Approche mixte par résonance magnétique nucléaire, dichroïsme circulaire électronique et modélisation moléculaire pour l'étude structurale de petits peptides bioactifs.

Sebastien Menant

Normandie Université

UNIVERSITÉ DE ROUEN NORMANDIE

THÈSE

Pour obtenir le diplôme de doctorat

Spécialité **CHIMIE**

Préparée au sein de l'**Université de Rouen Normandie**

Approche mixte par résonance magnétique nucléaire, dichroïsme circulaire électronique et modélisation moléculaire pour l'étude structurale de petits peptides bioactifs.

Présentée et soutenue par

SEBASTIEN MENANT

Thèse soutenue le 20/12/2024

devant le jury composé de :

MME LAURE GUILHAUDIS	Maître de Conférences - Université de Rouen Normandie (URN)	Directeur de thèse
MME ISABELLE MILAZZO	Professeur des Universités - Université de Rouen Normandie (URN)	Co-directeur de thèse
M. CHRISTOPHE MORELL	Professeur des Universités - Université Claude Bernard - Lyon 1	Membre du jury
M. VINCENT TOGNETTI	Maître de Conférences HDR - Université de Rouen Normandie (URN)	Membre du jury
MME ELISE DUMONT	Professeur des Universités - Universite Cote d'Azur	Rapporteur du jury
M. OLIVIER LEQUIN	Professeur des Universités - UNIVERSITE PARIS 6 PIERRE ET MARIE CURIE	Rapporteur du jury

Thèse dirigée par **LAURE GUILHAUDIS** (CHIMIE ORGANIQUE, BIOORGANIQUE, REACTIVITE, ANALYSE) et **ISABELLE MILAZZO** (Université de Rouen Normandie)

ED 508 NC Normande de Chimie

Laboratoire Cobra — Chimie organique, Bioorganique, Réactivité et Analyse.

THÈSE

Pour obtenir le diplôme de doctorat

Spécialité CHIMIE

Préparée au sein de l'Université de Rouen-Normandie

Approche mixte par résonance magnétique nucléaire, dichroïsme circulaire électronique et modélisation moléculaire pour l'étude structurale de petits peptides bioactifs

Présentée et soutenue par

Sébastien Menant

Thèse soutenue le 20/12/2024

devant le jury composé de		
Mme Élise Dumont	Professeure, Université Côte d'Azur	Rapporteure
M. Olivier Lequin	Professeur, Université Pierre et Marie Curie	Rapporteur
M. Christophe Morell	Professeur, Université Claude Bernard	Examinateur
M. Vincent Tognetti	Maître de Conférences, Université de Rouen-Normandie	Co-encadrant
Mme Isabelle Milazzo	Professeure, Université de Rouen-Normandie	Co-directrice de thèse
Mme Laure Guilhaudis	Maître de Conférences, Université de Rouen-Normandie	Directrice de thèse

Thèse dirigée par Laure Guilhaudis, Isabelle Milazzo et Vincent Tognetti – UMR 6014 COBRA

Remerciements :

Je tiens d'abord à profondément remercier tous les membres jurys, dont les rapporteurs Pr. Olivier Lequin et Pr. Élise Dumont, pour avoir accepté d'évaluer mes travaux de thèse, et l'examinateur, le Pr. Christophe Morell pour avoir accepté d'être mon enseignant référent et de présider le jury.

Ensuite je voudrais véritablement remercier mes encadrants de thèse, à savoir le Dr. Laure Guilhaudis pour avoir dirigé ma thèse, le Pr. Isabelle Milazzo, pour l'avoir co-dirigée, et le Dr. Vincent Tognetti pour son co-encadrement. Merci pour tout le temps que vous m'avez consacré, parfois même pendant vos vacances, parfois tard le soir, et toujours dans le but de m'aider. Vous avez fait de cette thèse l'une de vos priorités, ce dont je vous suis reconnaissant. Merci à vous trois pour avoir accepté de me faire confiance et de m'avoir donné une véritable autonomie. Merci également de m'avoir aidé à prendre confiance en moi et pour toutes ces discussions scientifiques toujours très enrichissantes.

Je remercie le Pr. Pierre-Yves Renard, pour m'avoir accueilli au sein du laboratoire COBRA.

J'aimerais remercier le CRIANN, pour m'avoir permis d'effectuer une grande partie de mon travail de thèse, et tout particulièrement le très compétent Patrick Bousquet-Melou.

Je remercie Brigitte Deschrevel, pour m'avoir laissé l'opportunité d'accomplir des missions d'enseignement.

Je remercie également tous les permanents de l'équipe RMN, le Pr. Hassan Oulyadi, le Dr. Gaël Coadou, le Dr. Muriel Sebban et Lina pour notamment m'avoir aidé à construire de meilleures présentations orales lors des réunions.

Je voudrais aussi remercier Benoit Gaüzère et Florian Yger, pour les quelques discussions sur le Machine Learning que nous avons pu avoir.

Merci Victoria pour ta bonne humeur, même si tu étais pour l'Espagne à la coupe du monde. Merci Khaoula pour ces quelques discussions que nous avons pu avoir. J'espère que tu vas enfin en finir avec ces TS ! Merci Morgane, pour ta franchise et ta gentillesse, j'espère que tu te plairas ici. Merci à Aurélien, le plus gentil des hommes, merci pour tous ces moments que je n'oublierai pas. Ces soirées au Fut et ces sessions d'escalade, d'escapâtes et d'escaflette, ont toutes été illuminées par ta présence avec ton humour aux petits oignons. Merci à Gautier, pour

toutes ces discussions autour du ptit caf', qui font oublier l'espace de quelques minutes la dure réalité de la thèse. Merci pour ton soutien. Merci à Matthieu, à la fois le pilier du labo et le gars sûr. Merci pour ta pédagogie, ton humour et ton rire chaleureux. Merci à Fida, pour ta gentillesse et ta bienveillance, je te souhaite bon courage pour la poursuite de ta thèse, tu vas y arriver j'en suis convaincu. Merci à Salomé, le dernier pilier de la course. Merci pour tous ces bons moments de vie, souvent autour de gâteaux ou de jeu de société. Merci Antoine d'être un compétiteur hors pair. Merci de toujours nous rappeler que si le dépassement de soi est important, le dépassement des autres est primordial. Merci Delphine de m'avoir fait reprendre le sport, de m'avoir invité dans tous moments festifs et conviviaux, de m'avoir conseillé et d'avoir pris régulièrement de mes nouvelles. Tu es vraiment quelqu'un de bien et je suis très heureux d'avoir pu partager ces moments avec toi. Merci à Maria, d'avoir participé aux discussions très sérieuses du midi, merci pour toutes ces anecdotes folkloriques. Merci à Teddy d'apporter un peu de nouveauté dans ce labo, et de me remplacer pour le plein d'azote... Merci Guillaume pour donner un repère quand on s'est perdu dans le labo, grâce à ton rire j'ai toujours su retrouver le chemin de mon bureau. Merci aussi de nous rappeler que si à un moment donné, quelqu'un doit avoir de la chance, ça sera forcément toi. Merci à Théo d'être une personne juste formidable. Merci pour toutes ces rigolades avec ton rire si communicatif. Ton seul défaut, et pas des moindres, c'est malheureusement de toujours supporter les mauvaises équipes : G2, Lens, Zverev … o_o. Merci au nouveau Théo d'avoir participé à ces petits moments de convivialité le midi. Merci à Estelle, sans qui on croirait tous qu'on sait bien chanter. Merci pour l'organisation de tous ces moments que je n'oublierai pas. Enfin merci à Jason, pour ces soirées jeux et surtout pour m'avoir permis de me remettre en question sur mon main smash, j'espère que tu ne regrettes pas trop :/ ? Ah oui et allez les Verts !

Merci à toutes les personnes que j'ai pu croiser au laboratoire : Aël, Solweig, Guillaume, Valentina, Lisa, Olivier, Hend, Oscar, Maxime, Charlotte, Nathaniel, Anaïs.

Merci à mes amis du Master Chimie de Lyon et mes amis de prépa.

Merci à mes amis d'enfance de Sainté, merci pour tous ces souvenirs inoubliables … on se sait.

Merci à toute ma famille, mes oncles et mes tantes, mes cousins et ma cousine. Merci pour votre soutien et toutes ces franches rigolades autour de moments festifs.

Merci à mes deux grand-mères Denise et Jeannette, qui doivent se demander pourquoi je suis toujours à l'université alors que j'ai 28 ans. Merci pour l'amour que vous m'avez donné durant toutes ces années.

Merci à mon grand frère, d’avoir toujours cru en moi alors, même quand je n’avais pas encore trouvé ma voie.

Merci à mes parents, pour leur soutien inconditionnel, merci d’avoir été là depuis le début et de m’avoir soutenu dans tout ce que je voulais entreprendre. J’ai vraiment eu de la chance de vous avoir à mes côtés, et j’espère vous rendre un peu fier.

Enfin, merci à Camille. Merci d’avoir eu la patience de m’attendre ces 3 années. Merci de m’avoir supporté aussi bien dans les bons moments que dans les mauvais. Tu as été mon repère pendant tout ce temps. C’est grâce à toi si j’en suis là aujourd’hui. Merci pour ton soutien sans faille durant toutes ces années.

Table des matières générale

I. Introduction

Table des matières

I. A. Contexte général et motivations

Les protéines sont omniprésentes dans les systèmes biologiques. Leur fonction dans les cellules vivantes et dans les tissus est très variée et elles sont fondamentales dans une multitude de phénomènes. Leur diversité fonctionnelle est remarquable, allant de la catalyse des réactions biochimiques à des fonctions structurelles et de régulation.[1,2] Par exemple, des enzymes comme l'amylase catalysent l'hydrolyse de l'amidon et du glycogène en sucres simples, facilitant ainsi la digestion.[3] D'autres protéines, telles que la tubuline, contribuent à la formation du cytosquelette, offrant une structure et un support mécanique aux cellules.[4] Certaines protéines, comme la myosine, sont impliquées dans des mécanismes moteurs, tels que la contraction musculaire.[5]

Les protéines jouent également un rôle essentiel dans la régulation et la transmission des signaux cellulaires. Par exemple, les récepteurs couplés aux protéines G (RCPG) jouent un rôle clé dans la transmission des signaux à l'intérieur des cellules. Ce phénomène de transduction active des voies de signalisation essentielles pour la régulation de nombreuses fonctions biologiques.[6]

La taille de ces protides peut varier considérablement. Elle est de quelques kilodaltons (kDa) pour les petites protéines jusqu'à plusieurs milliers de kDa pour les plus grandes. Par exemple, la connectine est la plus grande protéine connue chez l'être humain, avec une taille de plus de 3000 kDa.[7] Les peptides, quant à eux, sont des séquences de taille plus petite et sont généralement composées de moins de 20 résidus.

Ces molécules biologiques sont toutes constituées de polymères d'acides aminés, reliés entre eux par des liaisons peptidiques. Leur fonction biologique dépend directement de leur structure tridimensionnelle, qui elle-même découle d'un repliement complexe organisé en plusieurs niveaux de structuration.[8–10] Chaque niveau (primaire, secondaire, tertiaire et quaternaire) influence les niveaux supérieurs. Ainsi, même des modifications mineures dans la séquence d'acides aminés peuvent avoir des répercussions sur la fonction protéique, parfois avec des conséquences biologiques significatives.[11,12]

Dans cette introduction, les structures des protéines étudiées au cours de cette thèse ont été redéfinies en partant d'une vue d'ensemble avant de se concentrer plus précisément sur les structures secondaires avec une attention particulière portée sur les coudes, qui constituent un des axes centraux de cette recherche. Cette revue contextuelle est suivie de l'énoncé des trois principaux objectifs de la thèse. Deux chapitres méthodologiques suivent ce chapitre

introductif, apportant les bases techniques nécessaires à la compréhension des résultats présentés dans les trois chapitres suivants, chacun étant dédié à des objectifs spécifiques.

I. A. 1. Structures des protéines

L'unité de base des protéines est l'acide aminé. Il existe 20 acides aminés protéinogènes standards, auxquels on peut ajouter la sélénocystéine et la pyrrolysine, souvent considérés comme à part car ils sont spécifiques à certaines protéines. Ces molécules sont composées des éléments C, H, O, N et parfois S (comme dans la cystéine). Chaque acide aminé se distingue par sa chaîne latérale (ou résidu R), qui lui des propriétés chimiques (hydrophobes, hydrophiles, chargées, etc.) spécifiques ou particulières.

La chaîne principale des acides aminés forme le squelette peptidique, composé de liaisons amides. Les angles dièdres φ (phi) et ψ (psi) décrivent le repliement de cette chaîne principale, tandis que les angles χ (chi) caractérisent la disposition des chaînes latérales (**Figure I.1**). Le carbone α désigne le carbone de la chaîne principale, porteur de la chaîne latérale et participent à chacun des angles cités précédemment.

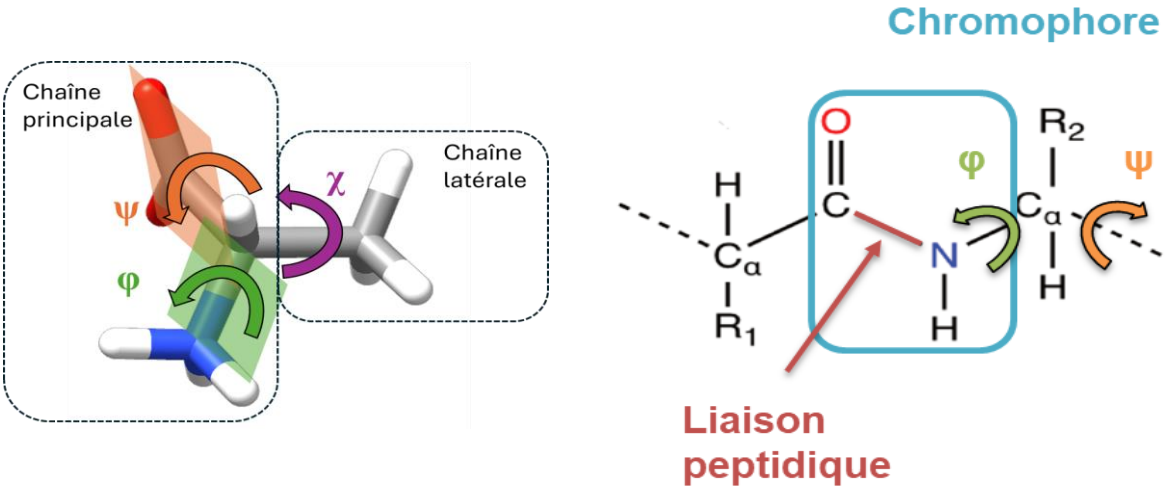
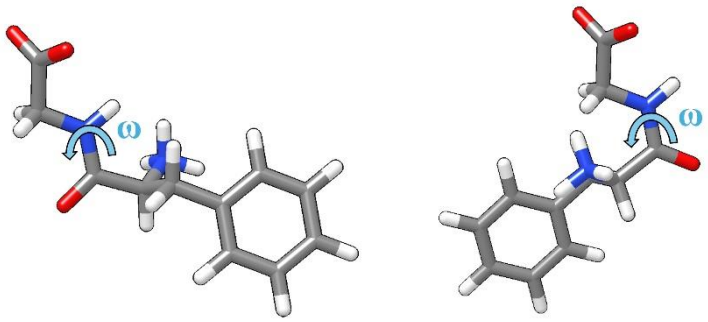


Figure I.1 *: Représentation d'un acide aminé (Alanine) à gauche et d'une liaison peptidique entre deux résidus à droite.*

La liaison peptidique entre deux acides aminés se forme par une réaction de condensation entre le groupe carboxyle d'un résidu et le groupement amine du suivant. Cette nouvelle fonction amide constitue un chromophore, entité qui va par exemple déterminer les caractéristiques optiques des protéines. L'angle dièdre associé à cette liaison est l'angle ω (**Figure I.2**), qui est souvent plus stable sous la forme *trans*.



trans *cis*

Figure I.2 *: Représentation de l'angle ω avec la conformation cis (à gauche) et trans (à droite) du peptide Phénylalanine-Glycine*

Ces polymères s'agencent entre eux dans une structure tridimensionnelle possédant plusieurs niveaux de structuration.

I. A. 1. a) Structure primaire

La structure primaire est la séquence linéaire d'acides aminés qui la compose et constitue le premier niveau de structuration des protéines. Elle définit complètement la séquence d'acides aminés en associant à chaque monomère sa place dans l'enchaînement polymérique et le nombre de résidus total (**Figure I.3**). La séquence primaire peut être écrite avec un code à trois lettres, et lorsque la séquence concerne les acides aminés naturels, il est fréquent d'utiliser la nomenclature associant à chaque acide aminé, une seule lettre.

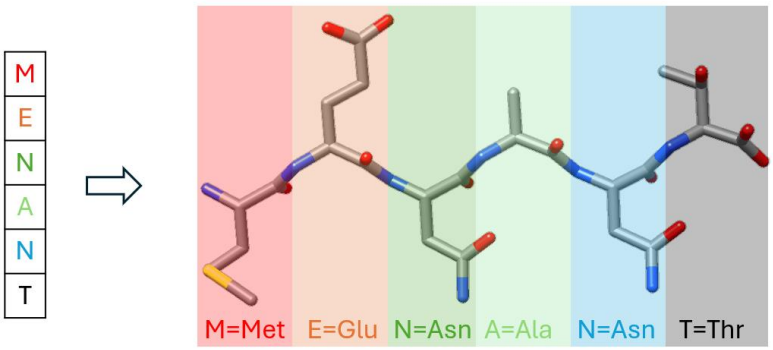


Figure I.3 *: Représentation de la structure primaire des protéines avec à gauche le code à 1 lettre et à droite, la représentation de la chaîne carbonée correspondante et son code à 3 lettres équivalent.*

I. A. 1. b) Structure secondaire

La structure secondaire constitue le premier niveau de repliement des chaînes peptidiques. Des peptides de même structure primaire peuvent adopter des repliements locaux différents en fonction des conditions d'observation. Il en existe quatre grandes familles que nous allons détailler. Ces structures sont définies principalement en fonction des angles dièdres φ et ψ adoptés le long de la chaîne peptidique. On peut distinguer les structures avec une répétition de ces angles sur de longs enchaînements d'acides aminés (hélice et feuillet) des structures uniques définis sur quelques résidus (coudes). Les structures régulières sont souvent reliées entre elles par des boucles riches en structure irrégulières.

Hélice

La famille des hélices est diversifiée. Elle est notamment composée des hélices α, 3_{10} et π (**Figure I.4**). Dans la majorité des cas, ces hélices tournent vers la droite du premier acide aminé vers le dernier, mais elles peuvent également tourner vers la gauche dans certains cas. Elles sont caractérisées par un nombre de résidus par tour ainsi qu'une distance par tour.

Les hélices α sont les structures secondaires les plus communes et sont définies pour des angles canoniques φ autour de -57 ° et ψ autour de -47 °. Elles sont stabilisées par une liaison hydrogène i, $i + 4$ sur chaque résidu i à l'exception des bords, ce qui permet un agencement intermédiaire avec 3,6 résidus par tour avec un pas de 5,4 Å.

Les hélices 3_{10} sont plus étirées et étroites avec seulement 3 résidus par tour et avec un pas de 6 Å. Ceci est possible grâce à une liaison hydrogène entre les résidus i et $i + 3$. Les angles dièdres φ et ψ associés à ce type d'hélice sont de -49 ° et -26 ° respectivement.

Les hélices π, beaucoup plus rares, sont plus larges avec 4,4 résidus par tour et un pas de 5 Å. La liaison hydrogène stabilisatrice est alors entre le résidu i et $i + 5$. L'angle φ est le même que pour les hélices α mais l'angle ψ descend à -70°.

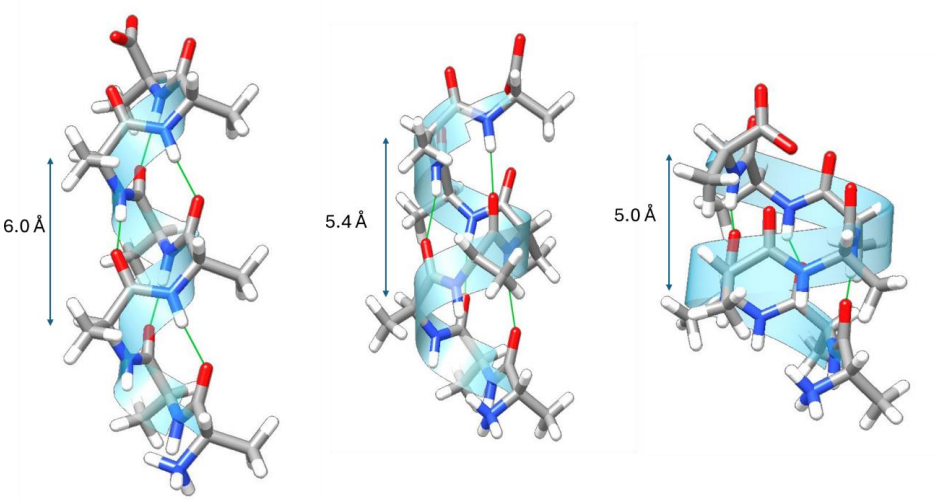


Figure I.4 *: Représentation d'une hélice 3_{10} à gauche, α au centre et π à droite avec en vert les liaisons hydrogène. Le pas de l'hélice est indiqué pour chaque hélice.*

Feuillet

Les feuillets sont construits par agencement latéral parallèle ou anti-parallèle de brins β, stabilisés par des liaisons hydrogène inter-brin (**Figure I.5**). Ces liaisons ont un angle plus favorable pour les formes anti-parallèles conduisant à une conformation généralement plus stable. Les brins sont constitués de résidus dans une conformation étendue. Les feuillets qui en résultent sont des repliements de ces brins, constituant des structures très denses. Ils sont caractérisés par des angles φ et ψ de -135 ° et 135 ° respectivement pour les feuillets anti-parallèles et -119 ° et +113 ° pour les feuillets parallèles.

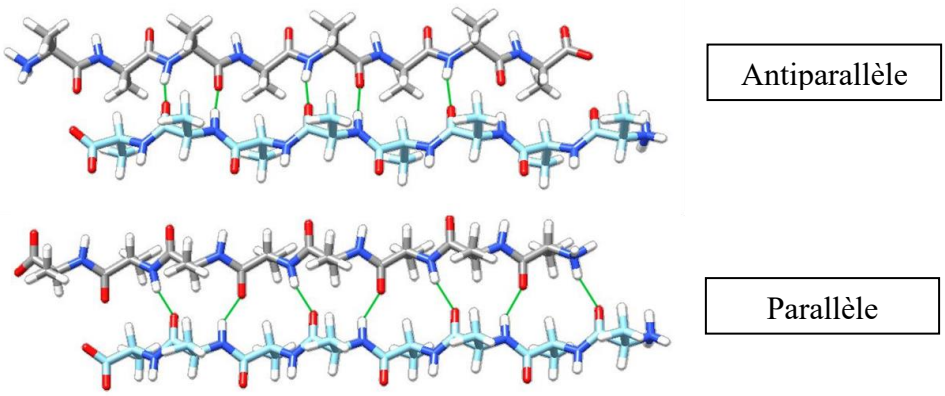


Figure I.5 *: Représentation des feuillets β antiparallèle et parallèle avec en vert les liaisons hydrogène.*

Polyproline-II

La structure polyproline-II (PPII) est à l'origine la conformation hélicoïdale qu'adoptent les enchaînements répétés de proline (**Figure I.6**). Elle est caractérisée par des angles dièdres φ et ψ de -75 ° et 150 ° respectivement.

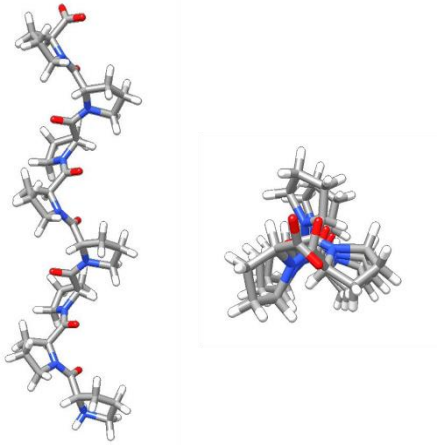


Figure I.6 *: Représentation d'une hélice polyproline-II. Vue longitudinale à gauche et axiale à droite.*

Puis, ces mêmes angles ont également été observés dans des séquences autres mais la nomenclature PPII est restée, désignant maintenant les structure avec ces angles dièdres.

Diagramme de Ramachandran

Les hélices, les feuillets et les conformations PPII font parties des motifs qui se répètent. Les brins β peuvent être formés jusqu'une dizaine d'acides aminés et les hélices peuvent aller au-delà. Classiquement, elles sont caractérisées sur le diagramme de Ramachandran qui représente les angles ψ en fonction des angles φ.[13] Les zones du diagramme de Ramachandran associées aux angles dièdres de chacune de ces conformations sont décrites **Figure I.7**.

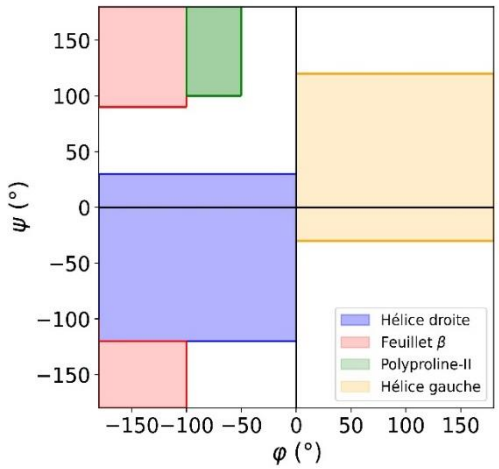


Figure I.7 *: Zones associées aux structures répétitives sur le diagramme de Ramachandran décrite d'après les références* ***14*** *et* ***15****.*

Coudes

Les coudes sont les structures les plus diversifiées. Ces derniers sont différenciés par le nombre d'acides aminés inclus dans la structure et par les angles dièdres φ et ψ adoptés. Les types δ, γ, β, α et π correspondent respectivement à une séquence de deux, trois, quatre, cinq et six acides aminés. Parmi eux, les coudes β sont les plus présents dans les protéines.[16] Les coudes sont la troisième structure secondaire la plus présente dans les protéines et représentent entre 20 % et 30 % des structures secondaire.[17] Ils constituent des zones de flexibilité, ce qui leur permet d'être responsables d'activités biologiques spécifiques. Cette structure étant au cœur de la thèse, deux autres sections lui sont consacrées l'introduction (**I. A. 2** et **I. A. 3**).

Toutes ces structures secondaires modifient les repliements locaux le long de la chaîne polypeptidique. Elles s'alternent avec aussi des séquences non structurées pour former le repliement global de la chaîne polypeptidique

I. A. 1. c) Structure tertiaire

La structure tertiaire (**Figure I.8**) représente le repliement complet d'une seule chaîne polypeptidique. Elle comprend souvent plusieurs structures secondaires qui peuvent être de famille différente. Elle fait intervenir des interactions de différentes natures (Hydrophobe, ionique, van der Waals, covalente) qui permettent de rendre la chaîne peptidique stable sous une conformation pour exercer une fonction biologique, par exemple la formation de sites actifs.

***Figure I.8**: Représentation de la structure tertiaire d'une protéine à travers l'exemple de l'α-amylase.*

I. A. 1. d) Structure quaternaire

Lorsque la protéine contient plusieurs chaînes polypeptidiques, elle est dite multimérique et possède une structure quaternaire **Figure I.9**. On distingue ces protéines par le nombre de sous-unités qu'elles contiennent. On parle ainsi de dimère, trimère, tétramère etc … La protéine tétramérique la plus connue est sans doute l'hémoglobine, qui est en fait hétéro-tétramérique (sous-unité identique deux à deux) et dont le fonctionnement dans le corps humain pour la captation d'oxygène dans les poumons et de son relâchement dans les tissus est bien connu.

Ces différents niveaux de structuration, très dépendants les uns des autres, sont donc primordiaux pour le repliement des protéines et conditionnent complètement leur activité biologique. Parmi les structures secondaires, les coudes ne sont pas encore correctement caractérisés. Ces motifs de petites tailles, sont difficiles à analyser en solution alors qu'ils participent aux repliements et sont des cibles de fonctions cellulaires. Dans cette thèse, nous parlerons davantage des coudes γ et β.

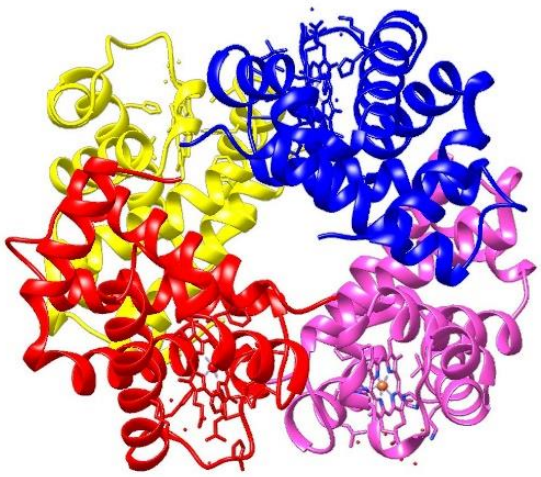


Figure I.9 *: Représentation de la structure quaternaire d'une protéine à travers l'exemple de l'hémoglobine. Chaque chaîne polypeptidique est associée à une couleur différente.*

I. A. 2. Classification des coudes

I. A. 2. a) Coudes γ

Les coudes γ concernent uniquement 3 résidus dont la conformation est stabilisée par une liaison hydrogène 1, 3. Ce sont de très petites structures classées en 2 types : le type γ classique ou seulement γ et le type γ inverse noté γ_{inv} **Figure I.10**.[18,19]

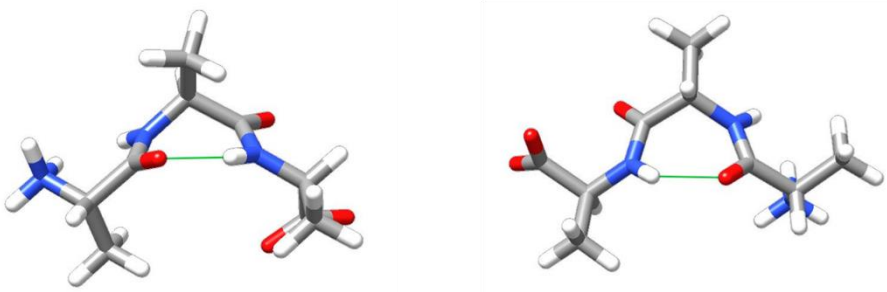


Figure I.10 *: Représentation des structures coudes de types γ à gauche et γ inverse à droite avec en vert les liaisons hydrogène.*

Les angles dièdres associés à ces types de structures sont ceux de l'acide aminé central, avec des angles φ et ψ de +75 ° et -64° respectivement pour les types classiques, et -79 ° et +69 ° pour les types inverses **Figure I.11**.

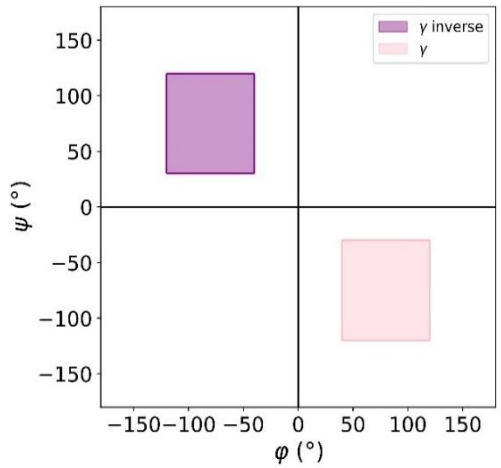


Figure I.11 *: Zones associées aux structures γ et γ inverse dans le diagramme de Ramachandran.*

I. A. 2. b) Coudes β

Les coudes β sont constitués de 4 résidus consécutifs et historiquement considérés stabilisés par une liaison hydrogène 1, 4 dont la taille du motif est inférieure à 7 Å, bien que la liaison hydrogène ne soit pas toujours présente. Ce sont les coudes les plus présents dans les protéines,

comptant pour plus de 20 % des résidus dans les protéines.[16,20,21] Au sein des boucles, ce sont deux tiers des résidus qui sont inclus dans un coude β.[22] Il en existe plusieurs types et les deux résidus centraux définissent totalement ces différents types.

La classification des coudes β a beaucoup évoluées au fil des années. Dès 1968, Venkatachalam avait exposé sa vision des coudes en analysant dans les coudes β, les valeurs que les angles dièdres des deux résidus centraux pouvaient prendre pour former une liaison hydrogène entre le premier et le dernier résidu.[23] Il en découle les 3 premiers types de coudes β-I, β-II et β-III avec leur image miroir β-I', β-II' et β-III'. Depuis, beaucoup de modifications ont été apportées. La classification la plus récente et complète est celle de De Brevern,[24] répertoriant 12 types coudes β (**Tableau I.1**), même si d'autres tentatives de classification ont été faites depuis.[22,25] Ces 12 types recouvrent une grande majorité du diagramme de Ramachandran (**Figure I.12**) et sont parfois superposés aux zones des autres structures secondaires, ce qui peut rendre leur identification complexe.

Tableau I.1 *: Angles dièdres φ et ψ associés à chaque catégorie des coudes de type β selon De Brevern.[24]*

Type	φ_{i+1} (°)	ψ_{i+1} (°)	φ_{i+2} (°)	ψ_{i+2} (°)
β-I	-60	-30	-90	0
β-I'	+60	+30	+90	0
β-II	-60	+120	+80	0
β-II'	+60	-120	-80	0
β-IV$_1$	-120	+130	+55	+41
β-IV$_2$	-85	-15	-125	+55
β-IV$_3$	-71	-30	-72	-47
β-IV$_4$	-97	-2	-117	-11
β-VI$_{a1}$	-60	+120	-90	0
β-VI$_{a2}$	-120	+120	-60	0
β-VI$_b$	-135	+135	-75	+160
β-VIII	-60	-30	-120	+120

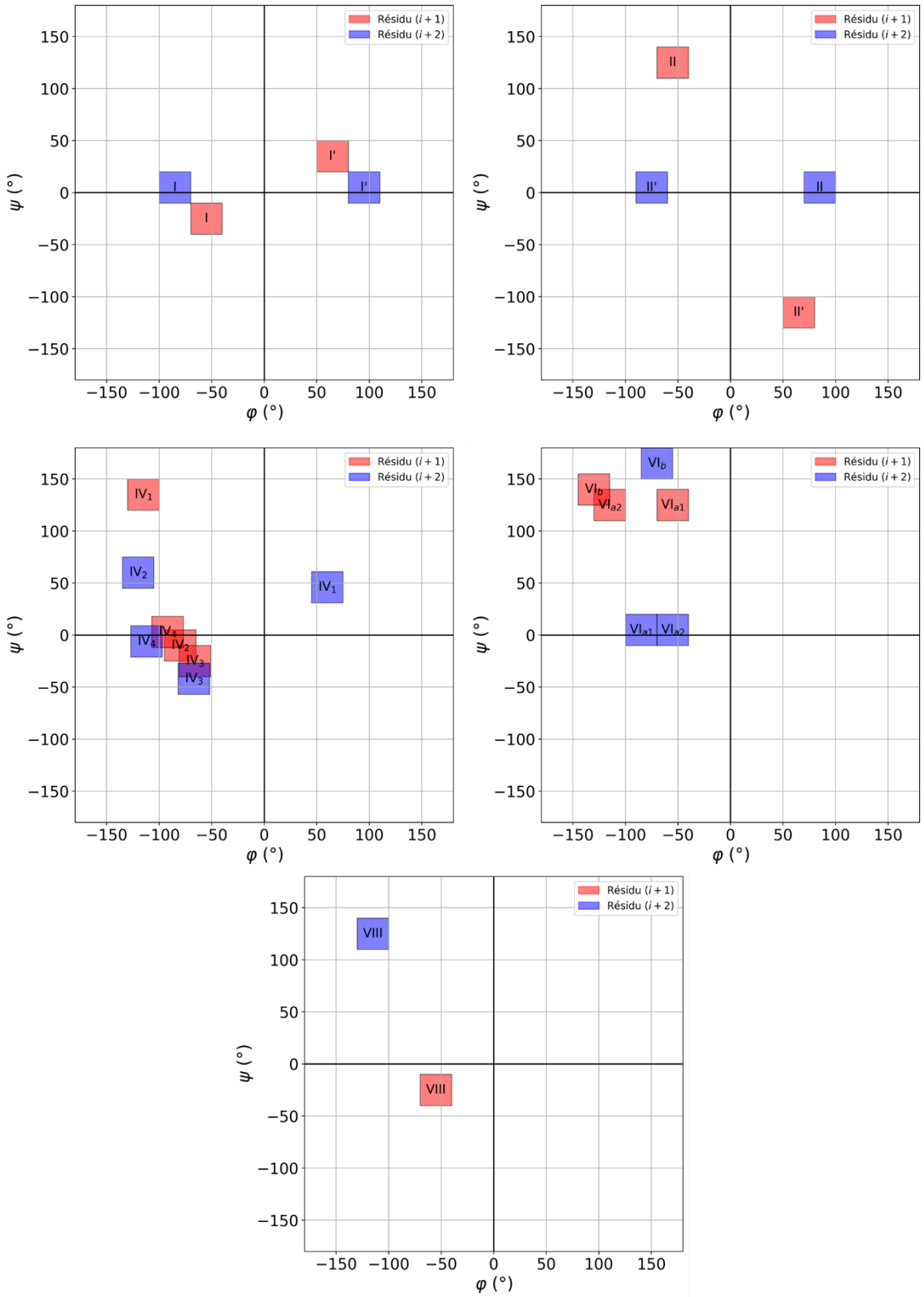


Figure I.12 *: Zones du diagramme de Ramachandran associées aux différentes catégories (**Tableau I.1**) des coudes de type β selon la classification de De Brevern.*[24]

I. A. 3. Importance des coudes

Les coudes, et en particulier les coudes β et γ, jouent un rôle essentiel dans le repliement des protéines, mais aussi dans la stabilité des structures tridimensionnelle et dans la fonction des protéines. Ces motifs structuraux, généralement constitués de quelques acides aminés, sont responsables des changements de direction dans la chaîne polypeptidique, facilitant ainsi la formation de structures compactes. Leur flexibilité et leur capacité à stabiliser les structures secondaires sont essentielles pour la fonctionnalité et la dynamique des protéines.

Les coudes confèrent de la flexibilité aux protéines, permettant des changements abrupts de direction au niveau des chaînes polypeptidiques et facilitant le repliement compact des protéines. Les coudes β sont très souvent situés à la surface des protéines,[26,27] reliant des éléments de structures secondaires comme les hélices et les feuillets à travers les boucles dans lesquelles ils représentent plus des deux tiers des résidus.[22]

Malgré leur flexibilité, les coudes jouent aussi un rôle important dans la stabilité conformationnelle des protéines. Les coudes stabilisent les structures en permettant la formation de liaisons hydrogène contribuant ainsi à la cohésion de la structure. Ils agissent aussi en favorisant la formation d'interactions à longue distance, en facilitant par exemple création de ponts disulfures ou d'interactions hydrophobes.

Les coudes participent également à la fonction biologique des protéines. En tant que régions souvent situées à la surface des protéines, ils sont bien placés pour participer à la reconnaissance moléculaire et aux interactions protéine-protéine. Ils peuvent jouer un rôle dans la formation de sites actifs enzymatiques et dans les interactions avec les ligands.[28,29]

Enfin ils jouent un rôle important dans la peptidomimétique, stratégie visant à imiter les structures peptidiques naturelles, en particulier dans le but de moduler ou de reproduire l'activité biologique des protéines. Les coudes, en tant que motifs structuraux clés dans les protéines, sont souvent la cible de cette approche.[30–32] Ces coudes, qui induisent des changements brusques de direction dans la chaîne polypeptidique, sont fréquemment impliqués dans la reconnaissance moléculaire et la stabilisation de l'architecture tridimensionnelle des protéines.[33] Les mimes de coude permettent de reproduire ces motifs dans des peptides de synthèse tout en améliorant leur stabilité ou leur biodisponibilité.

I. B. Objectifs

Les travaux de cette thèse visent à résoudre plusieurs problématiques liées à la caractérisation de la conformation des peptides très courts, dans lesquels la flexibilité est aujourd'hui un frein pour l'identification structurale. Dans ces molécules, les coudes occupent une place importante mais sont aussi des motifs très mal caractérisés.

I. B. 1. Développer la caractérisation des coudes en solution

En effet, bien que le Dichroïsme Circulaire Électroniques (DCE) soit la méthode de référence pour identifier les éléments de structures secondaires et déterminer leurs proportions en solution (DCE détaillé dans le chapitre II), il n'est aujourd'hui pas assez développé pour caractériser les différents types de coudes β et γ car il n'existe pas suffisamment de spectres DCE expérimentaux fiables. La complexité à prouver l'existence d'un coude en solution, sans que l'information soit masquée par la présence d'autres structures secondaires, est aujourd'hui le principal frein à l'utilisation du DCE pour la caractérisation des coudes. Les spectres DCE peuvent cependant être simulés par la théorie, et l'équipe du laboratoire COBRA a déjà proposé un protocole théorique pour simuler les spectres des coudes de types β-I, β-II, β-I' et β-II'.[34]

Le problème est maintenant de savoir à quel point ce protocole produit des spectres proches de la réalité et s'il est possible de se passer d'analyses expérimentales pour établir les spectres de référence de tous les types de coudes.

I. B. 2. Développer la caractérisation des molécules flexibles en solution

Ensuite, des techniques expérimentales et théoriques puissantes existent pour l'étude des structures protéiques, mais il demeure des limites importantes quant à leur capacité à caractériser précisément les molécules flexibles. Bien que les méthodes expérimentales actuelles, comme la RMN (Résonance Magnétique Nucléaire) ou le DCE, offrent des informations précieuses sur les conformations des peptides en solution, elles sont souvent limitées dans leur capacité à caractériser avec précision les structures dynamiques. Les données obtenues sont parfois insuffisantes pour capturer toute la complexité des échanges conformationnels de molécules flexibles. Les méthodes théoriques, comme la Dynamique Moléculaire (DM) ou les calculs quantiques de haute précision (DFT, TD-DFT), sont quant à elles très performantes pour prédire des conformations à l'échelle atomique. Toutefois, ces outils présentent également des limitations. En effet, ils sont souvent coûteux en temps de calcul

et leurs résultats peuvent manquer de précision si par exemple les interactions avec le solvant sont mal modélisées.

Un des objectifs de cette thèse sera donc de combler ces lacunes en développant des méthodes qui intègrent à la fois les données expérimentales et les modèles théoriques pour améliorer la prédiction des conformations des petits peptides en solution. L'idée est de trouver une synergie entre les approches théoriques, qui sont puissantes mais parfois coûteuses et complexes, et les approches expérimentales, plus directes mais souvent limitées dans la capture des états dynamiques.

Afin d'atteindre ces deux premiers objectifs, deux types de peptides seront étudiés :

- Un peptide structuré, ayant une conformation stable et bien définie. Ce peptide servira à optimiser le protocole théorique en testant la capacité des modèles à reproduire une structure stable et le spectre DCE qui en découle.
- Un peptide flexible, qui représente un défi plus complexe. Contrairement au peptide structuré, ce peptide adoptera plusieurs conformations en solution, ce qui nécessite une exploration plus large de l'espace conformationnel à l'aide de techniques de dynamique moléculaire et de calculs théoriques avancés.

I. B. 3. Amélioration de la vitesse et de la précision des prédictions théoriques

Un autre enjeu majeur est la vitesse d'exécution des protocoles théoriques actuels. Les calculs basés sur des méthodes de haute précision, comme les calculs de mécanique quantique couplés à des simulations de dynamique moléculaire, sont souvent lents et coûteux en ressources. Cette lenteur pose problème lorsque l'on cherche à modéliser des systèmes complexes ou des dynamiques sur des échelles de temps réalistes.

Le dernier objectif de ce travail sera donc de réduire le temps de calcul tout en maintenant la précision des prédictions, notamment par l'intégration d'approches de « *machine learning* ». Ces techniques permettront d'abord d'accélérer les simulations d'un paramètre RMN clé pour la détermination de structure peptidique : la constante de couplage $^{3}J_{HN}$-H_{α}.

Dans les deux chapitres suivants, nous exposerons en détail les méthodes expérimentales et théoriques utilisées lors de cette thèse, ainsi que les limites des méthodes actuelles pour parvenir à caractériser les coudes et les petites molécules flexibles.

I. C. Références

1. Videau, L. L., Arendall III, W. B. & Richardson, J. S. The cis-pro touch-turn: A rare motif preferred at functional sites. *Proteins: Struct., Funct., Bioinf.* **56**, 298–309 (2004).
2. Morris, R., Black, K. A. & Stollar, E. J. Uncovering protein function: from classification to complexes. *Essays Biochem.* **66**, 255–285 (2022).
3. Abedi, E., Kaveh, S. & Mohammad Bagher Hashemi, S. Structure-based modification of a-amylase by conventional and emerging technologies: Comparative study on the secondary structure, activity, thermal stability and amylolysis efficiency. *Food Chem.* **437**, 137903 (2024).
4. Hammond, J. W., Cai, D. & Verhey, K. J. Tubulin modifications and their cellular functions. *Curr. Opin. Cell Biol.* **20**, 71–76 (2008).
5. Hartman, M. A. & Spudich, J. A. The myosin superfamily at a glance. *J. Cell Sci.* **125**, 1627–1632 (2012).
6. Dorsam, R. T. & Gutkind, J. S. G-protein-coupled receptors and cancer. *Nat. Rev. Cancer* **7**, 79–94 (2007).
7. Maruyama, K. Connectin/titin, giant elastic protein of muscle. *FASEB J.* **11**, 341–345 (1997).
8. Easson, L. H. & Stedman, E. Studies on the relationship between chemical constitution and physiological action. *Biochem. J.* **27**, 1257–1266 (1933).
9. Orengo, C. A., Todd, A. E. & Thornton, J. M. From protein structure to function. *Curr. Opin. Struct. Biol.* **9**, 374–382 (1999).
10. Hegyi, H. & Gerstein, M. The relationship between protein structure and function: a comprehensive survey with application to the yeast genome 1. *J. Mol. Biol.* **288**, 147–164 (1999).

11. Liu, H. *et al.* A single residue substitution accounts for the significant difference in thermostability between two isoforms of human cytosolic creatine kinase. *Sci. Rep.* **6**, 21191 (2016).
12. Ben Soussia, I. *et al.* Mutation of a single residue promotes gating of vertebrate and invertebrate two-pore domain potassium channels. *Nat. Commun.* **10**, 787 (2019).
13. Ramachandran, G. N., Ramakrishnan, C. & Sasisekharan, V. Stereochemistry of polypeptide chain configurations. *J. Mol. Biol.* **7**, 95–99 (1963).
14. Pizzanelli, S. *et al.* Conformations of Phenylalanine in the Tripeptides AFA and GFG Probed by Combining MD Simulations with NMR, FTIR, Polarized Raman, and VCD Spectroscopy. *J. Phys. Chem. B* **114**, 3965–3978 (2010).
15. Avbelj, F. & Baldwin, R. L. Role of backbone solvation and electrostatics in generating preferred peptide backbone conformations: Distributions of phi. *Proc. Natl. Acad. Sci.* **100**, 5742–5747 (2003).
16. Koch, O. & Klebe, G. Turns revisited: A uniform and comprehensive classification of normal, open, and reverse turn families minimizing unassigned random chain portions. *Proteins: Struct., Funct., Bioinf.* **74**, 353–367 (2009).
17. Wang, J. *et al.* StraPep: a structure database of bioactive peptides. *Database* **2018**, bay038 (2018).
18. Matthews, B. W. The γ Turn. Evidence for a New Folded Conformation in Proteins. *Macromolecules* **5**, 818–819 (1972).
19. Milner-White, E. J. Situations of gamma-turns in proteins: Their relation to alpha-helices, beta-sheets and ligand binding sites. *J. Mol. Biol.* **216**, 385–397 (1990).
20. De Brevern, A. G. A Perspective on the (Rise and Fall of) Protein β-Turns. *Int. J. Mol. Sci.* **23**, 12314 (2022).

21. Bornot, A. & Brevern, A. G. D. Protein beta-turn assignments. *Bioinformation* **1**, 153–155 (2006).
22. Shapovalov, M., Vucetic, S. & Dunbrack, R. L. Jr. A new clustering and nomenclature for beta turns derived from high-resolution protein structures. *PLoS Comput. Biol.* **15**, e1006844 (2019).
23. Venkatachalam, C. M. Stereochemical criteria for polypeptides and proteins. V. Conformation of a system of three linked peptide units. *Biopolymers* **6**, 1425–1436 (1968).
24. De Brevern & G, A. Extension of the classical classification of β-turns. *Sci. Rep.* **6**, 33191 (2016).
25. Zhang, R., Stahr, M. C. & Kennedy, M. A. Introduction of a new scheme for classifying β-turns in protein structures. *Proteins: Struct., Funct., and Bioinf.* **90**, 110–122 (2022).
26. Rose, G. D., Glerasch, L. M. & Smith, J. A. Turns in Peptides and Proteins. in *Advances in Protein Chemistry* (eds. Anfinsen, C. B., Edsall, J. T. & Richards, F. M.) vol. 37 1–109 (Academic Press, 1985).
27. Guruprasad, K. & Rajkumar, S. gβ- and gγ-turns in proteins revisited: A new set of amino acid turn-type dependent positional preferences and potentials. *J. Biosci.* **25**, 143–156 (2000).
28. Brown, K. A. *et al.* Twists and turns: a tale of two shikimate-pathway enzymes. *Biochem. Soc. Trans.* **31**, 543–547 (2003).
29. Harris, K. L., Lim, S. & Franklin, S. J. Of Folding and Function: Understanding Active-Site Context through Metalloenzyme Design. *Inorg. Chem.* **45**, 10002–10012 (2006).
30. Caramiello, A. M., Bellucci, M. C., Marti-Rujas, J., Sacchetti, A. & Volonterio, A. Turn-Mimic Hydantoin-Based Loops Constructed by a Sequential Multicomponent Reaction. *J. Org. Chem.* **88**, 15790–15804 (2023).

31. Gou, F. *et al.* One-Pot Cyclization to Large Peptidomimetic Macrocycles by In Situ-Generated β-Turn-Enforced Folding. *J. Am. Chem. Soc.* **145**, 9530–9539 (2023).
32. Darapaneni, C. M., Ghosh, P., Ghosh, T. & Maayan, G. Unique β-Turn Peptoid Structures and Their Application as Asymmetric Catalysts. *Chem. – Eur. J.* **26**, 9573–9579 (2020).
33. Lahiri, P., Verma, H., Ravikumar, A. & Chatterjee, J. Protein stabilization by tuning the steric restraint at the reverse turn. *Chem. Sci.* **9**, 4600–4609 (2018).
34. Migliore, M. *et al.* Characterization of β-turns by electronic circular dichroism spectroscopy: a coupled molecular dynamics and time-dependent density functional theory computational study. *Phys. Chem. Chem. Phys.* **22**, 1611–1623 (2020).

II. Principes et limites

des méthodes expérimentales actuelles

Table des matières

II. A. Méthodes pour l'identification de structure en solution

La conformation adoptée par une molécule influence directement ses interactions avec d'autres molécules ou son environnement. Dans les protéines, cette relation structure-activité est conditionnée par l'édifice tri-dimensionnel, lui-même façonné par les différentes structures secondaires présentes. La détermination de la structuration est donc primordiale pour établir un lien entre la conformation adoptée par une protéine et son activité.[1] Les méthodes expérimentales actuelles telles que la RMN, le DC, la FT-IR et le Raman sont des outils extrêmement performant pour résoudre ces structures en solution à des échelles de précisions différentes. Par exemple, la RMN permet d'obtenir des informations à l'échelle inter-atomique, alors que les autres méthodes fournissent des informations moins locales sur les torsions du squelette peptidique ou sur l'environnement des liaisons hydrogènes.

En solution, les molécules évoluent dans leur solvant dans lequel des interactions sont possibles. Ces interactions peuvent être de nature intermoléculaire (solvant – soluté, soluté-soluté) ou intramoléculaire et peuvent conduire à une évolution de la conformation par rapport à celle en phase cristalline. Ceci est d'autant plus vrai lorsque que l'on parle de petites molécules qui peuvent être dotées d'une grande flexibilité. Nos travaux sont intégralement portés sur de très petits peptides ($< 500\ g.mol^{-1}$), dont certain pouvant procéder à des réarrangements en solution ce qui fait de la détermination de la conformation un réel défi.

Dans ce chapitre, nous allons d'abord aborder le principe des deux méthodes expérimentales utilisées lors de nos travaux que sont le DC et la RMN avec leur application pour l'élucidation de la structure des peptides et des protéines. Ensuite, nous présenterons les limites de ces techniques dans le cas de peptides présentant certaines structures secondaires ou dans le cas de peptides flexibles et les raisons pour lesquelles nous avons fait le choix de combiner ces approches expérimentales avec des approches théoriques

II. A. 1. Dichroïsme Circulaire (DC)

Le DC est la méthode la méthode de choix pour l'analyse des structures secondaires. C'est également une méthode très rapide, en solution et non destructive. Son signal, pouvant être positif ou négatif, offre une grande diversité de spectres même sur des gammes de longueurs d'onde restreintes.

II. A. 1. a) Principe

Le Dichroïsme Circulaire (DC) est une propriété optique que possèdent les molécules lorsqu'elles n'absorbent pas la Polarisation Circulaire Gauche (PCG) et la Polarisation Circulaire Droite (PCD) de la lumière de la même façon.[2] Cette différence électronique ou absorption est présente dans les molécules chirales et par conséquent dans les protéines en raison de l'asymétrie des carbones α. De plus, chaque liaison peptidique constitue un chromophore, entité qui interagit avec la lumière par délocalisation électronique. Cette différence, que l'on peut noter ΔA, est reliée aux absorptivités molaires de la PCG ε_g et de la PCD ε_d par la loi de Beer-Lambert :

$$\Delta A = (\varepsilon_g - \varepsilon_d) lC \tag{II.1}$$

Le dichroïsme circulaire $\Delta\varepsilon$ est la différence $(\varepsilon_g - \varepsilon_d)$. Elle est proportionnelle à l'ellipticité θ qui est la grandeur mesurée expérimentalement par les polariseurs. En supposant $\Delta A \ll 1$, ce qui est très souvent vérifié, le DC est proportionnel à l'ellipticité :

$$\Delta\varepsilon \propto \frac{\theta}{Cl} \tag{II.2}$$

qui peut être directement reliée à l'ellipticité molaire $[\theta]$:

$$[\theta] = 3298\, \Delta\varepsilon \tag{II.3}$$

avec $[\theta]$ en $deg\ cm^2 dmol^{-1}$. Le DC est dit Électronique (DCE) lorsque son application concerne les transitions électroniques situées dans le domaine UV-visible et Vibrationnel, (DCV) lorsqu'il concerne la région infrarouge.

II. A. 1. b) Signatures DC des structures secondaires

Le DCE étant très sensible aux changements de conformation, il constitue une méthode très largement utilisée aujourd'hui pour l'analyse du repliement des protéines.[3] Son principal intérêt est de permettre l'estimation de la proportion de structures secondaires en solution de manière très rapide et également d'étudier les changements rapides de conformation, ce qui est utile pour les études de cinétiques et les analyses de transitions conformationnelles en temps réel.[4]

Les régions classiquement utilisées pour les protéines sont la région UV lointaine $(170 - 250\ nm)$, dans laquelle sont observées les transitions électroniques $\pi \rightarrow \pi^*$ $(\sim 200\ nm)$ et $n \rightarrow \pi^*$ (~ 225 nm) fournissant des informations sur les structures secondaires. La région UV proche $(250 - 330\ nm)$ est plutôt utilisée pour l'analyse des aromatiques et des ponts disulfures.[5]

Les différentes grandes familles de structures secondaires ont des signatures CD différentes et facilement reconnaissable (**Figure II.1**) :

Les hélices α : 3 pics majoritaires dont un positif vers 190 nm puis deux négatifs à 208 nm et 222 nm. Le pic à 222 nm dépend de l'environnement autour des liaisons hydrogènes stabilisatrices $i, i+4$ caractéristiques des hélices α et n'est pas très sensible au nombre de résidus que comporte l'hélice contrairement aux deux autres pics.[3,6,7] Les autres types d'hélices 3_{10} ou π ont des allures similaires avec parfois des shifts ou des baisses d'intensité.

Les feuillets β : 2 pics majoritaires dont un positif entre 190 et 200 nm puis un négatif et large et moins intense entre 210 et 225 nm. Cependant, il existe une diversité de spectres pour cette famille du fait des arrangements parallèles, antiparallèles et mixtes.[7]

La conformation aléatoire : 1 unique signal majoritaire négatif entre 190 et 200 nm.

La polyproline-II : 2 pics majoritaires dont un négatif à 198 nm et un positif à 218 nm.[8,9]

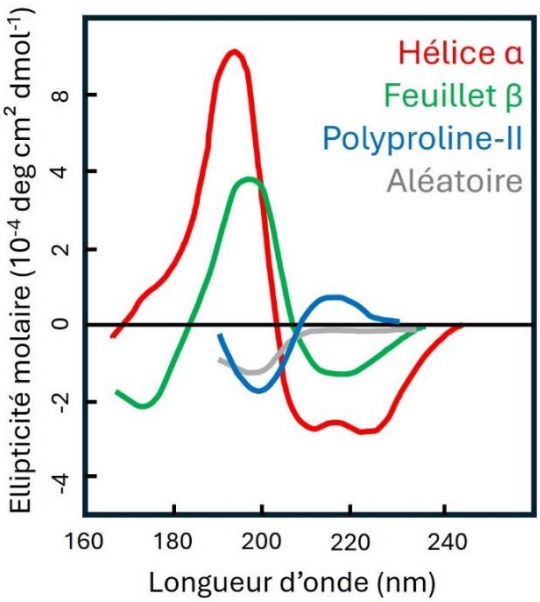


Figure II.1*: Spectres DCE de référence pour les hélices (courbe rouge), les feuillets (courbe verte), les conformations polyproline-II (courbe bleue) et les conformations aléatoires (courbe grise).*

Concernant les coudes, bien qu'il en existe une grande variété, il n'y a pas ou peu de spectres caractéristiques disponibles dans la littérature. Il n'existe par conséquent pas de spectre de référence pour les différents types de coudes.

II. A. 1. c) Méthodes d'analyse CD

La déconvolution des spectres CD est une méthode puissante pour déterminer les proportions de différentes structures secondaires telles que les hélices α, les feuillets β et les coudes dans les protéines. Bien que chaque méthode de déconvolution possède ses spécificités et convienne mieux à certaines structures, elles partagent toutes un principe commun : le spectre CD mesuré est une combinaison linéaire des spectres des structures secondaires, à laquelle s'ajoute un bruit ϵ_λ (contributions des groupes aromatiques, prosthétiques, etc.). Ce modèle peut s'exprimer ainsi :

$$\theta_\lambda = \sum_i c_i \theta_{\lambda,i} + \epsilon_\lambda \qquad \text{(II.4)}$$

avec θ_λ l'éllipticité mesurée de la protéine recherchée à la longueur d'onde λ, c_i la proportion de la structure secondaire i, $\theta_{\lambda,i}$ est l'ellipticité référence de la structure i à la longueur d'onde λ et ϵ_λ est l'erreur ou le bruit à la longueur d'onde λ. De ce principe découle plusieurs approches pour parvenir à estimer les coefficients c_i.[10] Ces approches sont intégrées dans des programmes de déconvolution, qui combinent l'algorithme de la méthode utilisée avec une base de données de spectres de référence.[11] Le choix du programme de déconvolution dépend principalement de la compatibilité entre la conformation de la molécule étudiée et des conformations représentées dans les bases de données. Ainsi, plusieurs ensembles de spectres de référence existent, mais ils ne sont pas tous compatibles avec chaque programme. En théorie, il est possible d'utiliser n'importe quel jeu de données avec n'importe quelle méthode de déconvolution, mais en pratique, certains programmes sont optimisés pour des bases de données spécifiques (par exemple des limitations sur la longueur d'onde), ce qui limite leur interopérabilité.

- Régression linéaire contrainte et non contrainte

C'est la technique de base pour la prédiction de structure secondaire et elle utilise le plus souvent la méthode des moindres carrés pour minimiser la fonction :

$$min \sum_\lambda \left[\theta_\lambda - \sum_i c_i \theta_{\lambda,i} + \epsilon_\lambda \right]^2 \qquad \text{(II.5)}$$

La base de données est souvent restreinte et focalisée sur des molécules similaires. Le calcul des spectres DCE de la poly-L-lysine avec différentes proportions d'hélice α et de feuillet β à

l'aide des spectres de référence poly-L-lysine 100 % hélice α, 100% feuillet β et 100% aléatoire en est un exemple très concret.[12]

Lorsque la régression est contrainte, la somme des coefficients doit en plus être égale à 1 en conservant des coefficients positifs, ce qui permet d'éviter des incohérences. Le programme LINCOMB[13] permet ce type d'approche.

- Régression ridge

Cette méthode, utilisée dans le programme CONTIN[4], est appliquée pour des ensemble de données plus large où la multicolinearité entre spectres peut être élevée. Elle ajoute une contrainte α (hyperparamètre) pour régulariser les coefficients c_i, permettant d'éviter que certains spectres de références dominent excessivement l'ajustement :

$$min\left[\sum_{\lambda}\left[\theta_\lambda - \sum_i c_i\theta_{\lambda,i} + \epsilon_\lambda\right]^2 + \alpha\sum_i {c_i}^2\right] \quad \text{(II.6)}$$

- Décomposition en valeurs singulières (SVD)

La décomposition en valeurs singulières (SVD) est une méthode mathématique qui permet de décomposer un ensemble de spectres de dichroïsme circulaire (CD) en une série de vecteurs singuliers (ou spectres de base).[14] En partant d'une matrice M contenant les spectres de i protéines mesurées sur j longueurs d'onde, on peut multiplier M par sa transposée M^T pour obtenir une matrice carrée de dimension $i \times i$. D'après le théorème spectral, cette matrice est diagonalisable et on peut donc écrire l'équation aux valeurs propres suivante :

$$MM^TU = UE \quad \text{(II.7)}$$

où U est la matrice des vecteurs propres et E est la matrice diagonale contenant les valeurs propres. Chaque vecteur propre U_i est une combinaison linéaire des spectres de référence, et les valeurs propres représentent l'importance relative de chaque vecteur propre.

Les spectres de base orthogonaux (ou composants principaux) sont générés à partir des vecteurs propres en utilisant l'opération suivante :

$$B = U^TM \quad \text{(II.8)}$$

où B contient les spectres de base. Chaque ligne de B correspond à un spectre de base, qui est une combinaison linéaire des spectres originaux dans M.

Ainsi, les spectres mesurés peuvent être décomposés en termes de ces spectres de base. Pour analyser un nouveau spectre T (de dimensions $j \times 1$) on le projette sur les spectres de cette base B. Les coefficients a_i obtenus quantifient combien chaque spectre de base contribue au spectre T, et sont donnés par l'équation :

$$a_i = B_i^T T \tag{II.9}$$

où a_i représente la contribution du i-ème spectre de base au spectre mesuré T.

La SVD fournit les meilleures estimations pour les hélices α, mais elle est moins performante pour les feuillets β et les coudes notamment si les données ne recouvrent pas une large gamme de longueur d'onde. Elle est surtout utilisée pour les protéines globulaires et est peu adaptée pour l'analyse de spectre de peptides ou de fragments protéiques.

- Sélection variable

Dans cette méthode, on commence par utiliser un large ensemble de données de référence. La structure de la protéine est d'abord estimée par SVD, puis certains spectres de références sont systématiquement éliminés pour créer des ensembles de données plus petits.[15] Ce processus est répété jusqu'à ce que des critères de performance acceptables soient atteints. Cette approche fournit des ajustements précis pour les protéines globulaires. Deux approches différentes sont fréquemment utilisées dans les programmes VARSLC[15] et CDSSTR[16].

- Méthode auto-cohérente

Dans cette approche initialement connue sous l'acronyme SELCON[17], le spectre mesuré est ajouté à l'ensemble des spectres de référence. Une estimation initiale des proportions de chaque structure secondaire est faite, puis ces proportions sont ajustées en réappliquant la SVD jusqu'à convergence (ou jusqu'à un nombre maximal d'itérations). Cette méthode est efficace pour les protéines globulaires. Les différentes versions du programme (SELCON2,[18] et SELCON3[19]) permettent d'évaluer plusieurs types de structures secondaires, y compris des conformations polyproline II, des hélices et des feuillets distordus. Cependant, SELCON reste moins performant pour l'estimation des coudes et n'est pas adapté à l'analyse des protéines autres que les protéines globulaires.

- Réseaux de neurones (Neural Networks)

Les réseaux de neurones sont des modèles d'intelligence artificielle capables de trouver des corrélations dans les données. Ils sont d'abord entraînés sur un ensemble de spectres CD de protéines dont la structure secondaire est connue. Une fois le réseau neuronal entraîné, il peut être utilisé pour analyser de nouveaux spectres et en déduire les proportions de structures secondaires. Les programmes comme CDNN[20] et K2D[21] utilisent cette approche et fournissent de bonnes estimations des hélices α et des feuillets β, mais ils sont moins adaptés à la détection des coudes. La programme K2D fournit de bonne estimation à la fois pour les protéines globulaires et pour les polypeptides.

- Algorithme de contrainte convexe (Convex Constraint Algorithm)

L'algorithme de contrainte convexe (CCA)[22] décompose un ensemble de spectres en un nombre défini de spectres de base. Ces spectres de base, lorsqu'ils sont recombinés, reproduisent l'ensemble des données avec une déviation minimale par rapport aux spectres originaux. CCA est particulièrement utile pour identifier les différents états structurels en fonction des changements induits par un ligand ou une variation de température. Toutefois, cette méthode est moins précise que la régression par moindres carrés, la SVD, ou les réseaux de neurones pour estimer la conformation protéique.

- Méthodes disponibles en ligne : DichroWeb et BeStSel

DichroWeb et BeStSel sont deux plateformes en ligne permettant d'analyser les spectres CD à l'aide des méthodes décrites précédemment. Elles n'utilisent pas les mêmes bases de données et ont des buts différents.

DichroWeb[23] : Cette plateforme utilise plusieurs des méthodes décrites ci-dessus, notamment la régression linéaire contrainte, CONTIN, VARSLC, et SELCON. L'utilisateur peut soumettre ses spectres et choisir la méthode qui convient le mieux pour extraire les structures secondaires. Il peut également choisir la base de données à condition qu'elle soit conforme au programme utilisé.

BeStSel[24] : Ce programme est spécialement conçu pour estimer les proportions de feuillets β parallèles et antiparallèles, en plus des autres structures secondaires, et est particulièrement efficace pour les protéines riches en feuillets β. Il utilise la méthode des moindres carrés avec une base de données en meilleur adéquation avec ces types de conformation.

Dans la pratique, plusieurs méthodes sont utilisées pour montrer la convergence vers une ou plusieurs structures secondaires majoritaires.

L'estimation de la proportion des différents coudes n'est pas proposée car les programmes de déconvolution présentés ci-dessus, lorsqu'ils prennent en compte le coude comme structure secondaire, se basent sur un spectre unique (**Figure II.2**). Ainsi ils utilisent des bases de données comme SP175,[25] où sont répertoriées des protéines aux structures et aux spectres DCE connus, pour optimiser des spectres de base ou spectres de références utilisés dans les analyses de déconvolution. Par exemple, l'algorithme DSSP (Define Secondary Structure of Proteins)[26] largement utilisé, inclut les coudes dans une catégorie unique de « *turn* » sans distinguer les sous-types de coudes. Cette classification unique des coudes ne reflète pas la diversité aujourd'hui reconnue, ce qui limite la précision des analyses basées sur ces spectres de base.

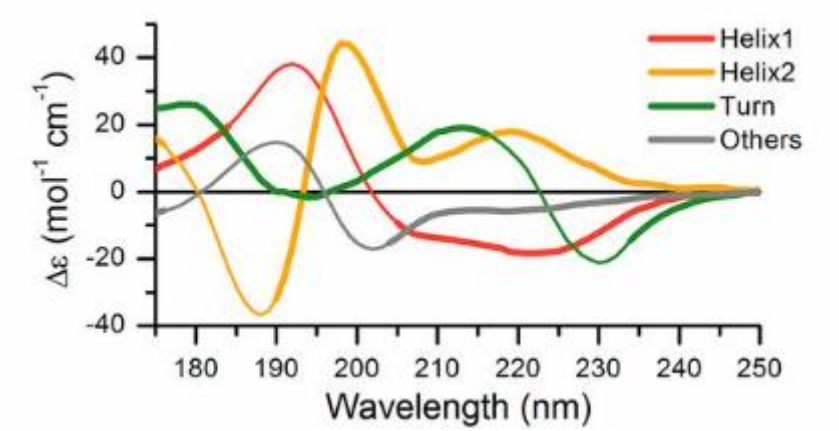


Figure II.2 *: Spectres de référence optimisés d'après la base de données SP175+ dans le programme BestSel.[27] Un seul spectre de référence est utilisé (courbe verte) pour la catégorie coude (« turn »).*

II. A. 2. La Résonance Magnétique Nucléaire (RMN)

La RMN est une propriété qu'ont certains noyaux qui, placés dans un champ magnétique, s'alignent sur ce champ en suivant un mouvement de précession autour de lui. C'est un outil très puissant pour identifier les molécules en solution et les repliements qu'elles adoptent dans leur milieu. Dans le cas de l'analyse de peptides et de petites protéines, c'est la RMN du proton qui est principalement utilisée, bien que d'autres noyaux puissent être utiles (^{13}C, ^{15}N). La RMN constitue une méthode non destructive, communiquant à la fois des informations très locales et globales, et permettant des analyses très diversifiées. Cependant, le temps de traitement est long, surtout pour les grandes molécules où les signaux deviennent rapidement très complexes.

II. A. 2. a) Principe de la RMN ^1H

- La RMN ^{1}H

Lorsque des noyaux atomiques de spin nucléaire $\frac{1}{2}$ sont soumis à un champ magnétique externe $\vec{B}_0$, les moments magnétiques des noyaux s'alignent sur la direction de $\vec{B}_0$, dans le même sens et dans le sens contraire. Les moments magnétiques dans le même sens que $\vec{B}_0$ sont dans un état E_α et ceux de sens opposé sont dans un état E_β. L'état E_α étant plus stabilisé et donc plus peuplé que l'état E_β d'après la distribution de Boltzmann, la somme des moments magnétiques, aussi appelée aimantation nucléaire $\vec{M}$, est donc dirigée dans le sens de $\vec{B}_0$. Ces noyaux sont également munis d'un mouvement de précession autour de $\vec{B}_0$ dont la fréquence est la fréquence de Larmor :

$$\omega_0 = \gamma_i B_0 \tag{II.10}$$

avec γ_i le rapport gyromagnétique de l'élément i. Pour l'atome d'hydrogène, on a donc

$$\omega_0 = 42{,}576\, B_0 \tag{II.11}$$

avec ω_0 en MHz et B_0 en T. Les électrons, de charges négative, en précession autour de $\vec{B}_0$ génèrent à leur tour un champ magnétique $\vec{B}_e$ de sens opposé à $\vec{B}_0$. Le champ magnétique effectif ressenti par les noyaux a donc pour amplitude :

$$B_{eff} = B_0(1 - \sigma_n) \tag{II.12}$$

avec σ_n la constante d'écran qui dépend du nuage électronique du noyau n. Autrement dit, pour que deux noyaux ressentent le même champ magnétique, ils doivent avoir le même environnement chimique. Lorsque c'est le cas, ils sont dits équivalents. Il en résulte que la fréquence de résonance de chaque noyau est différente selon l'environnement, ce qui va permettre de différencier les hydrogènes au sein d'une même molécule. En RMN, c'est le déplacement chimique δ qui est utilisé plutôt que la fréquence. Il est défini par :

$$\delta = 10^6 \frac{\nu - \nu_{ref}}{\nu_0} \tag{II.13}$$

où ν est la fréquence de résonance du noyau, ν_0 est la fréquence du spectromètre et ν_{ref} est la fréquence d'une référence dans le milieu. En chimie, c'est le Tétraméthylsilane (TMS) qui est le plus souvent employé dans les solvants organiques et le 4,4-Diméthyl-4-silapentane-1-sulfonate (DSS) dans l'eau.

- Le couplage scalaire

Dans une molécule, lorsque deux hydrogènes ayant des environnements chimiques différents sont proches en nombre de liaisons covalentes, généralement d'une distance égale ou inférieure à trois liaisons chimiques, la densité électronique des noyaux dans les états E_β et E_α jouent un rôle sur l'allure du spectre mesuré. En effet, chaque noyau d'hydrogène est affecté par son propre nuage électronique, lequel est à son tour influencé par le nuage électronique de l'autre hydrogène dont le noyau peut être soit dans l'état E_α, soit dans l'état E_β. Ainsi, dans environ 50 % des cas, le champ magnétique perçu par chaque hydrogène est diminué en raison de l'influence de son voisin, et dans les 50 %, il est augmenté. Ces variations provoquent des différences dans les déplacements chimiques observés, ce qui se manifeste par un couplage spin-spin dans le spectre RMN, caractérisé par la multiplication des signaux. La constante de couplage est donc la différence de fréquence entre les deux signaux du même hydrogène, elle est notée ${}^nJ$ avec n le nombre de liaisons séparant les deux atomes analysés. La densité électronique étant fortement liée à la structure des molécules, il est possible de déduire des paramètres géométriques par la mesure de cette constante de couplage.

II. A. 2. b) Determination de la structure

La détermination de la conformation adoptée par les peptides en solution par la RMN se fait à l'aide d'une combinaison de plusieurs analyses. Chacune d'elles apporte une information différente et donc complémentaire sur le repliement de la molécule. Par exemple, la recherche d'une éventuelle protection des protons amides au solvant apporte une information locale sur les repliements pouvant induire des liaisons hydrogène intramoléculaires, qui peut être complétée par la mesure de constante de couplage H_N-H_α introduisant une information localisée sur un angle proche. En revanche, la détermination des distances inter protons par l'intermédiaire des corrélations NOEs contribue quant à elle à l'analyse de la structure du peptide dans sa globalité.

- Le déplacement chimique secondaire

La fréquence de résonance des noyaux étant influencée par la densité électronique environnante, le déplacement chimique de certains protons des acides aminés est influencé par la structure secondaire localement adoptée par la molécule. En prenant comme référence le déplacement chimique de protons de protéines dénaturées, on peut mesurer la déviation du déplacement chimique $\Delta\delta_{CSD}$:

$$\Delta\delta_{CSD} = \delta_{obs} - \delta_{ref} \quad \text{(II.14)}$$

Il a été montré que les structures secondaires de type pure hélice α pouvait avoir une déviation de $-0.39\ ppm$ rapport aux structures dénaturées et les feuillets β $jusqu'à$ $+0.37\ pp$m sur les protons H_α. Il est donc possible d'en extraire une indication sur la structure secondaire majoritaire de la molécule ciblée. En pratique, on considère pour les noyaux hydrogènes que si pour 4 résidus consécutifs, $\Delta\delta_{CSD} < -0.1$ ppm alors la molécule a une probabilité plus forte d'être en hélice, et si pour 3 résidus successifs $\Delta\delta_{CSD} > 0.1$ ppm, elle a une forte probabilité d'adopter une structuration en brin β. Des valeurs intermédiaires sont associées à des conformations aléatoires ou des mélanges lorsque l'hypothèse est envisageable. Cependant cette analyse n'est pas discriminante, elle donne une indication sur la probabilité d'une structuration en hélice ou en feuillet.

- La constante de couplage ${}^3J$

D'après la relation de Karplus, l'angle dièdre Θ entre deux atomes d'hydrogène i et j séparés de trois liaisons est directement relié à la constante de couplage ${}^3J_{i,j}(\Theta)$ par la relation :

$${}^3J_{i,j}(\Theta) = Acos^2(\Theta) + Bcos(\Theta) + C \quad \text{(II.15)}$$

dans laquelle A, B et C sont des coefficients déterminés à l'aide d'expériences menées sur des molécules similaires et où Θ est contraint souvent à l'aide de cycles, ou de doubles liaisons.[28–30] L'analyse du repliement de la chaîne peptidique repose sur la détermination des angles dièdres φ et ψ de chaque résidu. L'angle φ est géométriquement lié à l'angle Θ entre le proton amide et celui en alpha d'un même résidu, par la relation :

$$\Theta = |\varphi \pm 60| \quad \text{(II.16)}$$

où la chiralité du résidu impose une rotation de $-60°$ pour les résidus L et $+60°$ pour les résidus D. La constante de couplage ${}^3J_{HN-H\alpha}$ est d'abord mesurée puis comparée qualitativement aux valeurs possibles dans les feuillets β et les hélices α. En effet, cette constante est particulièrement utile pour différencier les structures hélicoïdales des structures en feuillet β en raison de sa sensibilité sur cette gamme d'angle. En pratique, on suppose que lorsque plusieurs résidus consécutifs ont une constante de couplage ${}^3J_{i,j}(\Theta) < 6$ Hz, alors ils sont inclus dans une structure de type hélice. A l'inverse, si les constantes de couplage sont plus élevées ${}^3J(\Theta) > 9\ Hz$, alors les résidus sont supposés inclus dans un feuillet. Il est possible d'aller plus loin dans l'analyse afin de déterminer l'angle dièdre φ pour chaque résidu, à condition d'utiliser des paramètres A, B et C adaptés. De nombreux jeux de coefficients existent

dans la littérature, parmi lesquels ceux établis par Pardi[31] et repris par Wüthrich[32] sont les plus connus :

$$^3J(\Theta) = 6.4\cos^2(\Theta) - 1.4\cos(\Theta) + 1.9 \qquad \text{(II.17)}$$

Cette équation est représentée graphiquement **Figure II.3** et met en évidence les zones des brin β et des hélices α. La constante de couplage seule n'est pas discriminante et constitue un paramètre indicatif de la présence d'hélice ou de feuillet.

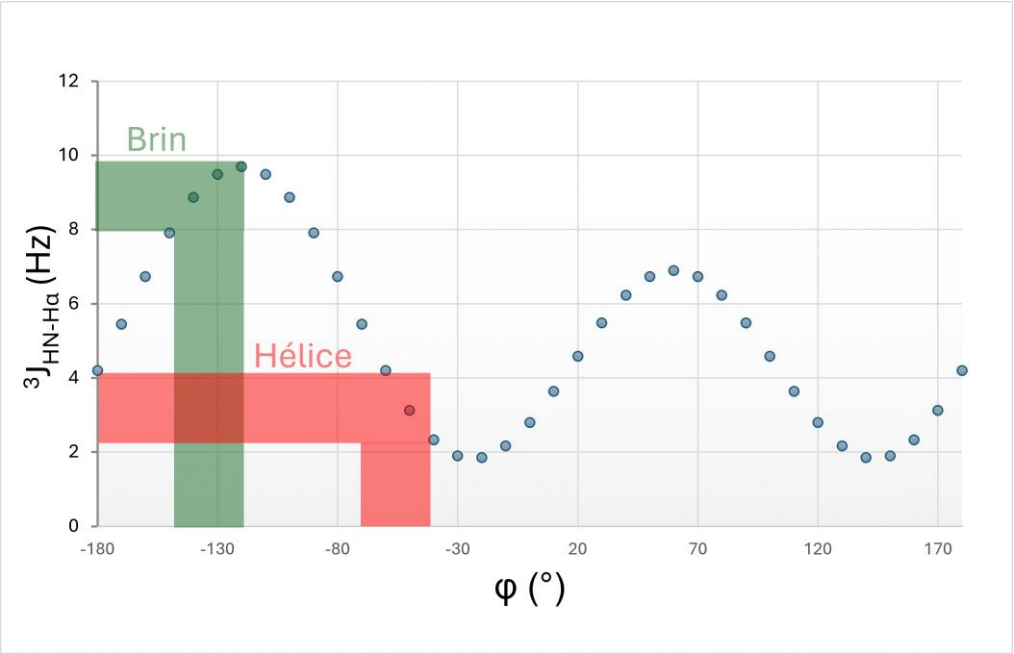


***Figure II.3**: Courbe de la constante de couplage $^3J_{HN-Hα}$ en fonction de l'angle dièdre φ par la relation de Karplus avec les coefficients de A, B et C déterminés par Pardi. La gamme d'angle dièdre des hélices sont reliés à leurs constantes de couplage (barre rouge). La gamme d'angle dièdre des brins sont reliés à leurs constantes de couplage (barre verte).*

- L'Effet Nucléaire Overhauser (NOE)

Le NOE est un phénomène où la relaxation croisée entre deux spins nucléaires proches provoque un transfert de polarisation. Cette interaction dipolaire est fonction de la distance entre les protons r_{ij}, et du temps de corrélation τ_c selon la relation suivante :

$$NOE \propto \frac{1}{{r_{ij}}^6} f(\tau_c) \qquad \text{(II.18)}$$

Le NOE est sensible dans l'espace pour des distances < 5 Å. Il est classiquement mesuré à l'aide d'expériences ^{1}H 2D NOESY dans laquelle la diagonale montre chaque signal proton dans le milieu et les éléments hors diagonale sont les corrélations NOEs entre les signaux concernés

(**Figure II.4**). Dans les peptides et les protéines, les intensités NOEs de certains protons sont reliées à des éléments de structure secondaire principaux comme les hélices α ou les brin β.[33]

Ainsi la présence de certaine corrélation peut dès lors être discriminante sur les structures secondaires présentes dans le milieu. Ces corrélations sont d'abord analysées qualitativement à l'aide du diagramme NOE, dans lesquels l'intensité de corrélation est traduite par des traits plus ou moins épais. Par exemple, la **Figure II.5** montre les diagrammes NOE correspondant à certaines structures secondaires.

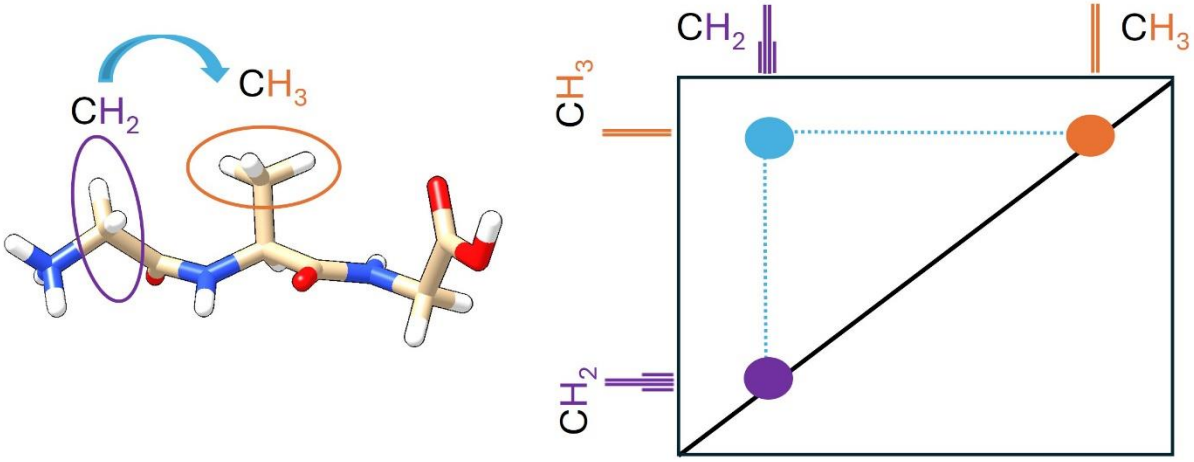


Figure II.4*: Schéma d'une corrélation sur une expérience ^{1}H 2D NOESY (en cyan) sur l'exemple du peptide Gly-Ala-Glyl. Les protons CH_2 (violets) corrèlent avec les protons CH_3 (oranges) situés à 6 liaisons d'écart (point cyan) mais proches dans l'espace d'une distance inférieure à 5 Å.*

Pour aller plus loin, on peut quantifier ces interactions. En effet, les corrélations NOEs peuvent être utilisées pour calculer les distance r_{ij} d'après la relation :

$$r_{ij} = r_{ref}\left(\frac{a_{ref}}{a_{ij}}\right)^{\frac{1}{6}} \qquad \text{(II.19)}$$

où $\frac{a_{ref}}{a_{ij}}$ est le rapport entre l'intensité de la référence et celle de la corrélation examiné et r_{ref} est la distance de référence, fixée entre 1.7 et 1.8 Å pour les protons géminaux d'un carbone sp^3. De l'Équation (II.19) découlent donc des distances, qui peuvent être calculées pour chaque structure secondaire (**Figure II.6**).

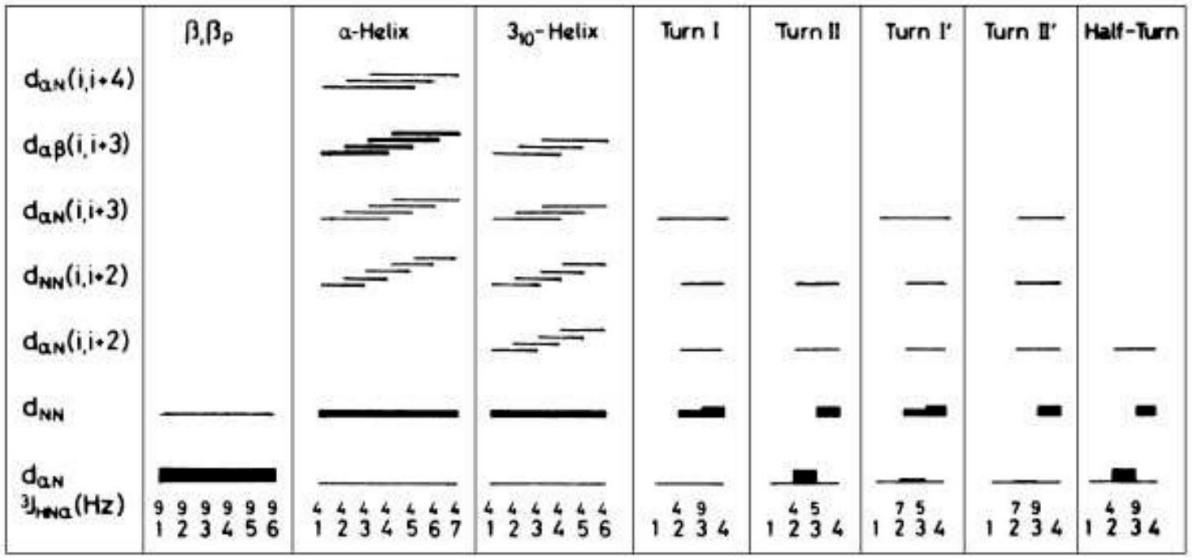


Figure II.5 *: Diagrammes NOE de différentes structures secondaires présentes dans les peptides et les protéines. Extrait de la ref.* ***32***

	α-Helix	3_{10} Helix	β Antiparallel	β Parallel	Type I turn	Type II turn
$d_{\alpha N}$	3·5	3·4	2·2	2·2	3·4/3·2	2·2/3·2
$d_{\alpha N}(i, i+2)$	4·4	3·8			3·6	3·3
$d_{\alpha N}(i, i+3)$	3·4	3·3			3·1–4·2	3·8–4·7
$d_{\alpha N}(i, i+4)$	4·2					
d_{NN}	2·8	2·6	3·3	4·0	2·6/2·4	4·5/2·4
$d_{NN}(i, i+2)$	4·2	4·1			3·8	4·3
$d_{\alpha\beta}(i, i+3)$	2·5–4·4	3·1–5·1				

Figure II.6 *: Distances NOEs de différentes structures secondaires présentes dans les peptides et les protéines. Extrait de la ref.* ***32***

- La protection des protons amides

La RMN est l'un des rares outils avec la FT-IR permettant l'analyse des liaisons hydrogène au sein des molécules en solution. Les protons amides des peptides et des protéines peuvent être impliqués dans ces interactions intramoléculaires N–H···O–C, qui peuvent être mises en évidence par les marquages isotopiques ^{13}C et ^{15}N couplés à de séquences RMN spécifiques.[34] Ces phénomènes se traduisent aussi indirectement par une variation plus (site ouvert) ou moins (site fermé) importante des déplacements chimiques des protons H_N en fonction de la température. En effet, une variation, aussi appelée coefficient de température, élevée en valeur absolue traduit une facilité pour le proton à s'échanger avec le solvant.

En fonction du solvant, une valeur seuil des coefficients de température est définie.[35–38] Les coefficients supérieurs à ces valeurs sont alors compatibles avec une protection des protons au solvant, et donc avec la présence de liaisons hydrogène intramoléculaires. Pour l'analyse des protéines dans l'eau, cette valeur seuil est de $-4.6\ ppb.K^{-1}$.[39]

Dans les structures hélicoïdales et les feuillets β, chaque proton amide du résidu i est supposé former une liaison hydrogène : le résidu $i+4$ pour les hélices α, ou avec un résidu d'un autre brin dans les feuillets β. Ainsi, à l'exception des extrémités, tous les protons amides devraient présenter un coefficient de température faible en valeur absolue. En pratique, ce paramètre est plutôt mesuré dans les petites molécules soit pour confirmer la présence d'une structure secondaire, soit pour monter l'évolution d'une confirmation par la non-linéarité du déplacement chimique en fonction de la température.

II. A. 3. La dynamique sous contrainte RMN

La RMN seule ne permet pas de déterminer la structure tridimensionnelle des molécules. Or, les dynamiques moléculaires, que nous aborderons dans le chapitre suivant, génèrent des représentations spatiales de la molécule et peuvent être utilisées en incorporant les paramètres expérimentaux issus de la RMN lors de simulations dites « sous contraintes RMN ». Elles sont donc couramment utilisées pour déterminer des structures tridimensionnelles de biomolécules. Pour les peptides ou les protéines, les distances mesurées par l'effet NOE et les constantes de couplage ^3J sont les principaux paramètres servant de contraintes. Ces mesures sont ensuite traduites en contraintes géométriques. Lors de la simulation de dynamique moléculaire, certains angles et distances sont ainsi contraints pour toujours rester proches des valeurs expérimentales.

Le principal avantage de cette approche est sa capacité à explorer des conformations moléculaires en accord avec les résultats expérimentaux, tout en exploitant pleinement les données de la RMN pour générer des géométries compatibles avec les mesures. Néanmoins, cette méthode présente aussi certaines limitations : certaines conformations, bien que valides selon les données expérimentales, peuvent ne pas être détectées en raison des contraintes imposées qui restreignent l'évolution de la structure pendant la simulation. En effet, ces contraintes peuvent limiter excessivement l'exploration de l'espace conformationnel, empêchant ainsi la découverte de certaines conformations alternatives. La méthode du recuit simulé permet en partie de pallier ces problèmes, en réduisant la probabilité que la conformation soit piégée dans un minimum local d'énergie.

II. B. Limites des méthodes expérimentales

Bien que les méthodes expérimentales actuelles, telles que la spectroscopie RMN et la dichroïsme circulaire (DC), soient souvent très performantes pour l'étude des protéines, elles présentent plusieurs limitations lorsqu'il s'agit d'étudier certaines structures secondaires spécifiques ou des peptides plus petits et flexibles. Ces limitations doivent être prises en compte pour améliorer la fiabilité de l'analyse structurale de protéines, en particulier dans le cas des structures secondaires complexes comme les coudes et de protéines ayant une flexibilité intrinsèque.

II. B. 1. Application aux coudes

Un des grands défis des méthodes actuelles réside dans l'identification précise des différentes structures secondaires de type coude. Ces structures jouent un rôle clé dans la conformation tridimensionnelle des protéines, mais elles sont souvent ignorées ou mal interprétées. En pratique, les coudes sont fréquemment modélisés de manière simplifiée, en tant que structures globales « coudes » ou « coudes β », ce qui ne reflète pas leur grande diversité. Il existe en effet une multitude de coudes différents, classifiés en fonction de leurs angles dièdres dont les types β sont résumés **Tableau I.1**. Mais cette complexité est rarement prise en compte dans les analyses RMN ou DC.

En RMN, aucun déplacement chimique secondaire spécifique n'est associé de manière claire aux structures de type coude. De même, la constante $^{3}J_{HN-H\alpha}$, ne peut pas caractériser correctement les différents coudes avec par exemple seulement deux constantes présentes dans les coudes β ce qui est insuffisant pour décrire les douze variétés de coudes β connues. Cependant, des progrès ont été réalisés dans la caractérisation des coudes à l'aide des diagrammes de corrélation NOE. Notamment, les travaux pionniers de Wüthrich ont démontré que les diagrammes NOE peuvent fournir des informations structurelles utiles sur les coudes, bien que ces études se soient principalement concentrées sur des acides aminés spécifiques et n'aient pas permis de caractériser tous les types de coudes avec une précision optimale. Le coefficient de température est un bon outil pour identifier la liaison hydrogène $1{,}4$ des coudes β et $1,3$ des coudes γ.

Le DC est la méthode de choix pour l'analyse des structures secondaires et pourrait permettre de surmonter certaines difficultés à caractériser les structures secondaires de type coude. Toutefois, la caractérisation des coudes, notamment des coudes β, reste complexe en raison de leur diversité et de la superposition des contributions spectrales des autres structures

secondaires. En effet, les coudes ne sont constitués que de quelques résidus, ce qui limite l'intensité de leurs signaux dans le spectre global de DC, rendant ainsi difficile leur identification précise. Pour établir des spectres expérimentaux de référence pour ces types de structures, il faut donc limiter les peptides à seulement quelques paires de base ce qui peut engendrer d'autres problèmes, notamment pour maintenir la structuration stable en solution.

Des tentatives théoriques et expérimentales ont été entreprises pour établir des spectres de références spécifiques aux coudes, mais elles aboutissent à des résultats contradictoires, limitant leur application pratique. Une partie étant dédiée sur les propositions de spectres simulés par théorie pour les coudes dans la partie Chapitre IV. B. 1, la section suivante décrit uniquement les différentes approches expérimentales utilisées pour fournir ces spectres de référence.

La première méthode a consisté soustraire, des spectres DCE expérimentaux de protéines à structure connue, les contributions des hélices, des feuillets et des conformations aléatoires. De cette façon, une étude a proposé que le spectre résultant pouvait être attribué aux structures secondaires de type coude β (**Figure II.7**).[40] Toutefois, ce spectre de coude représente une moyenne sur un grand nombre de conformations potentielles et n'est pas assez spécifique pour caractériser les sous-types de coudes.

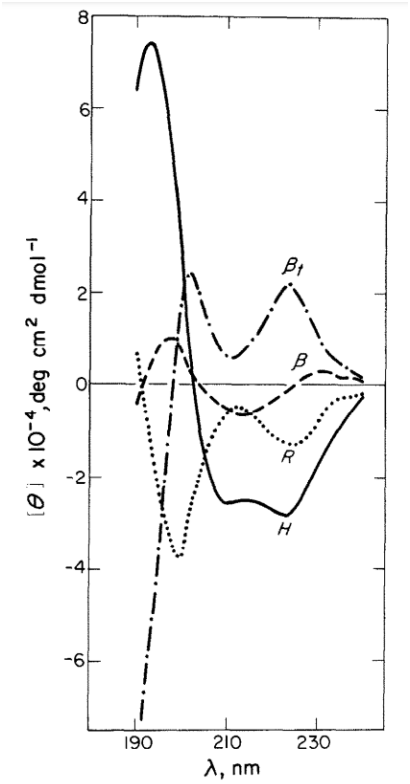


***Figure II.7** : Spectres de référence pour les hélices (H), les conformations aléatoires (R), les formes β (β) comprenant les feuillets, et les coudes β (βt).[40]*

Une seconde approche mixte visant à reproduire les motifs DCE proposés par des simulations théoriques a été utilisée pour proposer des spectres de référence de coudes de type β-I et β-II. Ainsi, le spectre DCE du peptide poly(Ala_2-Gly_2) reproduisant fidèlement le motif global (mais pas les intensités exactes) du spectre théorique des coudes de type β-I et β-II, proposé par Woody[42] (décrit dans le chapitre IV), les auteurs en ont conclu que ce spectre représentait ces catégories de coudes (**Figure II.8**).

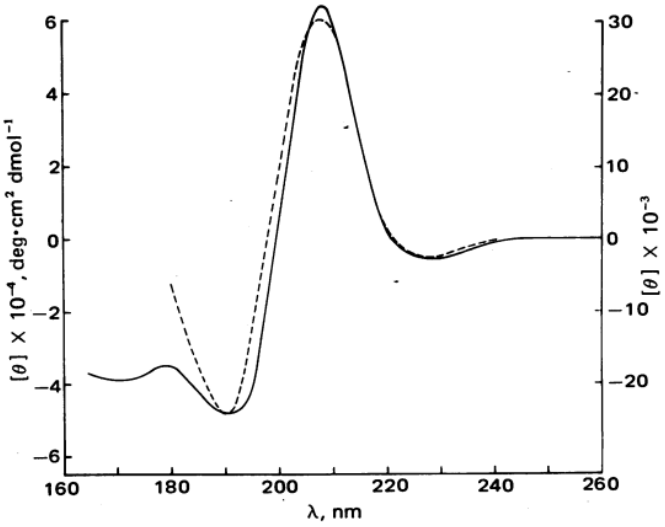


Figure II.8 *: Spectres DCE de du peptide poly (Ala_2-Gly_2) (ligne pleine) et spectre théorique proposé pour les coudes de type β-II (ligne pointillée).*

Enfin une approche alternative consistant à synthétiser des séquences conçues pour favoriser la formation de coude β, soit par la cyclisation,[43–45] soit par ajout d'acides aminés modifiés[46] et à réaliser leur étude structurale par diffraction RX et DCE a été utilisée. Après cristallisation, on identifie la conformation présente, puis on associe le spectre DCE à ce type de coude. Cependant, le problème majeur de cette méthode est le manque de preuve quant à la stabilité de la structuration en coudes en solution, où les mesures DCE sont réalisées. En conséquence, des contradictions apparaissent : par exemple, les spectres DCE de coude β-II obtenus pour des molécules cyclisées sont tous différents à une longueur d'onde donnée (**Figure II.9**), bien qu'un motif global semble se dégager. Ce motif ne correspond toutefois pas à ceux des molécules non cyclisées supposées en β-II en solution (**Figure II.10**).

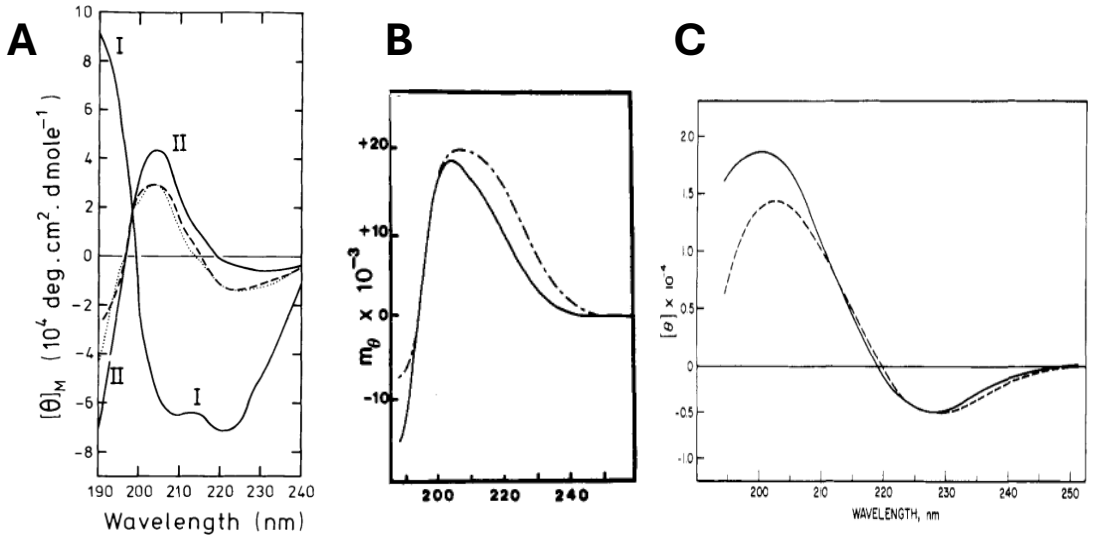


***Figure II.9**: Proposition de spectre DCE de coudes de type β-II dans des molécules cyclisées. A : Spectre expérimental du peptide cyclo (L-A-L-Ala-Aca) (ligne pleine I), cyclo (L-Ala-D-Ala-Aca) (ligne pleine II) représentant un coude β-II et cyclo (L-Ala-Gly-Aca) (ligne pointillée) dans le méthanol.[43] B : Spectre expérimental du peptide cyclo(D-Ala-L-Pro-D-Ala)$_2$ supposé β-II dans le TFE (ligne pleine) et dans l'eau (ligne pointellée).[44] C : Spectre expérimental du peptide cyclo (L-Orn-L-Pro-D-Phe)$_2$ (ligne pointiellée) et cyclo (L-Orn-L-Pro-D-Cha)$_2$, supposés de type β-II, dans l'hexafluoro2-propanol.[45]*

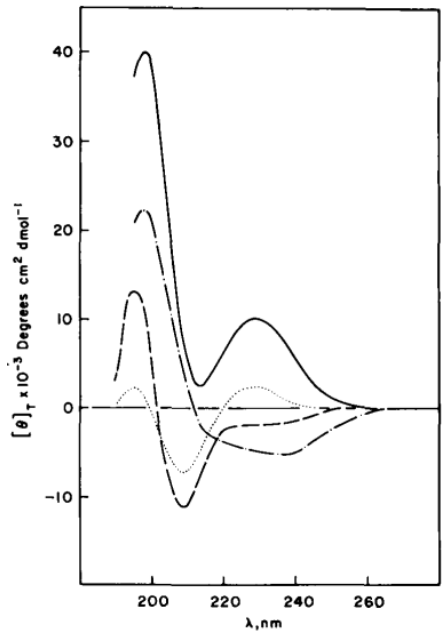


***Figure II.10** : Spectre expérimental du peptide Piv-L-Pro-Aib-NHMe, supposé de type β-II, dans le méthanol (ligne pleine) dans un mélange eau/méthanol 1:9 (ligne pointillée) dans le HFIP (tirets) et dans le cyclohexane (pointillés et tirets).[46]*

Ces méthodes soulignent les difficultés à obtenir des spectres de références fiables pour les coudes β, particulièrement en solution. L'absence de preuve concernant la stabilité des structures dans ce milieu est un obstacle majeur et génère des contradictions.

II. B. 2. Applications aux molécules petites et/ou flexibles

L'étude des petites molécules et des régions flexibles au sein des protéines pose également des défis considérables, notamment en raison de la dynamique intrinsèque de ces molécules. Les paramètres spectroscopiques mesurés par RMN, tels que les constantes de couplage, les déplacements chimiques, et les distances NOE, reflètent une moyenne dans le cas des échanges rapides entre les différentes conformations adoptées en solution. Dans le cas des molécules flexibles, cela complique leur interprétation, car une seule valeur expérimentale peut alors correspondre à plusieurs proportions pour chaque conformation. Plus la molécule est flexible, plus le nombre de conformations possibles augmente, et par conséquent, plus il devient difficile d'identifier la solution parmi les possibilités qui correspond parfaitement aux paramètres spectroscopiques mesurés et qui traduit la réalité.

En RMN, cela se manifeste par des difficultés à interpréter correctement des paramètres tels que les coefficients de température, les constantes de couplage ou les distances NOE. Ces paramètres, qui sont déterminants pour la structure de la molécule, peuvent ne pas être fiables en raison de la flexibilité moléculaire. De la même façon, la déviation du déplacement chimique en présence de plusieurs conformations peut entraîner une interprétation erronée des structures secondaires présentes.

Le problème est encore plus présent pour les peptides de très petites tailles, où la flexibilité est souvent accrue. La forte prévalence de coudes dans les peptides ajoute une couche supplémentaire de complexité dans l'interprétation des données. Un dernier problème réside dans la définition des structures secondaires dans les petites molécules. Les structures secondaires tels que les feuillets β et les hélices sont définies à partir d'un certain nombre de monomères, qui n'est pas forcément atteint dans les peptides très courts, ce qui peut amener des problèmes sur l'interprétation des données. Par exemple, certaines interactions, comme la liaison hydrogène $i, i+4$ de l'hélice α, ne sont pas présentes dans les peptides avec trois acides aminés. Or l'environnement de cette liaison influence les spectres optiques tel que le DC, ce qui pose nécessairement un problème lors de la déconvolution.

Aux vues des difficultés à interpréter correctement ces petits systèmes, l'appui d'une approche théorique semble nécessaire. Dans le chapitre suivant, nous allons voir les méthodes théoriques qui peuvent aider à la compréhension de ces molécules.

II. C. Références

1. Easson, L. H. & Stedman, E. Studies on the relationship between chemical constitution and physiological action: Molecular dissymmetry and physiological activity. *Biochem. J.* **27**, 1257–1266 (1933).
2. Andrews, S. S. & Tretton, J. Physical Principles of Circular Dichroism. *J. Chem. Educ.* **97**, 4370–4376 (2020).
3. Johnson, W. C. Protein secondary structure and circular dichroism: A practical guide. *Proteins Struct. Funct. Bioinforma.* **7**, 205–214 (1990).
4. Provencher, S. W. & Gloeckner, J. Estimation of globular protein secondary structure from circular dichroism. *Biochemistry* **20**, 33–37 (1981).
5. Woody, R. W. Chapter 2 - Circular Dichroism of Peptides. in *Conformation in Biology and Drug Design* (ed. Hruby, V. J.) vol. 7 15–114 (Academic Press, 1985).
6. Pelton, J. T. & McLean, L. R. Spectroscopic Methods for Analysis of Protein Secondary Structure. *Anal. Biochem.* **277**, 167–176 (2000).
7. Linhares, L. A. & Ramos, C. H. I. Unlocking Insights into Folding, Structure, and Function of Proteins through Circular Dichroism Spectroscopy—A Short Review. *Appl. Biosci.* **2**, 639–655 (2023).
8. Bochicchio, B. & Tamburro, A. M. Polyproline II structure in proteins: Identification by chiroptical spectroscopies, stability, and functions. *Chirality* **14**, 782–792 (2002).
9. Lopes, J. L. S., Miles, A. J., Whitmore, L. & Wallace, B. A. Distinct circular dichroism spectroscopic signatures of polyproline II and unordered secondary structures: Applications in secondary structure analyses. *Protein Sci.* **23**, 1765–1772 (2014).
10. Greenfield, N. J. Using circular dichroism spectra to estimate protein secondary structure. *Nat. Protoc.* **1**, 2876–2890 (2006).

11. Greenfield, N. J. Methods to Estimate the Conformation of Proteins and Polypeptides from Circular Dichroism Data. *Anal. Biochem.* **235**, 1–10 (1996).
12. Greenfield, N. J. & Fasman, G. D. Computed circular dichroism spectra for the evaluation of protein conformation. *Biochemistry* **8**, 4108–4116 (1969).
13. Perczel, A., Park, K. & Fasman, G. D. Analysis of the circular dichroism spectrum of proteins using the convex constraint algorithm: A practical guide. *Anal. Biochem.* **203**, 83–93 (1992).
14. Compton, L. A. & Johnson, W. C. Analysis of protein circular dichroism spectra for secondary structure using a simple matrix multiplication. *Anal. Biochem.* **155**, 155–167 (1986).
15. Manavalan, P. & Johnson, W. C. Variable selection method improves the prediction of protein secondary structure from circular dichroism spectra. *Anal. Biochem.* **167**, 76–85 (1987).
16. Johnson, W. C. Analyzing protein circular dichroism spectra for accurate secondary structures. *Proteins Struct., Funct. Bioinf.* **35**, 307–312 (1999).
17. Sreerama, N. & Woody, R. W. A Self-Consistent Method for the Analysis of Protein Secondary Structure from Circular Dichroism. *Anal. Biochem.* **209**, 32–44 (1993).
18. Sreerama, N. & Woody, R. W. Poly(Pro)II Helixes in Globular Proteins: Identification and Circular Dichroic Analysis. *Biochemistry* **33**, 10022–10025 (1994).
19. Sreerama, N. & Woody, R. W. Estimation of Protein Secondary Structure from Circular Dichroism Spectra: Comparison of CONTIN, SELCON, and CDSSTR Methods with an Expanded Reference Set. *Anal. Biochem.* **287**, 252–260 (2000).
20. Böhm, G., Muhr, R. & Jaenicke, R. Quantitative analysis of protein far UV circular dichroism spectra by neural networks. *Protein Eng., Des. Sel.* **5**, 191–195 (1992).

21. Andrade, M. A., Chacón, P., Merelo, J. J. & Morán, F. Evaluation of secondary structure of proteins from UV circular dichroism spectra using an unsupervised learning neural network. *Protein Eng., Des. Sel.* **6**, 383–390 (1993).
22. Perczel, A., Hollósi, M., Tusn´dy, G. & Fasman, G. D. Convex constraint analysis: a natural deconvolution of circular dichroism curves of proteins. *Protein Eng., Des. Sel.* **4**, 669–679 (1991).
23. Miles, A. J., Ramalli, S. G. & Wallace, B. A. DichroWeb, a website for calculating protein secondary structure from circular dichroism spectroscopic data. *Protein Sci.* **31**, 37–46 (2022).
24. Micsonai, A., Bulyáki, É. & Kardos, J. BeStSel: From Secondary Structure Analysis to Protein Fold Prediction by Circular Dichroism Spectroscopy. in *Structural Genomics* (eds. Chen, Y. W. & Yiu, C.-P. B.) vol. 2199 175–189 (Springer US, New York, NY, 2021).
25. Lees, J. G., Miles, A. J., Wien, F. & Wallace, B. A. A reference database for circular dichroism spectroscopy covering fold and secondary structure space. *Bioinformatics* **22**, 1955–1962 (2006).
26. Kabsch, W. & Sander, C. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22**, 2577–2637 (1983).
27. Micsonai, A. *et al.* Accurate secondary structure prediction and fold recognition for circular dichroism spectroscopy. *Proc. Natl. Acad. Sci.* **112**, E3095–E3103 (2015).
28. Cung, M. T., Marraud, M. & Neel, J. Experimental Calibration of a Karplus Relationship in Order to Study the Conformations of Peptides by Nuclear Magnetic Resonance. *Macromolecules* **7**, 606–613 (1974).
29. Ludvigsen, S., Andersen, K. V. & Poulsen, F. M. Accurate measurements of coupling constants from two-dimensional nuclear magnetic resonance spectra of proteins and determination of φ-angles. *J. Mol. Biol.* **217**, 731–736 (1991).

30. Vuister, G. W. & Bax, A. Quantitative J correlation: a new approach for measuring homonuclear three-bond J(HNH.alpha.) coupling constants in 15N-enriched proteins. *J. Am. Chem. Soc.* **115**, 7772–7777 (1993).
31. Pardi, A., Billeter, M. & Wüthrich, K. Calibration of the angular dependence of the amide proton-Cα proton coupling constants, 3JHNα, in a globular protein: Use of 3JHNα for identification of helical secondary structure. *J. Mol. Biol.* **180**, 741–751 (1984).
32. Wüthrich, K. NMR with Proteins and Nucleic Acids. *Europhys. News* **17**, 11–13 (1986).
33. Wagner, G. *et al.* Nuclear magnetic resonance identification of "half-turn" and 310-helix secondary structure in rabbit liver metallothionein-2. *J. Mol. Biol.* **187**, 131–135 (1986).
34. Cordier, F., Nisius, L., Dingley, A. J. & Grzesiek, S. Direct detection of N−H···O=C hydrogen bonds in biomolecules by NMR spectroscopy. *Nat. Protoc.* **3**, 235–241 (2008).
35. Merutka, G., Jane Dyson, H. & Wright, P. E. 'Random coil' 1H chemical shifts obtained as a function of temperature and trifluoroethanol concentration for the peptide series GGXGG. *J. Biomol. NMR* **5**, 14–24 (1995).
36. Miura, Y. α-proton Chemical Shift Index and Amide Proton Chemical Shift Temperature Coefficient of Melittin in Methanol: Indicators for a Helix Structure and an Intra-Molecular Hydrogen Bond? *Protein J.* **41**, 625–635 (2022).
37. Contreras, M. A., Haack, T., Royo, M., Giralt, E. & Pons, M. Temperature coefficients of peptides dissolved in hexafluoroisopropanol monitor distortions of helices. *Lett. Pept. Sci.* **4**, 29–39 (1997).
38. Comparison of helix-stabilizing effects of α,α-dialkyl glycines with linear and cycloalkyl side chains - Vijayalakshmi - 2000 - Biopolymers - Wiley Online Library.
39. Cierpicki, T. & Otlewski, J. Amide proton temperature coefficients as hydrogen bond indicators in proteins. *J. Biomol. NMR* **21**, 249–261 (2001).

40. Chang, C. T., Wu, C.-S. C. & Yang, J. T. Circular dichroic analysis of protein conformation: Inclusion of the β-turns. *Anal. Biochem.* **91**, 13–31 (1978).
41. Brahms, S., Brahms, J., Spach, G. & Brack, A. Identification of β,β-turns and unordered conformations in polypeptide chains by vacuum ultraviolet circular dichroism. *Proc. Natl. Acad. Sci.* **74**, 3208–3212 (1977).
42. Woody, R. W. Studies of theoretical circular dichroism of polypeptides: Contributions of β-turns. *Pept. Polypept. Proteins* 338–350 (1974).
43. Bandekar, J. *et al.* Conformations of cyclo(L-alanyl-L-alanyl-ε-aminocaproyl) and of cyclo(L-alany1-D-alanyl-ε-aminocaproyl); cyclized dipeptide models for specific types of β-bends. *Int. J. Pept. Protein Res.* **19**, 187–205 (1982).
44. Gierasch, L. M., Deber, C. M., Madison, V., Niu, C.-H. & Blout, E. R. Conformations of (X-L-Pro-Y)2 cyclic hexapeptides. Preferred .beta.-turn conformers and implications for .beta. turns in proteins. *Biochemistry* **20**, 4730–4738 (1981).
45. Bush, C. A., Sarkar, S. K. & Kopple, K. D. Circular dichroism of β turns in peptides and proteins. *Biochemistry* **17**, 4951–4954 (1978).
46. Crisma, M., Fasman, G. d., Balaram, H. & Balaram, P. Peptide models for β-turns. *Int. J. Pept. Protein Res.* **23**, 411–419 (1984).

III. Méthodes théoriques

Table des matières

Les méthodes théoriques jouent un rôle crucial en chimie et biochimie pour leur capacité à prédire les comportements des systèmes moléculaires et à soutenir l'interprétation des résultats expérimentaux. Dans cette partie, nous abordons plusieurs de ces méthodes en commençant par la méthode classique qu'est la Dynamique Moléculaire (DM), puis nous verrons les méthodes de calcul quantiques appliquées lors de cette thèse, la DFT et la TD-DFT et l'inclusion du solvant dans ces calculs.

III. A. Dynamique Moléculaire

La Dynamique Moléculaire (DM) est une technique de simulation très largement utilisée dans la recherche de la structuration des protéines. Elle permet notamment de simuler le comportement des molécules dans un environnement donné en suivant l'évolution au cours du temps de la vitesse et de la position des particules. Généralement régie par la mécanique classique dans les applications en biochimie, cette méthode de simulation permet donc de déterminer les trajectoires suivies.

III. A. 1. Principe général

Pour chaque particule i, on peut appliquer la deuxième loi de Newton :

$$m_i \overrightarrow{a_i} = \overrightarrow{F_i} \tag{III.1}$$

dans laquelle m_i est la masse de chaque particule i, $\overrightarrow{a_i}$ son accélération et $\overrightarrow{F_i}$ les forces subies par les particules. Ces forces dérivent elles-mêmes en général d'un potentiel d'interaction V:

$$\overrightarrow{F_i} = -\nabla_i V \tag{III.2}$$

Ce potentiel peut être exprimé comme la somme de deux termes,

$$V = V^{lié} + V^{non\ lié} \tag{III.3}$$

où $V^{lié}$ inclut les interactions intramoléculaires locales décrivant les liaisons covalentes, les angles de liaison et les torsions, tandis que $V^{non\ lié}$ représente les interactions non liées telles que les interactions de van der Waals et les interactions électrostatiques. Dans une simulation de dynamique moléculaire, les champs de forces définissent le potentiel d'interaction. La littérature offre une multitude de champs de forces, chacun ayant des degrés de complexité adaptés à différents types de systèmes moléculaires. Dans les champs de forces dit de classe 1, le potentiel V d'un champ de force est typiquement décrit par une somme de plusieurs contributions spécifiques :[1]

$$V = \underbrace{V^{liaison} + V^{angle} + V^{torsion}}_{V^{lié}} + \underbrace{V^{Coulomb} + V^{VdW}}_{V^{non\ lié}} \quad \text{(III.4)}$$

où le potentiel de liaison et le potentiel d'angle sont décrits par un potentiel harmonique :

$$V^{liaison} = \sum_{liaison} \frac{1}{2} k_b \, (l - l_0)^2 \,, \qquad V^{angle} = \sum_{angle} \frac{1}{2} k_a (\eta - \eta_0)^2 \quad \text{(III.5)}$$

avec k_b la constante de liaison, l la longueur d'une liaison donnée, l_0 sa longueur à l'équilibre, k_a la constante d'angle, η l'angle de liaison et η_0 sa valeur à l'équilibre. Le potentiel de torsion est quant à lui décrit par un potentiel périodique de la forme :

$$V^{torsion} = \sum_{angle} \frac{V_n}{2} (1 + \cos{(n\kappa - \eta)})^2 \quad \text{(III.6)}$$

avec V_n la constante de torsion, n le nombre de minima et de maxima entre 0 et 2π, κ l'angle de torsion et η la phase. Les interactions électrostatiques sont décrites par le potentiel de Coulomb :

$$V^{Coulomb} = \sum_{i>j} \frac{q_i q_j}{r_{ij}} \quad \text{(III.7)}$$

entre les charges q_i et q_j de la particule i et j respectivement, distantes d'une longueur r_{ij}. Les interactions de van der Waals sont décrites par la relation :

$$V^{VdW} = \sum_{i>j} 4 \in_{ij} \left(\frac{\sigma_{ij}^{12}}{r_{ij}^{12}} - \frac{\sigma_{ij}^{6}}{r_{ij}^{6}} \right) \quad \text{(III.8)}$$

dans laquelle $\in$ est l'énergie de Lennard-Jones, c'est-à-dire le paramètre définissant la profondeur du puit d'énergie et σ est la distance entre les particules lorsque le terme attractif (terme de droite) et répulsif (terme de gauche) sont égaux.

Les potentiels d'interaction varient selon le champ de force utilisé. Certains champs de force incorporent des termes supplémentaires, tels que le potentiel d'Urey-Bradley ou le potentiel d'angles impropres,[2] tandis que d'autres modifient les termes existants, par exemple en utilisant un potentiel de Morse plutôt qu'un potentiel harmonique pour décrire les interactions de liaisons. Chaque champ de force est conçu pour modéliser les systèmes moléculaires de manière spécifique, en tenant compte des particularités de leurs interactions. Cependant, il est souvent

nécessaire d'évaluer plusieurs champs de force pour déterminer celui qui offre la meilleure représentation du système étudié.

Parmi les champs de force classiques très populaires pour la modélisation moléculaire, on retrouve les déclinaisons de CHARMM[3], AMBER[4], GROMOS[5] et OPLS[6] qui sont largement utilisées pour simuler des systèmes biologiques. En complément de ces approches, plusieurs extensions et nouvelles catégories de champs de force ont vu le jour pour améliorer la précision des simulations, notamment les champs de force polarisables, les champs de force réactifs et les champs de force par apprentissage automatique ou Machine Learning – Force Fields (ML-FF).

Les champs de force polarisables permettent aux charges atomiques ou aux dipôles de s'ajuster dynamiquement en réponse à l'environnement électrostatique local, capturant ainsi les effets de polarisation électronique. Les champs de force comme AMOEBA[7] et les modèles de polarisation Drude Oscillator[8] sont des exemples courants de ces approches. Dans les champs de forces réactifs, des liaisons chimiques peuvent se briser ou se former ce qui permet de modéliser des réactions chimiques. ReaxFF[9] est l'un des champs de force les plus utilisés dans ce domaine, notamment pour les processus réactifs comme la combustion et la catalyse.[10] Enfin, avec les ML-FF, les potentiels d'interaction sont fondés sur les techniques avancées d'apprentissage automatique.[11] Des modèles comme ANI[12] ou SchNet[13] sont des ML-FF couramment utilisés.

III. A. 2. Les ensembles statistiques

Un modèle qui n'est gouverné que par les potentiels décrits précédemment ne suffit pas à reproduire fidèlement les conditions physiques réelles d'une simulation moléculaire. En effet, il pourrait présenter des fluctuations non désirées de température, de volume ou de pression. Ces fluctuations sont dues à l'absence de mécanismes de contrôle sur les grandeurs thermodynamiques entraînant des dérives indésirables du système. Pour contrer ces problèmes, on utilise des ensembles statistiques. Ces ensembles permettent de maintenir constantes certaines grandeurs thermodynamiques en fonction des conditions choisies. Parmi eux, les ensembles (N, V, T) et (N, P, T) sont particulièrement fréquents dans les simulations moléculaires.

Dans l'ensemble (N, V, T), aussi appelé ensemble canonique, le nombre de particules N, le volume V et la température T sont maintenus constants. La température est maintenue constante à l'aide d'un thermostat, dont le rôle est de compenser les fluctuations thermiques de façon à

ramener la température vers la valeur cible. Plusieurs thermostats sont couramment utilisés en DM, dont ceux de Berendsen[14], Nosé-Hoover[15], Andersen[16] et Bussi (v-rescaling)[17].

Dans l'ensemble (N, P, T), le nombre de particules N, la pression P et la température T sont maintenus constants. Dans cet ensemble, en plus du contrôle de température grâce à un thermostat, un barostat est utilisé pour maintenir la pression constante. Cela est réalisé par des ajustements dynamiques du volume, afin de compenser les variations de pressions internes. Parmi les barostats les plus couramment utilisés en DM, on retrouve les barostats de Berendsen[14] et celui de Parinello-Rahman[18].

III. A. 3. Les équations du mouvement

Pour simuler l'évolution temporelle des particules dans le système, on utilise des algorithmes d'intégration des équations du mouvement. Le temps est discrétisé en pas de temps Δt de manière à calculer la position et la vitesse des particules aux temps $t + \Delta t$. Plusieurs algorithmes ont été proposés pour intégrer ces équations. Dans le cas de l'algorithme de Verlet, les équations de propagation résultent de la somme du développement de Taylor à l'instant $(t + \Delta t)$ et $(t - \Delta t)$:

$$\begin{cases} \vec{r}_i(t + \Delta t) = 2\vec{r}_i(t) - \vec{r}_i(t - \Delta t) + \vec{a}_i(t)\Delta t^2 + o(\Delta t^4) \\ \vec{v}_i(t + \Delta t) = \dfrac{\vec{r}_i(t) - \vec{r}_i(t - \Delta t)}{\Delta t} + o(\Delta t^2) \end{cases} \quad \text{(III.9)}$$

Une autre méthode d'intégration est l'algorithme de « *leapfrog* », dans lequel les vitesses sont calculées avec un décalage de $\frac{\Delta t}{2}$ par rapport aux positions :

$$\begin{cases} \vec{r}_i(t + \Delta t) = \vec{r}_i(t) + \vec{v}_i\left(t + \dfrac{\Delta t}{2}\right)\Delta t \\ \vec{v}_i\left(t + \dfrac{\Delta t}{2}\right) = \vec{v}_i\left(t - \dfrac{\Delta t}{2}\right) + \vec{a}_i(t)\Delta t \end{cases} \quad \text{(III.10)}$$

Pour résumé, les simulations de dynamique moléculaire se déroulent de la façon suivante :

Étape 1 : Initialisation. À chaque atome de position $\vec{r}_i(t_0)$ est associé une vitesse $\vec{v}_i(t_0)$. Les vitesses initiales sont choisies en accord avec les ensembles utilisés de façon à prendre en compte les thermostats et barostats dans les équations de mouvements.

Étape 2 : Calcul des forces. La somme des forces exercées sur chaque particule est calculée à partir des potentiels décrits dans les champs de force.

Étape 3 : Intégration. L'algorithme d'intégration est utilisé pour calculer les nouvelles positions et vitesses au fil du temps.

Étape 4 : Propagation. Les étapes 2 et 3 sont répétées autant de fois que désiré afin de simuler l'évolution du système au cours du temps.

III. B. Théorie de la Fonctionnelle de la Densité (DFT)

III. B. 1. Équation de Schrödinger

Le fondement de la DFT est l'équation de Schrödinger qui décrit l'évolution spatiale et temporelle de la fonction d'onde d'un système quantique non relativiste :[19]

$$i\hbar\frac{\partial}{\partial t}\Psi(\{\vec{r_i}\},t) = \hat{H}\Psi(\{\vec{r_i}\},t) \qquad \text{(III.11)}$$

avec $\hbar$ la constante de Planck réduite, $\hat{H}$ l'opérateur Hamiltonien et $\Psi(\{\vec{r_i}\},t)$ la fonction d'onde contenant toutes les informations quantiques du système en fonction temps et de l'ensemble des vecteurs positions $\{\vec{r_i}\}$ des i particules. De cette équation découle la quantification de l'énergie E_n des états stationnaires $\psi_n(\{\vec{r_i}\})$ dans la version indépendante du temps :

$$\hat{H}\psi_n(\{\vec{r_i}\}) = E_n\psi_n(\{\vec{r_i}\}) \qquad \text{(III.12)}$$

Lorsque le système est un système moléculaire composé de N noyaux et de M électrons, l'hamiltonien, regroupant tous les opérateurs énergétiques, est exprimé de la manière suivante :

$$\hat{H} = \hat{T}_e + \hat{T}_n + \hat{V}_{ee} + \hat{V}_{ne} + \hat{V}_{nn} \qquad \text{(III.13)}$$

avec $\hat{T}_e$ (par la suite noté $\vec{T}$) l'opérateur d'énergie cinétique total des électrons, $\hat{T}_n$ l'opérateur d'énergie cinétique des noyaux, $\hat{V}_{ee}$ l'opérateur de répulsion biélectronique totale, $\hat{V}_{ne}$ l'opérateur du potentiel électrostatique entre les noyaux et les électrons et $\hat{V}_{nn}$ l'opérateur de répulsion électrostatique internucléaire. Sous cette forme, l'équation est trop complexe et est donc très peu utilisée pour des systèmes de très grande taille. En revanche, il est possible de faire quelques approximations. En effet, les opérateurs d'énergie cinétique étant inversement proportionnelle à la masse des particules et la masse des électrons étant considérablement négligeable par rapport à la masse des noyaux, le terme $\hat{T}_n$ peut donc être négligé et le potentiel $\hat{V}_{nn}$ peut être considéré comme une constante le temps de résoudre le problème électronique. Cette simplification est appelée l'approximation de Born-Oppenheimer.[20] De plus, l'approximation permet de résoudre un hamiltonien poly-électronique $\hat{H}^{el}$ (nommé par la suite hamiltonien) intervenant dans l'équation aux valeurs propres :

$$\hat{H}^{el}\psi_n^{el}(\{\vec{r_i}\}) = \left[\hat{T}_e + \hat{V}_{ee} + \hat{V}_{ne}\right]\psi_n^{el}(\{\vec{r_i}\}) = E_n^{el}\psi_n^{el}(\{\vec{r_i}\}) \quad \text{(III.14)}$$

Plusieurs méthodes de résolution de cette équation ont été proposées et parmi-elles, la Théorie de la Fonctionnelle de la Densité ou DFT visant à décrire le système à l'aide de la densité électronique $\rho(\vec{r})$ représentant le nombre moyen d'électrons par unité de volume en une position $\vec{r}$. Cette théorie repose sur deux théorèmes fondamentaux.

III. B. 2. Théorèmes de Hohenberg-Kohn

1er théorème de Hohenberg-Kohn *:*

« Chaque observable d'un système quantique stationnaire (en particulier l'énergie) peut être calculée, en principe exactement, seulement à partir de la densité électronique de l'état fondamental. C'est-à-dire que toutes les observables peuvent s'écrire sous la forme d'une fonctionnelle de la densité électronique de l'état fondamental.»[21]

L'énergie électronique totale du système est donc une fonctionnelle de la densité électronique :

$$E[\rho] = F_{HK}[\rho] + \int v_{ext}(\vec{r})\,\rho(\vec{r})d^3r \quad \text{(III.15)}$$

avec F_{HK} la fonctionnelle universelle (indépendante du potentiel externe) regroupant les termes d'interactions de répulsions biélectroniques et l'énergie cinétiques électroniques.

- 2e théorème de Hohenberg-Kohn

« La densité électronique de l'état fondamental peut être calculée, en principe exactement, en utilisant une méthode variationnelle qui n'implique que la densité électronique.»

Ainsi, le recours à la DFT pour la recherche de l'état fondamental revient à minimiser la fonctionnelle de la densité. Or la formulation mathématique de $F_{HK}[\rho]$ est inconnue sous forme analytique. Pour remédier à ce problème, on peut considérer à la place d'un système d'électrons en interactions, un système d'électrons indépendants évoluant dans un potentiel externe effectif générant la densité électronique du système réel. Ceci permet d'établir les équations de Kohn-Sham (KS) dans lesquelles la fonction d'onde du système fictif est définie sous la forme d'un déterminant de Slater d'orbitales monoélectroniques. Ces orbitales de KS ϕ_i sont donc celles d'un système fictif d'électrons décorrélés générant la même densité électronique que le système réel :

$$\left[\hat{T}_i + \hat{V}_H + \hat{V}_{ne} + \hat{V}_{xc}\right]\phi_i(\vec{r}) = \epsilon_i\phi_i(\vec{r}) \quad \text{(III.16)}$$

où $\hat{V}_H$ est le potentiel de Hartree et $\hat{V}_{xc}$ désigne le potentiel d'échange et corrélation contenant la correction d'énergie cinétique, l'échange et la corrélation électroniques. Ce système d'équations, qui est la base de la DFT de KS, est résolu par un processus itératif. Après le calcul d'une densité électronique initiale, le potentiel effectif est calculé et est utilisé pour déduire une nouvelle estimation de la densité électronique. Cette boucle est répétée jusqu'à convergence par rapport à un seuil ou un nombre maximum d'itérations.

Les équations en l'état sont exactes, mais le potentiel $\hat{V}_{xc}$ reste inconnu. Les approximations qui suivent ont donc pour but d'expliciter ce potentiel.

III. B. 3. Approximations du potentiel effectif

La DFT utilise différentes approximations pour décrire le potentiel d'échange-corrélation qui est indispensable pour déterminer la densité électronique du système et donc en déduire ses propriétés électroniques.

- Approximation de la densité locale (LDA)

La LDA est l'une des approches les plus simples. Dans cette méthode, les électrons sont modélisés comme un gaz uniforme autour des noyaux. L'énergie d'échange-corrélation prend alors la forme :

$$E_{xc}^{LDA}[\rho] = \int \rho(\vec{r})\varepsilon_{xc}(\rho(\vec{r}))d^3r \qquad \text{(III.17)}$$

où ε_{xc} désigne l'énergie d'échange-corrélation par électron. Dans cette formulation, les variations de la densité électronique sont locales et l'énergie d'échange-corrélation en un point ne dépend que de la densité électronique en ce point. L'énergie d'échange-corrélation peut être décomposée en deux contributions distinctes : une partie d'échange ε_x et une partie de corrélation ε_c. Dans le cadre de la LDA, la partie d'échange est connue sous le nom de la fonctionnelle d'échange de Slater-Dirac et s'exprime mathématiquement par :

$$E_x^{LDA}[\rho] = -\frac{3}{4}\left(\frac{3}{\pi}\right)^{1/3}\int \rho(\vec{r})^{4/3}d^3r \qquad \text{(III.18)}$$

En revanche, la partie corrélation n'est pas connue analytiquement et a donnée naissances a plusieurs fonctionnelles de corrélation, telles que VWN (Vosko-Wilk-Nursair)[22], PZ81 (Perdrew-Zunger)[23] et PW92c (Perdew-Wang)[24]. La fonctionnelle SVWN, combinant l'échange de Slater avec la corrélation VWN fait partie des premières fonctionnelles d'échange-corrélation développées.

- Approximation du Gradient Généralisé (GGA)

Une manière plus élaborée pour décrire le potentiel d'échange-corrélation consiste à ne plus décrire le nuage électronique comme uniforme et de prendre en compte les différences de densité électronique dans l'espace, permettant une description plus fidèle. Ceci se traduit mathématiquement par l'incorporation du gradient ($\vec{\nabla}$) de la la densité à l'énergie d'échange-corrélation :

$$E_{xc}^{GGA} = \int \rho(\vec{r})\varepsilon_{xc}\left(\rho(\vec{r}), \vec{\nabla}\rho(\vec{r})\right) d^3r \qquad \text{(III.19)}$$

Comme dans l'approche LDA, l'échange et la corrélation peuvent être traités séparément. Dans les méthodes GGA, la fonctionnelle d'échange de Becke[25] est couramment utilisée et souvent associée avec des fonctionnelles de corrélation comme LYP (Lee-Yang-Parr)[26] pour constituer la fonctionnelle BLYP, ou encore PW91 (Perdrew-Wang)[27] dans le cadre de la fonctionnelle BPW91. On peut également citer la fonctionnelle d'échange-corrélation PBE (Perdew-Burke-Ernzerhof)[28] parmi les fonctionnelles type GGA les plus fréquemment utilisées.

- Fonctionnelles Méta-GGA (m-GGA)

Le système peut être décrit de manière plus précise en incluant en plus la paramétrisation du Laplacien (∇^2) de la densité électronique, ou de la densité d'énergie cinétique de KS τ_s à l'énergie d'échange-corrélation :

$$E_{xc}^{m-GGA} = \int \rho(\vec{r})\varepsilon_{xc}\left(\rho(\vec{r}), \vec{\nabla}\rho(\vec{r}), \nabla^2\rho(\vec{r})\right) d^3r \qquad \text{(III.20)}$$

Parmi les fonctionnelles de type m-GGA les plus couramment utilisées, on retrouve les fonctionnelles TPSS (Tao-Perdew-Staroverov-Scuseria)[29] et M06-L[30].

- Fonctionnelles Hybrides

Les fonctionnelles hybrides introduisent une composante d'échange exacte E_x^{XX} à l'énergie d'échange-corrélation. Ce terme est calculé comme dans la méthode Hartree-Fock en considérant les orbitales de KS dans le but d'améliorer la précision du terme d'échange. La plus célèbre des fonctionnelles hybrides est incontestablement la fonctionnelle B3LYP,[31] qui incorpore 20 % d'échange exact dans le terme d'échange. D'autres exemples incluent des

pourcentages plus élevés tels que la fonctionnelle hybride PBE0[32] avec 25 % d'échange exact ou la fonctionnelle M06-2X[33], plus récente, incorporant 54 % d'échange exact.

- Fonctionnelles à séparation de portée (RSH)

Les fonctionnelles à séparation de portée ont été développées pour pallier les faiblesses des fonctionnelles hybrides, notamment leur incapacité à prédire avec précision des phénomènes tels que les transferts de charge ou les états de Rydberg dans les calculs de TD-DFT. Ces limitations sont attribuées à une description insuffisante des interactions électroniques à longue portée. Pour remédier à ce problème, ces fonctionnelles modulent la proportion d'échange exact en fonction de la distance interélectronique par la sommation d'Ewald d'une interaction à courte et à longue portée du potentiel de répulsion électronique :[34]

$$\frac{1}{r_{ij}} = \frac{1 - erf(\mu r_{ij})}{r_{ij}} + \frac{erf(\mu r_{ij})}{r_{ij}} \qquad \text{(III.21)}$$

avec μ le paramètre d'atténuation et erf la fonction erreur. La fonctionnelle CAM-B3LYP[35] (Coulomb Attenuatted Model-B3LYP) bénéficie d'une correction plus élaborée sous la forme :

$$\frac{1}{r_{ij}} = \frac{1 - [\alpha + \beta\, erf(\mu r_{ij})]}{r_{ij}} + \frac{\alpha + \beta\, erf(\mu r_{ij})}{r_{ij}} \qquad \text{(III.22)}$$

où α et β permettent respectivement d'incorporer la contribution d'échange exact et la contribution de l'équivalent DFT sur toute la gamme de distance interatomique. A cette séparation de portée peut s'ajouter une description des effets de corrélation à longue portée dans le cadre de la DFT-D avec comme exemple la fonctionnelle ωB97X-D[36].

- Densité non-locale

Enfin, certaines fonctionnelles dites non-locales virtuelles introduisent l'utilisation des orbitales de KS inoccupées. Cela permet de combiner la DFT avec approches issues de la théorie de perturbations du second ordre (PT2) ou des méthodes de type « *coupled-cluster* » (CC). Ces nouveaux types de fonctionnelles sophistiquées incluent les fonctionnelles d'échange-corrélation telles que B2PLYP[37] et ses variantes, et PBE0-DH[38].

- Échelle de Jacob

Ces différentes approximations peuvent être situées sur l'échelle de Perdew (**Figure III.1**) où chaque échelon représente un niveau de sophistication différent. En général, plus on monte dans cette échelle, plus la précision des calculs augmente, mais aux dépens du temps de calcul.[39]

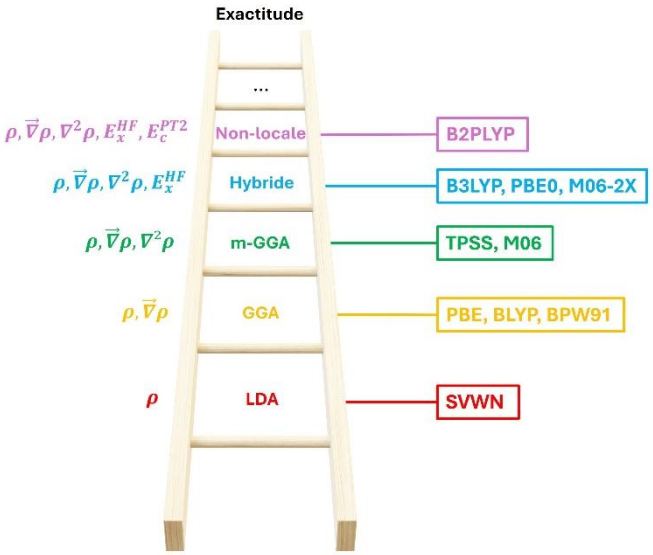


Figure III.1 *: Échelle de Perdew symbolisant la précision des méthodes d'approximation de la fonctionnelle d'échange-corrélation.*

- Fonction de base

Parmi les manières d'exprimer mathématiquement les orbitales atomiques, il existe deux grandes familles de fonction de base : les fonctions de type Slater (STO)[40] ϕ^{SF} et les fonctions de type gaussien (GTO) ϕ^{GF}. Les orbitales de KS sont décrites comme une combinaison linéaire de fonction de base $\chi_{\mu i}$:

$$\phi_i(\vec{r}) = \sum_{\mu=1}^{N_b} c_{\mu i}\chi_{\mu i}(\vec{r}) \tag{III.23}$$

où N_b désigne la taille de la base. La distribution angulaire de ces deux types de fonction est contrôlée par une fonction de type harmonique sphérique et les approximations mathématiques pour décrire l'orbitale $1s$ centrée en $\vec{R}_A$ sont :[41]

$$\chi_{1s}^{SF}(\zeta, \vec{r} - \vec{R}_A) = \left(\frac{\zeta^3}{\pi}\right)^{1/2} e^{-\zeta|\vec{r}-\vec{R}_A|} \tag{III.24}$$

$$\chi_{1s}^{GF}(\alpha, \vec{r} - \vec{R}_A) = \left(\frac{2\alpha}{\pi}\right)^{3/4} e^{-\alpha|\vec{r}-\vec{R}_A|^2} \tag{III.25}$$

Bien qu'elles aient des formulations similaires, ces fonctions de base ont un comportement différent lorsque $\vec{r} \to \vec{0}$ et $\vec{r} \to \infty$. De manière générale, les fonctions de Slater sont plus précises que les GTO. Par exemple la solution exacte de l'orbitale 1s de l'atome d'hydrogène est une fonction de Slater : $\left(\frac{1}{\pi}\right)^{1/2} e^{-r}$.

Cependant, d'un point de vue pratique, ce sont majoritairement les fonctions de type Gaussien qui sont utilisées, principalement car les combinaisons linéaires de GTO sont des GTO, ce qui diminue drastiquement le coût computationnel. Ainsi, pour décrire une STO, on peut par exemple utiliser une GTO, auquel cas la base est dite « minimale », ou plusieurs GTO, auquel cas la base est dite « multiple ». Chaque GTO est alors une gaussienne primitive.

Lorsque les bases sont à valence séparée, plusieurs STO sont utilisées avec des coefficients ζ différents pour décrire les orbitales de valence. Par exemple, pour des bases double ζ deux fonctions STO sont utilisées, pour des triple ζ trois fonctions et ainsi de suite.

Plus ζ ou α est petit, plus la fonction est large et diffuse. Ces fonctions dites « diffuses » peuvent être ajoutées pour décrire la densité électronique éloignée des noyaux de manière plus fidèle. De la même façon, des fonctions de polarisation, avec moment angulaire élevé permettant une meilleure description de la densité électronique, peuvent être ajoutées. Par exemple pour l'hydrogène, on ajoute des fonctions de type p, pour les éléments de la deuxième période, on ajoute des fonctions de type d et ainsi de suite. Les bases les plus courantes sont des bases à valences séparées et décrites dans la formulation de Pople[42] :

$$N - MOP + + G(d, p \ldots) \qquad \text{(III.26)}$$

dans laquelle N est le nombre de Gaussiennes primitives pour décrire les orbitales de cœur, M, O, P désignent le nombre de primitives pour décrire les orbitales de valence. Les « $+$ » désignent l'ajout de fonction diffuse. Le G symbolise l'utilisation de fonctions Gaussiennes. Enfin, les lettres d, p et parfois f symbolisent les types de fonction de polarisations rajoutées, elles peuvent aussi notées « * ».

Plus récemment, Dunning[43] a proposé une alternative avec les bases à corrélations consistantes polarisées ($cc - p$) optimisées à l'aide de méthodes plus récentes et plus précises (post-Hartree-Fock) que les bases de Pople (Hartree-Fock). Ces bases ont une autre notation :

$$aug - cc - pVNZ \qquad \text{(III.27)}$$

dans laquelle uniquement les orbitales de valences (V) sont décrites à l'aide d'une base multiple ζ (NZ) et les fonctions diffuses sont ajoutées par l'acronyme aug. Les bases de Sauer[44] sont des bases de type Dunning et optimisées pour les calculs RMN. Elles ont pour acronyme aug-cc-pVNZ-J.

III. C. Théorie de la Fonctionnelle de la Densité Dépendante du Temps (TD-DFT)

III. C. 1. Fondement de la TD-DFT

La TD-DFT est une extension de la DFT permettant d'étudier les propriétés électroniques des systèmes en fonction du temps. Elle est essentielle pour calculer les propriétés optiques des molécules car elle permet de déterminer les énergies de transition entre états électroniques. Sous la forme de l'Équation (III.11), la densité électronique n'apparaît pas explicitement. Runge et Gross ont alors proposé un théorème analogue au premier théorème de Kohn-Sham pour la DFT.[45]

- Théorème de Runge-Gross

« Soit un système de M électrons dans un potentiel externe dépendant du temps. Notons $\rho_1(\vec{r},t)$*, la densité électronique dépendante du temps de ce système dans le potentiel externe* $v_{ext,1}(\vec{r},t)$*, et* $\rho_2(\vec{r},t)$ *dans* $v_{ext,2}(\vec{r},t)$*, alors* ρ_1 *et* ρ_2 *seront différents si et seulement si* $v_{ext,1}$ *et* $v_{ext,2}$ *sont différents de plus d'une constante* $C(t)$ *»*

$$\rho(\vec{r},t) \leftrightarrow v_{ext}(\vec{r},t) + C(t) \tag{III.28}$$

Dans le cadre de la DFT, l'état fondamental est obtenu par minimisation de l'énergie totale. Ceci ne peut pas être appliqué dans les systèmes dépendants du temps car l'énergie totale n'est pas conservée. On fait donc appel à la fonctionnelle d'action quantique définie par :

$$A[\rho] = \int_{t0}^{t1} dt \left\langle \Psi(\{\vec{r_i}\},t) \middle| i\frac{\partial}{\partial t} - \hat{H}(t) \middle| \Psi(\{\vec{r_i}\},t) \right\rangle \tag{III.29}$$

laquelle est équivalente à l'équation de Schrödinger lorsque sa dérivée fonctionnelle est nulle :

$$\frac{\delta A[\rho]}{\delta \rho} = i\hbar\frac{\partial}{\partial t}|\Psi(\{\vec{r_i}\},t)\rangle - \hat{H}(t)|\Psi(\{\vec{r_i}\},t)\rangle = 0 \tag{III.30}$$

De plus, de la même façon qu'à l'état stationnaire, il est possible de considérer un système fictif d'électrons indépendants et évoluant dans un potentiel $\hat{V}_s(\vec{r},t)$ pour établir les équations de Kohn-Sham dépendantes du temps.[46] Dans ce nouveau système, on a donc :

$$\left[\hat{T}_i + \hat{V}_H + \hat{V}_{ne} + \hat{V}_{xc}\right]\phi_i(\{\vec{r_i}\},t) = i\hbar\frac{\partial\phi_i(\{\vec{r_i}\},t)}{\partial t} \tag{III.31}$$

dans lequel

$$V_{xc}(r,t) = \frac{\delta A_{xc}}{\delta\rho(\vec{r},t)} \tag{III.32}$$

et où A_{xc} est la partie d'échange-corrélation de l'action quantique décrite dans l'équation (III.29). Cependant, l'expression exacte de V_{xc} n'est pas connue analytiquement. Il est donc possible de faire certaines approximations pour l'expliciter. Par exemple, l'approximation de la densité locale adiabatique fondée sur l'hypothèse que la densité ne varie que très lentement dans le temps, permet une écriture de la fonctionnelle d'action sous la forme :

$$A_{xc} = \int_{t_0}^{t_1} dt E_{xc}[\rho] \tag{III.33}$$

On peut désormais expliciter le potentiel d'échange-corrélation sous la forme :

$$V_{xc}[\rho](\vec{r},t) = \frac{\delta A_{xc}[\rho]}{\delta\rho(\vec{r},t)} \approx \frac{\delta E_{xc}[\rho]}{\delta\rho(\vec{r})} = V_{xc}[\rho](\vec{r}) \tag{III.34}$$

Et donc d'utiliser la fonctionnelle d'énergie d'échange et de corrélation de la DFT stationnaire dans la version dépendante du temps. Cette approximation est surtout utilisée pour le calcul des états excités.

III. C. 2. Théorie de la réponse linéaire

Lorsque le potentiel externe est suffisamment faible, on peut retrouver le comportement du système à l'aide de la méthode perturbatrice dans laquelle la variation de densité électronique est considérée comme du premier ordre au cours du temps.[47]

Supposons que pour un temps $t < t_0$, le système n'est pas perturbé. Il est donc soumis uniquement au potentiel des noyaux $v_{ext} = v^{(0)}$. On a donc :

$$\rho(\vec{r}, t < t_0) = \rho^{(0)}(\vec{r}) \tag{III.35}$$

avec $\rho^{(0)}$ la densité électronique de l'état fondamental. Ensuite, en appliquant une perturbation $v^{(1)}$à partir de t_0, le potentiel externe devient $v_{ext} = v^{(0)} + v^{(1)}$. La densité électronique peut s'exprimer sous la forme d'une série perturbative :

$$\rho(\vec{r},t) = \rho^{(0)}(\vec{r}) + \rho^{(1)}(\vec{r},t) + \rho^{(2)}(\vec{r},t) + \cdots \quad \text{(III.36)}$$

avec $\rho^{(1)}$ la dépendance linéaire par rapport à $v^{(1)}$ et $\rho^{(2)}$ la dépendance quadratique et ainsi de suite. Dans le cadre de la théorie de la réponse linéaire la perturbation est considérée comme suffisamment faible pour que seul le terme linéaire soit retenu. La variation du premier ordre de la densité électronique peut être exprimée en fonction de la susceptibilité ou fonction réponse χ:

$$\rho^{(1)}(\vec{r},\omega) = \int \chi(\vec{r},\vec{r}',\omega) v^{(1)}(\vec{r}',\omega) d^3r' \quad \text{(III.37)}$$

Dans le cadre des équations de Kohn-Sham dépendantes du temps, on a donc :

$$\rho^{(1)}(\vec{r},\omega) = \int \chi_{KS}(\vec{r},\vec{r}',\omega) v_{KS}^{(1)}(\vec{r}',\omega) d^3r' \quad \text{(III.38)}$$

où χ_{KS} est la fonction réponse d'un système d'électrons indépendants et dans lequel la variation du potentiel de Kohn Sham est définie par

$$v_{KS}^{(1)}(\vec{r},t) = V_H^{(1)}(\vec{r},t) + v_{ext}^{(1)}(\vec{r},t) + V_{xc}^{(1)}(\vec{r},t) \quad \text{(III.39)}$$

On a donc :

$$\rho^{(1)}(\vec{r},\omega) = \int \chi_{KS}(\vec{r},\vec{r}',\omega)\{\delta V_H(\vec{r}',\omega) + \delta v_{ext}(\vec{r}',\omega) + \delta V_{xc}(\vec{r}',\omega)\} d^3r' \quad \text{(III.40)}$$

D'où

$$\begin{aligned} \chi(\vec{r},\vec{r}',\omega) = \chi_{KS}(\vec{r},\vec{r}',\omega) \\ + \int d^3r_1 \int \chi(\vec{r},\vec{r}_1,\omega)\{v_c(\vec{r}_1,\vec{r}_2) \\ + f_{xc}(\vec{r}_1,\vec{r}_2,\omega)\}\chi_{KS}(\vec{r}_2,\vec{r}',\omega) d^3r_2 \end{aligned} \quad \text{(III.41)}$$

avec $f_{xc}(\vec{r},t,\vec{r}',t)$ le noyau d'échange-corrélation et v_c la variation du potentiel de Hartree par rapport à la variation de densité électronique :

$$f_{xc}(\vec{r},t,\vec{r}',t') = \frac{\delta V_{xc}(\vec{r},t)}{\delta\rho(\vec{r}',t)} \quad \text{(III.42)}$$

$$v_c(\vec{r},t,\vec{r}',t') = \frac{\delta V_H(\vec{r},t)}{\delta\rho(\vec{r}',t')} \quad \text{(III.43)}$$

L'équation (III.41) peut être résolue par un processus itératif, permettant de déterminer la fonction réponse d'un système aux électrons non corrélés puis la densité électronique d'un système perturbé. Cette approche a montré son efficacité car la TD-DFT offre un compromis entre précision et coût computationnel, rendant possible l'étude des spectres d'absorption, d'émission et des états excités des systèmes de taille variable avec une bonne précision.[48]

III. C. 3. Application au Dichroïsme Circulaire Électronique

La TD-DFT est très utilisée pour la prédiction de propriétés optiques et permet notamment de calculer le signal de DCE des molécules. L'intensité des bandes de DCE est exprimée par le terme de force rotationnelle qui expérimentalement est décrit par la relation :[49]

$$R_{exp} = 2.297.10^{-39} \int_{\lambda_1}^{\lambda_2} \frac{\Delta\varepsilon(\lambda)}{\lambda} d\lambda \qquad \text{(III.44)}$$

dans laquelle λ (cm^{-1}) est le nombre d'onde, R_{exp} est exprimé en unité c.g.s $(10^{-40} esu\ cm\ erg\ G^{-1})$ et $\Delta\varepsilon$ $(M^{-1}\ cm^{-1})$ est le signal CD définit par la différence des absorptivités molaires de la PCG ε_g et de la PCD ε_d :

$$\Delta\varepsilon = \varepsilon_g - \varepsilon_d \qquad \text{(III.45)}$$

Ce signal CD peut également être décrit de manière quantique. En effet, l'absorptivité d'une molécule découle de la règle d'or de Fermi :

$$\varepsilon = \frac{\pi N_A \omega\ \delta(\omega - \omega_{fi})}{\varepsilon_0 \hbar c\ ln(10)} \frac{1}{E_0^2} \left|\langle f|\hat{H}|i\rangle\right|^2 \qquad \text{(III.46)}$$

avec N_A le nombre d'Avogadro, ω la pulsation, $\delta(\omega - \omega_{fi})$ la distribution de Dirac, ε_0 la permittivité diélectrique du vide, E_0 l'amplitude du champ électrique de la lumière polarisée circulairement et $|f\rangle$ et $|i\rangle$ désignent les états quantiques entre lesquels est décrite la probabilité de transition. A partir de l'équation (III.46), on peut montrer que la force rotationnelle s'exprime comme la partie imaginaire du produit scalaire entre le moment de transition électrique et le moment de transition magnétique :[50]

$$R_{f\to i} \propto Im\{\vec{\mu}_{fi} \cdot \vec{m}_{fi}\} \qquad \text{(III.47)}$$

Pour simuler le spectre DC mesuré expérimentalement, le signal DC de chaque transition est ensuite convolué par une fonction gaussienne selon :[49]

$$\Delta\varepsilon(E) = \frac{\Delta E_{i\to f}}{2.297.10^{-39}} \frac{R_{f\to i}}{\Delta\sqrt{\pi}} e^{-\left(\frac{E-\Delta E_{i\to f}}{\Delta}\right)^2} \qquad \text{(III.48)}$$

Le spectre DC final $\Delta\varepsilon_f$ n'est autre que la somme de toutes les transitions :

$$\Delta\varepsilon_f(E) = \frac{1}{2.297.10^{-39}} \frac{1}{\Delta\sqrt{\pi}} \sum_{i=1}^{n} \Delta E_{0\to i} R_{0\to i} e^{-\left(\frac{E-\Delta E_{0\to i}}{\Delta}\right)^2} \quad \text{(III.49)}$$

III. D. Inclusion du solvant

L'environnement d'une molécule peut avoir un effet considérable sur son comportement, sa géométrie et ses propriétés optiques. Il est donc primordial de le prendre en compte lors des calculs quantiques. L'environnement est en général réduit à des molécules de solvant, mais peut aussi inclure des contre-ions et des ligands ou des environnements plus complexes comme des membranes, des micelles SDS (chargées) ou DPC (zwitterion) ou des protéines. Ces molécules peuvent être considérées de manière implicite ou explicite lorsqu'elles sont incluses dans la description quantique du système complet. Il existe de nombreuses façons de décrire le solvant. Dans cette partie, uniquement deux types de modèles seront abordés : le continuum polarisable et l'inclusion polarisable.

III. D. 1. Modèle du Continuum Polarisable (PCM)

Dans ce modèle implicite, le solvant est considéré comme un milieu continu décrit par une constante diélectrique entourant la molécule ciblée.[51,52] Cette molécule est placée dans une cavité couramment construite à l'aide de sphères entourant ses atomes. Ainsi, le champ électrique généré par la densité de charges de la molécule va polariser le continuum, qui va à son tour impacter la distribution de charge de la molécule. D'un point de vue mathématique, le continuum est défini par le système :

$$\epsilon(\vec{r}) = \begin{cases} 1, \vec{r} \in C \\ \epsilon_r, \vec{r} \notin C \end{cases} \quad \text{(III.50)}$$

où $\epsilon(\vec{r})$ est la permittivité du milieu au rayon r par rapport au centre d'un atome, ϵ_r la permittivité du solvant et C la cavité dans laquelle est placée la molécule. L'information quantique du sytème est alors calculée à l'aide d'un nouvel hamiltonien :

$$\hat{H}^{eff} = \hat{H}^0 + \hat{V}^R \quad \text{(III.51)}$$

avec $\hat{H}^0$ l'opérateur hamiltonien dans le vide et $\hat{V}^R$ le potentiel d'interaction molécule-solvant.

Ce type de modèle pour inclure l'environnement a l'avantage d'être peu couteux en temps de calcul, mais ne prend pas en compte spécifiquement les interactions locales possibles comme les liaisons hydrogène

III. D. 2. Modèle de l'Inclusion Polarisable (PE)

Pour pallier ce manque de d'informations, il est possible d'avoir recourt à l'inclusion polarisable.[53,54] Ce second modèle fait partie des méthodes multi-échelle Quantum Mechanics/Mechanical Mechanics (QM/MM) dans lequel le système moléculaire est scindé en une partie MM qui décrit généralement le solvant éloigné de la molécule ciblée, et une partie QM qui contient généralement la molécule ciblée et parfois des molécules éloignées de la sphère de la solvatation. Dans ce type d'approche, différentes énergies sont prises en compte dans le calcul de l'énergie totale E_{tot} :

$$E_{tot} = E_{QM}^{PE} + E_{es}^{PE} + E_{ind}^{PE} \tag{III.52}$$

où E_{QM}^{PE} est l'énergie de la partie QM dans le vide, E_{es}^{PE} l'énergie due aux interactions électrostatiques et E_{ind}^{PE} l'énergie induite par les différents fragments sur l'environnement. Un nouvel hamiltonien effectif est alors défini pour prendre en compte les termes E_{es}^{PE} et E_{ind}^{PE}:

$$\hat{H}^{eff,PE} = \hat{H}^{0} + \hat{v}^{PE} \tag{III.53}$$

Le nouveau terme $\hat{v}^{PE}$ est défini de la manière suivante :[55]

$$\hat{v}^{PE} = \sum_{s=1}^{S}\sum_{k=0}^{K}\frac{(-1)^{(k+1)}}{k!}Q_s^{(k)}\sum_{pq}T_{s,pq}^{(k)}\hat{E}_{pq} - \sum_{s=1}^{S}\vec{\mu}_s^{ind}(F[\rho])\sum_{pq}T_{s,pq}^{(1)}\hat{E}_{pq} \tag{III.54}$$

L'originalité de cette approche réside dans le fait que ce nouveau terme contient à la fois l'information sur les charges avec le terme $Q_s^{(k)}$ représentant le multipôle d'ordre k du site s de l'environnement et également les polarisabilités, via le terme $\vec{\mu}_s^{ind}$ défini comme le vecteur contenant les moments dipolaires induits de chaque site s de l'environnement par les champs électriques F. Les indices pq désignant un élément de matrice de l'opérateur dans la base orbitale de KS. Les tenseurs d'interaction sont donc définis comme :

$$T_{s,pq}^{(k)} = \nabla^{k}\left[\frac{1}{|\vec{r}_s - \vec{r}_{pq}|}\right] \tag{III.55}$$

et le terme $\hat{E}_{pq}$ désigne un opérateur d'excitation exprimé sous forme d'un opérateur de création et d'annihilation. Cette formulation permet donc de prendre en compte l'effet de l'environnement sur le système quantique, mais également la polarisation de l'environnement par la partie QM. Cette méthode semble a priori adaptée à des systèmes dynamiques dont le réarrangement moléculaire est important et dont la réponse rapide du solvant à des changements de conformation du soluté doit être prise en compte. Elle permet la considération plus complexe et plus spécifique des interactions entre un solvant et son soluté que les méthodes PCM et a

notamment montré son efficacité dans les calculs des propriétés optiques (Fluorescence, DC, UV)[56–58].

III. E. Références

1. González, M. A. Force fields and molecular dynamics simulations. *Éc. Thématique Société Fr. Neutron.* **12**, 169–200 (2011).
2. Urey, H. C. & Bradley, C. A. The Vibrations of Pentatonic Tetrahedral Molecules. *Phys. Rev.* **38**, 1969–1978 (1931).
3. MacKerell, A. D. Jr. *et al.* All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins. *J. Phys. Chem. B* **102**, 3586–3616 (1998).
4. Cornell, W. D. *et al.* A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *J. Am. Chem. Soc.* **117**, 5179–5197 (1995).
5. Oostenbrink, C., Villa, A., Mark, A. E. & Van Gunsteren, W. F. A biomolecular force field based on the free enthalpy of hydration and solvation: The GROMOS force-field parameter sets 53A5 and 53A6. *J. Comput. Chem.* **25**, 1656–1676 (2004).
6. Jorgensen, W. L., Maxwell, D. S. & Tirado-Rives, J. Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids. *J. Am. Chem. Soc.* **118**, 11225–11236 (1996).
7. Ponder, J. W. *et al.* Current Status of the AMOEBA Polarizable Force Field. *J. Phys. Chem. B* **114**, 2549–2564 (2010).
8. Lamoureux, G. & Roux, B. Modeling induced polarization with classical Drude oscillators: Theory and molecular dynamics simulation algorithm. *J. Chem. Phys.* **119**, 3025–3039 (2003).
9. van Duin, A. C. T., Dasgupta, S., Lorant, F. & Goddard, W. A. ReaxFF: A Reactive Force Field for Hydrocarbons. *J. Phys. Chem. A* **105**, 9396–9409 (2001).
10. Senftle, T. P. *et al.* The ReaxFF reactive force-field: development, applications and future directions. *Npj Comput. Mater.* **2**, 1–14 (2016).
11. Unke, O. T. *et al.* Machine Learning Force Fields. *Chem. Rev.* **121**, 10142–10186 (2021).

12. Smith, J. S., Isayev, O. & Roitberg, A. E. ANI-1: an extensible neural network potential with DFT accuracy at force field computational cost. *Chem. Sci.* **8**, 3192–3203 (2017).
13. Schütt, K. T., Sauceda, H. E., Kindermans, P.-J., Tkatchenko, A. & Müller, K.-R. SchNet – A deep learning architecture for molecules and materials. *J. Chem. Phys.* **148**, 241722 (2018).
14. Berendsen, H. J. C., Postma, J. P. M., van Gunsteren, W. F., DiNola, A. & Haak, J. R. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* **81**, 3684–3690 (1984).
15. Hoover, W. G. Canonical dynamics: Equilibrium phase-space distributions. *Phys. Rev. A* **31**, 1695–1697 (1985).
16. Andersen, H. C. Molecular dynamics simulations at constant pressure and/or temperature. *J. Chem. Phys.* **72**, 2384–2393 (1980).
17. Bussi, G., Donadio, D. & Parrinello, M. Canonical sampling through velocity rescaling. *J. Chem. Phys.* **126**, 014101 (2007).
18. Parrinello, M. & Rahman, A. Polymorphic transitions in single crystals: A new molecular dynamics method. *J. Appl. Phys.* **52**, 7182–7190 (1981).
19. Schrödinger, E. Quantisierung als Eigenwertproblem. *Ann. Phys.* **384**, 361–376 (1926).
20. Born, M. & Oppenheimer, R. Zur Quantentheorie der Molekeln. *Ann. Phys.* **389**, 457–484 (1927).
21. Hohenberg, P. & Kohn, W. Inhomogeneous Electron Gas. *Phys. Rev.* **136**, B864–B871 (1964).
22. Vosko, S. H., Wilk, L. & Nusair, M. Accurate spin-dependent electron liquid correlation energies for local spin density calculations: a critical analysis. *Can. J. Phys.* **58**, 1200–1211 (1980).

23. Perdew, J. P. & Zunger, A. Self-interaction correction to density-functional approximations for many-electron systems. *Phys. Rev. B* **23**, 5048–5079 (1981).
24. Perdew, J. P. & Wang, Y. Accurate and simple analytic representation of the electron-gas correlation energy. *Phys. Rev. B* **45**, 13244–13249 (1992).
25. Becke, A. D. Density-functional exchange-energy approximation with correct asymptotic behavior. *Phys. Rev. A* **38**, 3098–3100 (1988).
26. Lee, C., Yang, W. & Parr, R. G. Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density. *Phys. Rev. B* **37**, 785–789 (1988).
27. Perdew, J. P. *et al.* Atoms, molecules, solids, and surfaces: Applications of the generalized gradient approximation for exchange and correlation. *Phys. Rev. B* **46**, 6671–6687 (1992).
28. Perdew, J. P., Burke, K. & Ernzerhof, M. Generalized Gradient Approximation Made Simple. *Phys. Rev. Lett.* **77**, 3865–3868 (1996).
29. Tao, J., Perdew, J. P., Staroverov, V. N. & Scuseria, G. E. Climbing the Density Functional Ladder: Nonempirical Meta--Generalized Gradient Approximation Designed for Molecules and Solids. *Phys. Rev. Lett.* **91**, 146401 (2003).
30. Zhao, Y. & Truhlar, D. G. A new local density functional for main-group thermochemistry, transition metal bonding, thermochemical kinetics, and noncovalent interactions. *J. Chem. Phys.* **125**, 194101 (2006).
31. Becke, A. D. Density-functional thermochemistry. I. The effect of the exchange-only gradient correction. *J. Chem. Phys.* **96**, 2155–2160 (1992).
32. Adamo, C. & Barone, V. Toward reliable density functional methods without adjustable parameters: The PBE0 model. *J. Chem. Phys.* **110**, 6158–6170 (1999).
33. Zhao, Y. & Truhlar, D. G. The M06 suite of density functionals for main group thermochemistry, thermochemical kinetics, noncovalent interactions, excited states, and

transition elements: two new functionals and systematic testing of four M06-class functionals and 12 other functionals. *Theor. Chem. Acc.* **120**, 215–241 (2008).

34. Tawada, Y., Tsuneda, T., Yanagisawa, S., Yanai, T. & Hirao, K. A long-range-corrected time-dependent density functional theory. *J. Chem. Phys.* **120**, 8425–8433 (2004).
35. Yanai, T., Tew, D. P. & Handy, N. C. A new hybrid exchange–correlation functional using the Coulomb-attenuating method (CAM-B3LYP). *Chem. Phys. Lett.* **393**, 51–57 (2004).
36. Chai, J.-D. & Head-Gordon, M. Long-range corrected hybrid density functionals with damped atom–atom dispersion corrections. *Phys. Chem. Chem. Phys.* **10**, 6615–6620 (2008).
37. Grimme, S. Semiempirical hybrid density functional with perturbative second-order correlation. *J. Chem. Phys.* **124**, 034108 (2006).
38. Brémond, E. & Adamo, C. Seeking for parameter-free double-hybrid functionals: The PBE0-DH model. *J. Chem. Phys.* **135**, 024106 (2011).
39. Perdew, J. P. & Schmidt, K. Jacob's ladder of density functional approximations for the exchange-correlation energy. *AIP Conf. Proc.* **577**, 1–20 (2001).
40. Slater, J. C. Atomic Shielding Constants. *Phys. Rev.* **36**, 57–64 (1930).
41. Szabo, A. & Ostlund, N. S. *Modern Quantum Chemistry: Introduction to Advanced Electronic Structure Theory*. (Courier Corporation, 2012).
42. Ditchfield, R., Hehre, W. J. & Pople, J. A. Self-Consistent Molecular-Orbital Methods. IX. An Extended Gaussian-Type Basis for Molecular-Orbital Studies of Organic Molecules. *J. Chem. Phys.* **54**, 724–728 (1971).
43. Dunning, T. H., Jr. Gaussian basis sets for use in correlated molecular calculations. I. The atoms boron through neon and hydrogen. *J. Chem. Phys.* **90**, 1007–1023 (1989).
44. Provasi, P. F., Aucar, G. A. & Sauer, S. P. A. The effect of lone pairs and electronegativity on the indirect nuclear spin–spin coupling constants in CH2X (X=CH2, NH, O, S): Ab

initio calculations using optimized contracted basis sets. *J. Chem. Phys.* **115**, 1324–1334 (2001).

45. Runge, E. & Gross, E. K. U. Density-Functional Theory for Time-Dependent Systems. *Phys. Rev. Lett.* **52**, 997–1000 (1984).
46. Casida, M. E. & Chong, D. P. Recent advances in density functional methods. (1995).
47. Gross, E. K. U. & Burke, K. Basics. in *Time-Dependent Density Functional Theory* (eds. Marques, M. A. L. et al.) 1–13 (Springer, Berlin, Heidelberg, 2006).
48. Adamo, C. & Jacquemin, D. The calculations of excited-state properties with Time-Dependent Density Functional Theory. *Chem. Soc. Rev.* **42**, 845–856 (2013).
49. Stephens, P. J. & Harada, N. ECD cotton effect approximated by the Gaussian curve and other methods. *Chirality* **22**, 229–233 (2010).
50. Andrews, S. S. & Tretton, J. Physical Principles of Circular Dichroism. *J. Chem. Educ.* **97**, 4370–4376 (2020).
51. Miertuš, S., Scrocco, E. & Tomasi, J. Electrostatic interaction of a solute with a continuum. A direct utilizaion of AB initio molecular potentials for the prevision of solvent effects. *Chem. Phys.* **55**, 117–129 (1981).
52. Tomasi, J., Mennucci, B. & Cammi, R. Quantum Mechanical Continuum Solvation Models. *Chem. Rev.* **105**, 2999–3094 (2005).
53. Olsen, J. M. H. & Kongsted, J. Chapter 3 - Molecular Properties through Polarizable Embedding. in *Adv. Quantum Chem.* (eds. Sabin, J. R. & Brändas, E.) vol. 61 107–143 (Academic Press, 2011).
54. List, N. H., Olsen, J. M. H. & Kongsted, J. Excited states in large molecular systems through polarizable embedding. *Phys. Chem. Chem. Phys.* **18**, 20234–20250 (2016).
55. Olsen, J. M., Aidas, K. & Kongsted, J. Excited States in Solution through Polarizable Embedding. *J. Chem. Theory Comput.* **6**, 3721–3734 (2010).

56. Bonvicini, A. *et al.* Rational design of novel fluorescent analogues of cholesterol: a “step-by-step” computational study. *Phys. Chem. Chem. Phys.* **21**, 15487–15503 (2019).
57. Migliore, M. *et al.* Characterization of β-turns by electronic circular dichroism spectroscopy: a coupled molecular dynamics and time-dependent density functional theory computational study. *Phys. Chem. Chem. Phys.* **22**, 1611–1623 (2020).
58. Creutzberg, J. & Hedegård, E. D. A method to capture the large relativistic and solvent effects on the UV-vis spectra of photo-activated metal complexes. *Phys. Chem. Chem. Phys.* **25**, 6153–6163 (2023).

IV. Résultats – Étude du peptide structuré Piv-Pro-D-Ser-NHMe

Table des matières

IV. A. Choix du peptide

La première partie de ce travail de thèse est destinée à mettre en place un protocole mixte, mêlant théorie et expérience, pour déterminer les spectres DCE de références des coudes de type β. Ces travaux sont fondés sur la preuve de concept, établie préalablement à mon arrivée au laboratoire, qui a permis de montrer la possibilité d'établir les spectres DCE théoriques des coudes de types β-I, β-I', β-II et β-II'.[1] Cette étude, purement théorique, montre la génération de spectres DCE par la TD-DFT sur des peptides modèles Ace-X-Ala-NHMe et Ace-X-Ala-NHMe dont les géométries ont été obtenues par des dynamiques moléculaires munies de fortes contraintes sur les angles dièdres φ et ψ des résidus centraux. Nous souhaitons adapter ce protocole à un peptide réel, adoptant une structure coudée majoritaire en solution. Notre objectif à plus long terme est, une fois ce protocole validé, d'établir par une approche théorique des spectres de référence pour les principaux coudes rencontrés dans les protéines (12 types β, 2 types γ). Une fois ces références établies, elles pourront être utilisées pour caractériser la structure secondaire de petits peptides comportant des coudes via l'enregistrement de spectres DCE expérimentaux, ce qui n'est pas possible à l'heure actuelle pour les raisons évoquées dans les chapitres I et II.

Pour parvenir à cet objectif, nous avons entrepris une démarche avec d'un côté, l'élaboration d'un modèle théorique de simulation de spectres DCE en solution, et de l'autre, des analyses expérimentales de DCE et de RMN également en solution. Pour que notre approche soit valide, le choix du peptide est primordial car ce dernier doit remplir plusieurs critères. D'abord, il doit être soluble dans un solvant facilement modélisable, ne comporter qu'un seul type de structure secondaire et se structurer selon un type de coude de manière majoritaire dans ce solvant. Ensuite, bien qu'il soit possible que certains coudes aient une signature DCE proche des spectres de référence des hélices ou des feuillets, nous souhaitons d'abord choisir un peptide pour lequel le signal DCE dans ce solvant est différent de celui associé aux autres familles de structure secondaire pour qu'il soit probable que le peptide se replie sous la forme d'un coude en solution. Enfin, il doit être assez court pour que les calculs TD-DFT soit réalisables.

Notre choix s'est porté sur le peptide Piv-Pro-D-Ser-NHMe (**Figure IV.1**). Cette séquence a déjà été étudiée par Marraud et Aubry et devrait réunir tous les atouts nécessaires pour parvenir à notre objectif.[2,3] En effet, des études par diffraction des rayons X montrent que le peptide possède une structure cristallographique de type coude β-II et qu'il possède une signature DCE

différente des autres structures secondaires dans l'eau, un solvant modélisable. De plus il est constitué des groupements pivaloyl et proline, qui sont des éléments très structurants de la chaîne peptidique. La structure pourrait donc se conserver en solution. Ce peptide très court de 46 atomes est une séquence qui convient pour limiter le temps de calcul.

Nous avons donc dans un premier temps développé notre protocole mixte sur ce peptide. Une analyse DCE a permis de vérifier la reproductibilité des spectres et d'avoir un appui expérimental, actualisé par des outils de DCE plus avancés, pour adapter notre modèle théorique. Une analyse RMN a été réalisée afin de vérifier la conservation de la structure en solution et de trouver des indices de structuration en coude β-II avec notamment : i) l'analyse des Effets Nucléaires Overhauser (NOE) entre les hydrogènes proches dans l'espace permettant de calculer les distances entres ces atomes ; ii) la détermination des coefficients de température des protons H_N qui traduisent la protection de ces hydrogènes et la possibilité d'une liaison hydrogène ; iii) la mesure de la constante de couplage H_N-H_α qui, en vertu de la relation de Karplus, est directement liée à la valeur de l'angle dièdre φ.

Parallèlement, l'approche théorique s'est appuyée sur le modèle établi, préalablement à ce travail, par l'équipe du laboratoire COBRA.[1] Le protocole comprend une dynamique moléculaire, suivie du calcul des énergies de transition électroniques du peptide avec la prise en compte du solvant par une méthode multi-échelle QM/MM d'inclusion polarisable. Le champ de forces, le temps de dynamique et le modèle d'eau ont été optimisés pour valider un protocole théorique cohérent avec les résultats expérimentaux.

Cette étude a fait l'objet d'une publication dans le journal J. Phys. Chem. B :

Menant, S., Tognetti, V., Oulyadi, H., Guilhaudis, L. & Ségalas-Milazzo, I. A Joint Experimental and Theoretical Study on the Structural and Spectroscopic Properties of the Piv-Pro-d-Ser-NHMe Peptide. *J. Phys. Chem. B* **128**, 6704–6715 (2024).

Le « *Supporting Information* » est disponible dans l'annexe. L'article dans la partie suivante est suivi d'une partie de conclusions et perspectives

IV. B. Article

A joint experimental and theoretical study on the structural and spectroscopic properties of the Piv-Pro-D-Ser-NHMe peptide

Sébastien MENANT, Vincent TOGNETTI,* Hassan OULYADI, Laure GUILHAUDIS,* Isabelle SEGALAS-MILAZZO*

Univ Rouen Normandie, INSA Rouen Normandie, CNRS, Normandie Univ, COBRA UMR 6014, INC3M FR 3038, F-76000, Rouen, France

Abstract

In this paper, we investigate the secondary structure of the Piv-Pro-D-Ser-NHMe peptide by means of nuclear magnetic resonance and electronic circular dichroism (ECD) experiments, in conjunction with theoretical simulations based on molecular dynamics and time-dependent density functional theory calculations including polarizable embedding to account for solvent effects. The various experimental and theoretical protocols are assessed and validated, and are shown to provide a consistent description of the turn structure adopted by this peptide in solution. In addition, a simple fitting procedure is proposed to make the simulated and experimental ECD almost perfectly match. This full methodology is finally tested on another small peptide, enlightening its efficiency and robustness.

Keywords: peptide structure, β-turn conformation, electronic circular dichroism, nuclear magnetic resonance, time-dependent density functional theory, polarizable embedding

Corresponding authors: vincent.tognetti@univ-rouen.fr, laure.guilhaudis@univ-rouen.fr, isabelle.milazzo@univ-rouen.fr

IV. B. 1. Introduction

Electronic Circular Dichroism (ECD) is largely used today for polypeptides and protein folding analysis.[4] Its main interest is to quickly provide the relative amounts of secondary structures in solution (helix, sheet and turn)[5] that are related to the three-dimensional arrangement of biomolecules and linked to their biological activity.[6] In addition, the secondary structure itself may be the target of biological activity. This is the case of turns during specific cellular recognition processes.[7] Thus, identifying the conformation in solution and the secondary structures involved in the folding is of paramount importance, and it is highly desirable that it can be achieved using ECD spectroscopy for any type of secondary structure elements.

Among them, the already mentioned turns are the third most common type (behind helices and sheets) in proteins and represent more than 30% of secondary structures.[8] They are versatile structures, allowing them to be responsible for specific biological activities, but making ECD spectrum characterization an intricate task due to their wide diversity. Besides, in small peptides, it is hard to maintain them stable in solution, while, in proteins, there is often too much superposition on ECD spectrum. Indeed, the measured experimental spectrum is the result of all secondary structure contributions in the polypeptide, and the spectrum intensity is depending on the number of amino acids involved in the secondary structure. Since turns are only composed of few residues (whereas helices or sheets are often made of several tens of monomers), there are currently few reliable references of ECD spectra for turns.

For such reasons, research on turn spectroscopic signatures is still in development, and the current trend is to tackle this issue with a complementary theoretical aspect, in particular for β-turns that constitute one of the most common turn categories.[9,10] More specifically, six β-turns types (namely β-I, β-II, β-III with their mirror images β-I', β-II', β-III') were initially defined by Venkatachalam,[11] depending on the values for the φ and ψ dihedral angles associated to the peptide backbone, often characterized by an intramolecular hydrogen bond between residues i, $i + 3$ (see **Figure IV.1**).

Then, simulated ECD spectra for β-I, -II and -III were computed by Woody by means of semi-empirical calculations, which afforded similar ECD patterns for β-I and β-II types,[12] which can be rationalized in terms of the nature of involved electronic transitions. Indeed, Woody proposed that the n→π* transition generates a negative band at 225 nm, while the π→π* transition creates a positive band at 205 nm and a negative band at 190 nm. Experimental studies were then carried out to confirm these results, but contradictory results were obtained.[13] It

should be mentioned that Woody did not explicitly consider the β-III types as he considered them being canonically close to α-helices for which ECD patterns are well known.

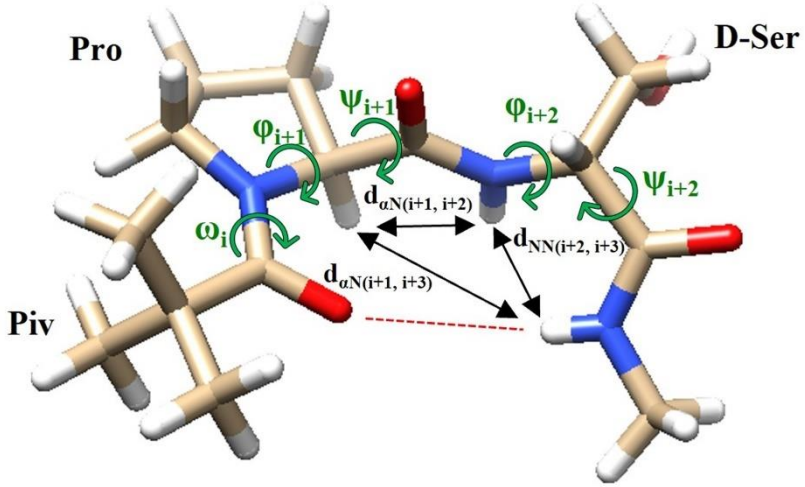


Figure IV.1*: View of the Piv-Pro-D-Ser-NHMe peptide with valuable distances and dihedral angles for β-turn identification. The red dotted line is the specific hydrogen bond of β-turns.*

One decade later, and in order to challenge Woody's statements, Sathyanarayana and Applequist[14] studied a model peptide (Ace-X-X-NHMe with X=Ala, Gly) still using semi-empirical theoretical methods. They found that β-I and β-III types shared similar spectra, with a negative band at 205 nm and a positive one at 190 nm and the mirror spectrum was obtained for β-II types. Such findings fully contradicted Woody's conclusions,[12] calling for a detailed reappraisal of the proposed ECD characterization of β-turns. This can be nowadays achieved using state-of-the-art Time Dependent-Density Functional Theory (TD-DFT) to simulate ECD spectra.

Indeed, such Quantum Chemistry (QC) method is often used to predict optical properties by computing the transition energies between electronic states, and affords an accuracy that clearly outperforms that of semi-empirical methods mentioned in the previous paragraph. In 2011, TD-DFT was notably coupled with classical Molecular Dynamics (MD) on a helix by Kaminsky.[15] The feasibility of such a simulation protocol was demonstrated by generating a ECD spectrum consistent with experimental data, and the importance of taking the solvent into account was spotlighted. In this seminal paper, global hybrid exchange-correlations were chosen, but it was later shown that range-separated hybrid ones are more suited.[16,17]

This led us to propose in 2019 a new MD/TD-DFT protocol[1] on a model peptide Ace-X-X-NHMe, focusing on simulated ECD spectra for β-I, β-II, β-I' and β-II' types. At variance with older papers, canonical dihedral angles were chosen according to the De Brevern's up-to-date classification (**Table IV.1**)[18] which defines 12 types of β-turns through the φ_{i+1}, ψ_{i+1}, φ_{i+2} and ψ_{i+2} dihedral angles (represented in **Figure IV.1**), where *i+1* and *i+2* are the central residues (see Table 1 for the typical values). Our previous QC calculations were performed within a state-of-the-art multiscale QM/Molecular Mechanics (MM) framework, in which the environment effects (solvent molecules and counterions) were described using a sophisticated polarizable embedding scheme,[19] and the final ECD spectrum was calculated by averaging over 1000 representative geometries. This model still has to be confronted with experimental data. In this work, our objective was thus to adapt this model to realistic experimental conditions to predict ECD signatures of β-turns of real-life peptides.

Type	φ_{i+1} (°)	ψ_{i+1} (°)	φ_{i+2} (°)	ψ_{i+2} (°)
β-I	-60	-30	-90	0
β-I'	+60	+30	+90	0
β-II	-60	+120	+80	0
β-II'	+60	-120	-80	0
β-IV$_1$	-120	+130	+55	+41
β-IV$_2$	-85	-15	-125	+55
β-IV$_3$	-71	-30	-72	-47
β-IV$_4$	-97	-2	-117	-11
β-VI$_{a1}$	-60	+120	-90	0
β-VI$_{a2}$	-120	+120	-60	0
β-VI$_b$	-135	+135	-75	+160
β-VIII	-60	-30	-120	+120

Table IV.1*: Values for canonical dihedral angles φ and ψ in the i+1 and i+2 residues according to De Brevern β-turn classification.16*

To this aim, the theoretical protocol has to be applied on a well-chosen peptide, according to the following specifications:

(*i*) the peptide should be in β-turn conformation in its crystallographic structure,

(*ii*) it should feature structural elements that could help it to remain structured in solution,

(*iii*) it should display an ECD spectrum that differs from those obtained for helices and sheets.

For these three reasons, peptide Piv-Pro-D-Ser-NHMe was selected **Figure IV.1**. Indeed, it was identified as a β-II turn by X-Ray-Diffraction experiments,[2] while the pivaloyl and proline residues are assumed to provide additional stability to maintain a strong structuration in common solvents. Besides, the experimental ECD spectrum measured by Marraud and Aubry clearly departs from those that are characteristic for more common secondary structures.[3]

The present paper, which is a joint experimental and theoretical study of this peptide, will be structured in two main parts. The first one, coming after the "material and methods" section (that describes both experimental and theoretical protocols), will consist in the experimental structural characterization by means of ECD and Nuclear Magnetic Resonance (NMR) techniques. Its purpose is to give definitive evidence of the β-turn structure adopted by this peptide in solution by performing additional experiments (with a particular focus on two-dimensional NMR experiments) to those already carried out by Marraud and Aubry.[19]

This structural characterization will be supplemented in a second part by QM/MM simulations that will allow for a better understanding of the conformational equilibrium. Besides, we will assess the influence of each simulation parameter and the performances of our computational protocol to predict the experimental ECD spectrum. It will be also validated on a second peptide, namely Gly-Ala-Gly that has already been investigated by other groups.[20–23] All of the obtained results will then be discussed within the general framework of structural biology before a general conclusion.

IV. B. 2. Material and methods

IV. B. 2. a) Chemical compounds

All experimental analyses were performed on the Piv-Pro-D-Ser-NHMe peptide (depicted in **Figure IV.1**) synthesized by Genepep (Saint-Jean-de-Védas, France) and solvents were supplied by CND Isotopes (Sainte Foy la Grande, France).

IV. B. 2. b) Electronic circular dichroism experiments

The ECD spectra were recorded between 185 and 260 nm at room temperature in acetonitrile, methanol, and water on a MOS 500 ECD (Biologic, Seyssinet-Pariset, France) spectrometer using a Xenon source and between 6 °C and 36 °C in water. The peptide concentration was set to 0.333 mg mL^{-1} (1.10 mM) for the three solvents with a cell of 0.1 cm length. For each sample, three scans were recorded using different ranges for the ECD signal detection (100 mdeg, 300 mdeg, 1000 mdeg) to check the reproducibility and determine the lowest wavelength allowing to obtain reliable experimental data.

The final ECD spectrum was obtained by subtracting the average of three blanks to the average of three sample spectra, and the intensity was fixed to 0 mdeg at 260 nm, where no significant ECD signal was expected. Once the minimum wavelength determined, we chose to keep for each sample the spectra recorded with the lower range (100 mdeg) to maximize the signal-to-noise ratio. To compare experimental and theoretical results, our experimental ECD spectra were plotted in the rotatory strength (M^{-1} cm^{-1}) formalism.

IV. B. 2. c) Nuclear magnetic resonance experiments

NMR experiments were recorded on a Bruker Avance III 600 MHz spectrometer, equipped with a cryoprobe. Two peptide concentrations were used: 0.333 and 1.667 mg mL^{-1} for aqueous solution containing 10% D_2O. Sodium 2,2-dimethyl-2-silapentane-5-sulphonate (DSS) was added as an internal reference. The temperature was set at 293 K and the following conventional two-dimensional experiments were carried out at 600 MHz: 2D COSY, TOCSY and NOESY (respectively Bruker's *cosygpprqf*, *dipsi2esgpph* and *noesyesgpph* pulse programs). NOESY experiments were also recorded using the noesyesgpphzs pulsequence developed by Thrippleton and Keeler to suppress zero quantum cross-peaks and to improve the accuracy of NOE cross-peaks integration.[24] TOCSY spectra were acquired with a 60 ms spin-lock time using a DIPSI sequence.[25] NOESY spectra were recorded with mixing times from 100 ms to 800 ms. TOCSY and NOESY experiments were performed in the phase-sensitive mode, using the time proportional phase incrementation method for quadrature detection (States-TPPI).[26]

For the TOCSY and NOESY experiments, water suppression was achieved using the excitation sculpting method.[27] For the COSY experiments, water resonance was suppressed with a low-power presaturation.

The spectra were recorded with 512 (t_1) × 4096 (t_2) data points and with a proton spectral width of 6000 Hz. The data were processed using TOPSPIN 4.1.4 (Bruker). The f1 dimension was zero-filled to 1024 real data points with the f1 and f2 dimensions being multiplied by shifted sine-bell functions prior to Fourier transform. Proton chemical shifts were reported relative to DSS taken as an internal reference, following the IUPAC[28] (see Table S1). Spectra used for chemical shift assignment were recorded before the addition of DSS to ensure the absence peptide/reference compound interaction.

To determine H_N temperature coefficients, 1D ^{1}H spectra were recorded at 400 MHz from 279 K to 307 K using a 2 K step.

The H_N-H_α, ^3J coupling constants was measured on an NMR ^{1}H 1D experiment (*zgesgp*) recorded at 600 MHz (resolution 0.09 Hz/pt) at 293 K, and the Karplus equation was used to infer the corresponding θ dihedral angle according to:

$$^3J = A\cos^2\theta + B\cos\theta + C, \quad (1)$$

with A, B and C three coefficients, whose values are taken according to either Cung *et al.* (set 1), Pardi *et al.* (set 2), Ludvigsen *et al.* (set 3) or Vuister *et al.* (set 4).[29–32]

$d_{\alpha N(i+1,\, i+2)}$, $d_{\alpha N(i+1,\, i+3)}$ and $d_{NN(i+2,\, i+3)}$ experimental distances (**Figure IV.1**) were calculated by using NOE correlations from a NOESY spectrum recorded at $\tau_m = 500\,ms$ with a zero quantum filter and at 600 MHz according to:

$$r_{ij} = r_{ref}\left(\frac{a_{ref}}{a_{ij}}\right)^{\frac{1}{6}}, \quad (2)$$

where r_{ij} is the chosen calculated distance, a_{ref} and a_{ij} are respectively the reference and the chosen NOE correlation intensities; r_{ref} is the reference distance of 1.75 Å between the two $H_{\beta,Pro}$.[33] This equation is valid for rigid molecules and relies on the assumption of uniform isotropic molecular tumbling.[34] To take into account possible internal mobility and integration errors due to residual zero-quantum contamination, an error of 10% was associated to each experimental distances.

IV. B. 2. d) Computational details: peptide building

Initial Cartesian coordinates of the Piv-Pro-D-Ser-NHMe crystal structure were taken from the mol file available at the CCDC[35] website deposited by Aubry *et al*. Missing hydrogen atoms were added using Chimera 1.16 software.[36]

IV. B. 2. e) Computational details: Molecular dynamics simulations

MD simulations were performed using the GROMACS 5.1.4 software.[37] Natural amino acids were described by the OPLS-AA[38] or CHARMM27[39] (CHARMM22 + CMAP for protein) force fields. The partial charges for all atoms of the Pivaloyl $(CH_3)_3CCO$ residue were set to the ones for the similar acetyl residue CH_3-CO and the charge was set to 0 on the added tertiary carbon. Then the molecule was solvated in a water cubic box (around 27 nm^3) using several water models (TIP3P,[40] TIP4P,[41] TIP5P,[42] or TIPS3P[43]) and the LINCS algorithm on bond constraints.

An energy minimization step was run, followed by an NVT equilibration at 300 K steps for 0.1 ns with Bussi's velocity rescale thermostat (0.02 ps time-step)[44] and a subsequent 0.1 ns NPT equilibration with Parrinello-Rahman barostat for 0.1 ns.[45] Electrostatics were calculated with the Particle Mesh Ewald (PME) method and the short-range van der Waals cutoff was set to 1 nm. The production run consists in NPT simulation of 10, 100 and 1000 ns with geometries saved every 0.01 ns. Only 100 or 1000 geometries, which were chosen at a regular time step, were used for the additional QC calculations.

To compare our experimental distances to the ones expected in β-I, β-II and β-VIII conformation, supplementary MD simulations were performed using a similar protocol in which φ_{i+1}, ψ_{i+1}, φ_{i+2} and ψ_{i+2} dihedral angles were constrained (see the last section in the supporting information (SI) file).

IV. B. 2. f) Computational details: Quantum Mechanics/Molecular Mechanics calculations

In this hybrid description of the full system, the peptide was described at the QM level, while the solvent molecules defined the MM part. The ECD spectra were calculated on each saved geometry from the MD simulation, for the 20 first excited states by linear-response TD-DFT using the DALTON 2018 software,[46] the range-separated CAM-B3LYP exchange-correlation functional, and the 6-31+G* basis set.[47] The influence of the environment was taken into account by the Polarizable Embedding (PE) model involving the effective Hamiltonian

$$\widehat{H}^{eff} = \widehat{H}^{KS} + \hat{v}^{PE}, \quad (3)$$

where $\widehat{H}^{KS}$ denotes the vacuum Kohn-Sham Hamiltonian, and $\hat{v}^{PE}$ is the operator describing the interaction between the MM environment and the QM molecule. It includes the electrostatic interaction between permanent charges and the QM region, but also the polarization of the environment by the QM part, which is not standard in common QM/MM modelling. The potential file was generated by the PyFraME library.[48] Initially, each of the n transition peaks was convoluted by Gaussian functions:[49]

$$\Delta\varepsilon_{total}(E) = \frac{1}{2.297.10^{-39}} \frac{1}{\Delta\sqrt{\pi}} \sum_{i=1}^{n} \Delta E_{0\to i} R_{0\to i} e^{-\left(\frac{E-\Delta E_{0\to i}}{\Delta}\right)^2}, \quad (4)$$

where $\Delta\varepsilon_{total}$ is the molar ellipticity, here expressed in M^{-1} cm^{-1}. $\Delta E_{0\to i}$ and $R_{0\to i}$ are respectively the transition energy and the rotatory strength, and Δ the bandwidth set to the 0.333 eV default value.

A four-parameter shifting strategy was implemented to improve the matching between experimental and theoretical results according to:

$$\Delta\varepsilon_{total}(E) = \frac{1}{2.297.10^{-39}} \frac{1}{\Delta\sqrt{\pi}} \left[\sum_{i=1}^{n} \left(\Delta E_{0\to i} - \Delta E_{<} \Theta(E_b - \Delta E_{0\to i}) - \Delta E_{>} \Theta(\Delta E_{0\to i} - E_b) \right) \times \right.$$

$$\left. R_{0\to i} \times \left(e^{-\left(\frac{E-(\Delta E_{0\to i} - \Delta E_{<})}{\Delta}\right)^2} \Theta(E_b - \Delta E_{0\to i}) + e^{-\left(\frac{E-(\Delta E_{0\to i} - \Delta E_{>})}{\Delta}\right)^2} \Theta(\Delta E_{0\to i} - E_b) \right) \right]. \quad (5)$$

where $\Delta E_{<}$ (resp. $\Delta E_{>}$) are the shifts applied to the transition energy when it is higher (resp. lower) in energy than the boarder energy E_b, and Θ denotes the Heaviside step function. These three parameters were optimized concomitantly with the bandwidth by three fitting methods (Levenberg-Marquardt,[50] Conjugate Gradient,[51] Differential Evolution[52]) implemented in the *lmfit* library.[53] The last one gave the best fit. All these post-treatment steps were processed by homemade Python scripts.

IV. B. 3. Results and discussion

IV. B. 3. a) Secondary structure from NMR experiments

As explained in the introduction, the first step consisted in ensuring that the peptide was actually a β-turn in solution. This task was mainly achieved through NMR analyses.

1D, 2D COSY and 2D TOCSY experiments allowed for a clear signal assignment, and also confirmed that very few impurities were present in the medium (see **Figure S1**, **Figure S2** and **Figure S3**). According to 1D spectrum, the proline adopted only one conformation in solution (**Figure S1**). Besides, 2D NOESY experiments (**Figure S4**, **Figure S5** and **Figure S6**) showed highly intense correlations between protons of the Piv residue and the delta hydrogens of the proline, indicating a close proximity between the two groups whereas no correlation was found between the Piv residue and the proline $H_{\alpha,}$ which should appear in a *cis* conformation (**Figure S6**). Consequently, the pivaloyl residue fulfills its anticipated role as a structuring element, which is consistent with the literature and the proline folds into only one conformation which is the *trans* one, that also be determined theortically adterwawrds (Section 3.3 and **Figure S7**)

Furthermore, the NMR analyses provided three major experimental parameters that convey more structural insight. The first one is the temperature coefficient that provides information on the protection of each H_N proton. Let us recall that this coefficient actually measures the variation in the chemical shift with respect to the temperature variation. A high value (typically higher than 4.6 ppb K^{-1} in absolute value in water) is usually linked to an opened hydrogen site, whereas a low value is more compatible with a well-protected site.[54]

In our specific case, two coefficient values are available: that of the amide proton in the serine and that in the NHMe residue. We found that the decrease of the chemical shift for both protons is highly linear (with a coefficient of determination R^2 higher than 0.999), so that the temperature coefficients are well defined in this temperature range (see **Figure S8**). A linear fit afforded the following values in water: -10 ppb K^{-1} for the serine residue and -4 ppb K^{-1} for the NHMe. We can thus conclude that this last hydrogen is well protected (< 4.6 ppb K^{-1} in absolute value[54]). This fact suggests the existence of an intramolecular hydrogen bond between the first and the last residue, in agreement with a β-turn conformation.

Noteworthy, previous results concluded that proton H_N in the serine residue was also well protected in acetonitrile.[19] In this paper, a temperature coefficient for this proton compatible with a protection was found in this organic solvent. Although a direct comparison with the value determined herein in water cannot be directly carried out as the reference threshold value is

highly dependent on the nature of the solvent[55–58], this proton is not protected in water. Thus, water seems to stabilize the conformation, but not as much as CH_3CN. This difference could be ascribed to the fact that, by bonding competitively with the peptide, protic solvents may destabilize the intramolecular hydrogen bonds and may lead to a shift of the conformational equilibrium. This could also explain the decrease in intensity from the ECD spectrum in acetonitrile to the one in the water (see below).

Then, because of the presence of only one proline residue in the *i*+1 position, which restraints the φ_{i+1} around -60°, only three β-turn types out of the 12 ones (**Table IV.1**) are possible for the Piv-Pro-D-Ser-NHMe peptide: β-I, β-II and β-VIII. To go further in the identification, one should scrutinize relevant intramolecular distances. They could be determined using the measured NOE correlations from the ^{1}H 2D NOESY experiments, as shown by eq. 2. Three distances are actually useful to determine the β-turn type: $d_{\alpha N(i+1,\, i+2)}$, $d_{\alpha N(i+1,\, i+3)}$ and $d_{NN(i+2,\, i+3)}$, as depicted in **Figure IV.1**. It can be however noticed that the presence of the corresponding three NOE correlations may constitute by itself a proof of a β-turn folding (see **Figure S4**). For peptides composed of natural amino acids, NOE cross-peaks between the H_α of the first residue and H_N proton of the third residue could also provide an additional proof of a β-turn conformation. The Piv residue does not possess such H_α proton, but very intense NOE cross-peaks between the terbutyl protons and the amide protons in the residues in position i+2 and i+3 are in agreement with a β-turn conformation.

Nevertheless, going further is actually not straightforward. Indeed, we need to compare our experimental distances to those expected in β-I, β-II and β-VIII turn types containing a D-residue at $i+3^{th}$ position. These three references values should also be determined using the same level of theory and based on the De Brevern[16] classification. These combined conditions create a very specific problem. To overcome this difficulty, we calculated these three distances (see the last section in the SI file) by imposing constraints on φ_{i+1}, ψ_{i+1}, φ_{i+2} and ψ_{i+2} dihedral angles in 10 ns molecular dynamics and starting from each of the three β-turn types.

Subsequently, we derived the average values for the $d_{\alpha N(i+1,\, i+2)}$, $d_{\alpha N(i+1,\, i+3)}$ and $d_{NN(i+2,\, i+3)}$ distances from 100 geometries. These results provided a theoretical support (the so-called "canonical" (can.) values), as detailed in **Table IV.2**, where experimental distances are also presented. Among the β-turn types, the β-II conformation most closely matches experimental values for each of these three distances, whereas the two other turn types produced at least one distance significantly different from those expected in a water environment.

This conjecture was further corroborated by inspecting the ^{3}J H_N-H_α spin-spin coupling constant results. Indeed, according to the Karplus formula (eq. 1), the φ dihedral angle between these two protons is linked to the corresponding ^{3}J coupling constant. In the Piv-Pro-D-Ser-NHMe peptide, only one H_N-H_α coupling constant is measurable, which belongs to the serine residue. This value was measured by us to be equal to 7.7 Hz. In a D-residue, the φ and θ dihedral angles obey the following relation:[59]

$$\theta = |\varphi + 60|. \quad (6)$$

The Karplus relationship, now expressed as a function of φ (instead of θ), is represented in **Figure IV.2** for four different parameterizations (see the "Material and Methods" sections for more details). Using canonical values for the φ_{i+2} dihedral angle, the Karplus equation (whatever the set of parameters) does not give a coupling constant above 3 Hz for a β-VIII turn and 5.5 Hz for a β-I turn type; both these turns were thus univocally excluded. Then, from the experimental coupling constant value, two φ_{i+2} values can be obtained for each parameterization, φ_{i+2a} and φ_{i+2b}. They respectively belong to the [82°, 91°] range for φ_{i+2a} and to the [149°, 158°] range for φ_{i+2b}, as represented in **Figure IV.2**, excluding the β-I and β-VIII types, thus confirming the previous β-II assignment. However, it should be noted that this approach is relevant only if one conformation is represented or is predominant in solution. This point has been discussed in the MD section 3. c).

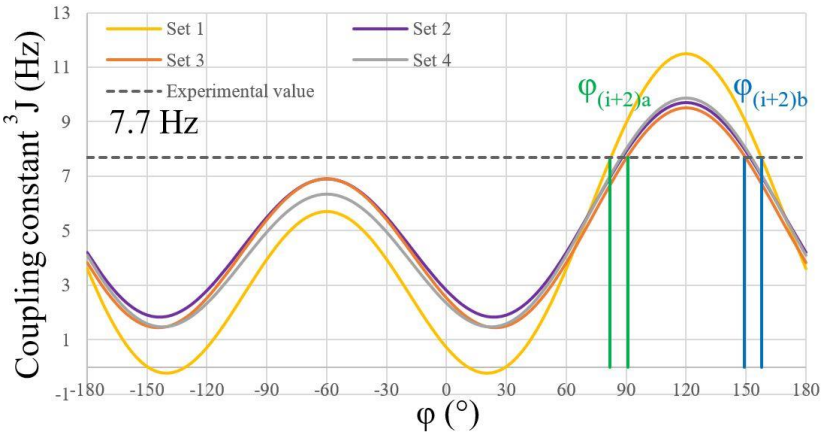


***Figure IV.2**: Plots of Karplus equation with several set of coefficients (set 1: Cung et al. (yellow), set 2: Pardi et al. (purple), set 3: Ludvigsen et al. (orange) and set 4: Vuister et al. (grey)).[30-32] The experimental $^{3}J(H_N\text{-}H_\alpha)$ value in the Ser residue of Piv-Pro-D-Ser-NHMe (7.7 Hz) crosses the Karplus equations at two ranges, [82°, 91°] (green) and [149°, 158°] (blue) corresponding to the two possible inverse images φ_{i+2} dihedral angles*

IV. B. 3. b) ECD experimental spectra

ECD analyses were performed in three common solvents: acetonitrile, water and methanol, and the obtained spectra are represented in **Figure IV.3**. At first glance, the patterns in these three environments are the same, and only the intensity decreases from the spectrum in acetonitrile to the one in methanol. For each spectrum, two major positive bands can be identified between 215 nm and 230 nm, and around 200 nm, this last one being the most intense. Under 190 nm, the spectra were not repeatable, except for water where a third negative band appeared around 185 nm. However, the reproducibility was confirmed in every case, and our data cover a larger wavelength range than the previous study.[19] Although the previous study mentioned a 190-280 nm frequency range, spectra were displayed only down to approximately 195 nm, likely due to reproducibility issues. These limited spectra made the authors[19] to overlook the negative band below this wavelength in water. So we also acquired additional spectra by changing the temperature in water, and no change was noted.

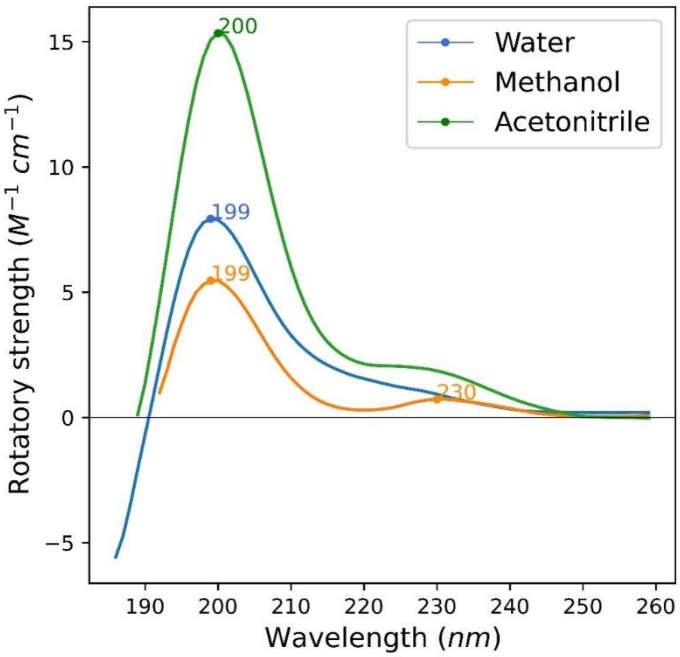


***Figure IV.3**: Experimental ECD spectra of Piv-Pro-D-Ser-NHMe in acetonitrile (green), water (blue) and methanol (orange).*

As expected, the experimental ECD pattern of the Piv-Pro-D-Ser-NHMe is very different from those for helices, sheets, or random conformation. The similarity of the spectra in the three solvents also points out that the conformation adopted by the peptide is stable since even a poor stabilizing solvent such as water can keep the secondary structure. Thus, the conclusion drawn from the NMR analysis is also perfectly supported by the ECD analysis.

IV. B. 3. c) Assessment of the MD protocol and theoretical validation of the peptide structure

We complemented this experimental work by a theoretical study. The first step consisted in the optimization of the computational protocol that was divided into two parts: a MD simulation that will give insight into the structure adopted by the peptide in solution, followed by TD-DFT calculations on selected snapshots to generate a theoretical average ECD spectrum.

The first change compared to our previous protocol[15] was the removing of restraints in the MD simulation that had been imposed so that the model structure remained in a specific β-turn type. Our purpose, here, is to model the real peptide behavior, so that the biomolecule should be free to explore the potential energy surface.

Secondly, we tested two force fields: OPLS-AA, which was used in our previous work, and CHARMM27, which is generally more suitable for proteins and peptides. Ramachandran plots, which represent angle φ against angle ψ for each amino acid, were generated to assess the influence of the force fields on the peptide structure. **Figure IV.4** points out that the serine residue is not as much confined as the proline is, and OPLS-AA datapoints are more scattered, while CHARMM27 conversely generates localized angles.

The optimal choice for the force field will use the information collected from the NMR studies, in particular the interproton distances (see **Table IV.2**). Although results were found quite similar after 10 ns of MD propagation, the measured distances became different by increasing the MD time length. At first glance, all distances were overestimated but the distance $d_{\alpha N(i+1,\ i+3)}$ is notably overestimated with OPLS-AA, with a mean of 4.9 Å against 3.8 Å for the experimental value. This discrepancy may be explained by the distribution of the ψ_{i+2} angle values, as a significant portion measured in the OPLS-AA simulation clustered around -150°, that considerably extends $d_{\alpha N(i+1,i+3)}$. This hypothesis was confirmed by two observations. Firstly, the distribution of $d_{\alpha N(i+1,\ i+3)}$ (see panel A in **Figure S9**) revealed two distinct populations: one around $d_{\alpha N(i+1,\ i+3)}$ = 6.2 Å and one other centered around 3.5 Å.

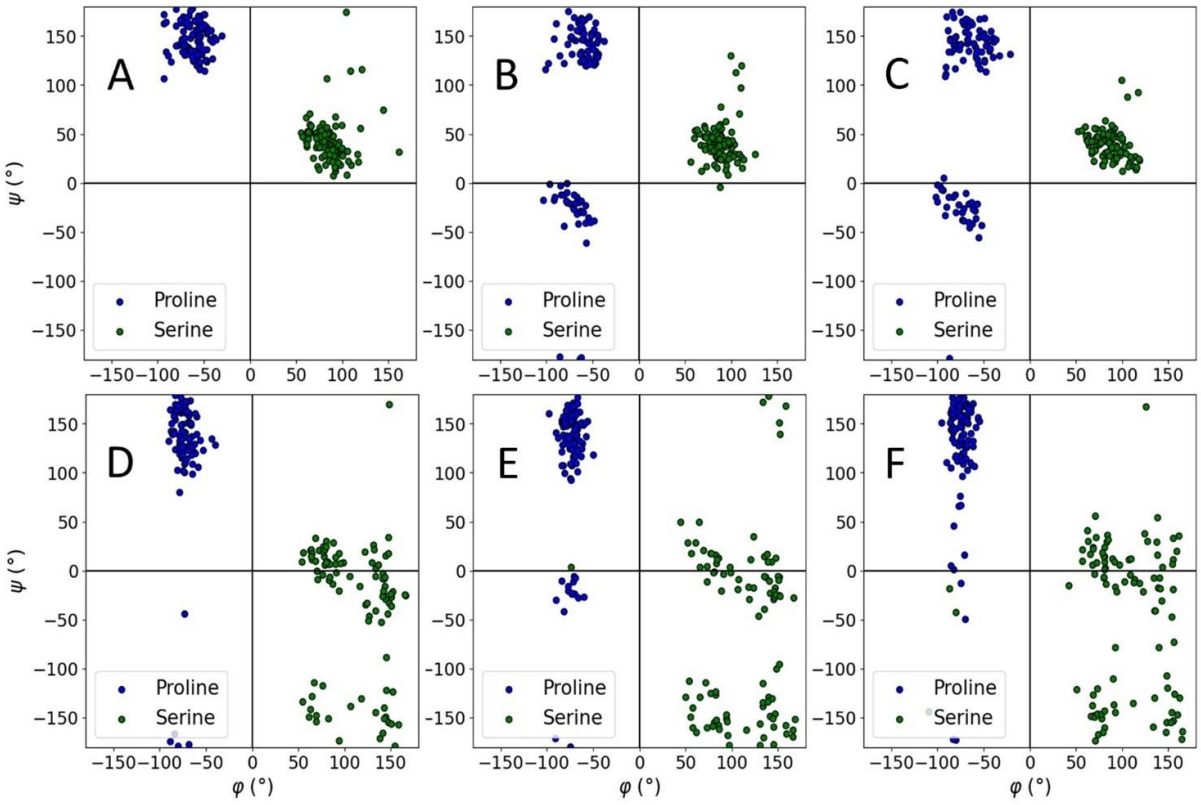


***Figure IV.4**: Ramachandran plots of φ and ψ dihedral angles (in °) in the serine and proline residues at 300 K with CHARMM27 (panels A, B and C) or OPLS-AA (D, E and F) force fields. Geometries were obtained by 10 ns (A and D), 100 ns (B and E), or 1000 ns (C and F) MD simulations within the TIP3P water model*

.

Secondly, the same separation was notably evident in the Ramachandran plot (see panel B in **Figure S9**) where all shorter distances (< 4.5 Å) correlated with ψ_{i+2} dihedral angles ranging between -78 and 56°. For these reasons, CHARMM27 was definitely chosen for the subsequent MD simulations. The MD time length was then set to 1000 ns to secure the convergence of both the dihedral angles and the distances.

Force field	**OPLS-AA**		
Distance	$d_{\alpha N(i+1, i+2)}$ (Å)	$d_{\alpha N(i+1, i+3)}$ (Å)	$d_{NN(i+2, i+3)}$ (Å)
10 ns	2.3 (0.2)	4.2 (1.2)	2.9 (1.0)
100 ns	2.4 (0.5)	4.9 (1.4)	2.4 (1.1)
1000 ns	2.4 (0.3)	4.6 (1.2)	3.1 (1.0)
Force field	**CHARMM27**		
Distance	$d_{\alpha N(i+1, i+2)}$ (Å)	$d_{\alpha N(i+1, i+3)}$ (Å)	$d_{NN(i+2, i+3)}$ (Å)
10 ns	2.2 (0.2)	3.9 (0.5)	2.6 (0.4)
100 ns	2.7 (0.6)	4.3 (0.9)	2.5 (0.3)
1000 ns	2.7 (0.6)	4.2 (0.8)	2.5 (0.3)
	Exp.		
Distance	$d_{\alpha N(i+1, i+2)}$ (Å)	$d_{\alpha N(i+1, i+3)}$ (Å)	$d_{NN(i+2, i+3)}$ (Å)
	2.0 (0.2)	3.8 (0.4)	2.5 (0.3)
	Can. β-I		
Distance	$d_{\alpha N(i+1, i+2)}$ (Å)	$d_{\alpha N(i+1, i+3)}$ (Å)	$d_{NN(i+2, i+3)}$ (Å)
	3.5 (0.1)	3.8 (0.2)	2.6 (0.2)
	Can. β-II		
Distance	$d_{\alpha N(i+1, i+2)}$ (Å)	$d_{\alpha N(i+1, i+3)}$ (Å)	$d_{NN(i+2, i+3)}$ (Å)
	2.2 (0.1)	3.5 (0.2)	2.6 (0.2)
	Can. β-VIII		
Distance	$d_{\alpha N(i+1, i+2)}$ (Å)	$d_{\alpha N(i+1, i+3)}$ (Å)	$d_{NN(i+2, i+3)}$ (Å)
	3.5 (0.1)	5.9 (0.3)	4.2 (0.1)

Table IV.2*: Average distances over 100 geometries generated during MD simulations of different time lengths (10, 100 and 1000 ns) using two different force fields (CHARMM27 and OPLS-AA), experimental ones (from our own NMR analyses (Figures S4 and S5)) for the Piv-Pro-D-Ser-NHMe peptide in water, and theoretical canonical ones according to the β-turn type. The number in parenthesis corresponds to the considered error: 10% for each experimental distance and the standard deviation for values form MD simulations*

The water model was then discussed in a similar way (**Figure IV.5**, **Table IV.3**) by testing TIP3P, TIP4P, TIP5P and TIPS3P. The results are quite similar for all water models, except for TIP3P. Hence, we have discarded this model due to the significantly larger $d_{\alpha N(i+1,\ i+2)}$ (2.7 Å instead of the experimental value of 2.0 Å) as a consequence of the prevalence of proline population at ψ_{i+1} around -20°. Such angles resulted in an increase of $d_{\alpha N(i+1,\ i+2)}$, deviating further from the experimental distance. This assertion was supported by **Figure S10**, demonstrating that all geometries with small absolute ψ_{i+1} values were associated with longer $d_{\alpha N(i+1,\ i+2)}$ distances. To go further in the selection, we have investigated the behavior of the peptide during the MD simulation. By only looking at central amino acids involved in the β-turn, we obviously missed what occurs at its ends. Thus, we also calculated the ω dihedral angle that defines the *trans* or *cis* conformation of the proline residue (represented in **Figure S6**) in each water model to assess their influence on this segment of the peptide (**Figure S11**).

	CHARMM27		
Distance	$d_{\alpha N(i+1,\ i+2)}$ (Å)	$d_{\alpha N(i+1,\ i+3)}$ (Å)	$d_{NN(i+2,\ i+3)}$ (Å)
Experimental	2.0 (0.2)	3.8 (0.4)	2.5 (0.3)
TIP3P	2.7 (0.6)	4.2 (0.8)	2.5 (0.3)
TIP4P	2.4 (0.4)	4.1 (0.7)	2.6 (0.3)
TIP5P	2.5 (0.5)	4.1 (0.8)	2.6 (0.3)
TIPS3P	2.5 (0.5)	4.2 (0.8)	2.6 (0.4)

Table IV.3*: Average distances over 100 geometries generated during 1000 ns MD simulations with the CHARMM27 force field and several water models (TIP3P, TIP4P, TIP5P and TIPS3P), and the experimental ones (from our own NMR analyses) for the Piv-Pro-D-Ser-NHMe peptide in water. The number in parenthesis corresponds to the considered error: 10% for each experimental distance and the standard deviation for values form MD simulations.*

The values of the ω angle near to -180° or 180°, as seen in the TIP3P, TIP4P and TIP5P water models, correspond to a *trans* conformation. But an additional conformation of the proline residue, the *cis* one, emerged in the geometries generated using the TIPS3P model. Specifically, in this simulation, the proline maintained the *cis* conformation with ω angle values around 0° during the last 200 ns. However, as explained in section 3. a), our NMR analyses have excluded this. Besides, a scan performed on the dihedral angle ω_i indicate a high barrier energy of 17.71 kcal mol^{-1} for such rotation to reach the *cis* conformation from the *trans* one (**Figure S7**). Therefore, a simulation providing more than 20% of a cis conformation cannot reasonably be

kept and only the TIP4P and TIP5P water models could be safely used. Our final choice was to keep the most recent and sophisticated one, namely TIP5P.

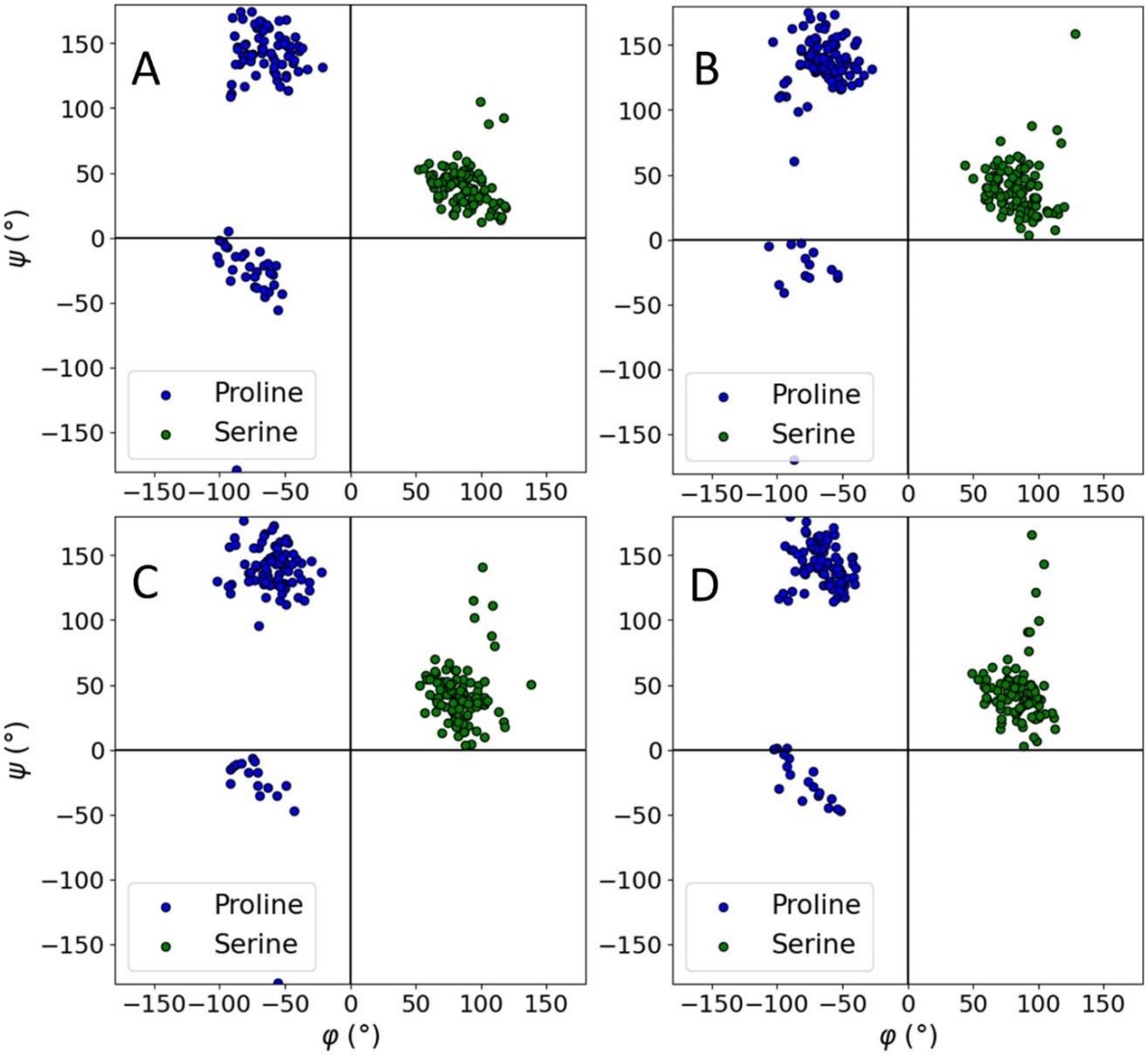


Figure IV.5*: Ramachandran plots of φ and ψ dihedral angles in the serine and proline residues with the CHARMM27 force field and four water models (TIP3P (A), TIP4P (B), TIP5P (C), TIPS3P (D)). Geometries were obtained by 1000 ns MD simulation at 300 K.*

Populations generated by our optimized MD simulation (CHARMM27-TIP5P, 1000 ns) were collected from 1000 snapshots (instead of 100 previously) and displayed on a heatmap of the Ramachandran plot (**Figure IV.6**). Only two β-turn types were close to these geometries: the β-II and β-IV types, the first one being the most populated one. Moreover, MD simulations confirm that there is only one strongly dominant structure in solution, which justifies the use of the Karplus equation to deduce the dominant structure in solution. The average distances of each population were summarized in Table S2 in which the influence of the minor population on the total was considered negligible. These theoretical geometries are actually fully consistent

with our previous experimental conclusions that suggested a major β-II conformation with only a *trans* conformation for the proline residue. This agreement made us particularly confident regarding the geometries used for our subsequent TD-DFT calculations.

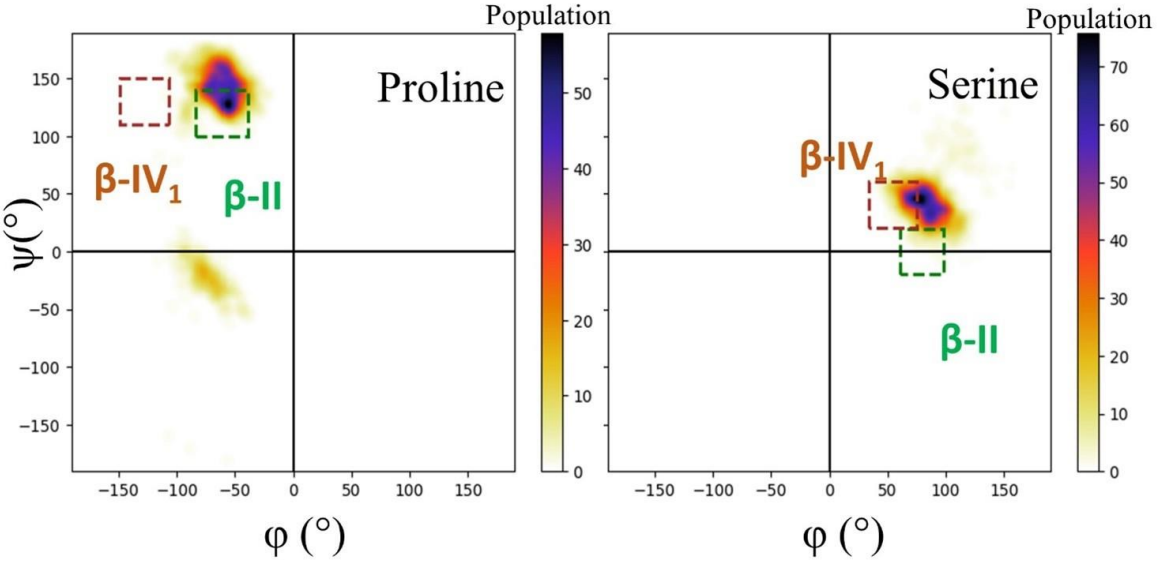


Figure IV.6*: Ramachandran plots of φ and ψ dihedral angles in proline (left) and serine (right) residues. Geometries were obtained by a 1000 ns MD simulation using the CHARMM27 force field and the TIP5P water model at 300 K.*

IV. B. 3. d) Theoretical simulation of ECD spectra

The final predicted ECD spectrum was calculated by averaging the spectra on all extracted snapshots (**Figure IV.7A**). Pleasantly, it exhibited the same pattern as the experimental one. Three major bands can be identified as on the experimental spectrum, with also comparable global shape and intensities. It can be also noticed that the transitions are slightly blue-shifted. The fact that range separated functionals (such as CAM-B3LYP) could overestimate the transition energies were already reported.[13] However, not all bands in the spectra seem to be shifted in the same way.

Accordingly, we proposed to correct the theoretical curves using a four-parameter model (see eq. 5 in the computational details section). The optimization of their values was performed with and without normalization of the signal, in order to provide both a comparison in intensity and sign, and a comparison of the global pattern. Indeed, through normalization (defined by $X_{Norm} = \frac{X - X_{min}}{X_{max} - X_{min}}$), the motif is described more precisely but at the cost of the absolute intensity. Panels B and C in **Figure IV.7** show the obtained results with the following optimized parameters: $\Delta = 0.29\ eV, \Delta E_{<} = 0.36\ eV, \Delta E_{>} = -0.25\ eV, E_b = 186.94\ nm$(without normalization), $\Delta = 0.23\ eV, \Delta E_{<} = 0.36\ eV, \Delta E_{>} = -0.14\ eV, E_b = 191.42\ nm$ (with

normalization). In both cases, the corrected spectrum fitted much more to the experimental one, providing considerable improvement for the ECD spectrum prediction.

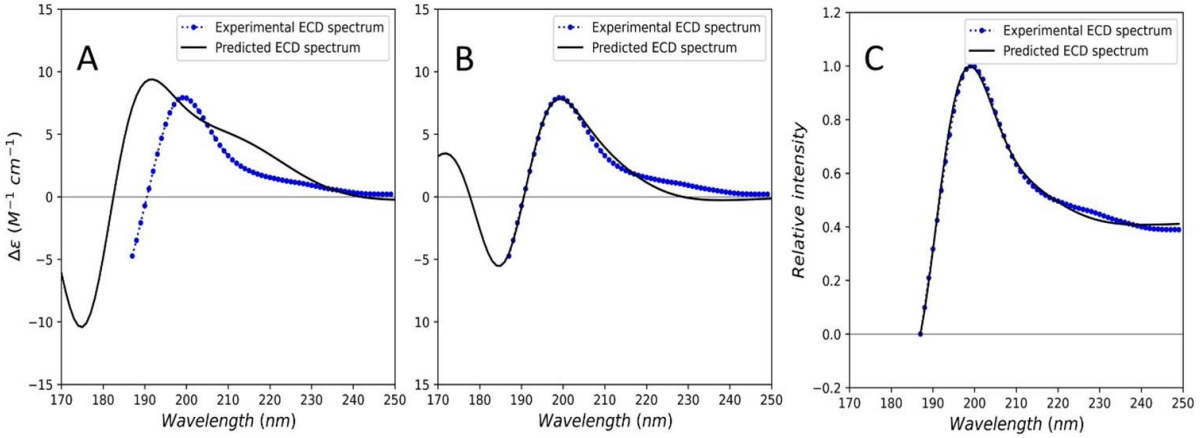


Figure IV.7*: Comparison of the predicted ECD spectrum (solid black line) of the Piv-Pro-D-Ser-NHMe peptide with the experimental one (blue points) without any fitting procedure (A), with optimized parameters without normalization (B), and with optimized parameters for normalized data (C). Level of theory: TD-DFT CAM-B3LYP/6-31+G* with polarizable embedding.*

In conclusion, the Piv-Pro-D-Ser-NHMe peptide, which is now known to fold mainly close to a β-II turn in water, exhibits an ECD spectrum as follows: a negative band around 185 nm before two positive bands at 200 nm and between 215 and 230 nm. Accordingly, the present protocol faithfully reproduced this pattern, aligning well with both the ECD prediction of a β-II turn as proposed by Migliore et al.[15] and by Sathyanarayana and Applequist[11].

IV. B. 3. e) Validation on second peptide

Finally, to check the versatility of our theoretical methodology, we tested it on another small peptide, namely the cationic form of Gly-Ala-Gly (GAG^+). In this way, we ensure that the protocol can be applied to other very short peptides. The choice of this peptide GAG^+, which was already studied experimentally and theoretically, was also driven to compare our theoretical protocol to other methods.[20-23] In particular, the recent theoretical protocol proposed by Monti *et al.*[23] differed from ours by the following aspects:

(i) MD were performed using the OPLS-AA force field, which was proved as unsuitable for the previous peptide;

(ii) The methodology involved Essential Dynamics (ED) steps before the QC calculations, extracting peptide conformers with their own weight;

(iii) The final ECD spectrum was generated by TD-DFT using the ωB97X-D exchange-correlation functional;

(iv) The solvent was considered at various levels (**Figure IV.8**). After one (**Figure IV.8**, red dotted line) or two ED simulations (**Figure IV.8**, yellow dotted line) and explicitly in TD-DFT calculations (**Figure IV.8**, green dotted line).

Our MD results were depicted in **Figure S12**, in which the polyporoline-II conformation is the major population in water. This result is consistent with the experimental and theoretical finding. [21-23] The application of our computational protocol generated the predicted ECD spectrum represented in panel A in **Figure IV.8**, which does not include any fitting procedure, and which can be compared with the experimental and the theoretical ones reported in ref. 23. Though both simulations have generated the same pattern, that is a large negative band around 190 nm with a positive one after 205 nm, our protocol afforded a spectrum closer to the experimental and also consistent with a dominant polyproline II population in solution. We then used the fitting parameters determined for Piv-Pro-D-Ser-NHMe without further refinement (see **Figure IV.8B** and **Figure IV.8C**). A good agreement with the experiment was reached, hinting that the efficiency of the original fitting parameters can be safely applied to other peptides.

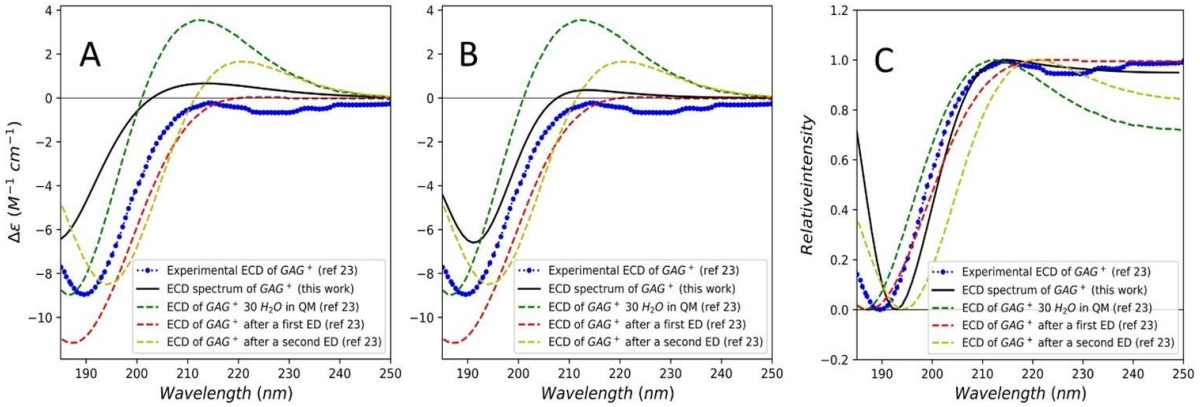


Figure IV.8*: Comparison of the predicted ECD (solid black line) spectrum and the experimental one (blue points) of the Gly-Ala-Gly+ cation without any fitting procedure (A), with optimized parameters without normalization (B), and with optimized parameters for normalized data (C). Level of theory for our computations: TD-DFT CAM-B3LYP/6-31+G* with polarizable embedding. Green, yellow and red dotted lines have been plotted using the data frame ref 23. (extracted by the Webplotdigitizer tool). Each of these last three lines was generated by the various theoretical protocol detailed in ref. 23.*

IV. B. 3. f) Comparison with a static approach

In this last part, the efficiency of a pure static approach will be discussed. The calculation of the ECD spectrum directly on a single optimized structure could also be possible, thus we have simulated the ECD spectrum of both, the X-Ray structure, and the optimized structure which we have compared to our MD + TD-DFT results (**Figure S13**). The PE model is not compatible with a static approach since the solvent molecules would not be optimized, thus we have chosen

to describe the environment implicitly with a continuum model, namely the C-PCM. The Mean Absolute Error was also calculated for each method (See supporting Information). The simulated spectra were very far from the experimental one (Pannel A and B, **Figure S13**) with a MAE of 3.25 M^{-1} cm^{-1} for the TD-DFT on the RX structure approach, and of 2.59 M^{-1} cm^{-1} for the TD-DFT on the optimized structure approach, whereas the MAE have decreased to 1.94 M^{-1} cm^{-1} for our MD + TD-DFT protocol (Pannel C, **Figure S13**). These results suggest that information is lost by considering only one structure in solution even in the presence of a strong dominant structure in solution.

Moreover, we have calculated optimized shifts (**Table S3**) to check whether an improvement of the MAE is still possible with this static method. The application of these optimized shifts was shown **Figure S13** (Pannel D, E and F). These shifts are rather large for the two static approaches and the MAE on the TD-DFT on the RX structure was only 1.98 M^{-1} cm^{-1}. The MAE has reached a convincing value of 0.51 M^{-1} cm^{-1} for the approach on the optimized structure whereas the MAE of our approach with optimized shifts has reached 0.46 M^{-1} cm^{-1}. However, it should not be forgotten that the initial ECD spectra (panels A and B **Figure S13**) are almost opposite to the experimental spectrum suggesting that our approach is the most suitable. Besides, if more than one conformation is possible in solution, the static approach should also be avoided.

IV. B. 4. Conclusion

In this paper, we introduced a new theoretical model capable of predicting both the conformation in solution and the ECD spectra of small peptides. The optimization process involved MD simulations, the force field, water model and dynamics parameters being chosen from comparison with NMR analyses. Each of these features significantly influenced the resulting geometries, thereby impacting the subsequent ECD spectrum.

For the Piv-Pro-D-Ser-NHMe, the MD results were much closer to the NMR analyses by applying the CHARMM27 force field coupled with the TIP5P water model and a dynamic time length of 1000 ns. These geometries were employed in quantum calculations of the ECD spectrum, which was compared to the experimental one. A notable agreement between the two spectra was observed and because of the proximity between the generated geometries and the NMR distances, we can affirm that the ECD pattern is the one of a β-II turn secondary structure. However, the transition energies were blue-shifted.

Thus, a modification of the convolution function was carried out, which considerably improved the accuracy of the predicted spectrum. Our optimized theoretical protocol was also performed on the GAG^+ peptide. The resulting ECD spectrum more closely resembled experimental data compared to recent attempts on this peptide, demonstrating the broader applicability of our approach.

Acknowledgments

This work has been partially supported by University of Rouen Normandy, INSA Rouen Normandy, the "Centre National de la Recherche Scientifique" (CNRS), the European Regional Development Fund (ERDF), Labex SynOrg (ANR-11-LABX-0029), Carnot Institut I2C, the graduate school for research XL-Chem (ANR-18-EURE-0020 XL CHEM), and the "Région Normandie". The Centre Régional Informatique et d'Applications Numériques de Normandie (CRIANN) is acknowledged for providing access to computational resources.

Supporting information

Experimental ECD, NMR spectra, and additional computational details.

IV. C. Conclusions et perspectives

Cette étude a permis de mettre en place un nouveau protocole théorique pour prédire le spectre DCE du peptide le Piv-Pro-D-Ser-NHMe structuré en solution selon un coude de type β-II. Ce protocole a également été testé sur le peptide GAG^+ chargé positivement avec un très bon accord avec le spectre expérimental dans les deux cas.

L'étude RMN a permis de vérifier la conservation de la structure du peptide en solution avec une structure de coude de type β-II majoritaire, ce qui montre que le spectre DCE mesuré est celui d'un coude de type β-II. La RMN a également servi de support pour optimiser le protocole théorique. Ce dernier comprend une Dynamique Moléculaire de 1 µs, effectuée avec le champ de force CHARMM27 et le modèle d'eau TIP5P. Sans contrainte sur les angles, les simulations de 10 ns ne sont pas suffisantes pour que la molécule explore l'ensemble des conformations. Des dynamiques suffisamment longues doivent être effectuées. L'utilisation du champ de force OPLS-AA génère des géométries trop éloignées de la réalité et bien plus éparpillées, tandis que certains modèles d'eau comme le modèle TIPS3P génère des géométries peu probables en solution.

La production du spectre DCE théorique est très satisfaisante uniquement par la moyenne des spectres convolués à l'aide de fonctions Gaussiennes. Pour une parfaite adéquation avec le spectre expérimental, nous avons eu recourt à l'utilisation de deux shifts lors de la convolution. Bien que les mêmes « *shifts* » aient aussi été utilisés sur le second peptide GAG^+, il est peu probable que ces « *shifts* » soient généralisables car ils sont certainement adaptés pour corriger certains types de transitions électroniques.

Concernant nos objectifs à plus long terme, il semble difficile à ce stade d'imaginer une prédiction des spectres purement théoriques pour tous les types de coude. En effet, les analyses expérimentales RMN et DCE permettent l'ajustement du protocole théorique et sont encore nécessaires pour être certain que le spectre simulé correspond au spectre mesuré. Pour franchir cette barrière, il faudrait appliquer notre démarche à d'autres peptides. Certains paramètres comme le champ de force doivent être testé de nouveau car ils exercent une grande influence sur les géométries générées et le spectre simulé qui en découle et il n'existe pas non plus de moyen pour prédire quel champ de force utilisé pour chaque molécule à tester. De plus les shifts d'énergies pourraient être améliorés soit en étant optimisés pour faire correspondre le spectre théorique avec le spectre expérimental de molécules différentes, soit en étudiant plus précisément quels types de transition électronique doivent être corrigés. Des études similaires

pourraient être faites pour augmenter le nombre de peptides testés. Par exemple, les peptides Z-Aib-Pro-Aib-Pro-OMe, Piv-Pro-Aib-NHMe, Piv-Pro-D-Ala-NHMe et Piv-Pro-Val-NHMe, se replient sous la forme d'un coude β en phase cristalline (β-I, β-II, β-II et β-I respectivement) et possèdent des motifs DCE particuliers.[60] Ces peptides sont de très bons candidats dans la démarche d'établir des spectres de référence pour le spectre DCE.

Un autre point à considérer par rapport à nos objectifs est que les petits peptides n'ont souvent qu'une faible proportion de structures stables dans l'eau. De ce fait, il est très commun d'utiliser des solvants organiques pour stabiliser la structure majoritaire et les spectres DCE ou RMN sont souvent enregistrés dans ces solvants ou dans un mélange eau/solvant organique. Des études menées dans ces types de solvants pourraient être faites pour élargir l'applicabilité du protocole que nous avons développé. En effet, pour les autres éléments de structures secondaires principaux (hélice et feuillet), il n'a pas été mis en évidence d'impact significatif du solvant sur l'allure du spectre DCE, mais on ne peut pas exclure ce type d'impact pour des éléments non répétitifs et courts comme les coudes. Ainsi, nous avons mesuré des spectres DCE différents pour le peptide Piv-Pro-D-Ser-NHMe dans l'acétonitrile et le méthanol. La différence d'intensité entre ces spectres, si elle était interprétée de manière habituelle, suggèrerait que la proportion de structuration sous la forme d'un coude β-II est plus importante dans l'acétonitrile et donc que les conditions de solvatation influenceraient la structuration du peptide. Mais cette différence pourrait aussi être due à l'impact du solvant sur les extrémités du motif coude, pour lesquelles la géométrie n'est pas fixée ou à l'effet d'un environnement différent sur la molécule. Nous avons fait plusieurs observations allant dans le sens de cette dernière hypothèse lors de l'établissement de notre protocole.

Notamment, nous avons constaté qu'au sein d'une même population de coudes de type β-II, bien que la moyenne des spectres théoriques sur cette population génère un spectre en accord avec l'expérience, les spectres individuels présentent des variations significatives. Certaines géométries bien que très similaires, aussi bien au niveau des angles dièdres centraux qu'au niveau de la chaîne peptidique entière, génèrent des spectre DCE totalement opposés. Ceci suggère que d'une part, les angles dièdres centraux ne sont plus les seuls paramètres gouvernant le spectre DCE simulé pour cette taille de peptide, et que d'autre part, l'environnement exerce une influence significative sur le spectre DCE, même sans provoquer de modifications de la géométrie. Il pourrait donc être intéressant de simuler les conformations et le spectre DCE dans d'autres conditions expérimentales dans le but d'éclaircir ce point.

Pour la suite des travaux, nous avons choisi de nous consacrer à un autre aspect. Nous voulons maintenant nous intéresser à des peptides flexibles qui adoptent plusieurs conformations en solution, et voir s'il est possible, en s'appuyant sur une stratégie similaire à celle développée dans ce chapitre, d'améliorer la prédiction des conformations des petits peptides en solution.

Pour développer le premier protocole, nous avons dû, comme expliqué en introduction de ce chapitre, choisir un peptide possédant tous les atouts nécessaires pour rester structuré en solution. Ce peptide, bien que réel, a été élaboré dans le but de stabiliser une structuration de type coude et ne possède donc pas d'activité biologique. Nous souhaitons désormais étendre ce protocole à des peptides bioactifs, dont la détermination des conformations en solution reste un réel défi. En effet, les peptides bioactifs naturels ont souvent des courtes séquences et sont caractérisés par une grande flexibilité conformationnelle. Il est de ce fait très difficile d'étudier leur conformation en solution, car ils existent presque invariablement sous la forme d'un mélange complexe de nombreux conformères.

IV. D. References

1. Migliore, M. *et al.* Characterization of β-turns by electronic circular dichroism spectroscopy: a coupled molecular dynamics and time-dependent density functional theory computational study. *Phys. Chem. Chem. Phys.* **22**, 1611–1623 (2020).
2. Aubry, A., Ghermani, N. & Marraud, M. Backbone side chain interactions in peptides. *Int. J. Pept. Protein Res.* **23**, 113–122 (1984).
3. Marraud, M. & Aubry, A. Backbone side chain interactions in peptides. *Int. J. Pept. Protein Res.* **23**, 123–133 (1984).
4. Johnson, W. C. Protein secondary structure and circular dichroism: A practical guide. *Proteins Struct. Funct. Genet.* **7**, 205–214 (1990).
5. Provencher, S. W. & Gloeckner, J. Estimation of globular protein secondary structure from circular dichroism. *Biochemistry* **20**, 33–37 (1981).
6. Easson, L. H. & Stedman, E. Studies on the relationship between chemical constitution and physiological action. *Biochem. J.* **27**, 1257–1266 (1933).
7. Videau, L. L., Arendall III, W. B. & Richardson, J. S. The cis-pro touch-turn: A rare motif preferred at functional sites. *Proteins Struct. Funct. Bioinforma.* **56**, 298–309 (2004).
8. Wang, J. *et al.* StraPep: a structure database of bioactive peptides. *Database* **2018**, bay038 (2018).
9. Chou, K.-C. Prediction of Tight Turns and Their Types in Proteins. *Anal. Biochem.* **286**, 1–16 (2000).
10. Wilmot, C. M. & Thornton, J. M. Analysis and Prediction of the Different Types of B-Turn in Proteins. *J. Mol. Biol.* **203**, 221-232 (1988)
11. Venkatachalam, C. M. Stereochemical criteria for polypeptides and proteins. V. Conformation of a system of three linked peptide units. *Biopolymers* **6**, 1425–1436 (1968).

12. Woody, R. W. Studies of theoretical circular dichroism of polypeptides: Contributions of β-turns. *Pept. Polypept. Proteins* 338–350 (1974).
13. Bandekar, J. *et al.* Conformations of cyclo(L-alanyl-L-alanyl-ε-aminocaproyl) and of cyclo(L-alany1-D-alanyl-ε-aminocaproyl); cyclized dipeptide models for specific types of β-bends. *Int. J. Pept. Protein Res.* **19**, 187–205 (1982).
14. Sathyanarayana, B. k. & Applequist, J. Theoretical π-π* absorption and circular dichroic spectra of β-turn model peptides. *Int. J. Pept. Protein Res.* **27**, 86–94 (1986).
15. Kaminský, J., Kubelka, J. & Bouř, P. Theoretical Modeling of Peptide α-Helical Circular Dichroism in Aqueous Solution. *J. Phys. Chem. A* **115**, 1734–1742 (2011).
16. Gattuso, H., García-Iriepa, C., Sampedro, D., Monari, A. & Marazzi, M. Simulating the Electronic Circular Dichroism Spectra of Photoreversible Peptide Conformations. *J. Chem. Theory Comput.* **13**, 3290–3296 (2017).
17. Brkljača, Z., Mališ, M., Smith, D. M. & Smith, A.-S. Calculating CD Spectra of Flexible Peptides: An Assessment of TD-DFT Functionals. *J. Chem. Theory Comput.* **10**, 3270–3279 (2014).
18. De Brevern & G, A. Extension of the classical classification of β-turns. *Sci. Rep.* **6**, 33191 (2016).
19. Olsen, J. M., Aidas, K. & Kongsted, J. Excited States in Solution through Polarizable Embedding. *J. Chem. Theory Comput.* **6**, 3721–3734 (2010).
20. Toal, S. E., Verbaro, D. J. & Schweitzer-Stenner, R. Role of Enthalpy–Entropy Compensation Interactions in Determining the Conformational Propensities of Amino Acid Residues in Unfolded Peptides. *J. Phys. Chem. B* **118**, 1309–1318 (2014).
21. Kumar, A., Schweitzer-Stenner, R. & Wong, B. M. A new interpretation of the structure and solvent dependence of the far UV circular dichroism spectrum of short oligopeptides. *Chem. Commun.* **55**, 5701–5704 (2019).

22. Kumar, A., Toal, S. E., DiGuiseppi, D., Schweitzer-Stenner, R. & Wong, B. M. Water-Mediated Electronic Structure of Oligopeptides Probed by Their UV Circular Dichroism, Absorption Spectra, and Time-Dependent DFT Calculations. *J. Phys. Chem. B* **124**, 2579–2590 (2020).
23. Monti, M., Stener, M. & Aschi, M. A computational approach for modeling electronic circular dichroism of solvated chromophores. *J. Comput. Chem.* **43**, 2023–2036 (2022).
24. Thrippleton, M. J. & Keeler, J. Elimination of Zero-Quantum Interference in Two-Dimensional NMR Spectra. *Angew. Chem. Int. Ed.* **42**, 3938–3941 (2003).
25. Shaka, A. J., Lee, J. & Pines, A. Iterative Schemesfor Bibear Operators; Application to Spin Decoupling. *J. Magn. Reson. 1969* **77**, 274-293 (1988)
26. Marion, D., Ikura, M., Tschudin, R. & Bax, A. Rapid recording of 2D NMR spectra without phase cycling. Application to the study of hydrogen exchange in proteins. *J. Magn. Reson. 1969* **85**, 393–399 (1989).
27. Hwang, T. L. & Shaka, A. J. Water Suppression That Works. Excitation Sculpting Using Arbitrary Wave-Forms and Pulsed-Field Gradients. *J. Magn. Reson. A* **112**, 275–279 (1995).
28. Markley, J. L. *et al.* Recommendations for the presentation of NMR structures of proteins and nucleic acids. IUPAC-IUBMB-IUPAB Inter-Union Task Group on the Standardization of Data Bases of Protein and Nucleic Acid Structures Determined by NMR Spectroscopy. *J. Biomol. NMR* **12**, 1–23 (1998).
29. Cung, M. T., Marraud, M. & Neel, J. Experimental Calibration of a Karplus Relationship in Order to Study the Conformations of Peptides by Nuclear Magnetic Resonance. *Macromolecules* **7**, 606–613 (1974).

30. Pardi, A., Billeter, M. & Wüthrich, K. Calibration of the angular dependence of the amide proton-Cα proton coupling constants, 3JHNα, in a globular protein: Use of 3JHNα for identification of helical secondary structure. *J. Mol. Biol.* **180**, 741–751 (1984).
31. Ludvigsen, S., Andersen, K. V. & Poulsen, F. M. Accurate measurements of coupling constants from two-dimensional nuclear magnetic resonance spectra of proteins and determination of φ-angles. *J. Mol. Biol.* **217**, 731–736 (1991).
32. Vuister, G. W. & Bax, A. Quantitative J correlation: a new approach for measuring homonuclear three-bond J(HNH.alpha.) coupling constants in 15N-enriched proteins. *J. Am. Chem. Soc.* **115**, 7772–7777 (1993).
33. Wüthrich, K. NMR with Proteins and Nucleic Acids. *Europhys. News* **17**, 11–13 (1986).
34. Boros, S., Gáspári, Z. & Batta, G. Chapter One - Accurate NMR Determinations of Proton–Proton Distances. in *Annual Reports on NMR Spectroscopy* (ed. Webb, G. A.) vol. 94 1–39 (Academic Press, 2018).
35. CCDC contains the supplementary crystallographic data for this paper. These data can be obtained free of charge from The Cambridge Crystallographic Data Centre via www.ccdc.cam.ac.uk/structures (accessed May 30, 2024).
36. Pettersen, E. F. *et al.* UCSF Chimera—A visualization system for exploratory research and analysis. *J. Comput. Chem.* **25**, 1605–1612 (2004).
37. Hess, B., Kutzner, C., van der Spoel, D. & Lindahl, E. GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation. *J. Chem. Theory Comput.* **4**, 435–447 (2008).
38. Kaminski, G. A., Friesner, R. A., Tirado-Rives, J. & Jorgensen, W. L. Evaluation and Reparametrization of the OPLS-AA Force Field for Proteins via Comparison with Accurate Quantum Chemical Calculations on Peptides. *J. Phys. Chem. B* **105**, 6474–6487 (2001).

39. MacKerell Jr., A. D., Banavali, N. & Foloppe, N. Development and current status of the CHARMM force field for nucleic acids. *Biopolymers* **56**, 257–265 (2000).
40. 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.* **79**, 926–935 (1983).
41. Jorgensen, W. L. & Madura, J. D. Temperature and size dependence for Monte Carlo simulations of TIP4P water. *Mol. Phys.* **56**, 1381–1392 (1985).
42. Mahoney, M. W. & Jorgensen, W. L. A five-site model for liquid water and the reproduction of the density anomaly by rigid, nonpolarizable potential functions. *J. Chem. Phys.* **112**, 8910–8922 (2000).
43. Mark, P. & Nilsson, L. Structure and Dynamics of the TIP3P, SPC, and SPC/E Water Models at 298 K. *J. Phys. Chem. A* **105**, 9954–9960 (2001).
44. Bussi, G., Donadio, D. & Parrinello, M. Canonical sampling through velocity rescaling. *J. Chem. Phys.* **126**, 014101 (2007).
45. Parrinello, M. & Rahman, A. Polymorphic transitions in single crystals: A new molecular dynamics method. *J. Appl. Phys.* **52**, 7182–7190 (1981).
46. Aidas, K. *et al.* The Dalton quantum chemistry program system. *WIREs Comput. Mol. Sci.* **4**, 269–284 (2014).
47. Yanai, T., Tew, D. P. & Handy, N. C. A new hybrid exchange–correlation functional using the Coulomb-attenuating method (CAM-B3LYP). *Chem. Phys. Lett.* **393**, 51–57 (2004).
48. Olsen, J. M. H. & Reinholdt, P. PyFraME: Python framework for Fragment-based Multiscale Embedding. Zenodo https://doi.org/10.5281/ZENODO.4899311 (2021).
49. Stephens, P. J. & Harada, N. ECD cotton effect approximated by the Gaussian curve and other methods. *Chirality* **22**, 229–233 (2010).

50. Moré, J. J. The Levenberg-Marquardt algorithm: Implementation and theory. in *Numerical Analysis* (ed. Watson, G. A.) vol. 630 105–116 (Springer Berlin Heidelberg, Berlin, Heidelberg, 1978).
51. Polak, E. & Ribiere, G. Revue française d'informatique et de recherche opérationnelle. *Sér. Rouge* **3**, 35–43 (1969).
52. Storn, R. & Price, K. Differential Evolution – A Simple and Efficient Heuristic for Global Optimization over Continuous Spaces. *J. Glob. Optim.* **11**, 341–359 (1997).
53. Newville, M., Stensitzki, T., Allen, D. B. & Ingargiola, A. LMFIT: Non-Linear Least-Square Minimization and Curve-Fitting for Python, 2014
54. Cierpicki, T. & Otlewski, J. Amide proton temperature coefficients as hydrogen bond indicators in proteins. *J. Biomol. NMR* **21**, 249–261 (2001).
55. Contreras, M. À., Haack, T., Royo, M., Giralt, E. & Pons, M. Temperature coefficients of peptides dissolved in hexafluoroisopropanol monitor distortions of helices. *Lett. Pept. Sci.* **4**, 29–39 (1997).
56. Merutka, G., Jane Dyson, H. & Wright, P. E. 'Random coil' 1H chemical shifts obtained as a function of temperature and trifluoroethanol concentration for the peptide series GGXGG. *J. Biomol. NMR* **5**, 14–24 (1995).
57. Miura, Y. α-proton Chemical Shift Index and Amide Proton Chemical Shift Temperature Coefficient of Melittin in Methanol: Indicators for a Helix Structure and an Intra-Molecular Hydrogen Bond? *Protein J.* **41**, 625–635 (2022).
58. Vijayalakshmi, S., Rao, R. B., Karle, I. L. & Balaram, P. Comparison of helix-stabilizing effects of α,α-dialkyl glycines with linear and cycloalkyl side chains. *Biopolymers* **53**, 84–98 (2000).
59. Steiner, D., Allison, J. R., Eichenberger, A. P. & Van Gunsteren, W. F. On the calculation of 3 J αβ-coupling constants for side chains in proteins. *J. Biomol. NMR* **53**, 223–246 (2012).

60. Crisma, M., Fasman, G. d., Balaram, H. & Balaram, P. Peptide models for β-turns. *Int. J. Pept. Protein Res.* **23**, 411–419 (1984).

V. Résultats – Étude structurale du peptide flexible Ala-Phe-Ala (AFA)

Table des matières

V. A. Choix du peptide et motivations

Cette seconde partie des résultats porte sur l'étude des peptides flexibles. Comme exposé dans les parties I et II, la flexibilité des petites molécules pose aujourd'hui un défi majeur pour la détermination précise de leurs conformations. En solution, les techniques actuelles de caractérisation structurale, telles que la RMN et le DCE, mesurent des paramètres qui reflètent une moyenne des conformations présentes en solution, surtout en cas d'échange rapide entre celles-ci.

Dans le chapitre précédent, nous avons montré l'intérêt de combiner ces méthodes expérimentales avec des simulations théoriques, notamment de la DM et du DCE, pour affiner la détermination des structures en solution. Dans ce chapitre, notre objectif est d'exploiter cette approche combinée pour caractériser les conformations de peptides flexibles en solution.

Pour cette étude, nous avons choisi le peptide Ala-Phe-Ala (AFA) pour plusieurs raisons (**Figure V.1**). En plus de son intérêt biologique dans l'étude des interactions aromatiques et des agrégats présents dans les fibres amyloïdes, ce peptide présente plusieurs avantages méthodologiques.[1] Tout d'abord, il a fait l'objet de nombreuses études antérieures, y compris des tentatives de détermination de sa structure, avec des proportions de structures secondaire variant considérablement selon les études (**Table V.1**), ce qui souligne sa flexibilité conformationnelle.[2–8]

Par ailleurs, le spectre DCE de AFA ne correspond pas à ceux des grandes familles de structures secondaires (hélices, feuillets, etc.), ce qui laisse penser que ce peptide adopte soit un ensemble de conformations en solution, soit une conformation majoritaire difficile à caractériser. Dans ce dernier cas, il pourrait s'agir d'une structure de type coude, comme les conformations γ ou γ_{inv}, avec une signature DCE spécifique à cette géométrie.

Comme dans le chapitre précédant, nous avons combiné des analyses expérimentales par DCE et RMN dans l'eau, avec pour cette dernière la détermination de distances caractéristiques et l'évaluation de la protection du proton amide, avec des simulations théoriques par DM puis TD-DFT. Afin d'obtenir une analyse plus robuste et de renforcer la convergence des données vers une même conclusion, nous avons également calculé par DFT la constante de couplage $^3J_{HN-H\alpha}$, qui est un indicateur des angles φ caractéristiques du repliement de la chaîne principale. Cette constante de couplage fournit une information supplémentaire contribuant ainsi à confirmer ou affiner l'identification des conformations adoptées par AFA en solution.

Notre approche combinant la RMN, le DCE et les simulations théoriques vise donc à élucider les divergences de la littérature en déterminant les conformations dominantes d'AFA en solution.

Cette introduction est suivie de l'article présenté sous la forme d'un manuscrit (non soumis), accompagné des « *supporting information* » en annexe. Enfin, une dernière section est dédiée aux conclusions et perspectives, venant clôturer le chapitre.

V. B. Article (Manuscrit)

V. B. 1. Introduction

The determination of protein and peptide tertiary structures in solution is a well-established process, particularly when a predominant conformation, with the folding of helices and sheets, is present or when the molecule is globular. In these cases, which are relatively common, the structure elucidation can be easily overhauled right to the end. To this aim, Nuclear Magnetic Resonance (NMR) and Electronic Circular Dichroism (ECD) are the most efficient tools that provide accurate information about the secondary structures in solution, which can lead to the three-dimensional structure. ECD spectrum pattern can be deconvoluted, using reference data from known secondary structures, such as helices or sheets established for polypeptides to predict the conformations in solution. Complementary, NMR can be used either to calculate interproton distances, to highlight which amide protons are protected from the solvent and could participate to an intramolecular hydrogen bond or to estimate the dihedral angle value linked to the ^{3}J H-H coupling in the peptide backbone (ψ for ^{3}J H_N-H_α) or in the side chain (χ for ^{3}J H_α-H_β).

When multiple conformations coexist in solution, the structure determination might be significantly complex because in such cases, the measured NMR parameters represent average values across all conformations. The issue is even more pronounced for very small peptides, where flexibility is often increased. The high prevalence of turns in small peptides adds an additional layer of complexity to data interpretation. Indeed, for these motifs, there are few or no reference data from NMR or reference ECD spectra available. This can lead to contradictions and mistakes.

Ala-Phe-Ala (AFA) represented in **Figure V.1** is an interesting peptide, particularly its intriguing solubility and conformational flexibility. Indeed, aromatic peptides are often highly prone to aggregation, playing a significant role in the formation of amyloid fibrils.[1] These fibrils are insoluble aggregates often rich in β-sheet structures and are associated with several degenerative diseases, including Alzheimer‘s disease. For instance, the β-amyloid peptide (Aβ), contains an aggregation-prone core segment $_{16}$KLVFFAE$_{22}$ which is rich in aromatic residues.[9,10]This aggregation tendency typically makes such peptides poorly soluble, posing challenges in pharmaceutical applications where solubility is essential.

Surprisingly, despite containing the aromatic phenylalanine residue, AFA does not exhibit the same propensity for aggregation. It has been reported to be highly soluble in water, with a critical concentration reaching up to 10 mM, and shows no signs of self-assembly.[2–5]

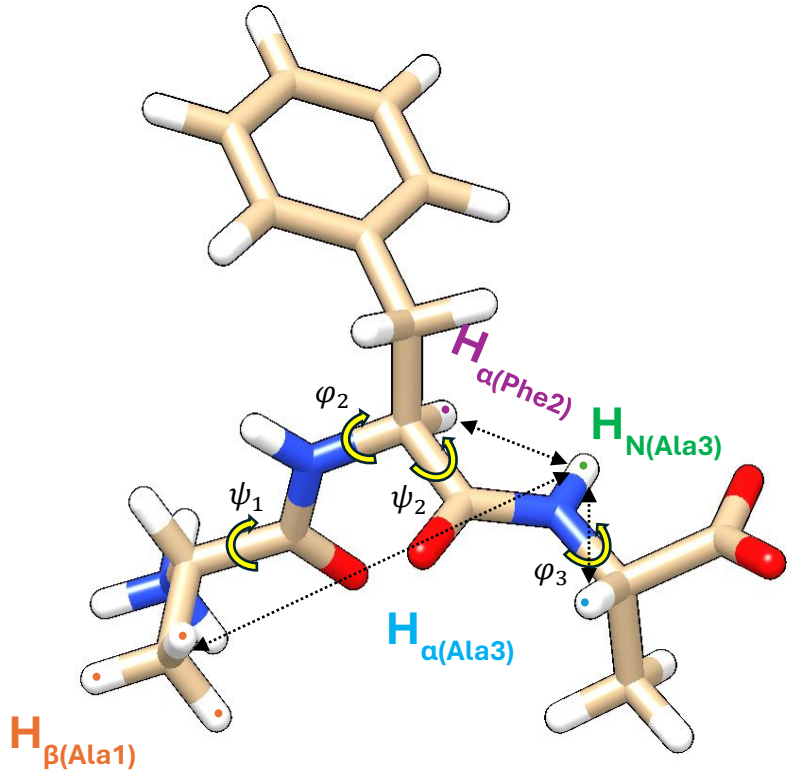


***Figure V.1**: Ala-Phe-Ala peptide. The three measured distances are displayed between the $H_{N(Ala3)}$ proton (in green) and the $H_{\alpha(Phe2)}$ (in purple), the $H_{\beta(Ala1)}$ (in orange) and the $H_{\alpha(Ala3)}$ (in cyan) protons. The dihedral angles of the peptide backbone are displayed in yellow.*

Elucidating the precise conformation of AFA is, therefore, of great interest to give insight to this solubility issue. However, various articles have been published with totally contradictory conclusions regarding the conformation of this peptide in solution (see **Table V.1**). The peptide was initially established as preferentially adopting the β-strand structure in water by Fourier Transform-InfraRed (FT-IR), Raman, Vibrational Circular Dichroism (VCD) and ECD spectroscopies at pH = 1.[2] Only one year later, it was reported to fold into the γ_{inv} turn (60 %) and the extended conformations (40%) in water at pH = 7.2 by ECD, NMR, FT-IR and molecular modeling under NMR constraints.[3] These conformational proportions are at odds with a theoretical study, which calculated the conformational propensities of residues in various host-guest peptides.[6] Their findings indicated that phenylalanine displays a distribution with approximately 0.35 Polyproline II (PPII), 0.20 helix and 0.20 β-strand conformations.

A few years later, different proportions and conformational preferences have been reported by Molecular Dynamic (MD) simulations coupled with NMR, FT-IR, polarized Raman and VCD. In this study, the AFA peptide was found to fold mainly into a PPII (37 % - 57 %) and extended conformations (31 % - 50 %) and weakly into an helix (8 % - 13 %) and a γ_{inv} (0 % - 4 %) conformation.[4] This results were supported by another study on the GFG peptide, a peptide very similar to AFA, in which they proposed a distribution between β-strand and PPII conformations by vibrational spectroscopy and NMR.[7]

More recently, the surprising solubility of this peptide was studied.[5] Using ECD, they have suggested that AFA appears to be solvated mainly as β-strand structures, which in fact turn out to be predominantly monomers, leading to an increase of its solubility.

In parallel, another theoretical study has suggested a non-negligible proportion of the α-helical conformation of the AFA peptide in solution.[8] They have reported proportions of 61 %, 14 %, 2 % and 20 % for the extended chain, PPII, γ_{inv} and the helix conformation respectively, using continuum model for water. However, when using an explicit solvation model with five water molecules, these proportions change drastically to 5 %, 79 % and 8 % for the extended chain, PPII and helix conformations respectively. Despite these findings, no clear preference for a particular solvent model has been established.

Overall, it seems that AFA is a highly versatile peptide which may fold into several conformations. But this peptide is also of biological interest because of its inability to produce amyloid fibrils despite the presence of the central aromatic residue. We have therefore chosen to revise its structure in solution by a combined experimental and theoretical approach.

		Ref. 2	Ref. 3	Ref. 4	Ref. 5	Ref. 8	
Experimental Methods in solution		**FT-IR Raman VCD ECD**	**FT-IR ECD NMR**	**FT-IR Raman VCD NMR MD**	**ECD**	**PCM**	**Explicit**
Conf.	**α-helix**	-	0 %	8-13 %	-	20 %	8 %
	β-strand	**Major**	40 %	31 % 50 %	**Major**	**60 %**	5 %
	PPII	-	0 %	**37 %-57 %**	-	14 %	**79 %**
	γ_{inv}	-	**60 %**	0 %	-	2 %	0 %
pH		1	7.2	Mix 4 and 7	7	-	

***Table V.1**: Secondary structure proportions reported from solution structural analyses of AFA from literature*

In this work, using the overall strategy summarised in **Figure V.2**, we assess the robustness of a mixed approach to determine the structure or the different conformations of very short peptides. The strategy combines an experimental part involving the measurement of NMR and ECD parameters with a sophisticated theoretical part. Geometries were generated by MD simulations and were compared with the experimental NOE distances to validate the MD protocol. Then, the choice of the force field was guided by the simulation of the ECD spectrum and NMR parameters, in which the environment was taken into account through the Polarizable Embedding model (PE). The final protocol considers all the experimental data and ensures a very good agreement with them. This mixed approach is totally suitable for very small or flexible peptides, for which reference spectra and spectroscopic signatures of secondary structures have not been established.

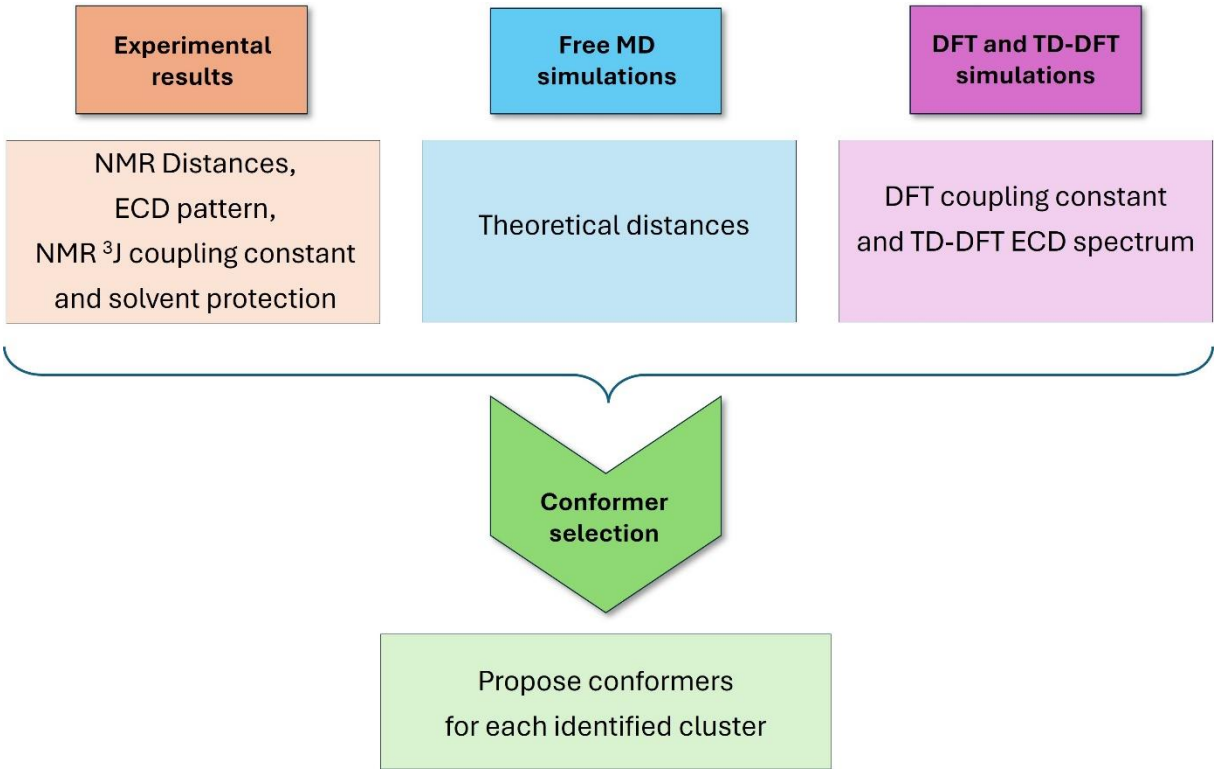


***Figure V.2**: Global Method comprising experimental compared with MD simulations and quantum calculations for conformer selection.*

V. B. 2. Material and Method

V. B. 2. a) Peptide and solvent supplier

Experimental analyses were performed on the Ala-Phe-Ala peptide synthetized by Abbexa (Cambridge, United Kingdom). D_2O were supplied by CND Isotopes (Sainte Foy la Foy la Grande, France) and Ultra-pure water was used for the peptide solvation.

V. B. 2. b) Electronic circular dichroism experiments

The ECD spectra of the AFA peptide were measured from 185 to 260 nm at room temperature in water using a MOS 500 ECD spectrometer (Biologic, Seyssinet-Pariset, France) with a Xenon light source. Measurements were taken at pH 7.2 (293 K and 281 K) and 4.0 (293 K) with a concentration of 0.5 mM in a 1 mm path length cell. The pH was adjusted by adding NaOH and HCl solutions. Three scans were accumulated and averaged for each sample. All spectra were corrected by subtraction of the background obtained for each peptide-free solution. For comparison between experimental and theoretical results, the ECD spectrum was plotted in terms of rotatory strength (M^{-1} cm^{-1}).

V. B. 2. c) Nuclear magnetic resonance experiments

All NMR experiments were conducted using a Bruker Avance III 600 MHz spectrometer equipped with a cryoprobe. The peptide concentration was set to 0.5 mM in a water solution containing 10% D_2O. Sodium 2,2-dimethyl-2-silapentane-5-sulphonate (DSS) was added as an internal reference. Two separate samples, whose pH was adjusted by addition of NaOH or HCl solutions to 7.2 and 4 respectively, were used for the NMR studies. Experiments were recorded at 293 K and 300 K.

Chemical shift assignment was performed using the following conventional two-dimensional proton experiments: 2D COSY, 2D TOCSY and 2D NOESY (respectively Bruker's *cosygpprqf*, *dipsi2esgpph* and *noesyesgpph19* pulse programs). Proton chemical shifts were reported relative to DSS taken as an internal reference.

TOCSY spectra were obtained with a 60 ms spin-lock time using a DIPSI sequence.[11] NOESY spectra were recorded with mixing times from 200 ms to 700 ms. Both TOCSY and NOESY experiments were conducted in the phase-sensitive mode, employing the time proportional phase incrementation (States-TPPI) method for quadrature detection.[12] Water suppression in the 1D and TOCSY experiments was achieved using the excitation sculpting method[13] whereas WATERGATE method[14] was performed for 2D NOESY experiments. In the COSY

experiments, water resonance was suppressed with a low-power presaturation. 2D Spectra were collected with 4096 and 512 complex data points in t1 and t2 dimensions. Data were processed using TOPSPIN 4.1.4 software (Bruker, Wissembourg, France). The f1 dimension was zero-filled to 1024 real data points. Time domain data were multiplied by shifted sinebell or Gaussian functions in both dimensions prior to Fourier transformation.

Amide temperature coefficients were determined by recording 1D ^{1}H experiments from 275 K to 309 K using a 2K step.

$^{3}JH_N$-H_α coupling constant value was measured on a 1D ^{1}H experiment recorded with a FID resolution of 0.78 Hz by point and the deconvolution was performed by 2 Lorentzian functions.

Distances were derived from a NOESY spectrum recorded with a mixing time of 600 ms with 64 scans. NOE cross peak volumes a_i were converted into r_i distances by the following expression from Wüthrich:[15]

$$r_i = r_{ref}\left(\frac{a_{ref}}{a_i}\right)^{1/6} \quad (1)$$

The NOE volumes were calibrated using geminal H_β cross-peak of the phenylalanine residue, which corresponds to a reference distance r_{ref} if 1.8 Å. An overall error on the calculated distance was set to 20 % to take into account the potential presence of interfering effects such as internal mobility, chemical exchange and experimental interferences. The derived distances were first compared to reference distances of each secondary structures from constrained MD simulations (see section V. B. 2. e)) and then used to optimize the theoretical protocol and finally determine the structures in solution.

V. B. 2. d) Unrestrained MD simulations

The peptide (**Figure V.1**) was built with the Chimera Software and the φ_2 and ψ_2 dihedral angles of the central residue were adjusted to fit the γ_{inv} (found in ref. **3**), the α-helix, the β-strand and the polyproline-II conformations.[16]

Then, MD simulations were performed with GROMACS 2022[17,18]. They were performed with both OPLS-AA[19] and CHARMM27[20] force fields within the TIP5P[21] water model and utilizing the LINCS[22] algorithm on bond constraints. The peptide was solvated in a periodic cubic box of about 3x3x3 nm^3.

The MD simulations were organised in four parts. First, an energy minimization was performed for 0.1 ns. Then a first step of equilibration was conducted in the NVT ensemble for 0.1 ns with

a 0.02 ps time step, in which the temperature reached 300 K with the Bussi's velocity thermostat.[23] The second equilibration was then carried out in the NPT ensemble for 0.1 ns within the Parrinello-Rahman barostat and the Particle Mesh Ewald (PME) method was used for electrostatic calculations with a van der Waals cutoff of 1 nm.[24] Finally, a production step of 1 µs was run and increased to 2 µs for the final simulation chosen for additional calculations, in which snapshots of the molecule geometry were saved each nanosecond. The theoretical distances were then calculated as the average over the 1000 extracted geometries, with the standard deviation providing a measure of the distribution of values.

V. B. 2. e) Reference distances for the secondary structures

This section has been added solely for the purpose of calculating reference distances for each secondary structure and comparing them with the experimental values (**Figure V.4**). For this section, MD was performed with the OPLS-AA force field within the TIP5P water model. The approach is very similar to the last section except that constraints of 10° on dihedral angles of the central residue were added and the production step was reduced to 10 ns in which geometries were extracted every 0.1 ns. The reference distances for each secondary structure were then calculated over 100 extracted geometries, with the standard deviation indicating the distribution of values.

V. B. 2. f) Clustering

The clustering was first achieved using the following area[4,25] on the Ramachandran plot: $-180° < \varphi_{\alpha-helix} < 0°$, $-120° < \psi_{\alpha-helix} < 30°$; $-180° < \varphi_{\beta-sheet} < -100°$, $90° < \psi_{\beta-sheet} < 180°$; $-180° < \varphi_{\beta-sheet} < -100°$, $-180° < \psi_{\beta-sheet} < -120°$; $-100° < \varphi_{PPII} < -50°$, $100° < \psi_{PPII} < 180°$; $0° < \varphi_{Left-handed\ helix} < 180°$, $-30° < \psi_{Left-handed\ helix} < 120°$; $-120° < \varphi_{\gamma inv} < -30°, 30° < \psi_{\gamma inv} < 100°$. The γ_{inv} area were reduced to avoid too much overlap with the β-strand and the PPII areas.

The k-means algorithm[26] was also applied on the φ_2 and ψ_2 dihedral angles of the phenylalanine residue using four, five and six clusters. The elbow method was used to assess the Within-Cluster Sum of Squares (WCSS) convergence. To account for the angular periodicity, these angles were projected onto unit circles using polar coordinates. Thus, the clustering was based on four descriptors derived from φ_2 and ψ_2: $\cos(\varphi_2)$, $\sin(\varphi_2)$, $\cos(\psi_2)$ and $\sin(\psi_2)$.The conformer of each cluster was chosen as the structure with the lowest Root Mean Square Deviation (RMSD) with the other intra-cluster geometry concerning the peptide backbone.

V. B. 2. g) Theoretical ECD spectrum

The geometries of the unrestrained MD simulations were used to calculate the ECD spectrum of the AFA peptide. The solvent was included within the polarizable embedding. This sophisticate approach have demonstrated its effectiveness to similar problematics[27,28] and is completely implemented in the DALTON 2018 program[29] which was used for the quantum mechanics/molecular mechanics calculations (QM/MM). In our approach, only the peptide was included in the QM part, as it has been demonstrated that the number of solvent molecules considered in the QM part has no impact on the ECD spectrum. In the PE approach,[30–32] the effective Hamiltonian is defined as followed:

$$\hat{H}^{eff} = \hat{H}^{KS} + \hat{v}^{PE} \tag{2}$$

where $\hat{H}^{KS}$ is the Kohn-Sham Hamiltonian in vacuum and $\hat{v}^{PE}$ is the PE operator defined as:

$$\hat{v}^{PE} = \sum_{s=1}^{S}\sum_{k=0}^{K}\frac{(-1)^{(k+1)}}{k!}Q_s^{(k)}\sum_{pq}T_{s,pq}^{(k)}\hat{E}_{pq} - \sum_{s=1}^{S}\mu_s^{ind}(F[\rho])\sum_{pq}T_{s,pq}^{(1)}\hat{E}_{pq} \tag{3}$$

in which the charges and the polarizabilities of the MM part are represented by the terms $Q_s^{(k)}$ and μ_s^{ind} respectively which denotes the k^{th} order multipole for the first one and the vector containing the induced dipole moments by the electric field F for the second one. The index pq is a matrix element in the Kohn-Sham orbitals. T represents the interaction tensors and $\hat{E}$ is the excitation operator. This method enables to take into account both the environment's effect on the quantum core and the polarization of the MM part by the QM part. This method should be well-suited for our work since it offers a complex consideration of the solute-solvent interaction which can lead the conformational state by hydrogen bonds. The DALTON input files were generated by the PyFrame library.[33]

The ECD spectra were generated by the TD-DFT calculations of the 20 first excited states i which were then included in a convolution by gaussian functions[34]:

$$\Delta\varepsilon_f(E) = \frac{1}{2.297.10^{-39}}\frac{1}{\Delta\sqrt{\pi}}\sum_{i=1}^{20}\Delta E_{0\to i}R_{0\to i}e^{-\left(\frac{E-\Delta E_{0\to i}}{\Delta}\right)^2} \tag{4}$$

where the $\Delta\varepsilon_f$ is the molar ellipticity, Δ is the gaussian bandwidth, $R_{0\to i}$ is the rotatory strength and $\Delta E_{0\to i}$ is the transition energy. The CAM-B3LYP functional with the 6-31+G* basis set were employed.[35] The convolution was processed by homemade Python scripts and the final

ECD is the average over all extracted geometry (1000 for 1 µs simulations and 2000 for the 2 µs simulation)

V. B. 2. h) Theoretical Coupling constant

In a theoretical point of view, the indirect spin-spin coupling constant J between the atoms K and L can be expressed as the second derivative of the energy E with respect to the nuclear magnetic moments M:

$$J_{K,L} = h\frac{\gamma_K}{2\pi}\frac{\gamma_L}{2\pi}\frac{\partial^2 E}{\partial M_K \partial M_L} \quad (5)$$

where γ is the nuclear magnetogyric ratio and h is the Planck constant. This equation can be solved within the linear response formalism.[36] Coupling constant calculations were first performed by DFT with the CAM-B3LYP, SVWN[37], BLYP, B3LYP[38] and the PBE0[39] functionals and the 6-311++G** basis set still within the PE on the 1000 geometries generated by the unrestrained MD simulations and considering only the Fermi Contact (FC) contribution. This calculation was made to ensure the accuracy and consistency of the computed coupling constants across different functionals. However, a more robust basis set is required to predict a reliable coupling constant. Thus, the calculation was also performed with the CAM-B3LYP functional with the aug-cc-pVTZ-J basis set, which was designed for this type of calculation, considering all contributions to J.[40] This calculation taking a long time, it was performed only every 10 geometries. A very good correlation between the two types of calculation was found, allowing the prediction of the 1000 coupling constants with the more robust one (**Figure S10**).

V. B. 3. Results

V. B. 3. a) Conformation analyses by Electronic Circular Dichroism and NMR

These studies were conducted in water at neutral pH and 293 K. This pH was chosen to correspond to a neutral physiological pH, and to avoid MD simulations with a cation/neutral mixture.

The ECD spectrum of AFA is shown **Figure V.3**. Two positive transitions can be observed under 300 nm. The first one around 216 nm is rather weak and wide, and the second one is almost twice more intense at 196 nm. This spectrum shows a highly unusual pattern and does not exhibit the features of helical, β sheet or random coil conformations. Although it could be due to the presence of a single conformation, such as a turn for which a reference spectrum has not yet been assigned, such spectrum is in agreement with the flexibility and the mixing of conformations reported for AFA in the literature. We have not noticed significant change with

pH and temperature (data non shown) suggesting that the structure equilibrium appears rather stable over a reasonable range of temperature and pH. Our experimental ECD spectra are also consistent with the literature although different conditions were sometimes used. [2,3] .

At this point, even if some studies have postulated that this kind of pattern could reflect the presence of a combination of the PPII and the β-strand conformations,[5] no conclusion on AFA conformational mixture can be drawn based on the ECD spectrum.

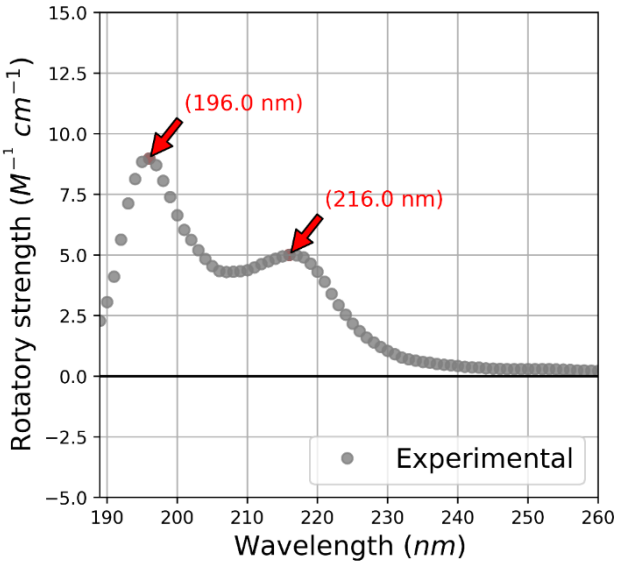


Figure V.3*: Experimental ECD spectrum of the AFA peptide. The spectrum was recorded in a mixture H_2O/D_2O (90/10), at a concentration of 0.5 mM and a temperature of 20°C. The pH was adjusted to 7.2. The red arrows indicate the two maxima at 196 and 216 nm.*

AFA conformation was further investigated by NMR. Its ^{1}H 1D spectrum at neutral pH and 293 K is shown in **Figure S1**. Only one set of resonances is present on the spectrum indicating that the peptide multiple conformations in aqueous solution are fast exchanging conformations on the NMR time scale.

Proton chemical shift assignment was carried out using the strategy published by Wüthrich and its collaborators.[15] Spins systems were assigned (**Table S1**) using the COSY (**Figure S2**) and the TOCSY (**Figure S3**) experiments.

At neutral pH, only the $H_{N(Ala3)}$ proton was visible on the NMR spectra (**Figure S1**, **Figure S2** and **Figure S3**). The temperature dependence of amide proton chemical shift, which is usually used to probe the existence of hydrogen bonds in small peptides, was thus measured from 275 K to 305 K at pH 4.0 and pH 7.2 (Figure S4). For both pH, linear curves were obtained

indicating that no significant conformational equilibrium changes in the 275-305 K temperature range.

The temperature coefficients determined for $H_{N(Phe2)}$ and $H_{N(Ala3)}$ protons are reported at pH 7.2 (**Table V.2**) and pH 4.0 (**Table S2**). At the two pH values, both protons exhibited a temperature coefficient more negative than the threshold value of -4.6 ppb K^{-1} in water.[41] Such values are usually linked to an opened hydrogen site, whereas values more positive than the threshold are more compatible with the presence of a well-protected site. Despite AFA flexibility, this reflects the absence of hydrogen bonds in some of the conformations adopted by the peptide in water.

Table V.2*: NMR spectroscopic parameters (interproton distances; coupling constant J and amide proton temperature coefficient) at pH 7.2 in water.*

	pH 7.2
$H_{N(Ala3)}$-$H_{\alpha(Phe2)}$	2.4 Å
$H_{N(Ala3)}$-$H_{\alpha(Ala3)}$	2.9 Å
$H_{N(Ala3)}$-$H_{\beta(Ala1)}$	4.2 Å
$^3JH_{N(Ala3)}$-$H_{\alpha(Ala3)}$	7.5 Hz
$\frac{d\delta_{ppm}}{\Delta T}$	- 7.5 ppb K^{-1}

Although NMR experiments were performed at 300 K and 293 K, we chose to investigate the parameters (interproton distances and ^3JH_N-H_α coupling constants) classically used for structural studies at 293 K as the temperature lowering allowed to reduce amide proton exchange with water. Although this temperature was slightly different from the 300 K used for the subsequent Molecular Dynamics simulations, this should not present any issues. Indeed, AFA ECD spectrum was unaffected by temperature change, and linear variations were observed for amide proton chemical shifts in this temperature range.

Only one coupling ^3JH_N-H_α constant, $H_{N(Ala3)}$-$H_{\alpha(Ala3)}$, could be measured at neutral pH (**Table V.2**). Let us recall that this coupling constant is related to the dihedral angle φ_3, one of the two angles used to describe the secondary structures in peptides. A value lower than 6 Hz is generally found in helical secondary structures, while a value higher than 9 Hz is observed in extended structures.[15,42] In the case of a small peptide like AFA, distortions may be observed compared to canonical secondary structure elements, and therefore, the absence of values for the first two amino acids makes the interpretation difficult. We can nevertheless note that a value of 7.5 Hz is an average value compatible with the existence of a conformational mixture.

To determine intramolecular distances between hydrogens, we have recorded a NOESY spectrum with a high signal to noise ratio (**Figure S5**). Despite the limited number of NOE correlations at neutral pH (**Figure S6** and **Figure S7**), three distances ($H_{N(Ala3)}$-$H_{\beta(Ala1)}$, $H_{N(Ala3)}$-$H_{\alpha(Phe2)}$ and $H_{N(Ala3)}$-$H_{\alpha(Ala3)}$) which are of interest for secondary structure investigation, were calculated (**Table V.2**), **Figure S6** shows that there is no correlation between the amide proton of the third residue and the α proton of the first residue. Besides, we can already see that there is a correlation between the amide proton of the third residue and the β protons of the first residue (**Figure S7**). This suggests that there are not only extended conformations in solution. However, with so little information and a complex mixture to interpret, we need further information to extract reliable populations. These distances were thus used in conjunction with theoretical data.

For the structural analysis of a peptide with a predominant conformation, the experimental distances are compared to reference distances in order to identify the secondary structure elements. As mentioned in the introduction (**Table V.1**), several contradictory studies have been conducted on AFA, some of which were partially based on NMR. However, for a few of these secondary structure elements, there are no reference distances available in the literature. Moreover, since AFA is a tripeptide, some reference distances may not be present. In order to make a reliable comparison, we have determined the expected distances for AFA adopting different secondary structures through molecular dynamics. Distances were calculated by imposing constraints on φ_2 and ψ_2 dihedral angles in 10 ns simulations and starting from each of the following secondary structure element: α-helix, β-strand, polyproline-II and inverse γ-turn.

Subsequently, we derived the average values for the three distances $H_{N(Ala3)}$-$H_{\beta(Ala1)}$, $H_{N(Ala3)}$-$H_{\alpha(Phe2)}$ and $H_{N(Ala3)}$-$H_{\alpha(Ala3)}$ distances from 1000 geometries with a 0.01 ns time step. These results provided theoretical distances as shown in **Figure V.4** where experimental distances are also presented.

A rapid comparison shows that experimental values do not match a single conformation, except the γ_{inv} structure in which all the distances are in line with the experimental. Nevertheless, the absence of an experimental correlation between the amide proton of the third residue and the α proton of the first residue contradicts this matching. Both β-Strand and PPII conformations exhibit very similar distances and only the α-helix conformation is significantly different over the $H_{N(Ala3)}$-$H_{\beta(Ala1)}$ and the $H_{N(Ala3)}$-$H_{\alpha(Phe2)}$ distances. Then, these two distances are also

precisely those for which the experimental values are different from the β-strand and the PPII conformation. These findings support the idea of multiple conformations, likely including α-helix, β-strand and PPII populations.

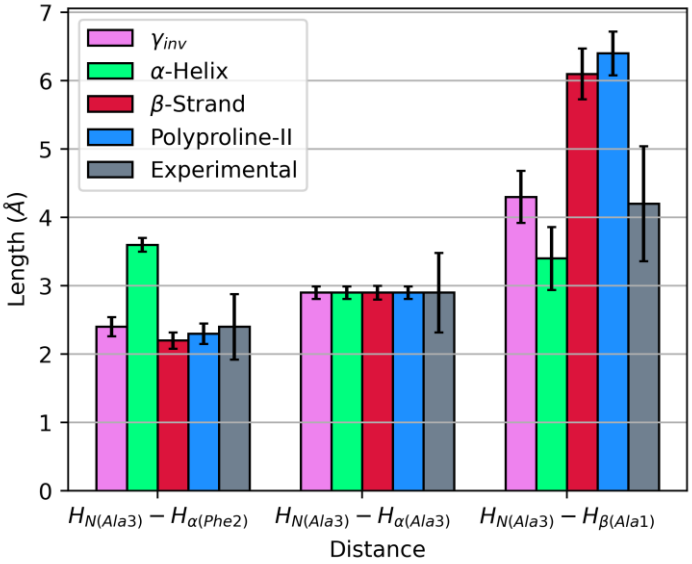


Figure V.4*: Histogram of the experimental and theoretical distances. The theoretical ones were calculated in constrained MD simulations with the OPLS-AA force field. The φ and ψ dihedral angles of the central residue were chosen to fit γ_{inv} (purple), α-helix (green), β-strand (red) and the Polyproline-II (blue) conformations as initial structures. Error bars have been added for each experimental distance, corresponding to 20 % of the measured values and the standard deviation for each simulated distance.*

V. B. 3. b) Generation of AFA low energy conformations by Molecular Dynamic simulations.

V. B. 2. b) i) MD simulations

Unrestrained MD simulations of 1 μs were used to generate all the low energy conformers. The peptide was thus free to populate the entire Ramachandran diagram for the central amino acid in each MD simulation. We have tested two frequently used force fields, namely OPLS-AA and CHARMM27. In addition, to ensure that the choice of the initial structure had no influence on the populations, four initial structures were tested: α-helix, β strand, PPII and γ_{inv}. To evaluate the quality of the MD, theoretical distances from the average over the geometries generated by each simulation were calculated for the three measured distances: $H_{N(Ala3)}$-$H_{\beta(Ala1)}$, $H_{N(Ala3)}$-$H_{\alpha(Phe2)}$ and $H_{N(Ala3)}$-$H_{\alpha(Ala3)}$.and compared to the experimental values.

The MD results of the eight simulations (four initial structures for each force field) are shown in the Ramachandran plots in **Figure S8**, **Figure S9** and in **Figure V.5**. In this latter, the three theoretical distances, $H_{N(Ala3)}$-$H_{\beta(Ala1)}$, $H_{N(Ala3)}$-$H_{\alpha(Phe2)}$ and $H_{N(Ala3)}$-$H_{\alpha(Ala3)}$, averaged over the 1000 geometries generated with OPLS-AA and CHARMM27 from each initial structures, were compared with the experimental ones. Whatever the initial conformation, MD simulations led to similar geometries (**Figure S8** and **Figure S9**) and theoretical distances (**Figure V.5**).

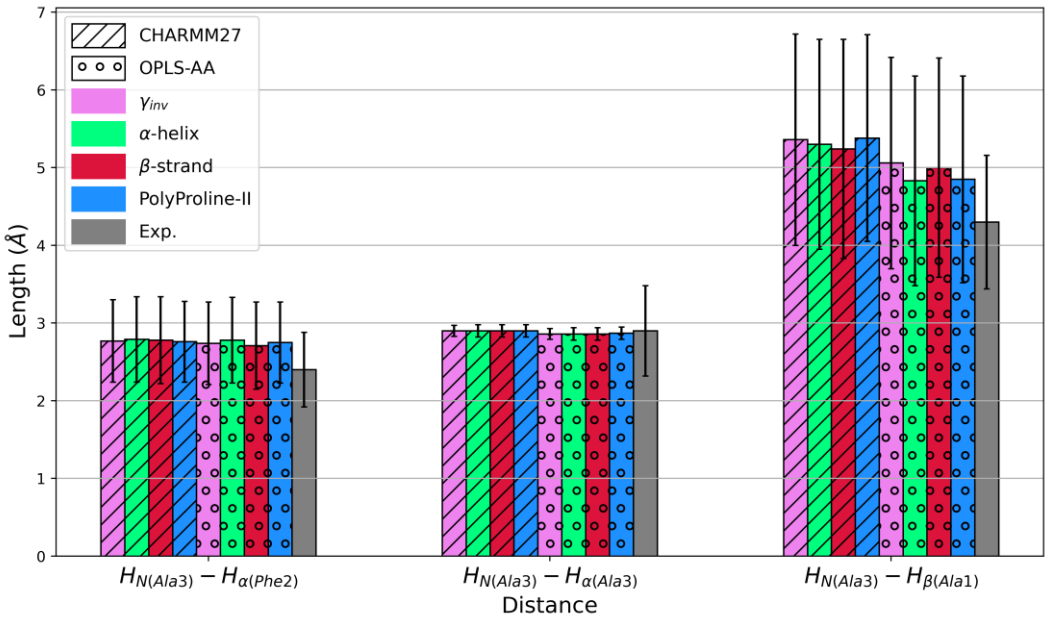


Figure V.5*: Histogram of the experimental and theoretical distances generated with the CHARMM27 (hashed bars) and OPLS-AA (circled bars) forcefield. The φ and ψ dihedral angles of the central residue were chosen to fit γ_{inv} (purple), α-helix (red), β-strand (blue) and the PPII conformations as initial structures. Error bars have been added for each experimental distance, corresponding to 20 % of the measured values and the standard deviation for each simulated distance.*

This is likely due to the duration of our simulations. Although 1 µs could be long for MD simulations on small peptides, we deemed it essential for a peptide that is presumably quite flexible. The $H_{N(Ala3)}$-$H_{\alpha(Phe2)}$ and $H_{N(Ala3)}$-$H_{\alpha(Ala3)}$ distances obtained from both OPLS-AA and CHARMM27 are very similar. The main difference between the two force fields is the theoretical distances obtained for the $H_{N(Ala3)}$-$H_{\beta(Ala1)}$. Taking into account errors on the experimental and calculated distance, the agreement with the experimental results is satisfactory despite the fact that for both force fields, the $H_{N(Ala3)}$-$H_{\beta(Ala1)}$ appears to be a bit overestimated. This could be due to a simulated helix population that is slightly too large.

The distances obtained from both OPLS-AA and CHARMM27 are almost equally close to the experimental values, although a slightly better agreement was found with OPLS-AA. This might be attributed to minor differences in the Ramachandran plot and the distribution of the

conformation. We thus decided to look in more detail in the geometries obtained by the simulations. To simplify the presentation, we focused our analyses on the geometries obtained with simulations using an inverse gamma turn as initial conformation. This choice was based on the fact that the MD simulation with this initial structure gives Ramachandran plots very similar to those obtained with the three other initial structures (**Figure S8** and **Figure S9**) and seems to explore more conformations.

Figure V.6 shows the distribution of the dihedral angle of the central residue of AFA in the Ramachandran plot, in which the areas corresponding to angles of an α-helix, β-strand, γ-inv and a PPII conformation are highlighted for geometries generated with the CHARMM27 (**Figure V.6A**) and the OPLS-AA force fields (**Figure V.6B**). Although no additional constraint was applied, the peptide appears to populate predominantly areas of the conformation α-helix, β-strand and PPII. Besides, the left-handed helix conformation is sparsely populated for both force fields. We can also notice that the α-helix area is not populated in the same way, but in both cases, two clusters seem to emerge.

An important point is that the angles generated with OPLS-AA are much more scattered than the ones generated with CHARMM27. This difference leads to a higher percentage of points outside the defined zones which are labeled “other” in **Figure V.7** in which the differences in conformation distribution between the two force fields are more clearly apparent. Moreover, there is an inversion of the distribution between the β-strand and the PPII conformations with 26 % and 19 % respectively with CHARMM27 and 19 % and 25 % respectively with OPLS-AA. For both force fields, the α-helix proportion has reached more than 30 %.

While OPLS-AA seems more suitable for our study, no definitive conclusion about the optimal choice of force field can be drawn at this stage. Therefore, additional theoretical data that can be compared to experimental results are needed to make a more informed decision regarding the choice of the force field. To carry out these additional analyses, we decided to use the geometries of the molecular dynamics obtained by using the γ_{inv} as initial structure.

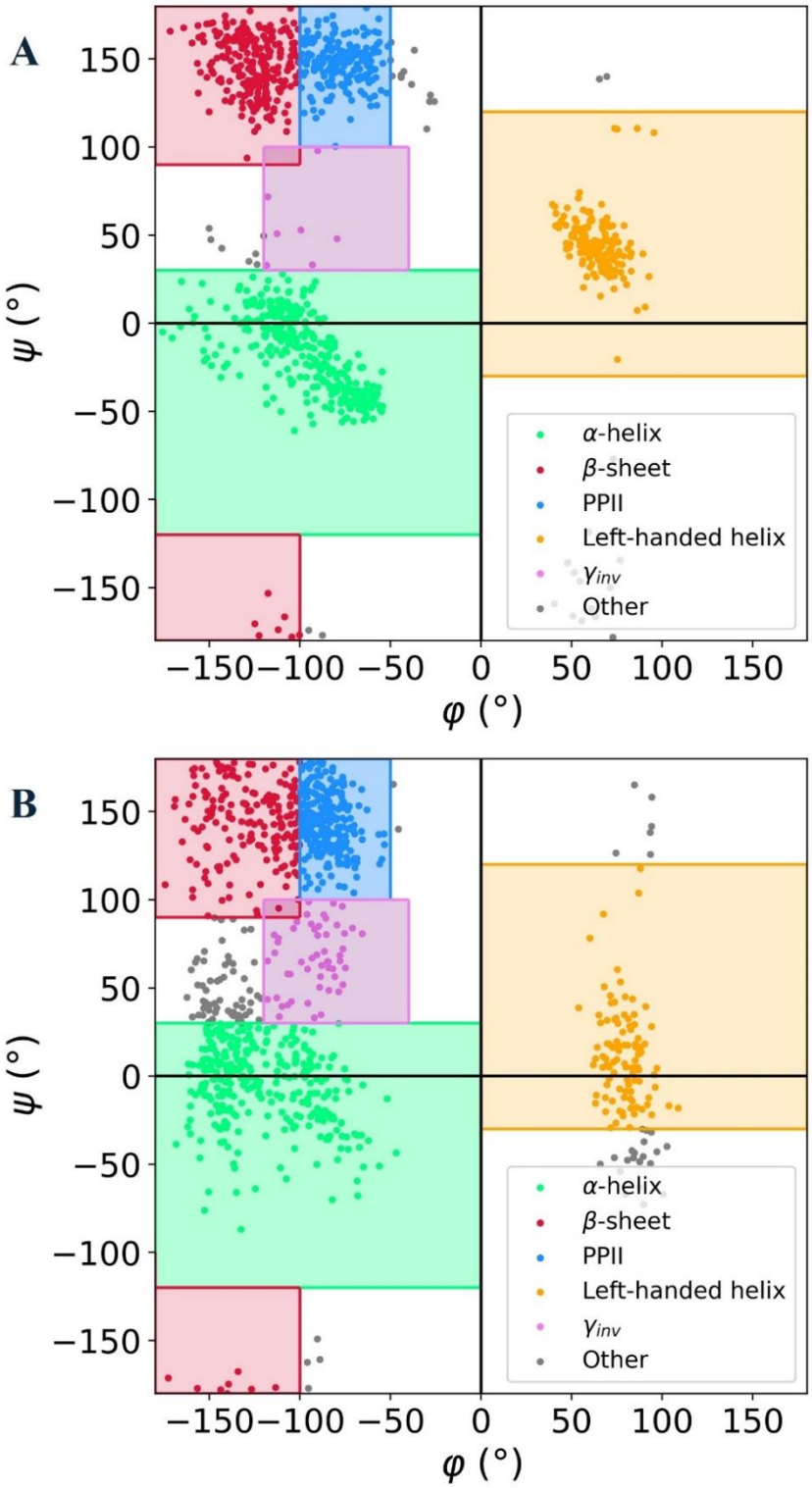


***Figure V.6**: Ramachandran plots of the AFA peptide by MD simulations with the CHARMM27 (up) and the OPLS-AA force fields. The area corresponding to the α-helix (green), β-strand (red), PPII (blue) and left-handed helix (orange) are highlighted.*

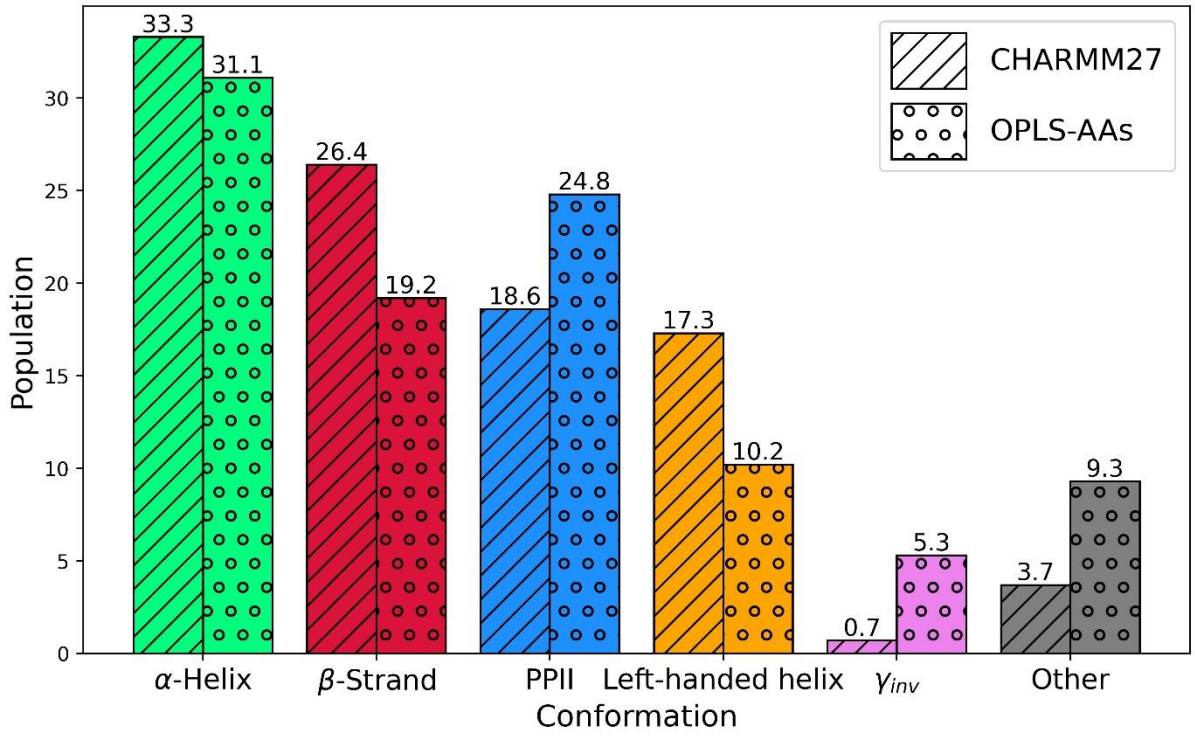


***Figure V.7**: Population (Percentage) of geometries inside the α-helix (green), β-strand (red), PPII (blue), left-handed helix (orange) and γ_{inv} (purple) conformations with the two different force fields (OPLSS-AA and CHARMM27).*

V. B. 2. b) ii) NMR indirect spin-spin coupling constant calculation

In this section, we have compared the two force fields by simulating a coupling constant of the peptide backbone, which is related to the geometry adopted by the molecule in solution. Since the three distances alone do not allow us to comfortably discriminate between OPLS-AA and CHARMM27, we have included the comparison of this spectroscopic parameter in **Table V.3**.

The $^{3}JH_{N(Ala3)}\text{-}H_{\alpha\,(Ala3)}$ coupling constant was computed by DFT. Although aug-cc-pVTZ-J is specifically designed for this type of calculation, it is very computationally intensive. We therefore carried out preliminary calculations using the 6-311++G** basis set, testing different exchange functionals and taking into account only the FC contribution over the 1000 geometries. This choice was motivated by both the high speed of such calculations (few min for each frame) and the prevalence of this contribution on the final value of the coupling constant in most cases.[43] The SVWN, BLYP, B3LYP, CAM-B3LYP and the PBE0 functionals (**Table S5**) were tested. The results demonstrate the convergence of the value by increasing the DFT level of theory for each force field, and the CAM-B3LYP has been chosen for further calculations. However, the prediction performed with this basis is not enough precise, and no conclusion can be made concerning the coupling constant value.

Indeed, in this type of calculations, a special attention must be paid concerning the basis set used. We have therefore considerably increased the accuracy of our calculations by using the aug-cc-pVTZ-J basis, taking into account not only the FC contribution but all contributions. This new calculation taking a long time, it was performed only on 100 geometries. A very good correlation between the two basis sets was found (**Figure S10**) and thus used to predict the coupling constant with the larger basis set over the 1000 geometries. **Table V.3** shows these revised values for the two force fields. Interestingly, the value with OPLS-AA (7.9 Hz) differs significantly from the one with CHARMM27 (9.2 Hz). Moreover, the value with OPLS-AA aligns much more closely with the experimental value (7.5 Hz) than the one with CHARMM27. This represents the first major difference between the two force fields.

Table V.3*: Comparison between experimental measures and simulated values on geometries from 1 µs MD simulations on a* γ_{inv} *structure with OPLS-AA and CHARMM27 force fields at 300 K.*

	Exp	CHARMM27	OPLS-AA
$H_{N(Ala3)}$-$H_{\alpha(Phe2)}$ (Å)	2.4 (0.5)[a]	2.77 (0.53)[b]	2.74 (0.49)[b]
$H_{N(Ala3)}$-$H_{\alpha(Ala3)}$ (Å)	2.9 (0.6)[a]	2.90 (0.07)[b]	2.86 (0.08)[b]
$H_{N(Ala3)}$-$H_{\beta(Ala1)}$ (Å)	4.2 (0.8)[a]	5.36 (1.36)[b]	5.06 (1.41)[b]
$^3J H_{N(Ala3)}$-$H_{\alpha(Ala3)}$ (Hz)	7.5 (1.6)[a]	9.5 (2.3)[b]	7.7 (2.6)[b]

a: The brackets indicate de experimental error (20 %)

b: The brackets indicate the standard deviation for simulated parameters.

V. B. 2. b) iii) Theoretical ECD simulation

To further refine our choice of the force field, we have then simulated the theoretical ECD spectra from geometries generated with the OPLS-AA and the CHARMM27 force fields. **Figure V.8** shows the two spectra, generated in the same way, except for the choice of the force field used for generating the geometries. The spectrum generated using geometries obtained with CHARMM27 shows a significant deviation from the experimental one. Though a positive peak was observed, that could be assigned to the peak at 216 nm, but at 220 nm, the second peak at 196 nm, which was the most intense, is completely absent. In contrast, the spectrum generated with geometries from OPLS-AA is very close to the experimental spectrum. While perfectly reproducing the pattern, it appears to be slightly shifted towards higher energies, with peaks at 193 and 214 nm.

In conclusion, the good convergence between the spectroscopic experimental data and the ones simulated using the geometries obtained with OPLS-AA indicates that the geometries generated with the OPLS-AA force field closely resemble the actual experimental results.

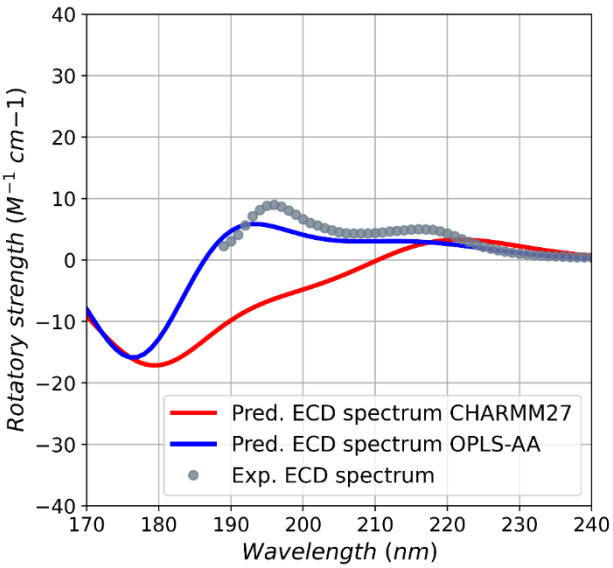


Figure V.8*: Predicted ECD spectra calculated on the geometries generated with the OPLS-AA (blue) and the CHARMM27 (red) force fields in a 1 µs simulation at 300K.*

V. B. 3. c) Analysis of ECD signatures and conformers determination

In this study, ECD spectrum was the key spectroscopic parameter for determining structures in agreement with experiment. In contrast, the experimental distances were not decisive, nor was the coupling constant, which did not provide enough information about the whole structure. To explore theoretical spectrum information in detail, a final ECD calculation was performed using geometries generated by the OPLS-AA force field with a dynamic simulation time increased from 1 to 2 µs. This extended calculation also aimed to obtain geometries matching as closely as possible the experimental ones. A minor change in the theoretical distances and very little change in conformation proportions were observed (**Table S4**). The MD results are shown in the heatmap (**Figure V.9**). The ECD results are shown in **Figure V.10** where the final predicted ECD spectrum (solid black line) matches the experimental one almost perfectly.

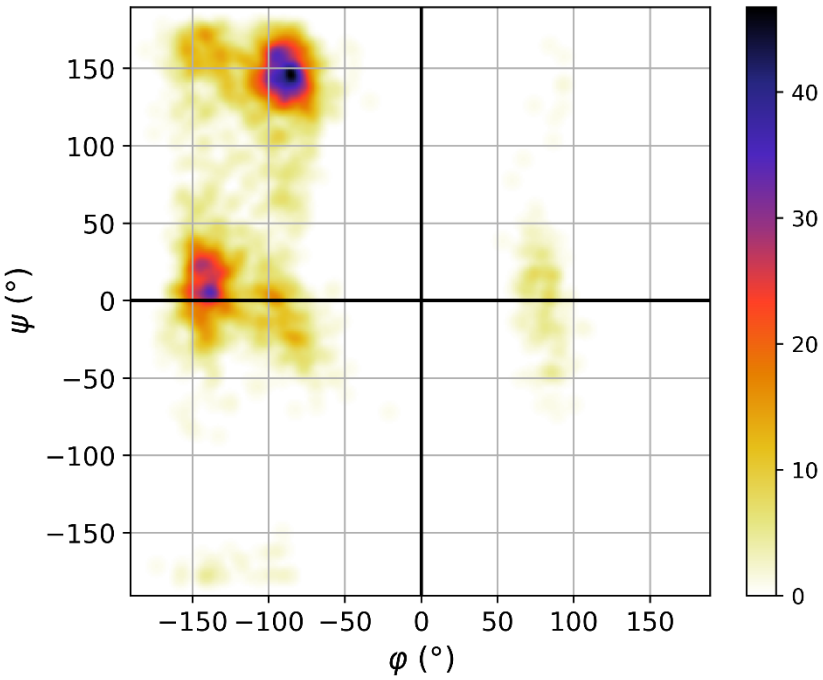


Figure V.9: *Heatmap of the Ramachandran plot generated with a 2 µs simulations with the OPLS-AA force fields at 300 K with the γ_{inv} conformation as initial structure.*

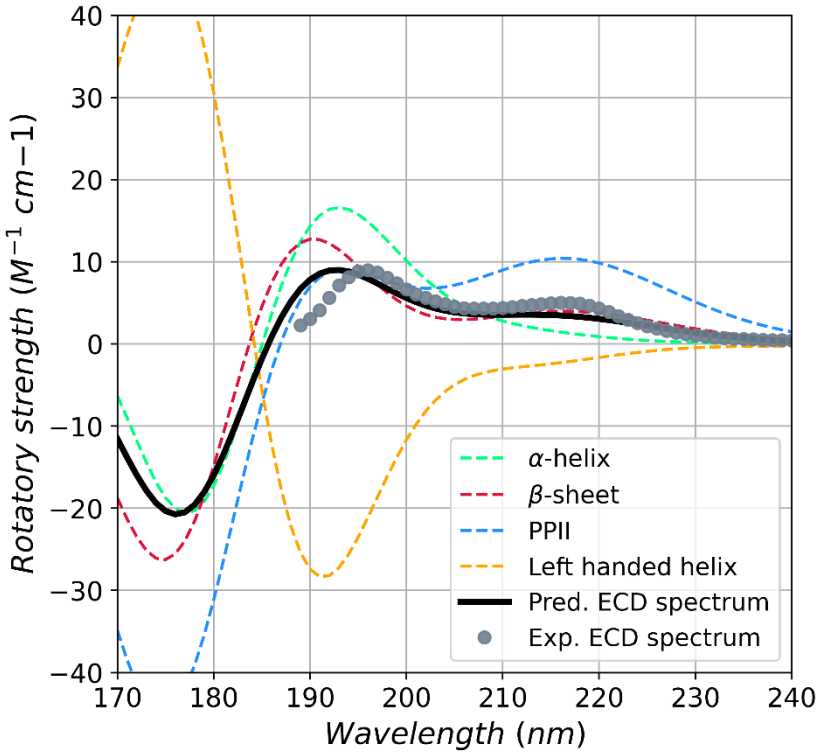


Figure V.10: *Predicted ECD spectra (solid black line) on the geometries generated with the OPLS-AA in a 2 µs simulation at 300K. The averaged spectra of each Ramachandra clusters (dashed line) are displayed: α-helix (in blue), β-Strand (in red), PPII (in green) and in left-handed conformations.*

Additionally, the average predicted spectrum for each defined area was plotted. At first glance, the results align with some experimental expectations. As anticipated, the peak at 216 nm is primarily attributed to the strong presence of the PPII population while the α-helix spectrum exhibits a pattern that is completely opposite to that of the left-handed helix. It is also noteworthy that the β-strand structure produces an ECD signal that is consistent with the experimental expectations. However, the α-helix structures exhibit an unusual ECD spectrum. For now, we have considered two explanations. The right-hand helix zone might be too large, leading to an overlap in the ECD spectrum. This idea is plausible, as we observe two distinct conformations in the Ramachandran diagram (**Figure V.9**), or the second conjecture, the geometries in the helix areas do not truly correspond to helices.

We have thus decided to use a different method to describe the conformations adopted by AFA as the identification of conformations based on areas displayed on the Ramachandran diagram seemed to be too much reductive. We have employed a different clustering method using de k-means algorithm. In this way, we can establish data-driven clusters which better reflect the actual distribution of the conformational space, rather than relying on predefined zones that may not fully capture the nuances of the data. By clustering based on the distribution of the φ_2 and ψ_2 dihedral angles, we can more precisely differentiate between the two helix clusters and verify if one of them produced an ECD spectrum similar than those expected for such secondary structure. Additionally, the structure with the lowest RMSD within each cluster is used to assess the relevance of the assigned secondary structures to these conformers.

Clustering based on dihedral angles using the k-means algorithm was thus tested with four (**Figure S11**), five (**Figure S12**) or six (**Figure S13**) clusters. With four clusters, the distribution and the corresponding ECD spectra are roughly similar to those obtained with the zone-based approach. However, the cluster corresponding to the PPII region is actually overlapping the β-strand area. When comparing this with the heatmap (**Figure V.9**) and the zones in **Figure V.6**, we observe that it forms a single cluster at the boundary between the PPII and β-strand regions. This also applies to higher number of clusters. As a result, zone-based discrimination is not appropriate because the cluster here is clearly on the boundary between the two secondary structure areas.

When using five or six clusters, the helix zone splits into the two expected sub-clusters, resulting in two markedly different spectra. However, neither spectrum resembles the known helix pattern. With six clusters, the intermediate zone between the right-handed helix and the

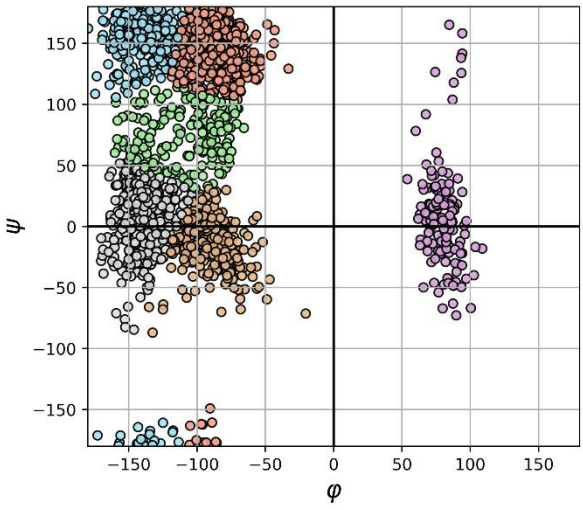
β-strand/PPII regions is represented by a cluster from which the ECD spectrum reflects a very different pattern. Thus, by dividing the clusters in the right-hand helix area into smaller clusters, which should cover less ECD patterns, none of these clusters exhibit the ECD spectrum typically expected of a classical helix. In contrast, the spectra for the β-strand and PPII zones have similarities with the expected patterns for these conformations.

We then checked whether the conformers matched the expected secondary structure assignments. Using the k-means algorithm on the φ_2 and ψ_2 dihedral angles of the central residue revealed little change in inertia after the identification of 6 clusters. We therefore extracted the conformers in this six-cluster distribution (see **Figure V.11**). The dihedral angles of the peptide backbone were reported in the **Figure V.11C** along with the proportion of each cluster. The proportions are 13 % (Cluster I), 13 % (Cluster II), 7 % (Cluster III), 9 % (Cluster IV), 32 % (Cluster V) and 26 % (Cluster VI).

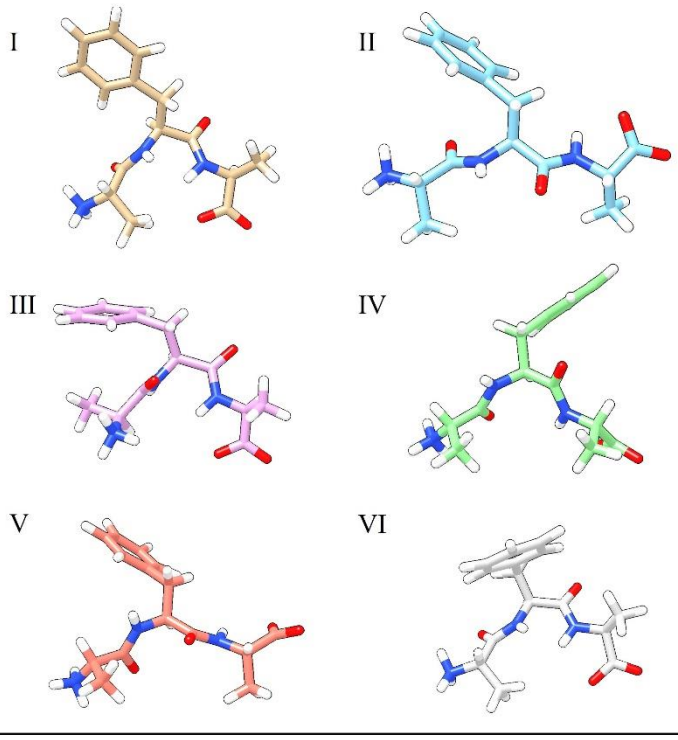
The conclusion is that the most populated cluster is actually the cluster V (32 %), located in an area rather close to the PPII zone, as predicted in the heatmap (see **Figure V.9**). At first glance, this may suggest a simple PPII conformation. However, the end angles are incompatible with this structure, aligning instead with those characteristics of an extended conformation.

Indeed, except for Cluster I, where the extended conformation is also observed at the ends, no other cluster replicates the central amino acid conformation at the terminal positions. Thus, associating a secondary structure with these conformations is not as straightforward as merely visualizing the position of the central amino acid on the Ramachandran plot.

A



B



C

Conformer	I	II	III	IV	V	VI
ψ_1 (°)	+ 166	+ 162	+ 161	+ 157	+ 158	+ 164
φ_2 (°)	- 101	- 146	+ 84	- 104	- 88	- 136
ψ_2 (°)	- 10	+ 166	+ 5	+ 83	+ 145	+ 4
φ_3 (°)	- 144	- 153	- 155	- 144	- 134	- 144
Proportion (%)	13	13	7	9	32	26

***Figure V.11**: **A**: Ramachandran plots of the AFA peptide by 2 µs MD simulations with the OPLS-AA force fields. Produced clusters are colored in beige (I), cyan (II), purple (III), light green (IV), light red (V) and white (VI). **B**: The conformer of each cluster in which carbons are colored in the cluster color. **C**: The dihedral angles of the peptide backbone for each conformer with the proportion of each cluster.*

V. B. 3. Discussion

A comparison with the literature (**Table V.1**) from a Ramachandran plot point of view with a simple description using the Ramachandran has shown that the study which has produced the closest results was the of the ref. **4,** in which the NMR experiments were carried out at pH 4 and the vibrational spectroscopies at pH 7. They found proportions of central residue located in the α-helix, β-strand and PPII areas of respectively 8-13 %, 31-50 % and 37-57 % whereas we found proportions of 31 %, 26 % and 19 %. The discrepancies can be explained by the fact that the proportion of helices might increase slightly with pH. Indeed, at pH 4, their NMR results are compatible with 13 % of the central residue located in the helix area. On the other hand, their VCD result at pH 7 does not allow them, as they noticed, to correctly predict the total proportion of phenylalanine in such area because the two populations of helices they found have opposite VCD signals. Moreover, it is compatible with the decrease of the $H_{N(Ala3)}$-$H_{\beta(Ala1)}$ and the $H_{N(Ala3)}$-$H_{\alpha(Phe2)}$ distances we found which are, as shown in **Figure V.4**, more compatible with a higher helix ratio for this residue and with a decrease of the extended and PPII conformations, with respect to their values. Besides, this was not interpretable on the ECD spectrum, until we observed that the spectra generated by these populations were very different from those expected.

For the other works, the results of ref. **3** have never been reproduced, either by ourselves by adding salt as described in their ECD legend, or by the same authors (ref. **5**) who, in their latest work, did not control the pH in the NMR tube. A simple dilution as they did gave us an acid pH, which may explain why their NMR results are very close to ref. **4**.

However, we have seen in this work that it is inappropriate to associate a secondary structure with a peptide as short and flexible as AFA. In this section, we want also to point out why small molecules should be reconsidered especially in term of spectroscopic signatures of their secondary structure. As a reminder, the typical ECD pattern of an α-helix displays three key features: a very intense positive peak around 190 nm, followed by two moderate negative peaks at 208 nm and 222 nm. However, the ECD spectrum we obtained for the α-helix region is significantly different, even after refining the clustering into increasingly smaller groups. We propose three main explanations for this discrepancy.

First, the two dihedral angles of the central amino acid may no longer dominate the ECD signal. Other structural elements, such as the two terminal residues and the aromatic side chain, which

also possess optical activity, may play a more significant role in shaping the ECD spectrum in this very small peptide.[44]

Second, the α-helix is traditionally defined by the presence of a stabilizing hydrogen bond between residues i and $i + 4$, which is impossible to form in a tripeptide like AFA. This hydrogen bond environment is particularly responsible for the intensity of the characteristic negative band at 222 nm.[45,46] Additionally, the length of the peptide chain has a substantial impact on the two other peaks of the ECD signal.[47] For a peptide as short as AFA, which does not even complete a full helix turn, the ECD spectrum could thus deviate from the typical α-helix pattern found in longer polypeptides.

Third, the other dihedral angles of the peptide backbone do not correspond to a helix structure. We therefore see no reason to associate the name helix with this structure.

Although the ECD spectrum for the helix region in AFA is unconventional, the presence of this conformational population enables us to simulate distances, coupling constants, and an ECD spectrum that are all in excellent agreement with the experimental data. Therefore, we believe there is no reason to expect the well-known 3-peak spectrum in this case, and that the simulated pattern for these dihedral angles appears plausible.

Finally, the last point to discuss is the clustering methods. It is also clear that Ramachandran's clustering by zone is insufficient, since 2 clusters are present in the helical area and the more populated cluster is located on the border between PPII and β-strand area. Clustering by centroids should therefore be preferred. In addition, it would appear that the dihedral angles of the central amino acid are not sufficient, so we could easily imagine including other geometric descriptors.

V. B. 4. Conclusion

Our study demonstrates that the AFA peptide adopts a mixture of conformations accurately simulated by the OPLS-AA force field. This finding suggests that OPLS-AA may be particularly well-suited for modelling flexible peptides, as it produces geometries whose calculated distances, coupling constant and simulated ECD spectrum align closely with the experimental measures. However, clustering by regions of the Ramachandran plot is insufficient for effectively differentiating the generated populations. Instead, k-means clustering is more appropriate and has produced a major cluster (Cluster V **Figure V.11**) which accounts for one third of the total population. Although the central amino acid in this cluster is

represented by a conformer in which the dihedral angles are fairly close to the PPII conformation, the other angles diverge significantly from this structure. The second predominant conformation observed in our simulations is the conformer VI (26 %), features a central amino acid in a helical form, while the terminal angles resemble those of a β-strand.

V. C. Conclusion et perspectives

Cette étude a permis de mettre en évidence plusieurs aspects importants. Tout d'abord, comme envisagé dans l'introduction de ce chapitre, nous avons confirmé que le peptide AFA adopte diverses conformations en solution. Cependant, contrairement à l'étude du chapitre précédent où les effets NOE étaient suffisants pour déterminer la structure en solution, ils ne sont ici pas assez discriminants, en raison de leur faible nombre. En effet, seules trois distances interprotons pertinentes pour le repliement de la chaîne principale ont pu être identifiées. De même, la constante de couplage mesurée et simulée ne correspond qu'au dernier résidu de la séquence, limitant son utilité pour une caractérisation complète.

En revanche, le DCE s'est révélé déterminant pour identifier les populations en solution. Grâce à des simulations précises de paramètres RMN et DCE, nous avons pu démontrer la pertinence du mélange conformationnel proposé. Les résultats obtenus valident l'approche expérimentale et théorique, confirmant que l'ensemble des conformations proposées reflète bien la réalité de la dynamique en solution du peptide AFA et répondant à notre objectif global d'utiliser une combinaison de méthodes pour déterminer la structure de petits peptides flexibles.

Les simulations montrent également que le champ de force OPLS-AA est particulièrement efficace pour générer des géométries proches de la structure réelle, en permettant de calculer avec précision les distances interatomiques, la constante de couplage et le spectre DCE toujours en accord avec les valeurs expérimentales. Cette efficacité semble être liée à la capacité de OPLS-AA à générer des géométries dispersées, visibles sur le diagramme de Ramachandran, contrairement au champ de force CHARMM27, qui, dans les deux peptides testés, tend à restreindre les conformations à des zones bien définies du diagramme. Il semblerait donc que OPLS-AA soit mieux adapté pour simuler des géométries de peptides très flexibles, tandis que CHARMM27 présente une tendance à limiter les conformations possibles.

Contrairement aux méthodes habituelles, nous ne cherchons pas à associer une structure secondaire spécifique au peptide AFA. En effet, nos travaux ont mis en lumière les limites de la classification traditionnelle des structures secondaires pour ce type de peptides flexibles et courts. Bien que les angles de l'acide aminé central puissent être localement proche d'une structure secondaire, en réalité très peu de conformères extraits montrent une répétition de ces angles. Par exemple, le conformère majoritaire dans le cluster V, qui représente environ un tiers des populations, est un hybride : il présente des angles dièdres proches de la conformation PPII en son centre, mais adopte une géométrie de type brin β aux extrémités. De même, le

conformère VI, représentant la seconde structure prédominante avec 26 % des géométries, montre au centre des angles dièdres proches de ceux des hélices droites, tandis que les extrémités adoptent des angles typiques des conformations étendues.

Ces résultats montrent que les catégories de structures secondaires usuelles et les classifications de coude (γ ou γ_{inv} pour trois résidus) ne suffisent pas pour décrire les conformations du peptide AFA. Nous avons identifié des structures stables qui mêlent les angles dièdres des grandes familles de structures secondaires, sans correspondre exactement à l'une d'elles.

Pour aller plus loin, il semble pertinent d'utiliser d'autres descripteurs que les seuls angles dièdres du résidu central pour partitionner les populations. En outre, le diagramme de Ramachandran ne semble pas parfaitement adapté pour cette méthodologie, car pour un tripeptide comme AFA, il ne peut prendre en compte que le résidu central, ce qui peut conduire à un mélange des clusters dans une même zone du diagramme, masquant les variations structurelles au sein de chaque population.

V. D. Références

1. Gazit, E. Self Assembly of Short Aromatic Peptides into Amyloid Fibrils and Related Nanostructures. *Prion* **1**, 32–35 (2007).
2. Eker, F., Griebenow, K., Cao, X., Nafie, L. A. & Schweitzer-Stenner, R. Preferred peptide backbone conformations in the unfolded state revealed by the structure analysis of alanine-based (AXA) tripeptides in aqueous solution. *Proc. Natl. Acad Sci.* **101**, 10054–10059 (2004).
3. Motta, A., Reches, M., Pappalardo, L., Andreotti, G. & Gazit, E. The Preferred Conformation of the Tripeptide Ala-Phe-Ala in Water Is an Inverse γ-Turn: Implications for Protein Folding and Drug Design. *Biochemistry* **44**, 14170–14178 (2005).
4. Pizzanelli, S. *et al.* Conformations of Phenylalanine in the Tripeptides AFA and GFG Probed by Combining MD Simulations with NMR, FTIR, Polarized Raman, and VCD Spectroscopy. *J. Phys. Chem. B* **114**, 3965–3978 (2010).
5. Bera, S. *et al.* Solid-state packing dictates the unexpected solubility of aromatic peptides. *Cell Rep. Phys. Sci.* **2**, 100391 (2021).
6. Tran, H. T., Wang, X. & Pappu, R. V. Reconciling Observations of Sequence-Specific Conformational Propensities with the Generic Polymeric Behavior of Denatured Proteins. *Biochemistry* **44**, 11369–11380 (2005).
7. Hagarman, A., Measey, T. J., Mathieu, D., Schwalbe, H. & Schweitzer-Stenner, R. Intrinsic Propensities of Amino Acid Residues in GxG Peptides Inferred from Amide I′ Band Profiles and NMR Scalar Coupling Constants. *J. Am. Chem. Soc.* **132**, 540–551 (2010).
8. Hernández, B., Pflüger, F., Kruglik, S. G. & Ghomi, M. Multiconformational analysis of tripeptides upon consideration of implicit and explicit hydration effects. *J. Mol. Graph. Model.* **102**, 107790 (2021).

9. Hsieh, M.-C., Liang, C., Mehta, A. K., Lynn, D. G. & Grover, M. A. Multistep Conformation Selection in Amyloid Assembly. *J. Am. Chem. Soc.* **139**, 17007–17010 (2017).
10. Williams, A. D., Shivaprasad, S. & Wetzel, R. Alanine Scanning Mutagenesis of Aβ(1-40) Amyloid Fibril Stability. *J. Mol. Biol.* **357**, 1283–1294 (2006).
11. Shaka, A. J., Lee, J. & Pines, A. Iterative Schemesfor Bibear Operators; Application to Spin Decoupling. *J. Magn. Reson. 1969* **77**, 274-293 (1988)
12. Marion, D., Ikura, M., Tschudin, R. & Bax, A. Rapid recording of 2D NMR spectra without phase cycling. Application to the study of hydrogen exchange in proteins. *J. Magn. Reson (1969)* **85**, 393–399 (1989).
13. Hwang, T. L. & Shaka, A. J. Water Suppression That Works. Excitation Sculpting Using Arbitrary Wave-Forms and Pulsed-Field Gradients. *J. Magn. Reson, Series A* **112**, 275–279 (1995).
14. Piotto, M., Saudek, V. & Sklenář, V. Gradient-tailored excitation for single-quantum NMR spectroscopy of aqueous solutions. *J. Biomol. NMR* **2**, 661–665 (1992).
15. Wüthrich, K., Billeter, M. & Braun, W. Polypeptide secondary structure determination by nuclear magnetic resonance observation of short proton-proton distances. *J. Mol. Biol.* **180**, 715–740 (1984).
16. Pettersen, E. F. *et al.* UCSF Chimera—A visualization system for exploratory research and analysis. *J.Comput. Chem.* **25**, 1605–1612 (2004).
17. Abraham, M. J. *et al.* GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **1–2**, 19–25 (2015).
18. Bauer, P., Hess, B. & Lindahl, E. GROMACS 2022 Manual. (2022) doi:10.5281/zenodo.6103568.

19. Kaminski, G. A., Friesner, R. A., Tirado-Rives, J. & Jorgensen, W. L. Evaluation and Reparametrization of the OPLS-AA Force Field for Proteins via Comparison with Accurate Quantum Chemical Calculations on Peptides. *J. Phys. Chem. B* **105**, 6474–6487 (2001).
20. MacKerell Jr., A. D., Banavali, N. & Foloppe, N. Development and current status of the CHARMM force field for nucleic acids. *Biopolymers* **56**, 257–265 (2000).
21. Mahoney, M. W. & Jorgensen, W. L. A five-site model for liquid water and the reproduction of the density anomaly by rigid, nonpolarizable potential functions. *J. Chem. Phys.* **112**, 8910–8922 (2000).
22. Hess, B., Bekker, H., Berendsen, H. J. C. & Fraaije, J. G. E. M. LINCS: A linear constraint solver for molecular simulations. *J. Comput. Chem.* **18**, 1463–1472 (1997).
23. Bussi, G., Donadio, D. & Parrinello, M. Canonical sampling through velocity rescaling. *J. Chem. Phys.* **126**, 014101 (2007).
24. Parrinello, M. & Rahman, A. Polymorphic transitions in single crystals: A new molecular dynamics method. *J. Appl. Phys.* **52**, 7182–7190 (1981).
25. Avbelj, F. & Baldwin, R. L. Role of backbone solvation and electrostatics in generating preferred peptide backbone conformations: Distributions of phi. *Proc. Natl. Acad Sci* **100**, 5742–5747 (2003).
26. Lloyd, S. Least squares quantization in PCM. *IEEE Trans. Inf. Theory* **28**, 129–137 (1982).
27. Menant, S., Tognetti, V., Oulyadi, H., Guilhaudis, L. & Ségalas-Milazzo, I. A Joint Experimental and Theoretical Study on the Structural and Spectroscopic Properties of the Piv-Pro-d-Ser-NHMe Peptide. *J. Phys. Chem. B* **128**, 6704–6715 (2024).
28. Migliore, M. *et al.* Characterization of β-turns by electronic circular dichroism spectroscopy: a coupled molecular dynamics and time-dependent density functional theory computational study. *Phys. Chem. Chem. Phys.* **22**, 1611–1623 (2020).

29. Aidas, K. *et al.* The Dalton quantum chemistry program system. *WIREs Comput. Mol. Sci.* **4**, 269–284 (2014).
30. Olsen, J. M., Aidas, K. & Kongsted, J. Excited States in Solution through Polarizable Embedding. *J. Chem. Theory Comput.* **6**, 3721–3734 (2010).
31. Olsen, J. M. H. & Kongsted, J. Chapter 3 - Molecular Properties through Polarizable Embedding. in *Advances in Quantum Chemistry* (eds. Sabin, J. R. & Brändas, E.) vol. 61 107–143 (Academic Press, 2011).
32. List, N. H., Olsen, J. M. H. & Kongsted, J. Excited states in large molecular systems through polarizable embedding. *Phys. Chem. Chem. Phys.* **18**, 20234–20250 (2016).
33. Olsen, J. M. H. & Reinholdt, P. PyFraME: Python framework for Fragment-based Multiscale Embedding. Zenodo https://doi.org/10.5281/ZENODO.4899311 (2021).
34. Stephens, P. J. & Harada, N. ECD cotton effect approximated by the Gaussian curve and other methods. *Chirality* **22**, 229–233 (2010).
35. Yanai, T., Tew, D. P. & Handy, N. C. A new hybrid exchange–correlation functional using the Coulomb-attenuating method (CAM-B3LYP). *Chem. Phys. Lett.* **393**, 51–57 (2004).
36. Vahtras, O. *et al.* Indirect nuclear spin–spin coupling constants from multiconfiguration linear response theory. *J. Chem. Phys.* **96**, 6120–6125 (1992).
37. Vosko, S. H., Wilk, L. & Nusair, M. Accurate spin-dependent electron liquid correlation energies for local spin density calculations: a critical analysis. *Can. J. Phys.* **58**, 1200–1211 (1980).
38. Becke, A. D. Density-functional exchange-energy approximation with correct asymptotic behavior. *Phys. Rev. A* **38**, 3098–3100 (1988).
39. Adamo, C. & Barone, V. Toward reliable density functional methods without adjustable parameters: The PBE0 model. *J. Chem. Phys.* **110**, 6158–6170 (1999).

40. Provasi, P. F., Aucar, G. A. & Sauer, S. P. A. The effect of lone pairs and electronegativity on the indirect nuclear spin–spin coupling constants in CH2X (X=CH2, NH, O, S): Ab initio calculations using optimized contracted basis sets. *J. Chem. Phys.* **115**, 1324–1334 (2001).
41. Cierpicki, T. & Otlewski, J. Amide proton temperature coefficients as hydrogen bond indicators in proteins. *J. Biomol. NMR* **21**, 249–261 (2001).
42. Smith, L. J. *et al.* Analysis of Main Chain Torsion Angles in Proteins: Prediction of NMR Coupling Constants for Native and Random Coil Conformations. *J. Mol. Biol.* **255**, 494–506 (1996).
43. Helgaker, T., Watson, M. & Handy, N. C. Analytical calculation of nuclear magnetic resonance indirect spin–spin coupling constants at the generalized gradient approximation and hybrid levels of density-functional theory. *J. Chem. Phys.* **113**, 9402–9409 (2000).
44. Chakrabartty, A., Kortemme, T., Padmanabhan, S. & Baldwin, R. L. Aromatic side-chain contribution to far-ultraviolet circular dichroism of helical peptides and its effect on measurement of helix propensities. *Biochemistry* **32**, 5560–5565 (1993).
45. Pelton, J. T. & McLean, L. R. Spectroscopic Methods for Analysis of Protein Secondary Structure. *Anal. Biochem.* **277**, 167–176 (2000).
46. Linhares, L. A. & Ramos, C. H. I. Unlocking Insights into Folding, Structure, and Function of Proteins through Circular Dichroism Spectroscopy—A Short Review. *Appl. Biosci.* **2**, 639–655 (2023).
47. Woody, R. W. Chapter 2 - Circular Dichroism of Peptides. in *Conformation in Biology and Drug Design* (ed. Hruby, V. J.) vol. 7 15–114 (Academic Press, 1985).

VI. Résultats - Prédiction de la constante de couplage $^{3}J_{HN-H\alpha}$ DFT par Machine Learning

Table des matières

VI. A. La constante de couplage

Le dernier objectif de cette thèse vise à améliorer la vitesse des calculs théoriques. Dans l'ensemble de nos travaux, la lenteur des simulations théoriques provient essentiellement des calculs TD-DFT, que nous avons effectués pour générer des spectres DCE, et des calculs DFT sur la constante de couplage $^{3}J_{HN-H\alpha}$ lors de l'étude du peptide AFA. Ces deux processus pourraient être améliorés par des modèles de Machine Learning (ML), permettant la prédiction de ces paramètres spectroscopiques.

La prédiction des spectres DCE par ML serait très intéressante à implémenter, mais relève d'une grande complexité. En effet, le problème n'est pas encore suffisamment bien posé et plusieurs questions restent encore en suspens quant aux objectifs auxquels le modèle de ML devrait répondre. Par exemple, est-ce que l'on veut être capable de générer seulement un motif, sans réellement se fier aux intensités, ni à la longueur d'onde à laquelle le motif apparait, où est-ce que, au contraire, on voudrait une estimation précise de l'intensité du spectre à une longueur d'onde donnée ? Également, on pourrait vouloir prédire le signe à une certaine longueur d'onde. Ces approches sont très différentes et nécessitent des modèles bien spécifiques.

Une autre question qui reste en suspens est quels sont les descripteurs d'entrée du modèle à choisir pour prédire ce spectre DCE. Une étude a montré qu'il était possible de simuler des spectres IR à l'aides d'un échantillonnage de la Fonction de Densité Radiale (FDR).[1] Nous avons essayé de reproduire ces résultats avec les spectres DCE en suivant la même méthodologie et un réseau de neurones similaire, sans succès. L'échantillonnage et le calcul de la FDR constituent un problème en soi, puisque la FDR inclut plusieurs paramètres et que toute la fonction n'est pas utile pour capturer l'information nécessaire à la prédiction du spectre DCE.

Nous nous sommes donc tournés vers la prédiction de la constante de couplage la plus discriminante dans l'analyse de la structure secondaire des protéines, qui a l'avantage d'être beaucoup plus simple à conceptualiser car il n'y a qu'une seule valeur à prédire contrairement au spectre DCE. Lors de nos travaux, nous avons pu observer que les calculs de la constante de couplage, pouvaient être particulièrement longs lorsqu'ils étaient effectués avec une grande base de type aug-cc-pVTZ-J.[2] Ces calculs sont d'autant plus longs que dans notre cas ils sont couplés à des simulations de dynamique moléculaire, qui produisent un nombre élevé de conformations, ce qui multiplie les calculs à effectuer. Cependant, ce type de base semble

nécessaire pour prédire avec précision les constantes de couplage, puisque dans le cas d'AFA, nous avons pu prédire une constante de couplage de 7.7 Hz pour une valeur mesurée de 7.5 Hz, au lieu de 6.3 Hz avec la base 6-311++G**.

Notre objectif est donc le suivant : prédire la constante de couplage $^3J_{HN-H\alpha}$ DFT à partir d'un modèle de ML. Le modèle pourrait ainsi permettre d'évaluer quasiment instantanément la constante de couplage moyenne d'un large ensemble de géométries, en adéquation avec la valeur expérimentale. Nos calculs MD et DFT prenant en compte l'environnement par l'inclusion polarisable, ce modèle serait donc adapté pour prédire la constante de couplage $^3J_{HN-H\alpha}$ de peptides en solution.

Dans le chapitre précédent, nous avons vu que cette grande base était très fortement corrélée à la base 6-311++G** ne prenant en compte que la contribution du contact de fermi (FC), pour le calcul de la constante de couplage $^3J_{HN-H\alpha}$ de l'alanine en position 3 de AFA. Cette seconde base, qui a permis d'accélérer les calculs de manière considérable, servira à générer l'ensemble de notre jeu de données. La corrélation sera également calculée sur la constante $^3J_{HN-H\alpha}$ de la sérine du peptide Piv-Pro-D-Ser-NHMe, et les simulations théoriques seront confrontées à la valeur expérimentale mesurée lors de ces travaux (7.7 Hz cf chapitre IV)

Puisqu'aucune introduction au machine learning n'a encore été faite lors cette thèse, la partie suivante y est consacrée afin de mieux appréhender les travaux présentés sous forme d'article dans la partie d'après. Les conclusions et les perspectives de ces travaux clôtureront ensuite ce chapitre.

VI. B. Introduction au Machine Learning (ML)

L'apprentissage automatique, aussi appelé « *Machine Learning* » (ML), est une branche de l'Intelligence Artificielle (IA) qui vise à donner aux systèmes informatiques la capacité d'apprendre et de s'améliorer automatiquement à partir de données, sans être explicitement programmés pour cette tâche.[3] Pour accomplir cela, le processus de ML repose sur plusieurs éléments :

- Un ensemble de données : qui servent de fondement à l'apprentissage en fournissant les informations nécessaires à l'entraînement du modèle. Elles sont généralement séparées en ensembles d'entraînement et de test pour évaluer la généralisation du modèle.
- Un modèle : Le modèle est la représentation mathématique choisie pour saisir les relations présentes dans les données. La complexité du modèle varie en fonction de la tâche, allant des modèles linéaires avec quelques paramètres jusqu'aux réseaux de neurones profonds pouvant contenir plusieurs milliers de paramètres.
- Une fonction de coût, aussi appelée fonction perte (« *loss function* »), dont l'optimisation fournira les valeurs des paramètres du modèle. Un algorithme d'optimisation, comme la descente de gradient, est ensuite utilisé pour minimiser cette fonction de coût et ainsi améliorer la performance du modèle.
- Une mesure de performance : la qualité du modèle est finalement évaluée selon des critères spécifiques à la tâche appelés métriques, tels que l'erreur quadratique moyenne pour les problèmes de régression ou la vraisemblance « *likelihood* » pour les classifications. Ces mesures de performance permettent de quantifier l'efficacité de l'algorithme.

Les méthodes de ML peuvent se diviser en trois grandes catégories : l'apprentissage supervisé, l'apprentissage non supervisé et l'apprentissage par renforcement.

VI. B. 1. Apprentissage supervisé

L'apprentissage supervisé est une méthode où l'algorithme est entraîné sur un ensemble de données contenant des variables d'entrée X (ou descripteurs) et des variables de sortie Y ou étiquettes) dans le but de modéliser la relation entre les deux. La relation trouvée entre X et Y est ensuite utilisée pour faire des prédictions sur de nouvelles données. Deux sous-types de problèmes principaux peuvent être résolus avec cette approche :

- Problèmes de régression : dans ce contexte, la variable de sortie est continue et le but est de prédire une valeur numérique.
- Problèmes de classification : ici, la variable de sortie est discrète et l'objectif est de prédire la classe à laquelle une observation appartient.

VI. B. 2. Apprentissage non supervisé

L'apprentissage non supervisé est une approche qui ne dispose que de variables d'entrée X, sans variable de sortie Y. L'objectif est de permettre à l'algorithme de découvrir de façon autonome des structures sous-jacentes dans les données. Ce type d'apprentissage est particulièrement adapté pour les tâches de partitionnement également appelé « *clustering* ».

- Clustering : le clustering regroupe les observations en sous-groupes homogènes appelés clusters, basés sur des similitudes entre les caractéristiques de données. Des algorithmes comme le k-means (utilisé au chapitre précédent) ou « *density-based spatial clustering of applications with noise* » (DBSCAN) sont couramment utilisés pour ce type de tâche.

VI. B. 3. Apprentissage par renforcement

Enfin, l'apprentissage par renforcement est une méthode où l'algorithme apprend de manière dynamique. Lors de la création du modèle de prédiction des variables Y à partir des X, un certain nombre de paramètres sont créés en plus, avec un poids standard. Le poids de chacun de ces paramètres est ensuite ajusté à chaque prédiction, en fonction de l'erreur commise. Ce processus est répété jusqu'à convergence de la performance.

VI. B. 4. Pré-traitement

Une fois le jeu de données et le type d'approche sélectionnés, les données doivent être traitées avant d'être utilisées dans le modèle. C'est l'étape de pré-traitement ou « *preprocessing* ». Le traitement le plus simple est la normalisation, qui consiste à redimensionner les caractéristiques des données pour assurer leur comparabilité. Elle est particulièrement importante si les échelles de données sont très différentes. Deux processus de normalisation ont été utilisés dans ces travaux. La remise à l'échelle ou « *rescaling* », aussi appelé « *min-max normalization* » qui consiste à donner à chaque x_i des X la valeur x_i' définie comme :

$$x_i' = \frac{x_i - min(x_i)}{max(x_i) - min(x_i)} \qquad \text{(VI.1)}$$

et la mise à l'échelle robuste ou « *robust Scaling* » où x_i' prend la valeur :

$$x_i{}' = \frac{x_i - Q_2(x_i)}{Q_3(x_i) - Q_1(x_i)} \quad \text{(VI.2)}$$

avec Q_1, Q_2, Q_3 les premiers, deuxièmes et troisième quartiles respectivement. Contrairement à la normalisation min-max, cette formulation permet d'être moins sensible aux valeurs aberrantes ou « *outliers* ».

Vient ensuite le choix du modèle. Dans cette partie, nous allons détailler uniquement les modèles testés lors de ces travaux. Supposons que nous avons un jeu de données de (X,Y). X est le jeu de données d'entrée, décrivant n échantillons avec p paramètres et Y sont les données de sortie des échantillons. Les modèles cherchent à trouver la fonction qui permet de passer de X à Y. La description du modèle seul n'est pas suffisante puisqu'il est indissociable dans la plupart des cas de la fonction de coût qui lui est associée. Ces fonctions seront donc également présentées dans cette partie.

VI. B. 5. Modèles linéaires

Dans ces modèles, la fonction de prédiction recherchée pour chaque échantillon i est sous la forme :

$$f(\vec{\beta}, \overrightarrow{x_i}) = \beta_0 + \beta_1 x_{i,1} + \cdots + \beta_p x_{i,p} \quad \text{(VI.3)}$$

La fonction de coût va alors prendre différentes formes en fonction du type de régression choisi.

- Régression linéaire

La fonction de coût à minimiser pour une régression linéaire est la somme des moindres carrés :

$$\min \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{i,j} \right)^2 \quad \text{(VI.4)}$$

- Régression ridge

La régression ridge ajoute une pénalité de norme $\mathcal{L}_2$ à la fonction de coût :

$$\min \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{i,j} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \quad \text{(VI.5)}$$

- La régression Lasso

La régression de type lasso ajoute une pénalité $\mathcal{L}_1$ à la fonction de coût

$$\min \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{i,j} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \qquad \text{(VI.6)}$$

Ces pénalités se visualisent sur la fonction coût (**Figure VI.1**) avec un cercle de contraintes pour la régression ridge et un losange pour la régression lasso en présence de deux paramètres β_0 et β_1. Ces deux paramètres ne peuvent pas dépasser cette limite, ce qui a tendance à faire de la méthode lasso une approche qui inhibe les paramètres de petites tailles. Ces deux paramètres ne peuvent pas dépasser cette limite, ce qui a tendance à faire de la régression lasso une approche qui inhibe les paramètres de petites tailles. En effet, elle tend à les faire converger vers zéro comme cela est illustré pour le paramètre β_0 sur la **Figure VI.1** à droite.

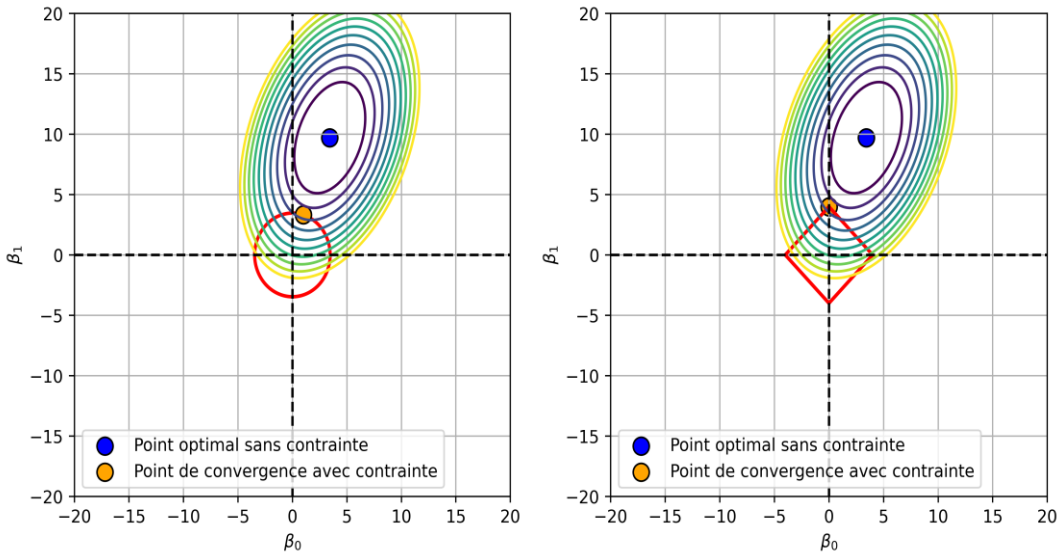


Figure VI.1 *: Schéma de visualisation des contraintes supplémentaires ridge (à gauche) et lasso (à droite) sur la fonction coût de deux paramètres β_0 et β_1. Le point bleu représente la convergence dans le cas d'une minimisation sans pénalité et le point orange celui avec la pénalité. La zone rouge représente la zone maximale de convergence de la fonction coût*

VI. B. 6. Support Vector Regression (SVR)

Dans les « *Support Vector Regression* » (SVR), qui font partie des modèles des « *Support Vector Machines* » (SVM), [4] l'objectif est de trouver une fonction de coût qui minimise l'erreur, tout en restant dans une marge ϵ autour des valeurs de y_i. Dans ces problèmes, on cherche à minimiser :

$$min \frac{1}{2} \|\beta\|^2 + C \sum_{i=1}^{n} (\zeta_i + \zeta_i^*) \qquad \text{(VI.7)}$$

avec comme contraintes :

$$\begin{cases} y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{i,j} \leq \epsilon + \zeta_i, \\ \sum_{j=1}^{p} \beta_j x_{i,j} + \beta_0 - y_i - \leq \epsilon + \zeta_i^*, \\ \zeta_i, \zeta_i^* \geq 0, \qquad \forall i \end{cases} \tag{VI.8}$$

Ce type de régression est particulièrement efficace pour ignorer les outliers présents dans l'ensemble de données, contrairement à la régression linéaire sous sa forme la plus simple (**Figure VI.2**).

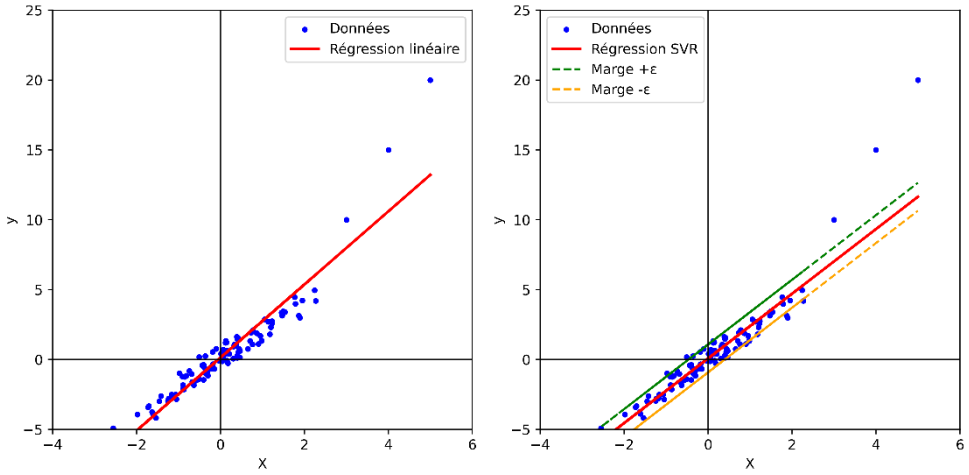


Figure VI.2 *: Différence entre une régression linéaire (à gauche) et une régression type SVR (à droite) en présence d'outliers.*

Sous la forme duale, la fonction de prédiction est :

$$f(x) = \sum_{i=1}^{n} (a_i - a_i^*) K(x_i, x) + \beta_0 \tag{VI.9}$$

qui permet de définir la fonction noyau $K(x_i, x)$, avec les multiplicateurs de Lagrange a_i et a_i^*. La définition du noyau permet notamment de transformer l'espace des données d'entrée, où la séparation linéaire est impossible, en un espace de plus grande dimension. En utilisant des noyaux, on peut effectuer des opérations de classification et de régression sur des données complexes, sans avoir besoin d'explicitement calculer les coordonnées dans cet espace de haute dimension, ce qui est souvent couteux en temps de calcul. Cette méthode permet ainsi de projeter les données $x_{i,j} \in \mathbb{R}^p$ dans un espace de dimension plus élevée, grâce à une fonction de transformation souvent notée $\phi(x_i)$ (**Figure VI.3**).

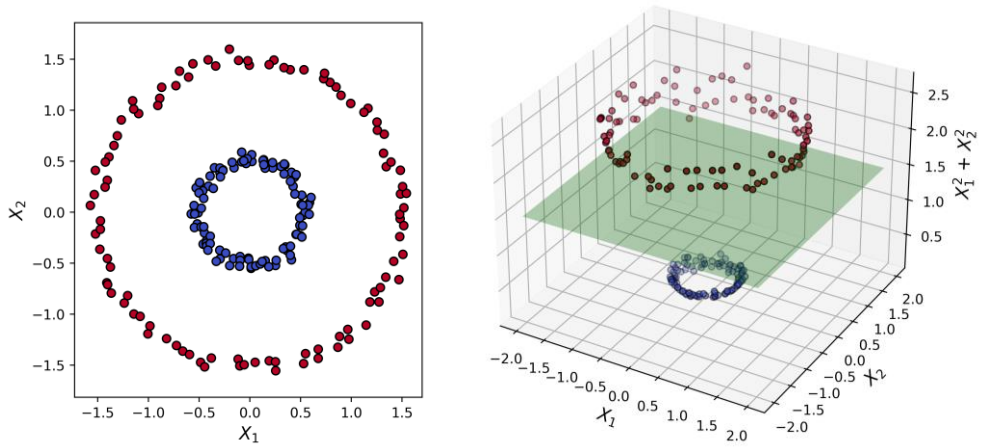


Figure VI.3 *: Représentation d'une classification non linéairement séparable (à gauche) entre les points rouges et les points bleus, puis devenue séparable par un hyperplan en vert (à droite) avec une dimension supplémentaire.*

Le choix du noyau impacte directement l'efficacité des modèles. Des exemples de noyaux incluent :

Les noyaux linéaires : $K(x_i, x_j) = x_i \cdot x_j$

Les noyaux polynomiaux : $K(x_i, x_j) = (x_i \cdot x_j + c)^d$

Les noyaux RBF (« *Radial Basis Function* ») : $K(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right)$

VI. B. 7. Régression ridge à noyau

La régression ridge à noyau est également utilisée pour traiter des problèmes qui ne sont pas linéairement séparables dans l'espace des dimensions d'entrée. La fonction de prédiction est alors :

$$f(x) = \sum_{i=1}^{n} a_i K(x_i, x) \qquad \text{(VI.10)}$$

et la fonction de coût à minimiser combine la régression linéaire avec la fonction coût de la régression ridge avec l'approche du noyau où f s'exprime selon l'équation (VI.10) :

$$min \sum_{i=1}^{n} \left(y_i - f(x)\right)^2 + \lambda \sum_{j=1}^{p} a_j^2 \qquad \text{(VI.11)}$$

VI. B. 8. Approche ensembliste

Les méthodes d'apprentissage d'ensemble, comme le « *bagging* » et le « *boosting* », visent à améliorer la performance des modèles de prédiction en combinant plusieurs modèles pour former des modèles plus compétents.

Le « *bagging* » consiste à entraîner plusieurs modèles indépendamment, sur des sous-ensembles d'échantillons construits par des tirages aléatoires avec remise dans le jeu de données original. Soit $D = \{(x_i, y_i)\}_{i=1}^{n}$ le jeu de données d'entraînement. Chaque modèle $f^{(b)}$ est entraîné sur un échantillon (« boostrap ») $D^{(b)}$ de taille n. La prédiction finale pour une entrée x est généralement obtenue par la moyenne des prédictions $\hat{y}^{(b)}$ sur le nombre de modèles B :

$$\hat{y} = \frac{1}{B}\sum_{b=1}^{B} \hat{y}^{(b)}$$

Un exemple bien connu de bagging est la régression par un modèle de forêt d'arbres aléatoires (« *random forest regression* »), où chaque arbre de décision est construit en sélectionnant aléatoirement des sous-ensembles.

Le « *boosting* » est une technique qui combine des modèles souvent moins robustes de manière séquentielle. Chaque modèle est ajusté en fonction des erreurs des modèles précédents, ce qui signifie que les observations mal prédites reçoivent plus de poids lors de l'entraînement du modèle suivant. Cela permet de corriger les erreurs de prédiction de manière itérative. La régression par gradient boosting est un exemple de type d'approche. Ce modèle construit des arbres de décision de manière séquentielle, où chaque arbre corrige les erreurs de prédiction des arbres précédents. L'algorithme minimise la fonction de perte et ajuste les prédictions de manière itérative, ce qui améliore la précision du modèle global.

VI. B. 9. Perceptrons multicouches

Dans le cadre de cette thèse, nous avons également utilisé des réseaux de neurones de type perceptrons multicouches (« *Multi Layer Perceptron* » ou MLP) pour modéliser la relation entre un vecteur d'entrée et une sortie scalaire. Le vecteur d'entrée x_i de dimension $\mathbb{R}^p$ (composé de p paramètres) est associé à une sortie y_i de dimension $\mathbb{R}^1$ (**Figure VI.4**). Un MLP constitué de h couches contenant k_j neurones est résolu par une fonction de prédiction f qui relie l'entrée x_i à la sortie y_i en plusieurs étapes :

Pour chaque couche j, les activations des neurones sont calculées selon les formules suivantes :

$$\begin{cases} z^{(j)} = W^{(j)}A^{(j-1)} + b^{(j)} \\ A^{(j)} = f_j\left(z^{(j)}\right) \end{cases} \quad \text{(VI.12)}$$

$W^{(j)}$ est la matrice de poids de la couche j, de dimension $k_j \times k_{j-1}$, et $b^{(j)}$ est le biais correspondant. Pour la première couche, $W^{(1)}$ est de dimension $k_1 \times p$. Les activations initiales de la première couche sont initialisées avec $A^{(0)} = x_i$

La fonction d'activation f_j est appliquée élément par élément à $z^{(j)}$ pour chaque couche j pour obtenir les activations $A^{(j)}$. La tangente hyperbolique est une fonction d'activation couramment utilisée et, pour le neurone l de la couche j, prend la forme :

$$A_l^{(j)} = f_j\left(z_l^{(j)}\right) = tanh\left(z_l^{(j)}\right) = \frac{e^{z_l^{(j)}} - e^{-z_l^{(j)}}}{e^{z_l^{(j)}} + e^{-z_l^{(j)}}}$$

La sortie finale y_i est obtenu par le même processus, à partir des activations de la dernière couche $A^{(h)}$ sur la combinaison linéaire une combinaison linéaire $z^{(h+1)} = W^{(h+1)}A^{(h)} + b^{(h+1)}$.

Ensuite, une fonction coût est appliquée de façon à évaluer l'écart entre la prédiction et la valeur réelle. Par exemple, l'erreur quadratique moyenne est souvent utilisée dans ce cas. Puis, les poids des perceptrons sont ajustés à l'aide d'un algorithme d'optimisation tel que la descente de gradient stochastique (SFD) ou l'algorithme Adam pour améliorer la prédiction. Ce processus est répété jusqu'à ce qu'un critère de convergence soit atteint, tel qu'un seuil de tolérance sur la fonction coût ou un nombre maximal d'itérations.

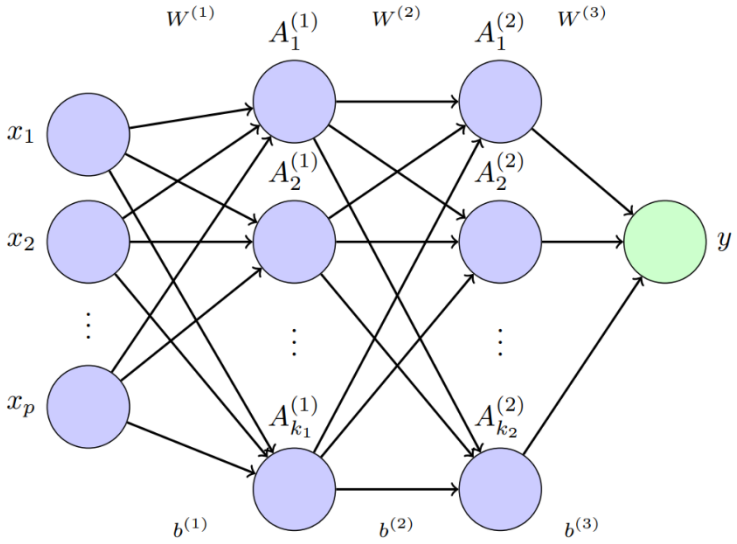


Figure VI.4 *: Représentation d'un réseau de neurones type MLP avec une couche d'entrée (x_i) de p paramètres, deux couches cachées de neurones avec k_1 et k_2 neurones, puis la sortie y avec un seul paramètre à prédire.*

VI. C. Article (Manuscrit)

VI. C. 1. Introduction

The structure of proteins in solution is commonly determined experimentally by Nuclear Magnetic Resonance (NMR).[5] These analyses involve measuring the ${}^3J_{HN-H\alpha}$ coupling constant, which is representative of an average in the case of rapid conformational exchanges. This constant can either be interpreted directly to characterise the local structure of the amino acid concerned or used as a constraint in Molecular Dynamic (MD) simulations to maintain a conformation in agreement with the experimental value during MD simulation and extract a more global structure.[6]

In proteins, the φ and ψ angles describe the dihedral angles around the C_α-N and C_α-C bonds in the peptide backbone, respectively, which are useful to identify the secondary structure. But ${}^3J_{HN-H\alpha}$ coupling constant is actually described by the θ dihedral angle which is related to φ (**Figure VI.5**) by:[7,8]

$$\begin{cases} \theta = |\varphi + 60|, & for\ D\ residue \\ \theta = |\varphi - 60|, & for\ L\ residue \end{cases} \qquad \text{(VI.13)}$$

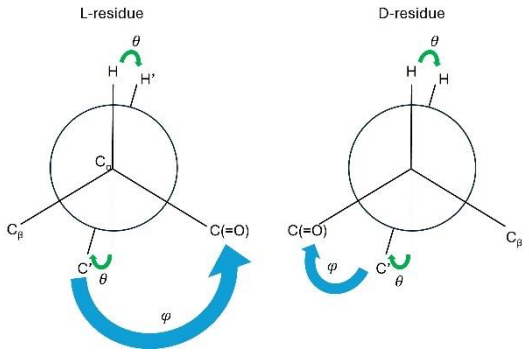


Figure VI.5*: Schematic representation of φ and θ dihedral angles through the Newman projection of the C_α-N bond.*

The mathematical relationship between dihedral angles θ of vicinal hydrogens and their NMR scalar coupling constants, has historically been described by the Karplus equation. Originally introduced in 1959, it provides a relationship between θ and the observed ${}^3J_{HH}$ coupling constant separated by three bonds according to:

$${}^3J_{HH} = Acos^2(\theta) + Bcos(\theta) + C \qquad \text{(VI.14)}$$

where θ is the dihedral angle between the two successive planes formed by the four atoms bonded sequentially, and A, B, and C are empirically derived parameters that depend on the

considered molecule. This equation has been widely used both in protein and other organic systems, offering a simple and practical tool to link the molecular structure with NMR parameters.

However, the use of the Karplus equation generates a significant degree of approximation. Indeed, the equation does not explicitly consider the electronic effects surrounding the nuclei, or the influence of local structural variations, such as hydrogen bonding, solvent effects, and other non-covalent interactions that can alter the coupling constants.[9] Although some of these effects could be indirectly taken into account in the parameters, they are often found by experiments on constrained molecules, generally very different from the molecule on which the equation is used. There are large number of sets of coefficients in the literature and it is therefore a complex issue to know which ones are suitable for a specific molecule.[10–16]

To overcome these limitations, quantum chemical methods, particularly those based on Density Functional Theory (DFT), have been developed to compute this indirect spin-spin scalar coupling constants. Unlike the Karplus equation, DFT is not limited to proton-proton couplings. Indeed, it can calculate coupling constants between any pair of atoms by considering the full electronic structure of the molecule. The calculation of scalar couplings using DFT is generally grounded in Ramsey's equations,[17] which provides a theoretical framework for understanding how the electronic environment influences the spin-spin coupling between nuclei.

However, the computational cost of DFT remains a significant drawback. Calculating scalar couplings for large systems, such as entire proteins or thousands of peptide conformations, is computationally intensive and time-consuming. This limitation restricts the practical application of DFT to small systems or selected regions of larger molecules, making it challenging to apply these methods on a large scale.

In response to these challenges, Machine Learning (ML) has emerged as a promising approach to predicting NMR parameters efficiently while maintaining high accuracy.[18–21] ML techniques are well-suited for capturing complex, non-linear relationships in data. By training models on large datasets, ML can learn to predict coupling ${}^{3}J_{HC-CH}$ constants rapidly and accurately suggesting that this could be applied to the ${}^{3}J_{HN-H\alpha}$ coupling constant with direct application to protein folding.[22]

The use of ML in predicting scalar coupling constants offers several advantages. ML models can incorporate a wide range of molecular features of different natures. For example, they can account for geometric parameters, like those in the Karplus equation, as well as electronic

parameters. Once trained, ML models can make rapid predictions for new molecules or conformations, making them ideal for high-throughput applications and large-scale studies.

Nevertheless, several challenges remain. One key issue is the need for extensive and diverse training datasets that represent a wide range of molecular geometries. Molecular Dynamics (MD) simulations provide a solution by generating conformational snapshots. DFT-calculated coupling constants from these structures can then be used to train various ML models.

Moreover, we want the predicted DFT values to agree with the experimental ones. In a previous study, we showed that it was possible to predict the $^3J_{HN-H\alpha}$ constant with high accuracy for the AFA peptide in water, a medium in which the peptide folds as a conformational mixture. These DFT calculations included the CAM-B3LYP functional coupled to the Dunning-derived base set, aug-cc-pVTZ-J,[2] which is optimized for NMR calculations but very time consuming. The accuracy of this calculations was improved by taking into account solvation effects through the polarisable embedding.

This same study, based on MD simulations and in which many geometries were extracted, proposed to generate much more ^3J values possible using a basis that is less suitable for this type of approach, 6-311++G**, and taking only the contribution of the Fermi Contact (FC). This decision was initially motivated by a time saving of a factor of 100 between the two types of calculation, and the fact that the FC contribution often largely dominates this type of interaction. Besides, the correlation we found between the two basis sets allows us to simulate the results of the large basis on the whole data set to finally recover the experimental value.

In this present work, we aim to develop an approach for predicting the $^3J_{HN-H\alpha}$ coupling constant by combining the accuracy of DFT calculations with the efficiency of ML. In order to use a maximum of diversified data, we used data from the MD of AFA, to which we added geometries of the Piv-Pro-D-Ser-NHMe peptide from a previous study.[23] However, as the latter peptide structures very quickly in water, greatly limiting the range of ^3J constant covered, we also used other shorter MDs of this peptide at different starting points.

After reparametrizing the correlation between the two basis sets on the whole geometry dataset (1000 for AFA, 3000 for Piv-Pro-D-Ser-NHMe), we checked the agreement between the ^3J coupling constant from the original MD of the Piv-Pro-D-Ser-NHMe peptide (1000 geometries), with the value experimentally measured previously by NMR. Then, all the geometries and their associated coupling constant, i.e. the ones calculated with the 6-311++G** basis set and considering solely the FC contribution, were used in the ML models.

Our work focuses on generating suitable descriptors from MD simulations of the Piv-Pro-D-Ser-NHMe and the Ala-Phe-Ala peptides. Several sets of descriptors were tested taking into account local and less local geometric descriptors of the ^3J coupling constant, as well as nuclear descriptors. This research has implications for structural biology and NMR spectroscopy, where accurate and efficient predictions of NMR parameters are crucial. Moreover, our findings contribute to the broader field of ML in quantum chemistry and molecular modelling.

VI. C. 2. Computational Details

VI. C. 2. a) Generated geometries

In this study, a total of 4000 geometries were used, organized into four sets of 1000 geometries each. Three of these sets were derived from the Piv-Pro-D-Ser-NHMe peptide while the remaining set was based on the Ala-Phe-Ala peptide:

- Piv-Pro-D-Ser-NHMe: 1000 geometries were extracted from a 1 µs simulation. These geometries were also used to check the adequacy with the measured values of 7.7 Hz. As this peptide is highly structured, 2000 other geometries were generated using 50 µs simulations with two different starting points to obtain a wide range of coupling constants.
- AFA: 1000 geometries were extracted from a 1 µs simulation. As the peptide is highly flexible, a wide range of constants had already been produced.

The geometry creation was performed with the Gromacs 2022 software,[24] with the initial structures constructed with Chimera.[25] More details are available in the Supporting Information.

VI. C. 2. b) Descriptor selection

Three sets of descriptors were defined for our analysis:

ML1: This set utilizes only the cosine and squared cosine of the θ dihedral angle (see **Figure VI.6**) as descriptors. It allows for a comparison between a linear regression model, equivalent to the Karplus equation, and other models.

ML2: This set is based on purely geometric descriptors focusing on the dihedral angle θ ($H_1C_1NH_2$), and extend to the nearest neighbouring atoms ($C_3C_2H_1C_1NH_2C_4$). It includes:

- 7 interatomic distances H_1H_2, H_1C_1, C_1N, NH_2, C_2C_1 C_3C_1 and C_4N.
- 3 angles H_1C_1N, C_1NH_2 and $C_2C_1C_3$, with their cosines.

- 4 dihedral angles $H_1C_1NH_2$, $C_2C_1NH_2$, $C_3C_1NH_2$, $H_1C_1NC_4$ with both cosines and the sines. The squared cosines of the θ dihedral angle ($H_1C_1NH_2$) were also computed.

This set comprises a total of 19 descriptors

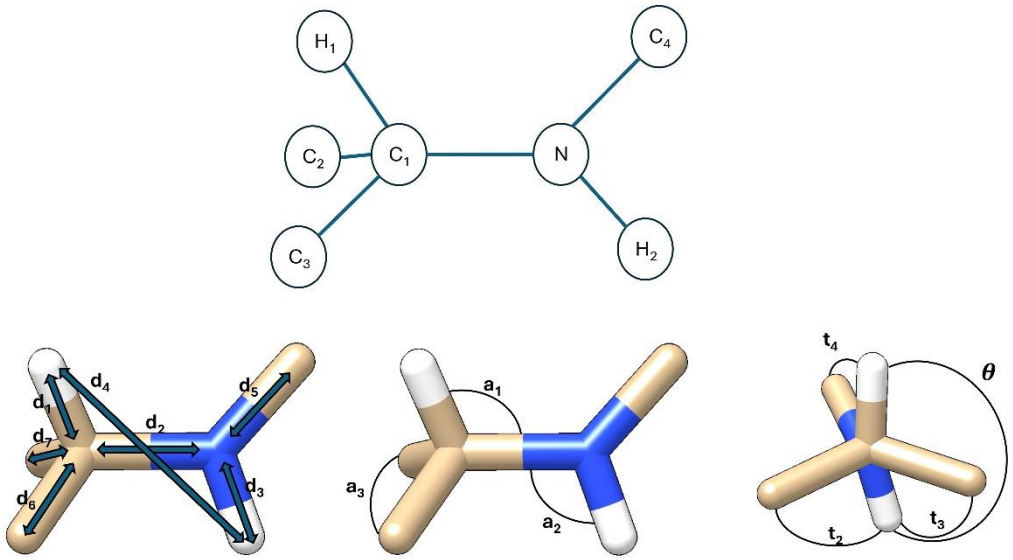


***Figure VI.6**: Schematic representation of geometrical descriptors chosen on the $C_3C_2H_1C_1NH_2C_4$ sequence. The seven distances (left) the three angles (middle) and the four dihedral angles used are displayed.*

ML3: This set extends to ML2 by including descriptors derived from the Coulomb matrix.[26] This matrix is a one of the many ways often used in ML to encode chemical information. It captures both the chemical composition and information about the geometric structure of a molecule without the need for additional descriptors.[27] It is useful in ML models, particularly for predicting molecular properties such as electronic ground an excited states propreties.[28,29] The matrix is invariant to translations and rotations. Sorting the eigenvalues also results in permutation invariance.[30] These invariances make it possible to describe each geometry in a unique way. The Coulomb matrix M_{ij}^{C} was computed for the sequence $C_3C_2H_1C_1NH_2C_4$:

$$M_{ij}^{C} = \begin{cases} 0.5Z_i^{2.4} & for\ i = j \\ \dfrac{Z_iZ_j}{R_{ij}} & for\ i \neq j \end{cases} \qquad \text{(VI.15)}$$

where Z represents atomic numbers and R_{ij} the interatomic distances. Diagonalizing these matrices generated 7 eigenvalues, which we sorted in ascending order (denoted λ_i in the following) and then added to the ML2 descriptors to produce a total of 26 descriptors.

VI. C. 2. c) Coupling Constant Calculations

The reduced nuclear spin-spin indirect coupling tensor $\underline{\underline{K}}_{K,L}$ between two atoms K and L can be computed through DFT from a second-order Taylor expansion of the energy according to:

$$\underline{\underline{K}}_{K,L} = \frac{\partial^2 E}{\partial M_K \partial M_L} \tag{VI.16}$$

where M_K and M_L are the magnetic moment of nuclei K and L.[31] The direct contribution, the classical dipolar coupling tensor, have been ignored since in an isotropic medium, rapid rotation results in the cancellation of traceless tensors.[32] NMR measurements are more directly related to the indirect coupling tensor $\underline{\underline{J}}_{K,L}$ than the reduced one, $\underline{\underline{K}}_{K,L}$ and it is defined as:

$$\underline{\underline{J}}_{K,L} = h \frac{\gamma_K}{2\pi} \frac{\gamma_L}{2\pi} \underline{\underline{K}}_{K,L} \tag{VI.17}$$

with h the Planck constant, γ the magnetogyric ratios and which leads to the measured value:

$$J_{K,L} = \frac{1}{3} Trace\left(\underline{\underline{J}}_{K,L}\right) \tag{VI.18}$$

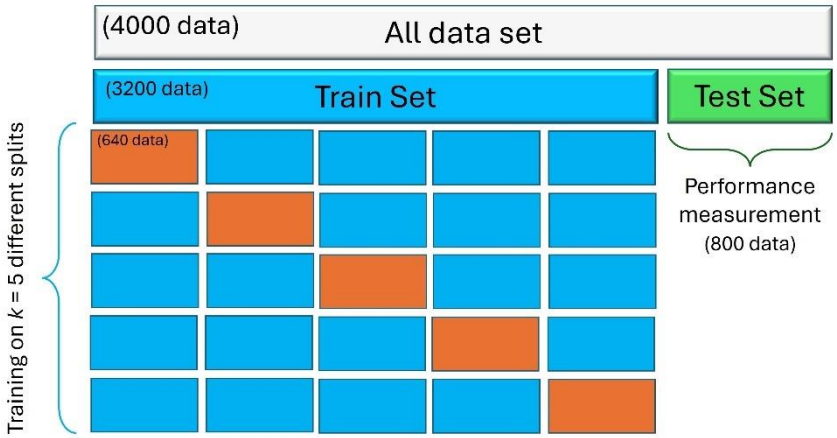
The Ramsey equations enable to define four contributions: the diamagnetic spin-orbit (DSO), the paramagnetic spin-orbit (PSO), the Fermi-contact (FC), and the spin-dipole (SD) operators.[33] All calculations were performed using the CAM-B3LYP[34] functional in the DALTON2018 program,[35] coupled with two types of basis sets: aug-cc-pVTZ-J[2] with all contribution and 6-311++G** with solely the FC contribution.

As explained in the introduction, the 6-311++G** basis with the FC contribution calculations were performed to generate many ^3J coupling constant values to be used as scalar output in our ML models. Besides, the correlation with the larger basis set was reparametrized and the regression is applied to confirm consistency with the experimental measured value (7.7 Hz) on the 1 µs MD dataset. The correlation was achieved by extracting 50 geometries, of a regular time step, in which both basis sets were used.

VI. C. 2. d) Machine learning models

A variety of ML models were tested: linear, ridge, lasso, random forest, gradient boosting, support vector, kernel ridge, gaussian process and multilayer perceptron regressions. Initial model were computed using the sci-kit learn library[36] with the default initial parameters, followed by hyperparameter optimization (see Supporting Information). A five-fold cross-validation (**Figure VI.7**) was performed using 3200 geometries (80 % of the dataset). The

principle of this method is to divide the dataset into five equal parts and, at each iteration, use four parts to train the model and the fifth to validate it. This is repeated five times so that each part serves as validation once. Performance was evaluated on the test sets using the model trained on each fold and then average to obtain a measure of the model's effectiveness. The standard deviation between performances was used as an error bar. The normalization was achieved using a robust scaler, in which the x_i descriptors were transformed to



$(x_i - Q_2)/(Q_3 - Q_1)$ with Q_j the j^{th} quartile. Before the single draw for the training set, all the data set is randomly shuffled.

Figure VI.7*: Schematic representation of a 5 five-fold algorithm. Each split of the global train set is divided into four train sets (blue) and a validation test (orange). The test set (green) is used to measure the performance.*

VI. C. 2. e) Scoring

Model performance was evaluated using the Mean Absolute Error (MAE), the Root Mean Square Error (RMSE) and the coefficient of determination R^2, calculated on test sets and on the whole dataset.

VI. C. 3. Results and Discussions

VI. C. 3. a) Basis set correlation

We initially took advantage of the ${}^3J_{HN-H\alpha}$coupling constant predictions for the AFA peptide from the prior study, which demonstrated that the ${}^3J$ values calculated with 6-311++G** basis set with only the FC contributions, were fully correlated with those from the aug-cc-pVTZ-J set, accurately predicting the experimental ${}^3J$ coupling constant. The choice of basis change is motivated by the large cost saving of taking both a People basis set and only the FC contribution, in order to generate as much data as possible for the ML Models. We observed an

10-fold reduction in computation time by switching to People basis set, and a 10-fold reduction by taking solely the FC contribution.

To ensure that this correlation is consistent on other peptides, we added coupling constants computed on the Piv-Pro-D-Ser-NHMe peptide geometries with both basis sets. As this peptide is highly structure in solutions, two additional very short MD additional MD were performed at different starting points from the one use of the 1µs MD simulation from the previous study, leading to a total of 200 points. The results are shown in **Figure VI.8** and confirm the high correlation between the two basis sets.

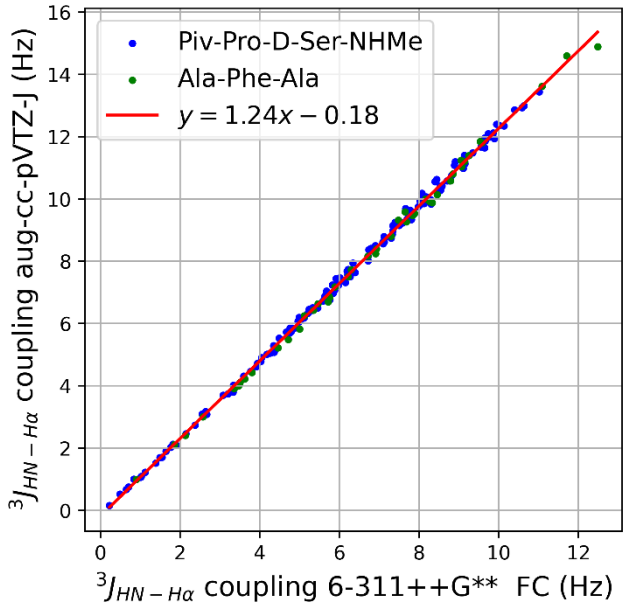


***Figure VI.8**: Validation of coupling constant prediction with the measured values and validation of the correlation between the aug-cc-pVTZ-J and 6-311++G** basis sets (MAE=0.10 Hz, R^2=0.998).*

We then verified if the 1000 coupling constant from the geometries of the Piv-Pro-D-Ser-NHMe peptide from the optimized 1 µs MD simulation (which reproduce the experimental conformation) and computed with the 6-311++G** with the FC contribution can predict the measured value of 7.7 Hz. The uncorrected 6-311++G** values (6.3 Hz for AFA and 6.4 Hz for Piv-Pro-D-Ser-NHMe) are not accurate enough in both peptides, but the use of the correlation with the aug-cc-pVTZ-J allow to simulate $^3J_{HN-H\alpha}$ constants (7.7 Hz for AFA and 7.7 Hz for Piv-Pro-D-Ser-NHMe) very closely the measured values (7.5 Hz for AFA and 7.7 Hz for Piv-Pro-D-Ser-NHMe) in both cases.

The results indicate a consistent correlation across different peptides and on a large range of ^3J coupling constant, which allow to reproduce the experimental in both peptides. Consequently,

the 4000 coupling constants calculated with the slower basis set were employed in our machine learning models.

VI. C. 3. b) Karplus Descriptors (ML1)

The Karplus equation was established to predict the ^3J coupling constants between two protons, based on the local geometry of the dihedral angle θ. Given this, we initially evaluated the performance of a simple Karplus-like model to predict the DFT calculated values. For this evaluation, we thus chose to start with only two geometric descriptors $cos^2(\theta)$ and $cos(\theta)$. We were interested in seeing whether a model other than Karplus's would predict better with the same descriptors.

Surprisingly, with just these two descriptors, the linear model performed as well as more complex models (**Figure VI.9**), achieving a MAE of 0.87 Hz and a RMSE of 1.13 Hz. Further analysis revealed some other limitations on those models. When comparing predictions from the linear model to DFT-calculated values (**Figure VI.10**) a clear pattern, of deviations merged. As the DFT $^3J_{HN-H\alpha}$ values increased, the predictions exhibited increasing divergence, with errors ranging from + 4.2 Hz to - 4.0 Hz, indicating poor reliability for extreme values. Additionally, prediction limits appeared, particularly for large (> 9.7 Hz) and small (< 0.7 Hz) ^3J coupling constants, suggesting that the model struggled to make accurate predictions outside a mid-range interval. This problem was not isolated to the linear model, as a similar behaviour was observed with other tested models (Figure S3). The ML1 set, although simple, was thus found to provide inaccurate predictions due to the limited number of descriptors considered.

VI. C. 3. c) Addition of geometric descriptors (ML2)

To address the shortcomings of ML1, we moved to ML2 which integrates more geometric descriptors. By adding 17 geometric descriptors, the prediction performance improved significantly across all tested models (**Figure VI.11**). This is particularly true for non-linear models, which benefited more from the additional data complexity. For instance, the Kernel Ridge Regression (KRR) with a polynomial kernel of degree two, achieved an MAE of 0.45 Hz with a polynomial kernel, while a linear regression model only improved to 0.55 Hz. Notably, some models, like random forests and gradient boosting, displayed a marked discrepancy between the MAE on the training set and the test set, hinting at slight overfitting. These findings emphasize that while more descriptors help, they also increase the risk of overfitting, particularly in tree-based models.

Further improvement in the prediction was evident when comparing to DFT values, as seen in the closer alignment of the prediction curve (**Figure VI.12**). The error distribution also shifted to a more satisfactory pattern, forming a single Gaussian distribution centered around zero, with maximal errors ranging from -1.8 Hz to +1.7 Hz. Thus, the incorporation of 16 other geometric descriptors significantly increases the prediction efficiency of the coupling constant for all models, particularly the kernel ridge regression.

VI. C. 3. a) Addiction of MC descriptors (ML3)

Given the success of ML2 we decided to integrate nuclear descriptors to further refine the prediction accuracy. We have decided to incorporate the sorted eigenvalues of the Coulomb matrix for each structure to the descriptor set. Since the Coulomb matrix is calculated for the four atoms of the θ dihedral angle and their first neighbours, this adds a total of 7 descriptors to the previous model. The inclusion of these nuclear descriptors significantly boosted the prediction accuracy for all models, with the KRR standing out: it reached an MAE of 0.32 Hz and an RMSE of 0.4 Hz (**Figure VI.13**).

In this model, the predictions are even more closely aligned with the DFT values. The error distribution is tighter, displaying a single, clear Gaussian profile with errors ranging from -1.3 Hz to + 1.4 Hz (**Figure VI.14**). This indicates that the model is highly effective at predicting coupling constants and the inclusion of nuclear descriptors has significantly enhanced its accuracy.

However, analysis of the correlation matrix revealed that many descriptors are correlated, which is not unexpected given the chosen set of descriptors (**Figure VI.15**). The more correlated descriptors are λ_1, λ_2, $cos(\theta)$, d_4 and $cos(t_4)$ and are all strongly correlated together, either positively (λ_1/d_4, $\lambda_1/cos(t_4)$, $\lambda_2/cos(\theta)$ and $d_4/cos(t_4)$ or negatively (λ_1/λ_2, $\lambda_1/cos(\theta)$, λ_2/d_4, $\lambda_2/cos(t_4)$, $cos(\theta)/d_4$ $cos(\theta)/cos(t_4)$). These five descriptors therefore present a redundancy of information, which could at best be neutral, but at worst harm the performance of our models. It is therefore essential to select descriptors without too much multicollinearity.

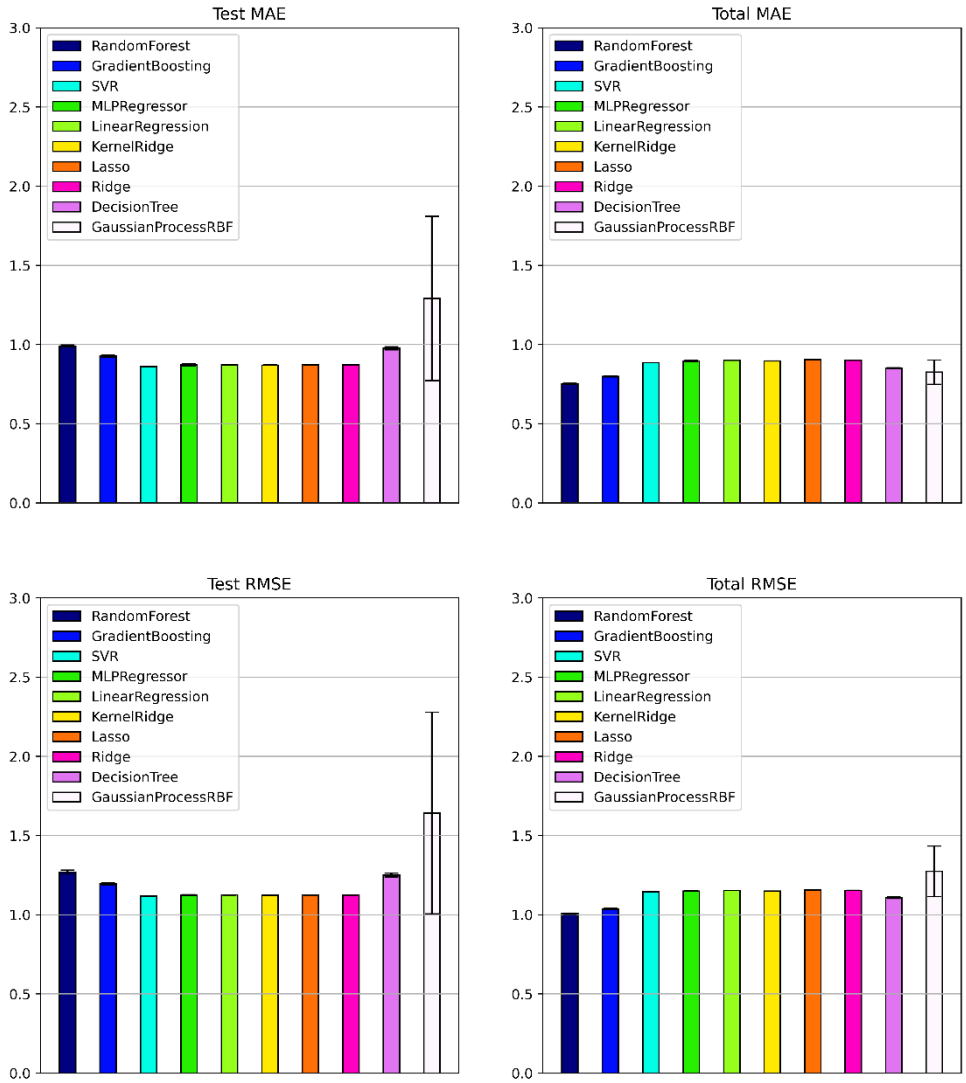


***Figure VI.9**: Performance of all tested model on the ML1 set of descriptors. The MAE (up) and the RMSE (down) were calculated on the test set (left) and the total set (right)*

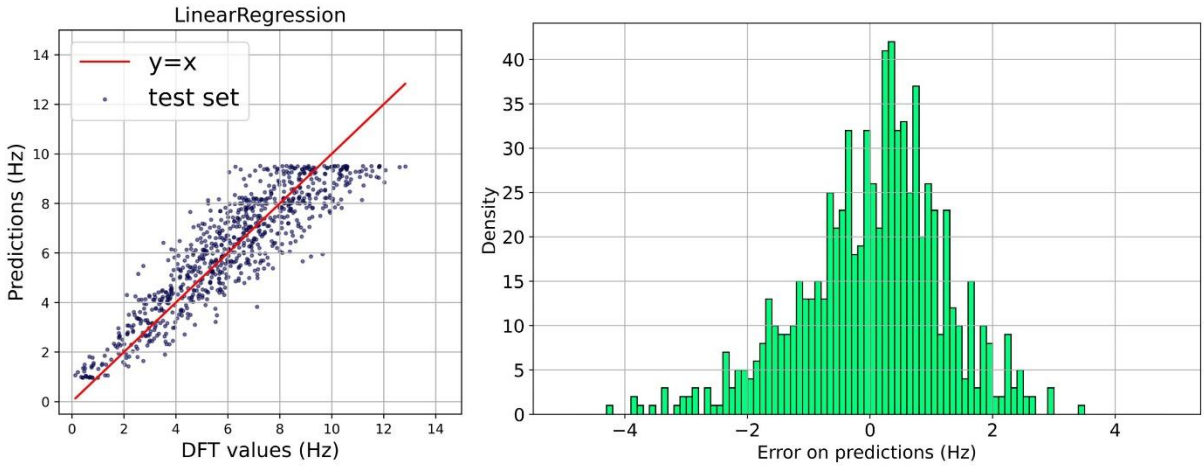


***Figure VI.10**: Predicted values against the DFT values (6-311++G** basis set with the FC contribution only) with the linear regression (left) and the associated error distribution (right) with the ML1 set of descriptors*

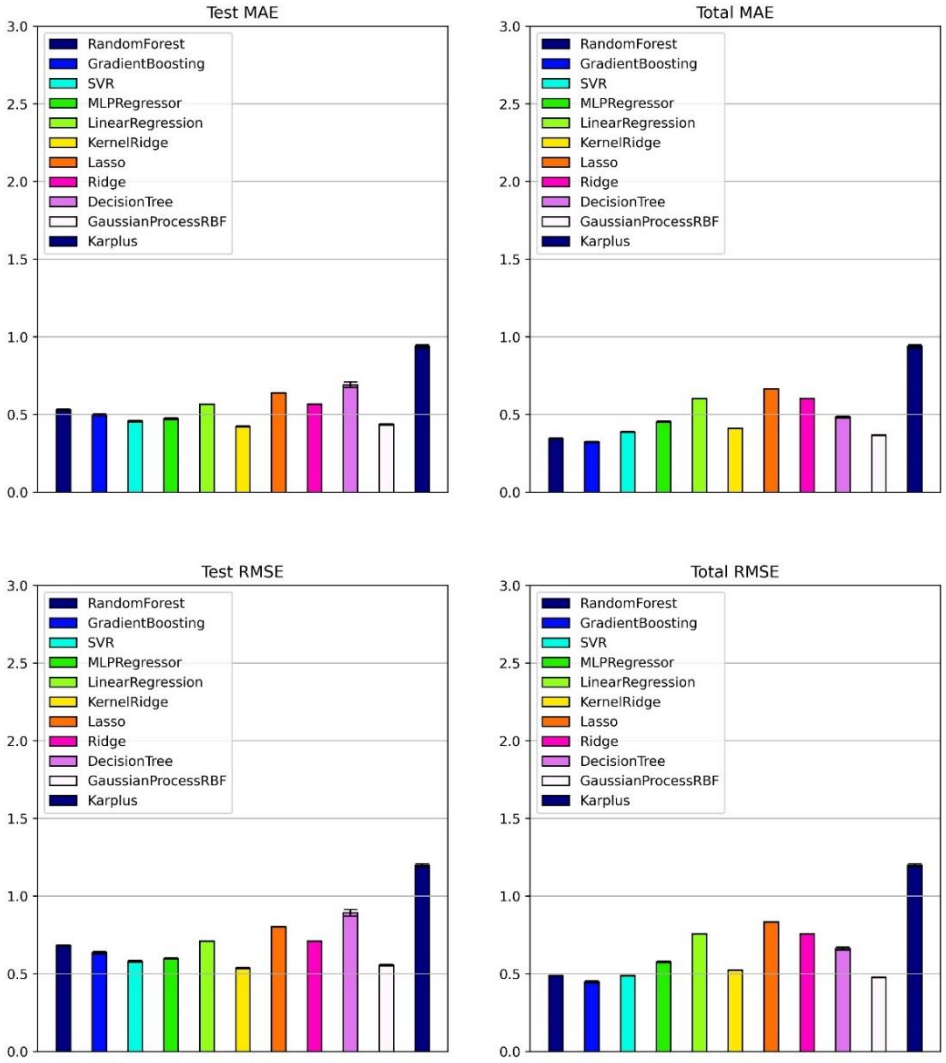


Figure VI.11*: Performance of all tested model on the ML2 set of descriptors. The MAE (up) and the RMSE (down) were calculated on the test set (left) and the total set (right)*

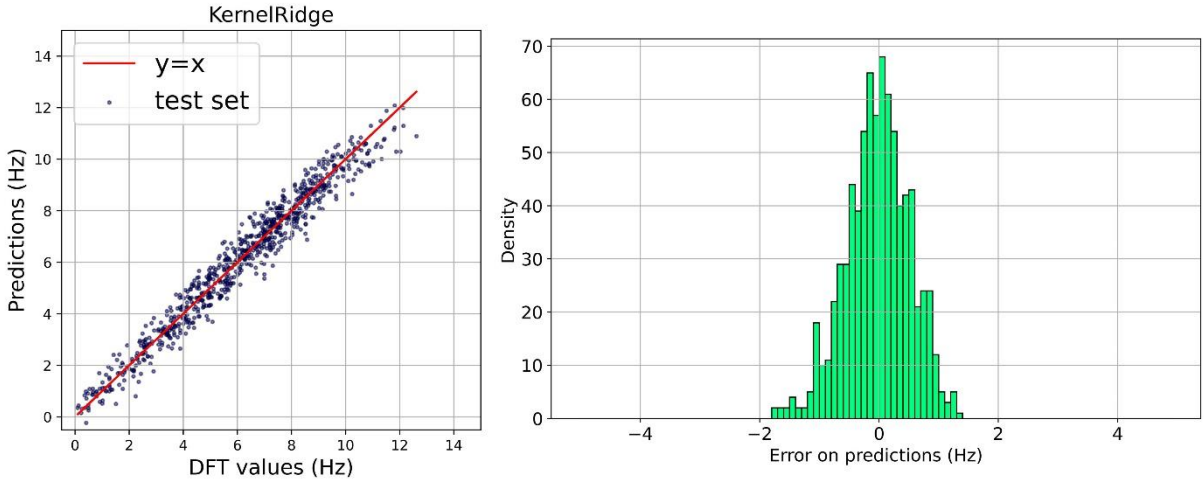


Figure VI.12*: Predictions against the DFT (6-311++G** basis set with the FC contribution only) values with the kernel ridge regression (left) and the associated error distribution (right) with the ML2 set of descriptors.*

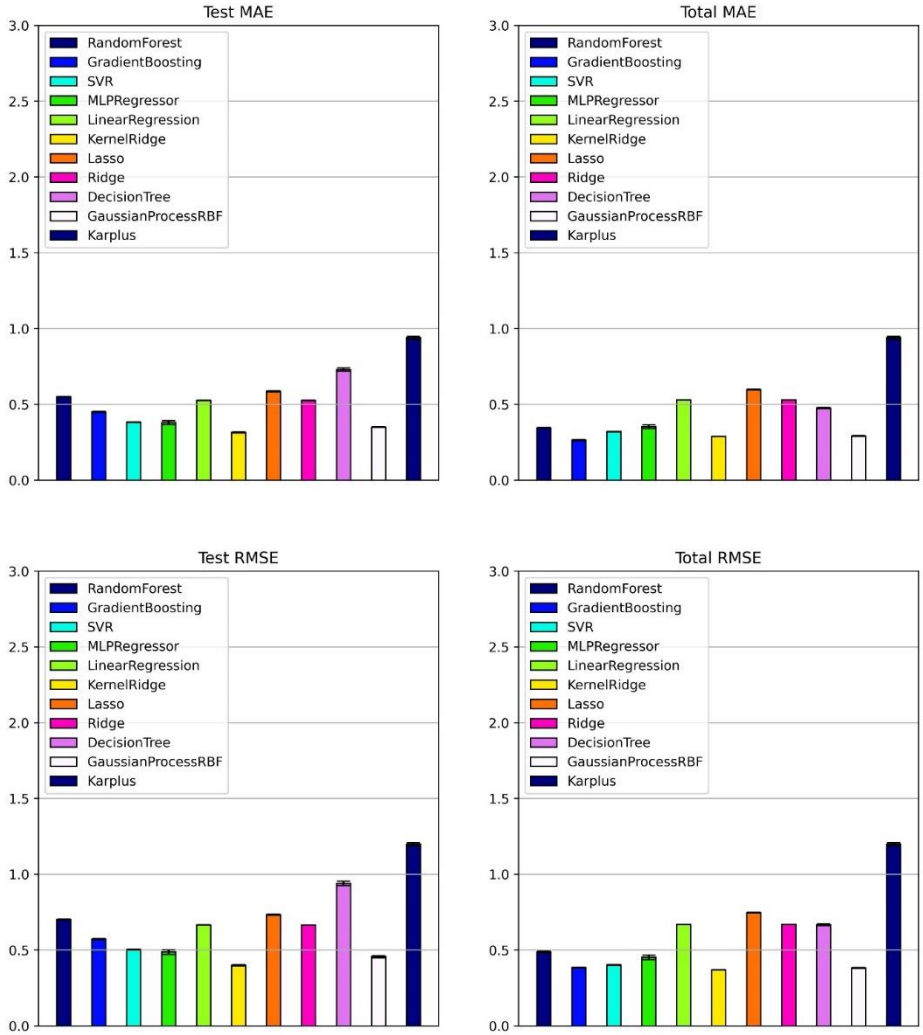


Figure VI.13: *Performance of all tested model on the ML3 set of descriptors. The MAE (up) and the RMSE (down) were calculated on the test set (left) and the total set (right)*

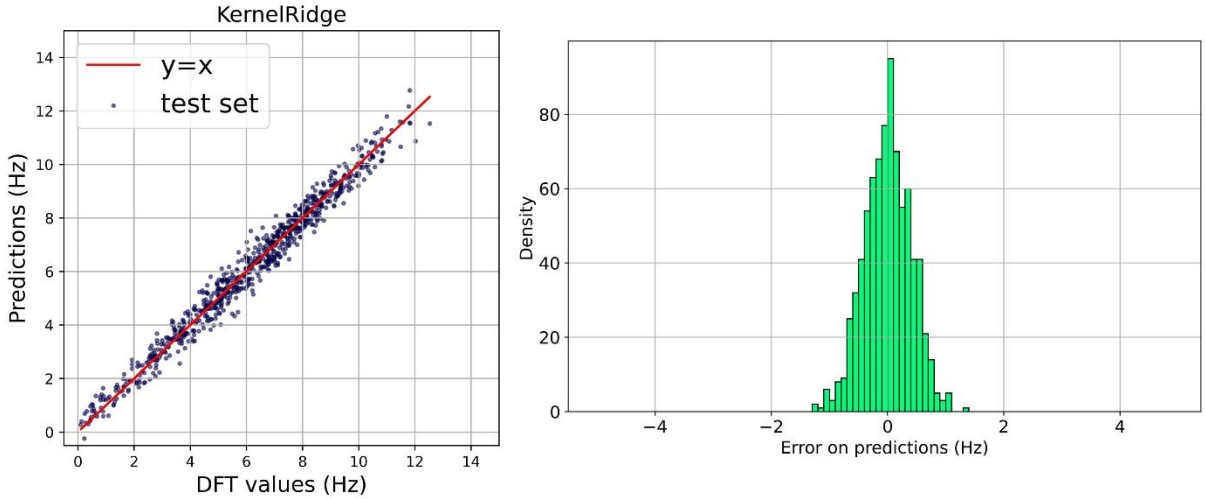


Figure VI.14: *Predictions against the DFT (6-311++G** basis set with the FC contribution only) values with the kernel ridge regression (left) and the associated error distribution (right) with the ML set of descriptors.*

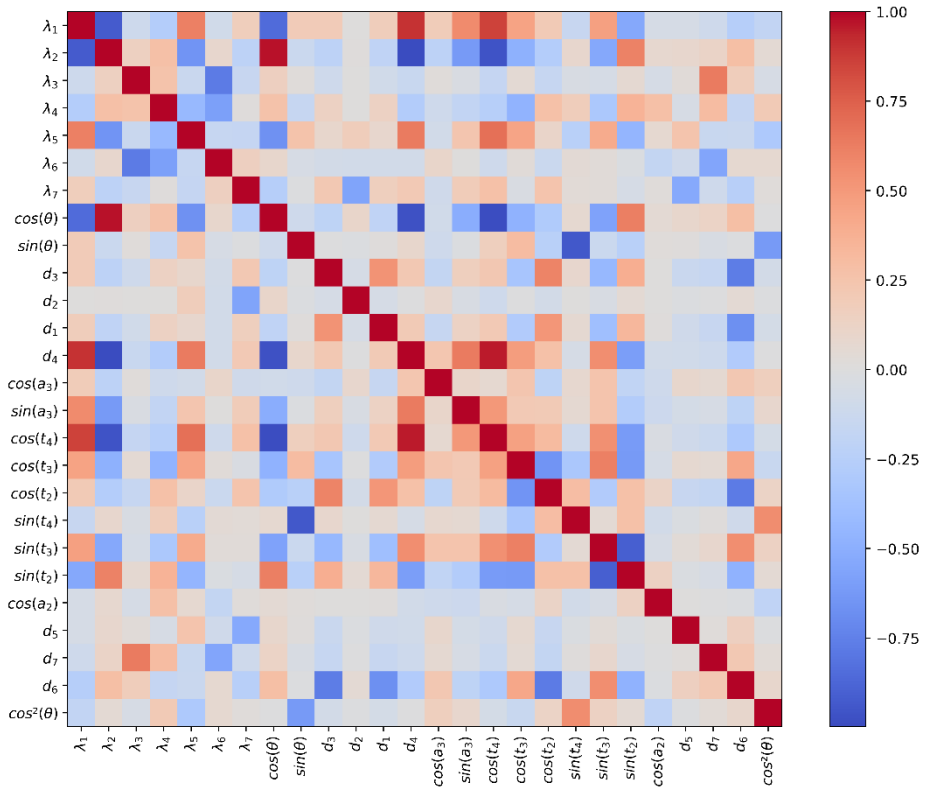


***Figure VI.15**: Correlation matrix of the ML3 set of descriptors.*

To refine this model, we carried out a descriptor selection. The aim of this method is to identify the minimal set of descriptors that still provided high-level prediction while avoiding overfitting. The approach involves first selecting the descriptor that achieves the largest reduction in MAE when used individually. Then, additional descriptors sequentially are included one by one, always choosing the one that improves the model's performance the most. To ensure that including descriptors improves performance, it is essential to examine the performance on both the training and test sets. If the performance remains unchanged on the test set, thus the newly introduced descriptors do not enhance the model. If, in parallel, performance on the training set continues to increase, then we are generally in a case of overfitting. In this case, the model is no longer able to generalise and over-specifies on its training set.

Figure VI.16 shows the performance improvement of the KRR model as descriptors were added. Performance increased steadily with the test set and the full dataset until a convergence was reached at 17 over 26 descriptors. Beyond this point, additional descriptors did not improve predictive accuracy and even led to slight overfitting, confirming the need to limit the model

complexity. Besides, the first six selected descriptors, which lead to a significant reduction in MAE, are from different nature: local geometric ($cos(\theta)$, d_2, d_4), geometric extension to the first neighbour ($cos(t_3)$, d_6) and nuclear (λ_5). This diversity suggests that combining these different types of descriptors enhances the model's performance

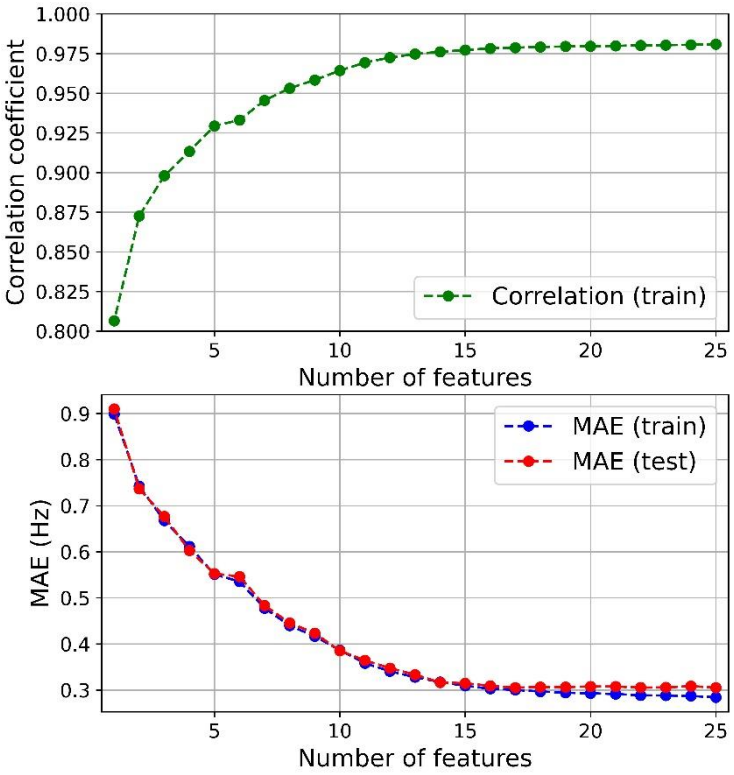


Figure VI.16 *: Variable selection with the ML3 set of descriptors utilizing the kernel ridge regression model. The correlation coefficient (up) and the MAE (down) are displayed at each stage of the selection.*

Then, we have performed the Shappley Additive explanation (SHAP) algorithm[37] for evaluating the contribution and the importance of each descriptor in our ML models, allowing us to understand how each feature influences the prediction of the ${}^3J_{HN-H\alpha}$ coupling constant (**Figure VI.17**). The horizontal axis shows the impact of the descriptors on the predictions. A negative SHAP value indicates a decrease in the output variable, while positive values tend to increase the value of the coupling constant. The $cos^2(\theta)$ descriptor, for example, has SHAP values between -1.5 and +1 which means that this descriptor has a large impact both on predicting low values of constants and on predicting higher values.

The descriptors are ranked according to their importance, with those that have the greatest impact on the model's decision at the top such as $cos^2(\theta)$ and $cos(\theta)$ very consistently. Then

two distances have a significant impact on prediction, d_4 and d_2 which are the interproton distance and the C_α-N distance, respectively. Finally, the red values show a positive impact on prediction and the blue values a negative impact. We observe that $cos^2(\theta)$ has a linear impact over a wide range but that other variables such as d_4 and d_2 have a positive and negative impact, respectively, more important than $cos^2(\theta)$ for predicting rather average variables. Thus, although $cos^2(\theta)$ and $cos(\theta)$ are dominant, some variables such as d_4 and d_2 are also dominant and the other additional descriptors allow the model to be better fitted to the set of predictions, certainly thanks to precise corrections and complex interaction, which justifies the use of more elaborate models.

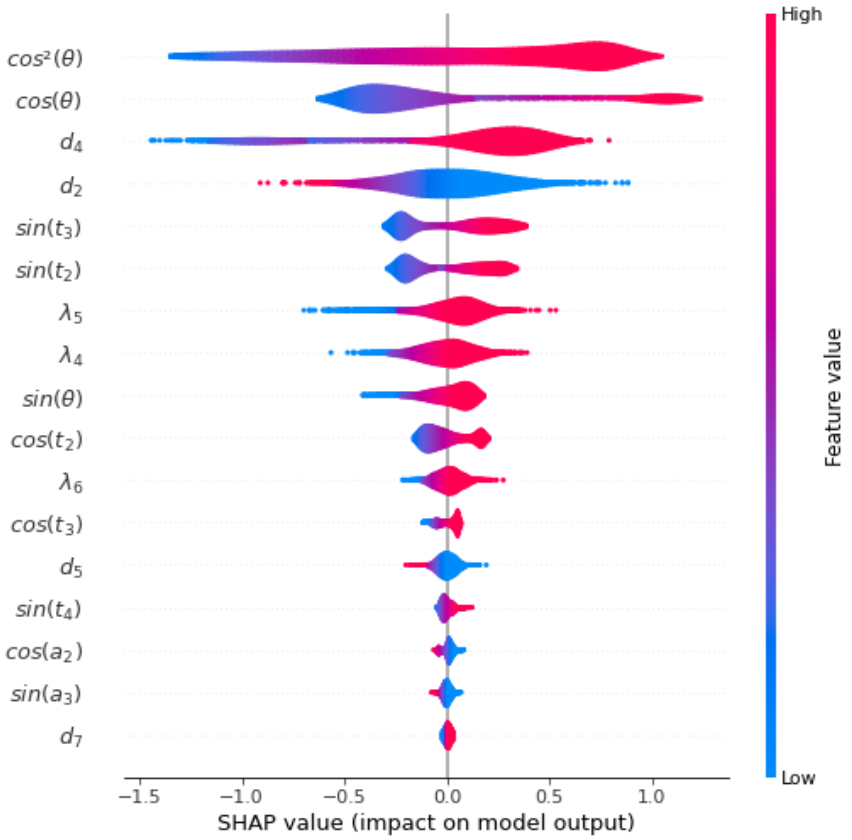


***Figure VI.17**: Summary plot of the SHAP algorithm on the 17 descriptors selected from the descriptor selection.*

VI. C. 4. Conclusion

In conclusion, we have developed a machine learning model, based on polynomial kernel ridge regression of degree 2, to predict the $^3J_{HN-H\alpha}$ DFT coupling constant resulting from a calculation with the 6-311++G** basis solely taking into account the FC contribution. This model, based on geometric descriptors and descriptors derived from the Coulomb matrix,

significantly outperforms our optimized Karplus model. These descriptors are all easily calculated from simple Cartesian coordinates of only seven atoms. In addition, the quasi-instantaneous prediction of this ^3J constant with a high accuracy, can then be coupled with the linear correlation of the DFT calculation with the aug-cc-pVTZ-J basis, whichs appear to be suitable for predicting the experimental value of the ^3J constant.

VI. D. Conclusion et perspectives

Cette étude a permis de répondre à notre troisième objectif global qui portait sur l'amélioration de la vitesse des protocoles théoriques pour la détermination de structure des peptides en solution. Les travaux montrent que le choix de descripteurs pertinents améliore tous modèles testés, et que le modèle de régression ridge à noyau polynomial de degré deux est le plus performant sur les tests effectués. Les descripteurs sélectionnés ont l'avantage d'être très facilement implémentés car ils ne requièrent que les coordonnés cartésiennes des sept atomes localement impliqués dans la constante de couplage $^3J_{HN-H\alpha}$.

Pour le moment, nous n'avons testé qu'un seul type de constante de couplage. Bien que celle-ci soit la plus pertinente pour le repliement de le chaîne peptidique, il pourrait être intéressant d'étendre la méthode à la fois aux autres constantes de couplages interproton, pour par exemple étudier le repliement de la chaîne latérale, mais aussi à d'autres type d'atomes, là où le modèle de Karplus n'est plus utilisé.

Un autre point qui nous semble pertinent à explorer du point de vue des expérimentateurs, est comment retrouver la conformation à partir de la constante de couplage mesurée avec une meilleure précision que le modèle de Karplus. Ce modèle pourrait être construit à partir de calculs théoriques de la même manière, avec en entrée les constantes de couplage, mais aussi d'autres descripteurs comme le déplacement chimique des atomes voisins cette constante, et en sortie l'angle dièdre θ.

VI. E. Références

1. Hemmer, M. C., Steinhauer, V. & Gasteiger, J. Deriving the 3D structure of organic molecules from their infrared spectra. *Vib. Spectrosc.* **19**, 151–164 (1999).
2. Provasi, P. F., Aucar, G. A. & Sauer, S. P. A. The effect of lone pairs and electronegativity on the indirect nuclear spin–spin coupling constants in CH2X (X=CH2, NH, O, S): Ab initio calculations using optimized contracted basis sets. *J. Chem. Phys.* **115**, 1324–1334 (2001).
3. Azencott, C.-A. Introduction au Machine Learning.
4. Smola, A. J. & Schölkopf, B. A tutorial on support vector regression. *Stat. Comput.* **14**, 199–222 (2004).
5. Wüthrich, K., Billeter, M. & Braun, W. Polypeptide secondary structure determination by nuclear magnetic resonance observation of short proton-proton distances. *J. Mol. Biol.* **180**, 715–740 (1984).
6. Sternberg, U., Tzvetkova, P. & Muhle-Goll, C. The simulation of NMR data of flexible molecules: sagittamide A as an example for MD simulations with orientational constraints. *Phys. Chem. Chem. Phys.* **22**, 17375–17384 (2020).
7. IUPAC-IUB Comm. on Biochem. Nomenclature. IUPAC-IUB Commission on Biochemical Nomenclature. Abbreviations and symbols for the description of the conformation of polypeptide chains. Tentative rules (1969). *Biochemistry* **9**, 3471–3479 (1970).
8. Marraud, M. & Aubry, A. Backbone side chain interactions in peptides. *Int. J. Pept. Protein Res.* **23**, 123–133 (1984).
9. Imai, K. & Ōsawa, E. An empirical extension of the karplus equation. *Magn. Reason. Chem.* **28**, 668–674 (1990).

10. Pardi, A., Billeter, M. & Wüthrich, K. Calibration of the angular dependence of the amide proton-Cα proton coupling constants, 3JHNα, in a globular protein: Use of 3JHNα for identification of helical secondary structure. *J. Mol. Biol.* **180**, 741–751 (1984).
11. Brueschweiler, R. & Case, D. A. Adding Harmonic Motion to the Karplus Relation for Spin-Spin Coupling. *J. Am. Chem. Soc.* **116**, 11199–11200 (1994).
12. Wang, A. C. & Bax, A. Determination of the Backbone Dihedral Angles φ in Human Ubiquitin from Reparametrized Empirical Karplus Equations. *J. Am. Chem. Soc.* **118**, 2483–2494 (1996).
13. Schmidt, J. M., Blümel, M., Löhr, F. & Rüterjans, H. Self-consistent 3J coupling analysis for the joint calibration of Karplus coefficients and evaluation of torsion angles. *J. Biomol. NMR* **14**, 1–12 (1999).
14. Vuister, G. W. & Bax, A. Quantitative J correlation: a new approach for measuring homonuclear three-bond J(HNH.alpha.) coupling constants in 15N-enriched proteins. *J. Am. Chem. Soc.* **115**, 7772–7777 (1993).
15. Cung, M. T., Marraud, M. & Neel, J. Experimental Calibration of a Karplus Relationship in Order to Study the Conformations of Peptides by Nuclear Magnetic Resonance. *Macromolecules* **7**, 606–613 (1974).
16. Ludvigsen, S., Andersen, K. V. & Poulsen, F. M. Accurate measurements of coupling constants from two-dimensional nuclear magnetic resonance spectra of proteins and determination of φ-angles. *J. Mol. Biol.* **217**, 731–736 (1991).
17. Ramsey, N. F. Electron Coupled Interactions between Nuclear Spins in Molecules. *Phys. Rev.* **91**, 303–307 (1953).
18. Huang, B. & von Lilienfeld, O. A. Ab Initio Machine Learning in Chemical Compound Space. *Chem. Rev.* **121**, 10001–10036 (2021).

19. Fang, J. *et al.* Predicting scalar coupling constants by graph angle-attention neural network. *Sci. Rep.* **11**, 18686 (2021).
20. Shibata, K. & Kaneko, H. Prediction of spin–spin coupling constants with machine learning in NMR. *Anal. Sci. Adv.* **2**, 464–469 (2021).
21. Gerrard, W. *et al.* IMPRESSION – prediction of NMR parameters for 3-dimensional chemical structures using machine learning with near quantum chemical accuracy. *Chem. Sci.* **11**, 508–515 (2020).
22. Navarro-Vázquez, A. A DFT/machine-learning hybrid method for the prediction of 3JHCCH couplings. *Magn. Res. Chem.* **59**, 414–422 (2021).
23. Menant, S., Tognetti, V., Oulyadi, H., Guilhaudis, L. & Ségalas-Milazzo, I. A Joint Experimental and Theoretical Study on the Structural and Spectroscopic Properties of the Piv-Pro-d-Ser-NHMe Peptide. *J. Phys. Chem. B* **128**, 6704–6715 (2024).
24. Abraham, M. J. *et al.* GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **1–2**, 19–25 (2015).
25. Pettersen, E. F. *et al.* UCSF Chimera—A visualization system for exploratory research and analysis. *J. Comput. Chem.* **25**, 1605–1612 (2004).
26. Rupp, M., Tkatchenko, A., Müller, K.-R. & von Lilienfeld, O. A. Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning. *Phys. Rev. Lett.* **108**, 058301 (2012).
27. Raghunathan, S. & Priyakumar, U. D. Molecular representations for machine learning applications in chemistry. *Int. J. Quantum Chem.* **122**, e26870 (2022).
28. Montavon, G. *et al.* Machine learning of molecular electronic properties in chemical compound space. *New J. Phys.* **15**, 095003 (2013).
29. Tchagang, A. B. & Valdés, J. J. Prediction of the Atomization Energy of Molecules Using Coulomb Matrix and Atomic Composition in a Bayesian Regularized Neural Networks. in

Artificial Neural Networks and Machine Learning – ICANN 2019: Workshop and Special Sessions (eds. Tetko, I. V., Kůrková, V., Karpov, P. & Theis, F.) 793–803 (Springer International Publishing, Cham, 2019).

30. Montavon, G. *et al.* Learning invariant representations of molecules for atomization energy prediction: 26th Annual Conference on Neural Information Processing Systems 2012, NIPS 2012. *Advances in Neural Information Processing Systems 25* 440–448 (2012).
31. Helgaker, T. & Pecul, M. Spin–Spin Coupling Constants with HF and DFT Methods. in *Calculation of NMR and EPR Parameters* 101–121 (John Wiley & Sons, Ltd, 2004).
32. Autschbach, J. & Le Guennic, B. Analyzing and Interpreting NMR Spin–Spin Coupling Constants Using Molecular Orbital Calculations. *J. Chem. Educ.* **84**, 156 (2007).
33. Cremer, D. & Gräfenstein, J. Calculation and analysis of NMR spin–spin coupling constants. *Phys. Chem. Chem. Phys.* **9**, 2791–2816 (2007).
34. Yanai, T., Tew, D. P. & Handy, N. C. A new hybrid exchange–correlation functional using the Coulomb-attenuating method (CAM-B3LYP). *Chemical Physics Letters* **393**, 51–57 (2004).
35. Aidas, K. *et al.* The Dalton quantum chemistry program system. *WIREs Comput. Mol. Sci.* **4**, 269–284 (2014).
36. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
37. Lundberg, S. M., Erion, G. G. & Lee, S.-I. Consistent Individualized Feature Attribution for Tree Ensembles. Preprint at https://doi.org/10.48550/arXiv.1802.03888 (2019).

Conclusion générale

Dans le cadre de cette thèse, nous avions trois objectifs principaux, auxquels nous avons en grande partie répondu. Ces objectifs visaient à résoudre des problématiques diverses et contribuent chacun à l'amélioration de l'étude structurale des petits peptides en solution.

Le premier objectif portait sur la caractérisation des coudes en solution en développant un protocole théorique capable de simuler les spectres DCE pour chaque type de coude. Nous nous sommes dans ce travail focalisés sur le peptide Piv-Pro-D-Ser-NHMe qui se structure en coude de type β-II en solution aqueuse. En utilisant une dynamique moléculaire suffisamment longue (1µs) pour permettre à la molécule d'explorer un maximum de conformations, avec le champ de force CHARMM27 et le modèle d'eau TIP5P, nous avons pu générer le spectre DCE du peptide. L'extraction des géométries est suivie de calculs TD-DFT avec la fonctionnelle CAM-B3LYP et le jeu de bases 6-31+G*. Dans ces calculs quantiques, l'environnement est pris en compte par une méthode d'inclusion polarisable.

Bien que nous ayons mis au point un protocole permettant de générer le spectre DCE pour un coude de type β-II sur le peptide Piv-Pro-D-Ser-NHMe, l'intégration de mesures expérimentales demeure encore indispensable pour l'appliquer à d'autres peptides se structurant en coude(s). Ce protocole est en effet fortement influencé par le choix du champ de force, qui conditionne les géométries prédites et, par conséquent, les paramètres spectroscopiques qui en découlent.

Le deuxième objectif concernait la détermination des conformations de peptides flexibles en solution. Nous avons utilisé le protocole développé dans la première partie du travail et l'avons adapté pour déterminer les conformations adoptées par le peptide AFA dans l'eau. L'applicabilité du protocole théorique est très fortement liée au choix du champ de force.

Dans notre première étude, le champ de force CHARMM27 s'est avéré totalement en accord avec les mesures expérimentales en générant majoritairement des conformations β-II, dont les distances extraites étaient très proches des valeurs expérimentales, et les angles φ de la sérine en adéquation avec la constante de couplage $^{3}JH_N\text{-}H_\alpha$ mesurée expérimentalement. En revanche, dans la seconde étude, le champ de force OPLS-AA a donné de meilleurs résultats. Cependant, bien que les distances simulées donnassent une simulation légèrement plus réaliste qu'avec CHARMM27 et que la constante de couplage calculée du troisième résidu allait dans le même sens, il n'était pas possible de conclure sur le choix du champ de force avec

uniquement la simulation des paramètres RMN. En effet, il a été nécessaire pour effectuer la sélection d'utiliser tous les paramètres spectroscopiques simulés, dont le spectre DCE était le plus déterminant, pour faire un choix sur le champ de force.

Une fois le protocole théorique optimisé, nous avons utilisé une méthode de clustering basée sur la méthode des *k*-means pour monter qu'AFA est présent sous la forme de six conformères et que les conformères V et VI sont les plus représentés. Nous pensons que cette description en conformères, similaire à celle utilisée pour les petites molécules organiques, est plus adaptée que celle utilisée actuellement dans la littérature qui propose pour le peptide des proportions de structures secondaires sans tenir compte de la géométrie adoptée par les trois acides aminés composant le peptide.

En conclusion, le protocole développé au cours de ce travail peut être utilisé pour simuler le spectre ECD d'un petit peptide structuré ou pour déterminer les conformations adoptées par un petit peptide dans l'eau. Il est toutefois nécessaire de sélectionner un champ de force adapté et de disposer de paramètres spectroscopiques expérimentaux. OPLS-AA génère des géométries beaucoup plus dispersées et moins contraintes dans des zones spécifiques du diagramme de Ramachandran, contrairement à CHARMM27. Nous pensons donc que OPLS-AA est plus adéquat pour explorer l'espace conformationnel des peptides flexibles, alors que CHARM27, en limitant les conformations en dehors de certaines zones, serait plus adapté aux molécules ayant des structures bien définies, dominantes dans ces régions spécifiques.

Enfin le dernier objectif concernait l'optimisation des protocoles théoriques en réduisant le temps de calcul. A cette fin, nous avons développé un modèle de régression ridge à noyau (ou « *kernel ridge regression* ») pour prédire les constantes de couplage $^{3}JH_{N}$-H_{α} calculées par DFT de façon quasi instantanée, avec une très haute précision. Ce modèle se base sur 17 descripteurs facilement calculables à partir des coordonnées cartésiennes et des numéros atomiques, incluant à la fois des descripteurs purement géométriques et des paramètres nucléaires dérivés de la matrice de Coulomb. Les calculs de constantes de couplage précis nécessitent d'importantes ressources, surtout lorsqu'ils sont couplés avec des simulations de DM, impliquant des milliers de calculs. Notre modèle permet de surmonter cette limitation en introduisant très peu d'erreurs (MAE = 0.3 Hz, RMSe = 0.4 Hz), rendant ainsi le processus bien plus efficient.

Perspectives

En perspective de ce travail, plusieurs pistes peuvent être envisagées pour améliorer la détermination de structures de petits peptides flexibles ou adoptant une conformation de type coude en solution.

Actuellement, nous disposons d'un protocole qui simule des spectres DCE en accord avec les valeurs expérimentales, à conditions de générer des géométries représentatives de la molécule en solution. Toutefois, l'utilisation de contraintes pour la simulation des spectres de coude spécifiques ne garantit pas l'obtention de spectres réalistes. Pour renforcer cette démarche, il est pertinent de réaliser des DM sans contraintes sur des peptides dont la géométrie en solution est bien caractérisée. Cette stratégie nous a permis de valider un spectre pour un coude β-II, mais il faudrait étendre cette approche à divers peptides pour confirmer la généralité de ces résultats et associer un spectre DCE précis pour cette conformation. Il faudrait également reproduire ces résultats pour d'autres types de coude. Des candidats potentiels pour ces études ont été identifiés, comme les peptides Z-Aib-Pro-Aib-Pro-OMe, Piv-Pro-Aib-NHMe, Piv-Pro-D-Ala-NHMe et Piv-Pro-Val-NHMe qui se replient sous la forme d'un coude β en phase cristalline (β-I, β-II, β-II et β-I respectivement) possédant des motifs DCE particuliers.

De plus, nous avons observé que le diagramme de Ramachandran est très limitant pour les très petits peptides. Le clustering à l'aide de k-means, tel que nous l'avons effectué, s'affranchit des limites des zones du diagramme pour former des clusters proches structurellement, mais est pour l'instant fondé sur les mêmes descripteurs géométriques. Une description à l'aide d'autres descripteurs pourrait être envisagée, notamment en ajoutant les angles dièdres des terminaisons.

Lors de ce travail, des observations intrigantes ont été faites : des peptides avec géométries très similaires (faible RMSD pour l'ensemble des atomes) peuvent générer des spectres DCE très différents. Un exemple illustré avec le peptide Piv-Pro-D-Ser-NHMe, qui présente deux conformations ayant un RMSD de 0,24 Å sur l'ensemble des atomes, montre que malgré cette similarité, les spectres DCE simulés divergent de manière significative (**Figure i**). Ce phénomène soulève la question sur la sensibilité des spectres DCE à des variations subtiles de conformation ou d'environnement et nécessite une exploration approfondie pour mieux comprendre les facteurs influençant ces différences spectroscopiques.

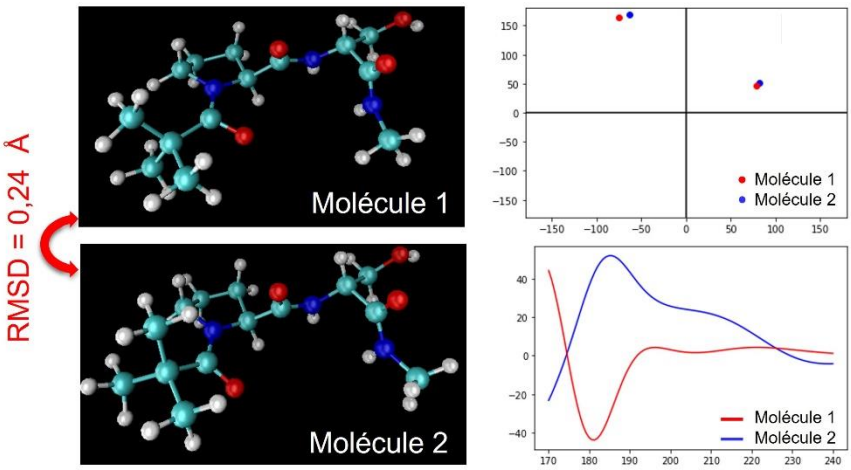


Figure i *: Représentation de deux conformations du peptide Piv-Pro-D-Ser-NHMe avec des RMSD très proche (0,24 Å) sur l'ensemble des atomes avec la représentation du Ramachandran des deux résidus et les spectres DCE simulés.*

Des pistes de développement en « *machine learning* » pourraient contribuer à affiner les prédictions. Pour améliorer notre approche, il pourrait être intéressant de prédire les géométries à l'aide de la constante de couplage plutôt que cette dernière, et d'utiliser d'autres paramètres spectroscopiques comme le déplacement chimique noyaux proches pour effectuer cette prédiction. L'utilisation de la relation de Karplus s'avère limitée en précision et le recours à des modèles plus sophistiqués, intégrant davantage de variables, pourrait offrir une meilleure estimation de la géométrie.

Enfin un des obstacles majeurs à la prédiction des spectre DCE par la TD-DFT reste la durée des calculs qui est élevée. Il serait intéressant de développer des modèles prédictifs permettant de générer ces spectres à partir de descripteurs géométriques ou autres. Des tests préliminaires ont été effectués dans le cadre de cette thèse pour prédire les spectres à partir d'un échantillonnage de la fonction de densité radiale, mais les résultats n'ont pas été concluants. Un échantillonnage plus poussé de cette fonction ou l'exploration de nouveaux descripteurs pourraient potentiellement offrir de nouvelles voies pour la prédiction rapide des spectres DCE.

Annexe : Chapitre IV

Table des matières

Figure S1. 1D ^{1}H NMR spectrum of the Piv-Pro-D-Ser-NHMe peptide (1.667 mg mL^{-1}) in water (90% H_2O/10% D_2O) with the *zgesgp* pulse program at 293 K on a 600 MHz spectrometer. Water suppression was achieved using the excitation sculpting method. Peaks corresponding to major impurities in solution are indicated by an asterisk.

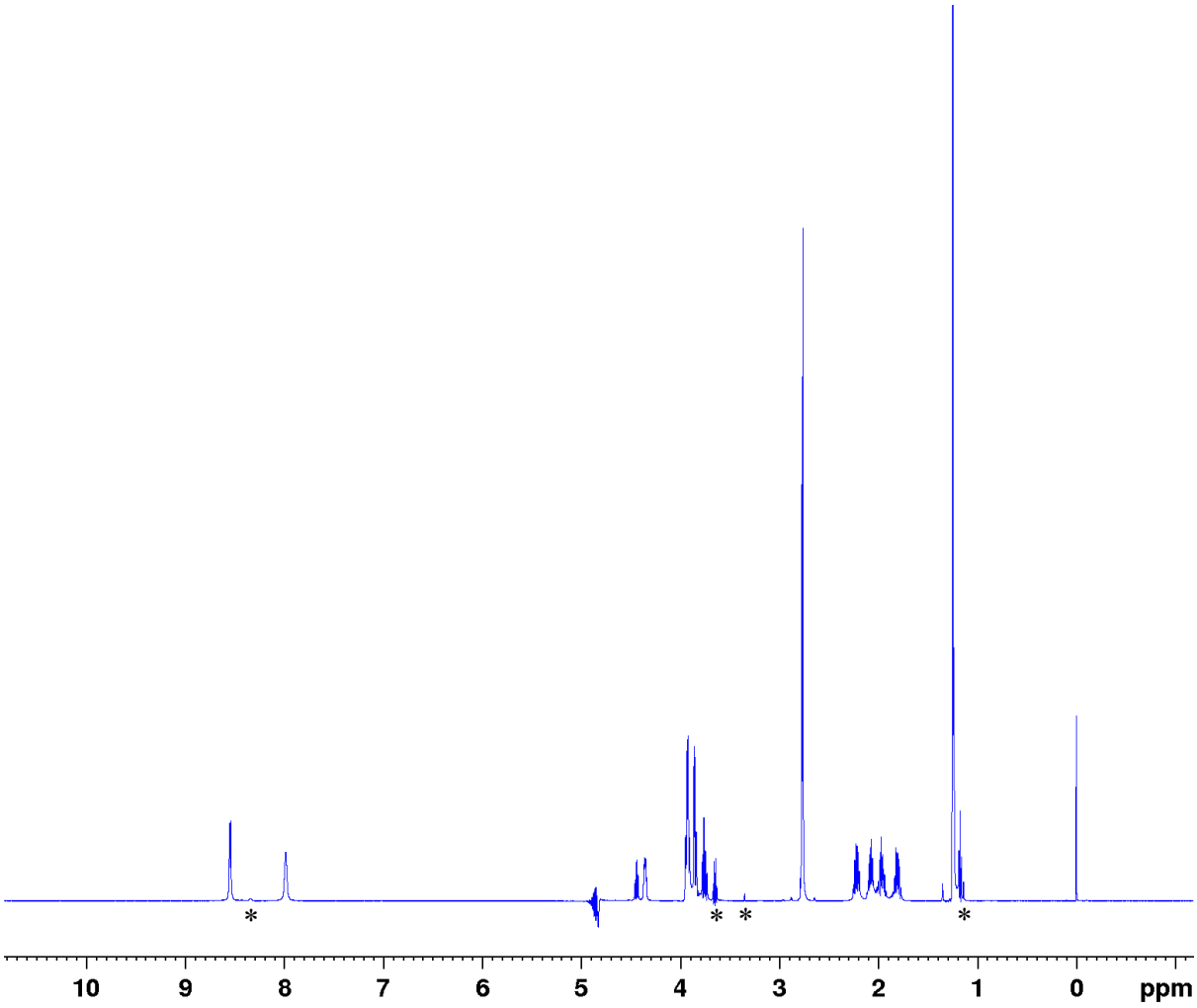


Table S1. Proton chemical shifts at 293 K.

Residue	H_N	H_α	H_β	Others
Piv	-	-	-	1.25 $H^t_{Bu(Piv)}$
Pro[a]	-	4.44	2.22 $H_{\beta3}$, 1.80 $H_{\beta2}$	3.93 $H_{\delta3}$, 3.76 $H_{\delta2}$, 2.08 $H_{\gamma2}$, 1.96 $H_{\gamma3}$
D-Ser[b]	8.55	4.36	3.94/3.85 $H_{\beta3/\beta2}$	
NHMe	7.99	-	-	2.76 $H_{CH3(NHMe)}$

Chemical shifts are reported relative to DSS and methylene protons are labeled (2 or 3) in agreement with the IUPAC convention for amino acids.[28] ; a) The stereospecific assignment of the proline protons was carried out by comparing the following NOE cross-peak intensities : $H_{\alpha(Pro)}$-$H_{\beta3(Pro)}$ with $H_{\alpha(Pro)}$-$H_{\beta2(Pro)}$ for β protons; $H_{\beta3(Pro)}$-$H_{\delta3(Pro)}$ with $H_{\beta3(Pro)}$-$H_{\delta2(Pro)}$ for δ protons; $H_{\delta3(Pro)}$-$H_{\gamma3(Pro)}$ with $H_{\delta3(Pro)}$-$H_{\gamma2(Pro)}$ for γ protons; b) no stereospecific assignment was made for the serine residue.

Figure S2. 2D COSY ^{1}H spectrum of the Piv-Pro-D-Ser-NHMe peptide (1.667 mg mL^{-1}) in water (90% H_2O/10% D_2O) with the COSY *gpprqf* pulse program at 293 K on a 600 MHz spectrometer. Water suppression was achieved using a low-power presaturation.

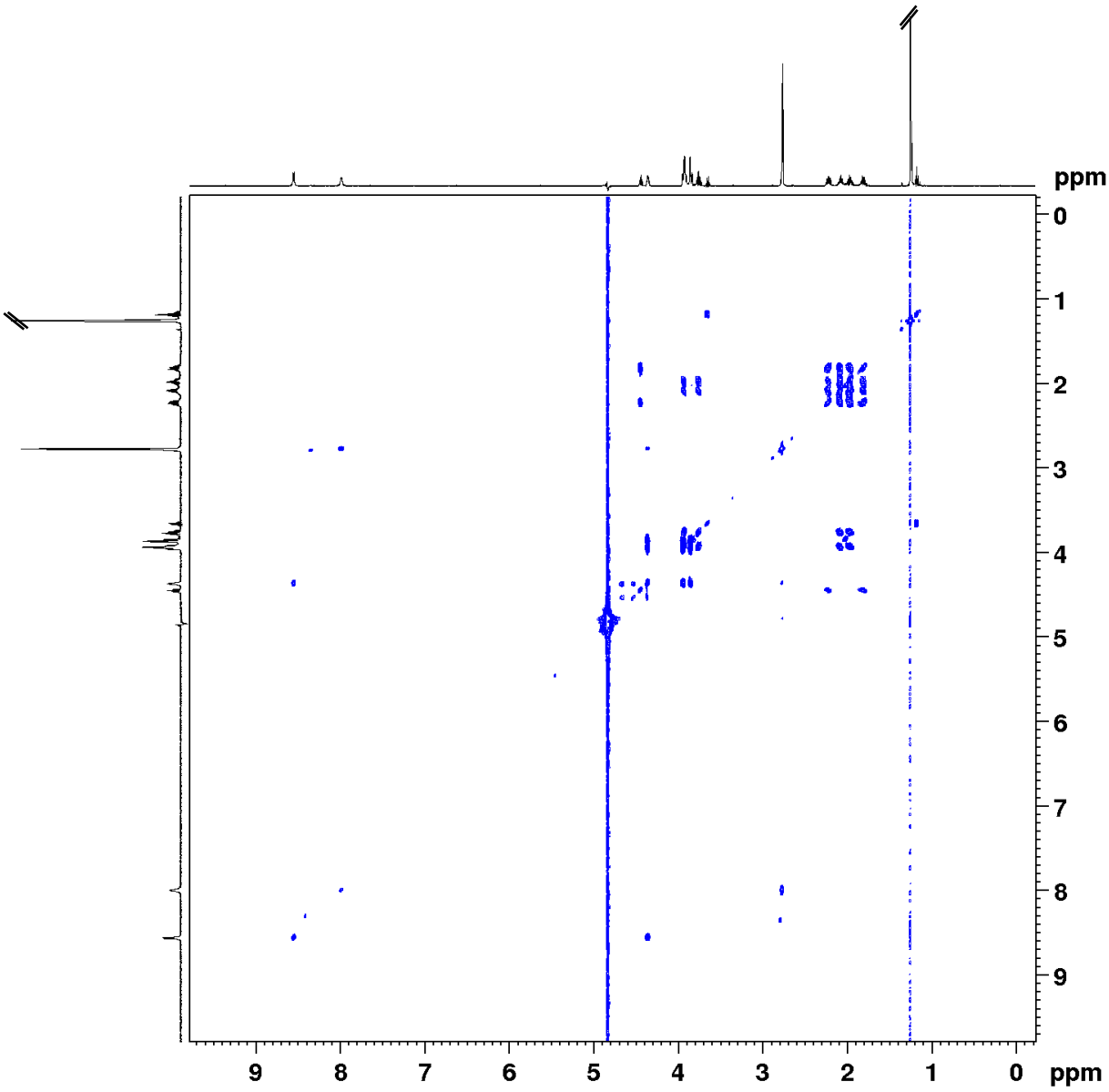


Figure S3. 2D TOCSY ^{1}H spectrum (blue: positive, red: negative) of the Piv-Pro-D-Ser-NHMe peptide (1.667 mg mL^{-1}) in water (90% H_2O/10% D_2O) with *dipsiesgpph* pulse program (60 ms spin-lock time) at 293 K on a 600 MHz spectrometer. Water suppression was achieved using the excitation sculpting method.

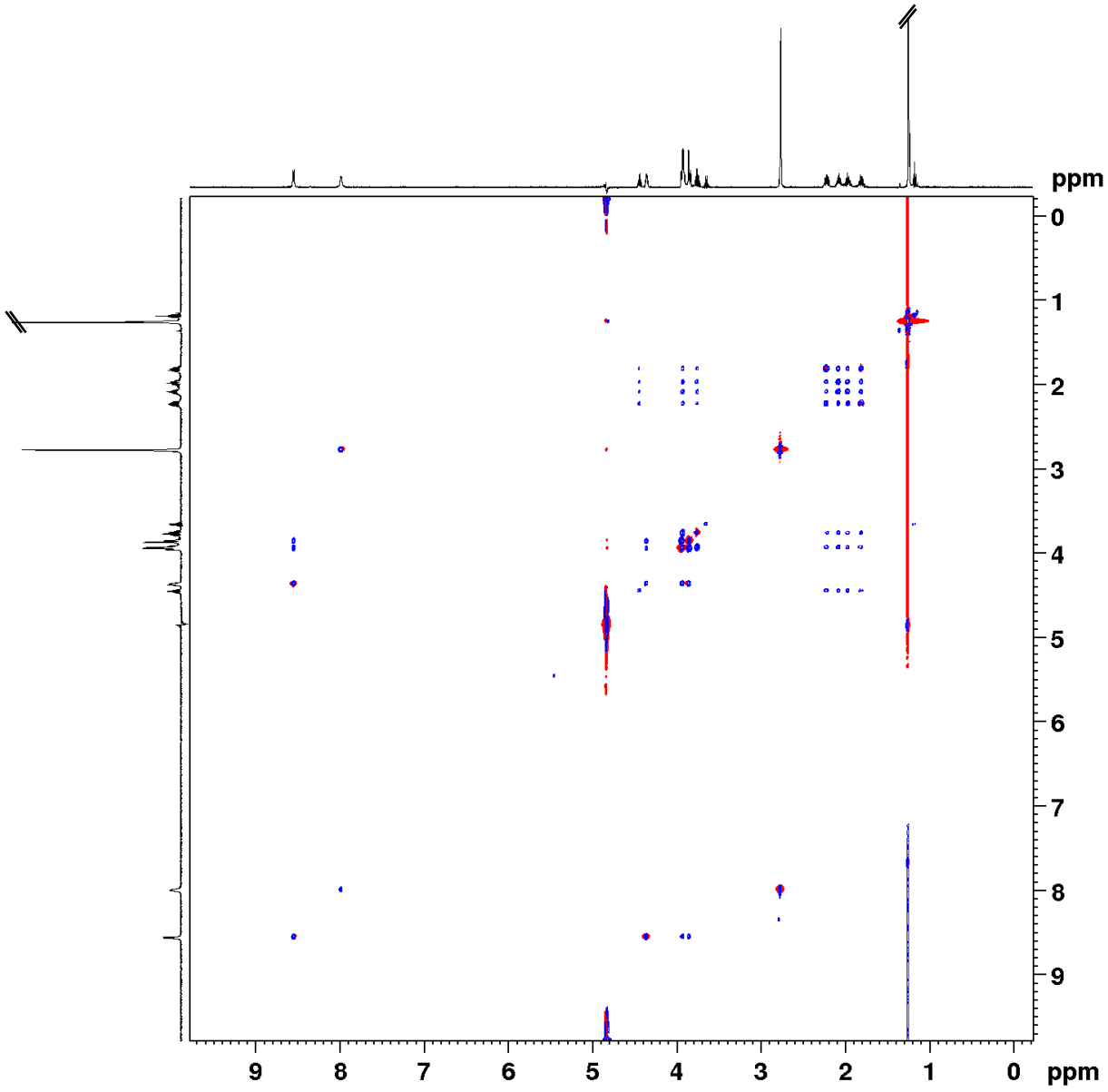


Figure S4. 2D NOESY ^{1}H spectrum (blue: positive, red: negative) of the Piv-Pro-D-Ser-NHMe peptide (1.667 mg mL^{-1}) in water (90% H_2O/10% D_2O) with the *noesyesgpphzs* (500 ms mixing time) pulse program at 293 K on a 600 MHz spectrometer. Water suppression was achieved using the excitation sculpting method. Three areas are depicted: those for the distance calculations (A and B) and the one for the discussion on the *trans/cis* conformation (C).

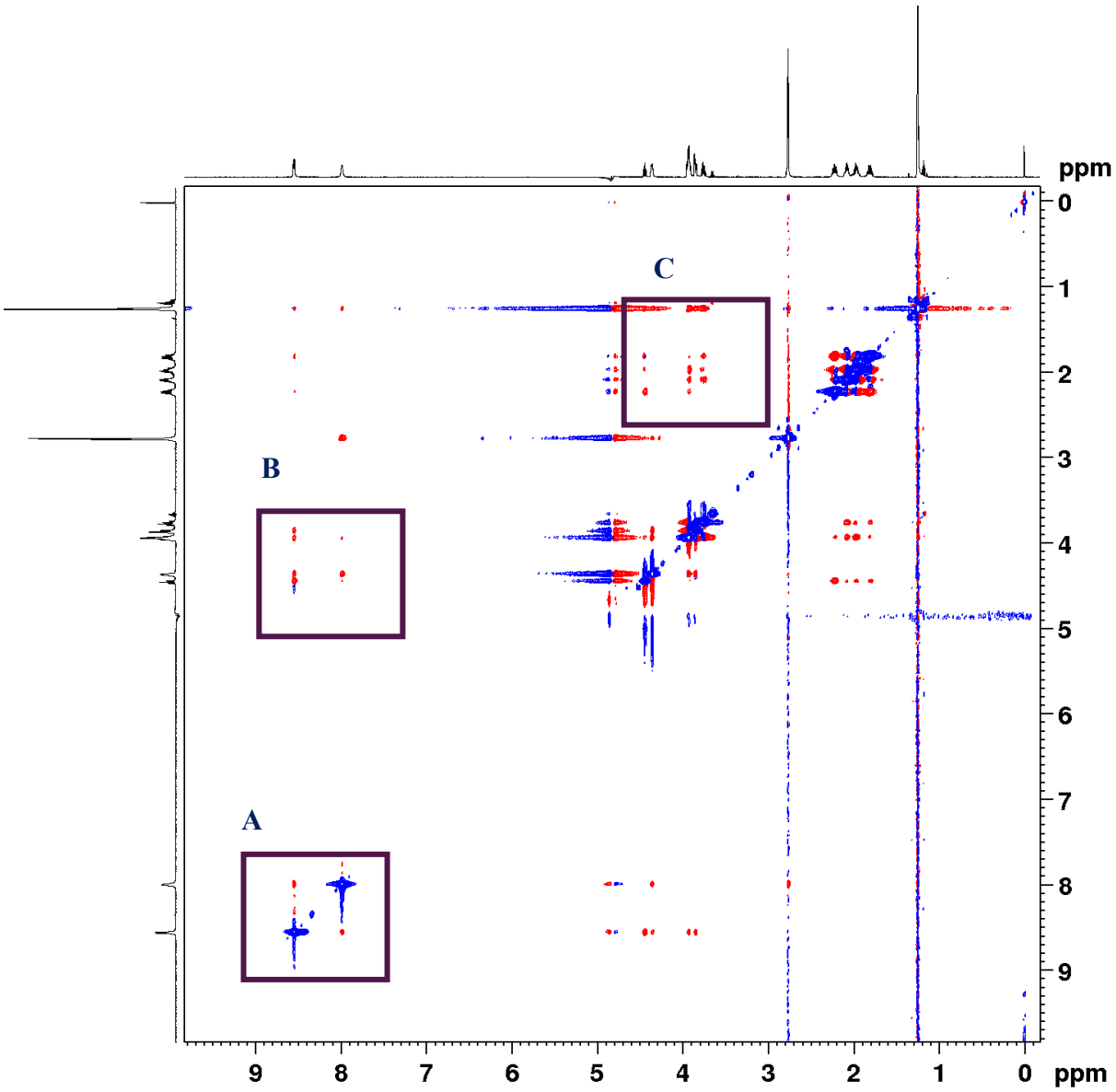


Figure S5. Focus on the correlations (A: $H_{N(Ser)}/H_{N(NHMe)}$, B: $H_{N(Ser)}/H_{\alpha(Pro)}$ and $H_{\alpha(Pro)}/H_{N(NHMe)}$) in the 2D NOESY ^{1}H spectrum (blue: positive, red: negative) of the Piv-Pro-D-Ser-NHMe peptide (1.667 mg mL^{-1}) in water (90% H_2O/10% D_2O) with the *noesyesgpphzs* (500 ms mixing time) pulse program at 293 K on a 600 MHz spectrometer. The three resulting distances are represented below.

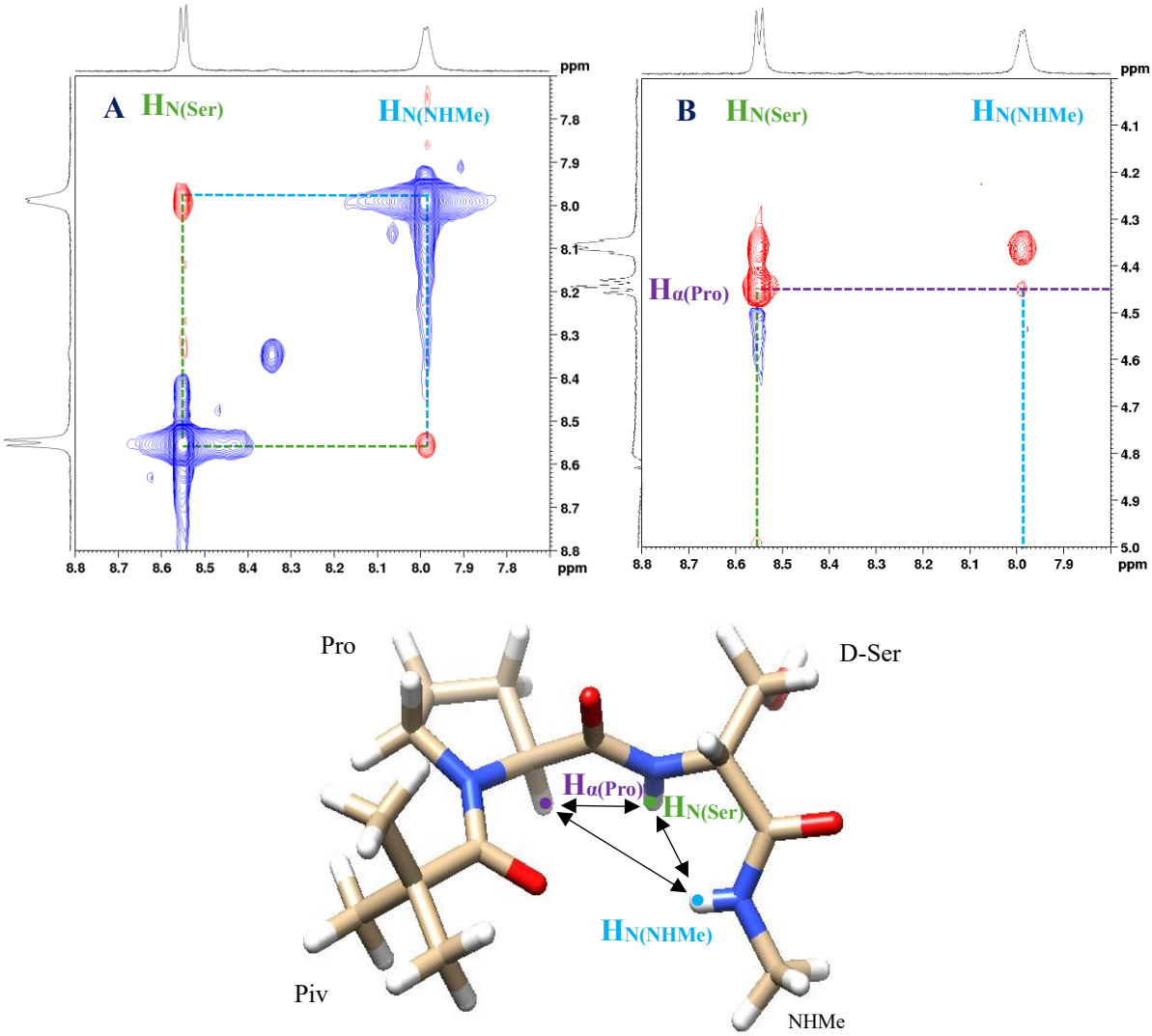


Figure S6. Piv-Pro-D-Ser-NHMe with the proline in *cis* (left) and *trans* (right) conformations. For each conformation, the protons likely to be found nearby are indicated by colored dots. Arrows are displayed to illustrate this spatial proximity.

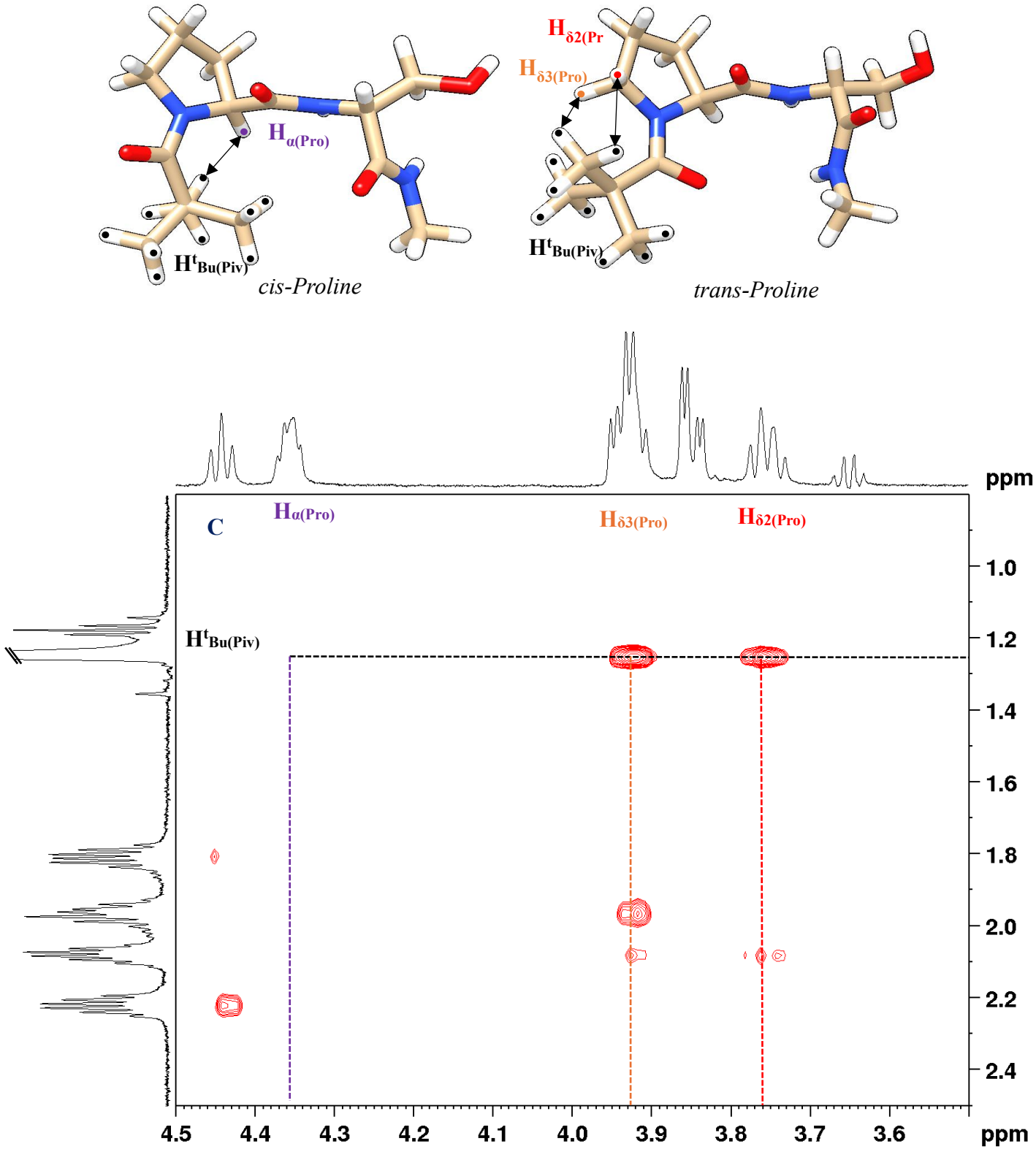


Figure S7. Potential energy surface for the *trans-cis* isomerization at the CAM-B3LYP/6-31+G* level of theory with water solvent described by C-PCM using the Gaussian 16 software.[1] Atomic color code: H in white, C in black, N in blue and O in red. Atoms involved in the ω_i dihedral angle are highlighted in cyan.

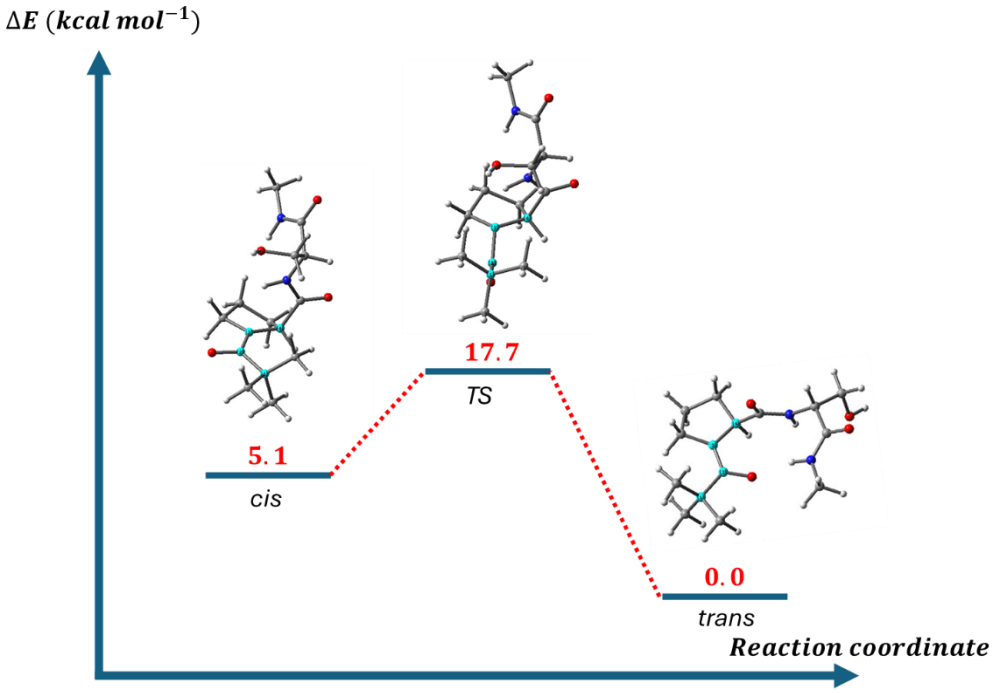


Figure S8. Chemical shifts (in ppm) against temperature (in K) for the two H_N hydrogens, the one in the Ser (blue) and the one in the NHMe residue (orange) for the Piv-Pro-D-Ser-NHMe peptide in water.

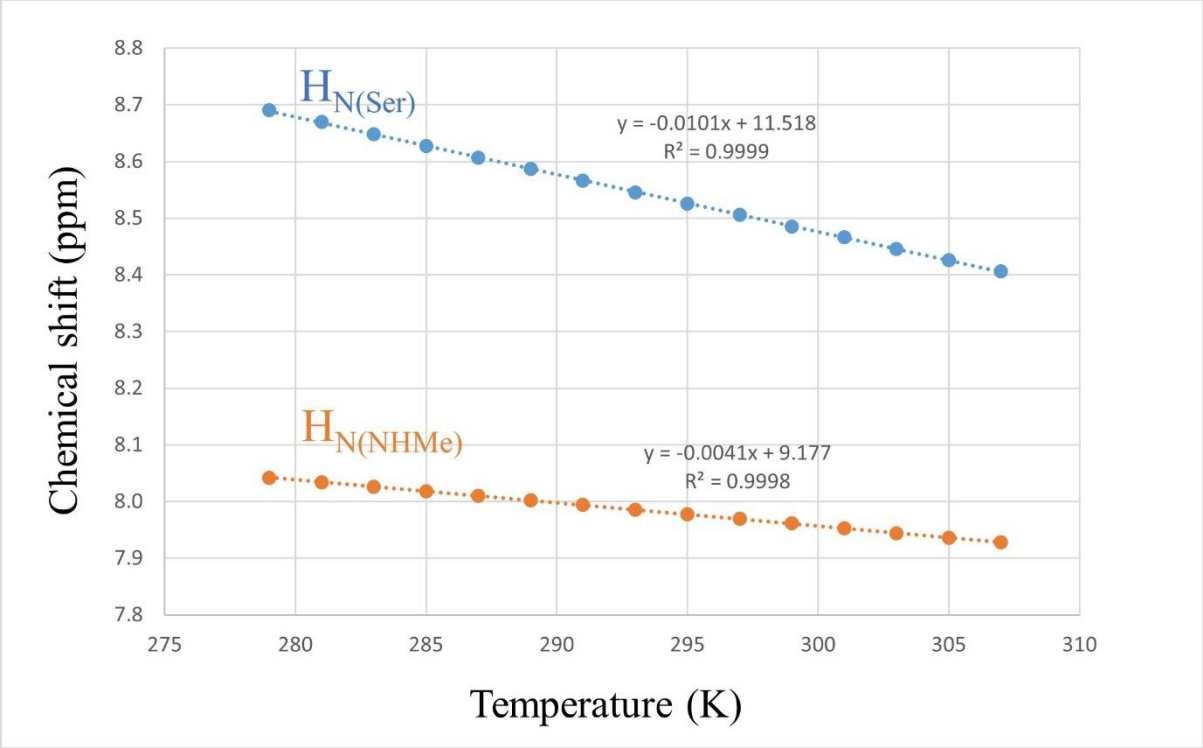


Figure S9. $d_{\alpha N(i+1,\ i+3)}$ distance distribution (in Å, A) and Ramachandran plot of φ and ψ dihedral angles in the serine (*i+2*) residue (B). Level of theory for the MD simulation: OPLS-AA, 1000 ns, 300 K.

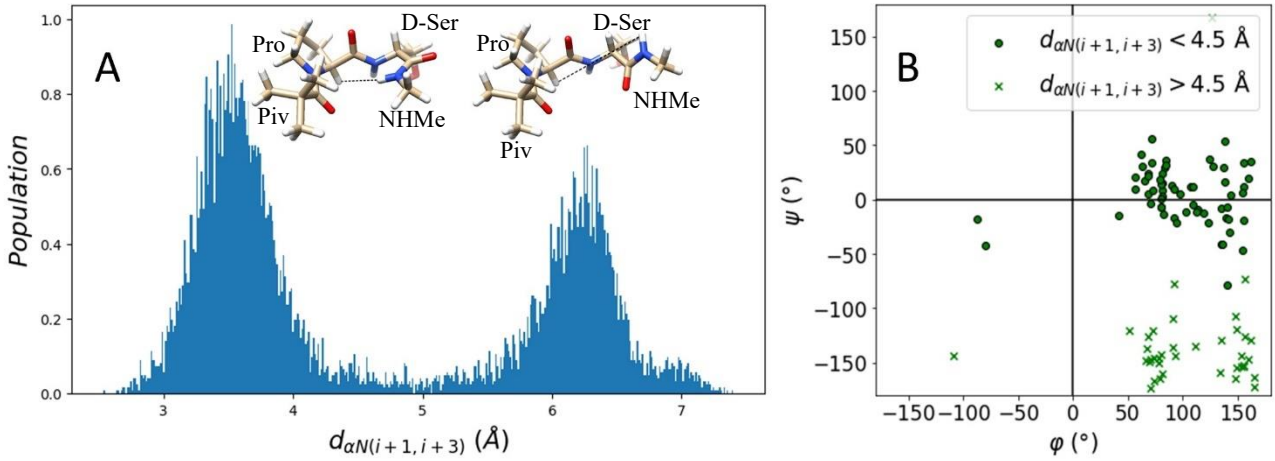


Figure S10. Distribution of the $d_{\alpha N(i+1,\, i+2)}$ values on the Ramachandran plot of φ and ψ dihedral angles in the proline $(i+1)$ residue. Level of theory for the MD simulation: CHARMM27, TIP3P, 1000 ns, 300 K.

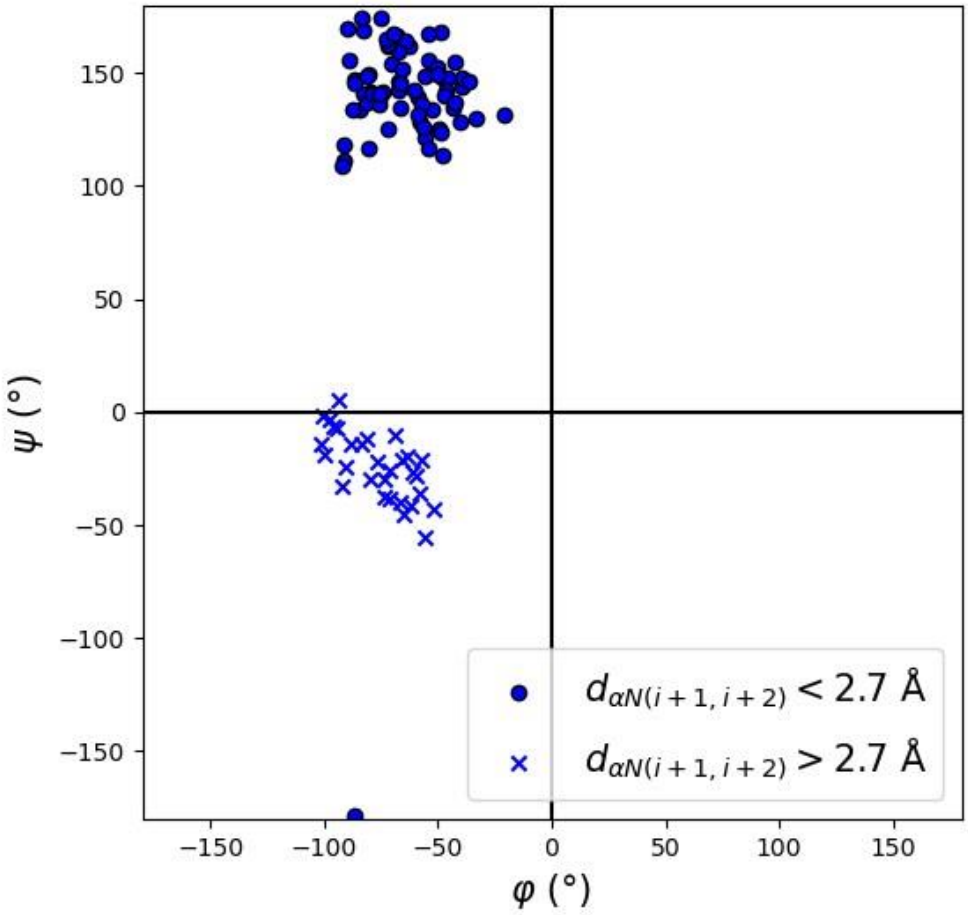


Figure S11. Distribution of the ω angle values for various water models: TIP3P (A), TIP4P (B), TIP5P (C), TIPS3P (D). Level of theory for the MD simulation: CHARMM27, 1000 ns, 300 K.

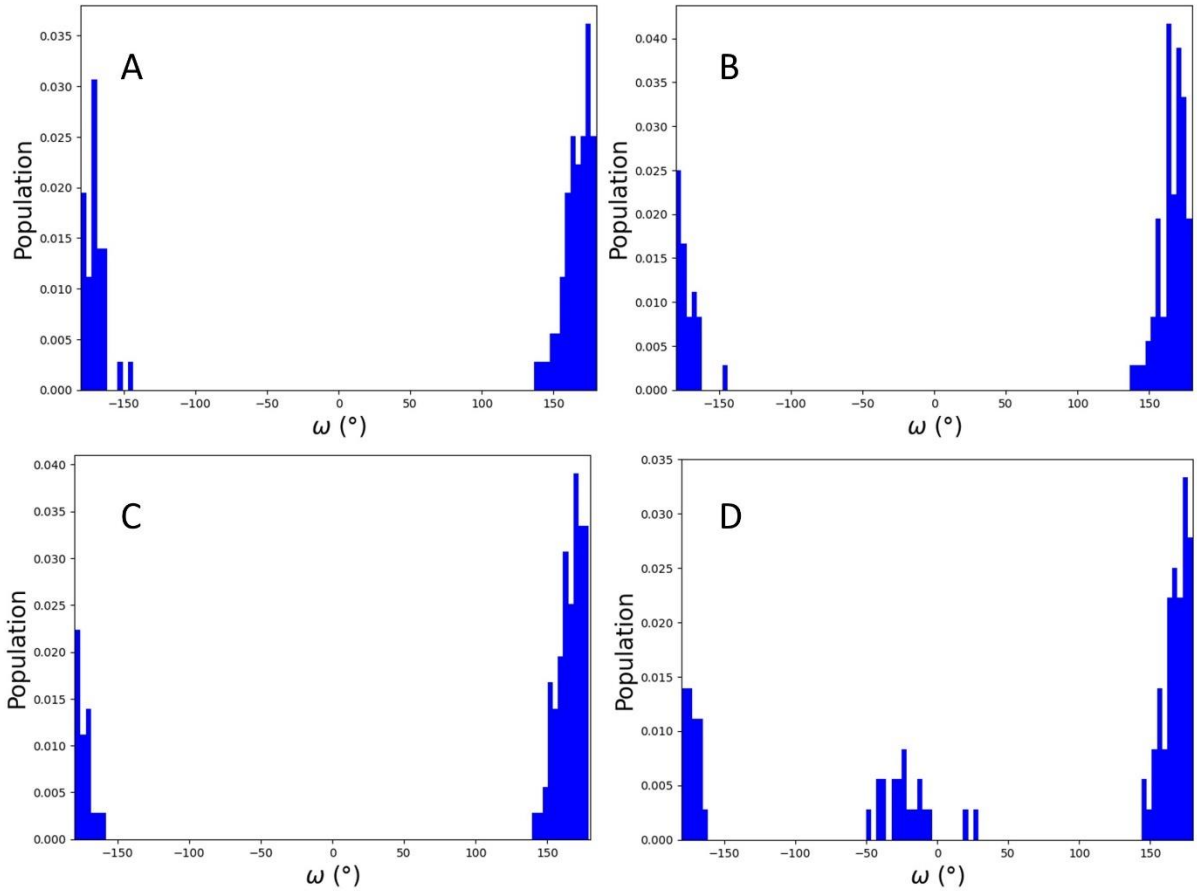


Table S2. Average distances over the dominant population Pop. 1 (corresponding to a ψ_{i+1} dihedral angle value between 100 ° and 180 °), over the week one Pop. 2 (corresponding to a ψ_{i+1} between -50 ° and 10 °) or over all 1000 geometries generated during MD using CHARMM27 force field within the TIP5P water model at 300K, and experimental ones (from our own NMR analyses (Figures S4 and S5)) for the Piv-Pro-D-Ser-NHMe peptide in water. The values in parentheses indicate the error. This error was set to 10 % for the experimental values. For theoretical ones, it is equal to the standard deviation.

	Pop. 1	**Pop. 2**	**Total**	**Exp.**
$\mathbf{d_{\alpha N(i+1,\, i+2)}}$ **(Å)**	2.3 (0.2)	3.5 (0.1)	2.5 (0.5)	2.0 (0.2)
$\mathbf{d_{\alpha N(i+1,\, i+3)}}$ **(Å)**	3.9 (0.5)	5.4 (0.3)	4.1 (0.7)	3.8 (0.4)
$\mathbf{d_{NN(i+2,\, i+3)}}$ **(Å)**	2.6 (0.4)	2.5 (0.3)	2.6 (0.3)	2.5 (0.3)

Figure S12. Ramachandran plots of φ and ψ dihedral angles of alanine residue in the GAG^{+} peptide. The peptide was initially constructed using the Chimera software with the φ and ψ dihedral angle in the PolyProline-II area and a Chloride ion was added. Then, geometries were obtained by the optimized MD protocol (1000 ns MD simulation using the CHARMM27 force field and the TIP5P water model at 300 K).

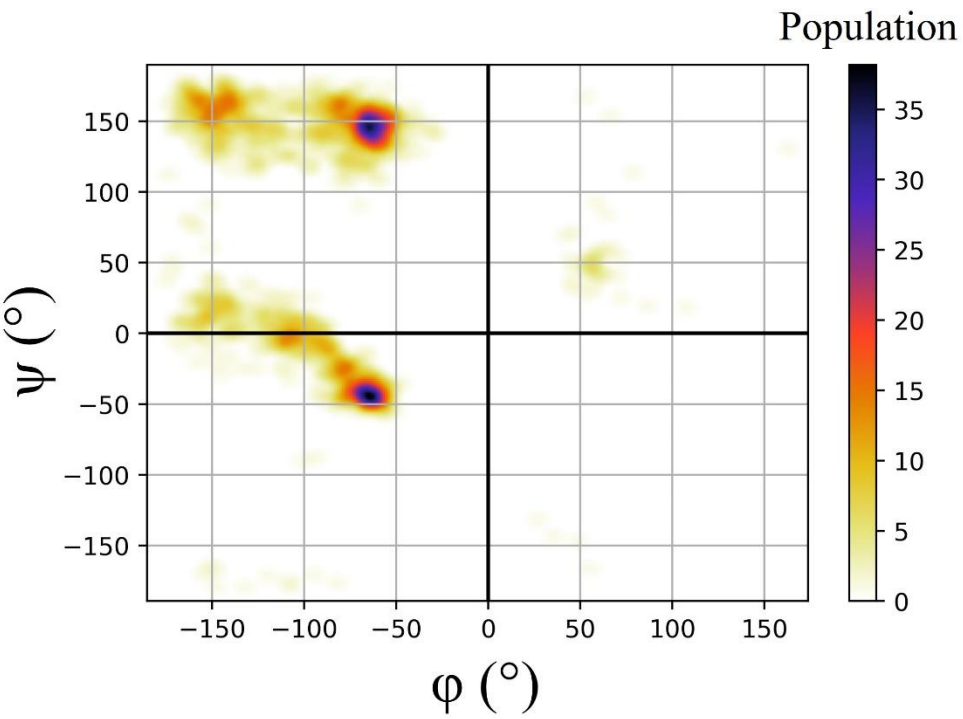
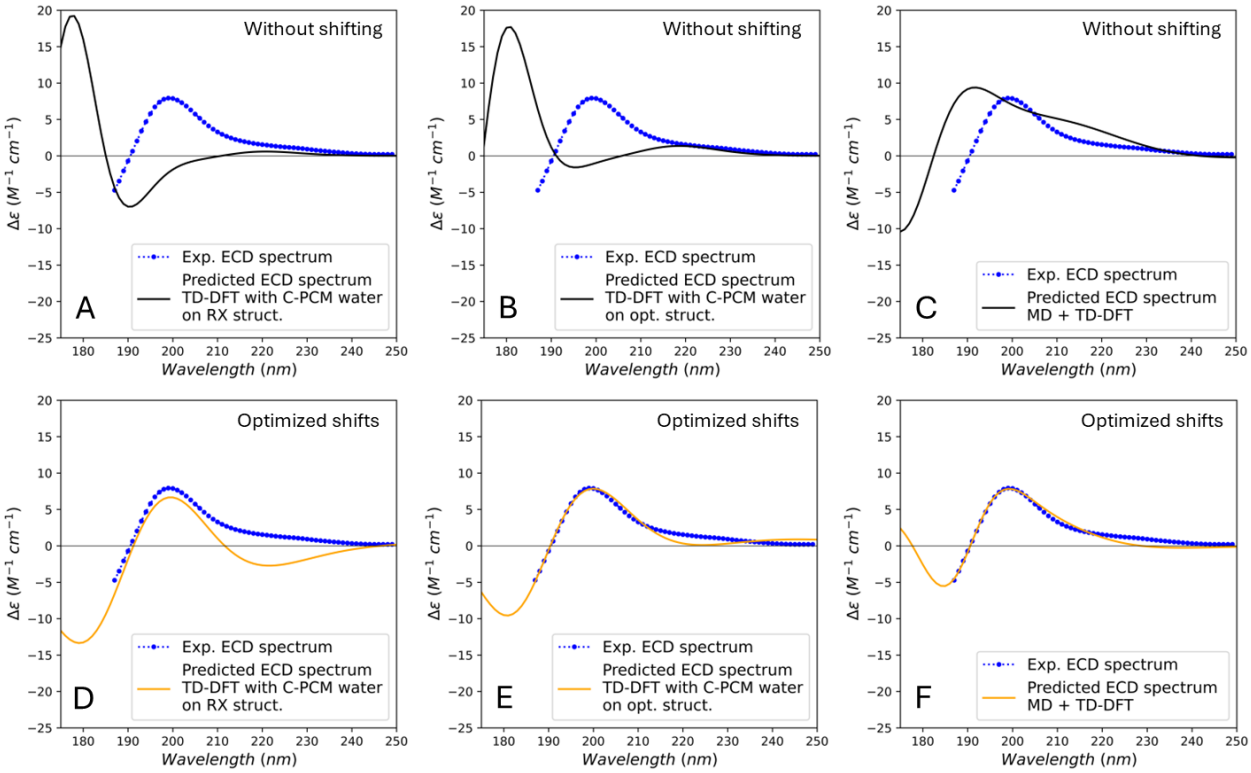


Figure S13. Comparison of the different theoretical protocols: TD-DFTa on the RX structure (left); TD-DFT on the optimizeda structure (middle); and MDb+TD-DFTb (right) without (up) or with (down) the application of optimized shifts to simulate ECD spectrum of the Piv-Pro-D-Ser-NHMe peptide.



a) Optimization and TD-DFT were performed by the Gaussian 16 software (CAM-B3LYP, 6-31+G*, C-PCM). The first 20 excited states were calculated by TD-DFT.
b) MD was carried out with Gromacs (CHARMM27, TIP5P, 300K) and TD-DFT (CAM-B3LYP, 6-31+G*) was done with the Dalton program on the 20 first excited states within the PE model for solvent effects.

Table S3. Optimized shifting parameters for the different theoretical protocols used in Figure S14

Optimized parameters	TD-DFT RX	TD-DFT opt. RX	MD +TD-DFT
Δ	0.50 eV	0.49 eV	0.29 eV
$\boldsymbol{\Delta E_<}$	0.71 eV	0.57 eV	0.36 eV
$\boldsymbol{\Delta E_>}$	0.54 eV	-0.1 eV	-0.25 eV
$\boldsymbol{E_b}$	218.59 nm	224 nm	186.84 nm

Mean Absolute Error (MAE) calculation on the different ECD theoretical protocols.

$$MAE = \frac{1}{n}\sum_{i}^{n}\left|y_i^{pred} - y_i^{exp}\right|$$

where n= 62 is the size of y^{pred} and y^{pred}. y_i^{exp} are the experimental values sampled every 1.0 nm from 187 to 259 nm and y_i^{pred} are the theoretical ones sampled on the experimental spectrum.

Theoretical protocol	MAE (M^{-1} cm^{-1})
TD-DFT RX	3.25
TD-DFT RX + optimized param.	1.98
TD-DFT opt.	2.59
TD-DFT opt. + optimized param.	0.51
MD+TD-DFT	1.94
MD+TD-DFT + optimized param.	0.46

Experimental ECD spectrum of the Piv-Pro-D-Ser-NHMe peptide in water.

First column corresponds to the wavelength in nm and the second one to the Rotatory strength in M^{-1} cm^{-1}. The spectrum was plotted in Figure 3

185;-5.69417829
186;-5.572437841
187;-4.730442693
188;-3.476046089
189;-2.075427532
190;-0.715448757
191;0.643223166
192;2.047255913
193;3.417343845
194;4.679047908
195;5.797301395
196;6.713826562
197;7.380412371
198;7.773681019
199;7.927016374
200;7.884718011
201;7.658247423
202;7.300667071
203;6.841328078
204;6.304426925
205;5.736325045
206;5.176652517
207;4.626561552
208;4.119011522
209;3.669345058
210;3.282929048
211;2.960918739
212;2.68681322
213;2.45219527
214;2.256391753
215;2.091391753
216;1.951782899
217;1.828511219
218;1.715709521
219;1.622701637
220;1.54753487
221;1.475015161
222;1.402859309
223;1.331837477
224;1.26504245
225;1.207613705
226;1.159396604
227;1.110375985
228;1.056079442
229;0.994232868
230;0.925006064
231;0.852501516
232;0.78012735
233;0.713204973
234;0.65189812
235;0.593853851
236;0.539693754
237;0.487437841
238;0.434536082
239;0.381831413
240;0.333192844
241;0.29324621
242;0.261532747
243;0.23753396
244;0.222976046

245;0.213133414
246;0.208189206
247;0.207142207
248;0.205425106
249;0.201898423
250;0.198457247
251;0.193949363
252;0.190811704
253;0.189732868
254;0.190991207
255;0.193501819
256;0.196352941
257;0.1982453
258;0.197494239
259;0.192027896
260;0.179388417

Experimental ECD spectrum of the Piv-Pro-D-Ser-NHMe peptide in acetonitrile.

First column corresponds to the wavelength in nm and the second one to the Rotatory strength in M^{-1} cm^{-1}. The spectrum was plotted in Figure 3

189;0.105510612
190;1.330639782
191;2.968523347
192;4.694906004
193;6.52865373
194;8.433201941
195;10.25812614
196;11.89766525
197;13.31540327
198;14.37389327
199;15.06701031
200;15.34035779
201;15.23629472
202;14.76434203
203;13.97704669
204;12.96325045
205;11.796604
206;10.55097029
207;9.307398423
208;8.116130988
209;7.005245603
210;6.015342632
211;5.159157065
212;4.438902365
213;3.852486355
214;3.379927229
215;3.01168587
216;2.728826562
217;2.509718011
218;2.343677987
219;2.220415403
220;2.136943602
221;2.089096422
222;2.065
223;2.055191025
224;2.049075197
225;2.0344906
226;2.018741662
227;1.993168587
228;1.959287447
229;1.917801698
230;1.859129776
231;1.786352335
232;1.704163129
233;1.608411158
234;1.502122498
235;1.388608247
236;1.266055185
237;1.142686477
238;1.021740449
239;0.900557914
240;0.786312917
241;0.679235901
242;0.580503335
243;0.49168587
244;0.409081261
245;0.331658581
246;0.261962705
247;0.197597635
248;0.140428745
249;0.092062159
250;0.053979381
251;0.028686568
252;0.015364099
253;0.007074864
254;-0.000754721

255;-0.006790237
256;-0.011888417
257;-0.016489418
258;-0.020616737
259;-0.02387735
260;-0.025461552

Experimental ECD spectrum of the Piv-Pro-D-Ser-NHMe peptide in methanol.

First column corresponds to the wavelength in nm and the second one to the Rotatory strength in M^{-1} cm^{-1}. The spectrum was plotted in Figure 3

192;0.983617344	213;0.817474227	234;0.634846574
193;1.848238326	214;0.651135537	235;0.598309582
194;2.878298969	215;0.522055791	236;0.560122195
195;3.7094906	216;0.430079139	237;0.515291086
196;4.445300182	217;0.363816798	238;0.466300182
197;4.98793208	218;0.320915797	239;0.41389721
198;5.301334142	219;0.299802608	240;0.358993329
199;5.445148575	220;0.293808975	241;0.302106125
200;5.466252274	221;0.29862826	242;0.24705852
201;5.280927835	222;0.316183141	243;0.197008187
202;5.010612492	223;0.346536204	244;0.156465737
203;4.652001213	224;0.398030625	245;0.123249545
204;4.219132808	225;0.466045482	246;0.096375379
205;3.748938751	226;0.541017283	247;0.074372347
206;3.259833232	227;0.609741358	248;0.056090055
207;2.780303214	228;0.665218314	249;0.040967253
208;2.335721649	229;0.70447544	250;0.031613099
209;1.931137053	230;0.725012129	251;0.02822923
210;1.578311098	231;0.722210431	252;0.029799879
211;1.280494239	232;0.702013341	253;0.034090358
212;1.02723772	233;0.67077926	254;0.037446938

255;0.04086416 257;0.050085203 259;0.072647968

256;0.044730139 258;0.058614615 260;0.095160703

Theoretical ECD spectrum of the Piv-Pro-D-Ser-NHMe peptide in water (no shift).

First column corresponds to the wavelength in nm and the second one to the Rotatory strength in M^{-1} cm^{-1}. The spectrum was plotted in Figure 7

170.0;-6.100338243	191.0;9.35456704	212.0;4.8231280259
171.0;-7.5476607922	192.0;9.3787191968	213.0;4.6697947908
172.0;-8.7909259361	193.0;9.2602028921	214.0;4.5085004598
173.0;-9.7325953872	194.0;9.0353019376	215.0;4.3386375504
174.0;-10.2940287588	195.0;8.7377567814	216.0;4.160090954
175.0;-10.422702011	196.0;8.3972030262	217.0;3.9731753976
176.0;-10.0966663772	197.0;8.038249903	218.0;3.7785716488
177.0;-9.3258571632	198.0;7.6801240701	219.0;3.5772631174
178.0;-8.1502636278	199.0;7.3367726776	220.0;3.3704738715
179.0;-6.6353433124	200.0;7.017305915	221.0;3.1596087773
180.0;-4.8653598703	201.0;6.7266601742	222.0;2.946196363
181.0;-2.9355087655	202.0;6.4663747196	223.0;2.7318350172
182.0;-0.9437538962	203.0;6.2353934082	224.0;2.5181431573
183.0;1.0167661035	204.0;6.0308248904	225.0;2.3067140238
184.0;2.8640683237	205.0;5.8486167824	226.0;2.0990757097
185.0;4.5319298968	206.0;5.6841192824	227.0;1.896656949
186.0;5.9725582734	207.0;5.5325301476	228.0;1.7007590371
187.0;7.1573601267	208.0;5.3892252174	229.0;1.5125340807
188.0;8.0760376639	209.0;5.2499867371	230.0;1.3329695713
189.0;8.7343750595	210.0;5.111146055	231.0;1.1628790817
190.0;9.1511445396	211.0;4.9696585794	232.0;1.0028986985

233.0;0.8534886477
234.0;0.7149394515
235.0;0.5873818687
236.0;0.4707998333
237.0;0.3650455983
238.0;0.2698563192
239.0;0.1848713715
240.0;0.1096497689
241.0;0.0436871431
242.0;-0.0135681553
243.0;-0.0627001721
244.0;-0.1043119219
245.0;-0.1390135612
246.0;-0.1674122304
247.0;-0.1901035543
248.0;-0.2076647425
249.0;-0.2206491801
250.0;-0.2295823738

Theoretical ECD spectrum of the Piv-Pro-D-Ser-NHMe peptide in water (optimized shifts).

First column corresponds to the wavelength in nm and the second one to the Rotatory strength in M^{-1} cm^{-1}. The spectrum was plotted in Figure 7

170.0;3.2450119939	191.0;0.6101029581	212.0;3.2511050117
171.0;3.431302695	192.0;2.1182149626	213.0;2.930757438
172.0;3.46382238	193.0;3.5449732826	214.0;2.628175618
173.0;3.3122495853	194.0;4.8205080263	215.0;2.3422506866
174.0;2.9574833398	195.0;5.8932242156	216.0;2.0722367138
175.0;2.3947642534	196.0;6.7321463662	217.0;1.8177414336
176.0;1.6355804487	197.0;7.3267047664	218.0;1.5786601406
177.0;0.7082079931	198.0;7.6843720229	219.0;1.3550820446
178.0;-0.3430818905	199.0;7.8268034713	220.0;1.1471908023
179.0;-1.4600343846	200.0;7.785230475	221.0;0.9551737261
180.0;-2.5735470727	201.0;7.5958137421	222.0;0.7791480347
181.0;-3.6078928446	202.0;7.2955207123	223.0;0.6191077589
182.0;-4.4859916797	203.0;6.9188938123	224.0;0.4748915341
183.0;-5.1355690265	204.0;6.4958706898	225.0;0.3461693296
184.0;-5.4957932661	205.0;6.0506392187	226.0;0.2324449311
185.0;-5.5236957926	206.0;5.6013806755	227.0;0.1330704718
186.0;-5.1994641554	207.0;5.1606811819	228.0;0.047269274
187.0;-4.5296710534	208.0;4.7363698481	229.0;-0.025836441
188.0;-3.5477119903	209.0;4.3325602194	230.0;-0.0871959383
189.0;-2.3111446797	210.0;3.9507148357	231.0;-0.1378008276
190.0;-0.896156291	211.0;3.5906067153	232.0;-0.1786581692

233.0;-0.2107670159
234.0;-0.2350990553
235.0;-0.2525835712
236.0;-0.2640965962
237.0;-0.2704538719
238.0;-0.2724070764
239.0;-0.2706426995
240.0;-0.2657829366
241.0;-0.2583880138
242.0;-0.2489594302
243.0;-0.2379436977
244.0;-0.225736259
245.0;-0.2126853576
246.0;-0.1990957207
247.0;-0.1852319827
248.0;-0.1713218318
249.0;-0.1575588971
250.0;-0.1441054166

Theoretical ECD spectrum of the GAG^+ peptide in water (no shift).

First column corresponds to the wavelength in nm and the second one to the Rotatory strength in M^{-1} cm^{-1}. The spectrum was plotted in Figure 8

183.0;-6.4577515665
184.0;-6.5077482388
185.0;-6.4335463527
186.0;-6.2470436584
187.0;-5.9641749
188.0;-5.6031768759
189.0;-5.1830801431
190.0;-4.7224970396
191.0;-4.2387226754
192.0;-3.7471293293
193.0;-3.2608144318
194.0;-2.7904546268
195.0;-2.3443189992
196.0;-1.9283994874
197.0;-1.5466228834
198.0;-1.201114958
199.0;-0.89249246
200.0;-0.6201630328
201.0;-0.3826167652
202.0;-0.1776964956
203.0;-0.002837323
204.0;0.1447308999
205.0;0.2678211076
206.0;0.3691666799
207.0;0.4513368258
208.0;0.5166854032
209.0;0.5673276401
210.0;0.6051386126
211.0;0.6317672957
212.0;0.6486604261
213.0;0.6570911727
214.0;0.658188574
215.0;0.652964747
216.0;0.6423379126
217.0;0.6271502186
218.0;0.6081801477
219.0;0.5861499102
220.0;0.5617286598
221.0;0.5355326143
222.0;0.5081232495
223.0;0.48000468
224.0;0.4516211889
225.0;0.4233556458
226.0;0.395529305
227.0;0.3684032231
228.0;0.34218131
229.0;0.3170148428
230.0;0.293008139
231.0;0.2702250068
232.0;0.2486955647
233.0;0.2284230365
234.0;0.2093901832
235.0;0.1915651022
236.0;0.1749062113
237.0;0.159366318
238.0;0.1448957554
239.0;0.1314446291
240.0;0.1189642707
241.0;0.1074080247
242.0;0.0967315137
243.0;0.0868925263
244.0;0.0778506662
245.0;0.0695668799

246.0;0.0620029608 248.0;0.0488835495

247.0;0.0551211026 249.0;0.0432523721

Theoretical ECD spectrum of the GAG^+ peptide in water (optimized shifts).

First column corresponds to the wavelength in nm and the second one to the Rotatory strength in M^{-1} cm^{-1}. The spectrum was plotted in Figure 8

183.0;-3.4732573077
184.0;-3.8988200686
185.0;-4.3931165876
186.0;-4.9168317308
187.0;-5.4273190796
188.0;-5.882954934
189.0;-6.2470842248
190.0;-6.4911475247
191.0;-6.5967229986
192.0;-6.5563738383
193.0;-6.373342331
194.0;-6.060259727
195.0;-5.637133047
196.0;-5.1289176757
197.0;-4.5629867469
198.0;-3.9667707391
199.0;-3.365774729
200.0;-2.7821005256
201.0;-2.233520362
202.0;-1.7330791827
203.0;-1.2891509898
204.0;-0.9058439398
205.0;-0.5836379188
206.0;-0.3201435247
207.0;-0.1108877878
208.0;0.0499455861
209.0;0.1688696595
210.0;0.2525674165
211.0;0.3074923675
212.0;0.3395908907
213.0;0.3541319441
214.0;0.3556253094
215.0;0.3478082866
216.0;0.333681734
217.0;0.3155786374
218.0;0.2952512726
219.0;0.2739660085
220.0;0.2525975808
221.0;0.2317171038
222.0;0.2116701406
223.0;0.19264282
224.0;0.1747153162
225.0;0.1579030143
226.0;0.1421864216
227.0;0.1275313561
228.0;0.1139012044
229.0;0.1012631
230.0;0.0895897837
231.0;0.0788587046
232.0;0.0690496475
233.0;0.0601418708
234.0;0.0521114414
235.0;0.0449291858
236.0;0.0385594571
237.0;0.0329597448
238.0;0.0280810448
239.0;0.0238688347
240.0;0.0202644694
241.0;0.0172068148
242.0;0.0146339491
243.0;0.0124847972
244.0;0.0107005919
245.0;0.0092260926

246.0;0.0080105232

247.0;0.0070082167

248.0;0.0061789753

249.0;0.005488173

Computational protocol to determine β-turn reference distances

Reference distances for $d_{\alpha N(i+1,\, i+2)}$, $d_{\alpha N(i+1,\, i+3)}$ and $d_{NN(i+2,\, i+3)}$ distances of Piv-Pro-D-Ser-NHMe in β-I, β-II and β-VIII turns were established employing a specific MD protocol. In this case, φ_{i+1}, ψ_{i+1}, φ_{i+2} and ψ_{i+2} values were adjusted to canonical angles according to the De Brevern classification. During the MD simulations, φ_{i+1}, ψ_{i+1}, φ_{i+2} and ψ_{i+2} angles were constrained within a range of 10°, following the same molecular dynamics process as described in the computational protocol, utilizing the CHARMM27 force field. A final production step of 10 ns was conducted, from which 100 geometries were extracted at a regular time step. The three reference distances $d_{\alpha N(i+1,\, i+2)}$, $d_{\alpha N(i+1,\, i+3)}$ and $d_{NN(i+2,\, i+3)}$ were calculated by averaging over these 100 geometries and are shown in Table 2.

References.

1 Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Petersson, G. A.; Nakatsuji, H.; Li, X.; Caricato, M.; Marenich, A. V.; Bloino, J.; Janesko, B. G.; Gomperts, R.; Mennucci, B.; Hratchian, H. P.; Ortiz, J. V.; Izmaylov, A. F.; Sonnenberg, J. L.; Williams-Young, D.; Ding, F.; Lipparini, F.; Egidi, F.; Goings, J.; Peng, B.; Petrone, A.; Henderson, T.; Ranasinghe, D.; Zakrzewski, V. G.; Gao, J.; Rega, N.; Zheng, G.; Liang, W.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Throssell, K.; Montgomery, J. A., Jr.; Peralta, J. E.; Ogliaro, F.; Bearpark, M. J.; Heyd, J. J.; Brothers, E. N.; Kudin, K. N.; Staroverov, V. N.; Keith, T. A.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A. P.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Millam, J. M.; Klene, M.; Adamo, C.; Cammi, R.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Farkas, O.; Foresman, J. B.; Fox, D. J. Gaussian 16, Revision C.01, Gaussian, Inc., Wallingford CT, 2016.

Annexe : Chapitre V

Table des matières

Figure S1: 1D ^{1}H NMR spectrum of the AFA peptide (0.5 mM) in water (90 % H_2O/10 % D_2O) recorded with the *zgesgp* pulse program at 293 K and pH 7.2

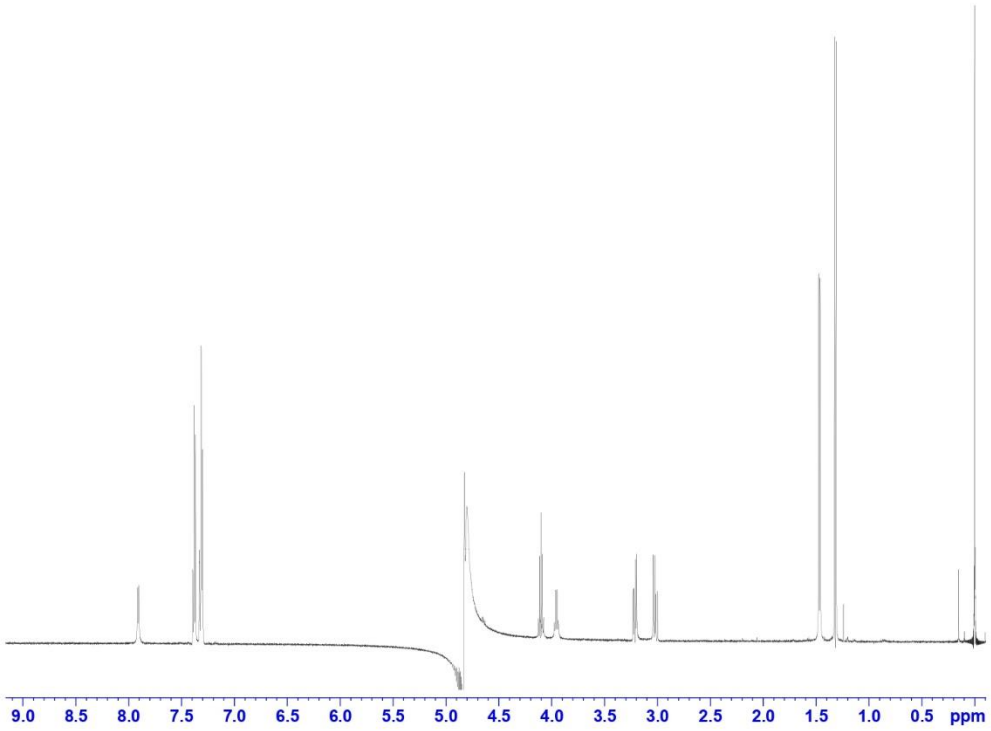


Table S1. Proton chemical shifts at 293 K and pH 7.2

Residue	H_N	H_α	H_β	Others
Ala1	-	3.96	1.47	-
Phe2	-	4.65	3.21 $H_{\beta3}$, 3.02 $H_{\beta2}$	7.38 H_δ 7.31 Hε,ζ
Ala3	7.91	4.10	1.32	-

Figure S2. 2D COSY ^{1}H spectrum of the AFA peptide (0.5 mM) in water (90 % H_2O/10 % D_2O) recorded with the *cosygpprqf* pulse program at 293 K and pH 7.2

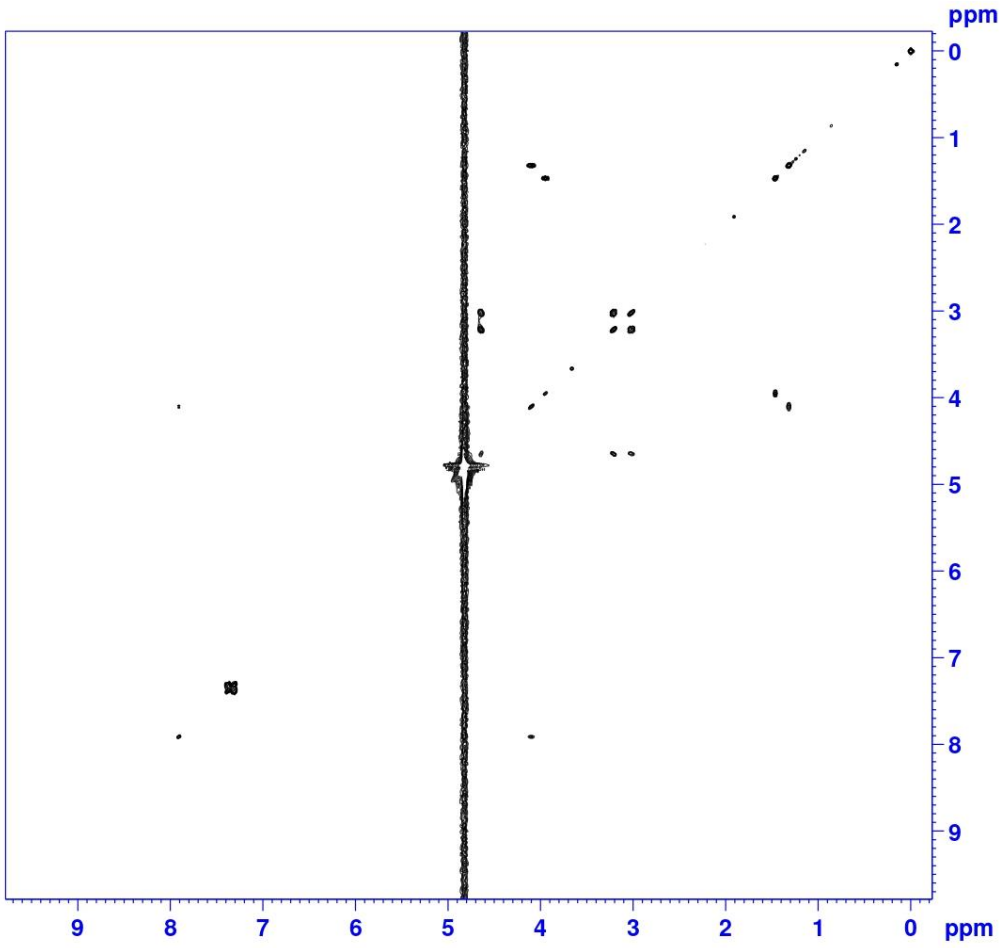


Figure S3. 2D TOCSY ^{1}H spectrum (blue: negative, red: positive) of the AFA peptide (0.5 mM) in water (90 % H_2O/10 % D_2O) recorded with the *dipsiesgpph* pulse program at 293 K and pH 7.2

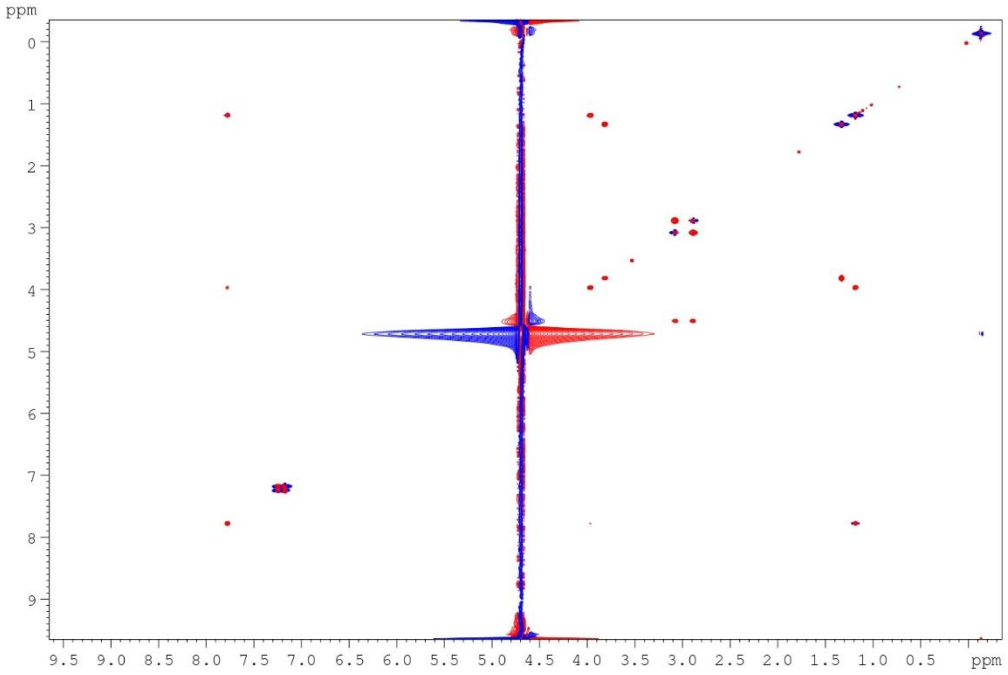


Figure S4. Variation of the HN chemical shifts with the Temperature between 275 K and 305 K at pH 4.0. 1D ^{1}H NMR spectra were recorded with the *zgesgp* pulse program.

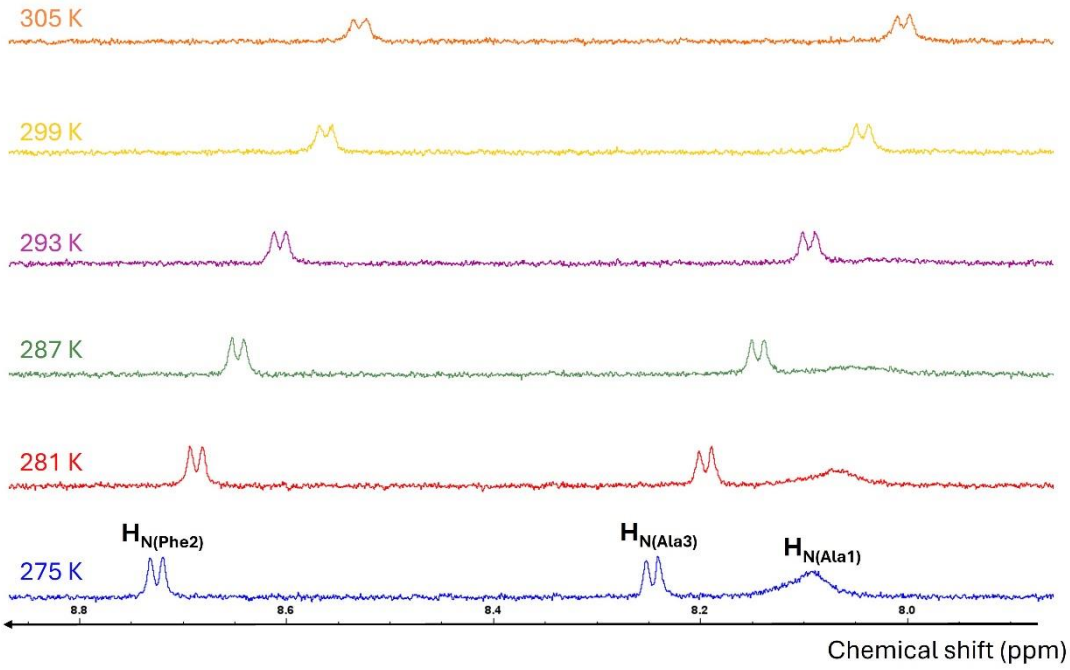


Table S2. Temperature coefficient of the amide protons

Proton	Temp. coef. at pH 7.2 (ppb K^{-1})	Temp. coef. At pH 4.0 (ppb K^{-1})
$H_{N(Ala1)}$	-	-
$H_{N(Phe2)}$	-	6.6
$H_{N(Ala3)}$	7.5	7.8

Figure S5. 2D NOESY ^{1}H spectrum (black: positive, red: negative) of the AFA peptide (0.5 mM) in water (90 % H_2O/10 % D_2O) recorded with the *noesygpph19* pulse program at 293 K and pH 7.2

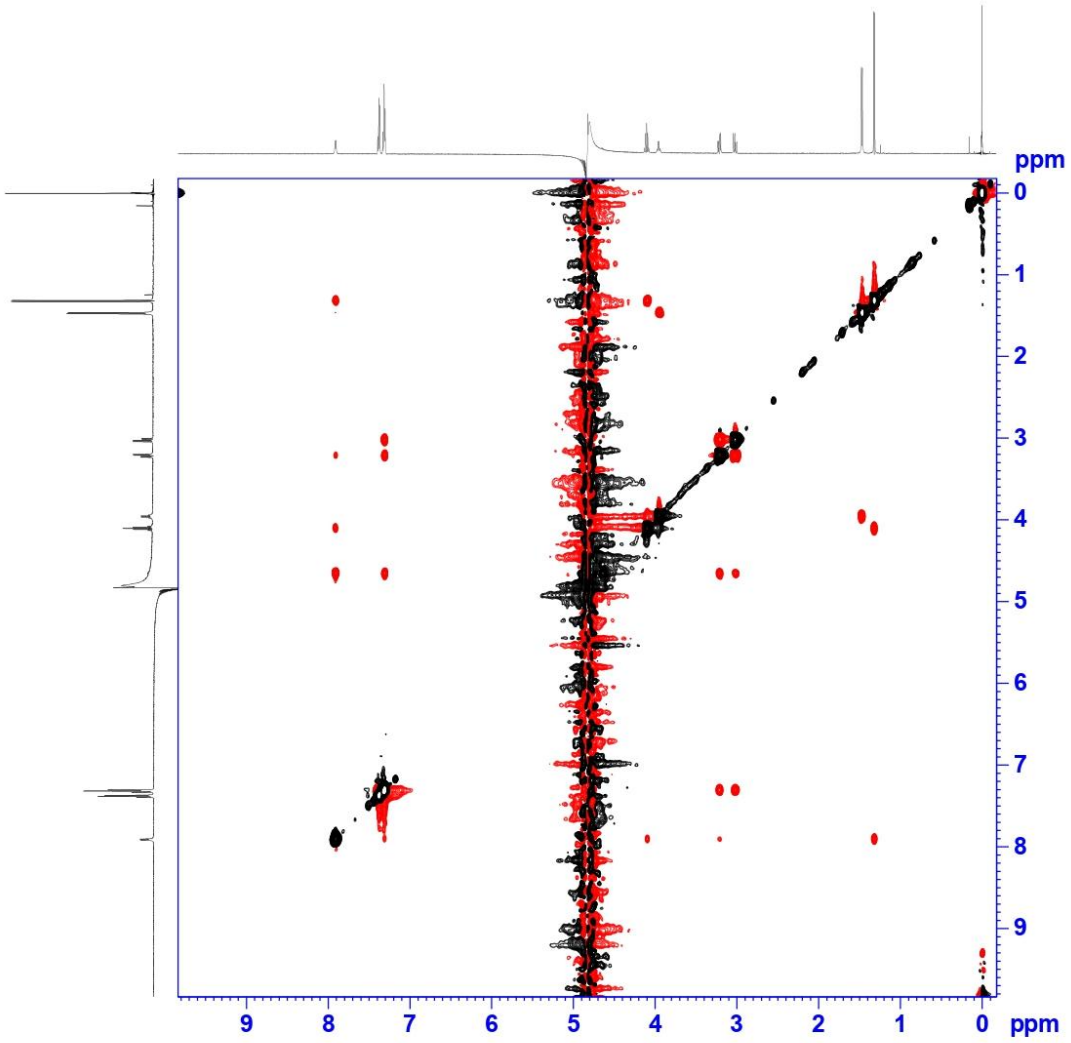


Figure S6. 2D NOESY ^{1}H spectrum (black: positive, red: negative) of the AFA peptide (0.5 mM) in water (90 % H_2O/10 % D_2O) recorded with the *noesygpph19* pulse program at 293 K and pH 7.2. Zoom on the correlations between the $H_{N(Ala3)}$ proton (in green) and the $H_{\alpha(Ala1)}$) (in black), the $H_{\alpha(Ala3)}$ (in cyan) and the $H_{\alpha(Phe2)}$ (in purple) protons depicted below.

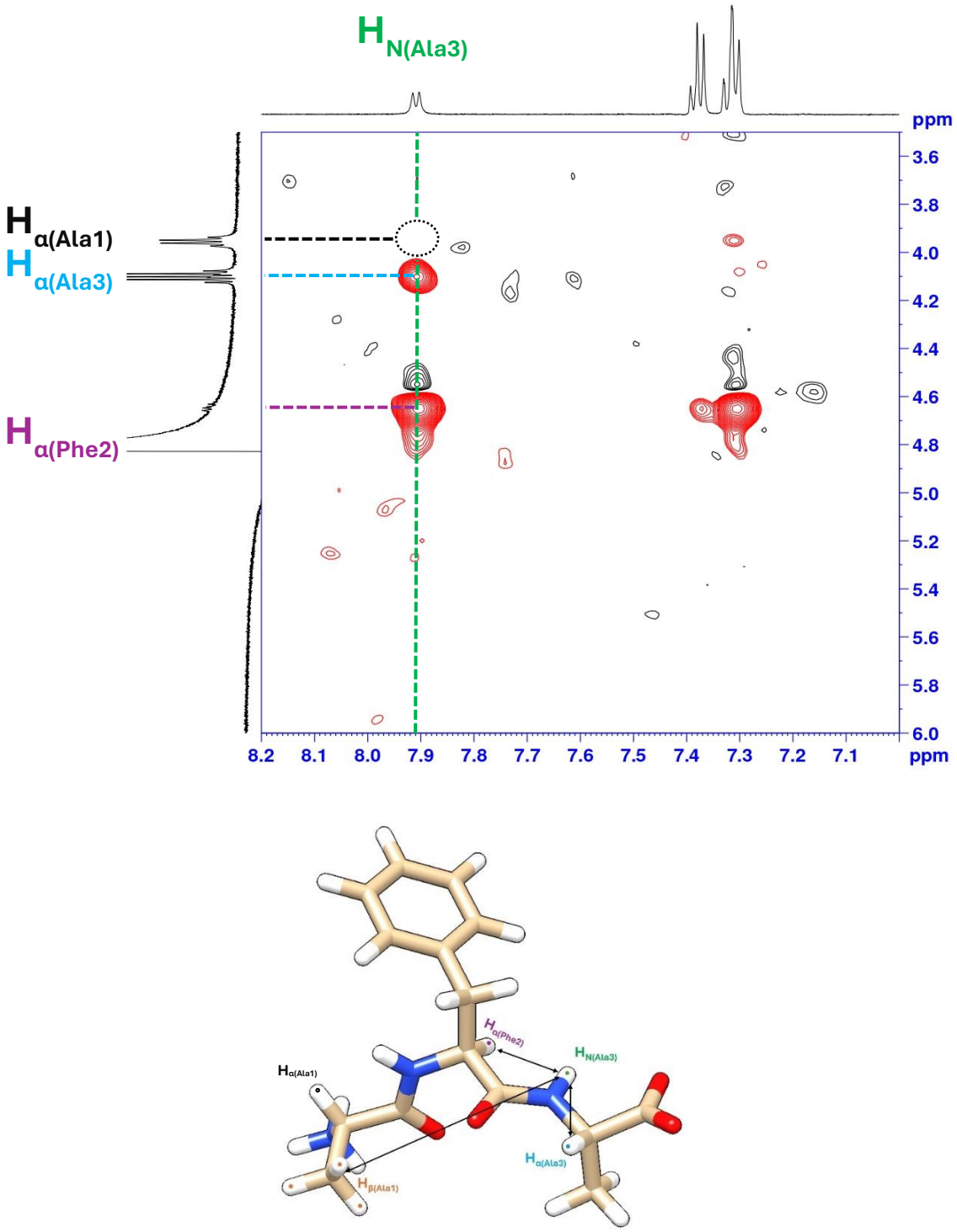
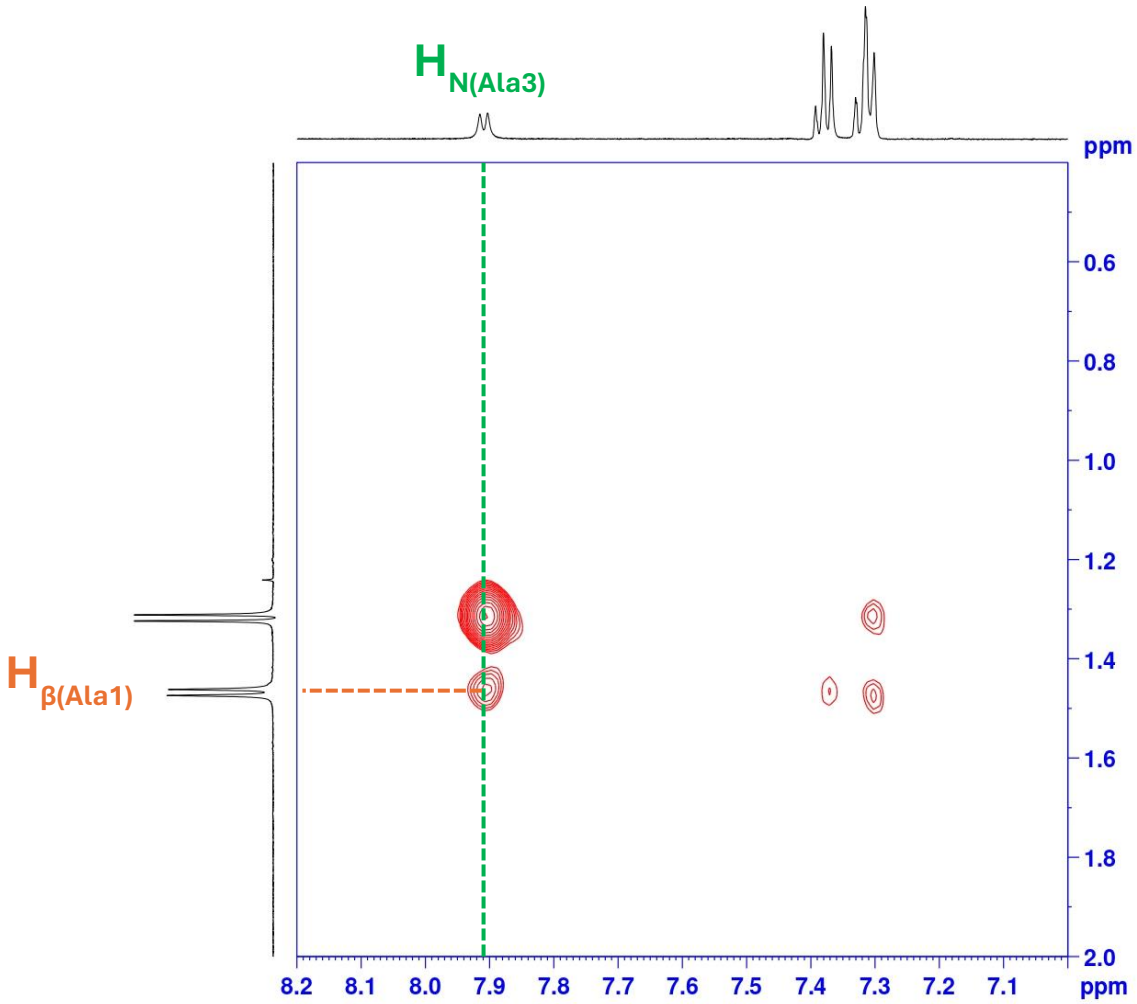


Figure S7. 2D NOESY ^{1}H spectrum (black: positive, red: negative) of the AFA peptide (0.5 mM) in water (90 % H_2O/10 % D_2O) recorded with the *noesygpph19* pulse program at 293 K and pH 7.2. Zoom on the correlation between the $H_{N(Ala3)}$ proton and the three $H_{B(Ala1)}$ protons depicted below.



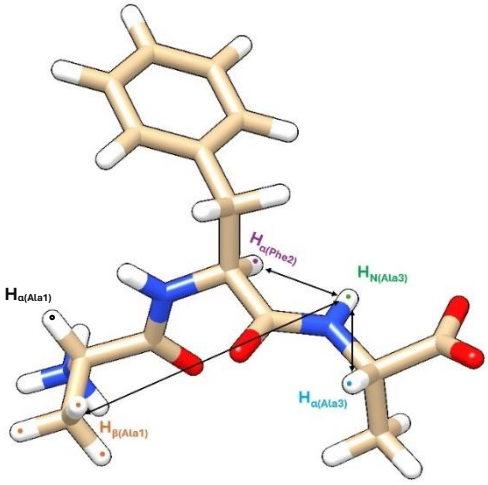


Figure S8. MD results on the Ramachandran plots using four different initial structures: α-helix (upper left) β-strand (upper right) PPII (lower left) and γ_{inv} (lower right). Simulations were performed with the OPLS-AA force fields at 300 K during 1 μs.

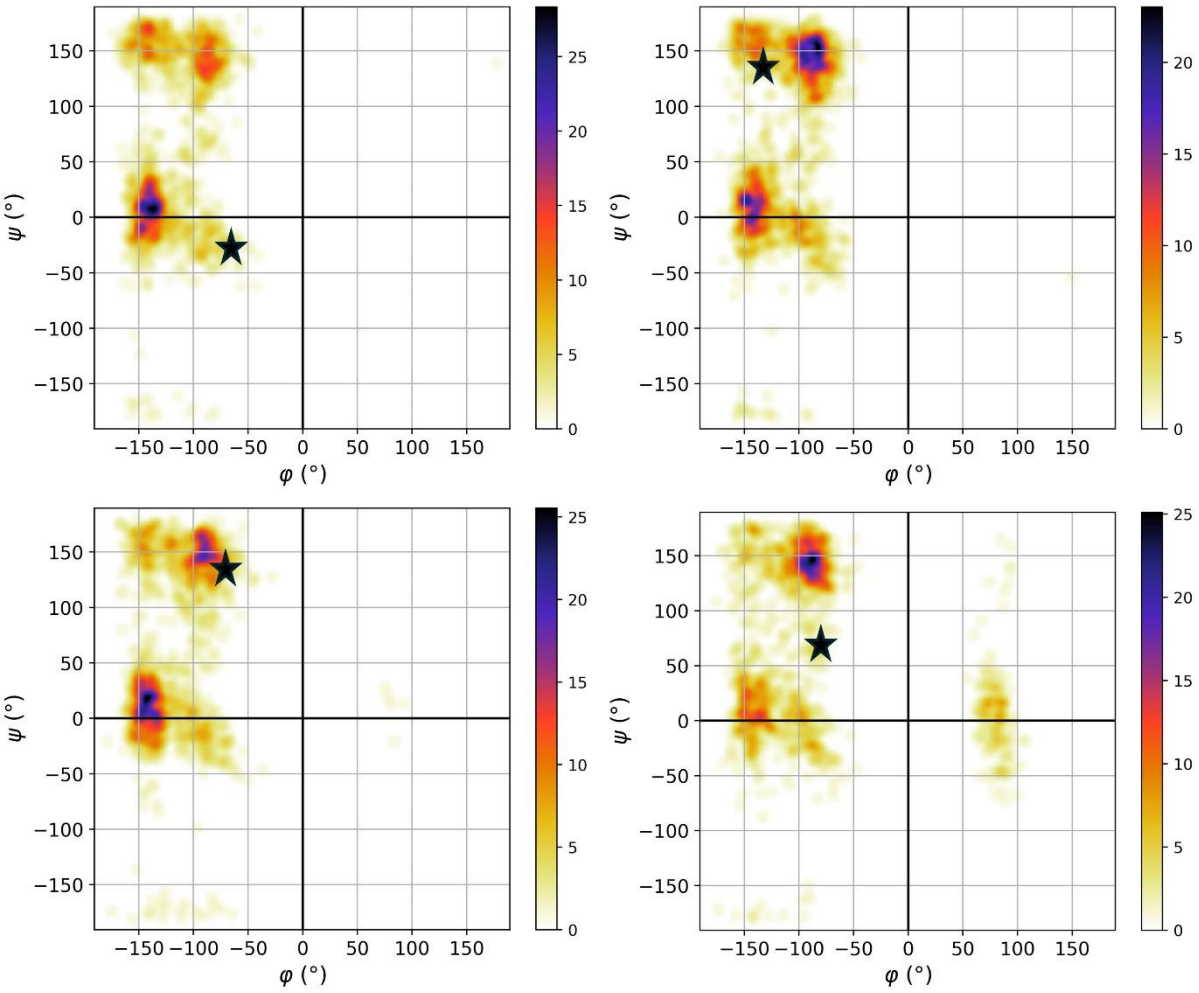


Figure S9. MD results on the Ramachandran plots using four different initial structures: α-helix (upper left) β-strand (upper right) PPII (lower left) and γ_{inv} (lower right). Simulations were performed with the CHARMM27 force fields at 300 K during 1 μs.

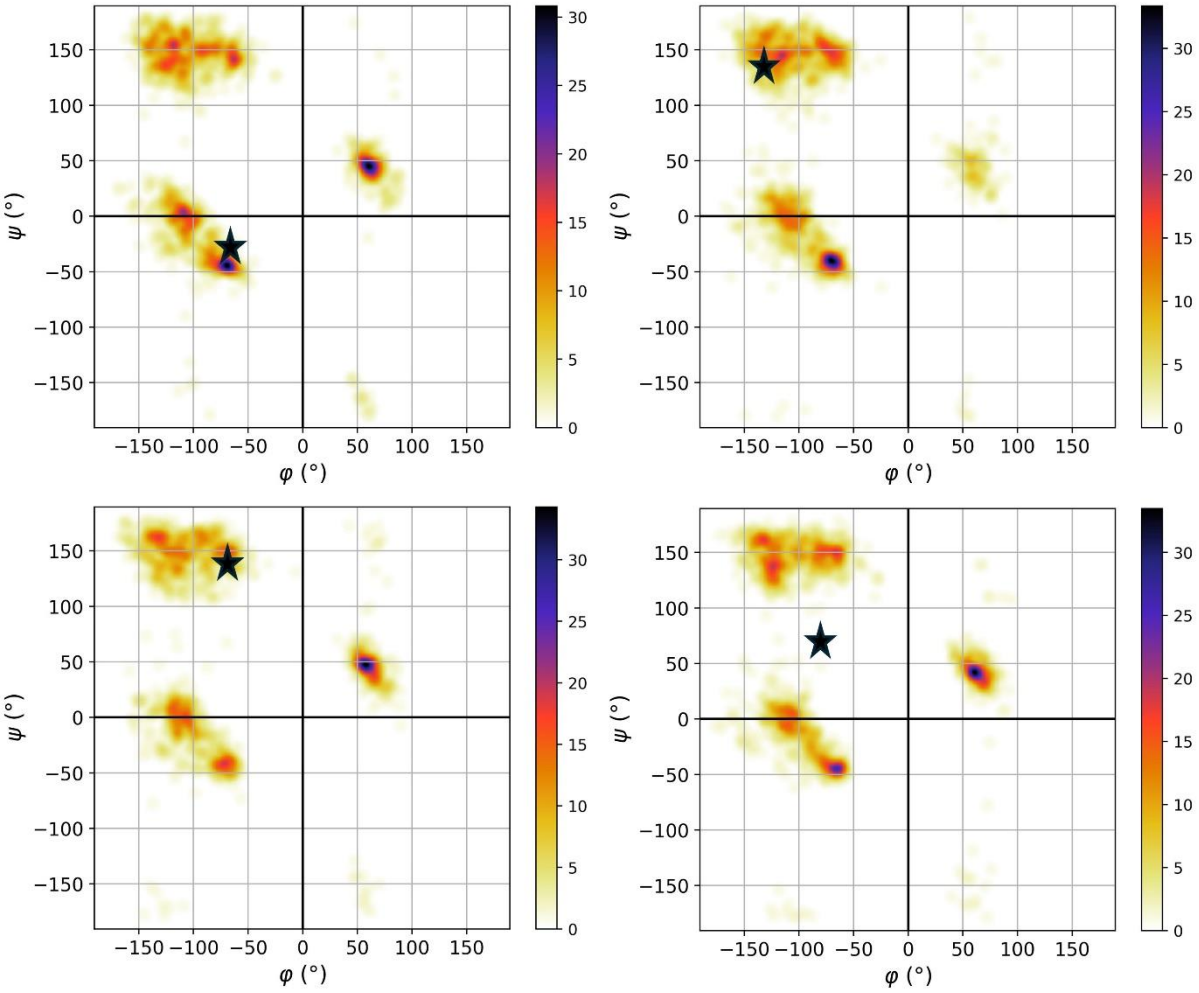


Table S3. Distances calculated on 1 μs MD simulations using four different initial structures. The experimental value includes a 20% error in brackets.

Distances	MD simulations								Exp.
	OPLS-AA				CHARMM27				
	α-helix	β-strand	PPII	γ_{inv}	α-helix	β-strand	PPII	γ_{inv}	
$H_{N(Ala3)}$-$H_{\alpha(Phe2)}$ (Å)	2.74	2.71	2.75	2.74	2.77	2.79	2.78	2.77	2.4 (0.48)
$H_{N(Ala3)}$-$H_{\alpha(Ala3)}$ (Å)	2.86	2.86	2.87	2.86	2.90	2.90	2.90	2.90	2.9 (0.85)
$H_{N(Ala3)}$-$H_{\beta(Ala1)}$ (Å)	5.06	5.00	4.85	5.06	5.30	5.24	5.38	5.36	4.2 (0.84)

Table S4. Distances calculated on 2μs MD simulations using the γ_{inv} conformation as initial structure. OPLS-AA force fields was used.

Distances	MD simulation	Experimental
	γinv	
$H_{N(Ala3)}$-$H_{\alpha(Phe2)}$	2.72 Å	2.4 (0.48) Å
$H_{N(Ala3)}$-$H_{\alpha(Ala3)}$	2.86 Å	2.9 (0.85) Å
$H_{N(Ala3)}$-$H_{\beta(Ala1)}$	5.00 Å	4.2 (0.84) Å

Table S5. $^{3}J\ H_{N(Ala3)}\text{-}H_{\alpha(Ala3)}$ Coupling constant calculation with several functionals and the 6-311++G** basis sets (FC contribution)

Functional	$^{3}J\ H_{N(Ala3)}\text{-}H_{\alpha(Ala3)}$ (Hz) CHARMM27	$^{3}J\ H_{N(Ala3)}\text{-}H_{\alpha(Ala3)}$ (Hz) OPLS-AA
SVWN	6.0	4.9
BLYP	8.1	6.6
B3LYP	7.9	6.5
CAM-B3LYP	7.7	6.3
PBE0	7.6	6.3

Figure S10. Correlation between the 6-311++G** basis set considering solely the FC contribution and the aug-cc-pVTZ-J basis set.

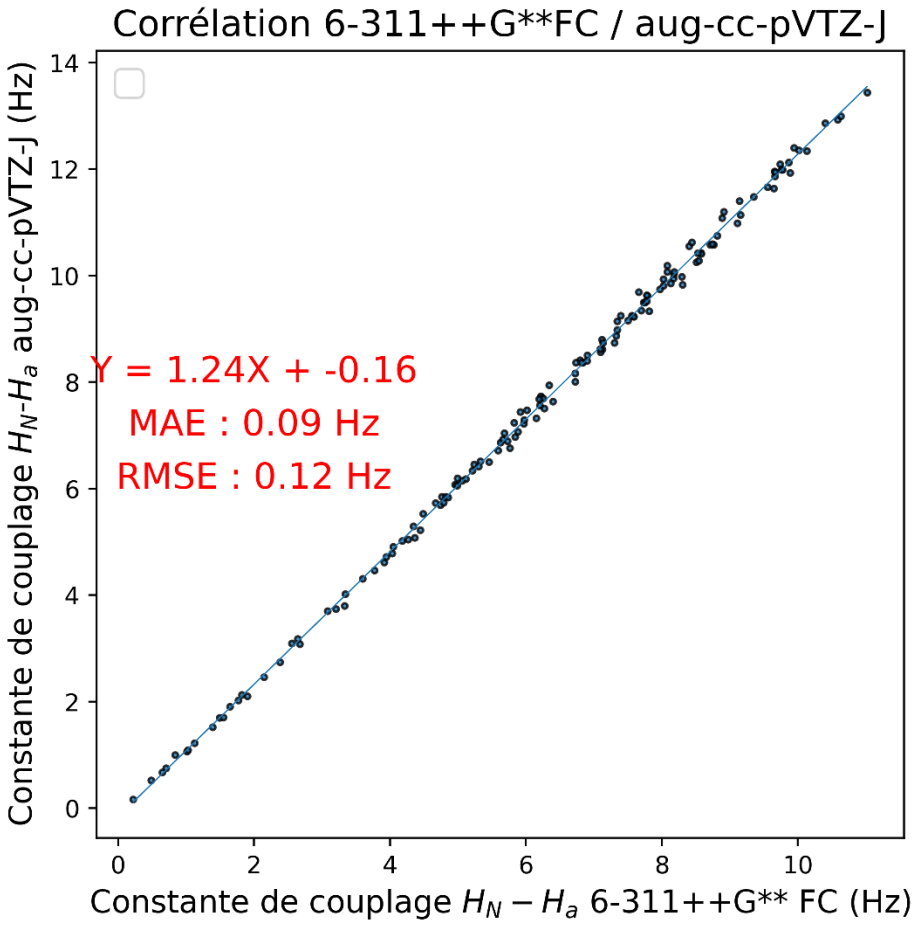


Figure S11. Clustering on dihedral angles using the k-means algorithm with 4 clusters. Ramachandran plots (on the left) and the corresponding average predicted spectrum (on the right).

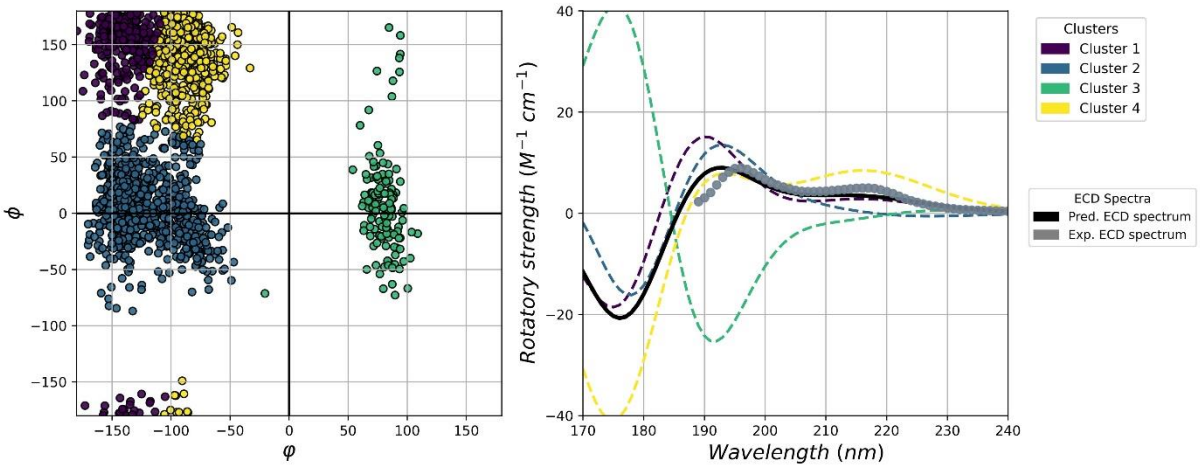


Figure S12. Clustering on dihedral angles using the k-means algorithm with 5 clusters. Ramachandran plots (on the left) and the corresponding average predicted spectrum (on the right).

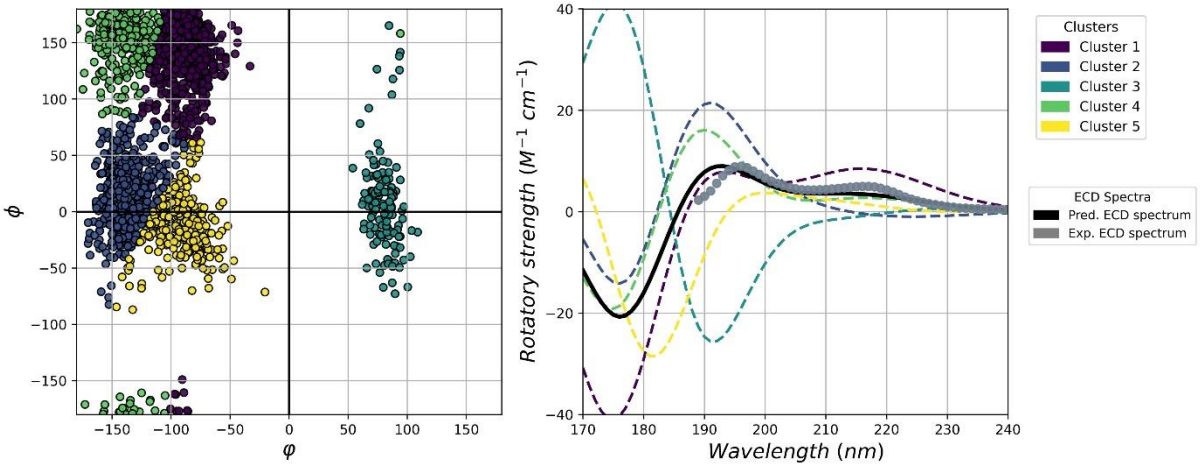
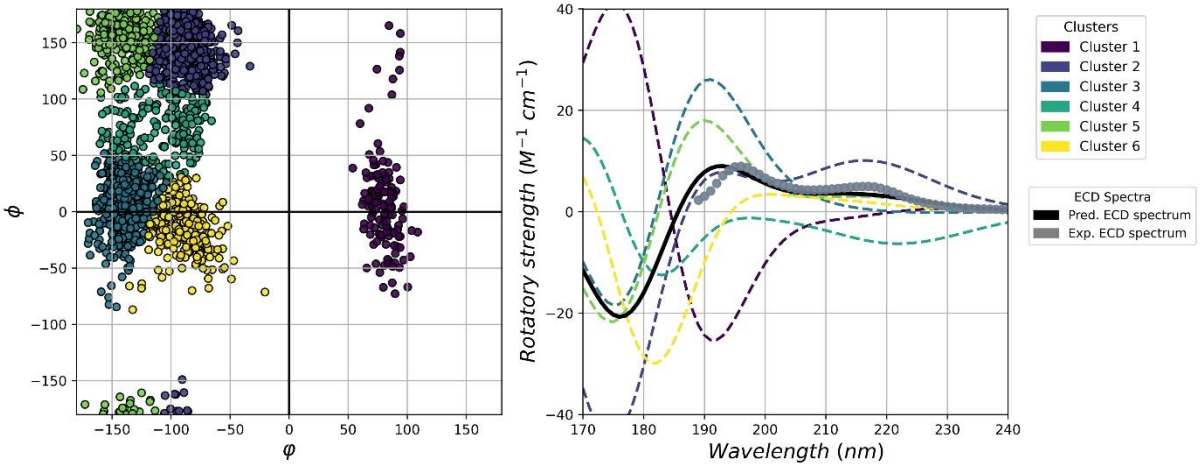


Figure S13. Clustering on dihedral angles using the k-means algorithm with 6 clusters. Ramachandran plots (on the left) and the corresponding average predicted spectrum (on the right).



Annexe : Chapitre VI

Table des matières

Figure S1: Supporting information for the MD simulations.

Molecular dynamic simulations were carried out on the Gromacs software, with the CHARMM27 and OPLS-AA force fields following the global method described in the Chapter IV for the Piv-Pro-D-Ser-NHMe peptide and in the Chapter V for AFA. The important parameters used and the number of coupling constant calculations are summarised below.

Peptide	In line with experimental structure	Force field	Time	Geometries	6-311++G** Calculations	Aug-cc-pVTZ-J Calculations
AFA	Yes	OPLS-AA	1 µs	1000	1000	50
Piv-Pro-D-Ser-NHMe	Yes	CHARMM27	1 µs	1000	1000	50
	No	CHARMM27	50 ps	1000	1000	50
	No	CHARMM27	50 ps	1000	1000	50

Figure S2 : Hyperparameter optimisation grid based on all the tested model and the parameters selected to produce the lowest MAE on the ML3 descriptors.

Random Forest Regression: 'n_estimators': [5, 10, 25], 'max_depth': [None, 10, 20, 30], 'min_samples_split': [2, 5, 10], 'min_samples_leaf': [1, 2, 4]

Best model: "max_depth": 30, "min_samples_leaf": 1, "min_samples_split": 5, "n_estimators": 25

Gradient Boosting Regression: 'n_estimators': [50, 100, 200], 'learning_rate': [0.01, 0.1, 0.2], 'max_depth': [3, 4, 5], 'min_samples_split': [2, 5, 10], 'min_samples_leaf': [1, 2, 4]

Best model: "learning_rate": 0.2, "max_depth": 4, "min_samples_leaf": 2, "min_samples_split": 10, "n_estimators": 200

SVR: 'C': [5, 10,20,40], ‘kernel': ['poly', 'rbf'], 'degree': [2, 3, 4], 'gamma': ['scale', 'auto'], 'epsilon': [0.1, 0.2, 0.3]

Best model: "C": 10, "degree": 2, "epsilon": 0.1, "gamma": "auto", "kernel": "rbf"

MLP Regression: 'max_iter': [3000], 'hidden_layer_sizes': [(20,),(50,), (100,), (200,), (20,20), (50, 50), (100, 100), (200, 200)], 'activation': ['relu', 'tanh', 'logistic'], 'alpha': [0.0001, 0.005, 0.001], 'early_stopping': [True], 'n_iter_no_change': [20]

Best model: activation": "tanh", "alpha": 0.0001, "early_stopping": true, "hidden_layer_sizes": [20, 20], "max_iter": 3000, "n_iter_no_change": 20

Kernel Ridge Regression: 'alpha': [0.001, 0.01, 0.1, 1], 'kernel': ['linear', 'poly', 'rbf'], 'degree': [2, 3, 4], 'gamma': [0.001, 0.01,0.1, 0.5, 1]

Best model: alpha": 0.001, "degree": 2, "gamma": 0.1, "kernel": "poly"

Lasso Regression: 'alpha': [0.01, 0.1, 1, 10]

Best model: "alpha": 0.01

Ridge Regression: 'alpha': [0.01, 0.1, 1, 10]

Best model: "alpha": 0.1

Decision Tree Regression: 'max_depth': [None, 10, 20, 30], 'min_samples_split': [2, 5, 10], 'min_samples_leaf': [1, 2, 4]

Best model: "max_depth": 10, "min_samples_leaf": 4, "min_samples_split": 2

Gaussian Process Regression (GPR) with RBF kernel: 'alpha': [0.0001,0.0005,0.001,0.01], kernel__length_scale': [0.1,0.5,1, 1.5, 2.0,5,10]

Best model: "alpha": 0.01, "kernel__length_scale": 1

Figure S3 : Predicted values against the DFT values (6-311++G** basis set with the FC contribution only) for all model tested with the ML1 descriptors.

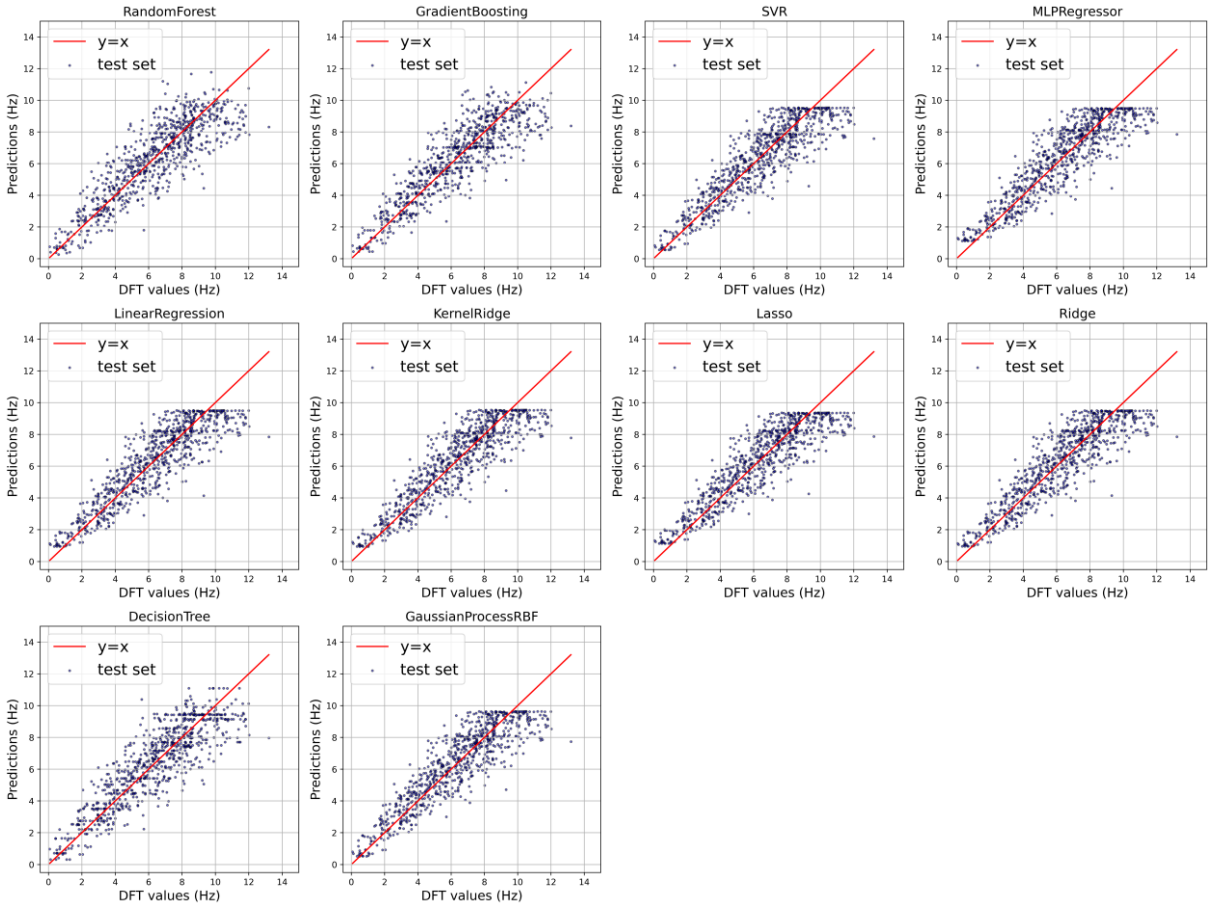


Figure S4 : Predicted values against the DFT values (6-311++G** basis set with the FC contribution only) for all model tested with the ML2 descriptors.

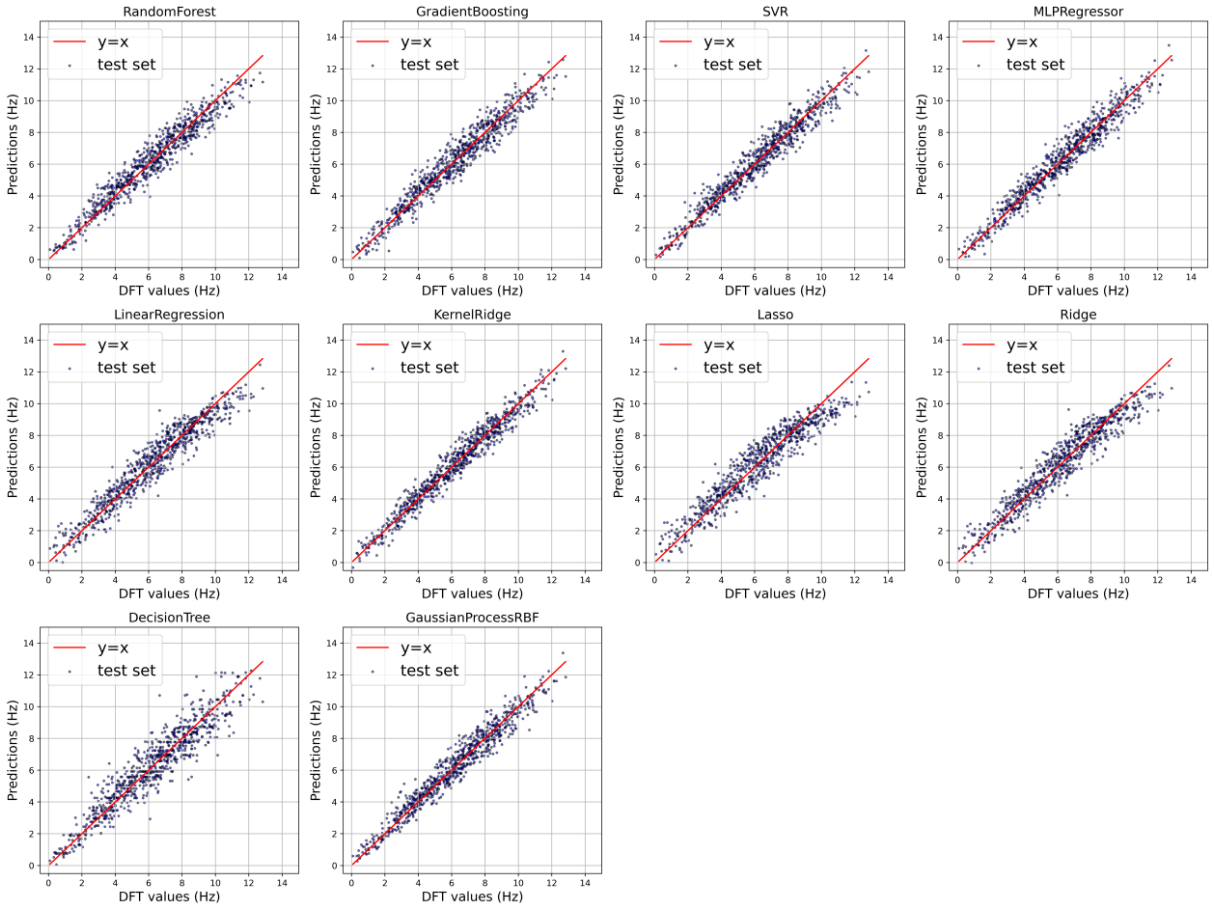


Figure S5 : Predicted values against the DFT values (6-311++G** basis set with the FC contribution only) for all model tested with the ML3 descriptors.

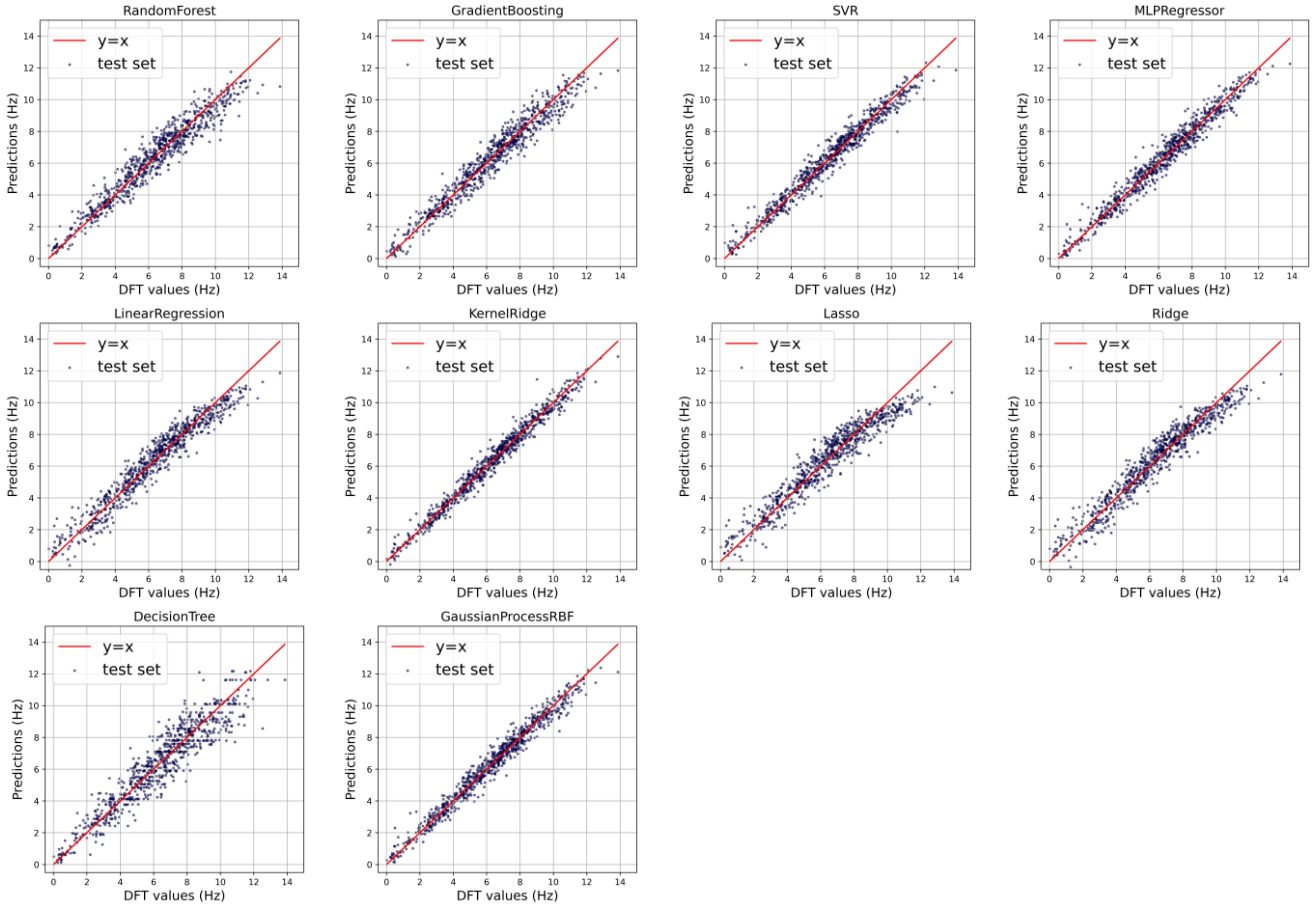
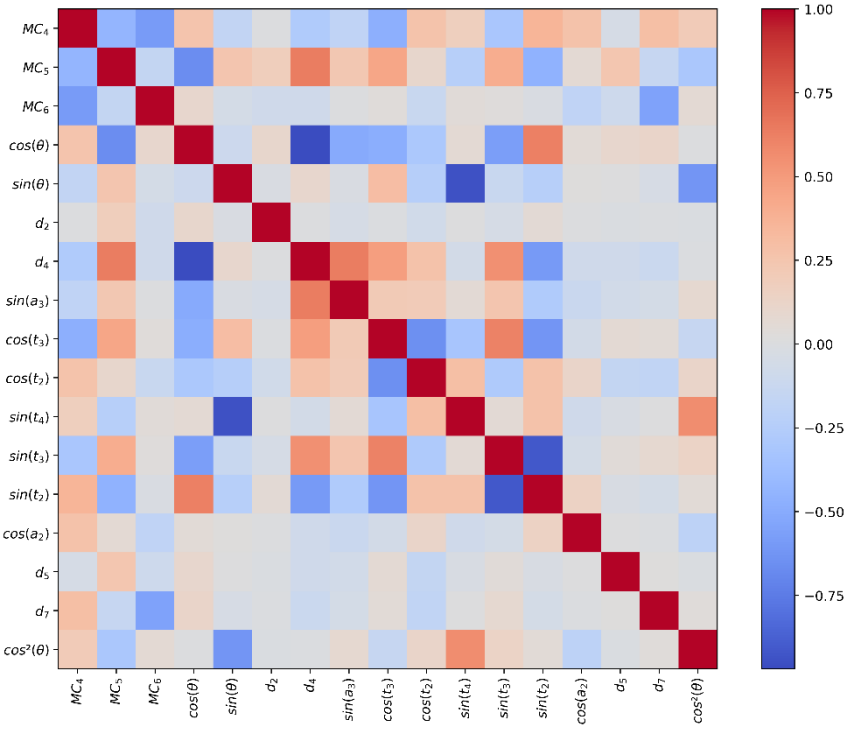


Figure S6 : Correlation matrix of the ML3 set of descriptors after descriptor selection



Approche mixte par résonance magnétique nucléaire, dichroïsme circulaire électronique et modélisation moléculaire pour l'étude structurale de petits peptides bioactifs.

Résumé

La structure secondaire des protéines et des peptides en solution est généralement résolue par la spectroscopie Résonance Magnétique Nucléaire (RMN) et par le Dichroïsme Circulaire Électronique (DCE). Ces deux méthodes d'analyse ont cependant des limitations lorsqu'elles sont utilisées sur des peptides très courts, pouvant contenir des motifs coudés souvent mal caractérisés. En solution, la flexibilité de ces petites molécules complexifie aussi la détermination de structures. Les méthodes théoriques comme la Dynamique moléculaire (DM), la simulation de paramètres RMN par la Théorie de la Fonctionnelle de la densité (DFT) et la production de spectre DCE par TD-DFT sont des outils adaptés pour pallier ces problèmes, bien qu'elles puissent avoir un coût computationnel important. A travers les exemples du peptide Piv-Pro-D-Ser-NHMe et du peptide Ala-Phe-Ala, deux molécules respectivement structurée et flexible, nous avons optimisé des protocoles mixtes, expérimentaux et théoriques, pour la détermination de structure de séquences très courtes d'acides aminés. Enfin, l'efficacité computationnelle a été améliorée par Machine Learning (ML) pour la simulation de la constante de couplage $^{3}JH_{N}$-H_{α} par DFT.

Mots-clés : Dichroïsme Circulaire Électronique, RMN, modélisation, peptide bioactif, machine learning

Abstract

The secondary structure of proteins and peptides in solution is generally resolved by Nuclear Magnetic Resonance (NMR) spectroscopy and Electronic Circular Dichroism (ECD). However, these two analytical methods have limitations when applied on very short peptides, which may contain turn motifs that are still not well characterised. In solution, the flexibility of these small molecules also increases the complexity of structure determination. Theoretical methods, such as Molecular Dynamics (MD), simulation of NMR parameters by Density Functional Theory (DFT) and the computation of DCE spectra by TD-DFT are suitable tools for overcoming these problems, although they can have a significant computational cost. Using the examples of the Piv-Pro-D-Ser-NHMe peptide and the Ala-Phe-Ala peptide, two molecules that are structured and flexible, respectively, we have optimized a mixed experimental and theoretical protocols for structure determination of very short amino acid sequences. Finally, computational efficiency has been improved by Machine Learning (ML) for simulating the $^{3}JH_{N}$-H_{α} coupling constant by DFT.

Keywords : Electronic Circular Dichroism, NMR, modelling, bioactive peptide, machine learning