉNARD, R. et STORER, A. C. (1992). Oxyanion hole interactions in serine and cysteine proteases. *Biological Chemistry*, 373(2):393–400.
- [Merz Jr *et al.*, 2010] MERZ JR, K. M., RINGE, D. et REYNOLDS, C. H. (2010). *Drug design : structure-and ligand-based approaches*. Cambridge University Press.
- [Meyer et Genoni, 2018] MEYER, B. et GENONI, A. (2018). Libraries of extremely localized molecular orbitals. 3. construction and preliminary assessment of the new databanks. *The Journal of Physical Chemistry A*, 122(45):8965–8981.

- [Meyer *et al.*, 2016a] MEYER, B., GUILLOT, B., RUIZ-LOPEZ, M. F. et GENONI, A. (2016a). Libraries of extremely localized molecular orbitals. 1. model molecules approximation and molecular orbitals transferability. *Journal of Chemical Theory and Computation*, 12(3):1052–1067.
- [Meyer *et al.*, 2016b] MEYER, B., GUILLOT, B., RUIZ-LOPEZ, M. F., JELSCH, C. et GENONI, A. (2016b). Libraries of extremely localized molecular orbitals. 2. comparison with the pseudoatoms transferability. *Journal of chemical theory and computation*, 12(3):1068–1081.
- [Misquitta et Stone, 2016] MISQUITTA, A. J. et STONE, A. J. (2016). Ab initio atom–atom potentials using camcasp : Theory and application to many-body models for the pyridine dimer. *Journal of chemical theory and computation*, 12(9):4184–4208.
- [Mocchetti *et al.*, 2023] MOCCHETTI, E., DIDIERJEAN, C. et GUILLOT, B. (2023). New descriptors of protein structures electrostatic properties based on the topography of electric field lines. *Not yet submitted*.
- [Mocchetti *et al.*, 2022] MOCCHETTI, E., MORETTE, L., MULLIERT, G., MATHIOT, S., GUILLOT, B., DEHEZ, F., CHAUVAT, F., CASSIER-CHAUVAT, C., BROCHIER-ARMANET, C., DIDIERJEAN, C. et HECKER, A. (2022). Biochemical and structural characterization of chi-class glutathione transferases : A snapshot on the glutathione transferase encoded by sll0067 gene in the cyanobacterium *synechocystis* sp. strain pcc 6803. *Biomolecules*, 12(10):1466.
- [Mohan *et al.*, 2013] MOHAN, N., SURESH, C. H., KUMAR, A. et GADRE, S. R. (2013). Molecular electrostatics for probing lone pair– π interactions. *Physical Chemistry Chemical Physics*, 15(42):18401–18409.
- [Møller et Plesset, 1934] MØLLER, C. et PLESSET, M. S. (1934). Note on an approximation treatment for many-electron systems. *Physical review*, 46(7):618.
- [Morokuma, 1971] MOROKUMA, K. (1971). Molecular orbital studies of hydrogen bonds. iii. $\text{c}=\text{o} \cdots \text{h}-\text{o}$ hydrogen bond in $\text{h}_2\text{co} \cdots \text{h}_2\text{o}$ and $\text{h}_2\text{co} \cdots 2\text{h}_2\text{o}$. *The Journal of Chemical Physics*, 55(3):1236–1244.
- [Mous *et al.*, 2022] MOUS, S., GOTTHARD, G., EHRENBERG, D., SEN, S., WEINERT, T., JOHNSON, P. J., JAMES, D., NASS, K., FURRER, A., KEKILLI, D. *et al.* (2022). Dynamics and mechanism of a light-driven chloride pump. *Science*, 375(6583):845–851.
- [Munshi et Guru Row, 2005a] MUNSHI, P. et GURU ROW, T. N. (2005a). Charge density based classification of intermolecular interactions in molecular crystals. *CrystEngComm*, 7(100):608–611.
- [Munshi et Guru Row, 2005b] MUNSHI, P. et GURU ROW, T. N. (2005b). Evaluation of weak intermolecular interactions in molecular crystals via experimental and theoretical charge densities. *Crystallography Reviews*, 11(3):199–241.
- [Murray et Politzer, 2003] MURRAY, J. S. et POLITZER, P. (2003). The use of the molecular electrostatic potential in medicinal chemistry. In *Quantum medicinal chemistry*, pages 233–254. Wiley-VCH.
- [Muzet *et al.*, 2003] MUZET, N., GUILLOT, B., JELSCH, C., HOWARD, E. et LECOMTE, C. (2003). Electrostatic complementarity in an aldose reductase complex from ultra-high-resolution crys-

- tallography and first-principles calculations. *Proceedings of the National Academy of Sciences*, 100(15):8742–8747.
- [Nakano *et al.*, 2000] NAKANO, T., KAMINUMA, T., SATO, T., AKIYAMA, Y., UEBAYASI, M. et KITAURA, K. (2000). Fragment molecular orbital method : application to polypeptides. *Chemical Physics Letters*, 318(6):614–618.
- [Nassour *et al.*, 2017] NASSOUR, A., DOMAGALA, S., GUILLOT, B., LEDUC, T., LECOMTE, C. et JELSCH, C. (2017). A theoretical-electron-density databank using a model of real and virtual spherical atoms. *Acta Crystallographica Section B*, 73(4):610–625.
- [Nelyubina *et al.*, 2009] NELYUBINA, Y. V., ANTIPIN, M. Y. et LYSSENKO, K. A. (2009). Hydrogen bonds between zwitterions : intermediate between classical and charge-assisted ones. a case study. *The Journal of Physical Chemistry A*, 113(15):3615–3620.
- [Nguyen *et al.*, 2018] NGUYEN, D., KISIEL, Z. et VOLKOV, A. (2018). Fast analytical evaluation of intermolecular electrostatic interaction energies using the pseudoatom representation of the electron density. i. the löwdin α -function method. *Acta Crystallographica Section A*, 74(5):524–536.
- [Pathak et Gadre, 1990] PATHAK, R. K. et GADRE, S. R. (1990). Maximal and minimal characteristics of molecular electrostatic potentials. *The Journal of chemical physics*, 93(3):1770–1773.
- [Patkowski, 2020] PATKOWSKI, K. (2020). Recent developments in symmetry-adapted perturbation theory. *Wiley Interdisciplinary Reviews : Computational Molecular Science*, 10(3):e1452.
- [Pellet-Many *et al.*, 2008] PELLET-MANY, C., FRANKEL, P., JIA, H. et ZACHARY, I. (2008). Neuropilins : structure, function and role in disease. *Biochemical Journal*, 411(2):211–226.
- [Petrungolo et Tomasi, 1975] PETRONGOLO, C. et TOMASI, J. (1975). The use of the electrostatic molecular potential in quantum pharmacology. i. ab initio results. *International Journal of Quantum Chemistry*, 9(S2):181–190.
- [Phipps *et al.*, 2015] PHIPPS, M. J., FOX, T., TAUTERMANN, C. S. et SKYLARIS, C.-K. (2015). Energy decomposition analysis approaches and their evaluation on prototypical protein–drug interaction patterns. *Chemical society reviews*, 44(10):3177–3211.
- [Pichon-Pesme *et al.*, 2000] PICHON-PESME, V., LACHEKAR, H., SOUHASSOU, M. et LECOMTE, C. (2000). Electron density and electrostatic properties of two peptide molecules : Tyrosyl-glycyl-glycine monohydrate and glycyl-aspartic acid dihydrate. *Acta Crystallographica Section B*, 56(4):728–737.
- [Pichon-Pesme *et al.*, 1995] PICHON-PESME, V., LECOMTE, C. et LACHEKAR, H. (1995). On building a data bank of transferable experimental electron density parameters applicable to polypeptides. *The Journal of Physical Chemistry*, 99(16):6242–6250.
- [Piquemal *et al.*, 2006] PIQUEMAL, J., CISNEROS, G. A., REINHARDT, P., GRESH, N. et DARDEN, T. A. (2006). Towards a force field based on density fitting. *The Journal of chemical physics*, 124(10):104101.
- [Piquemal *et al.*, 2007] PIQUEMAL, J.-P., CHEVREAU, H. et GRESH, N. (2007). Toward a separate reproduction of the contributions to the hartree-fock and dft intermolecular interaction

- energies by polarizable molecular mechanics with the sibfa potential. *Journal of Chemical Theory and Computation*, 3(3):824–837.
- [Politzer, 1988] POLITZER, P. (1988). Computational approaches to the identification of suspect toxic molecules. *Toxicology Letters*, 43(1-3):257–276.
- [Rackers et Ponder, 2019] RACKERS, J. A. et PONDER, J. W. (2019). Classical pauli repulsion : An anisotropic, atomic multipole model. *The Journal of chemical physics*, 150(8):084104.
- [Radisky et al., 2006] RADISKY, E. S., LEE, J. M., LU, C.-J. K. et KOSHLAND, D. E. (2006). Insights into the serine protease mechanism from atomic resolution structures of trypsin reaction intermediates. *Proceedings of the National Academy of Sciences*, 103(18):6835–6840.
- [Ranaghan et al., 2019] RANAGHAN, K. E., SHCHEPANOVSKA, D., BENNIE, S. J., LAWAN, N., MACRAE, S. J., ZUREK, J., MANBY, F. R. et MULHOLLAND, A. J. (2019). Projector-based embedding eliminates density functional dependence for qm/mm calculations of reactions in enzymes and solution. *Journal of chemical information and modeling*, 59(5):2063–2078.
- [Rezác et al., 2011a] REZÁC, J., RILEY, K. E. et HOBZA, P. (2011a). Extensions of the s66 data set : more accurate interaction energies and angular-displaced nonequilibrium geometries. *Journal of Chemical Theory and Computation*, 7(11):3466–3470.
- [Rezác et al., 2011b] REZÁC, J., RILEY, K. E. et HOBZA, P. (2011b). S66 : A well-balanced database of benchmark interaction energies relevant to biomolecular structures. *Journal of Chemical Theory and Computation*, 7(8):2427–2438.
- [Roetti et Clementi, 1974] ROETTI, C. et CLEMENTI, E. (1974). Simple basis sets for molecular wavefunctions containing atoms from $z=2$ to $z=54$. *The Journal of Chemical Physics*, 60(12):4725–4729.
- [Rybicka et al., 2022] RYBICKA, P. M., KULIK, M., CHODKIEWICZ, M. L. et DOMINIAK, P. M. (2022). Multipolar atom types from theory and statistical clustering (matts) data bank : Impact of surrounding atoms on electron density from cluster analysis. *Journal of Chemical Information and Modeling*, 62(16):3766–3783.
- [Sahu et Gadre, 2014] SAHU, N. et GADRE, S. R. (2014). Molecular tailoring approach : a route for ab initio treatment of large clusters. *Accounts of chemical research*, 47(9):2739–2747.
- [Saleh et al., 2012] SALEH, G., GATTI, C., LO PRESTI, L. et CONTRERAS-GARCÍA, J. (2012). Revealing non-covalent interactions in molecular crystals through their experimental electron densities. *Chemistry—A European Journal*, 18(48):15523–15536.
- [Salem, 1961] SALEM, L. (1961). The forces between polyatomic molecules. ii. short-range repulsive forces. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, 264(1318):379–391.
- [Schiebel et al., 2018] SCHIEBEL, J., GASPARI, R., WULSDORF, T., NGO, K., SOHN, C., SCHRAEDER, T. E., CAVALLI, A., OSTERMANN, A., HEINE, A. et KLEBE, G. (2018). Intriguing role of water in protein-ligand binding studied by neutron crystallography on trypsin complexes. *Nature communications*, 9(1):1–15.
- [Schmidt et al., 2003] SCHMIDT, A., JELSCH, C., ØSTERGAARD, P., RYPNIEWSKI, W. et LAMZIN, V. S. (2003). Trypsin revisited : crystallography at (sub) atomic resolution and quantum chemistry revealing details of catalysis. *Journal of Biological Chemistry*, 278(44):43357–43362.

- [Schmidt *et al.*, 2011] SCHMIDT, A., TEETER, M., WECKERT, E. et LAMZIN, V. S. (2011). Crystal structure of small protein crambin at 0.48 Å resolution. *Acta Crystallographica Section F*, 67(4):424–428.
- [Schrödinger, 1926] SCHRÖDINGER, E. (1926). An undulatory theory of the mechanics of atoms and molecules. *Physical review*, 28(6):1049.
- [Senn et Thiel, 2009] SENN, H. M. et THIEL, W. (2009). Qm/mm methods for biomolecular systems. *Angewandte Chemie International Edition*, 48(7):1198–1229.
- [Shi *et al.*, 2013] SHI, Y., XIA, Z., ZHANG, J., BEST, R., WU, C., PONDER, J. W. et REN, P. (2013). Polarizable atomic multipole-based amoeba force field for proteins. *Journal of chemical theory and computation*, 9(9):4046–4063.
- [Shirsat *et al.*, 1992] SHIRSAT, R. N., BAPAT, S. V. et GADRE, S. R. (1992). Molecular electrostatics. a comprehensive topographical approach. *Chemical physics letters*, 200(4):373–378.
- [Sironi *et al.*, 2007] SIRONI, M., GENONI, A., CIVERA, M., PIERACCINI, S. et GHITTI, M. (2007). Extremely localized molecular orbitals : theory and applications. *Theoretical Chemistry Accounts*, 117(5):685–698.
- [Slater et Kirkwood, 1931] SLATER, J. C. et KIRKWOOD, J. G. (1931). The van der waals forces in gases. *Physical Review*, 37(6):682.
- [Söderhjelm *et al.*, 2006] SÖDERHJELM, P., KARLSTRÖM, G. et RYDE, U. (2006). Comparison of overlap-based models for approximating the exchange-repulsion energy. *The Journal of chemical physics*, 124(24):244101.
- [Spackman, 1986a] SPACKMAN, M. A. (1986a). Atom–atom potentials via electron gas theory. *The Journal of chemical physics*, 85(11):6579–6586.
- [Spackman, 1986b] SPACKMAN, M. A. (1986b). A simple quantitative model of hydrogen bonding. *The Journal of chemical physics*, 85(11):6587–6601.
- [Spackman, 1987] SPACKMAN, M. A. (1987). A simple quantitative model of hydrogen bonding : application to more complex systems. *Journal of Physical Chemistry*, 91(12):3179–3186.
- [Spackman, 1998] SPACKMAN, M. A. (1998). Charge densities from x-ray diffraction data. *Annual Reports Section C*, 94:177–207.
- [Spackman, 2007] SPACKMAN, M. A. (2007). Comment on on the calculation of the electrostatic potential, electric field and electric field gradient from the aspherical pseudoatom model by volkov, king, coppens & farrugia (2006). *Acta Crystallographica Section A*, 63(2):198–200.
- [Spackman et Maslen, 1986] SPACKMAN, M. A. et MASLEN, E. N. (1986). Chemical properties from the promolecule. *The Journal of Physical Chemistry*, 90(10):2020–2027.
- [Sparrow *et al.*, 2021] SPARROW, Z. M., ERNST, B. G., JOO, P. T., LAO, K. U. et DiSTASIO JR, R. A. (2021). Nenci-2021. i. a large benchmark database of non-equilibrium non-covalent interactions emphasizing close intermolecular contacts. *The Journal of Chemical Physics*, 155(18):184303.
- [Sprang *et al.*, 1987] SPRANG, S., STANDING, T., FLETTERICK, R., STROUD, R., FINER-MOORE, J., XUONG, N., HAMLIN, R., RUTTER, W. et CRAIK, C. (1987). The three-dimensional

- structure of asn102 mutant of trypsin : role of asp102 in serine protease catalysis. *Science*, 237(4817):905–909.
- [Stalke, 2011] STALKE, D. (2011). Meaningful structural descriptors from charge density. *Chemistry–A European Journal*, 17(34):9264–9278.
- [Stone, 2013] STONE, A. (2013). *The theory of intermolecular forces*. Oxford university press, 2nd édition.
- [Suydam *et al.*, 2006] SUYDAM, I. T., SNOW, C. D., PANDE, V. S. et BOXER, S. G. (2006). Electric fields at the active site of an enzyme : Direct comparison of experiment with theory. *science*, 313(5784):200–204.
- [Svensson *et al.*, 1996] SVENSSON, M., HUMBEL, S., FROESE, R. D. J., MATSUBARA, T., SIEBER, S. et MOROKUMA, K. (1996). Oniom : a multilayered integrated mo+ mm method for geometry optimizations and single point energy predictions. a test for diels- alder reactions and pt (p (t-bu) 3) 2+ h2 oxidative addition. *The Journal of Physical Chemistry*, 100(50):19357–19363.
- [Sylvestre-Gonon *et al.*, 2022] SYLVESTRE-GONON, E., MORETTE, L., VILORIA, M., MATHIOT, S., BOUTILLIAT, A., FAVIER, F., ROUHIER, N., DIDIERJEAN, C. et HECKER, A. (2022). Biochemical and structural insights on the poplar tau glutathione transferase gstu19 and 20 paralogs binding flavonoids. *Frontiers in Molecular Biosciences*, 9.
- [Szalewicz, 2012] SZALEWICZ, K. (2012). Symmetry-adapted perturbation theory of intermolecular forces. *Wiley interdisciplinary reviews : computational molecular science*, 2(2):254–272.
- [Szalewicz et Jeziorski, 1979] SZALEWICZ, K. et JEZIORSKI, B. (1979). Symmetry-adapted double-perturbation analysis of intramolecular correlation effects in weak intermolecular interactions : the He-He interaction. *Molecular Physics*, 38(1):191–208.
- [Thomas *et al.*, 2017] THOMAS, S. E., MENDES, V., KIM, S. Y., MALHOTRA, S., OCHOA-MONTAÑO, B., BLASZCZYK, M. et BLUNDELL, T. L. (2017). Structural biology and the design of new therapeutics : from hiv and cancer to mycobacterial infections : a paper dedicated to john kendrew. *Journal of molecular biology*, 429(17):2677–2693.
- [Toukmaji et Board Jr, 1996] TOUKMAJI, A. Y. et BOARD JR, J. A. (1996). Ewald summation techniques in perspective : a survey. *Computer physics communications*, 95(2-3):73–92.
- [Tsirelson et Ozerov, 1996] TSIRELSON, V. G. et OZEROV, R. P. (1996). *Electron Density and Bonding in Crystals : Principles, Theory and X-ray Diffraction Experiments in Solid State Physics and Chemistry*. Taylor & Francis.
- [Tzeliou *et al.*, 2022] TZELIOU, C. E., MERMIGKI, M. A. et TZELI, D. (2022). Review on the qm/mm methodologies and their application to metalloproteins. *Molecules*, 27(9):2660.
- [Van Vleet *et al.*, 2018] VAN VLEET, M. J., MISQUITTA, A. J. et SCHMIDT, J. (2018). New angles on standard force fields : Toward a general approach for treating atomic-level anisotropy. *Journal of Chemical Theory and Computation*, 14(2):739–758.
- [Vennelakanti *et al.*, 2022] VENNELAKANTI, V., NAZEMI, A., MEHMOOD, R., STEEVES, A. H. et KULIK, H. J. (2022). Harder, better, faster, stronger : Large-scale qm and qm/mm for predictive modeling in enzymes and proteins. *Current opinion in structural biology*, 72:9–17.

- [Veszprémi et Fehér, 1999] VESZPRÉMI, T. et FEHÉR, M. (1999). *Quantum chemistry : fundamentals to applications*. Springer Science & Business Media.
- [Volkov et Coppens, 2007] VOLKOV, A. et COPPENS, P. (2007). Response to spackman's comment on the calculation of the electrostatic potential, electric field and electric field gradient from the aspherical pseudoatom model. *Acta Crystallographica Section A*, 63(2):201–203.
- [Volkov *et al.*, 2006] VOLKOV, A., KING, H. F., COPPENS, P. et FARRUGIA, L. J. (2006). On the calculation of the electrostatic potential, electric field and electric field gradient from the aspherical pseudoatom model. *Acta Crystallographica Section A*, 62(5):400–408.
- [Volkov *et al.*, 2004a] VOLKOV, A., KORITSANSZKY, T. et COPPENS, P. (2004a). Combination of the exact potential and multipole methods (ep/mm) for evaluation of intermolecular electrostatic interaction energies with pseudoatom representation of molecular electron densities. *Chemical physics letters*, 391(1-3):170–175.
- [Volkov *et al.*, 2004b] VOLKOV, A., LI, X., KORITSANSZKY, T. et COPPENS, P. (2004b). Ab initio quality electrostatic atomic and molecular properties including intermolecular energies from a transferable theoretical pseudoatom databank. *The Journal of Physical Chemistry A*, 108(19):4283–4300.
- [von Hopffgarten et Frenking, 2012] von HOPFFGARTEN, M. et FRENKING, G. (2012). Energy decomposition analysis. *Wiley Interdisciplinary Reviews : Computational Molecular Science*, 2(1):43–62.
- [von Laue, 1915] von LAUE, M. (1915). Concerning the detection of x-ray interferences. *Nobel lecture*, 13.
- [Vuković *et al.*, 2021] VUKOVIĆ, V., LEDUC, T., JELIĆ-MATOŠEVIĆ, Z., DIDIERJEAN, C., FAVIER, F., GUILLOT, B. et JELSCH, C. (2021). A rush to explore protein–ligand electrostatic interaction energy with charger. *Acta Crystallographica Section D : Structural Biology*, 77(10).
- [Wade *et al.*, 1998] WADE, R. C., GABDOULLINE, R. R., LÜDEMANN, S. K. et LOUNNAS, V. (1998). Electrostatic steering and ionic tethering in enzyme–ligand binding : Insights from simulations. *Proceedings of the National Academy of Sciences*, 95(11):5942–5949.
- [Wagner *et al.*, 2002] WAGNER, A., FLAIG, R., ZOBEL, D., DITTRICH, B., BOMBICZ, P., STRÜMPPEL, M., LUGER, P., KORITSANSZKY, T. et KRANE, H.-G. (2002). Structure and charge density of a c60-fullerene derivative based on a high resolution synchrotron diffraction experiment at 100 k. *The Journal of Physical Chemistry A*, 106(28):6581–6590.
- [Wahlgren *et al.*, 2011] WAHLGREN, W. Y., PÁL, G., KARDOS, J., PORROGI, P., SZENTHE, B., PATTHY, A., GRÁF, L. et KATONA, G. (2011). The catalytic aspartate is protonated in the michaelis complex formed between trypsin and an in vitro evolved substrate-like inhibitor : a refined mechanism of serine protease action. *Journal of Biological Chemistry*, 286(5):3587–3596.
- [Wang *et al.*, 2013] WANG, X., HE, X. et ZHANG, J. Z. H. (2013). Predicting mutation-induced stark shifts in the active site of a protein with a polarized force field. *The Journal of Physical Chemistry A*, 117(29):6015–6023.

- [Warshel, 2014] WARSHEL, A. (2014). Multiscale modeling of biological functions : from enzymes to molecular machines (nobel lecture). *Angewandte Chemie International Edition*, 53(38): 10020–10031.
- [Warshel et Levitt, 1976] WARSHEL, A. et LEVITT, M. (1976). Theoretical studies of enzymic reactions : dielectric, electrostatic and steric stabilization of the carbonium ion in the reaction of lysozyme. *Journal of molecular biology*, 103(2):227–249.
- [Watson et Crick, 1953] WATSON, J. D. et CRICK, F. H. (1953). Molecular structure of nucleic acids : a structure for deoxyribose nucleic acid. *Nature*, 171(4356):737–738.
- [Weatherly *et al.*, 2021] WEATHERLY, J., MACCHI, P. et VOLKOV, A. (2021). On the calculation of the electrostatic potential, electric field and electric field gradient from the aspherical pseudatom model. ii. evaluation of the properties in an infinite crystal. *Acta Crystallographica Section A*, 77(5):399–419.
- [Weiner *et al.*, 1982] WEINER, P. K., LANGRIDGE, R., BLANEY, J. M., SCHAEFER, R. et KOLLMAN, P. A. (1982). Electrostatic potential molecular surfaces. *Proceedings of the National Academy of Sciences*, 79(12):3754–3758.
- [Williams et Cox, 1984] WILLIAMS, D. E. et COX, S. R. (1984). Nonbonded potentials for aza-hydrocarbons : the importance of the coulombic interaction. *Acta Crystallographica Section B*, 40(4):404–417.
- [Yang, 1991] YANG, W. (1991). Direct calculation of electron density in density-functional theory. *Physical review letters*, 66(11):1438.
- [Yoffe et Maggiora, 1980] YOFFE, J. A. et MAGGIORA, G. M. (1980). The london approximation and the calculation of dispersion interactions as a sum of atom-atom terms. *Theoretica chimica acta*, 56:191–198.
- [Zarychta *et al.*, 2015] ZARYCHTA, B., LYUBIMOV, A., AHMED, M., MUNSHI, P., GUILLOT, B., VRIELINK, A. et JELSCH, C. (2015). Cholesterol oxidase : ultrahigh-resolution crystal structure and multipolar atom model-based analysis. *Acta Crystallographica Section D*, 71(4):954–968.
- [Zhang et Zhang, 2003] ZHANG, D. W. et ZHANG, J. Z. H. (2003). Molecular fractionation with conjugate caps for full quantum mechanical calculation of protein–molecule interaction energy. *The Journal of chemical physics*, 119(7):3599–3605.

Abréviations

ADN : acide désoxyribonucléique

aEP/MM : "analytical exact potential and multipole model" ou méthode analytique du potentiel exact et modèle multipolaire

AMBER : "assisted model building and energy refinement" ou construction de modèle assistée et affinement d'énergie

AMOEBA : "atomic multipole optimized energetics for biomolecular simulation" ou énergétique optimisée par les multipôles atomiques pour les simulation de biomolécules

ARN : acide ribonucléique

BCP : "bond critical point" ou point critique de liaison

BP : "bond path" ou chemin de liaison

CC : "coupled cluster" ou cluster couplé

CCP : "cage critical point" ou point critique de cage

CHARMM : "chemistry at Harvard molecular mechanics" ou chimie pour la mécanique moléculaire d'Harvard

CRM2 : laboratoire de cristallographie, résonance magnétique et modélisation

DFT : "density functional theory" ou théorie de la fonctionnelle de la densité

EDA : "energy decomposition analysis" ou analyse de la décomposition de l'énergie

EFP : "effective fragment potential" ou potentiel de fragments effectifs

ELF : "electron localization function" ou fonction de localisation des électrons

ELMAM : "experimental library mutlipolar atom model" ou librairie expérimentale du modèle d'atomes multipolaires

ELMO : "extremely localized molecular orbital" ou orbitale moléculaire extrêmement localisée

EP/MM : "exact potential and multipole model" ou potentiel exact et modèle multipolaire

FABP : "fatty acid binding protein" ou protéine de fixation d'acides gras

FMO : "fragment molecular orbital" ou orbitale moléculaire fragmentée

GEM : "Gaussian electrostatic model" ou modèle électrostatique gaussien

GGA : "generalized gradient approximation" ou approximation du gradient généralisé

GID : "generalized Invariom database" ou base de données d'atomes invariants (Invarioms) généralisée

GROMACS : "Groningen machine for chemical simulations" ou machine de Groningen pour les simulations chimiques

GST : glutathion transférase

GSTC : glutathion transférase de classe Chi

- HAR** : "Hirshfeld atom refinement" ou affinement atomique de Hirshfeld
- HF** : Hartree-Fock
- IAM** : "independent atom model" ou modèle des atomes indépendant
- IAS** : "interatomic scatterers" ou diffuseurs interatomiques
- KEM** : "kernel energy method" ou méthode d'énergie de cœur
- LDA** : "local density approximation" ou approximation de la densité locale
- LJ** : Lennard-Jonnes
- MAE** : "mean absolute error" ou erreur absolue moyenne
- MASTIFF** : "multipolar anisotropic Slater-type intermolecular force field" ou champ de force intermoléculaire anisotropique multipolaire de type Slater
- MATTS** : "multipolar atom types from theory and statistical clustering" ou types atomiques multipolaires théoriques et par regroupement statistique
- MFCC** : "molecular fractionation with conjugated caps" ou fragmentation moléculaire avec capsules conjuguées
- MP** : Møller-Plesset
- nc-BCP** : "non-covalent bond critical point" ou point critique de liaison non-covalente
- NCI** : "non-covalent interaction" ou interaction non-covalente
- NCP** : "nucleus critical point" ou point critique de noyau
- NENCI-2021** : "non-equilibrium non-covalent interaction 2021" ou jeu de données d'énergies d'interaction non covalente hors équilibre de 2021
- nEP/MM** : "numerical exact potential and multipole model" ou méthode numérique du potentiel exact et modèle multipolaire
- NmHR** : halorhodopsine de l'organisme *Nonlabens marinus*
- NRP1** : neuropiline 1
- OPLS** : "optimized potentials for liquid simulations" ou potentiels optimisés pour les simulations de liquides
- PDB** : "protein data bank" ou base de données des protéines
- PSB** : "protonated Schiff base" ou base de Schiff protonée
- QM/MM** : "quantum mechanics / molecular mechanics" ou mécanique quantique / mécanique moléculaire
- QTAIM** : "quantum theory of atoms in molecules" ou théorie quantique des atomes dans les molécules
- RCP** : "ring critical point" ou point critique de cycle
- RMSD** : "root-mean-square deviation" ou déviation quadratique moyenne
- RPSB** : "retinal protonated Schiff base" ou base de Schiff protonée du rétinol
- SAPT** : "symmetry-adapted perturbation theory" ou théorie de la perturbation adaptée par symétrie
- SARS-CoV-2** : "severe acute respiratory syndrome coronavirus 2" ou coronavirus 2 du syndrome respiratoire aigu sévère
- SBP** : "solute binding protein" ou protéine de fixation de soluté
- SCDS** : "semiclassic density sum" ou somme des densités semi-classiques

SGTI : "S gregaria trypsin inhibitor" ou inhibiteur de la trypine de schistocerca gregaria (criquet pèlerin)

SIBFA : "sum of interacting fragment ab initio" sommation ab initio de fragments en interaction

TAAM : "transferable aspherical atom model" ou modèle des atomes asphériques transférables

UBDB : "University at Buffalo pseudoatom databank" ou base de données de pseudo-atomes de l'Université de Buffalo

VDW : van der Waals

VEGF : "vascular endothelial growth factor" ou facteur de croissance de l'endothélium vasculaire

VIH : virus de l'immunodéficience humaine

VIR : "real and virtual atom refinement" ou affinement des atomes réels et virtuels

XCW : "x-ray constrained wavefunction" ou fonction d'onde contrainte par les rayons X

ZIE : zone d'influence électrophile

ZIN : zone d'influence nucléophile

Résumé / Abstract

Résumé

L'extension des approches de cristallographie quantique aux macromolécules biologiques vise à décrire les propriétés de ces systèmes complexes en combinant à la fois données expérimentales de cristallographie et informations de nature quantique. En particulier, les paramètres multipolaires de la librairie ELMAM2 permettent de reconstruire une densité électronique précise des protéines par transfert de données cristallographiques issues d'expériences de diffraction des rayons X par des cristaux de petits peptides. Dans ce travail, de nouvelles méthodologies ont été développées sur la base de cette densité électronique reconstruite, pour être appliquées au domaine de la biologie structurale.

Le premier type de méthodologie repose sur le développement de descripteurs originaux issus de la topologie du potentiel électrostatique. Cette quantité est classiquement représentée en biologie structurale comme projetée sur une surface moléculaire et interprétée en termes de complémentarité électrostatique. L'analyse topologique de ce champ scalaire permet quant à elle une description tridimensionnelle des propriétés électrostatiques, grâce à la détermination des points critiques et des bassins topologiques du potentiel électrostatique. Les descripteurs graphiques qui ont été développés permettent de localiser les sites électrophiles et nucléophiles d'une molécule et de déterminer les surfaces entourant les faisceaux primaires des lignes de champ électrique joignant ces sites. La zone d'influence d'un site électrophile ou nucléophile correspond à l'union des faisceaux primaires qui lui sont associés et contient toutes les lignes de champ électrique émergeant de, ou convergeant vers, ce site. Un outil graphique a été implémenté dans le logiciel MoProViewer pour représenter ces zones d'influence et les faisceaux primaires dans les structures de protéines. Ces approches originales, analysables et interprétables grâce à la topographie des lignes de champ électrique, proposent un nouveau paradigme considérant le potentiel électrostatique d'une protéine dans son ensemble, en trois dimensions et à l'échelle de l'atome. L'intérêt de ces descripteurs pour la compréhension des processus de reconnaissance moléculaire protéine-ligand, des mécanismes enzymatiques et structuraux a été montré au travers de leur application à différents systèmes protéiques.

Le second objectif de ce travail a été de construire un modèle d'énergie d'interaction basé sur les paramètres de densité électronique et les tenseurs de polarisabilité atomique transférables de la librairie ELMAM2. L'énergie d'interaction peut être divisée en quatre termes à savoir les termes électrostatique, d'induction, de dispersion et enfin d'échange-répulsion. Les contributions électrostatique et d'induction de ce modèle ELMAM avaient déjà été établies dans de précédents travaux. Le terme d'énergie de dispersion a été modélisé sur la base de l'approximation de

London utilisant les tenseurs de polarisabilités anisotropes atomiques transférés. La contribution d'échange-répulsion a été définie à partir du modèle de recouvrement des densités électroniques, basé sur les paramètres multipolaires transférés de la librairie ELMAM2. La somme de ces deux contributions constitue le potentiel d'interaction de van der Waals du modèle ELMAM. La validation de ce potentiel par comparaison aux valeurs d'énergie théoriques dans un jeu de données de benchmarking a montré une erreur absolue moyenne de 1,00 kcal/mol. Appliqués en biologie structurale, les estimations de ces différentes énergies d'interaction entre le site actif d'une enzyme et son substrat permettront d'identifier les principaux points d'ancrage de ce dernier et de discuter la nature de ces interactions.

Abstract

The extension of quantum crystallography approaches to biological macromolecules aims to describe the properties of these complex systems by combining both experimental crystallographic data and quantum information. In particular, the multipolar parameters from the ELMAM2 library enable the reconstruction of accurate protein electron density by transferring crystallographic data obtained from X-ray diffraction experiments on small peptide crystals. In this work, new methodologies have been developed based on this reconstructed electron density for application in the structural biology field.

The first methodology involves the development of original descriptors derived from the topology of the electrostatic potential. In classical structural biology, this quantity is represented as projected onto molecular surface and is interpreted in terms of electrostatic complementarity. The topological analysis of this scalar field enables a three-dimensional description of electrostatic properties by determining critical points and topological basins of the electrostatic potential. The graphical descriptors that have been developed allow for the localization of electrophilic and nucleophilic sites of a molecule and the determination of surfaces surrounding the primary bundles of electric field lines connecting these sites. The influence zone of an electrophilic or nucleophilic site corresponds to the union of the primary bundles associated with it and contains all electric field lines emerging from, or converging onto, that site. A graphical tool has been implemented in the MoProViewer software to represent these influence zones and primary bundles in protein structures. These original approaches, analyzable and interpretable through the topography of electric field lines, propose a new paradigm considering the electrostatic potential of the whole protein, in three dimensions and at the atomic scale. The relevance of these descriptors for understanding protein-ligand molecular recognition processes, enzymatic mechanisms, and structural phenomena has been demonstrated through their application to different protein systems.

The second objective of this work was to construct an interaction energy model based on the electron density parameters and transferable atomic polarizability tensors from the ELMAM2 library. The interaction energy can be divided into four terms : electrostatic, induction, dispersion, and exchange-repulsion. The electrostatic and induction contributions of the ELMAM model had already been established in previous works. The dispersion energy term was modeled on the basis of the London approximation using transferred anisotropic atomic polarizability tensors. The exchange-repulsion contribution was defined based on the electron density overlap

model, using multipolar parameters transferred from the ELMAM2 library. The sum of these two contributions constitutes the van der Waals interaction potential of the ELMAM model. The validation of this potential by comparing it to theoretical energy values in a benchmarking dataset showed an average absolute error of 1.00 kcal/mol. When applied in structural biology, the estimation of these different interaction energies between the active site of an enzyme and its substrate will help identify the key binding sites and discuss the nature of these interactions.

Résumé vulgarisé

Dans le monde du vivant, les processus biologiques sont assurés par quatre familles de molécules : les protéines, les glucides, les lipides et les acides nucléiques. Ces molécules, souvent composées de plusieurs milliers d'atomes, interagissent entre elles pour permettre le bon fonctionnement des organismes. Aussi, l'étude de leurs interactions est essentielle pour comprendre leurs mécanismes d'action. Les charges portées par les atomes pouvant être soit positives soit négatives, elles sont capables de s'attirer et de se repousser mutuellement. Dans ce travail, de nouvelles méthodes sont proposées pour décrire la force des interactions entre ces charges. En partant d'approches existantes pour des molécules simples, des outils ont été développés pour être appliqués à des biomolécules complexes. Ces méthodes innovantes fournissent une information rapide contrairement aux approches plus théoriques qui demandent des ressources computationnelles importantes.

Plain language summary

In the living world, biological processes are carried out by four families of molecules : proteins, carbohydrates, lipids and nucleic acids. These molecules, often constituted of several thousand atoms, interact with each other to enable organisms to ensure the proper functioning of organisms. Studying their interactions is essential to understand their mechanisms of action. As the charges carried by atoms can be either positive or negative, they are capable of attracting and repelling each other. In this work, new methods are proposed to describe the strength of interactions between these charges. Starting from existing approaches for simple molecules, tools have been developed for application to complex biomolecules. These innovative methods provide rapid information compared to more theoretical approaches that require significant computational resources.