Analyse de systèmes complexes par dynamique moléculaire polarisable, vers une conception rationnelle de médicaments : des quadruplexes de Guanine au SARS-CoV-2

Dina El Ahdab

- Manuscrit de Thèse -

Sorbonne Université
Université Saint-Joseph de Beyrouth

École doctorale 388
Laboratoire de Chimie Théorique UMR7616
Laboratoire Structure et Interactions des Macromolécules

Analyse de systèmes complexes par dynamique moléculaire polarisable, vers une conception rationnelle de médicaments: des quadruplexes de Guanine au SARS-CoV-2

Thèse
Présentée et soutenue publiquement le 18 Octobre 2021
pour l'obtention des
Doctorat de Sorbonne Université
et
Doctorat de l'Université Saint-Joseph de Beyrouth

par Dina El Ahdab

Composition du jury

Président	Professeur Maher Abboud	Université Saint-Joseph de Beyrouth
Rapporteurs	Professeur Antonio Monari	Université de Paris
	Professeur Mazen Al Ghoul	American University of Beirut
Examinatrice	Docteur Elodie Laine	Sorbonne Université
Co-directeurs de Thèse	Docteur Nohad Gresh	Sorbonne Université
	Docteur Zeina Hobaika Khoury	Université Saint-Joseph de Beyrouth
Directeurs de Thèse	Professeur Jean-Philip Piquemal	Sorbonne Université
	Professeur Richard Maroun	Université Saint-Joseph de Beyrouth

"Je contemple, comme du dehors, cet assemblage prodigieux de molécules, qui, pour quelque temps encore, est moi."

Roger Martin du Gard

Remerciements

Les travaux présentés dans cette Thèse font l'objet d'une convention de cotutelle internationale entre le Laboratoire de Chimie Théorique de Sorbonne Université et le Laboratoire Structure et Interactions des Macromolécules de la faculté des sciences de l'Université Saint-Joseph de Beyrouth.
Je tiens à remercier le doyen de la faculté des sciences de l'Université Saint-Joseph de Beyrouth, Pr Richard Maroun, et le directeur du Laboratoire de Chimie Théorique, Pr Jean-Philip Piquemal, pour m'avoir acceuillie dans leur Laboratoire.
Je souhaite également rendre hommage et exprimer ma profonde gratitude à tous ceux qui, de près ou de loin, ont contribué à sa réalisation et à son aboutissement.

Je remercie Pr Maher Abboud qui m'a honorée de présider le jury de cette Thèse. Je souhaite également adresser mes remerciements et ma gratitude aux Pr Antonio Monari et Pr Mazen Al Ghoul de l'honneur qu'ils m'ont fait en acceptant d'être rapporteurs de cette Thèse et de juger ce travail. J'associe à ces remerciements Dr Elodie Laine qui m'a gratifiée en acceptant d'examiner ce travail.

Je remercie mes directeurs de Thèse qui ont su trouver les mots justes pour m'orienter vers les bons axes de recherche, et ont su trouver les mots d'encouragement. Soyez assuré de mon attachement et de ma profonde gratitude.

Je tiens à remercier le Professeur Jean-Philip Piquemal, directeur du Laboratoire de Chimie Théorique à Sorbonne Université, de m'avoir accueilli dans son équipe et d'avoir accepté de diriger ce travail et d'avoir partagé ses brillants intuitions. Qu'il soit aussi remercié pour sa gentillesse, sa disponibilité permanente et pour les nombreux encouragements qu'il m'a prodigué.

Je remercie également le Professeur Richard Maroun, doyen de la Faculté des Sciences de l'Université Saint-Joseph de Beyrouth qui m'a aussi encadré pendant cette Thèse. Je lui suis reconnaissante de m'avoir fait bénéficier tout au long de ce travail de sa grande compétence, de sa rigueur intellectuelle et de son dynamisme. Je suis également sensible à la confiance qu'il m'a accordée.

J'adresse de chaleureux remerciements à mes co-encadrants de Thèse ;
Je remercie Dr Zeina Hobaika Khoury, pour son attention et son soutien. J'ai été extrêmement sensible à ses qualités humaines d'écoute et de compréhension tout au long de ce travail doctoral.
Je remercie également Dr Nohad Gresh, pour ses connaissances, ses critiques, ses conseils constructifs et les heures qu'il a consacrées et qui m'ont aider à bien mener ce travail. Je le remercie pour son accueil chaleureux à chaque fois que j'ai sollicité son aide.

Je tiens à remercier tous les membres du Laboratoire de Chimie Théorique UMR 7616, qui ont répondu avec calme et patience aux questions quotidiennes dont je les accablais, notamment Théo Jaffrelot Inizan, Frédéric Célerse, Louis Lagardère, Luc-Henry Jolly, Olivier Adjoua, Salem Tacine, Johanna Klein et Perla El Darazi. Je souhaite également bonne continuation à Nastasia Mauger et Johanna Klein.

Je remercie mes ami(e)s avec qui j'ai partagé mes études et notamment ces années de Thèse ;
Merci à Mona Barake et Rim Khayat pour votre amitié, pour toutes ces journées, soirées, week-end passés ensemble et à toutes vos petites attentions qui me vont droit au cœur. Je joins à ces remerciements Nadine Wanny et Carine Ayoubi.
Merci à Valérie Zgheib, avec qui j'ai pu découvrir un monde littéraire et développer l'esprit journalistique au sein de CAMPUS-J.
Merci à Rani Abbas, Hadi Hassan, Ahmad Abdelkader, Rami Massri, Paul Al Malak, Rami Jradi, Carel Antoun, Monah Nasr, Reina Dannaoui, Amar Naji, Sana Abdallah, Anhal Saab, Christian Beaini, Fernando Yahchouchi, David Al Moukachar et Elio Nassar pour tous les fous moments et les belles soirées.

Je désire grandement remercier Mme Fadia Alam Gemayel, directrice du Campus du Liban Nord de l'Université Saint-Joseph de Beyrouth, pour ses attentions discrètes et de m'avoir toujours supporté dans tout ce que j'ai entrepris.

Enfin, c'est avec une joie immense et un cœur ému que je dédie cette Thèse à ma chère famille pour ses affections inépuisables et ses précieux conseils. Un très grand merci à ma mère, mon père et mon frère, Jad, qui m'ont gratifié de leur amour et fourni les motivations qui ont permis l'aboutissement de mes entreprises. Je leur adresse, du fond du cœur, toute ma gratitude.

Table de matières

1 Introduction générale **18**
1.1 Introduction à la dynamique moléculaire (MD) . . . 19
1.1.1 Histoire de la dynamique moléculaire . . . 19
1.1.2 champs de force . . . 20
1.1.3 Minimisation d'énergie . . . 26
1.1.4 Algorithmes d'intégration en dynamique moléculaire . . . 27
1.1.5 Ensembles thermodynamiques . . . 28
1.1.6 Conditions périodiques . . . 28
1.1.7 Particle Mesh Ewald (PME) . . . 28
1.1.8 Limitations de la dynamique moléculaire . . . 29
1.1.9 Logiciel - Tinker . . . 29
1.2 Protocole générale de la dynamique moléculaire . . . 30
1.2.1 Structures initiales et construction des systèmes . . . 30
1.2.2 Minimisation . . . 30
1.2.3 Chauffage . . . 31
1.2.4 Production . . . 31
1.2.5 Stratégie d'échantillonnage adaptative non supervisé . . . 31
1.2.6 Débiaisage . . . 33
1.3 Outils d'analyse des trajectoires issues de la dynamique moléculaire . . . 34
1.3.1 Clusterisation des structures issues de la MD adaptative . . . 34
1.4 La Protéase Principale, M^{pro}, du SARS-CoV-2 . . . 35
1.4.1 Biologie des coronavirus . . . 41
1.4.2 Cycle viral du SARS-CoV-2 . . . 42
1.4.3 Importance de la protéase principale, M^{pro}, du SARS-CoV-2 . . . 43
1.5 Les quadruplexes de guanine (GQ) au niveau des acides nucléiques . . . 45
1.5.1 Les GQ au niveau du promoteur de l'oncogène c-kit 1 . . . 48
1.5.2 Les GQ au niveau de la région LTR-III du VIH-1 . . . 51
1.6 Objectifs de notre étude . . . 56

2 Modélisation à haute résolution de la dynamique structurelle de la Protéase Principale M^{pro} du SARS-CoV-2 59
2.1 Détails computationnels et protocole de simulation 60
2.2 Résultats et discussion 62
2.2.1 Performance de l'exploration par échantillonnage adaptatif : comparaisons avec d'autres simulations disponibles 62
2.2.2 Regroupement et extraction non supervisés de l'énergie libre relative débiaisée entre les domaines représentatifs 65
2.3 Corrélation avec les données expérimentales : marqueurs structurels de l'activité des protomères et nouvelles fonctionnalités 69
2.3.1 Marqueurs de la structuration du trou oxyanionique 69
2.3.2 Évaluation des volumes des cavités enzymatiques 77
2.3.3 Analyse des fluctuations locales : grande flexibilité de la région C-terminale . . 80
2.3.4 Analyse comparative de la "druggabilité" : recherche de poches cryptiques . . 82
2.3.5 Analyse de solvatation : l'importance des effets de polarisation explicites dans l'eau 85
2.4 Étude de la région de dimérisation 90
2.4.1 Stabilité de l'interface de dimérisation : études sur les liaisons hydrogènes . . . 90
2.4.2 Interactions allostériques entre l'interface de dimérisation avec le site catalytique 94
2.4.3 Importance des modèles de solvatation et des effets de polarisation dans les interactions allostériques entre les sites 97
2.5 Conclusions 107
2.6 Publications 112
2.6.1 Publication 1 112
2.6.2 Publication 2 132
2.6.3 Publication 2 - "Back Cover" 142

3 Étude par dynamique moléculaire des structures des quadruplexes de guanine au niveau du promoteur de l'oncogène c-kit et du LTR-III du VIH-1 144
3.1 Détails computationnels et protocole de simulation 145
3.1.1 Choix des structures de départ et construction des systèmes 145
3.1.2 Dynamique moléculaire conventionnelle 146
3.1.3 Algorithme adaptatif: longues dynamiques moléculaires polarisables 147
3.1.4 Analyse de composantes indépendantes du temps et de la structure (tICA) . . 147
3.2 Résultats et Discussion 148
3.2.1 Évaluation de la stabilité conformationnelle et étude des états de transition . . 148
3.2.2 Clusterisation et analyse conformationnelle des clusters 156
3.2.3 Stabilisation du GQ de c-kit 1 par les ions K^+ et les molécules d'eau polarisables 169

3.2.4 Poches cryptiques au niveau des GQ de c-kit 1 et du VIH-1 LTR-III 170
3.3 Conclusions 175

4 Conclusions générales et futures perspectives 178
4.1 Rappel de notre problématique 179
4.2 Spécificité et résultats de notre approche 179
4.3 Perspectives sur la modélisation des GQ 183
4.4 Conception des ligands à visée thérapeutique 183
4.5 Perspectives techniques : raffinements des champs de force polarisables sur la base des calculs SAPT-DFT 186
4.5.1 Le champ de force polarisable: SIBFA (Sum of Interactions Between Fragments *Ab initio* computed) 187
4.5.2 Construction d'une base de données pour les interactions intermoléculaires sur la base de calculs perturbationnels SAPT(DFT) 189
4.5.3 Raffinements supplémentaires du potentiel polarisable SIBFA sur la base des calculs SAPT(DFT) 191
4.5.4 Validation sur des oligomères 192

5 Références bibliographiques 200

6 Annexes 224

Liste des abréviations et symboles

Å: Angstrom
ACE : Enzyme de conversion de l'angiotensine
ADN : Acide DésoxyriboNucléique
aMD : Dynamique Moléculaire accélérée
AML : Leucémie Aigüe Myéloïde
ARN : Acide RiboNucléique
ARNg : Acide RiboNucléique génomique
ARNsg : Acide RiboNucléique sous-génomique
ARNm : Acide RiboNucléique messager
AMBER : Assisted Model Building with Energy Refinement
AMOEBA : Atomic Multipole Optimized Energetics for Biomolecular Applications
CHARMM : Chemistry at Harvard Macromolecular Mechanics
CD : Cluster of Differentiation
CG : Conjugated gradient
Cl^- : Ion chlorure
cMD : classical Molecular Dynamics
CPU : Central Processing Unit
CCSDT : Coupled Cluster Single-double and triple
D : Debaye
DBSCAN: Density-based spatial clustering of applications with noise
DFT : Density Functional Theory
MD : Molecular Dynamics
ELF : Electron Localization Function
FF : Force field
fs : femtoseconde
GaMD : Gaussian accelerated Molecular Dynamics
GDMA : Gaussian Distributed Multipole Analysis
GIST : Tumeur Gastro-Intestinale
GQ : Quadruplexes de guanines
GPU : Graphical Processing Unit
GROMACS : GROningen MAchine for Chemical Simulations
HAART : Highly Active Antiretroviral Therpay
HIPPO : Hydrogen-like Interatomic Polarizable POtential
HTMD : Dynamique moléculaire à haut débit
IFN : Intérferons

ISA : space-Iterated Stockholder Atoms
K : Kelvin
K^+ : Ion potassium
LF : Lethal Factor
LP : Longue boucle d'hélice
LTR : séquence Terminale Longue Répétée
MM : Molecular Mechanics
M^{pro} : Protéase principale du coronavirus du syndrome respiratoire aigu sévère
MPI : Message Passing Interface
MSM : Markov State Model
NC : NucléoCapside
n–PFF : non-Polarisable Force Field
ns : nanoseconde
OPLSAA : Optimized Potentials for Liquid Simulations
ORF : Open Reading Frame
PBC : Periodic Boundary Conditions
PCA : Analyse des composants principaux, de l'acronyme anglo-saxon "Principal Component Analysis"
PDB : Protein Data Bank
PFF : Polarisable Force Field
pH : potentiel d'Hydrogène
PLpro : Protéase de type papaïne
PME : Particle Mesh Ewald
ps : picoseconde
QC : Chimie Quantique
RESPA : Reversible Reference System Propagator Algorithms
RMN : Résonance Magnétique Nucléaire
RMSD : Déviation de la Racine de la Moyenne des Carrés, de l'acronyme anglo-saxon "Root-Mean-Square Deviation"
RMSF : Fluctuation de la Racine de la Moyenne des Carrés, de l'acronyme anglo-saxon "Root-Mean-Square Fluctuation"
RT : RétroTranscriptase
SAPT : Symmetry Adapted Perturbation Theory
SARS-CoV-1 : Coronavirus 1 du syndrome respiratoire aigu sévère
SARS-CoV-2 : Coronavirus 2 du syndrome respiratoire aigu sévère
SCF : Facteur de croissance des cellules souches, de l'acronyme ango-saxon "Stem Cell Factor"
SD : Steepest Descent

SIBFA : Sum of Interactions Between Fragments *Ab initio* computed

SIDA : Syndrome d'immunodéficience acquise

tICA : Analyse de composantes indépendantes du temps et de la structure, de l'acronyme anglo-saxon "Time-Lagged Independent Component Analysis"

5'-UTR : Région 5' non traduite

vdW : van der Waals

VIH-1 : Virus de l'Immunodéficience Humaine de type 1

W-C : Watson-Crick

Symboles des acides aminés

A Ala alanine
C Cys cystéine
D Asp acide aspartique
E Glu acide glutamique
F Phe phénylalanine
G Gly glycine
H His histidine
I Ile isoleucine
K Lys lysine
L Leu leucine
M Met méthionine
N Asn asparagine
P Pro porline
Q Gln glutamine
R Arg arginine
S Ser serine
T Thr thréonine
V Val valine
W Trp tryptophane
Y Tyr tyrosine

Symboles des bases nucléiques

A : Adénine
C : Cytosine
G : Guanine
T : Thymine

Résumé

La conception de nouvelles molécules thérapeutiques *in silico* nécessite une compréhension fine de la dynamique structurale des protéines et des acide nucléiques cibles. Les objectifs de cette Thèse sont d'appliquer les techniques de dynamique moléculaire couplées aux champs de force polarisables à des systèmes biologiques complexes d'importance primordiale dans le cadre d'une conception rationnelle de médicaments anti-cancéreux et anti-viraux (mise au point d'inhibiteurs). Nous avons choisi d'étudier les quadruplexes de guanine (GQ) au niveau du proto-oncogène c-kit 1 et de la région LTR-III du VIH-1 ainsi que la protéase principale (M^{pro}) du SARS-CoV-2. En effet, ces complexes moléculaires cibles présentent un fort intérêt biologique et médical du fait de leur implication dans des maladies graves : le cancer, le SIDA et le Covid-19. Les structures GQ sont considérées comme des cibles thérapeutiques émergentes vue que la répression transcriptionnelle des gènes par la stabilisation de ces structures pourrait être une nouvelle stratégie anti-cancéreuse et anti-virale. Par ailleurs, la M^{pro}, joue un rôle essentiel dans le cycle de vie et la réplication du SARS-CoV-2. Alors que cette protéase n'a pas d'homologue humain, il s'agit d'une cible anti-virale intéressante.

Grâce aux avancés technologiques développées au sein de notre Laboratoire, tant au niveau informatique qu'au niveau des techniques d'échantillonnage accéléré en dynamique moléculaire basée sur la PCA, nous avions la possibilité d'effectuer plusieurs μs de simulations en dynamique moléculaire à haute-résolution sur des systèmes de grande taille en solvant explicite et avec un champ de force polarisable permettant d'obtenir une haute précision dans la prédiction des conformations possibles des systèmes étudiés. Ceci nous a permis d'étudier finement la dynamique moléculaire de la protéase principale, M^{pro} du SARS-CoV-2 et des quadruplexes de guanine trouvés dans le promoteur de c-kit 1 et au niveau de la région LTR-III du VIH-1.

Les marqueurs structurels clés liés à la stabilité des trous oxyanioniques de la M^{pro} révèlent une asymétrie entre les protomères. Les résultats mettent en évidence la plasticité du site actif et la présence des poches cryptiques. L'analyse des schémas de solvatation au niveau du site actif et de la région de dimérisation démontre que les molécules d'eau polarisables AMOEBA confinées sont capables d'explorer une large gamme de moments dipolaires, conduisant à un nombre de molécules d'eau cohérent avec les données expérimentales. Les données montrent l'impact de la protonation sur la déstructuration de la boucle oxyanionique et de l'interface de dimérisation.

Sept clusters stables ont été identifiés au cours de MD du GQ de c-kit 1 par l'analyse tICA, où la stabilité des trois tétrades de guanine empilées et de la longue boucle d'hélice ont été mises en évidence. Cependant la flexibilité des deux boucles mononucléotidiques et d'une boucle dinucléotidique a été analysée en surveillant les distances inter-bases pertinentes, deux angles de torsion clés, χ et γ, et les indices d'empilement π-π. Deux poches cryptiques stables ont été identifiées dans les sept clusters, dont les sillons pourraient servir de sites de liaison pour les ligands pharmacologiques. Une simula-

tion cMD préliminaire sur la structure GQ au niveau de la LTR-III du VIH-1 a ensuite été effectuée. Deux poches cryptiques ont également été identifiées à son issue, mais leurs conformations diffèrent significativement de celles des poches c-kit 1, et seule la première s'est avérée possiblement viable pour la conception de ligands.

Les travaux effectués dans cette Thèse suggèrent que l'utilisation de PFF pourrait être critique dans la découverte de médicaments pour modéliser avec précision la complexité des interactions moléculaires. Les résultats offrent aux concepteurs de ligands thérapeutiques des modèles d'espaces conformationnels à très haute précision allant des quadruplexes de guanine jusqu'aux protéines du SARS-CoV-2.

Mots-clé : Quadruplexe de guanine, c-kit, VIH, SARS-CoV-2, Protéase principale M^{pro}, dynamique moléculaire.

Abstract

In silico design of new therapeutic molecules requires a detailed understanding of the structural dynamics of target proteins and nucleic acids. The objectives of this thesis are to apply molecular dynamics techniques coupled to polarizable force fields to complex biological systems of paramount importance within the framework of a rational design of anti-cancer and anti-viral drugs (inhibitors development). We chose to study guanine quadruplexes (GQ) present at the level of c-kit 1 proto-oncogene and HIV-1 LTR-III as well as the main protease (M^{pro}) of SARS-CoV-2. Indeed, these target molecular complexes are of great biological and medical interest due to their involvement in serious diseases: cancer, AIDS and Covid-19. GQ structures are considered as emerging therapeutic targets as the transcriptional repression of genes by stabilizing these structures could be a new anti-cancer and anti-viral strategy. In addition, M^{pro} plays an essential role in the life cycle and replication of SARS-CoV-2. While this protease does not have a human homolog, it is an interesting anti-viral target.

Thanks to the technological advances developed within our Laboratory, both at the computational level and at the level of accelerated sampling techniques in molecular dynamics based on PCA, we had the possibility to perform several μs of simulations in High-resolution molecular dynamics on large systems in explicit solvent and with a polarizable force field allowing to obtain a high precision in the prediction of the possible conformations of the studied systems. This allowed us to study the molecular dynamics of the main protease, M^{pro} of SARS-CoV-2 and of the guanine quadruplexes found in c-kit 1 promoter and HIV-1 LTR-III.

The key structural markers related to the stability of the oxyanion holes of the M^{pro} reveal asymmetry between the protomers. The results highlight the plasticity of the active site and the presence of cryptic pockets. Analysis of solvation patterns at the active site and dimerization region demonstrates that confined AMOEBA polarizable water molecules are able to explore a wide range of dipole moments, leading to a consistent water molecule number with respect to experimental data. The data show the impact of protonation on the destructuring of the oxyanionic loop and the dimerization interface.

Seven stable clusters were identified during MD of c-kit 1 GQ by tICA analysis, where the stability of the three stacked guanine tetrads and the long propeller loop were demonstrated. However the flexibility of the two single-nucleotide loops and one dinucleotide loop was analyzed by monitoring relevant inter-base distances, two key torsion angles, χ and γ, and π-π stacking index. Two stable cryptic pockets were identified in the seven clusters, the grooves of which could serve as binding sites for pharmacological ligands. A preliminary cMD simulation on the GQ structure at the level of HIV-1 LTR-III was then performed. Two cryptic pockets have also been identified at its outcome, but their conformations differ significantly from those of c-kit 1 pockets, and only the former has been shown to be possibly viable for ligand design.

The work done in this thesis suggests that the use of PFF could be critical in drug discovery to

accurately model the complexity of molecular interactions. The results provide designers of therapeutic ligands with very high precision conformational space models ranging from guanine quadruplexes to SARS-CoV-2 proteins.

Keywords: Guanine quadruplex, c-kit, HIV, SARS-CoV-2, M^{pro} main protease, molecular dynamics.

Chapitre 1

Introduction générale

1.1 Introduction à la dynamique moléculaire (MD)

La machinerie cellulaire repose en grande partie sur les protéines (et les acides nucléiques) qui peuvent interagir avec d'autres protéines, une petite molécule ou encore un acide nucléique à travers un mécanisme d'action moléculaire lors d'un processus biologique donné. Une bonne connaissance du repliement d'une molécule, c'est à dire l'agencement dans l'espace tridimensionnel des acides aminés ou des acides nucléiques formant sa séquence, est essentiel pour mieux comprendre son rôle, ses interactions possibles avec d'autres molécules et à un niveau plus global le fonctionnement d'un mécanisme biologique. Dans le cadre de la chasse à la "druggabilité" (i.e. la capacité d'un site moléculaire à se lier avec un médicament), la compréhension du repliement peut aussi être utile pour des applications à but thérapeutique direct comme la compréhension des effets d'une mutation de la séquence des acides aminés ou des acides nucléiques ou la recherche de nouvelles molécules à visée thérapeutique ("drug design"). Il convient dans ce contexte, d'avoir accès au plus grand nombre de conformations moléculaires stables possibles afin de pouvoir cibler la protéine ou l'acide nucléique en question. La dynamique moléculaire est un technique très puissante qui permet de simuler l'évolution de la conformation et le repliement d'une molécule au cours du temps.

Au sens large, par définition, la Dynamique Moléculaire (MD) est la science de la simulation d'un système contenant des particules. Elle peut être appliquée sur des systèmes allant de l'atome jusqu'à une galaxie. La MD consiste à étudier les conformations structurelles d'une molécule, qui est considérée comme un assemblage d'atomes dont les coordonnées spatiales varient au cours du temps, en lui appliquant les lois de la mécanique classique newtonienne. Ainsi, chaque particule est considérée comme une masse ponctuelle dont le mouvement est déterminé par l'ensemble des forces qui s'exercent sur elle. Il en résulte un mouvement atomique correspondant à des vibrations autour d'un minimum d'énergie, qui correspond à une conformation stable de la molécule, ou au passage d'un de ces minima à un autre. La dynamique moléculaire possède ainsi la faculté d'extraire la molécule d'un minimum local permettant de découvrir un large éventail de conformations capables d'être adoptées par la molécule. Ainsi, le résultat d'une simulation de MD consiste en la génération d'une série de positions atomiques, c'est à dire des coordonnées spatiales de tous les atomes du système qui évoluent en fonction du temps (trajectoire).

1.1.1 Histoire de la dynamique moléculaire

La simulation par dynamique moléculaire (MD), développée pour la première fois à la fin des années 70 [92, 223] est passée de la simulation de plusieurs centaines d'atomes à des systèmes ayant une pertinence biologique, y compris des protéines entières en solution avec des représentations de solvant explicites, des protéines membranaires ou de grands complexes macromoléculaires comme des nucléosomes [50] ou des ribosomes [103]. Récemment, des simulations MD ont été utilisées pour modéliser les comportement physico-chimiques des molécules étudiées.

La résolution numérique des équations décrivant le comportement atomique nécessite l'utilisation d'un pas de temps extrêmement petit, de l'ordre de la fs (10^{-15} s) de manière à prendre en considération les vibrations atomiques les plus rapides. Ainsi, jusqu'à très récemment, le temps maximum de simulation dépassait rarement les 10 à 100 ns si l'on voulait utiliser les champs de force les plus modernes (i.e. polarisables) et ceci malgré la puissance des ordinateurs actuels. En effet, pour simuler des systèmes réalistes, il convient de traiter des ensembles moléculaires complexes de plusieurs dizaines de milliers d'atomes et de simuler aussi de manière explicite la présence d'un solvant et/ou d'une membrane.

Pour pallier au surcoût computationnel lié à l'utilisation des champs de force polarisables, nous avons développé au sein de notre Laboratoire une nouvelle plateforme d'échantillonage adaptatif qui nous permet d'atteindre des simulations de l'ordre des μs en dynamique moléculaire en multipliant des répliques explorant l'espace des phases en parallèle [93].

1.1.2 champs de force

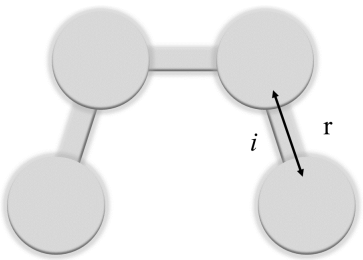
Lors des simulations de dynamique moléculaire, il convient de bien définir les paramètres du système, grâce à un champ de force optimisé pour reproduire le plus fidèlement possible la réalité. Les champs de force sont apparus dans les années 1970 et sont toujours en cours d'évolution, les paramètres s'affinant afin d'obtenir les simulations numériques les plus réalistes possibles. En effet, un champ de force est un ensemble d'équations paramétriques capables de modéliser l'énergie potentielle d'un ensemble d'atomes. Il décrit les interactions entre atomes liés (énergies de liaison, d'angle de valence et de torsion des angles dièdres) et non-liés (interactions de type van der Waals, électrostatique et liaisons hydrogène). Ainsi, l'ensemble de paramètres comprend des valeurs pour la masse atomique, le rayon de van der Waals et la charge partielle d'atomes individuels, et des valeurs d'équilibre de longueurs de liaison, de mesures d'angles plans et dièdres pour des paires, triplets et quadruplets d'atomes liées, et les valeurs de constante de force pour chaque potentiel. Les champs de force les plus courants utilisent un modèle de charge fixe dans lequel chaque atome se voit affecter une charge électrostatique qui n'est pas affectée par l'environnement local (i.e. champs de force non polarisable). Les développements proposés pour les champs de force de génération plus récentes incluent la polarisabilité permettant de mimer le fait que la densité de charge des particules sont influencées par celles de leurs voisines. L'effet de la polarisabilité peut, par exemple, être approximée par l'introduction de modèles dit à dipôles induits ; elle peut aussi être représentée par des particules de Drude. L'utilisation pratique de la polarisabilité dans les champs de force pour un usage courant fut historiquement ralentie par le coût important associé au calcul du champ électrique local.

Tous les champs de force classiques non polarisable partage une forme fonctionelle commune qui est la somme de ces différents termes :

$$E_{\mathrm{F}} = E_1 + E_\theta + E_\tau + E_{\mathrm{vdw}} + E_{\mathrm{él}} + E_{\mathrm{HB}} \tag{1.1}$$

Les termes de cette équation correspondent aux :

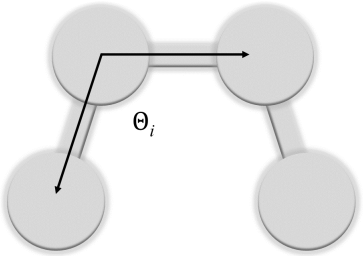
- Énergie d'élongation des liaisons ("Stretching"), E_1



$$E_1 = \frac{1}{2}\sum_{i=1}^{n} \mathrm{k}_{r,i}(r_i - r_i^0)^2 \tag{1.2}$$

où
r_i = longueur instantanée de la liaison i
$\mathrm{r}_i{}^0$ = longueur de la liaison i à l'équilibre
$\mathrm{k}_{r,\ i}$ = constante de force de la liaison i
n = nombre de liaisons dans la molécule

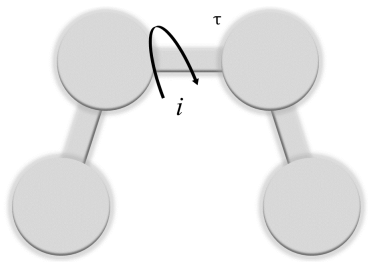
- Énergie de déformation des angles ("Bending"), E_θ



$$E_\theta = \frac{1}{2}\sum_{i}^{n} \mathrm{k}_{\theta,i}(\theta_i - \theta_i^0)^2 \tag{1.3}$$

où
θ_i = angle de liaison entre 3 atomes
$\theta_\mathrm{i}{}^0$ = angle de liaison entre 3 atomes à l'équilibre
$\mathrm{k}_{\theta,\ \mathrm{i}}$ = constante de force de flexion
n = nombre d'angles dans la molécule

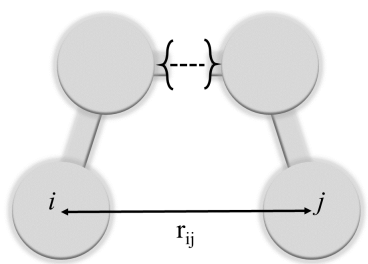
- Énergie de torsion des angles dièdres ("Torsion"), E_τ



$$E_\tau = \frac{1}{2}\sum_{i=1}^{n} \mathrm{A}_{i,m}[1 + cos(m\tau_i - \phi)] \tag{1.4}$$

où
τ_i = angle de torsion
ϕ = phase
$\mathrm{A}_{i,\ \mathrm{n}}$ = constante de force de courbure
m = périodicité de $\mathrm{A}_{i,\ \mathrm{m}}$
n = nombre d'angles dièdres

- Interactions de van der Waals, $\mathrm{E}_{\mathrm{vdw}}$

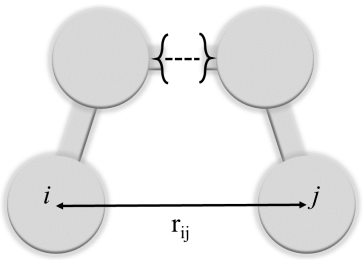


$$E_{vdw} = \sum_{i,j=1}^{n} [\mathrm{A}_{i,j}(\frac{r_{ij}^0}{rij})^{12} - \mathrm{B}_{i,j}(\frac{rij^0}{rij})^6] \tag{1.5}$$

où
A_{ij}, B_{ij} = constantes de van der Waals
i_{ij} = distance entre 2 atomes i et j non-liés
$\mathrm{r}_{ij}{}^0$ = somme des rayons de van der Waals de i et j
n = nombre de paires d'atomes non-liés

- Interactions électrostatiques, $\mathrm{E}_{\text{él}}$

$$E_{él} = \sum_{ij}^{n} (\frac{q_i q_j}{4\pi\epsilon_{ij} r_{ij}}) \tag{1.6}$$



où
q = charge partielle de l'atome
$\frac{1}{4\pi\epsilon_{ij}}$ = permittivité électrique du milieu
r_{ij} = distance entre les 2 atomes i et j
n = nombre de paires des atomes non-liés

Nous distinguons les champs de force classiques (ou non-polarisable, nPFF) comme AMBER, CHARMM, GROMACS des champs de force polarisable basés sur des multipôles distribués (PFF) comme SIBFA (Sum of Interactions Between Fragments Ab initio computed) et AMOEBA (Atomic Multipole Optimized Energetics for Biomolecular Applications). Particulièrement, le champ de force polarisable AMOEBA, qui fut développé par Ponder *et coll.* [194, 237] nous a largement servi pour nos études. Il est capable d'atteindre une précision chimique dans les propriétés thermodynamiques de protéines et des acide nucléiques [235] en solution, en s'appuyant sur une évaluation des énergies conformationnelles et d'interaction comparable aux modèles de mécanique quantique. Nous l'avons choisi dans nos études pour ses capacités à reproduire diverses propriétés physico-chimiques en phase aqueuse. Sa totale flexibilité intramoléculaire, alliée à son traitement fin de l'électrostatique (utilisation de multipôles distribués), est un atout pour la prédiction de spectres vibrationnels moléculaires. AMOEBA a été comparé à d'autres champs de force (OPLSAA, AMBER, CHARMM) par Jensen *et coll.* [178, 105] pour reproduire les énergies conformationnelles des acides aminés calculées en chimie quantique. Il a été démontré que les champs de force qui utilisent uniquement des charges ponctuelles pour le calcul de l'énergie électrostatique ne peuvent reproduire que la moitié des conformations étudiées, alors qu'AMOEBA atteint un taux de 80 % grâce à la distribution multipolaire et la prise en compte des effets de polarisation.

Le champ de force AMOEBA

Le champ de force AMOEBA a la forme fonctionnelle générale suivante [172] :

$$U = U_{bond} + U_{angle} + U_{b\theta} + U_{oop} + U_{torsion} + U_{vdW} + U_{ele}^{perm} + U_{ele}^{ind} \tag{1.7}$$

où les cinq premiers termes décrivent les interactions de valence à courte portée (étirement des liaisons ("bond"), flexion d'angle ("angle"), terme croisé d'angle de liaison ('bθ'), flexion hors du plan ("oop") et rotation de torsion ("torsion"), et les trois derniers termes sont les contributions vdW et électrostatiques non liées. Outre l'inclusion explicite de la polarisabilité, la principale différence entre AMOEBA et un n–PFF est le remplacement du modèle de charge partielle fixe par des multipôles atomiques allant jusqu'aux quadrupôles. L'utilisation de dipôles et de quadrupôles permanents permettent une reproduction plus précise des potentiels électrostatiques moléculaires et un réglage fin des effets directionnels dans le cadre de la liaison hydrogène et d'autres interactions (stacking etc...). L'inclusion de polarisation dipolaire explicite permet au modèle AMOEBA de répondre aux changements de microenvironnements et de pouvoir rester transférables dans des environnements moléculaires hétérogènes. De part son design, AMOEBA a été conçu à travers un paramétrage direct utilisant les données expérimentales en phase condensées et les résultats des calculs de chimie quantique de haut niveau réalisés en phase gaz. AMOEBA présente également un traitement cohérent de la polarisation intra- et intermoléculaire qui est obtenue grâce à un schéma d'amortissement motivé physiquement pour les effets de polarisation locaux. Un autre aspect attrayant de la stratégie de paramétrisation d'AMOEBA est son utilisation des moments multipolaires dérivés explicitement des densités d'électroniques issue d'un calcul *ab initio* réalisés sur des petites molécules ou fragments moléculaires. L'objectif de la conception d'AMOEBA était d'atteindre une "précision chimique" de 0.5 kcal/mol ou mieux pour les interactions entre les petites molécules et les protéines-ligands. Nous décrivons ci-dessous plus en détails la forme fonctionnelle du champ de force AMOEBA (équation 1.7) :

- Interactions de valence à courte distance :
  AMOEBA comprend une flexibilité intramoléculaire complète. Pour les atomes directement liés et séparés par deux liaisons, l'énergie covalente est représentée par des fonctions empiriques de longueurs et d'angles de liaison. Les formes fonctionnelles pour l'étirement des liaisons (Eq. 1.8), angle la flexion (Eq. 1.9), et le couplage entre l'étirement et la flexion (Eq. 1.10), sont ceux du champ de force MM3 [157] et incluent une prise en compte de l'anharmonicité grâce à l'utilisation d'écarts d'ordre supérieur par rapport aux longueurs de liaison idéales (b_0) et les angles (θ_0):

$$U_{bond} = K_b(b - b_0)^2[1 - 2.55(b - b_0) + (7/12)2.55(b - b_0)^2] \tag{1.8}$$

$$U_{angle} = K_\theta(\theta-\theta_0)^2[1-0.014(\theta-\theta_0)+5.6\times10^{-5}(\theta-\theta_0)^2-7.0\times10^{-7}(\theta-\theta_0)^3+2.2\times10^{-8}(\theta-\theta_0)^4] \tag{1.9}$$

$$U_{b\theta} = K_{b\theta}[(b - b_0) + (b^{'} - b_0^{'})](\theta - \theta_0) \tag{1.10}$$

$$U_{oop} = K_\chi \chi^2 \tag{1.11}$$

où la longueur de liaison, b ou b', et l'angle de liaison, θ, et les énergies sont en unités de Ångström (Å), degrés (°), et kcal/mol, avec les constantes de force, K données en unités correspondantes. Il est important de noter qu'un champ de force anharmonique est défini comme une expansion d'ordre supérieur de la surface d'énergie potentielle moléculaire autour d'une géométrie de référence, généralement choisie pour être une structure d'équilibre.

De plus, une fonction Wilson-Decius-Cross est utilisée à sp^2-centres trigonaux hybridés pour restreindre la flexion hors du plan (Eq. 1.11), où pour les centres liés séquentiellement i, j, k et l, χ fait référence à l'angle entre le vecteur jl et le plan ijk.

Une expansion de Fourier traditionnelle (une forme trigonométrique de 1 à 6 fois) fonctionnelle de torsion (Eq. 1.12) :

$$U_{torsion} = \sum_n K_{n\phi}[1 + cos(n\phi \pm \delta)] \tag{1.12}$$

est utilisée pour aider à fusionner les termes de "valence" à courte portée avec les interactions "non-liées" à longue portée. Les paramètres de torsion sont affinés après la détermination des paramètres non liés dans l'espoir que le modèle électrostatique intramoléculaire AMOEBA amélioré conduira à un meilleur équilibre physique entre les interactions locales (vdW + électrostatique + torsion) et à longue distance (vdW + électrostatique) dans l'énergie conformationnelle.

- Interactions de van der Waals
  L'interaction de van der Waals (vdW) additive par paires dans AMOEBA adopte la forme fonctionnelle tamponnée 14-7

$$U_{vdw}(ij) = \epsilon_{ij}(\frac{1.07}{\rho_{ij} + 0.07})^7(\frac{1.12}{\rho_{ij}^7 + 0.12} - 2) \tag{1.13}$$

où ϵ_{ij} en kcal/mol est la profondeur potentielle du puits, et $\rho_{ij} = R_{ij}/R_{ij}^0$ où R_{ij} en Å est la séparation réelle entre i et j, et R_{ij}^0 est la distance énergétique minimale. Pour les paires d'atomes hétérogènes, les règles de combinaison sont données par :

$$\epsilon_{ij} = \frac{4\epsilon_{ii}\epsilon_{jj}}{(\epsilon_{ii}^{1/2} + \epsilon_{jj}^{1/2})^2} et R_{ij}^0 = \frac{(R_{ii}^0)^3 + (R_{jj}^0)^3}{(R_{ii}^0)^2 + (R_{jj}^0)^2} \tag{1.14}$$

Les paramètres vdW dans AMOEBA sont dérivés en s'appuyant sur des données *ab initio* issues de la phase gaz et sont optimisés pour aussi reproduire les propriétés expérimentales de la phase liquide.

- Interactions électrostatiques permanentes :
  L'énergie électrostatique dans AMOEBA comprend les contributions des multipôles permanents et induits. Les multipôles atomiques permanents (PAM) à chaque centre atomique comprennent les moments monopolaires (charge), dipolaire (dipôle) et quadrupolaire (quadrupôle) :

$$M_i = [q_i, \mu_{ix}, \mu iy, Q_{ixx}, Q_{ixy}, Q_{ixz}, ...Q_{izz}]^t \tag{1.15}$$

  où q_i est la charge ponctuelle située au centre de l'atome i, μ est le dipôle et Q est le quadrupôle, le tout en représentation cartésienne, et t est la transposée. Le dipôle et le quadrupôle sont définis par rapport à un référentiel local défini par les atomes voisins.

- Polarisation électronique :
  La polarisation électronique fait référence à la distorsion de la densité électronique sous l'influence d'un champ externe. Elle représente une contribution majeure à la description énergétique globale à plusieurs corps des amas moléculaires et des phases condensées. Dans AMOEBA, un moment dipolaire ponctuel classique est induit à chaque site atomique polarisable en fonction du champ électrique ressenti par ce site. La polarisation moléculaire est calculée via un modèle d'induction utilisant des polarisabilités atomiques distribuées. Ce schéma d'induction mutuelle nécessite qu'un dipôle induit produit sur n'importe quel site i polarise davantage tous les autres sites, et une telle induction mutuelle se poursuivra jusqu'à ce que les dipôles induits sur chaque site atteignent la convergence. L'amortissement de l'interaction de polarisation à très courte distance évite ce qu'on appelle la catastrophe de polarisation, un artefact bien connu des modèles de polarisabilité ponctuelle. L'amortissement est effectivement réalisé en étalant l'un des moments multipolaires atomiques dans chaque paire de sites d'interaction. La fonction d'étalement des charges adoptée par l'AMOEBA a la forme fonctionnelle suivante (dite de Thole) :

$$\rho = \frac{3a}{4\pi} exp(-au^3) \tag{1.16}$$

  où $u = r_{ij}/(\alpha_i \alpha_j)^{1/6}$ est la distance effective en fonction de la séparation linéaire r_{ij} et des polarisabilités atomiques des sites *i*(α_i) et *j*(α_j). Le facteur " a" est un paramètre de largeur sans dimension de la distribution de charge étalée et contrôle efficacement la force d'amortissement.

1.1.3 Minimisation d'énergie

Après la préparation du système pour la simulation de dynamique moléculaire, il est important de minimiser l'énergie du système pour relaxer les structures des molécules. La minimisation d'énergie permet ainsi d'optimiser la géométrie de l'assemblage atomique jusqu'à l'obtention d'un minimum local sur la surface de potentiel. En pratique, pour une molécule de N atomes, la fonction à minimiser comprend 3N variables. À partir du jeu de coordonnées initiales, on recherche le jeu de coordonnées

cartésiennes qui réduit à son minimum la somme de toutes les interactions énergétiques. Les méthodes de minimisation d'énergie sont basées sur des algorithmes généraux d'optimisation. Les méthodes les plus utilisées sont le "steepest descent" (SD) ou méthode de la plus grande pente et le "conjugated gradient" (CG) ou la méthode du gradient conjugué. De façon générale, SD converge plus rapidement vers le minimum (la conformation de protéine la plus stable), tandis que CG y arrive en moins d'itérations [133]. Nous avons décidé d'opter pour la minimisation de nos systèmes dans nos études par la méthode SD.

1.1.4 Algorithmes d'intégration en dynamique moléculaire

L'algorithme de Verlet (Verlet, 1967) permet de calculer les vitesses et les positions des atomes à un temps t+δt lorsque l'on connaît ces dernières au temps t et à condition que l'on choisisse un pas de temps δt suffisamment petit. On procède ainsi par subdivision de la trajectoire en une série d'états discrets séparés par un intervalle de temps δt très court qui définit alors le pas d'intégration des équations précédentes.

Cependant le multi-pas de temps est essentiel afin de pouvoir accélérer la MD en préservant la dynamique. Ceci est possible en utilisant des intégrateurs réversibles de l'algorithme de propagation du système de référence (RESPA)[132] via un noyau Verlet avec le potentiel divisé en portions à évolution rapide et en portions à évolution lente. En effet, les énergies qu'il faut calculer pour intégrer la dynamique à chaque pas de temps, sont divisées en : des termes liés (interactions intramoléculaires) d'une part, et des termes non-liés (interactions intermoléculaires) d'autre part. Les termes d'énergie non liés varient rapidement et les forces dérivant de ces énergies varient également avec une fréquence élevée. Si nous voulons bien observer ces mouvements, nous aurons besoin d'un nouveau pas de temps δt. Pendant ce temps, les termes non liés (van der Waals, électrostatique, polarisation) varient beaucoup plus lentement (surtout à longue distance), de sorte que le calcul de leurs dérivées aussi souvent que l'on calcule les forces des termes liés conduit à une quantité considérable de temps passé sur des calculs non nécessaires. Choisir d'utiliser un pas de temps plus grand pour toute la simulation ne permettrait pas de reproduire les termes variant rapidement, et conduirait à une accumulation d'énergie dans ces modes. Ce n'est donc pas une solution viable.

Un intégrateur idéal jonglerait ainsi entre deux longueurs de pas de temps, l'une pour les termes énergétiques à haute fréquence, la seconde pour les termes à basse fréquence. C'est notamment le cas de l'intégrateur BAOAB-RESPA. Ce dernier est basé sur la mise en place d'un schéma BAOAB (détaillé dans [115]) en association avec le découpage RESPA. Cet intégrateur permettra de stabiliser la dynamique et de viser des pas de temps plus élevés.

1.1.5 Ensembles thermodynamiques

Trois ensembles thermodynamiques, NVE, NVT et NPT, sont majoritairement utilisés en dynamique moléculaire, avec N désignant le nombre de particules, toujours constant. Ces ensembles contrôlant chacun, avec ses propres paramètres, les propriétés physiques des systèmes permettant de produire des conditions physiques pour le système en question.

Historiquement, l'ensemble micro-canonique NVE (nombre de particules, volume et énergie totale constants) est l'ensemble naturel de la dynamique moléculaire d'un système conservatif, car l'énergie et le volume sont conservés au cours du temps. Cependant, cet ensemble se trouve relativement éloigné des conditions expérimentales d'un système réel puisque le système n'est pas soumis à d'autres restrictions et peut évoluer sans limites. Les simulations ont donc été par la suite réalisées dans différents ensembles statistiques (canonique, isobare-isotherme, etc...) selon les conditions dans lesquelles on veut la réaliser et afin de se rapprocher au mieux des observations expérimentales.

Pour l'ensemble NPT connu aussi comme isobare-isotherme, le système physique est entouré d'un bain de pression et de température. C'est l'ensemble le plus proche des conditions expérimentales qui se réalisent souvent sous des valeurs constantes de température et de pression.

Finalement, le troisième ensemble NVT ou canonique, mime aussi les conditions fréquentées en expérience, avec le volume et la température maintenus constants. Il est utilisé pour amener un système vers une température stable de 300 Kelvin. Cet ensemble est utilisé avant la phase de production donc pour l'équilibration du système notamment la réorientation et la relaxation du solvant.

1.1.6 Conditions périodiques

Les simulations de dynamique moléculaire en solution peuvent être effectuées dans des boîtes de solvants de différentes formes (cubique, parallélépipédique, octaédrique, etc...) détentrices de conditions aux limites telles que :

- Lorsque l'oxygène d'une molécule d'eau sort d'une des faces de la boîte, une copie de cette molécule d'eau est repositionnée du côté opposé de la boîte avec la même vitesse.

- Lorsqu'un atome de la molécule ou un ion sort d'un côté de la boîte, une copie de cet atome est placée dans la boîte, et ce afin que le nombre d'atomes soit constant dans la boîte et le système électriquement neutre

Ainsi, afin d'éviter les effets de bord dans le calcul des interactions entre atomes non-liés et les calculs soluté-solvant, la boîte de simulation est dupliquée dans toutes les directions.

1.1.7 Particle Mesh Ewald (PME)

En lien avec la notion de conditions périodiques vue précedemment, la technique de la somme d'Ewald [66] a pour finalité de calculer de façon sûre les interactions électrostatiques entre atomes dans un

environnement soumis à des conditions périodiques aux bornes. Les énergies électrostatiques comprenant à la fois des termes d'interactions de courtes et de longues portées, il convient de décomposer le potentiel d'interaction en termes de courte et de longue portée. L'extension de l'approche d'Ewald à son approximation PME (utilisation de transformées de Fourier rapide ou FFTs) permet en pratique d'accélérer les calculs en réduisant la complexité qui devient proportionnelle à nlog(n).

1.1.8 Limitations de la dynamique moléculaire

En elles-mêmes, les limitations de taille du système et de temps de simulation ne sont pas des obstacles incontournables à l'obtention d'information intéressante pour la compréhension de la fonction biologique. Comme le système est simulé en condition périodique, ceci résout partiellement le problème de la taille. L'extension à des protéines sur des échelles de temps de l'ordre de la nanoseconde, partant de leur structure cristallographique, est souvent très riche en enseignements sur les interactions et corrélations allostériques entre résidus. Enfin, il est maintenant courant d'effectuer des calculs d'énergie libre en forçant une transition conformationnelle dans une protéine selon un chemin choisi. Ces calculs sont lourds et limités à l'espace accessible à la protéine sur les échelles de temps de la nanoseconde le long de ce chemin mais ils donnent souvent de bons résultats directement comparables avec des valeurs expérimentales : δG, barrières d'activation, taux de transition, différences d'affinités... etc.

D'autre part, outre l'émergence de meilleures performances numériques (i.e. gràce aux processeurs graphiques GPUs, ou des processeurs dédiés FGPA), de nouvelles stratégie d'accélération de l'échantillonage statistique sont apparues comme la dynamique moléculaire accélérée (aMD) [232] et dynamique moléculaire accélérée Gausienne (GaMD) [144], mais aussi la métadynamique [117, 20]. Le concept de métadynamique est dérivé de la dynamique moléculaire classique, à la différence qu'un biais va être appliqué au système modélisé, sous forme de contraintes. La métadynamique a été particulièrement étudiée dans l'équipe du Pr. Parrinello (ETH Zurich et Université de Lugano). Ce type de MD permet d'accéder à des échelles de temps plus importantes, et donc l'observation de phénomènes "rares" tels que la liaison d'un ligand à son récepteur ou la dissociation d'un ligand à haute affinité.

1.1.9 Logiciel - Tinker

Le logiciel de modélisation moléculaire Tinker [176] developpé par Washington University (Jay Ponder), UT Austin (Pengyu Ren) et notre Laboratoire est une plateforme complète et générale pour réaliser des calculs de mécanique et de dynamique moléculaires. Tinker a la possibilité d'utiliser plusieurs jeux de paramètres courants, tels que Amber (ff94, ff96, ff98, ff99, ff99SB), CHARMM (19, 22, 22/CMAP), Allinger MM (MM2-1991 et MM3-2000) , OPLS (OPLS-UA, OPLS-AA), Merck Molecular Force Field (MMFF), modèle polarisable de Liam Dang, AMOEBA (2004, 2009, 2013, 2017, 2018) champs de force multipolaires atomiques polarisables, AMOEBA+ qui ajoute des effets

de pénétration de charge, et un nouveau champ de force HIPPO (Hydrogen-like Interatomic Polarizable POtential). Des ensembles de paramètres pour d'autres modèles de champ de force sont à l'étude pour des versions futures. En plus de Tinker classique, il existe des branches : Tinker-HP [114], où SIBFA a été récemment porté, et Tinker-OpenMM [138], conçues pour être utilisées respectivement sur des super-ordinateurs à mémoire distribuée parallèle MPI et sur des unités de traitement graphique (GPU) de pointe [176]

1.2 Protocole générale de la dynamique moléculaire

1.2.1 Structures initiales et construction des systèmes

Les structures de départ utilisées pour réaliser ces simulations sont des structures qui peuvent êtres obtenues par diffraction aux rayons X (cristallographie), par études RMN ou CryoEM, ou même elle peuvent provenir d'une modélisation *in silico* (le célèbre exemple d'Alphafold de Google). Afin de maintenir un environnement aqueux constant tout en diminuant le nombre d'atomes, les conditions périodiques aux limites ("Periodic Boundary Conditions" (PBC)) ont été appliquées. En pratique, lors d'une simulation de dynamique moléculaire, des contraintes sont appliquées pour maintenir la protéine d'intérêt dans une boîte de dimensions fixes remplie de molécules d'eau [15].

Deux méthodes existent pour définir la génération d'une boîte d'eau, soit de façon implicite soit explicite :

- La méthode implicite simplifie les calculs en considérant un continuum de solvatation pour alléger la simulation. Ainsi, une approximation des interactions électrostatiques de longue portée réduit considérablement le nombre de molécules à considérer lors de la MD [14].

- À l'inverse, la méthode explicite signifie que chacune des molécules d'eau est simulée dans le système. Bien que cela augmente la complexité de la simulation en augmentant le nombre de degrés de liberté (i.e. le système contient beaucoup plus de molécules [46]), ceci permet une meilleure description du système, notamment en gérant mieux les interactions à courte distance.

Dans notre étude nous avons choisit d'effectuer nos simulations de MD en utilisant l'eau explicite, pour permettre aux simulations d'être plus fidèles à la réalité. Ajoutons que les progrès technologiques de nos jours nous permettent de surmonter les limites computationnelles historiquement liées à la prise en compte explicite de la solvatation.

1.2.2 Minimisation

Une fois le système construit, il subit une étape de minimisation afin de permettre une organisation réaliste de l'eau et des ions autour de la molécule ainsi qu'une optimisation de la géométrie de la structure de départ.

1.2.3 Chauffage

Le chauffage est exécuté dans un ensemble NVT (nombre de particule, volume et température constants) par palier de 50 K entre 0 K et 300 K. Les vitesses sont distribuées aléatoirement à chaque palier de chauffage lequel est suivi d'une période d'équilibration de 1 ns sans contraintes sur les atomes du système.

1.2.4 Production

Les productions sont réalisées dans un ensemble NPT (nombre de particules, pression et température constants) et en utilisant les algorithmes PME pour le calcul des forces électrostatiques. Ces calculs ont été faites majoritairement sur des GPUs, le cas échéant sur des CPUs.

1.2.5 Stratégie d'échantillonnage adaptative non supervisé

L'échantillonnage adaptatif est utilisé depuis de nombreuses années et s'est avéré être un outil d'exploration puissant pour étudier le repliement et la dynamique des protéines, la liaison des ligands et une variété d'événements moléculaires rares [31, 244, 26, 80]. Pour cette famille d'approches, plusieurs itérations de simulations MD indépendantes sont effectuées, en basant les conditions initiales à chaque itération sur les résultats des étapes d'itération précédentes. Nous proposons une nouvelle stratégie d'échantillonnage adaptatif non supervisée (i.e. entièrement automatisé) dédiée à notre utilisation spécifique des PFF au sein de grands systèmes de supercalculateurs permettant l'utilisation simultanée de centaines ou de milliers de cartes GPU.

Les résultats présentés dans cette Thèse ont pu bénéficier d'une accélération GPU dans le nouveau code GPU Tinker-HP [7] qui a été utilisé pour la première fois pour nos simulations COVID-19. Cependant, la procédure est tout à fait générale et peut être appliquée à toute plateforme de calcul homogène ou hétérogène compatible avec Tinker-HP [114, 101] ou tout logiciel MD. Par conséquent, compte tenu de la répartition particulière des ressources numériques disponibles, les simulations sont organisées par itérations comme suit. Au début de chaque itération, des structures initiales sont sélectionnées parmi les configurations échantillonnées dans les itérations précédentes, à partir desquelles des simulations MD indépendantes sont exécutées, générant de nouvelles configurations. La sélection des structures initiales à chaque itération suit une procédure adaptative conçue pour améliorer l'exploration d'un espace de faible dimension de variables lentes.

Plus précisément, notons M_k le nombre de configurations disponibles au début de l'itération $k \geq 0$, et $(q_i)_{1 \leq i \leq M_k}$ les configurations. Ici, une configuration signifie les positions $q \in R^{3N}$ de tous les atomes du système. En particulier, au tout début de l'algorithme, nous supposons que nous partons de configurations $M_0 \geq 1$, obtenues à partir d'une première simulation MD. Au début de l'itération k, d'abord, la protéine est alignée dans toutes les configurations, en utilisant les atomes du squelette de la structure de référence. Une analyse de la composantes principales (PCA) [6] (qui sera détaillée

dans la section suivante) est ensuite réalisée, en utilisant les "packages" scikit-learn [164] et MDTraj [140], sur les atomes de protéines $(q_i)_{1\leq i\leq M_k}$, dont on considère les $n = 4$ modes principaux. On note $\xi_k : R^{3N} \to R^n$ la projection orthogonale sur ces n modes principaux et on écrit $x_i = \xi_k(q_i)$. Au début de l'itération k, cela représente l'estimation actuelle des variables lentes du système, et afin d'améliorer l'échantillonnage, une exploration de toutes les valeurs de ces variables lentes est envisagée. En d'autres termes, il convient idéalement que les valeurs de x échantillonnées soient uniformément distribuées sur un ensemble compact de R^n. La procédure de sélection est conçue pour pousser l'exploration en direction de cette cible idéale.
Parmi toutes les structures actuellement disponibles, q_i est choisie pour être la structure initiale d'une nouvelle simulation avec une probabilité inversement proportionnelle à sa densité (dans l'espace de faible dimension donné par les quatre premières composantes de la PCA).

La densité ρ_k des variables collectives est approximé par un noyau gaussien, i.e. pour $x \in R^n$

$$\rho_k(x) = \frac{1}{(2\pi\sigma^2)^{n/2} M_k} \sum_{i=1}^{M_k} \exp\left(-\frac{|x - x_i|^2}{2\sigma^2}\right) \tag{1.17}$$

pour $\sigma > 0$. Pratiquement, nous avons utilisé la méthode D.W. Scott, implementée dans Scipy [219], pour estimer une bande passante appropriée σ. Désignons par s_k le nombre des trajectoires MD qui vont être exécutés pendant l'itération k. Afin de pouvoir sélectionner les structures initiales $(q_{I_1}, \ldots, q_{I_{s_k}})$ de ces simulations, les indices $I_1, \ldots, I_{s_k}$ sont générés en tant que variables aléatoires indépendantes dans $\{1, \ldots, M_k\}$ répartis selon

$$P\left(I = i\right) = \frac{\rho_k^{-1}(x_i)}{\sum_{j=1}^{M_k} \rho_k^{-1}(x_j)} \tag{1.18}$$

Autrement dit, parmi toutes les structures actuellement disponibles, q_i est choisi pour être la structure initiale d'une nouvelle simulation avec une probabilité inversement proportionnelle à sa densité (dans l'espace de faible dimension donné par les quatre premières composantes de la PCA). L'effet de cette sélection peut intuitivement être illustré comme suit : si deux domaines de taille similaire (au sens de la mesure Lebesgue sur R^n) ont été visités, avec l'un qui concentre la plupart des trajectoires passées tandis que l'autre ne contient que quelques points, environ la moitié des nouvelles structures initiales seront sélectionnées dans chaque domaine; au contraire, une sélection uniforme parmi les configurations passées aurait mis beaucoup plus de poids sur le domaine dense.

A partir des structures initiales $(q_{I_1}, \ldots, q_{I_{s_k}})$, les simulations MD indépendantes sont échantillonnées, et l'état de chaque simulation est enregistré toutes les 0.1 ns (la structure initiale n'est pas enregistrée, puisqu'elle a déjà été enregistrée dans l'une des itérations précédentes). Ici, indépendant signifie que les vitesses initiales (échantillonnées selon la densité gaussienne d'équilibre) et les bruits blancs des thermostats Langevin sont indépendants (et, bien sûr, indépendants des itérations précédentes, de sorte qu'une trajectoire commençant à une certaine configuration q_i sera différente de la trajectoire

qui a initialement produit ce q_i). A la fin de cette k^{th} les structures d'itération $(q_j)_{M_k<j\leq M_{k+1}}$ ont été ajoutées, et l'itération $k+1$ démarre.

La procédure pénalise les zones déjà largement visitées, et rappelle en quelque sorte la méthode métadynamique [118] sauf que le biais statistique se fait par une étape de sélection entre chaque itération plutôt que par une force de biais mise à jour le long de la trajectoire. Par rapport à la métadynamique, cette étape de sélection non supervisée a l'avantage de s'affranchir du choix critique d'une variable collective initiale au début de la simulation renforçant l'automatisation du schéma d'échantillonnage.

Cette stratégie appartient à la famille des algorithmes d'échantillonnage adaptatif basés sur les comptages, où l'on n'exploite que le nombre de passages dans les différents états (micro ou macro) visités dans les itérations précédentes pour choisir à partir de quel état relancer les trajectoires. Ceux-ci sont connus pour être efficaces à des fins d'exploration pure, même si des algorithmes plus raffinés existent lorsque certaines informations sont disponibles quant à l'endroit où l'échantillonnage doit être guidé[244]. Cependant, contrairement à ce qui se fait habituellement dans le contexte des Markov State Models (MSM) [31] les états ne sont pas définis en appliquant un algorithme de clustering aux structures déjà explorées, mais sont la projection sur les n composantes principales générés par PCA (ici, $n = 4$ comme nous l'avons vu) de toutes les données précédentes. Cela a l'avantage de fournir une stratégie d'échantillonnage non supervisée qui ne repose pas sur un algorithme de clustering particulier (et donc de ses paramètres associés) et de traiter chaque point de cette représentation à 4 dimensions différemment.

1.2.6 Débiaisage

A la fin de la simulation, les configurations M_K doivent être échantillonnées avec K le nombre total d'itérations. Pour un K important, la distribution de ces configurations ne converge pas vers la distribution canonique à cause du biais statistique induit par la sélection. Pour calculer les grandeurs thermodynamiques, ce biais doit être pris en compte. Ainsi, il convient de calculer un score $\omega_i > 0$ pour chaque $i \in \{1, \ldots, M_K\}$ afin que la moyenne canonique d'un φ observable soit estimée par

$$\langle\varphi\rangle \simeq \frac{\sum_{i=1}^{M_K} \omega_i \varphi(q_i)}{\sum_{i=1}^{M_K} \omega_i} \tag{1.19}$$

Le score ω_i est le rapport entre les probabilités d'obtenir q_i dans la simulation biaisée et dans une simulation débiaisée (où, entre chaque itération, les prochaines conditions initiales sont uniformément choisies parmi toutes les configurations actuellement disponibles, c'est-à-dire toutes avec probabilité $1/M_k$).

Par conséquent, il est calculé comme suit : pour tout $i \leq M_0$, $\omega_i = 1$. Supposons par induction que ω_i a été calculé pour tous $i \leq M_{k-1}$ pour certaines k. $(i_1, \ldots, i_{s_k})$ sont les indices qui ont été choisis au hasard pour les conditions initiales au début de l'itération k. Pour chaque $h \in \{i_1, \ldots, i_{s_k}\}$,

nous calculons

$$\alpha_h = \frac{1}{M_k P(I=h)} = \frac{\rho_k(x_h)}{M_k} \sum_{j=1}^{M_k} \rho_k^{-1}(x_j) \tag{1.20}$$

Ensuite, le score de toutes les configurations générées lors de l'itération k à partir de la condition initiale q_h est $\alpha_h \omega_h$. De cette façon, ω_i est calculé pour tous les $i \leq M_k$.

Ce dernier point est important car il signifie que le temps total de simulation peut être utilisé pour calculer des propriétés statistiques moyennes débiaisées et donc explicables. Par exemple, il est possible de les comparer à ceux obtenus lors de la réalisation d'analyses MD conventionnelles.

Enfin, il convient de noter qu'à la place de la PCA, cette stratégie d'échantillonnage adaptatif peut être utilisée avec n'importe quelle autre variable collective et/ou algorithme de réduction de dimensionnalité. Dans l'ensemble, la procédure est entièrement non supervisée, rapide et peut être utilisée dans Tinker-HP de manière entièrement automatisée.

1.3 Outils d'analyse des trajectoires issues de la dynamique moléculaire

L'analyse des données issues de la MD nécessitent l'utilisation d'outils statistiques et informatiques, tel que l'analyse de la composante principale (PCA), l'analyse de composantes indépendantes du temps et de la structure (tICA) et la capacité d'effectuer la clusterisation. Afin de comprendre les fluctuations de conformations observés, il convient d'exploiter les résultats des calculs en analysant les trajectoires. Plusieurs propriétés structurelles et physico-chimiques peuvent expliquer la relation entre la structure, le dynamisme et l'activité d'une protéine ou d'un ADN/ARN. Les exploitations consistent à calculer par exemple les fluctuations atomiques, les corrélations entre les mouvements des parties distinctes d'une protéine, le potentiel électrostatique vu par un ligand s'approchant d'une protéine, les modes de vibrations d'une macromolécule... etc.).

1.3.1 Clusterisation des structures issues de la MD adaptative

Pour trier et extraire les conformations les plus représentatives parmi plusieurs millions de conformations qui peuvent être générées, il convient d'effectuer une clusterisation de ces données. Il existe de nombreux algorithmes de clusterisation, tels que la PCA et la tICA. Ces algorithmes sont détaillés ci-dessous :

Analyse de la Composante Principale

La PCA ("Principal Component Analysis"), est une méthode statistique d'analyse de données multivariables. Le but est d'identifier la relation entre ces données afin d'en obtenir une représentation

simplifiée. Pour cela, de nouvelles variables, appelées "composantes principales" (PC), sont créées afin de différencier aux mieux les données.

Il s'agit d'une approche à la fois géométrique (les variables étant représentées dans un nouvel espace, selon des directions d'inertie maximale) et statistique (la recherche portant sur des axes indépendants expliquant au mieux la variabilité - la variance - des données).

On représente généralement le résultat d'une PCA par une projection des données en deux dimensions en choisissant les deux premières PC comme axes permettant la plus grande dispersion des données dans ce nouveau plan. Pratiquement, la PCA est une technique statistique qui consiste à réduire la dimensionnalité d'un système pour cerner les facteurs dominants décrivant son dynamisme [51]. La dimensionnalité est décrite par le nombre de composantes principales (PC).

Analyse des composantes indépendantes du temps et de la structure

La méthode tICA ("Time-Lagged Independent Component Analysis") ou l'analyse de composantes indépendantes du temps et de la structure, a été introduite dans la dynamique moléculaire comme méthode permettant de trouver les degrés de liberté les plus lents dans un ensemble de données de séries chronologiques qui peuvent être formés à partir de combinaisons linéaires à partir d'un ensemble de degrés de liberté d'entrée [173, 193]. La tICA peut être utilisée comme méthode de réduction de la dimensionnalité et, à ce titre, elle est quelque peu similaire à la PCA. Cependant, alors que PCA trouve des combinaisons linéaires à haute variance des degrés de liberté d'entrée, tICA trouve des combinaisons linéaires à haute autocorrélation des degrés de liberté d'entrée. Ainsi, alors que la PCA repose sur des propriétés statiques et linéaires faisant en sorte que la décomposition de ces PC serait moins valide pour le dynamisme des molécules simulées [152], nous utilisons tICA, qui rend les PC le plus indépendantes possible afin de mieux décrire et analyser le dynamisme du système.

Suivie de classification par grappes (clusterisation), tICA permet de classifer les propriétés structurelles et physico-chimiques des structures au sein d'une trajectoire de MD. Ainsi, des macro-états ont été identifiés durant les simulations pour chacun des variants analysés. La classification par grappes permet d'observer les similarités entre les macro-états de chacun des variants pour discerner un changement structural ou dynamique en commun.

1.4 La Protéase Principale, M^{pro}, du SARS-CoV-2

L'émergence en 2019 de la maladie grave COVID-19 (COrona VIrus Disease 19) ; causée par le virus respiratoire SARS-CoV-2 qui induit un syndrome respiratoire aigu sévère, fut associée à une mortalité importante [215] qui a poussé l'Organisation mondiale de la santé (OMS) à déclarer le 11 mars 2020 que le COVID-19 est officiellement une pandémie [3]. Les chercheurs autour du monde se sont mobilisés afin d'étudier le nouveau virus dont la séquence génomique appartient à la lignée B [21] de la famille des beta-coronavirus. Ceci fut également possible au sein de notre équipe de

recherche dès le début de la période de confinement internationale en printemps 2020, grâce à nos techniques et avancés informatiques. Jusqu'à ce jour, il n'y a pas de traitements spécifiques contre le COVID-19 disponibles sur le marché. Cependant, certaines options thérapeutiques disponibles qui se limitent à une activité de précaution et à fournir un soutien, sont utilisées pour intercepter des conséquences supplémentaires [8]. Actuellement le 23 Mars 2021 Pfizer a lancé une étude de phase 1 sur un nouvel agent thérapeutique antiviral oral contre le SARS-CoV-2 [89]. Comme pour le SARS-CoV-1 et le syndrome respiratoire du Moyen-Orient (MERS), le génome viral du SARS-CoV-2 code pour des protéines non structurales comme la protéase principale M^{pro}, également appelée cystéine protéase de type 3-chymotrypsine (CCP ou 3CLPro), protéase de type papaïne, l'ARN polymérase ARN-dépendante (RdRp) et l'hélicase (Table 1.3). Il code également pour des protéines structurales comme la glycoprotéine de pointe "Spike" (S) et des protéines accessoires. Alors que les protéines structurales sont nécessaires pour les interactions du virus avec les récepteurs de la cellule hôte lors de l'entrée virale, les protéines non structurales jouent un rôle clé au cours du cycle de vie du virus [245]. Dans le cadre de la recherche des ligands à visée thérapeutique contre le COVID-19, les protéines structurales et non structurales sont considérées comme des cibles prometteuses pour la conception d'agents antiviraux contre le SARS-CoV-2 [245]. Certains médicaments commerciaux susceptibles de former des liaisons hydrogène avec des résidus clés dans la poche de liaison de la M^{pro} ont été identifiés [125, 127], et figurent dans le Tableau 1.1. Ils peuvent avoir une tolérance aux mutations élevée et peuvent également fonctionner comme inhibiteurs pour d'autres coronavirus avec des sites de liaison M^{pro} similaires.

Spécifiquement, la M^{pro} joue un rôle central dans le contrôle de la réplication virale et de la transcription par le traitement protéolytique des polyprotéines virales [60]. Alors que cette protéine n'a pas d'homologue humain, il s'agit d'une cible antivirale idéale. De nombreuses études sur le ligand inhibiteur sont basées sur le ciblage actif des poches de sites. Cependant, aucun n'a atteint les essais cliniques à ce jour [170, 112]. En pratique, en raison de la nature dimère de M^{pro}, une autre stratégie peut être employée pour inhiber son activité par le développement d'inhibiteurs de dimérisation [60, 29]. En effet, la conception d'inhibiteurs de dimérisation a déjà été rapportée pour de nombreuses enzymes virales telles que la transcriptase inverse du VIH, l'intégrase, la ribonucléotide réductase du virus de l'herpès simplex et l'ADN polymérase [246, 29]. En fait, le ciblage de l'interface de dimérisation pourrait potentiellement affecter la poche de substrat et ainsi inhiber l'activité de la M^{pro} en raison de la connectivité allostérique entre l'interface de dimérisation et le site catalytique. Cependant, il y a un manque de données expérimentales et théoriques sur l'inhibition de la dimérisation de la M^{pro} du SARS-CoV-2 puisque peu de rapports sont disponibles [60, 228]. Nous fournissons dans ce manuscrit des simulations approfondies sur la M^{pro} en utilisant le champ de force polarisable AMOEBA (PFF) [180, 236] et la nouvelle stratégie d'échantillonnage adaptatif non supervisé hautement parallèle disponible dans la version GPU [7] de la plateforme Tinker-HP [114]. Ces simulations de plusieurs μs et leur conformation dans l'espace associé ont été comparées à des

Protéine cible	Nom complet	Rôle pendant l'infection virale	Médicament candidat
3CLpro	Coronavirus main protease 3CLpro	Une protéase pour la protéolyse de polyprotéine virale en unités fonctionnelles	Lopinavir
PLpro	Papain-like protease PLpro	Une protéase pour la protéolyse de polyprotéine virale en unités fonctionnelles	Lopinavir
RdRp	RNA-dependent RNA polymerase	ARN polymérase ARN-dépendante pour la réplication du génome viral	Remdesivir, Ribavirin
Protéine "S"	Viral spike glycoprotein	Protéine de surface virale pour se lier au récepteur de cellule hôte ACE2	Arbidol
TMPRSS2	Transmembrane protease, serine 2	Protéase produite par la cellule hôte qui amorce la protéine S pour faciliter sa liaison à ACE2	Camostat mesylate
ACE2	Angiotensin-converting enzyme 2	Protéine de récepteur viral sur les cellules hôtes qui se lie à la protéine virale S	Arbidol
AT2	Récepteur à l'Angiotensin AT2	Effecteur important impliqué dans la régulation de la pression artérielle et du volume du système cardiovasculaire	L-163491

Table 1.1: Liste des médicaments candidats ciblant les différentes protéines structurales et non structurales du SARS-CoV-2

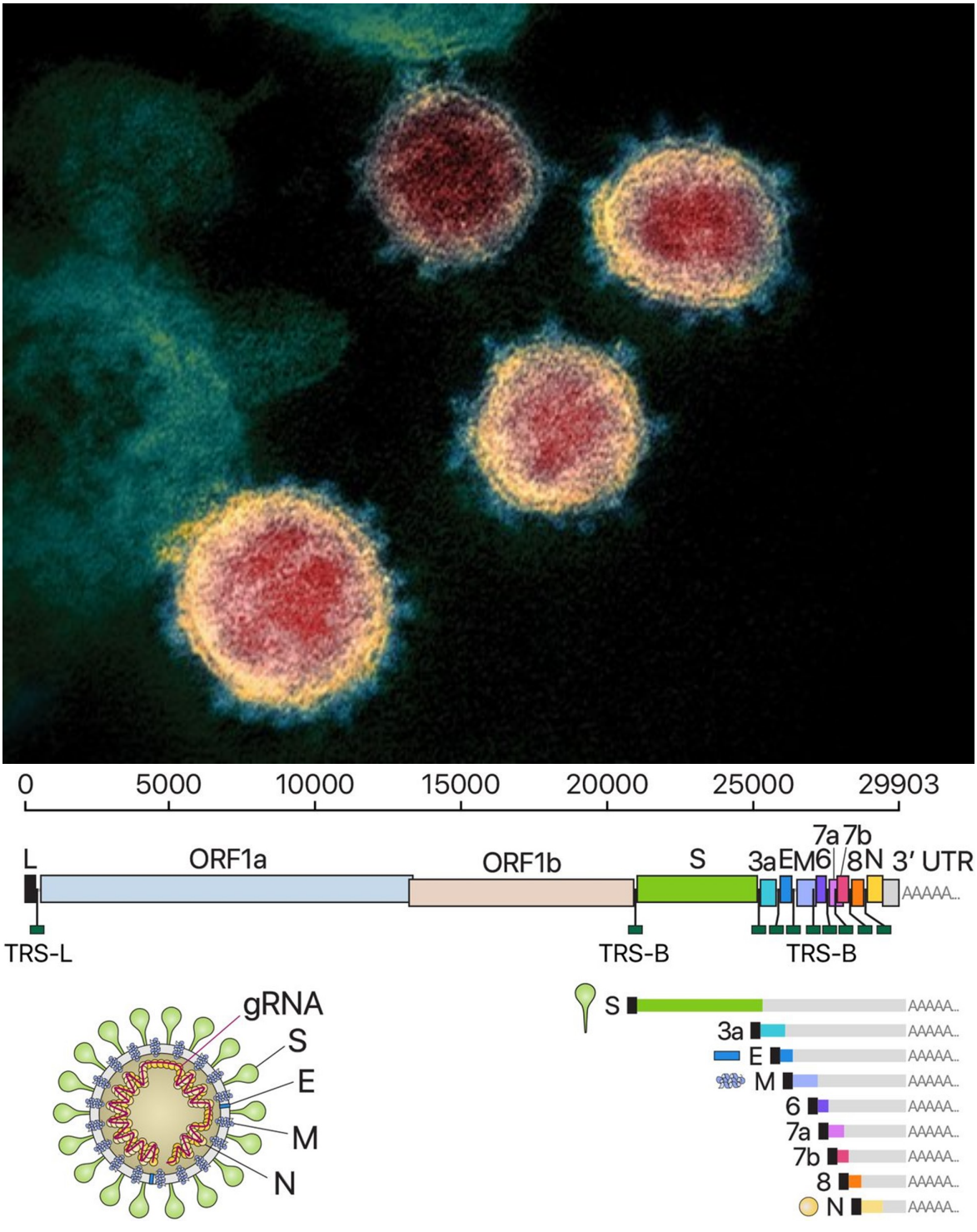
Protéine non structural (nsp)	Fonction
nsp 1 et 3	Inhibition de la signalisation IFN et blocage de la réponse immunitaire innée de l'hôte par la promotion de la dégradation cellulaire et bloquage la traduction de l'ARN de l'hôte.
nsp 2	Liaison à la protéine d'interdiction.
nsp 3	Promouvoir l'expression des cytokines et le clivage de la polyprotéine virale. C'est une protéine qui contient des domaines protéases homologues à la papaïne (papaïne-like protease PLpro).
nsp 5	Promouvoir l'expression des cytokines et le clivage de la polyprotéine virale. Protéase de type sérine qui appartient à la famille de la chymotrypsine (3C-like proteinase 3CLpro ou Main protease, M^{pro}).
nsp 4 et 6	Contribuer à la structure des DMV en tant que protéine d'échafaudage transmembranaire (formation de DMV).
Complexe nsp 7/8	Pince de processivité pour l'ARN polymérase par le complexe hexadécamérique des bras.
nsp9	Protéine phosphatase de liaison à l'ARN.
nsp 10, 16 et 14	Stimulation de l'activité ExoN et 2-O-MT.
nsp 12	Enzyme de réplication : ARN polymérase ARN-dépendante (RdRp).
nsp 13	ARN hélicase, 5'triphosphatase.
nsp 14	Relecture du génome viral.
nsp 15	Endoribonucléase virale et protéase de type chymotrypsine.
nsp 16	Éviter la reconnaissance de MDA5 et inhiber la régulation de l'immunité innée.

Table 1.2: Liste des protéines non structurales des coronavirus et leur fonction [18].

simulations de longue échelle de temps simulés en non PFF et disponible en ligne : DE Shaw Research (DESRES) [57] et RIKEN Center for Biosystems Dynamics Research [110]. Nous avons constaté que AMOEBA est étroitement corrélé avec les données expérimentales, soulignant l'importance de la polarisabilité dans les structures biologiques et donc adéquate pour la découverte de médicaments *in silico*. Afin de conduire d'autres études de "docking" moléculaire, nous présentons dans ce travail une étude de la structuration du site de dimérisation en fonction de différents pH et une analyse détaillée des motifs de solvatation et des effets de polarisation à l'interface de dimérisation. Nos résultats offrent de nouvelles perspectives qui pourraient aider à développer une stratégie d'inhibition de la dimérisation grâce à une stratégie de conception de médicaments afin de désactiver l'activité de la M^{pro} du SARS-CoV-2.

La M^{pro} est active sous sa forme d'homodimère, composé par dimérisation de deux protomères, désignés comme monomère A et monomère B, et la dyade catalytique sur chaque protomère est définie

a)



b)

Figure 1.1: a) Une image au microscope électronique à transmission montre la forme en couronne du SARS-CoV-2. *Crédit: NIAID-RML.* b) Présentation schématique de l'organisation du génome du SARS-CoV-2, des ARNm sous-génomiques canoniques et de la structure du virion. À partir de l'ARN génomique de pleine longueur (29 903 nt) qui sert également d'ARNm, ORF1a et ORF1b sont traduits. En plus de l'ARN génomique, neuf ARN sous-génomiques majeurs sont produits. Les tailles des boîtes représentant les petites protéines accessoires sont plus grandes que la taille réelle de l'ORF pour une meilleure visualisation. La boîte noire indique la séquence de tête [107]

Protéine non structural (nsp)	Fonction
Protéine N	La protéine N est une protéine de 50 kDa qui encapside le génome viral pour former la nucléocapside hélicoïdale. La nucléocapside est entourée d'une bicouche lipidique dérivée de la cellule hôte, dans laquelle sont enchâssées les trois protéines d'enveloppe S, M et E.
Protéine S	La protéine S joue un rôle majeur lors de l'entrée virale [23]. C'est une protéine transmembranaire de type I fortement N-glycosylée, de 180 à 200 kDa qui s'assemble en homo-trimères à la surface de la particule virale [55]. Elle possède un long domaine N-terminal et un court domaine C-terminal. Elle joue une double fonction dans l'entrée virale en permettant d'une part la liaison du récepteur cellulaire et d'autre part la fusion de l'enveloppe virale avec les membranes des cellules cibles. Elle a un rôle déterminant pour le tropisme cellulaire et pour la pathogénicité [81]. Le rôle de cette protéine sera développé ultérieurement.
Protéine M	La protéine de membrane M est la plus abondante de l'enveloppe et joue un rôle majeur dans l'assemblage du virion. Elle est capable d'interagir avec elle-même et toutes les autres protéines structurales : S, E et N [214]. Elle a un poids moléculaire de 25 à 30 kDa [136].
Protéine E	La petite protéine d'enveloppe E a un rôle dans la sécrétion des virions. C'est une protéine peu abondante de 9 à 10 kDa. Elle a une extrémité N-terminale courte, suivie d'un domaine transmembranaire puis d'une extrémité C-terminale constituant la majorité de la protéine. La protéine E ne possède pas de peptide signal clivable ce qui suggère que c'est une protéine transmembranaire de type II [175].

Table 1.3: Liste des protéines structurales des coronavirus et leur fonction.

par les résidues **His41** et **Cys145** [108, 188].

1.4.1 Biologie des coronavirus

Architecture structurale et transcriptomique du SARS-CoV-2

Comme les autres coronavirus (ordre Nidovirales, famille Coronaviridae, sous-famille Coronavirinae), le SARS-CoV-2 est un virus enveloppé avec un génome ARN simple brin de sens positif d'environ 30 kb. Le SARS-CoV-2 appartient au genre betacoronavirus, avec le SARS-CoV-1 et le coronavirus du syndrome respiratoire du Moyen-Orient (MERS-CoV) (avec 80% et 50% d'homologie respectivement) [107, 65]. Les coronavirus (CoV), sont des virus sphériques enveloppés d'un diamètre de 80 à 120 nm [136] qui provoquent principalement des infections enzootiques chez les oiseaux et les mammifères. Mais les épidémies récurrentes de SARS, de MERS et maintenant du COVID-19 ont clairement démontré la capacité remarquable des CoV à traverser les barrières entre les espèces et à se transmettre entre les humains [143]. Les CoV portent les plus grands génomes (26 à 32 kb) parmi toutes les familles de virus à ARN. Chaque transcrit viral a une structure 5'-cap et une queue 3 'poly(A) [116, 233]. Lors de l'entrée dans la cellule, l'ARN génomique est traduit pour produire des protéines non structurales (nsps) à partir de deux cadres de lecture ouverts (Open Reading Frame, ORF), ORF1a et ORF1b. L'ORF1a produit le polypeptide la (pp1a, 440-500 kDa) qui est clivé en 11 nsps. Le décalage de cadre du ribosome –1 se produit immédiatement en amont du codon d'arrêt ORF1a, ce qui permet la poursuite de la traduction de ORF1b, produisant un grand polypeptide (pp1ab, 740-810 kDa) qui est clivé en 16 nsps. Le clivage protéolytique est médié par les protéases virales nsp3 et nsp5 qui abritent un domaine de protéase de type papaïne et un domaine de protéase de type 3C, respectivement. Le génome viral est également utilisé comme modèle pour la réplication et la transcription, qui est médiée par nsp12 hébergeant l'activité ARN polymérase dépendante de l'ARN (RdRP) [198, 199]. Des ARN intermédiaires de sens négatif sont générés pour servir de modèles pour la synthèse d'ARN génomique de sens positif (ARNg) et d'ARN sous-génomique (ARNsg). L'ARNg est emballé par les protéines structurales pour assembler les virions descendants. Les ARNsg plus courts codent pour des protéines structurales conservées (protéine de pointe (S), protéine d'enveloppe (E), protéine de membrane (M) et protéine de nucléocapside (N)) et plusieurs protéines accessoires. SARS-CoV-2 est connu pour avoir six protéines accessoires (3a, 6, 7a, 7b, 8 et 10) selon l'annotation actuelle (Reference de la séquence NCBI : NC_045512.2). La protéine S, est à l'origine du nom de cette famille virale puisqu'elle donne un aspect de couronne en microscopie électronique, Figure 1.1.

1.4.2 Cycle viral du SARS-CoV-2

Les étapes initiales de l'infection du SARS-CoV-2 impliquent la liaison spécifique de la protéine (S) de coronavirus (S) aux récepteurs d'entrée cellulaire ; l'enzyme de conversion de l'angiotensine 2 (ACE2) grâce à une protéine de la surface cellulaire (TMPRSS2) qui facilitera la pénétration dans la cellule de l'hôte. L'expression et la distribution tissulaire des récepteurs d'entrée influencent par conséquent le tropisme viral et la pathogénicité. Au cours du cycle de vie intracellulaire, Figure 1.2, et suite à la décapsidation du virus à l'intérieur du cytoplasme cellulaire, les coronavirus répliquent leur ARN génomique pour produire des copies complètes qui sont incorporées dans des particules virales nouvellement produites. Les coronavirus possèdent des génomes d'ARN remarquablement grands flanqués de régions non traduites 5' et 3' qui contiennent des structures d'ARN secondaire à action cis essentielles à la synthèse de l'ARN. À l'extrémité 5', l'ARN génomique comporte deux grands cadres de lecture ouverts (ORF ; ORF1a et ORF1b) qui occupent les deux tiers du génome coiffé et polyadénylé. ORF1a et ORF1b codent pour 15 à 16 protéines non structurales (nsp), dont 15 composent le complexe de réplication et de transcription virale (RTC) qui comprend, entre autres, les enzymes de traitement et de modification de l'ARN et une fonction de relecture de l'ARN nécessaire au maintien de l'intégrité du génome du coronavirus ($\geq$ 30 kb) [71]. Les ORF qui codent pour des protéines structurales et les ORF intercalés qui codent pour des protéines accessoires sont transcrits à partir du tiers 3' du génome pour former un ensemble imbriqué d'ARNm sous-génomiques (ARNm sg). Les protéines accessoires du coronavirus sont des ensembles très variables de protéines spécifiques du virus qui affichent une conservation limitée même au sein des espèces individuelles, mais on pense principalement qu'elles contribuent à moduler les réponses de l'hôte à l'infection et sont des déterminants de la pathogénicité virale [166, 136]. Néanmoins, les fonctions moléculaires de nombreuses protéines accessoires restent largement méconnues en raison du manque d'homologies avec les protéines accessoires d'autres coronavirus ou avec d'autres protéines connues [60].

Dans un premier temps, comme montré dans la Figure 1.2, le coronavirus SARS-CoV-2 pénètre dans les cellules pulmonaires (étape d'infection) après attachement à leurs récepteurs ACE2 et grâce à une protéine de la surface cellulaire (TMPRSS2). Une fois le virus pénètre dans le cytoplasme de la cellule facilité par TMPRSS2, et suite à la décapsidation du virus, le contenu génomique viral y est libéré. À ce niveau, la polymérase virale catalyse la réplication de l'ARN. La traduction de l'ARNm viral en protéines non-structurales prend lieu au niveau des ribosomes. Ces protéines vont former les complexes de réplication et induire des réarrangements de membrane. La réplication de l'ARN virale est catalysé par l'ARN polymérase ARN dépendante et va permettre la formation de nouveaux ARN génomiques (ARNg) qui pourront être incorporés dans les particules virales néo-synthétisées. Les protéines de structure du virus (S, E, M et N) sont produites suite à la traduction des ARN sous-génomiques (ARNsg). Les nouvelles particules virales sont assemblées et maturées au niveau du compartiment intermédiaire entre le réticulum endoplasmique et le Golgi (REGIC). Elles sont ensuite sécrétées dans le milieu extra-cellulaire. La libération du virus engendrera l'infection d'autres cellules

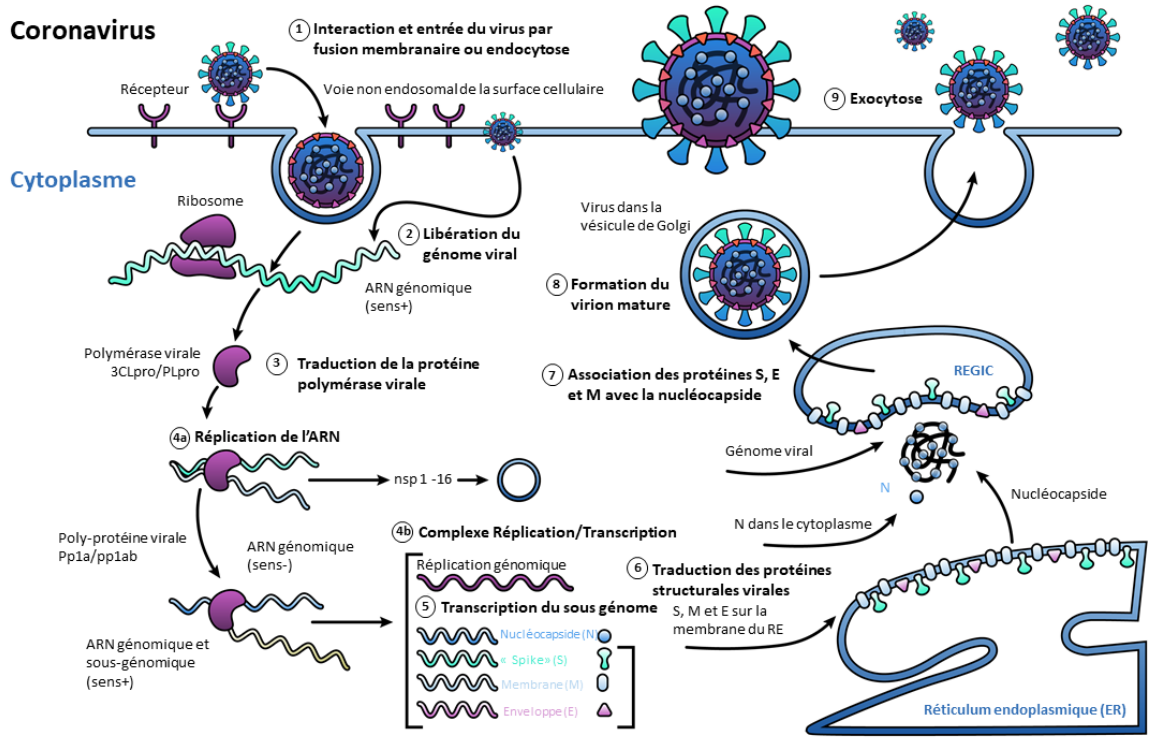


Figure 1.2: Représentation schématique de l'infection des cellules pulmonaires humaine par le SARS-CoV-2 [184].

et le déclenchement de la réponse immunitaire. Le virus va ensuite se multiplier dans la cellule pulmonaire, puis en ressortir pour infecter d'autres cellules.

1.4.3 Importance de la protéase principale, M^{pro}, du SARS-CoV-2

La protéase principale du SARS-CoV-2 (M^{pro}), émerge comme une cible thérapeutique très prometteuse. Cette protéine non structurale du coronavirus est responsable du traitement et donc de la maturation de la polyprotéine traduite à partir de l'ARN viral [79]. En effet, avant auto-activation et libération des polyprotéines virales pp1a et pp1ab, la M^{pro} (Nsp5) fait partie intégrante de ces polyprotéines (Figure 1.3). Le mécanisme d'auto-activation de l'enzyme n'est pas bien compris (voir [229] pour une revue). Plusieurs études ont utilisé des constructions portant des protéines fluorescentes aux deux extrémités du SARS-CoV M^{pro} et connectées à l'enzyme par des séquences peptidiques contenant des sites de clivage M^{pro} [38, 209]. De tels modèles de polyprotéines sont généralement monomériques, mais la formation de dimères lors de l'ajout de substrats a été observée [38]. Suite à des mutations des trois résidus **Arg4**, **Glu290** et **Arg298** (**Glu290Arg**, **Arg298Glu** et partiellement au niveau du **Arg4Glu**) impliqués dans l'interface monomère-monomère de la protéase mature, l'enzyme monomère résultante peut toujours effectuer un autoclivage N-terminal, tandis que l'activité

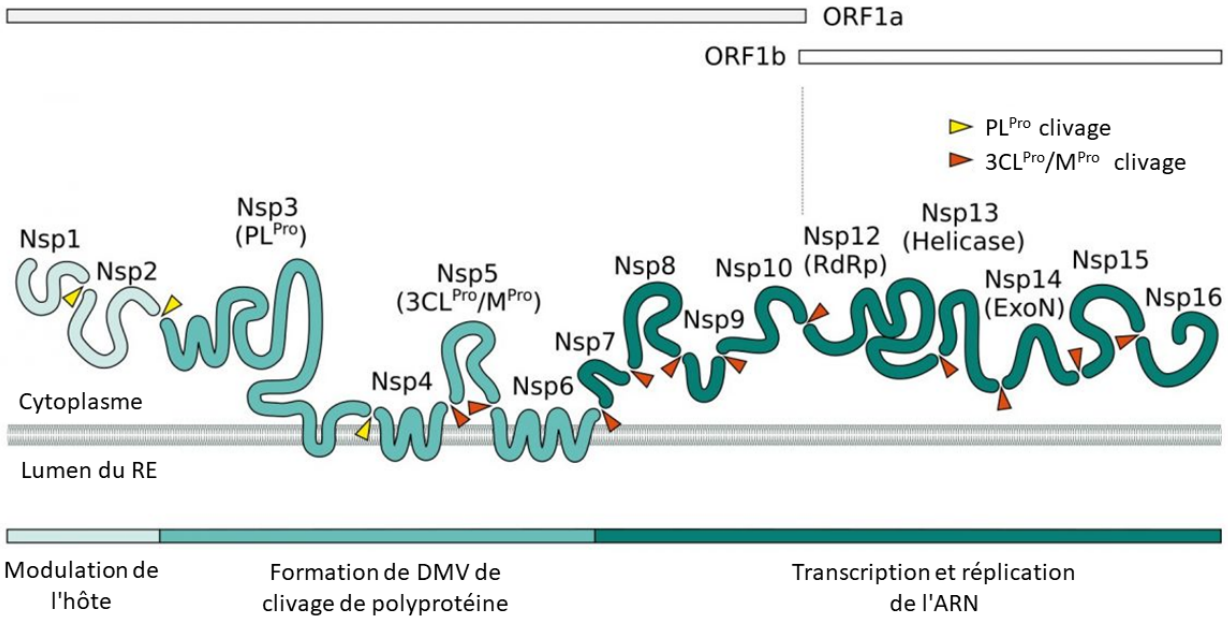


Figure 1.3: Représentation schématique des sites de clivage au niveau de la polyprotéine du SARS-CoV-2 [184]. Occupant les deux tiers du génome à partir de l'extrémité 5', le cadre de lecture ouvert 1 (ORF1) code deux grandes polyprotéines, pp1a et, par le biais d'un décalage du cadre ribosomique pendant la traduction, pp1ab. Ces polyprotéines sont transformées en Nsps matures suite à l'action des deux protéases (indiquées en jaune et en rouge). La protéase principale (M^{pro}, également appelée protéase de type 3C, 3CLpro) est la Nsp5, tandis que la protéase de type papaïne (PLpro) fait partie de la Nsp3. La "Papaine-like Protease" (PLpro) effectue trois réactions de clivage (flèches jaunes) pour libérer Nsp1, Nsp2 et Nsp3, tandis que la M^{pro} clive la polyprotéine sur 11 sites (flèches rouges) pour libérer Nsp4-Nsp16 [79].

de dimérisation et de trans-clivage est complètement inhibée. En outre, le mutant **Glu290Arg** mature peut reprendre une activité d'auto-clivage N-terminal lorsqu'il est mélangé avec une espèce M^{pro} inactive, alors que son activité de clivage trans reste absente. Par conséquent, l'autotraitement N-terminal de la M^{pro} semble ne nécessiter que deux monomères "immature" se rapprochant l'un de l'autre pour former une structure dimère "intermédiaire" et ne dépend pas de la conformation dimère active existant dans la protéase mature [186]. La forme octamère de la M^{pro} immature, qui présente un échange tridimensionnel du domaine hélicoïdal III de l'enzyme [188, 239], peut jouer un rôle dans le processus d'auto-activation.

Compte tenu de ce qui précède, la M^{pro} joue un rôle fonctionnel très important dans la maturation de la poly-protéine virale et la continuation du cycle viral en clivant la polyprotéine virale en unités fonctionnelles précises pour la réplication et la pathogenèse du virus. Comme cette protéase à cystéine est très conservé au niveau du site catalytique fonctionnel, elle est considéré comme une cible très attrayante pour les développeurs des molécules inhibitrices à visée thérapeutique. Il convient alors d'étudier en profondeur sa structure et de modéliser les différentes conformations structurales qu'elle peut adopter afin de bien exploiter les vulnérabilités du virus SARS-CoV-2 qui peuvent fournir des voies vers de nouvelles molécules thérapeutiques antivirales.

1.5 Les quadruplexes de guanine (GQ) au niveau des acides nucléiques

Le support de l'information génétique, l'acide désoxyribonucléique, ou l'ADN, occupe une place centrale dans la compréhension des mécanismes sous-jacents aux processus vitaux comme le code génétique ou encore l'hérédité. Bien que notre appréhension de l'ADN d'un point de vue structural et fonctionnel connaît actuellement un changement radical, depuis la découverte de sa structure en double-hélice en 1953 par Watson et Crick [224], la vision de l'ADN était celle d'une structure sophistiquée, sanctuarisée dans le noyau de la cellule, là où il est précieusement protégé. Or, il est apparu récemment que l'ADN n'est pas seulement un code qui attend patiemment d'être lu et copié au gré de l'activité cellulaire, mais plutôt une structure hautement dynamique qui est capable de gérer elle-même une partie de son activité.

En effet, notre bagage génomique se compose de près de 3 milliards de paires de bases azotées (nucléotides) dont seulement 1 à 2% constitue de l'ADN codant, qui chez l'homme se traduit à environ 20 à 25 000 gènes fonctionnels. La partie non-codante de notre génome, improprement nommée "ADN poubelle", se caractérise par de nombreuses séquences dites "séquences répétées" (qui représente environ 50% de notre génome). Ce n'est que récemment que nous commençons à comprendre l'étendue et la portée de leurs implications fonctionnelles sur le comportement du génome. Parmi ces répétitions, les séquences riches en guanine sont particulièrement au cœur d'une attention scientifique sans précédent. Effectivement les guanines offrent des possibilités d'appariements entre nucléotides bien au-delà de

l'appariement dit de Watson-Crick (normalement entre guanine et cytosine), permettant de former des triplets voire des quartets de nucléotides de guanine (ces derniers étant également nommés G-quartet, ou tétrade de guanine) [197].

Les longues séquences d'oligomères riches en GC sont abondamment représentées dans les oncogènes et les rétrovirus [49, 19]. Leur propension à se réorganiser d'une double à une quadruple hélice était reconnue depuis longtemps [76, 121]. Cela se produit après la séparation des brins et, dans le brin riche en G, la formation de tétramères de guanine cycliques empilés impliquant des bases G bien définies qui peuvent être distantes physiquement dans la séquence linéaire primaire [196]. De tels tétramères G cycliques ont été identifiés plus tôt [35] et sont stabilisés par des appariements de type Hoogsteen : chaque G a ses atomes O6 et N7 acceptant un proton de N1 et N2 d'un G voisin et donne à son tour un proton de son N1 et des atomes N2 à O6 et N7 de son autre voisin G [35] (figure 1.4). Un motif récurrent dans les G-quadruplexes (GQ) est celui de trois tétrades empilées reliées entre elles par une ou plusieurs boucles nucléotidiques, comme présenté sur la Figure 1.4, c).

Les cations de métaux alcalins, principalement K^+ et Na^+, peuvent être des contributeurs essentiels à la stabilité du GQ. Il y a deux cations situés le long de l'axe z des deux tétramères empilés successifs, à proximité de leurs centres géométriques. Les GQ ont été très tôt reconnus comme cibles pour la conception de ligands chimiothérapeutiques. C'est le cas des télomères humains qui dans les cellules cancéreuses ne subissent pas de raccourcissement après chaque division cellulaire, d'où l'immortalisation [49]. Un tel raccourcissement est empêché par l'enzyme télomérase [126]. Ainsi, son inhibition par des ligands de liaison GQ de haute affinité pourrait s'opposer à l'immortalisation [49]. Les GQ sont impliqués dans la régulation de l'expression des gènes aux niveaux transcriptionnel et traductionnel [49]. On les trouve également dans les promoteurs de gènes d'un large éventail de gènes essentiels à la signalisation cellulaire, qui module les six caractéristiques du cancer [33, 77, 85, 68] : l'autosuffisance en signaux de croissance, l'insensibilité aux signaux inhibiteurs de la croissance, la capacité à éviter l'apoptose, la capacité de se répliquer indéfiniment, l'induction de l'angiogenèse et la capacité à former des métastases.

Il est suggéré que les quadruplexes intramoléculaires peuvent être regroupés en deux grandes classes (i) ceux avec au moins une boucle mononucléotidique, montrant souvent des topologies singulières même si les boucles sont très flexibles, et (ii) avec toutes les boucles comprenant au moins deux nucléotides, conduisant à un dynamisme topologique. Les boucles peuvent avoir des structures secondaires empilées de base plus stables et moins dynamiques [226].

Les analyses bioinformatiques [177] ont identifié 375000 séquences candidates dans le génome humain qui pourraient adopter des structures de G4, bien que beaucoup soient probablement dues au hasard et puissent être relativement instables.

Au niveau de l'ADN, les quadruplexes de guanine sont rencontrés au niveau de plusieurs régions, notamment au niveau des télomères chromosomiques sous forme de répétitions (TTAGGG)n. Leur rôle à ce niveau s'inscrit dans le cadre de la protection des télomères contre le raccourcissement chro-

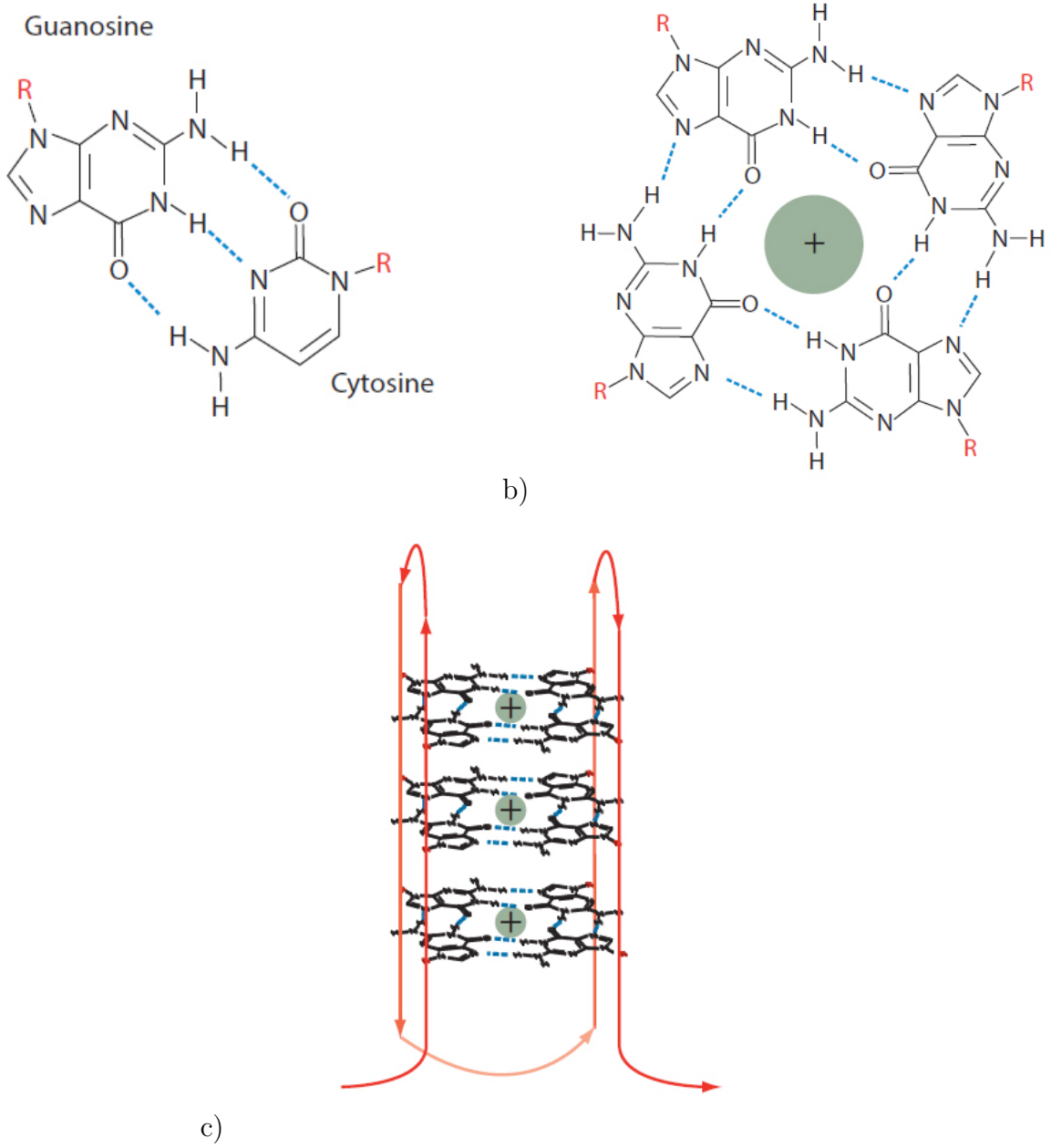


Figure 1.4: a) Appariement des bases de type Watson et Crick de la guanine et de la cytosine. "R" représente le squelette sucre-phosphate et qui peut être soit de l'ARN, soit de l'ADN. b) Représentation d'un tétraplexe de guanosine. En présence de cations monovalents, des guanosines rapprochées peuvent interagir les unes avec les autres pour générer un tétraplexe de guanosine. Un tétraplexe a 4 bases et peut être formé à partir d'une seule molécule nucléique en boucle, ou de jusqu'à 4 molécules d'ARN ou d'ADN distinctes qui se réunissent. c) Quadruplexe de guanoisine intramoléculaire. Deux tétrades ou plus peuvent s'empiler pour former une structure secondaire complexe appelée quadruplexe. L'exemple illustré est formé à partir d'un seul brin d'ADN (indiqué en rouge). Des bases intermédiaires sont nécessaires pour permettre le pliage du brin pour le positionnement des bases G de telle sorte qu'elles prennent les structures tétraplexes individuelles compactes dans l'empilement quadruplex [161].

mosomique qui prend lieu à chaque division cellulaire. Ils jouent alors un rôle contre le vieillissement cellulaire. Ils assurent également la diversité des immunoglobulines en agissant sur la recombinaison dans les régions de commutation des immunoglobulines et modulent l'activité promotrice de gènes tels que l'insuline ou le c-myc [58]. On les trouve également dans les promoteurs de gènes d'un large éventail de gènes essentiels à la signalisation cellulaire et qui module les six caractéristiques du cancer [33, 77, 85, 68]. Les GQ sont également rencontrés au niveau des rétrovirus [37, 36].

La formation/déformation des GQ peut entraîner une perte/un gain d'association(s) de facteur de transcription. La formation de GQ stable inhibera la transcription d'une séquence génomique alors que sa déformation en simple brin d'ADN permettra à la ARN polymérase de réaliser sa fonction de transcription de l'ADN en ARN messager et donc l'expression du gène concerné.

Dans le cadre de notre travail, nous prêtons une attention particulière aux GQ au niveau du proto-oncogène c-kit 1 impliqué dans une variété de types de cancer chez l'homme (particulièrement responsable du tumeur stromale gastro-intestinales) et au niveau de la région LTR-III du VIH-1 responsable du SIDA. Il convient d'évaluer la stabilité des GQ dans ces deux systèmes afin d'identifier des régions bien structurés qui seront des cibles spécifiques des molécules stabilisantes qui empêcheront la déformation des GQ dans le but d'inhiber complètement la transcription du proto-oncogène c-kit 1 et de l'ADN viral.

1.5.1 Les GQ au niveau du promoteur de l'oncogène c-kit 1

CD117/c-kit est un récepteur transmembranaire codé par le proto-oncogène c-kit et appartenant à la famille des récepteurs tyrosine-kinase de type III et ayant une masse atomique de 145 à 160 kDa. Il intervient dans la croissance tumorale et la progression de plusieurs cancers [56]. En pratique, le gène CD117, appelé c-kit, code un récepteur du facteur de croissance tyrosine kinase pour le facteur de cellules souches (SCF) de l'anglais "Stem Cell Factor". SCF appartient à la famille des cytokines et est aussi connu en anglais sous le nom de "steel factor" ou de "c-kit ligand". Lorsque le SCF se lie au récepteur CD117/c-kit, il induit la dimérisation de ce dernier qui active son activité protéine kinase intrinsèque, qui à son tour phosphoryle et active des molécules de transduction du signal qui propagent tout un signal dans la cellule. La transduction du signal par CD117/c-kit joue un rôle dans la prolifération, la différenciation, la migration et l'apoptose de nombreux types de cellules et jouent ainsi un rôle important dans l'hématopoïèse, le maintien des cellules souches, la gamétogenèse, la mélanogenèse et dans le développement, la migration et la fonction des mastocytes, Figure 1.5.

Physiologiquement, CD117/c-kit est exprimé par plusieurs types de cellules, à noter les mastocytes, les mélanocytes, les cellules souches hématopoïétiques, les cellules interstitielles de Cajal, les cellules germinales (ovocytes, spermatocytes) ainsi que par divers épithélia (couche basale de la peau, sein, glandes sudoripares, salivaires) [56].

Dans le contexte du cancer, les métastases sont la principale cause de morbidité et de mortalité des patients cancéreux, mais en raison des lacunes persistantes de nos connaissances, elles restent

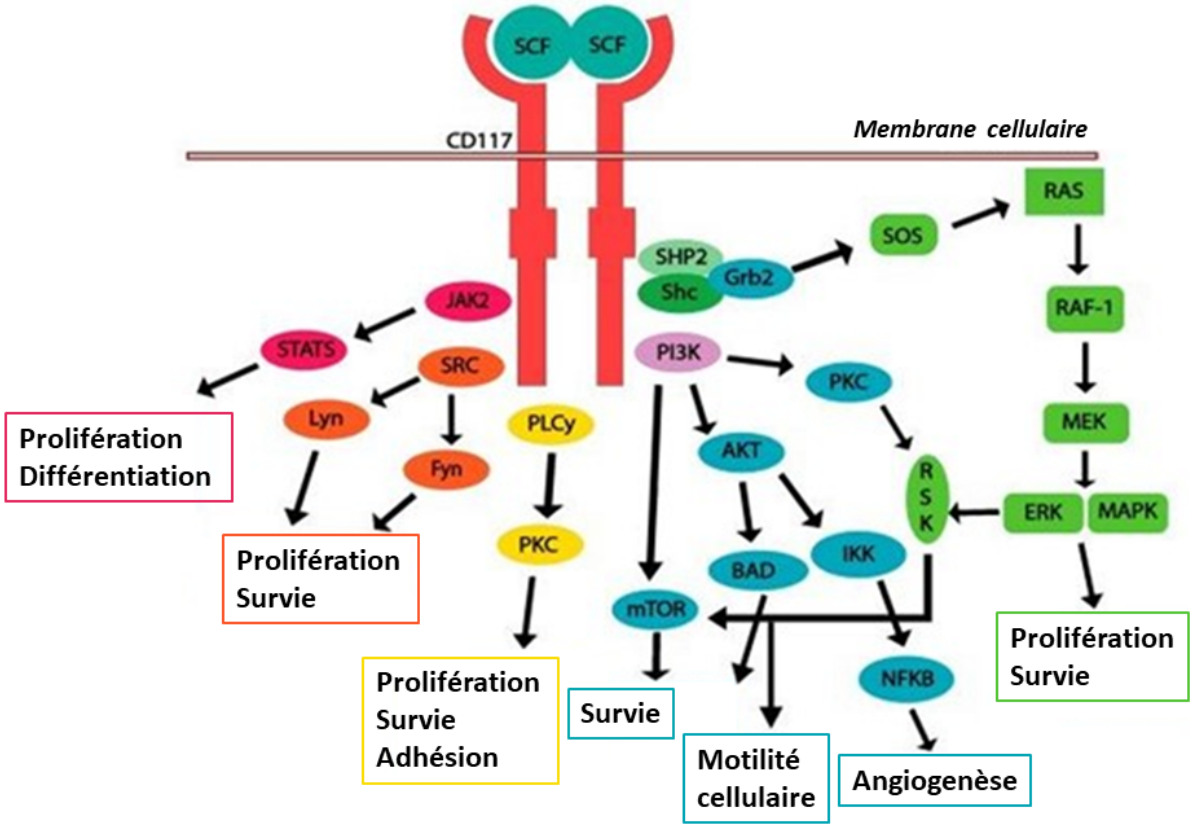


Figure 1.5: Représentation schématique de la voie de signalisation activé par CD117. L'activation de CD117 stimule de multiples voies de signalisation. La liaison du ligand du "Stem Cell Factor" (SCF) au récepteur CD117 induit une dimérisation de ce dernier et une signalisation en aval, entraînant la prolifération, la différenciation, la survie, l'adhésion, la motilité et l'angiogenèse de la cellule.

incurables. Les métastases surviennent souvent à mesure que les tumeurs du patient progressent ou réapparaissent après le traitement initial. La récidive tumorale au site primaire peut être provoquée par une cellule souche cancéreuse ou une cellule progénitrice tumorale, tandis que la récidive à un site secondaire est provoquée par des cellules souches cancéreuses métastatiques ou des cellules initiatrices de métastases. Les efforts en cours visent à identifier et à caractériser ces cellules souches à l'origine de la récidive et des métastases. Un marqueur potentiel pour la sous-population de cellules souches cancéreuses est CD117/c-kit, un récepteur de tyrosine kinase associé à la progression du cancer et au maintien normal des cellules souches. En outre, l'activation de CD117 par son facteur de cellules souches ligand (SCF ; ligand du kit) dans la niche des cellules progénitrices stimule plusieurs voies de signalisation entraînant la prolifération, la survie et la migration cellulaire comme montré sur la Figure 1.5. La voie de signalisation SCF/CD117 peut contribuer au contrôle de la progression du cancer par la régulation de la cellule souche et de la résistance aux inhibiteurs de la tyrosine kinase [67].

CD117/c-kit 1 et régulation de la progression du cancer

La sur-activation de CD117/c-kit est la principale cause observée dans plusieurs types de cancer, tels que les tumeurs gastro-intestinales (GIST), la mastocytose, la leucémie aiguë myéloïde (AML) et le mélanome [106, 200, 129]. Des études et des essais cliniques récents ont suggéré que la quantification du CD117/c-kit peut être utilisé efficacement pour le pronostic, en particulier pour prédire les métastases cancéreuses la planification du traitement et la réponse à la chimiothérapie [56]. Des biomarqueurs impliquant CD117/c-kit ont été identifiés et étudiés sur divers types de cellules tumorales [211, 141]. Dans une étude de classification, Longley *et coll.* ont démontré que CD117/c-kit a été exprimé dans 21 % des cancers du sein, 17 % des cancers colorectaux, 35 % des sarcomes, 36 % des carcinomes à cellules rénales, 17 % des cancers de l'ovaire et 17 % des tumeurs hépatocellulaires. Particulièrement, 89 à 100 % des patients GIST exprimaient CD117/c-kit [129]. La stabilisation des structures d'ADN G-quadruplexe par de petites molécules ayant une affinité pour les promoteurs d'oncogènes est apparue comme une stratégie anticancéreuse prometteuse, en raison d'un rôle potentiel dans la régulation de l'expression génique. La fréquence d'apparition des séquences formant des quadruplexes de guanine autres qu'au niveau des télomères dans le génome humain a été évaluée par des approches informatiques [212, 83, 84]. Ces études ont montré que de telles séquences sont particulièrement répandues dans les séquences promotrices [62] et dans les 5'-UTR [85]. Leur sur-représentation dans un certain nombre de séquences promotrices d'oncogènes [84] a conduit à l'hypothèse que les quadruplexes promoteurs peuvent être des cibles pour une intervention thérapeutique utilisant de petites molécules qui stabilisent une structure quadruplexe particulière au sein d'une séquence promotrice [19]. Ceci entraîne en principe une inhibition de la transcription de ce gène particulier. Le concept a été évalué avec de petites molécules ciblant en particulier les quadruplexes promoteurs c-myc et c-kit. Le défi reste de concevoir de petites molécules capables de cibler sélectivement un quadruplexe de guanine particulier en évitant les effets hors cible sur les autres quadruplexes. Il est important de noter que l'idée de considérer un quadruplexe en tant que cibles dans le cadre des thérapies anticancéreuses humain a été validé par leur visualisation directe dans les cellules et les tissus humains [27, 28].

De là vient l'idée d'une inhibition à un stade précoce de la transcription du CD117/c-kit en ARN (ARNm) et donc de la traduction de ce dernier au récepteur transmembrannaire CD117/c-kit par la stabilisation des structures GQ au niveau de l'ADN. En fait, les GQ pourraient nuire à l'initiation de la transcription par l'ARN polymérase ou, s'ils sont présents dans le brin antisens, inhiber la transcription [49], bloquant ainsi la transcription de l'ARNm. Par conséquent, la stabilisation du GQ au niveau du proto-oncogène c-kit 1 au niveau génomique pourrait être une nouvelle thérapie de transcription pour le cancer.

Afin de pouvoir concevoir de petites molécules stabilisantes du QG, il est important d'évaluer la stabilité structurale du quadruplexe et d'étudier le comportement des boucles. Il convient ensuite de rechercher des poches cryptiques stables au sein des GQ qui pourraient être ciblées par des ligands pharmacologiques.

Architecture génomique du promoteur de c-kit 1

Au niveau génomique, le proto-oncogène codant pour CD117/c-kit est localisé au niveau du bras long du chromosome 4, Figure 1.6. La séquence 22-mer, formée des 22 nucléotides : *d(AG3AG3CGCTG3A-GGAG3)*, en amont de 87 nucléotides du site d'initiation de la transcription du gène humain c-kit, forme une structure GQ stable, dont la structure 3D a été initialement élucidée par RMN en présence d'ions K^+ [167]. Ceci a révélé que le promoteur de c-kit adopte un échafaudage unique où, 16 guanines de la séquence participent à la formation du coeur du GQ1.7). Les guanines adoptent une conformation *anti*, ce qui implique qu'il s'agit d'un GQ à brins parallèles. La structure comprend quatre boucles qui relient les tétramères de guanine ; deux boucles à un seul résidu, une boucle latérale à deux résidus et une boucle à hélice latérale (LP) longue et formée de cinq résidus. La topologie sans précédent de ce promoteur de c-kit en GQ a ensuite été confirmée par plusieurs structures cristallines aux rayons X, également obtenues en présence de K^+ [225, 227]. À ce jour, il s'agit du seul promoteur GQ pour lequel des structures RMN et cristallines sont disponibles : ces études concourent à montrer un seul arrangement replié et démontrent la robustesse de cette topologie en GQ du promoteur c-kit. La première structure cristalline à être rapportée a révélé la présence d'ions et de molécules d'eau bien conservés dans la structure. Ces ions sont dans des positions pratiquement identiques dans les deux molécules de l'unité cristallographique asymétrique, indiquant leur rôle potentiel dans le maintien de l'intégrité structurelle de la boucle.

Notre stratégie de travail consiste à simuler pour plus de 7.5 μs par dynamique moléculaire classique et adaptative en phase aqueuse explicite, la structure GQ au niveau du promoteur de l'oncogène c-kit 1 afin de bien étudier sa stabilité structurale au niveau global et au niveau des poches cryptiques susceptible à être ciblés par des molécules stabilisantes. Une clusterisation des données produites nous aidera à mieux comprendre les différents états conformationnels du GQ.

1.5.2 Les GQ au niveau de la région LTR-III du VIH-1

VIH-1 et stratégies de répression virale

Le Virus de l'immunodéficience humaine (VIH) est responsable du syndrome de l'immunodéficience humaine acquise (SIDA). Il est caractérisé par une infection conduisant à une élimination progressive du système immunitaire de l'hôte infecté, le rendant vulnérable à de multiples infections opportunistes. Ce rétrovirus infecte l'homme et touche actuellement plus de 37,7 millions de personnes dans le monde (UNAIDS - AIDS 2020) [2].

Le VIH se transmet à l'homme par trois voies majoritaires : le sang, contact sexuel et voie verticale (*in-utero* via une transmission transplacentaire, *intra-partum* durant l'accouchement via un contact sanguin et les sécrétions vaginales et *postpartum* : suite à l'accouchement via l'allaitement maternel). Le VIH cible principalement les cellules immunitaires présentant les récepteurs CD4 à leur surface. Ces cellules comprennent les lymphocytes T CD4+, les monocytes et macrophages, ainsi que les cellules

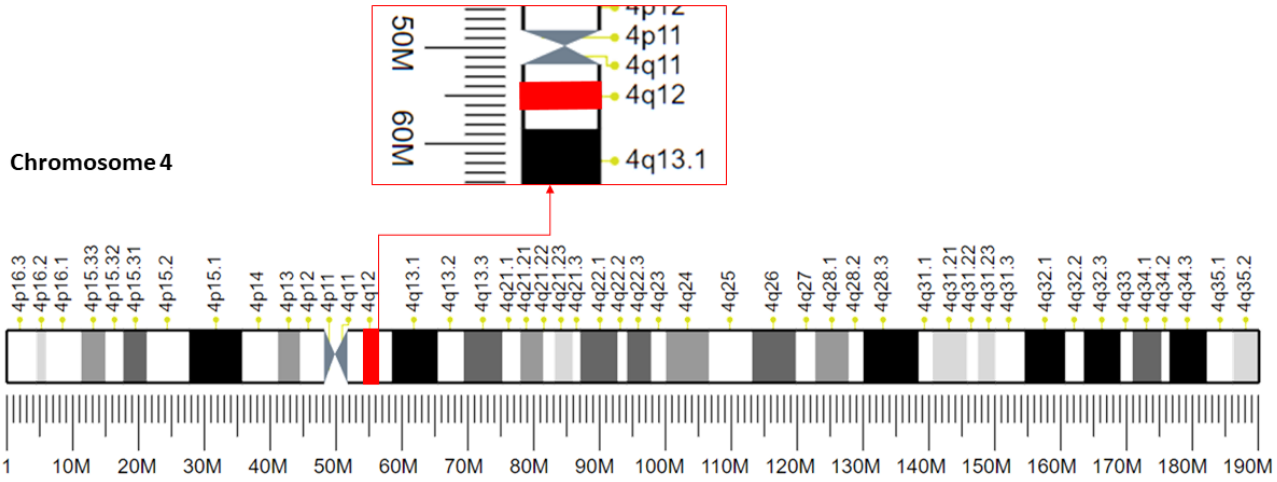


Figure 1.6: Représentation graphique du chromosome 4 humain. Le gène codant du CD117/c-kit est localisé sur la bande 12 du bras long (q) du chromosome 4.

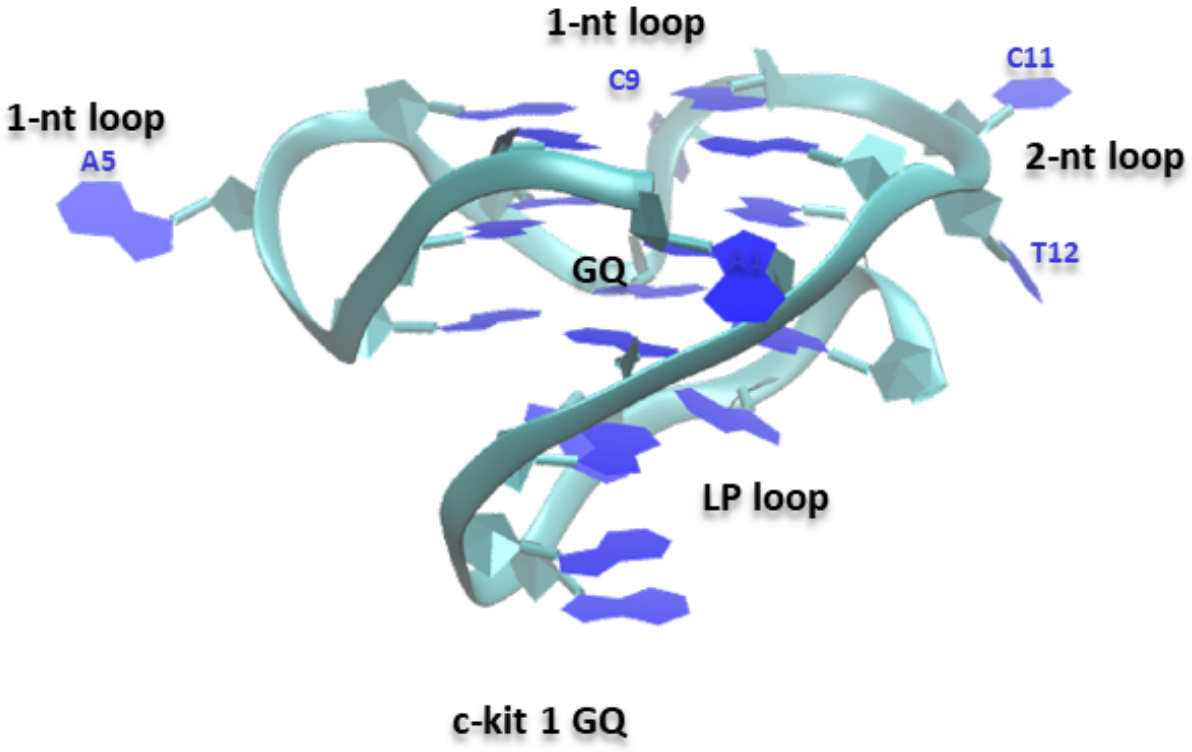


Figure 1.7: Représentation 3D du GQ de c-kit 1 sur VMD.

dendritiques. L'infection par VIH évolue en plusieurs étapes se succédant dans le temps et conduisant à une destruction du système immunitaire de l'hôte. On distingue trois phases caractéristiques à l'infection du VIH [69] :

- La primo-infection
  Cette étape commence après la transmission du virus. Elle est caractérisée par une réplication importante du virus entraînant la destruction des lymphocytes CD4+. Cette étape est suivie par une réponse inflammatoire importante caractérisée par l'apparition des anticorps anti-VIH par stimulation des lymphocytes B et une réponse cytotoxique des lymphocytes CD8+ permettant de contenir l'infection. La production d'anticorps anti-VIH rend le patient séropositif.

- La phase asymptomatique
  Durant la phase asymptomatique qui peut être de durée variable qui s'étend de quelques mois à plusieurs années, le virus se multiplie activement en diminuant progressivement le taux de lymphocytes CD4+. Le système immunitaire contrôle l'infection jusqu'à l'épuisement des CD4+.

- La phase symptomatique ou phase SIDA
  Cette phase est caractérisée par l'apparition des premiers signes cliniques de l'infection. Le sujet sans traitement est immunodéprimé et des pathologies infectieuses ne peuvent plus être contrôlées par le système immunitaire préalablement détruit qui conduit alors au décès du patient.

Le génome rétroviral du VIH est constitué de deux copies d'ARN simple brin de polarité positive composé de 9181 nucléotides. L'ARN viral est encadré de séquences répétées R, de la région U5 en 5' et U3 en 3' qui sera lors de la transcription inverse convertie en ADN permettant la création aux extrémités de séquences répétées non codantes appelées LTR (Long Terminal Repeat) composé des régions U3, R et U5 jouant un rôle crucial dans la réplication virale. Le VIH-1 comporte plusieurs gènes communs à l'ensemble des rétrovirus codants pour les protéines de structure et les enzymes de réplication ainsi que les gènes accessoires. On trouve le gène *gag* (group antigen) codant la polyprotéine structurale du virus Pr55gag maturée par la protéase virale en protéine de matrice MA (p17), de capside (p24), en nucléocapside NC (p7) et en peptide SP1, SP2 et p6. Le gène *pol* code pour les enzymes virales, protéase, transcriptase inverse et l'intégrase. Le gène *env* code quant à lui pour la polyprotéine gp160 précurseur des glycoprotéines de surface gp120 et gp41. On retrouve ensuite les gènes codants pour *tat, rev, vpr* et les autres régulateurs *vif, nef* et *vpu* [171].

Une fois que le génome est rétrotranscrit par l'enzyme virale transcriptase inverse, il s'intègre dans le chromosome de la cellule hôte sous la forme provirus. Le provirus peut alors subir un cycle réplicatif productif ou rester dans un état dormant appelé latence. La progression efficace du cycle viral repose sur le bon fonctionnement de la répétition terminale 5'-LT, qui est caractérisée par des sites de liaison au facteur de transcription et sert de promoteur viral unique [113]. Un schéma simplifié du cycle virale du VIH-1 est représenté dans la figure 1.8.

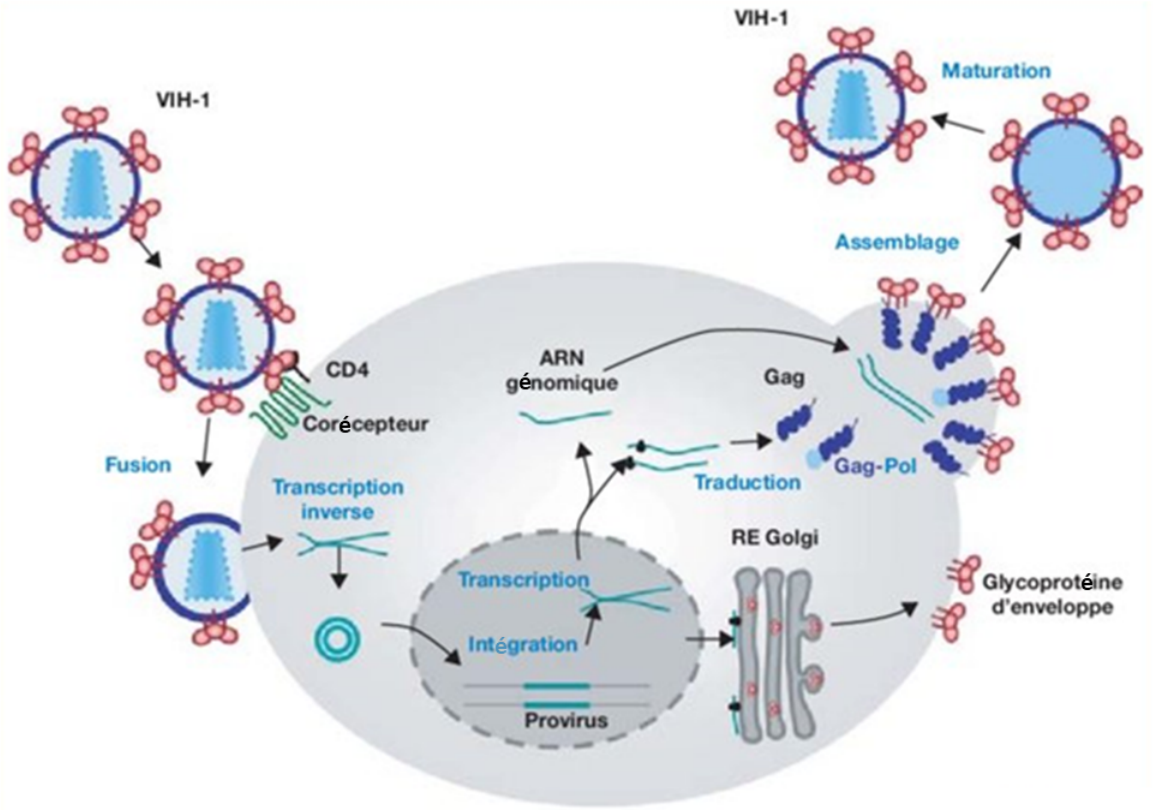


Figure 1.8: Schéma simplifié du cycle de multiplication du VIH-1. Les glycoprotéines d'enveloppe du VIH-1 interagissent avec leur récepteur CD4 et leur corécepteur (CCR5 ou CXCR4 principalement) permettant la fusion des membranes virales et cellulaires. Le core viral est libéré dans le cytoplasme de la cellule hôte. Le génome ARN viral est ensuite rétrotranscrit en une molécule d'ADN double brin. Cette molécule d'ADN est ensuite importée dans le noyau puis intégrée dans le génome de la cellule hôte. Cette forme intégrée est nommée provirus. Les gènes viraux sont alors transcrits. Les ARNm viraux sont épissés ou non, puis exportés dans le cytoplasme ou au niveau du réticulum endoplasmique (pour l'ARNm des glycoprotéines d'enveloppe), et, enfin, sont traduits. Les composants viraux s'assemblent pour former la particule virale qui bourgeonne à la surface de la cellule (dans le cas des cellules T). La protéase virale clive alors les polyprotéines virales Gag et Gag-Pol pour former les virus matures [87].

Il existe actuellement plusieurs traitements contre le VIH, cependant aucune thérapie ne permet de guérir de cette maladie. La stratégie actuelle consiste à l'utilisation d'inhibiteurs ciblant l'entrée du virus en empêchant l'interaction avec la cellule cible ou la fusion des deux bicouches lipidiques [234], mais aussi des inhibiteurs des enzymes virales comme la transcriptase inverse (RT) [213, 52], la protéase (PR) [242] et l'intégrase (IN). La thérapie anti-VIH actuelle tourne autour d'une combinaison de molécules HAART (Highly Active Antiretroviral Therapy) ciblant différentes étapes du cycle viral. L'objectif majeur étant de réduire le risque de développement de virus mutant résistant dû au taux d'erreur important généré par la RT du virus.

La formation de multiples GQ dans le génome viral et proviral [183, 61] et en particulier dans le promoteur LTR [183, 191] a été rapportée. Les GQ au niveau du LTR agissent comme des éléments répresseurs de l'initiation de la transcription virale : la stabilisation par les ligands spécifiques du GQ intensifie cet effet [182] tandis que les protéines cellulaires modulent la transcription virale en induisant/dépliant les GQ dans cette région [63]. La conservation des séquences qui forment les GQ au niveau de la région 5'-LTR est observée dans tous les lentivirus des primates, validant, en outre, les GQ viraux en tant que nouvelles cibles antivirales. Cependant, le ciblage sélectif des GQ viraux avec de petites molécules est difficile et peu de composés reconnaissent des structures spécifiques des GQ [145]. La dynamique moléculaire des structures à haute résolution des GQ viraux peuvent donner de nouvelles perspectives pour atteindre un niveau plus élevé de sélectivité et de spécificité.

Architecture génomique des GQ au niveau de la région LTR-III du VIH-1

Au sein de la séquence LTR riche en guanines, au niveau de la région U3 du génome proviral, la formation de multiples conformations GQ impliquant différentes G est possible [37]. Cette séquence a été divisée en trois composants principaux formant des GQ, à savoir LTR-II, LTR-III et LTR-IV (Figure 1.9). Dans des études précédentes, le GQ formé par la séquence LTR-III a montré la stabilité thermique la plus élevée dans les expériences de dichroïsme circulaire et de fusion FRET. De plus, le test d'arrêt de la polymérase Taq sur la séquence LTR pleine longueur, dans une solution K^+, a révélé un site d'arrêt se produisant principalement sur le site LTR-III et cet effet a été exacerbé avec les ligands spécifiques contre le GQ, tels que BRACO-19. La protéine cellulaire nucléoline est impliquée dans la régulation de l'activité du promoteur virale suite à sa fixation sur le GQ du LTR [63]. Plus précisément, la stabilisation du GQ du LTR se traduit par une diminution de l'activité du promoteur viral. En revanche, la protéine cellulaire hnRNP A2/B1 se lie et déplie les GQ du LTR, c'est-à-dire LTR-II et LTR-III, activant la transcription virale [137]. Cette preuve soutient le rôle clé du GQ du LTR-III dans le motif de repliement des GQ au niveau du LTR dans les événements régulateurs de la transcription du VIH-1. Ainsi, le ciblage sélectif de la conformation GQ du LTR-III avec des ligands stabilisants peut représenter une stratégie intéressante pour inhiber la production de virus.

Butovskaya *et coll.* ont rapporté une structure RMN à haute résolution du GQ du LTR-III formé des 28 nucléotides : 5'-d[<u>GGG</u>A<u>GG</u>CGGT<u>GG</u>CCT<u>GGG</u>C<u>GGG</u>ACT<u>GGGG</u>]-3', contenant six

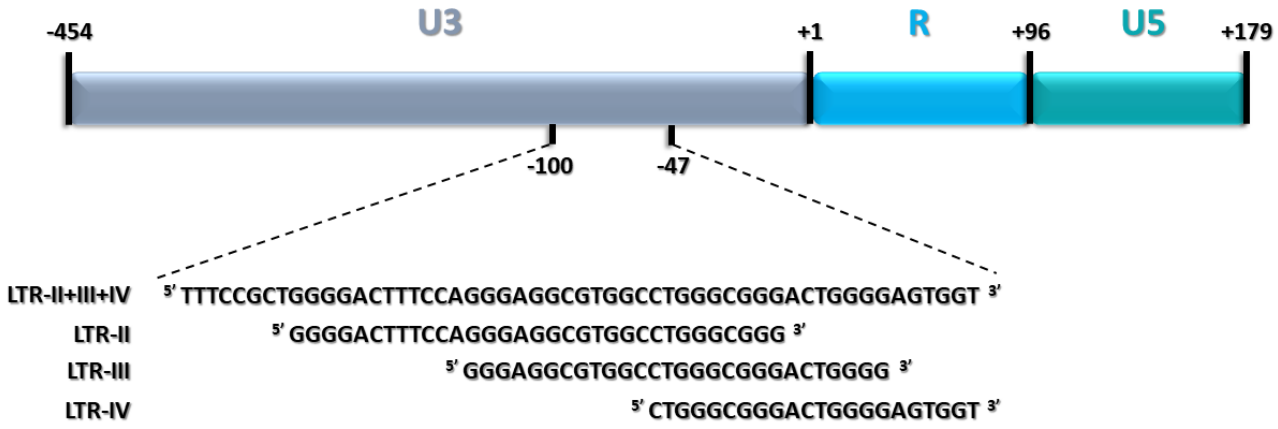


Figure 1.9: Représentation de la séquence riche en G au niveau du LTR de la région promotrice U3 du génome proviral du VIH-1 et les sous-séquences associées LTR-II, LTR-III et LTR-IV.

assemblage de 2 à 4 guanine (soulignés). LTR-III forme une topologie de repliement GQ (3 + 1) avec trois brins (G15 – G17, G19 – G21 et G26 – G28) pointant vers le bas et un brin (G1 – G2) pointant vers le haut ; le noyau de tétrades de guanine présente deux sillons moyens, un sillon large et sillon étroit. Quatre boucles relient les tétrades : une boucle d'hélice de 1 nt (résidu 18), une boucle latérale de 3 nt (du résidu 22 au résidu 24), une boucle en forme de V (entre le résidu 25 et le résidu 26) et une boucle diagonale de 12 nt (du résidu 3 au résidu 14) contenant trois paires de bases Watson - Crick formant une structure en épingle a cheveu ("hairpin") (Figure 1.10).

Afin de proposer un autre modèle de GQ de très haute résolution conçu pour la modélisation des ligands thérapeutiques, nous simulons par dynamique moléculaire classique (400 ns) la séquence du LTR-III capable d'adopter une structure particulière hybride en GQ et en boucle à cheveu. Nous examinons la stabilité globale de cette structure et nous montrons l'existence des poches cryptiques afin de démontrer leur stabilité au cours de la MD.

1.6 Objectifs de notre étude

Les objectifs de cette Thèse se basent sur la recherche des poches cryptiques au niveau de systèmes complexes par dynamique moléculaire polarisable. Nous chercherons ainsi à valider le choix de nos structures en évaluant leur stabilité au niveau des quadruplexes de guanine dans la région du proto-

oncogène c-kit et dans la région LTR-III du VIH-1 ainsi qu'au niveau de la zone de dimérisation de la protéase principale M^{pro} du SARS-CoV-2. Ces systèmes présentent des cibles très attrayantes pour la conception de petites molécules thérapeutiques. Durant cette Thèse nous avons choisi d'étudier les cibles suivantes: i) le proto-oncogène c-kit qui code pour le récepteur transmembranaire CD117/c-kit, et dont la sur-expression est associée à certains types de cancer, ii) le virus VIH responsable de la maladie du SIDA et qui est un problème majeur de santé publique a l'échelle mondiale, il a entraîné jusqu'ici près de 33 millions de décès et iii) le virus SARS-CoV-2 qui cause le COVID-19, maladie respiratoire responsable de la pandémie de COVID-19 ; on dénombre plus de 195 529 504 cas de coronavirus à travers le monde et 4 174 037 décès.

Afin d'aboutir à nos objectifs, nous profitons des progrès technologiques au sein de notre Laboratoire tant au niveau computationnel qu'au niveau des outils liés à l'analyse de l'exploration. Nous introduisons une méthode d'échantillonnage adaptatif non supervisé axée sur la densité basée sur des modèles statistiques et une analyse des composantes principales (PCA). Ceci a été déployé sur un supercalculateur généraliste. Le problème d'exploration globale étant décomposé en un ensemble de trajectoires MD distinctes, le processus peut être relancé à l'aide d'une méthode de sélection itérative, et divers calculs peuvent avoir lieu sur un grand nombre d'unités de traitement graphique (GPU) désormais disponibles dans les supercalculateurs généralistes. Une telle stratégie permet au logiciel Tinker-HP [114], dans une version très récemment opérationnelle sur cartes graphiques GPU [7], d'effectuer des simulations MD de plusieurs μs en quelques jours, là où des années auraient été nécessaires avec une seule carte GPU ou avec des simulations MD conventionnelles basées sur CPU.

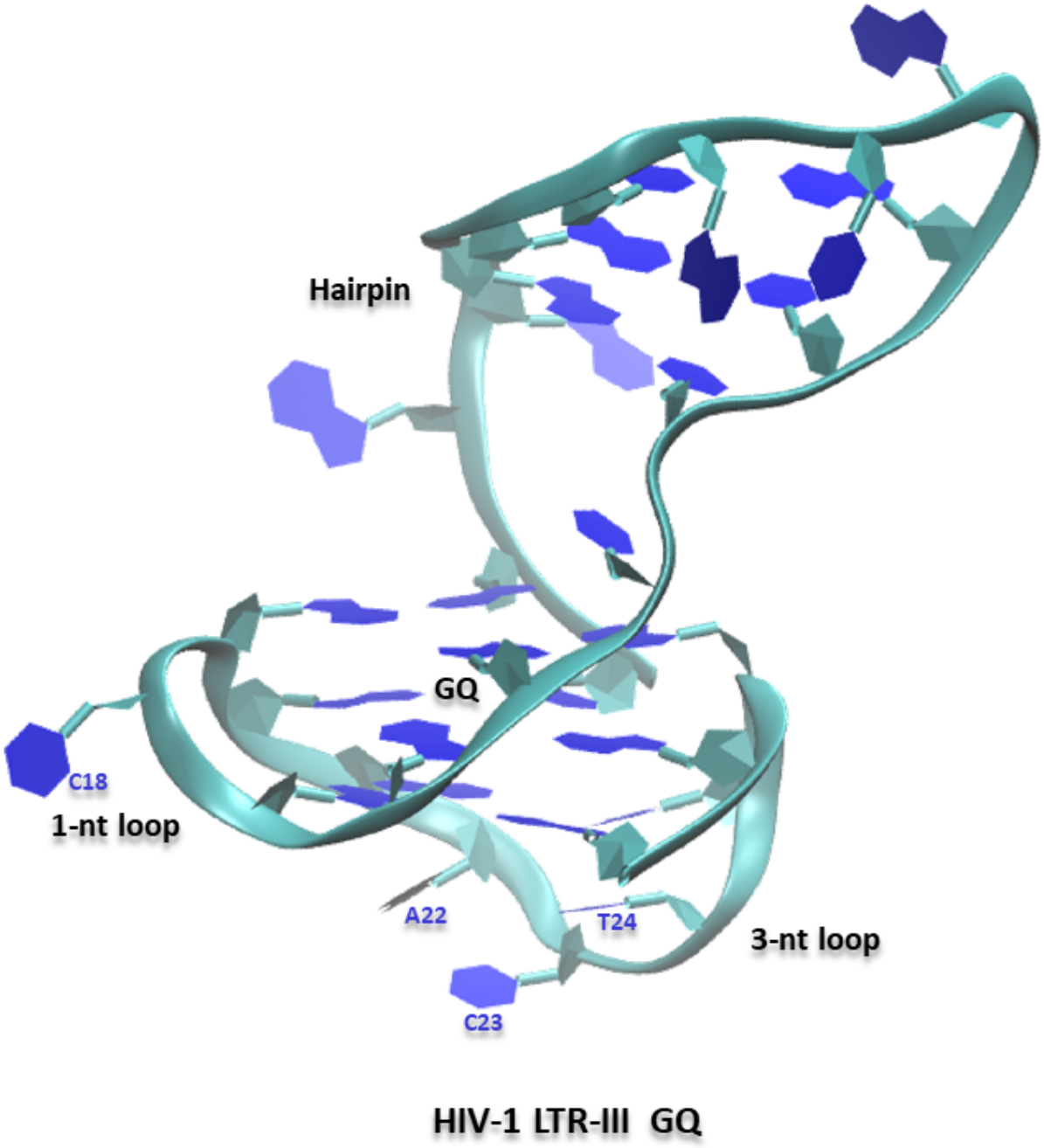


Figure 1.10: Représentation 3D du GQ au niveau du région LTR-III du VIH-1 sur VMD.

Chapitre 2

Modélisation à haute résolution de la dynamique structurelle de la Protéase Principale M^{pro} du SARS-CoV-2

2.1 Détails computationnels et protocole de simulation

Préparation des systèmes et choix des structures initiales.

Afin d'effectuer une simulation d'échantillonnage adaptatif non supervisé à grande échelle, les structures de départ doivent être sélectionnées à partir d'une simulation de MD conventionnelle issu d'un n–PFF ou d'un PFF. Nous avons choisi le jeu de données de RIKEN comme point de départ. De leur 10 μs de MD conventionnel (PDB: *6LU7*, pH=8) [98] issu du AMBER14ff n–PFF [134] et en utilisant PCA comme fil conducteur, nous avons soigneusement extrait 14 structures pertinentes qui représentent notre point de départ pour notre étude. Il convient de noter que la structure cristalline *6LU7* est une structure *holo* comprenant un inhibiteur lié de manière covalente. La structure *apo* non liée à l'inhibiteur a été initialement obtenue par RIKEN en supprimant l'inhibiteur et relâchée sur 10 μs de simulation (https://data.mendeley.com/datasets/vpps4vhryg/1). Chaque structure issue de Amber14ff a ensuite été minimisée avec le PFF AMOEBA [180, **?**, 172, 237] et un algorithme L-BFGS jusqu'à ce qu'une déviation de la valeur quadratique moyenne (RMS) de 1 kcal/mol sur le gradient soit atteinte. Il est important de noter que tous les résidus d'histidine ne sont pas protonés dans la structure RIKEN de la même manière que celle de DESRES. Comme il a été récemment démontré que le pKa le plus élevé pour les sites histidine de protonation possibles était plus faible dans le SARS-CoV-2 M^{pro} que dans le SARS-CoV-1, soit environ 6.6, [217] la présente simulation est donc cohérente avec les conditions de pH physiologiques (pH=7.4) [109].

Protocole de simulation

La simulation de tous les atomes a été réalisée à l'aide du module GPU nouvellement développé [7] dans la plateforme Tinker-HP [114], qui fait partie de la plateforme Tinker 8 [176]. Ce module récemment développé est capable d'exploiter efficacement la précision mixte [7] offrant une forte accélération des simulations utilisant des GPU. La structure initiale comportant 98 694 atomes qui forment le dimère de la M^{pro} entièrement solvaté a été extraite de la banque de données de protéines (PDB : *6LU7*). Le champ de force polarisable AMOEBA [180, 237] a été utilisé pour décrire le comportement de tous les atomes de la protéine de du solvant. Des conditions aux limites périodiques utilisant une boîte cubique d'une longueur de côté de 100 Å ont été utilisées. Les simulations de dynamique moléculaire de Langevin ont été réalisées à l'aide de l'intégrateur BAOAB–RESPA1 [115] en utilisant un pas de temps externe de 10 fs, un solveur de polarisation à gradient conjugué préconditionné (avec un seuil de convergence de 10^{-5}), repartitionnement de la masse des hydrogènes (HMR) et des vitesses initiales aléatoires. Les conditions aux limites périodiques (PBC) ont été utilisées en utilisant la méthode "Smooth Particle Mesh Ewald" (SPME) avec une grille de dimension 128 Å$\times$ 128 Å$\times$ 128 Å. Le seuil d'Ewald a été porté à 7 Å et le seuil de van der Waals à 9 Å.

Nous avons commencé la simulation en exécutant une cMD de 10 ns pour chacune des 14 structures représentatives de RIKEN. Une première sélection d'échantillonnage adaptatif est ensuite effectuée

sur ces structures initiales de 140 ns. Nous avons choisi d'utiliser les quatre premiers composants de la PCA comme espace conformationnel pour la méthode d'échantillonnage adaptatif. À chaque itération, la procédure d'échantillonnage adaptatif est ensuite utilisée sur ces quatre premières composantes PCA nouvellement calculées afin de sélectionner 100 structures. Ensuite, 100 simulations moléculaires indépendantes de 10 ns ont été réalisées dans l'ensemble NVT à 300K sur des cartes GPU NVIDIA V100 uniques. Chaque trajectoire appartenant à la même itération d'échantillonnage adaptatif a été exécutée simultanément sur le Supercalculateur HPE Jean Zay (IDRIS, GENCI, France).
La trajectoire complète de 15.14 μs avec et sans eau est accessible gratuitement via le Centre national suisse de calcul intensif (CSCS) [4] et a été liée au portail communautaire BioExcel/Molssi COVID–19.

Pour étudier l'interface de dimérisation, nous avons analysé en profondeur la conformation spatiale de tous les atomes de la M^{pro} produit durant les 15.14 μs de simulation en utilisant le champ de force polarisable AMOEBA (champ de force de la protéine AMOEBA [194, 237] et le modèle d'eau flexible AMOEBA03 [179]) ainsi que celui fourni par le groupe RIKEN [110] (champ de force AMBER ff14SB [134], modèle d'eau TIP3P [102]) et DESRES [57] (champ de force DES-AMBER [169], modèle d'eau TIP4P-D [168]). En suivant le même protocole de simulation, nous avons effectué des séries supplémentaires séparées de simulations adaptatives pour un total de 12 μs avec AMOEBA pour simuler à des pH plus bas. Dans ce cas, une protonation supplémentaire des résidus d'histidine s'est produite. Par conséquent, pour produire des données supplémentaires aux simulations pH=7.4 et pH=6 proposées dans notre jeu de données précédent [93], nous avons également successivement protoné (2 X 6μs exécutions) les deux résidus **His163** sur les deux chaînes de la protéine pour simuler un abaissement supplémentaire du pH (voir la discussion et le Tableau 2 dans la référence [210]). D'autres simulations de dynamique moléculaire conventionnelle AMOEBA et AMBER99SB à 800 ns (400 ns X2) ont été produites à un pH physiologique et en repartant des points de départ de notre ensemble de données précédent, en sauvegardant les structures toutes les 10 ps pour permettre une analyse approfondie du rôle du solvant : l'eau. Toutes les simulations supplémentaires ont été effectuées à l'aide du module GPU nouvellement développé [7] dans la plateforme Tinker-HP, [114] qui fait partie de la plateforme Tinker 8 [176]. Ce module récemment développé est capable d'exploiter efficacement la précision mixte [7], offrant une forte accélération des simulations utilisant des cartes GPU. Des conditions aux limites périodiques utilisant une boîte cubique d'une longueur de côté de 100 Åont été utilisées. Les simulations de dynamique moléculaire de Langevin ont été réalisées à l'aide de l'intégrateur BAOAB–RESPA1 [115] en utilisant un pas de temps externe de 10 fs, un solveur de polarisation à gradient conjugué préconditionné (avec un seuil de convergence de 10^{-5}), hydrogène–masse repartitionnement (HMR) et vitesses initiales aléatoires. Les conditions aux limites périodiques (PBC) ont été utilisées en utilisant la méthode Smooth Particle Mesh Ewald (SPME) avec une grille de dimension 128 Å$\times$ 128 Å$\times$ 128 Å. Le seuil d'Ewald a été porté à 7 Ået le seuil de van der Waals à 9 Å. L'analyse de post-traitement a été effectuée à l'aide de MDTraj [140], Scikit-Learn [165] et Scipy package [219]. Les matrices de corrélation croisée dynamique - DCCM, ont été générés sur la base de

l'atome C_α de chaque résidu en utilisant la fonctionnalité fournie dans le package MD-TASK [34].

2.2 Résultats et discussion

2.2.1 Performance de l'exploration par échantillonnage adaptatif : comparaisons avec d'autres simulations disponibles

Comme nous l'avons mentionné, nous avons utilisé la PCA [6] comme quantité intermédiaire pour orienter l'itération d'échantillonnage consécutive. Cependant, c'est aussi une bonne quantité pour évaluer rapidement les performances du schéma d'échantillonnage adaptatif pour l'exploration de l'espace conformationnel. En effet, l'analyse des trajectoires MD avec la PCA est une stratégie bien connue dans la communauté sous le nom de "dynamique essentielle" [12, 11, 24]. La PCA étant un algorithme de réduction de dimensionnalité qui évalue les directions maximisant la variance de l'ensemble de données, elle est ainsi révélatrice d'une diversité conformationnelle du système. Par conséquent, elle peut être considérée comme un moyen d'évaluer la quantité d'échantillonnage et peut également détecter des "mouvements essentiels" explicites, non discernables autrement à l'aide de variables collectives prédéfinies. Ainsi, il est intéressant de comparer la quantité d'échantillonnage sur l'espace de ces variables réduites. C'est pourquoi nous avons projeté la trajectoire de RIKEN, et de DESRES et les premiers jeu de données Tinker-HP (2 μs) sur les deux premières composantes de la PCA des 2 premières μs du jeu de données Tinker-HP (Figure 2.1, **a)** et **b)**). On peut voir que, dans cet espace, le schéma adaptatif Tinker-HP a déjà capturé les principales caractéristiques des PCA de RIKEN et DESRES. Il apparaît également que la trajectoire RIKEN a échantillonné une portion de l'espace conformationnel proche du jeu de données Tinker-HP alors que la trajectoire DESRES semble explorer uniquement la zone la plus échantillonnée par Tinker-HP. La même procédure a été appliquée pour les composants PCA et les données associées de l'ensemble de données Tinker-HP (Figure 2.1 **c)**) et il est frappant de constater qu'une portion beaucoup plus grande de l'espace conformationnel a été échantillonnée par notre schéma adaptatif.

En guise de conclusion préliminaire, nous pouvons dire que notre stratégie d'échantillonnage adaptatif nous a permis de générer une simulation MD polarisable de plusieurs microsecondes qui a échantillonné une vaste zone du paysage de l'énergie libre. De plus, nous avons analysé la déviation quadratique moyenne (RMSD) sur les squelettes protéiques par rapport au rayon de giration (voir la Figure 2.2) pour la trajectoire AMOEBA de 15.14 μs. Ceci a révélé d'importants changements de conformation. Les variations du rayon de giration sont d'environ de 2 Å, alors qu'il varie de 1 Å pour les MD conventionnels non polarisables.

De tels tracés sont très utiles pour comprendre une question clé : qu'est-ce qui rend les résultats AMOEBA différents ? Est-ce le choix d'un PFF versus un n–PFF ou est-ce le choix de la stratégie d'échantillonnage adaptatif ?

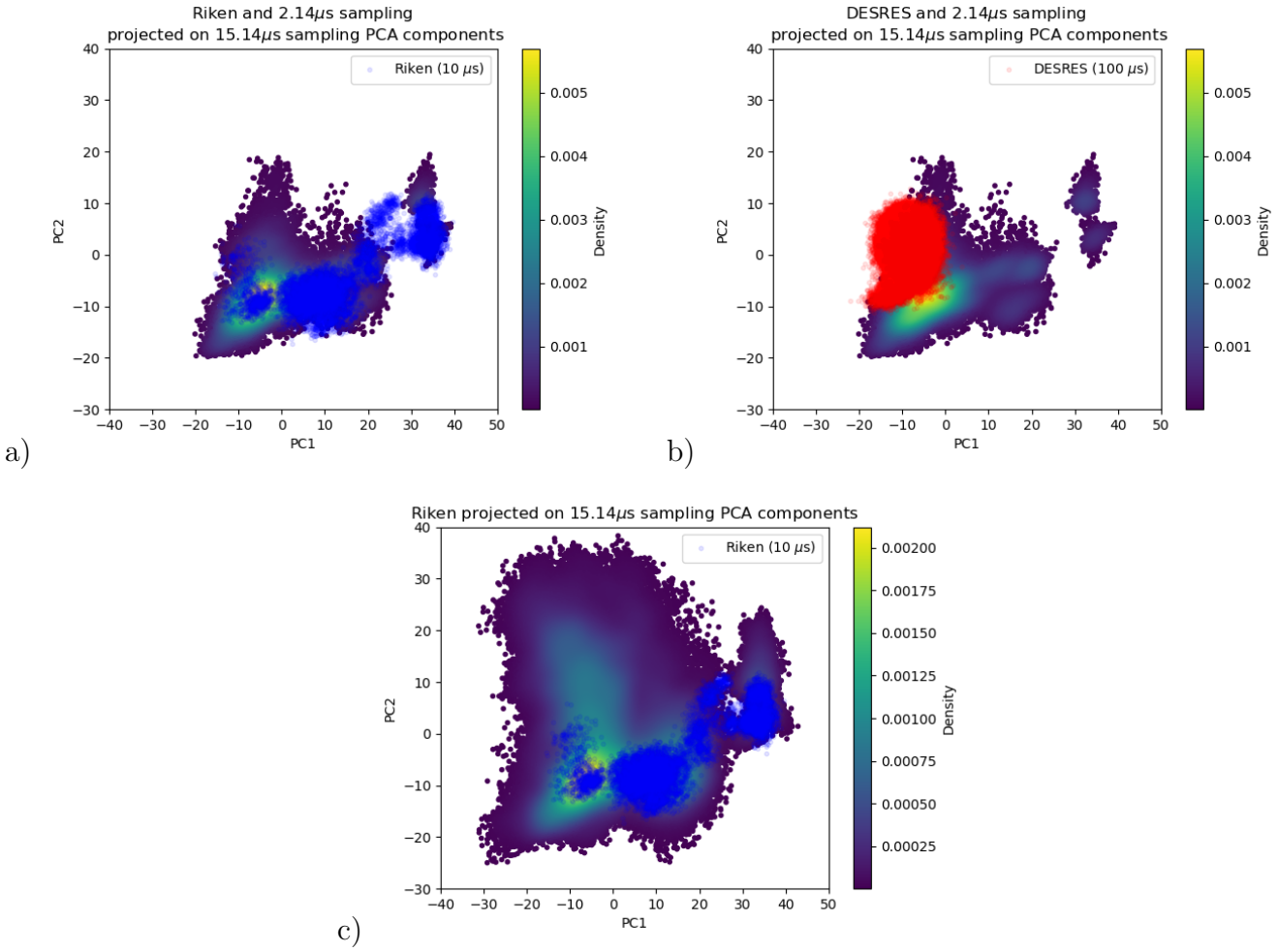


Figure 2.1: Jeux de données RIKEN et DESRES superposés au squelette protéique *6LU7* et projetés sur les deux premiers composants PCA adaptés respectivement aux 2μs (**a**), **b**)) et 15.14 μs (**c**)) de la simulation.

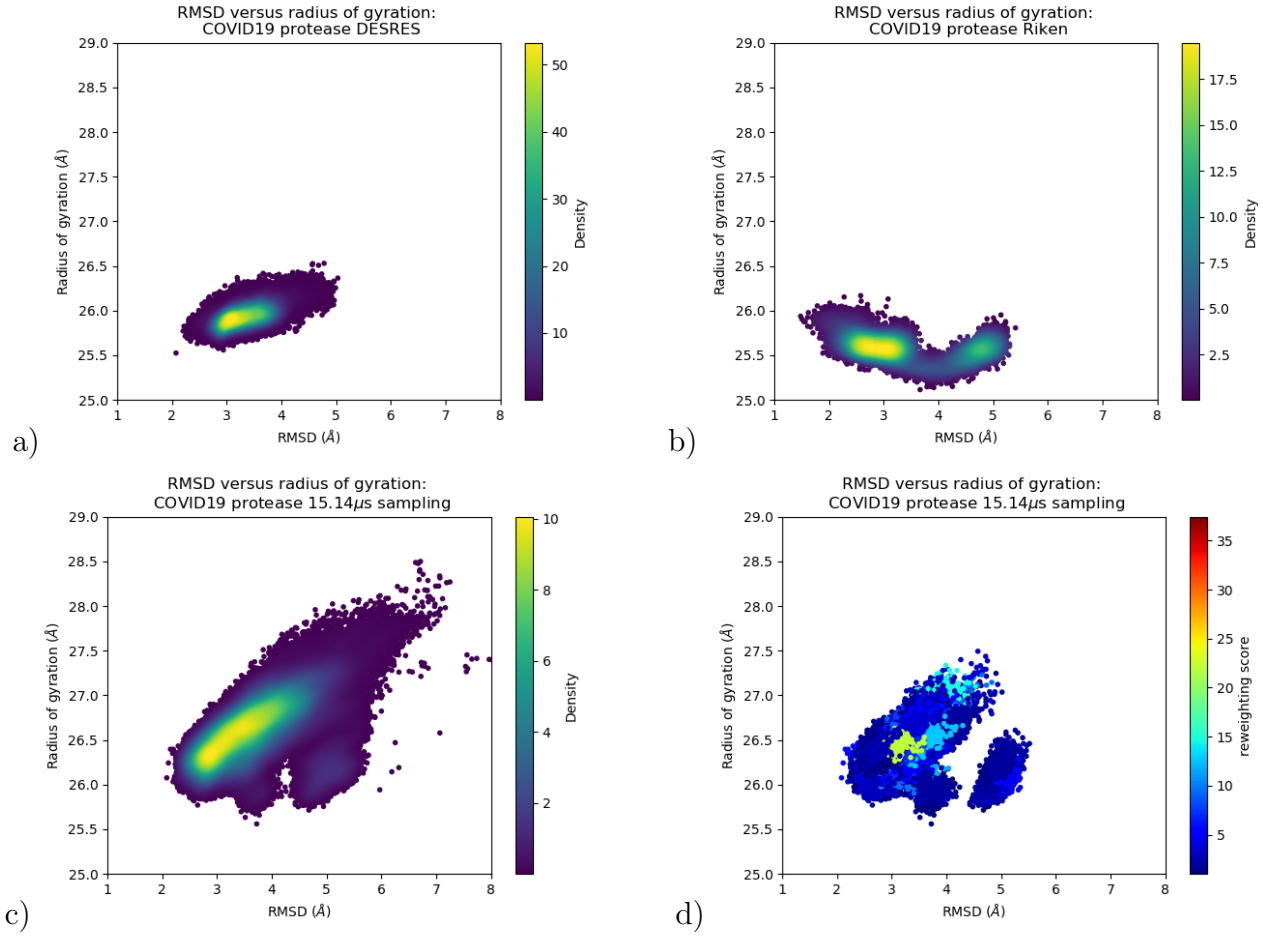
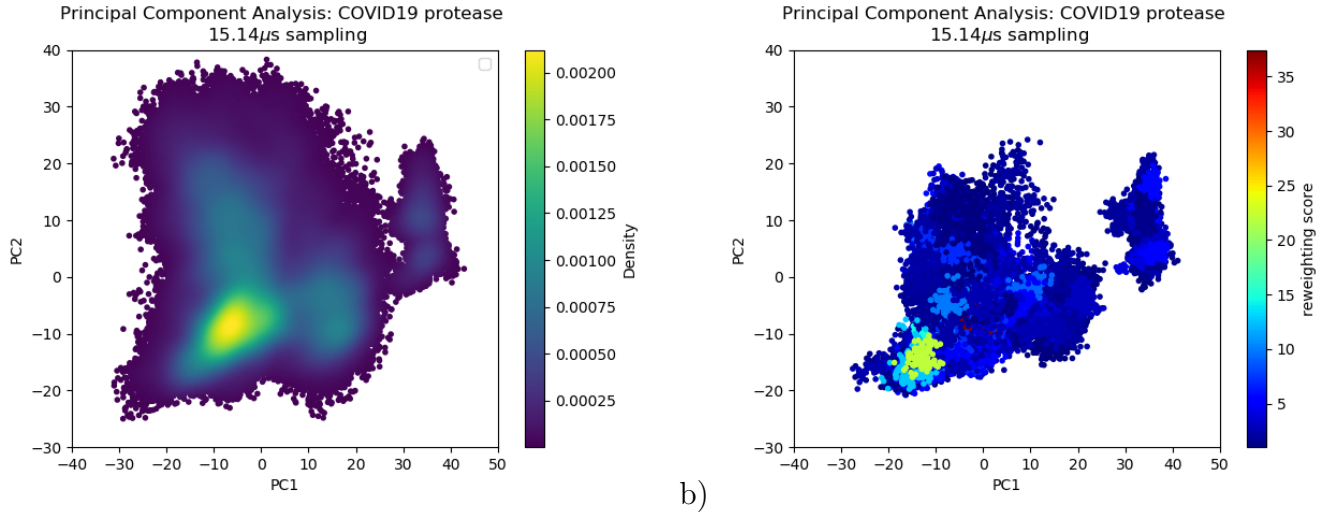


Figure 2.2: Déviation quadratique moyenne (RMSD) des atomes du squelette protéique par rapport au rayon de giration (Å) des ensembles de données DESRES **a)**, Riken **b)** et AMOEBA 15.14μs **c)**. Dans la Figure **d)**, nous avons coloré la simulation 15.14 μs en fonction de leur score de repondération. Nous ne montrons que les structures qui ont un score supérieur à 1 qui correspondent à des états probables très probables. De plus, cela réduit l'ensemble de données de 151400 structures à 33100, soulignant la grande exploration de surface d'énergie libre de notre simulation. Une telle Figure peut être comparée directement à celle de RIKEN pour évaluer l'effet des différents champs de force (voir texte)

Afin de fournir une comparaison juste et quantitative entre les FF et de décorréler les effets des FF eux-mêmes des gains dus à l'échantillonnage adaptatif, nous nous limitons aux structures avec un score de repondération supérieur à 1 car c'est le score des structures visitées lors d'une MD classique puisque les structures de score inférieur à 1 sont celles qui ont été privilégiées par l'algorithme adaptatif pour maximiser l'exploration. 3/4 des points sont donc supprimés selon ce critère offrant une vision de la performance de l'échantillonnage adaptatif. Le graphe représentant les points restants est présenté sur la Figure 2.3 pour AMOEBA et on peut directement le comparer au tracé de RIKEN par exemple. Il existe clairement des différences entre les résultats AMBER et AMOEBA, et elles proviennent également du choix du FF.



a) b)

Figure 2.3: Graphes 2D représentant la projection de la simulation AMOEBA 15.14μs montrant a) toutes les structures générées et b) seulement les structures ayant un score ≥ 1 colorées selon leur score de repondération. Nous remarquons que les scores de débiaisage peuvent être utilisés comme une carte exploratoire du paysage de l'énergie libre et de l'algorithme d'échantillonnage adaptatif. En effet, les scores faibles correspondent à des configurations atypiques qui n'auraient pas été observées par la MD classique. Cette Figure révèle de vastes zones de configurations à faible score, ce qui montre que la sélection adaptative a eu un effet important sur l'exploration des événements rares.

2.2.2 Regroupement et extraction non supervisés de l'énergie libre relative débiaisée entre les domaines représentatifs

Si l'analyse PCA révèle des informations utiles, un regroupement approprié des ensembles produits est un cadre plus précis et quantitatif pour discuter des différences entre les simulations et les nouvelles caractéristiques possibles capturées par le champ de force AMOEBA. Par conséquent, nous avons appliqué à toutes les trajectoires la méthode de la clusterisation spatial basée sur la densité des applications avec bruit (DBSCAN) "Density-based spatial clustering of applications with noise" [64].

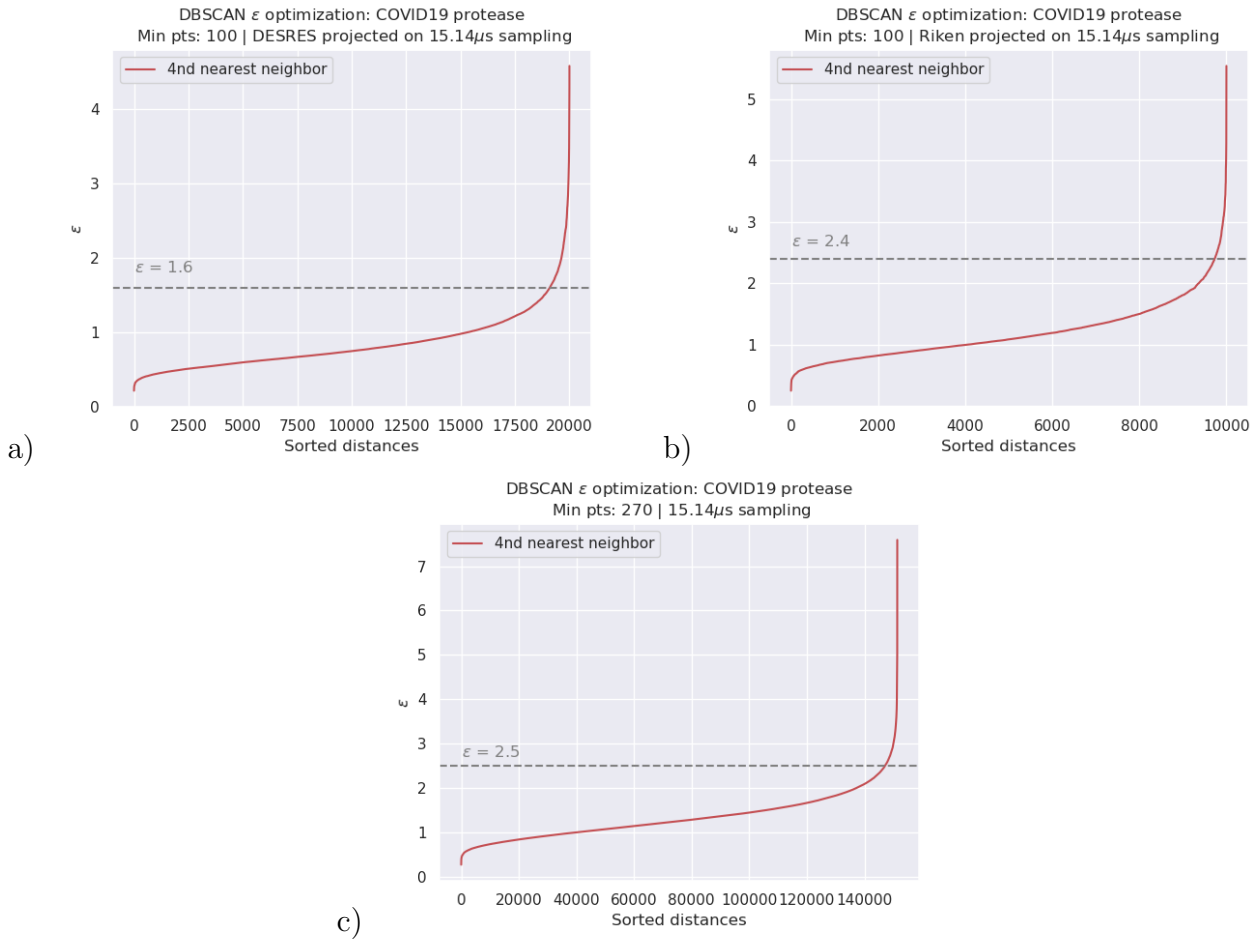


Figure 2.4: Paramétrage basé sur DBSCAN de **a)** DESRES (100μs), évalué toutes les 4 structures, **b)** jeux de données Riken (10μs) et **c)** la simulation Tinker-HP 15.14 μs.

DBSCAN est un algorithme d'apprentissage automatique non supervisé qui regroupe les données en clusters en fonction de leur densité. Cet algorithme est particulièrement bien adapté dans notre cas car il est spécialement conçu pour cibler des clusters de forme arbitraire. Pour évaluer la densité, DBSCAN utilise deux paramètres, ϵ la distance pour laquelle deux points sont considérés comme voisins et $MinPts$ le nombre minimum de points pour définir un cluster. ϵ a été choisi en utilisant la procédure du graphe du plus proche voisin, c'est-à-dire en traçant la distance au n-voisin le plus proche pour chaque point, ordonné de la plus grande à la plus petite valeur, et en évaluant ϵ pour lequel le graphe commence un coude. Pour un ϵ donné, nous avons ensuite scanné différentes valeurs de $MinPts$ jusqu'à ce que des clusters relativement grands couvrant une large plage de l'espace soient trouvés. En pratique, nous avons évalué la distance au 4ième voisin le plus proche sur les 4-Dimensions composées des quatre premières composantes principales de la PCA pour 15.14 μs, les résultats sont présentés dans la Figure 2.4. Pour DESRES et RIKEN, après avoir été alignées sur leur PDB respective, les structures ont été projetées sur cet espace 4D.

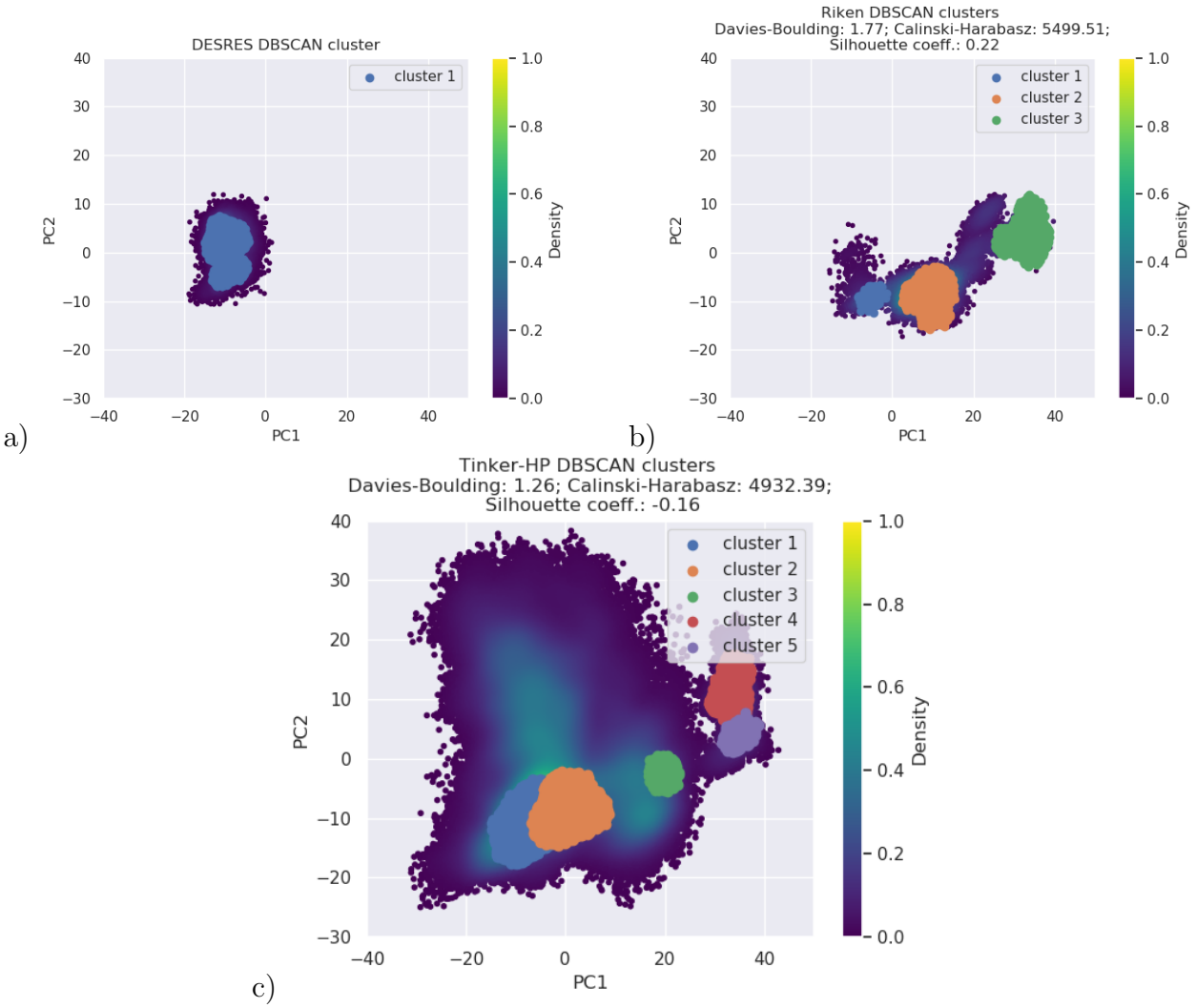


Figure 2.5: Regroupement par DBSCAN des ensembles de données **a)** DESRES (100μs), **b)** RIKEN (10μs) et **c)** Tinker-HP (15μs). Nous détectons respectivement 1, 3 et 5 clusters dans chacune des simulations.

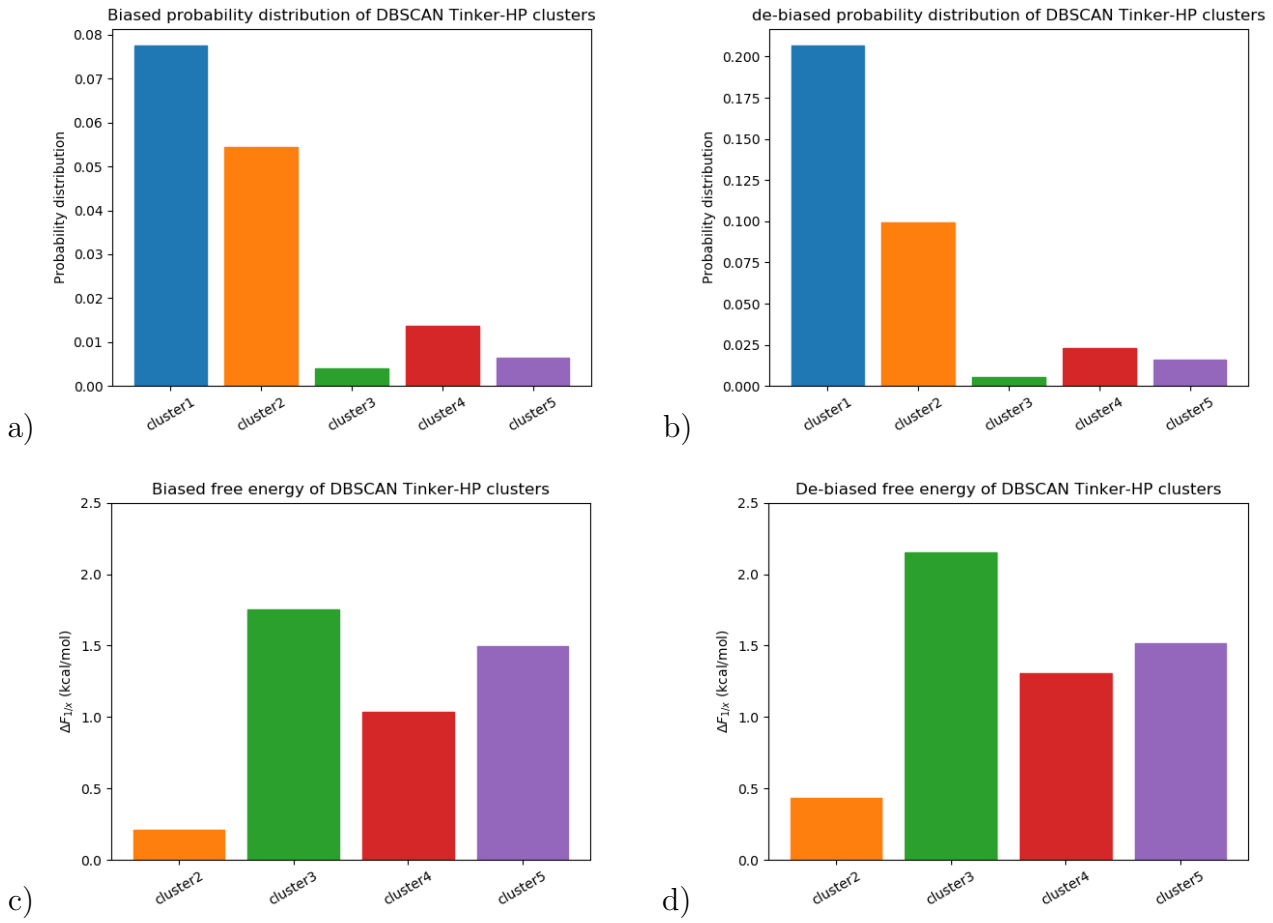


Figure 2.6: Distribution de probabilité des clusters biaisée **a)** et débiaisée **b)**. Énergies libres relatives des clusters DBSCAN Tinker-HP 15.14μs, par rapport au cluster 1, biaisé **c)** et débiaisé **d)**.

Notre choix d'utiliser la PCA comme point de départ de la clusterisation est motivé par la diversité conformationnelle apportée par le couplage du PFF et du schéma d'échantillonnage adaptatif. Pour la visualisation, les clusters sont ensuite projetés sur les deux premiers composants principaux (Figure 2.5). Pour évaluer la qualité de la clusterisation, nous avons utilisé trois méthodes de notation pour les données étiquetées inconnues [128] : le coefficient Silhouette, les indices de Calinski–Harabasz et de Davies–Bouldin. Ces indices ont confirmé notre procédure d'optimisation des paramètres et la haute qualité de la clusterisation. Notre nouveau schéma d'échantillonnage adaptatif a le principal avantage d'offrir l'accès à de véritables propriétés statistiques telles que les énergies libres.

Pour comprendre la stabilité des clusters, les énergies libres pour chaque cluster sont calculées (Figure 2.6 **c)**, **d)**) à travers l'évaluation de la distribution de probabilité sur le nombre total de structures. La distribution de probabilité débiaisée (Figure 2.6 **a)**, **b)**) est estimée avec la procédure de dé-biaisage expliquée dans la section précédente. L'étape de dé-biaisage préserve la tendance entre les clusters mais augmente les probabilités. Cela signifie que les cinq grappes ont été désavantagées par l'échantillonnage adaptatif. Par exemple, la simulation biaisée a évalué 8% de probabilité de présence

pour le cluster 1, qui aurait dû contenir, dans une simulation débiaisée, 20% des configurations. Par ailleurs, le cluster 1 est bien la région la plus explorée à la fois par DESRES et RIKEN. Ainsi, l'algorithme a réussi à désavantager cette partie de l'espace conformationnel ce à quoi on pouvait s'attendre car il a favorisé les zones de transition intermédiaires au détriment des régions denses afin de découvrir de nouvelles régions.

2.3 Corrélation avec les données expérimentales : marqueurs structurels de l'activité des protomères et nouvelles fonctionnalités

2.3.1 Marqueurs de la structuration du trou oxyanionique

Pour assurer la validité de nos simulations AMOEBA, nous avons comparé nos propriétés calculées avec les données expérimentales disponibles. Depuis le début de la pandémie de COVID-19, diverses structures aux rayons X ont été publiées (PDB: *6Y84, 6LU7, 6Y2G, ...*) [5, 98, 240]. Ils ont fourni des informations importantes sur les interactions spécifiques entre les résidus ainsi que des informations structurelles sur le site actif. Pour être cohérent avec les simulations RIKEN nous avons utilisé comme référence le même PDB : *6LU7* [98]. Notez que DESRES a utilisé un autre PDB, 6Y84, [5] que nous avons utilisé comme référence dans le calcul de ses propriétés. Des structures cristallines ont été projetées sur les deux premiers composants PCA des simulations Tinker-HP (voir Figure 2.7). Récemment, Zhou *et coll.* ont publié une étude expérimentale de la structure *apo* (PDB 1UJ1) [243] à pH physiologique. Ils ont trouvé plusieurs caractéristiques permettant de caractériser la présence de la structure du trou oxyanionique qui est un élément structurel clé de l'activité de chaque protomère. En particulier, ils ont proposé de surveiller la distance entre **Glu166** et **His172** et un empilement π–π entre **Phe140** et **His163**. La définition de ces marqueurs structurels n'est pas nouvelle et a été initialement discutée pour la M^{pro} du SARS-CoV-1 [230, 210]. Le trou oxyanionique est responsable de la stabilisation du substrat dans le site actif et est d'une importance cruciale pour la cinétique et l'activité de l'enzyme. En effet, le site de fixation du substrat est composé de 4 poches notées S1 à S4 avec la poche S1 impliquant des résidus très conservés tels que **Glu166, His172, His163 et Phe140**. Le trou oxyanionique de la protéase à cystéine englobe les amides du squelette (**Gly143, Ser144, Cys145**) tandis que les résidus 138 à 145 forment la boucle de liaison oxyanionique [231, 210, 240]. L'existence de ce dernier est responsable en partie de la structuration de la poche S1 [210]. Lorsque l'empilement et l'interaction **Glu166–His172** sont rompus, un ré-arrangement se produit conduisant éventuellement à l'effondrement du trou oxyanionique. Dans ce cas, **Glu166** interagit potentiellement avec **His163** au lieu de **His172**. En d'autres termes, de fortes interactions de **Glu166** avec **His172** associées à un empilement **Phe140-His163** sont cohérentes avec un trou oxyanionique structuré, et peuvent être utilisées comme marqueur de l'activation du protomère enzymatique. À l'inverse, une

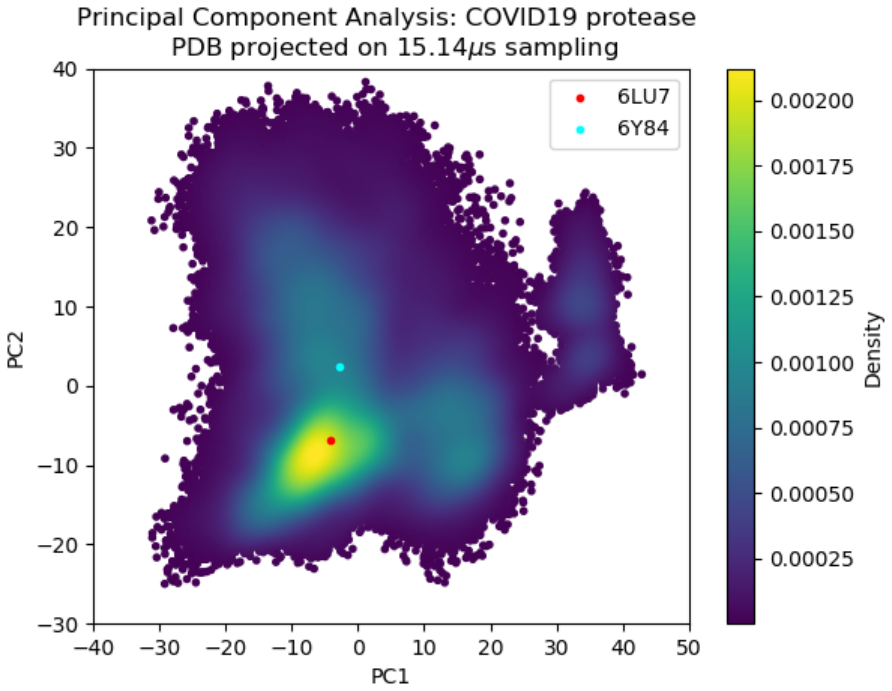


Figure 2.7: Projection des PDB *6LU7* et *6Y82* sur les deux premiers composants PCA ajustés à 15.14 μs de la simulation.

forte interaction de **Glu166** avec **His163** serait plutôt un marqueur de l'inactivation du protomère en lien avec un effondrement du substrat poche de liaison au substrat S1. Bien entendu, une telle analyse n'est qu'interprétative, la structuration des trous oxyanioniques étant bien plus complexe. Cependant, son utilité a été démontrée depuis les premières études sur la protéase principale du SARS-CoV-1 [210]. En pratique, l'absence d'un trou oxyanionique bien structuré conduit à l'inhibition de l'activité enzymatique. Expérimentalement, on sait que la forme monomérique de la M^{pro} est inactive alors que la forme active est un homodimère contenant deux protomères [241]. Dans l'état *holo* du SARS-CoV-1, le premier protomère est actif tandis que le second est trouvé inactif [43]. Pour le SARS-CoV-2, une structure cristalline pH=6 (PDB : 1UJ1), [241] a prédit une forte asymétrie des protomères avec une conformation inactive pour l'un des protomères liée à une interaction **Glu166** et **His172** rompue. Cependant, l'inactivité de l'un des protomères reste une hypothèse car les études cristallographiques du dimère rencontrent des difficultés pour capturer les détails de chaque protomère individuel. En effet, les données ne sont disponibles que sur l'un des protomères de l'unité asymétrique ce qui conduit toujours à la conformation la plus ordonnée et donc la plus active. Concernant l'état *apo*, des résultats expérimentaux récents conduisent à une faible activité potentielle du dimère *apo* en lien avec un trou oxyanionique déstructuré observé [243]. Il est important de souligner que les distances/marqueurs présentent une distribution de différentes valeurs centrée autour d'un maximum de fréquence en raison des conditions de liquide qui diffèrent de celles du cristal. Dès lors, nous avons étudié ces marqueurs. Pour étudier l'interaction d'empilement **Phe140–His163**, nous utilisons un

indice d'empilement développé par Branduardi et Parinello [32] qui l'ont décrit comme un produit de 2 fonctions de Fermi, l'une considérant la dépendance radiale, l'autre la dépendance angulaire de l'interaction. Le modèle fournit un indice allant de 0 pour une interaction non empilée à 0.6 pour une interaction parfaite. Les interactions **Glu166** et l'empilement π–π ont ainsi été calculés pour les deux chaînes de toutes les structures RIKEN, DESRES et Tinker-HP puis classés en histogrammes. Enfin, chaque histogramme a été dé-biaisé (c'est-à-dire repondéré) et extrapolé à l'aide d'un estimateur de densité à noyau univarié. Les résultats finaux sont donnés sur les Figures 2.8 et 2.9.

De plus, une simulation d'échantillonnage adaptatif de 6 μs a été réalisée (sur la machine Irene Joliot Curie (TGCC, GENCI, France)) sur le monomère (PDB: *6LU7*) et les mêmes caractéristiques que celles décrites ci-dessous (π- -π empilement entre **Phe140** et **His163**, interactions **Glu166** avec **His172** et **His163**) ont été calculés. Le monomère étant connu pour être dans une conformation inactive, cela nous aide à rationaliser le comportement observé dans nos simulations. Les résultats sont représentés sur les Figures 2.10 et 2.11.

Les protocoles de préparation et de simulation sont similaires à ce que nous avons fait pour le dimère. Par conséquent, **His172** et **His163** sont également non protonés, nous avons minimisé la structure jusqu'à un RMS du gradient de 1 kcal/mol et généré un cMD initial de 200 ns. Nous avons ensuite sélectionné 100 structures initiales aléatoires selon le protocole d'échantillonnage adaptatif de sélection de structures à l'aide de l'ACP, et nous avons effectué 6 itérations de 1 μs pour un temps de simulation total de 6 μs.

Pour l'interaction formée par Glu166, dans le cas de Tinker-HP, nous avons observé une asymétrie entre les deux protomères. Dans un protomère, le **Glu166–His172** est significativement plus faible que dans l'autre présentant un marqueur bien défini d'une plus petite activité du protomère. Cette non-interaction relative est en accord avec les résultats obtenus sur le monomère qui apparaît très similaire. La situation est plus complexe dans l'autre protomère où l'on observe une oscillation entre deux états, présentant soit une interaction **Glu166–His172** formée, soit son absence conduisant à seulement quelques marqueurs d'activité partiels. Cependant, l'état "en interaction" domine clairement les statistCes résultats démontrent que le trou d'oxyanion n'est que partiellement organisé dans l'autre protomère. Il est cohérent avec les données expérimentales sur l'état *apo* [243] et aussi avec les données sur le protomère actif de l'état *holo* qui montre des distances autour de 5 Å (voir la référence [217] et les références ici pour une discussion des différentes structures cristallines disponibles). Il ne s'agit bien sûr que d'un seul marqueur mais cela pourrait déjà corroborer l'asymétrie observée dans l'état *holo* où un seul protomère s'avère être actif [240], une caractéristique similaire qui avait été précédemment observée dans le SARS-CoV-1 [43]. Sur la base de l'analyse de ce marqueur unique, nous avons tendance à avoir un premier protomère inactif couplé à un deuxième protomère qui présente des caractéristiques d'activité partielles mais claires (deux états) par rapport à son homologue inactif et au monomère. Des interprétations similaires peuvent être déduites des simulations DESRES et RIKEN malgré une image moins claire des interactions **His172–Glu166** qui apparaissent extrêmement flexibles avec des

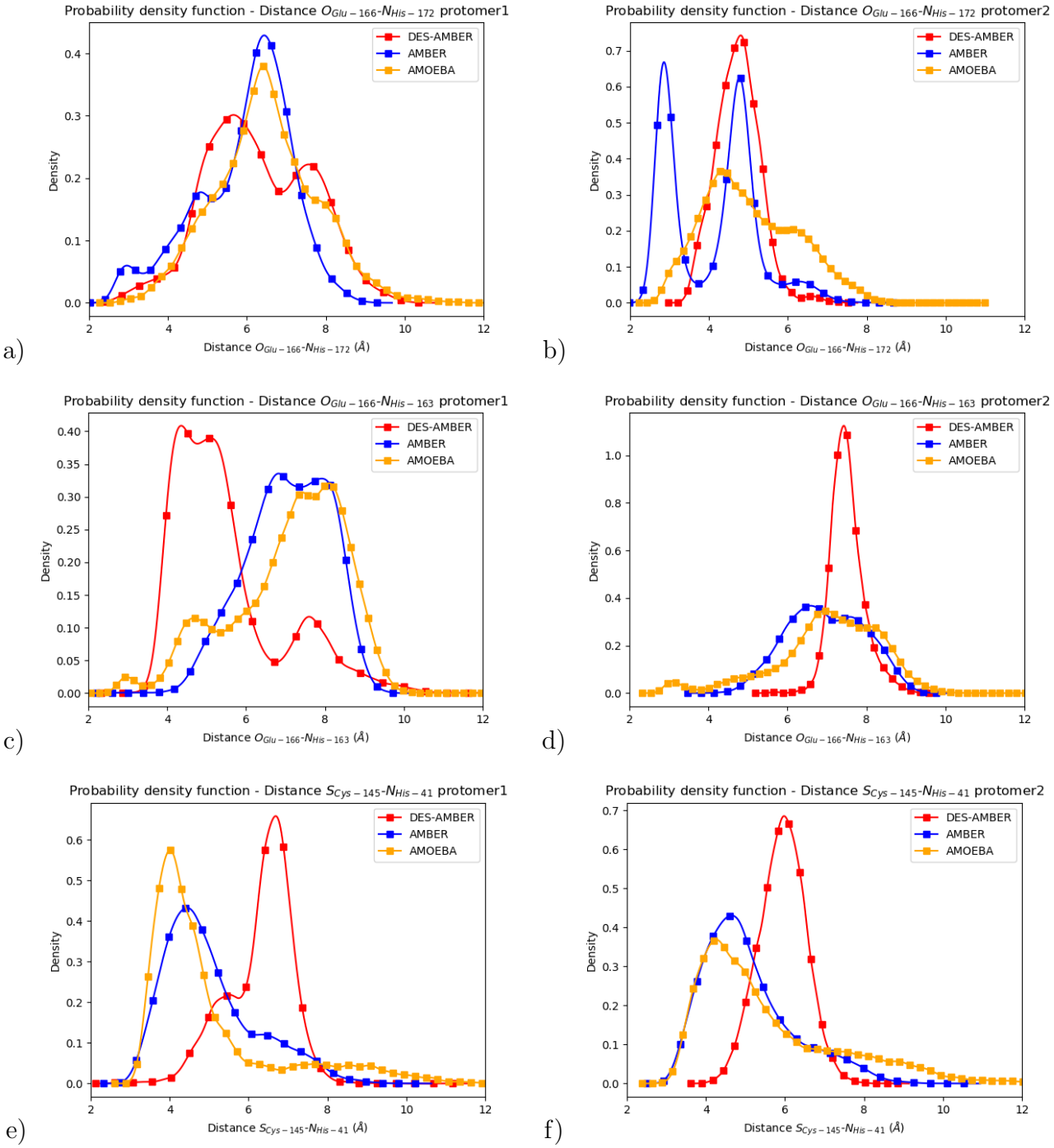
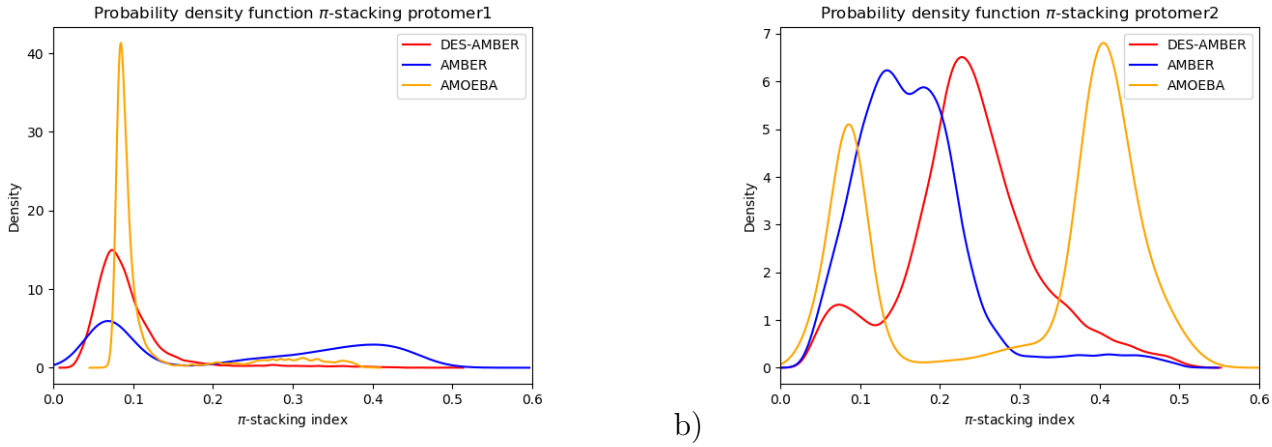


Figure 2.8: Représentation de plusieurs interactions d'intérêt de **Glu166** (avec **His163** et **His172**), au sein du cadran catalytique. Chaque calcul a été effectué sur les deux chaînes du système pour chaque simulation. Toutes les simulations sont réalisées à pH=7,4 (résidus d'histidine neutres)



a) b)

Figure 2.9: Représentation de l'empilement π–π entre **Phe140** et **His163**. Chaque calcul a été effectué sur les deux chaînes du système pour chaque simulation. Toutes les simulations réalisées à pH=7,4 (résidus d'histidine neutres)

états plus mixtes, en particulier pour AMBER. Ce n'est pas surprenant car les interactions **Glu–His** peuvent être classées comme des liaisons H, une classe d'interactions faibles directionnelles qui sont connues pour être difficiles à modéliser en utilisant des n–PFF [78, 163] car la polarisabilité contribue de manière significative à la précision des simulations de structures stabilisées par des liaisons hydrogène [142, 119]. Cependant, une seule distance ne suffit pas pour conclure et doit être combinée avec d'autres marqueurs comme la distance **Glu166–His163**. On note ici une plus forte asymétrie de telles distances dans les protomères pour DESRES alors que dans le cas de RIKEN et Tinker-HP on a pu à nouveau observer un mélange entre états interagissants / non interagissants. Cependant, ce deuxième marqueur doit être soigneusement considéré comme une comparaison directe avec notre simulation de monomère (voir Figures 2.10 et 2.11) qui montre que ce critère de distance est moins bien défini pour discuter de l'activité du protomère que de la distance **Glu166–His172**. Notre monomère étant connu pour être inactif, on pourrait en déduire que ce marqueur doit toujours être associé à l'évaluation de la distance **Glu166–His172**. En pratique, il faut considérer la force relative de ces interactions et la distance **Glu166–His163** apparaît ici nettement plus longue que celle entre **Glu166–His172**. Les distances **Glu166–His163** semblent cohérentes avec les données sur le protomère actif de l'état *holo* qui montre des distances allant au-delà de 6–8 Å (voir la référence [217] et les références marquées pour une discussion des différentes structures cristallines disponibles). À ce propos, une meilleure conservation du cadran catalytique est observée dans les simulations RIKEN et Tinker-HP avec une distance **Cys145—His41** plus petite par rapport à DESRES (voir Figure 2.8). Le site actif de la protéase principale, M^{pro}, comprend une dyade catalytique composée des résidus **Cys145** et **His41**. Les structures cristallines aux rayons X du SARS-CoV-1 [210, 231] ont trouvé une distance **Cys145—His41** entre 3–3.9 Å. En comparaison, nos simulations ont révélé des distances autour de 4 Å alors que AMBER et DES–AMBER sont respectivement autour de 4.5 et 6-7 Å. En

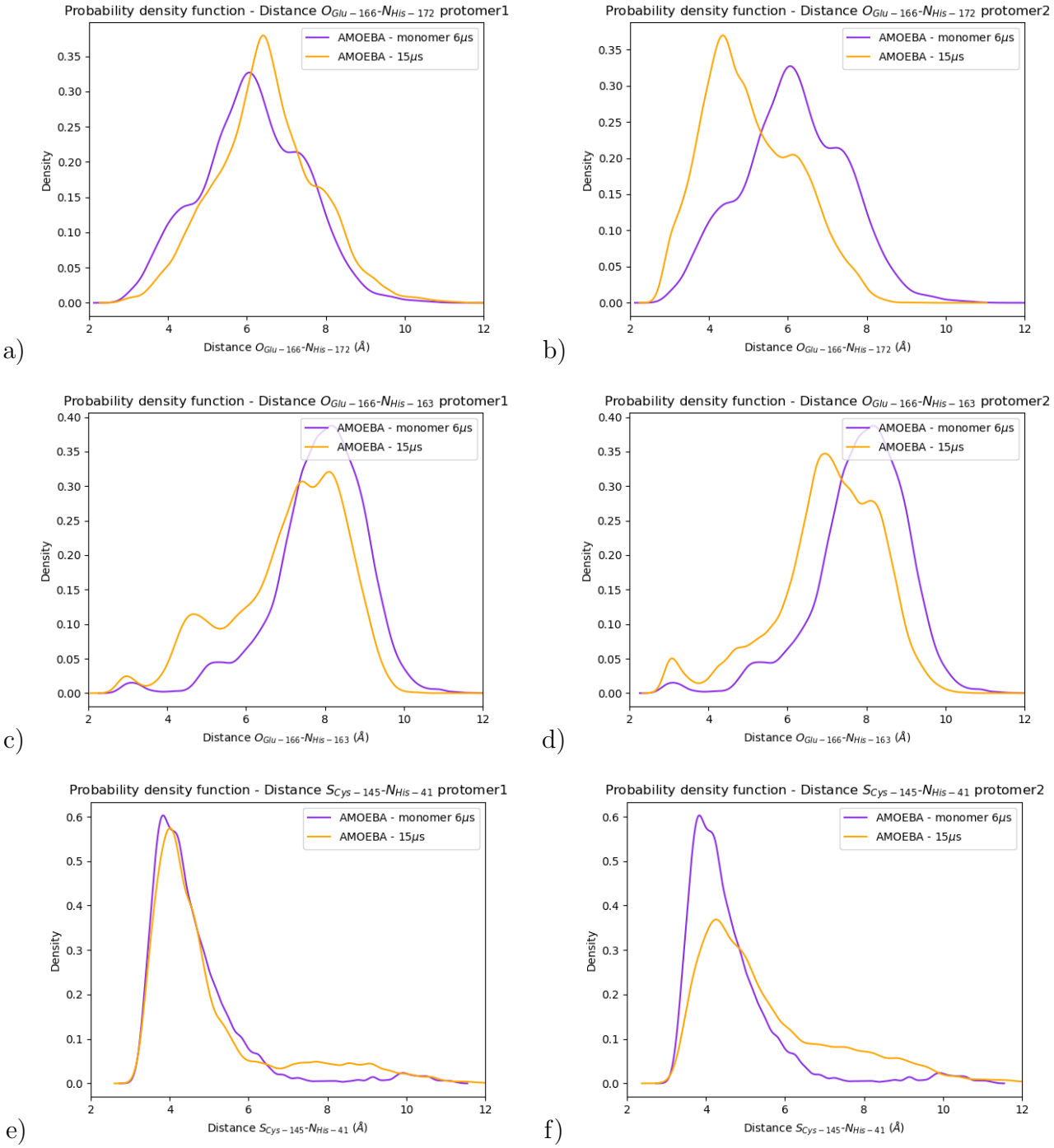
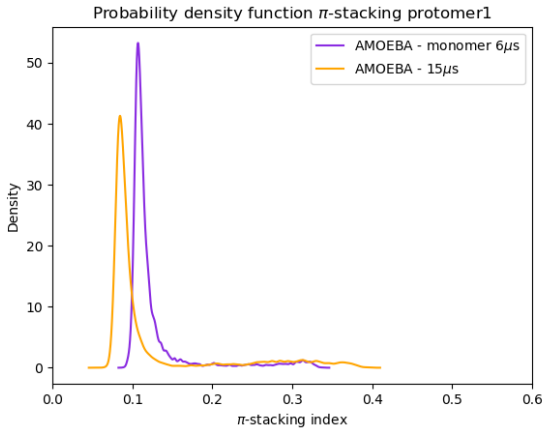
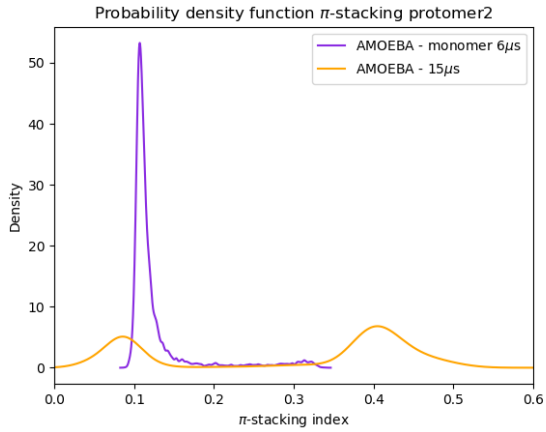


Figure 2.10: Représentation de plusieurs interactions d'intérêt de **Glu166** (avec **His163** et **His172**), au sein du cadran catalytique. Chaque calcul a été effectué sur les deux chaînes du système pour la simulation du dimère 15.14 μs et comparé à la simulation 6μs du monomère isolé (toutes les simulations ont été effectuées à pH=7.4 avec des résidus d'histidine neutres).



a)



b)

Figure 2.11: Représentation de l'empilement π–π entre **Phe140** et **His163**. Chaque calcul a été effectué sur les deux chaînes du système pour la simulation du dimère 15.14 μs et comparé à la simulation 6μs du monomère isolé (toutes les simulations ont été effectuées à pH=7,4 avec des résidus d'histidine neutres).

ce qui concerne les petites différences relatives entre les protéases principales du SARS-CoV-1 et du SARS-CoV-2, AMOEBA semble plus proche de l'expérimentation.

Enfin, un dernier marqueur est étudié pour confirmer nos observations : l'empilement π–π entre **Phe140** et **His163**. Les résultats sont représentés sur la Figure 2.9. Tinker-HP ne capture pas cet empilement dans un protomère alors qu'à nouveau deux états coexistants (empilé et non empilé) sont observés dans l'autre protomère. Les mêmes observations peuvent être faites pour DESRES et RIKEN bien que les états soient moins bien définis en rapport avec la difficulté bien connue de capturer l'empilement π–π avec les n–PFF [39]. Malgré ces différences, les 3 simulations semblent cohérentes. Dans l'ensemble, notre conclusion initiale reste valable : nous décrivons une situation asymétrique dans laquelle un protomère est totalement inactif et l'autre présente des caractéristiques d'activité partielle. Il est important de préciser que ces résultats ne sont pas artificiels et liés à notre structure de départ. La Figure 2.12 montre la convergence du marqueur d'empilement sur la simulation 15.14 μs. Si le protomère 1 n'évolue clairement pas au cours de la simulation, le protomère 2 évolue lentement vers l'organisation à 2 états discutée. Dans l'ensemble, nos résultats sont compatibles avec la description de Zhou *et coll.* [243] de la structure cristallographique *apo* qui a observé un trou oxyanionique structuré incomplet présentant plusieurs états mixtes de structuration. Cela met en évidence la grande flexibilité de l'enzyme discutée dans la littérature expérimentale à température ambiante [109]. Nos données soutiennent également l'éventuelle forte asymétrie entre les protomères discutés dans l'état *holo* [241].

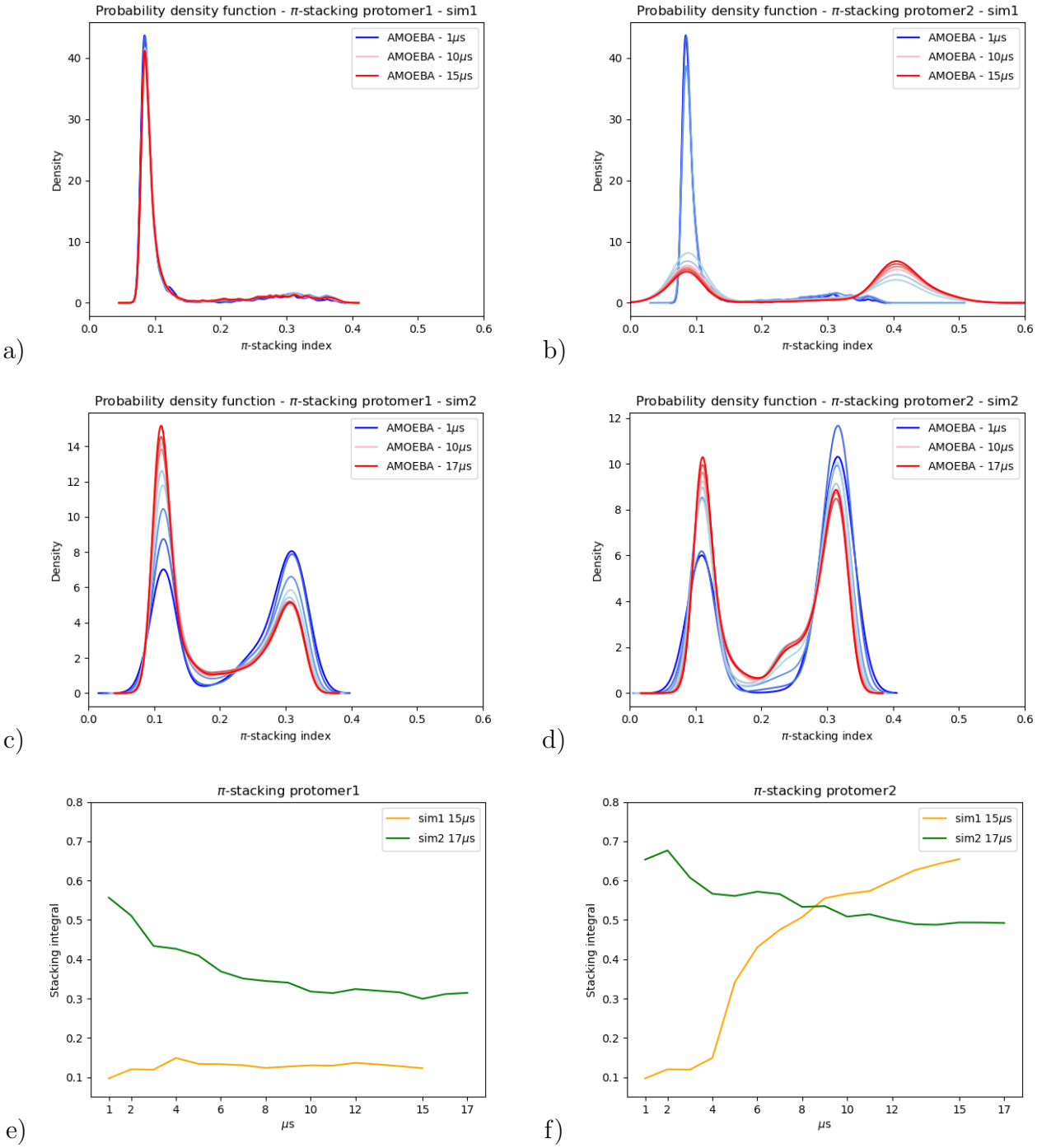


Figure 2.12: Convergence d'empilement π-π pour les deux protomères dans la simulation 1 (pH=7.4, histidines neutres) et la simulation 2 (pH=6, **His172** protoné pour les deux protomères) **a)**, **b)** et **c)**, **d)**. Intégrale de la fonction de densité de probabilité de l'empilement π-π (choisi supérieur à 0.25) sur le temps de simulation pour les deux protomères. **e)** et **f)**

2.3.2 Évaluation des volumes des cavités enzymatiques

Une façon de mesurer certaines différences globales potentielles entre les différentes simulations consiste à mesurer le volume du site actif dans chaque groupe et à décrire la tendance observée de la même manière que l'empilement π–π précédemment discuté. À côté de la cavité principale du site actif, la protéase principale présente 2 autres cavités : le site distal et le site de dimérisation. Bien que le site de dimérisation pourra être primordial à la stabilité de la forme dimérique de la M^{pro}, la fonction du site distal reste inconnue. Représentées sur la Figure 2.13, ces cavités sont considérées comme des cibles potentielles pour l'inhibition par des médicaments [72, 122]. Une description précise de chacune de ces cavités est essentielle à l'estimation d'inhibiteurs efficaces. Pour chaque cluster, nous avons donc estimé ces 3 volumes de cavités. Les volumes ont été calculés pour chaque amas isolé à l'aide du logiciel POVME 3.0 [221]. Pour chaque cavité, un pas de grille de 1 Å a été choisi. Les résidus 7–198, 198–306 et tous les résidus à moins de 3.5 Å de l'autre protomère ont été sélectionnés pour les sites actifs, distaux et de dimérisation avec respectivement 12 Å, 10 Å et 10 Å. 1 000 structures ont été choisies au hasard par cluster pour l'analyse. Lorsqu'un cluster compte moins de 1 000 structures, nous choisissons toutes les structures. Des informations détaillées sont données dans le Tableau 2.1 sur la taille de chaque cluster ainsi que leur taille réduite relative.

	Taille du cluster	Taille réduite
cluster1 DESRES	11 731	1 000
cluster1 RIKEN	347	299
cluster2 RIKEN	3 018	1 000
cluster3 RIKEN	2 463	1 000
cluster1 Tinker-HP	11 735	1 000
cluster2 Tinker-HP	8 256	1 000
cluster3 Tinker-HP	622	599
cluster4 Tinker-HP	2 062	1 000
cluster5 Tinker-HP	958	899

Table 2.1: Taille des clusters complets et réduits utilisés dans les volumes de cavité, études de flexibilité du terminal C et de poches cryptiques.

De manière similaire à l'empilement π–π et aux distances **Glu166**, nous avons utilisé l'estimateur de densité à noyau univarié (KDE) sur les volumes. Les volumes finaux sont représentés sur la Figure 2.13. Chaque cluster a une distribution normale validant la qualité des clusters issus du DBSCAN. Les différentes tendances qui apparaissent sont représentées par des flèches noires. Pour les 3 cavités, nous avons observé une similitude entre l'unique cluster de DESRES, les clusters 1 et 2 de RIKEN et les clusters 1 et 2 de Tinker-HP. Un accord est également trouvé avec les volumes obtenus par Stzain *et coll.* en utilisant la "Gaussian Accelerated MD (GaMD)": une stratégie d'échantillonnage améliorée couplée à AMBER ff14SB [208] qui correspondent également à ces résultats confirmant l'importance de simuler suffisamment longtemps en MD conventionnelle. Globalement, alors que les

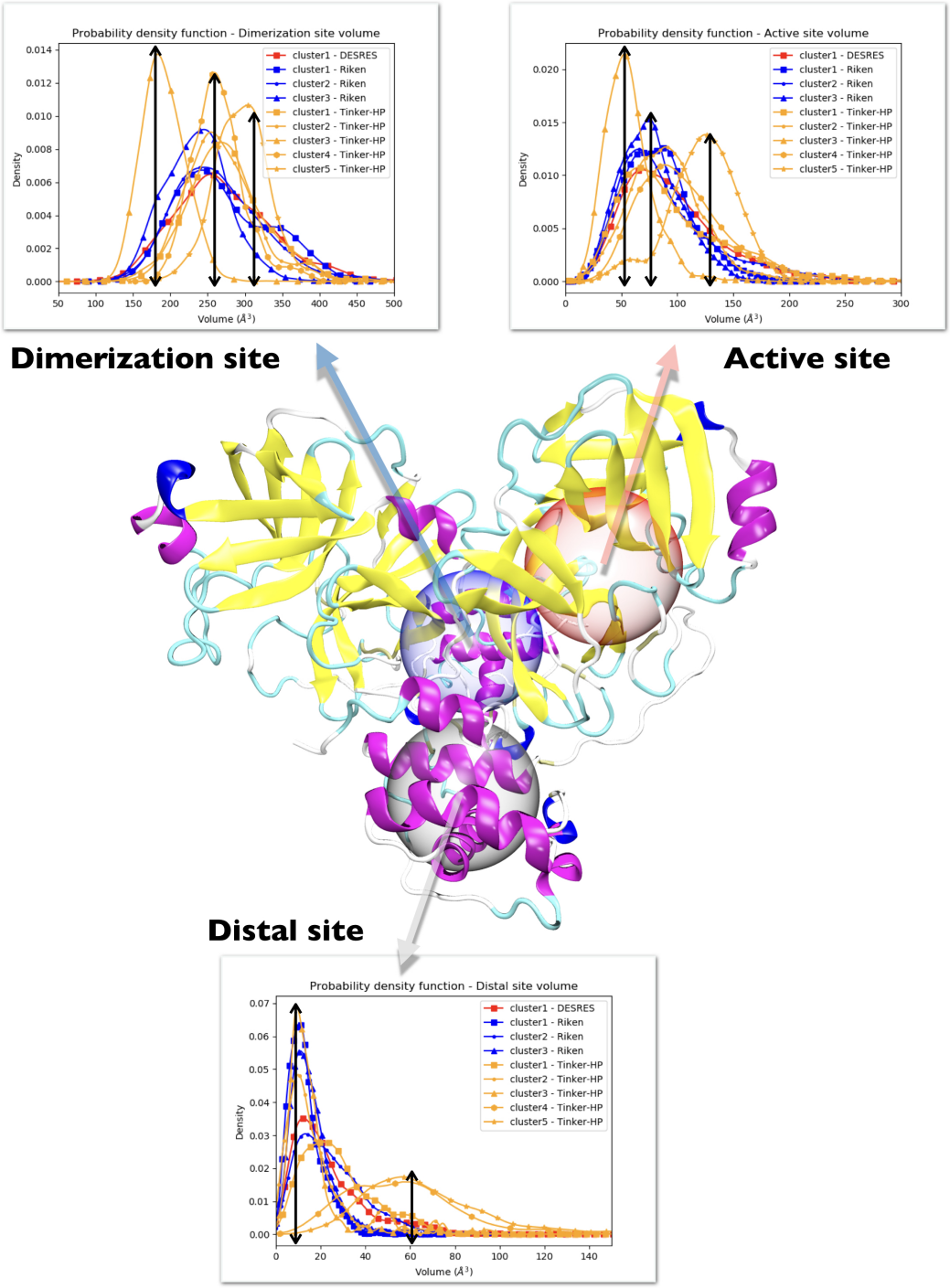


Figure 2.13: Représentation des 3 cavités considérées dans cette étude : le site de dimérisation, le site actif et le site distal. Pour chaque cavité, les tendances provenant de chaque cluster sont représentées et superposées dans trois graphiques différents. Chaque courbe a été débiaisée selon l'approche de repondération décrite dans ce travail. Les volumes de cavité sont la somme des volumes trouvés dans les deux protomères. Les flèches noires relient les maxima de fréquence à l'axe de volume pour mettre en évidence la différence entre les clusters.

clusters 1 et 2 de Tinker-HP sont en bon accord avec les clusters RIKEN et DESRES, nos clusters 3, 4 et 5 semblent différents et spécifiques soulignant l'importance du choix du PFF, c'est-à-dire que ces données n'étant pas obtenues en utilisant des échantillonnage couplé au non-PFF [208]. Comme nous l'avons souligné précédemment, des différences apparaissent en effet entre les clusters et entre les différentes trajectoires, allant dans le même sens que l'analyse précédente de l'empilement π–π entre les résidus **Phe140** et **His163** dans les chaînes A et B. Pour Tinker-HP, nous avons observé une contraction pour les trois cavités du cluster 3 alors que dans le cluster 4 et surtout le cluster 5, nous avons observé une forte différence avec une augmentation non négligeable des volumes des cavités. Les cavités des clusters 4/5 présentent des fluctuations de volume plus importantes lors de l'utilisation du PFF AMOEBA. Alors que les volumes de cavité provenant des simulations AMBER/DES-AMBER et des clusters 1 et 2 des simulations AMOEBA sont en accord, les résultats AMOEBA capturent clairement une caractéristique supplémentaire non capturée par les simulations DES-AMBER et AMBER. Cette information pourrait être importante pour la conception de nouveaux inhibiteurs potentiels. Par conséquent, étant donné que de fortes différences entre les méthodes sont observées dans les évaluations de volume des différents clusters, il est intéressant d'estimer les volumes globaux des protomères si l'on veut essayer de capturer davantage l'asymétrie discutée. Les volumes des cavités étudiées peuvent être trouvés sur la Figure 2.14. Le protomère 1 (par exemple,

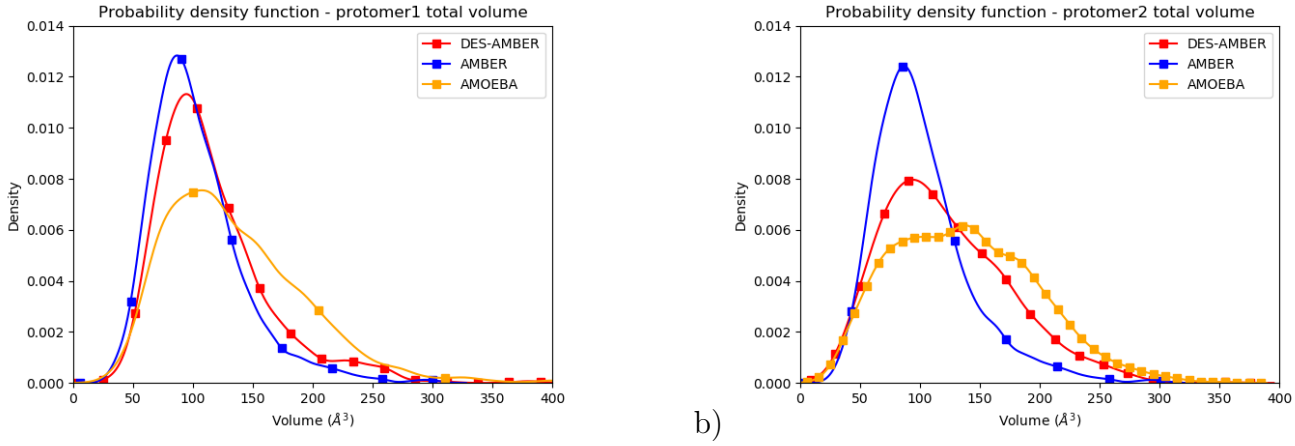


Figure 2.14: Représentation graphique des sites actifs pour le protomère 1 (**a**)) et le protomère 2 (**b**)) pour les simulations DESRES, RIKEN et Tinker-HP.

prédit pour être non-actif) dépeint un comportement gaussien fort tandis que le protomère 2 (par exemple montré pour osciller entre un état actif et un état non-actif) est caractérisé par un étalement gaussien avec un volume associé plus important par rapport au protomère 1. Cette augmentation de volume est donc concomitante à l'asymétrie précédente liée aux différents marqueurs structuraux discutés. Il est à noter que cette asymétrie se retrouve également pour la simulation DESRES mais dans une moindre mesure par rapport aux simulations AMOEBA Tinker-HP. Concernant l'ensemble de données RIKEN, cette caractéristique n'est pas découverte car les deux protomères décrivent une

tendance gaussienne similaire avec des valeurs très similaires.

2.3.3 Analyse des fluctuations locales : grande flexibilité de la région C-terminale

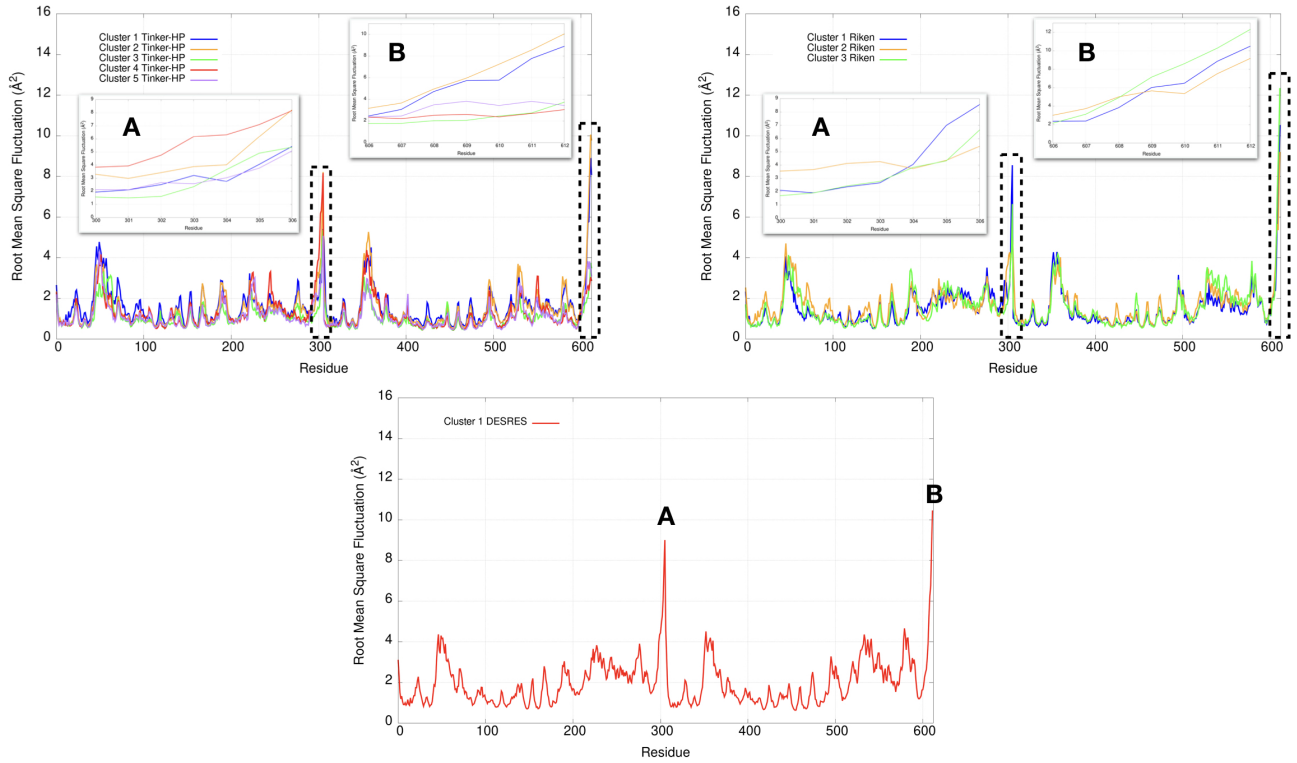


Figure 2.15: Représentation du RMSF pour chaque cluster de chaque simulation (Tinker-HP, RIKEN et DESRES). Les zooms sur les deux chaînes (A et B) sont représentés en sous-graphiques et correspondent à l'extrémité C-terminale où se trouvent les fluctuations les plus importantes (résidus 300 à 306 pour les chaînes A et B).

Enfin, il est également possible d'étudier les fluctuations locales de la dynamique structurelle du système de dimères de la M^{pro} pour découvrir d'autres types de différences entre les jeux de données. Nous avons calculé la fluctuation des résidus dans chaque cluster sur les mêmes structures précédemment choisies au hasard en se basant sur les "Root Mean Square Fluctuation" (RMSF). Ceux-ci ont été calculés sur les 5 clusters de Tinker-HP (AMOEBA), les 3 clusters de RIKEN (AMBER) et le cluster unique de DESRES (DES-AMBER). Les résultats sont représentés sur la Figure 2.15. La fluctuation la plus intéressante ainsi que les principales différences entre les clusters proviennent d'un ré-arrangement spatial différent de la région C-terminale de la protéine (les résidus 300 à 306 sur les chaînes A et B du dimère). En fait, cette région est très dynamique, ce qui est conforme aux observations expérimentales aux rayons X où la densité électronique du domaine C-terminal était

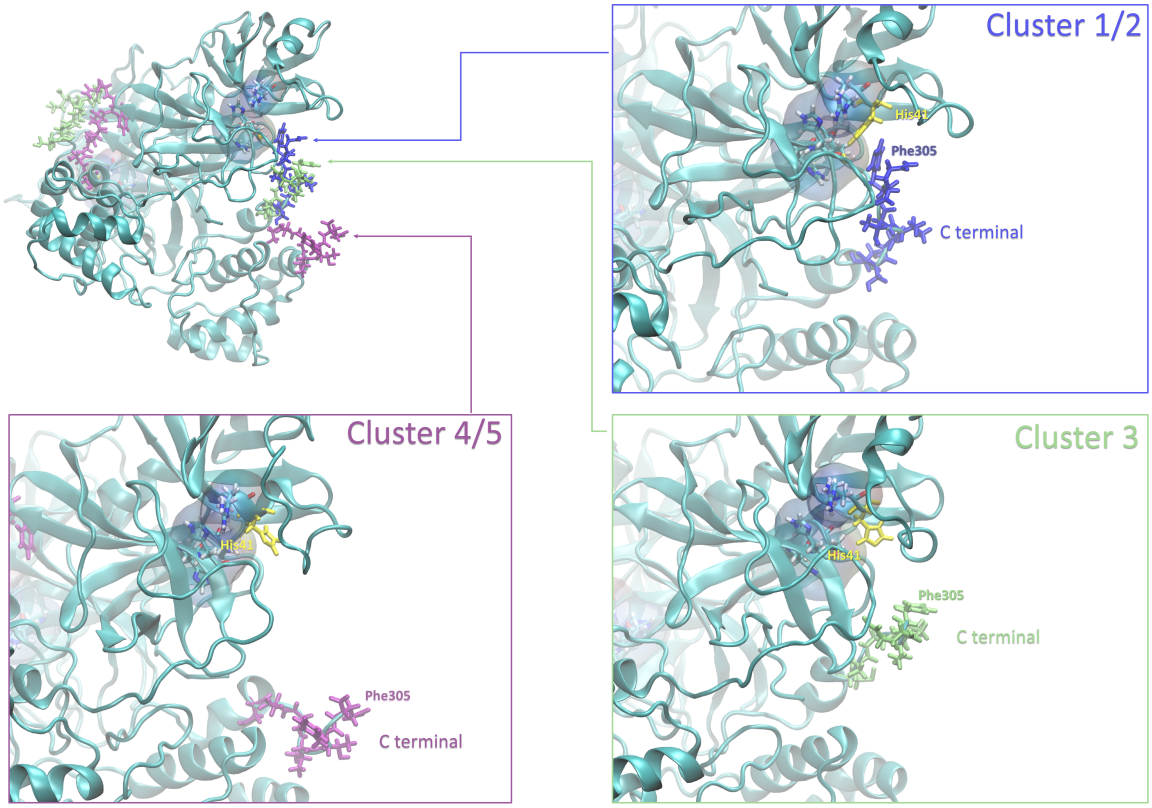


Figure 2.16: Représentation des 3 états possibles de la région C-terminale. La protéine entière est présentée en bleu glacier. L'extrémité C-terminale présentée en bleu saphir représente la plupart des états des clusters 1 et 2, où le résidu **Phe305** de la région C-terminale est empilé avec **His41** du site catalytique. L'extrémité C-terminale présentée en citron vert représente la plupart des états du cluster 3, et celle présentée en violet représente la plupart des états des clusters 4 et 5.

insuffisante pour le traçage du squelette, suggérant la flexibilité de cette région [243]. Des agrandissements visuels de cette région sont fournis dans les sous-graphiques de la Figure 2.15 pour les chaînes A et B qui ne diffèrent pas de manière significative. Le cluster 1 de la simulation DESRES présente la même fluctuation que le cluster 1 de la simulation RIKEN. Ce comportement de la région C-terminale dans ces deux clusters est caractérisé par une interaction π–π entre **Phe305** et **His41**, bloquant éventuellement l'accès de tout ligand au site actif. Lorsque la région C-terminale n'interagit pas avec **His41**, elle adopte une configuration dépliée qui montre une grande flexibilité de ces acides aminés terminaux. Les représentations structurelles peuvent être trouvées dans la Figure 2.16. Cet événement étant observé sur le site actif d'une seule chaîne et non des deux, il pourrait s'agir d'un autre marqueur de l'inactivation du protomère précédemment évoquée. Nous avons également observé de telles fluctuations dans les clusters 1 et 2 extraits de nos simulations Tinker-HP/AMOEBA. Cependant, dans le cluster 1, alors que l'interaction par empilement π–π **Phe305–His41** est bien observée, on mesure une plus faible fluctuation de la chaîne A pour le cluster 1. Elle correspond à une interaction plus faible entre **Phe305** et **His41** car les configurations où la région C-terminale est moins structurée sont préférées. Une caractéristique similaire est observée pour le cluster 2 de RIKEN, mais avec une inversion des pics de fluctuation entre A et B. Globalement, les clusters 1 et 2 obtenus à partir des simulations Tinker-HP et RIKEN apparaissent relativement similaires dans l'espace PCA. Ils correspondent à des clusters où la région C terminale peut osciller entre deux états : un avec une interaction d'empilement π–π entre **Phe305** et **His41**, et un autre avec une région C-terminale moins structurée et plus flexible. Les clusters 4 et 5 de nos simulations Tinker-HP et dans une moindre mesure le cluster 3 de RIKEN correspondent à une autre configuration de la région C-terminale. Des images représentatives sont fournies dans la Figure 2.16 pour chaque groupe de conformations C-terminales. Dans ces amas, la région C-terminale apparaît plus préservée/organisée car elle est localisée plus loin du site actif. Pour résumer la discussion concernant cette particularité, la grande flexibilité de la région C-terminale observée par rayons X remontent à un accès modulé au site actif lié à une absence de l'empilement π–π entre **Phe305** et **His41**. En d'autres termes, la région C-terminale du protomère totalement inactif oscille entre plusieurs états et l'un d'eux interagit directement avec l'autre site actif du protomère. Une telle interaction a tendance à bloquer l'accès au site actif et donc à moduler vers le bas l'activité du site potentiellement le plus actif. Cette grande flexibilité est capturée à la fois par RIKEN et Tinker-HP, illustrant l'importance de l'échantillonnage conformationnel local et confortant l'analyse expérimentale d'une inactivation complète de l'état *apo* [243].

2.3.4 Analyse comparative de la "druggabilité" : recherche de poches cryptiques

Afin de vérifier si toutes les fonctionnalités précédentes pouvaient affecter la "druggabilité" du système de dimères de la M^{pro}, nous avons cherché si de nouvelles poches cryptiques sont détectées dans chaque

cluster. En prenant en compte les mêmes ensembles que pour l'analyse des volumes de cavités, les

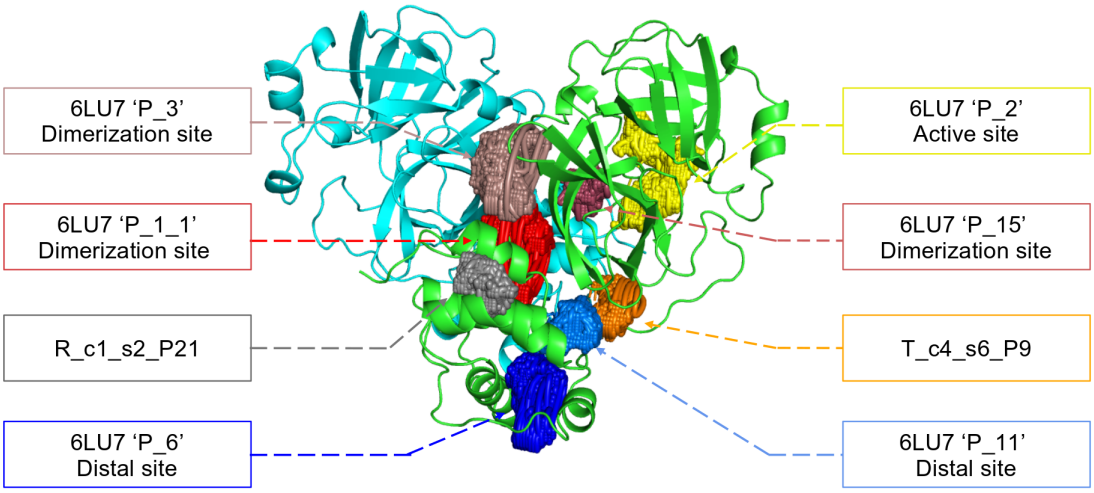


Figure 2.17: Représentation des emplacements des poches sur la structure de la protéase principale *6LU7* SARS-CoV-2

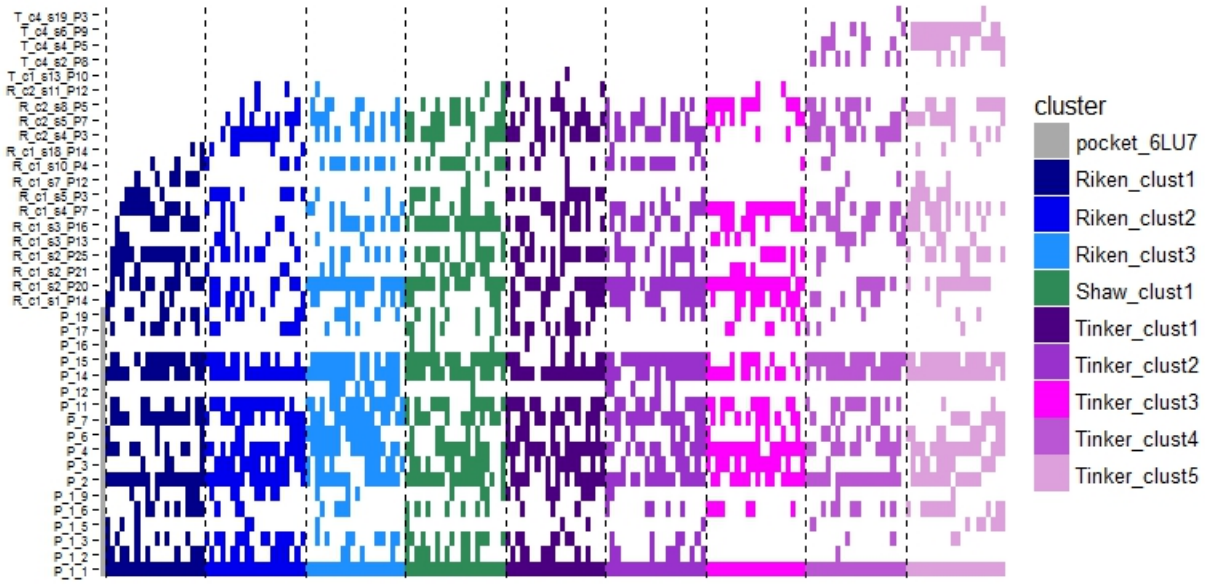


Figure 2.18: Représentation schématique des poches DoGSite Score détectées au sein de la structure *6LU7* (première colonne à gauche, représentée en gris) et 20 structures extraites de chaque cluster identifié au sein de RIKEN (gradient bleu), DESRES (vert) et Tinker-HP (gradient magenta) simulations.

poches cryptiques ont été recherchées à l'aide du logiciel DoGSite Scorer [220], un outil automatisé de détection de poches et de calcul de descripteurs de poches. DoGSite Scorer a détecté 18 poches situées sur la chaîne A ou à l'interface des chaînes A et B de la structure cristalline (PDB : *6LU7*)

de la protéase principale du SARS-CoV-2 *6LU7*. Parmi ces poches, 6 sont déjà décrites dans la littérature [208, 72] : les poches 'P_1_1', 'P_3' et 'P_15' correspondant au site de dimérisation ; la poche 'P_2' correspondant au site actif et les poches 'P_6' et 'P_11' situées dans la région distale. Ces 18 poches ont servi de référence et toutes les poches détectées sur les structures sélectionnées DESRES, RIKEN et Tinker-HP ont été affectées à ces poches de référence en comparant la liste des résidus des différentes poches et en sélectionnant la poche de référence avec le nombre maximum de résidus communs. Lorsque le nombre maximum de résidus communs était inférieur à 5 et que le rapport entre le nombre maximum de résidus communs et le nombre de résidus dans la poche prédite était inférieur à 0.25, la poche n'était affectée à aucune poche de référence et était définie comme un nouvelle poche cryptique. De nouvelles poches cryptiques ont été nommées d'après la première structure dans laquelle elles ont été détectées et ajoutées à l'ensemble des poches de référence. Par exemple, le 'R_c1_s1_P14' mentionné dans la Figure 2.18, est la poche P_14 détectée par DoGSite Scorer dans la structure 1 (s1) du cluster 1 (c1) des simulations RIKEN (R) . Les résultats de l'assignation des poches et de l'identification de nouvelles poches cryptiques sont présentés dans la Figure 2.18. Nous avons observé que les poches de référence précédemment mises en évidence comme "site actif", "site de dimérisation" et "site distal", sauf "P_6", sont particulièrement conservées et détectées dans une grande majorité des structures analysées. Cependant, un nombre conséquent d'autres poches ont également été détectées : i) dans quelques structures telles que 'R_c1_s2_P21', 'R_c1_s18_P14' ou 'T_c4_s19_P3' ou ii) dans de nombreuses structures, telles que 'R_c1_s2_P20', 'R_c1_s2_P25' ou 'R_c1_s4_P7'. Fait intéressant, 3 poches n'ont été récupérées que dans les clusters 4 et 5 des simulations Tinker-HP : 'T_c4_s2_P8', 'T_c4_s5_P5' et 'T_c4_s6_P9'. Le dernier, 'T_c4_s6_P9' est particulièrement intéressant puisque son volume est égal à 199 $Å^3$ et son score de druggability, DrugScore [192], atteint 0.62. Nous avons répété la procédure de détection et d'analyse des poches sur 100 structures sélectionnées au hasard (20 pour chacun des 5 clusters identifiés dans les simulations Tinker-HP. Nous avons observé que les 3 poches précédemment identifiées 'T_c4 _s2_P8', 'T_c4_s5_P5' et 'T_c4_s6_P9' ont également été détectés sur les structures sélectionnées aléatoirement dans le cluster 4 et 5 des simulations Tinker HP mais aussi partiellement dans le cluster 3. Nous avons ensuite évalué si toutes les poches affectées à la poche 'T_c4_s6_P9' présentaient des propriétés similaires. Nous avons observé que le volume moyen de ces poches était de 215 $Å^3$ mais que peu de structures présentaient des valeurs extrêmes bien supérieures à ce volume moyen (Figure 2.19). De même, la valeur moyenne de DrugScore était de 0.37 mais avec de grandes variations entre les structures et les clusters (voir Figure 2.20). À titre de comparaison, nous avons également calculé la distribution des valeurs DrugScore pour chaque poche nouvellement identifiée, c'est-à-dire les poches qui n'ont pas été détectées dans la structure *6LU7* (Figure 2.21). Une poche, 'R_c1_s2_P21' présente des propriétés particulières avec une valeur moyenne de druggabilité de 0.6 et une valeur de volume moyenne de 150 $Å^3$, ce qui semble indiquer que cette poche ne peut contenir que de très petits composés. La découverte de la poche 'T_c4_s6_P9' est donc un résultat très prometteur, mais qui souligne la nécessité de sélectionner avec soin une ou plusieurs structure(s)

dans laquelle les propriétés de la poche sont optimales pour la suite *in silico* recherches pour identifier de petites molécules capables de moduler l'activité de la protéase SARS-CoV-2. Toutes les poches décrites ici sont représentées dans la structure *6LU7* sur la Figure 2.17.

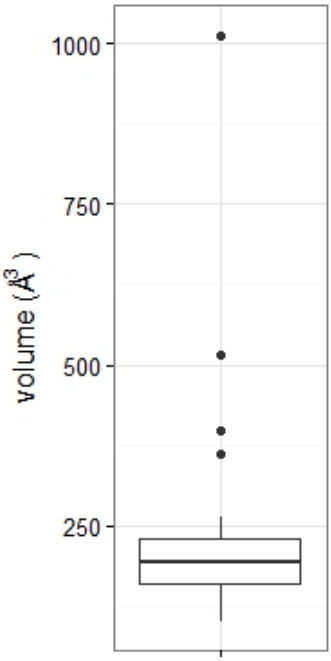


Figure 2.19: Boxplot de la répartition des volumes (calculée automatiquement avec DoGSite Scorer, en Å^3) parmi les poches correspondant à la poche 'T_c4_s6_P9'. (En raison de la différence dans leurs algorithmes, les volumes calculés par DoGSite Scorer sont différents de POVME.)

2.3.5 Analyse de solvatation : l'importance des effets de polarisation explicites dans l'eau

Les molécules d'eau jouent un rôle essentiel dans le fonctionnement des enzymes et des protéines. En fait, l'eau peut être un produit ou un réactif dans les réactions de condensation et d'hydrolyse, un intermédiaire d'état de transition dans les réactions chimiques et un élément structurel au niveau moléculaire. Dans ce dernier cas, l'eau inter-connecte la protéine par des liaisons hydrogène afin de maintenir et de stabiliser les positions des résidus et du repli [120]. Des études expérimentales antérieures sur le SARS-CoV-1 et le SARS-CoV-2 ont montré qu'une molécule d'eau structurelle était conservée dans la protéase principale des deux virus et interagissait avec l'azote cyclique de l'**His41** [210, 231, 109]. Une étude cristallographique récente sur le SARS-CoV-2 suggère qu'une autre molécule d'eau pourrait être observée autour de **His163** [243]. Afin de calculer le nombre de molécules d'eau à l'intérieur du site actif et à proximité de **His41** et **His163** des deux protomères, nous avons créé une sphère virtuelle de rayon 4 Å, centrée sur l'azote de chacune des deux histidines

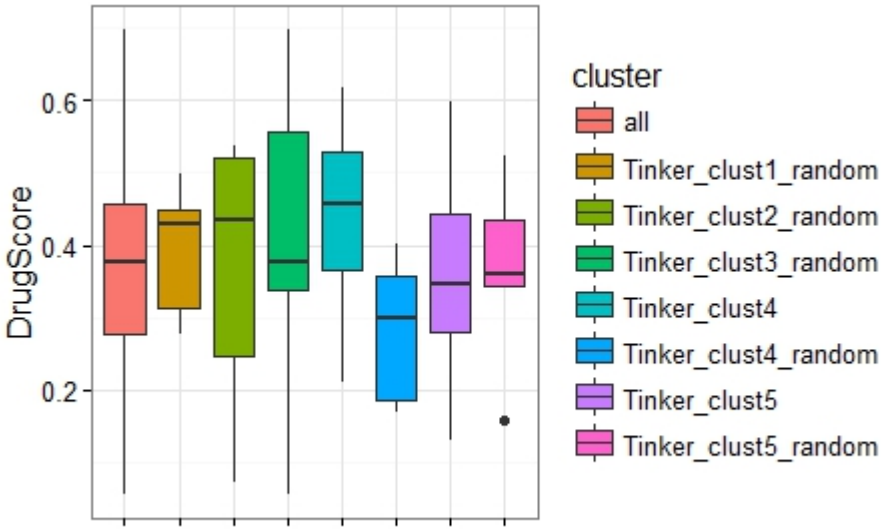


Figure 2.20: Boxplots de la distribution du DrugScore (calculé automatiquement avec DoGSite Scorer) parmi les poches correspondant à la poche 'T_c4_s6_P9'.

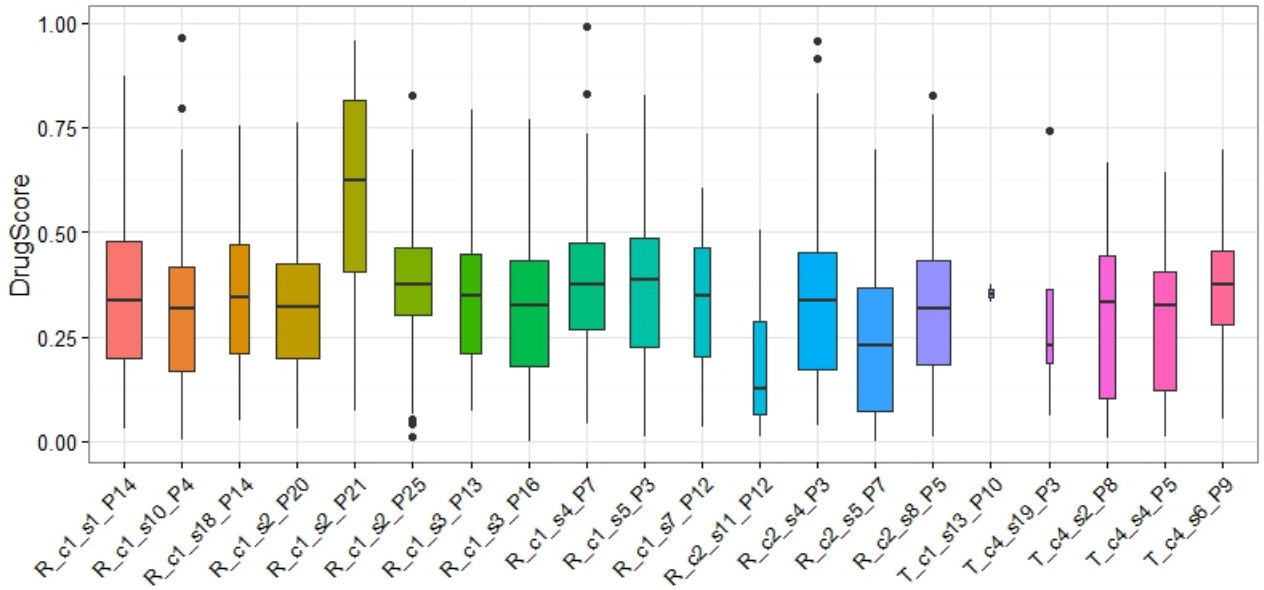


Figure 2.21: Boxplots de la distribution de DrugScore (calculé automatiquement avec DoGSite Scorer) pour chaque poche nouvellement identifiée, c'est-à-dire les poches qui n'ont pas été détectées dans la structure *6LU7*.

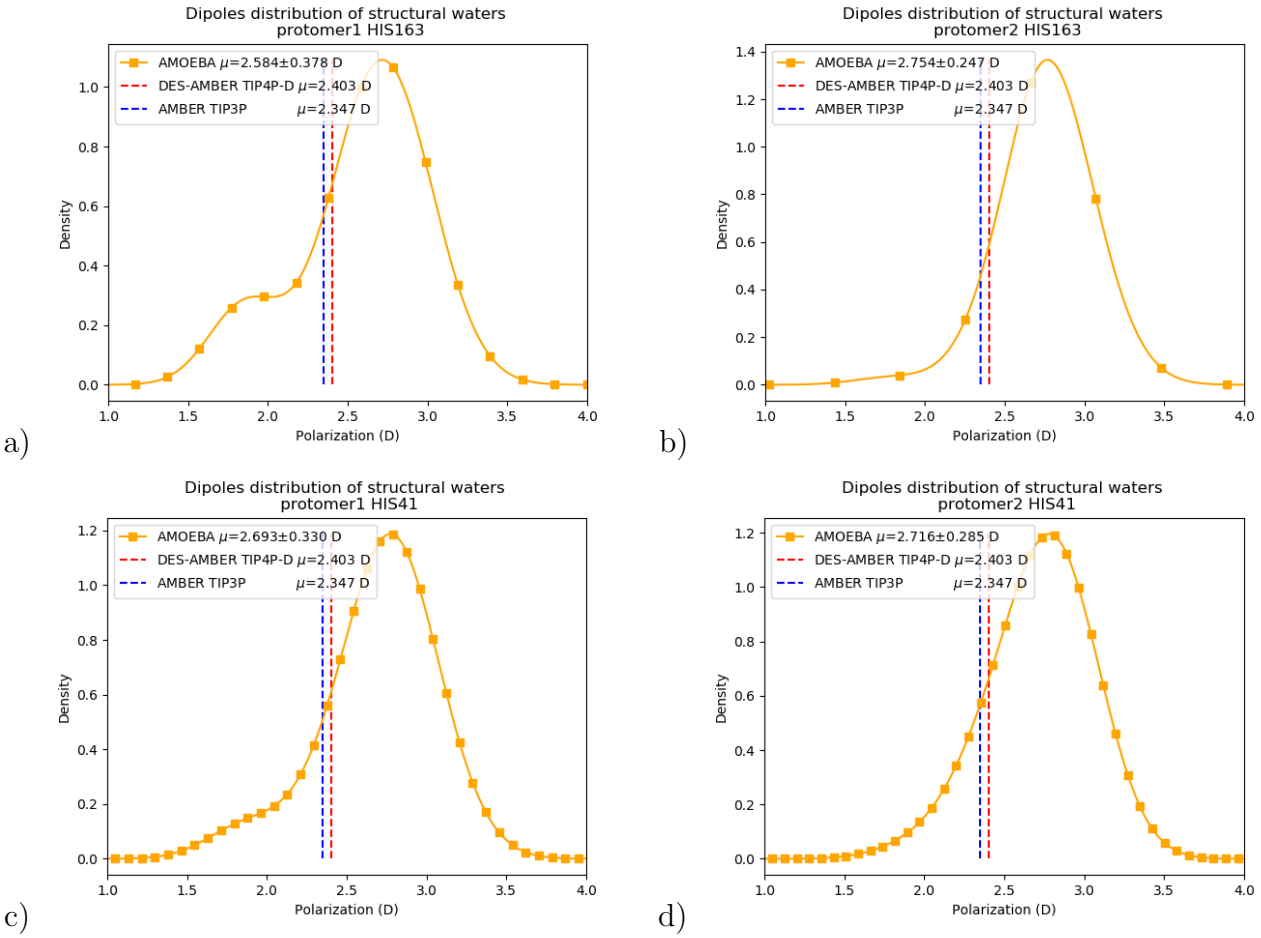


Figure 2.22: Distribution des moments dipolaires des molécules d'eau pour les protomères 1 et 2 autour de l'**His163 a)**, **b)** et autour de l'**His41 c)**, **d)**.

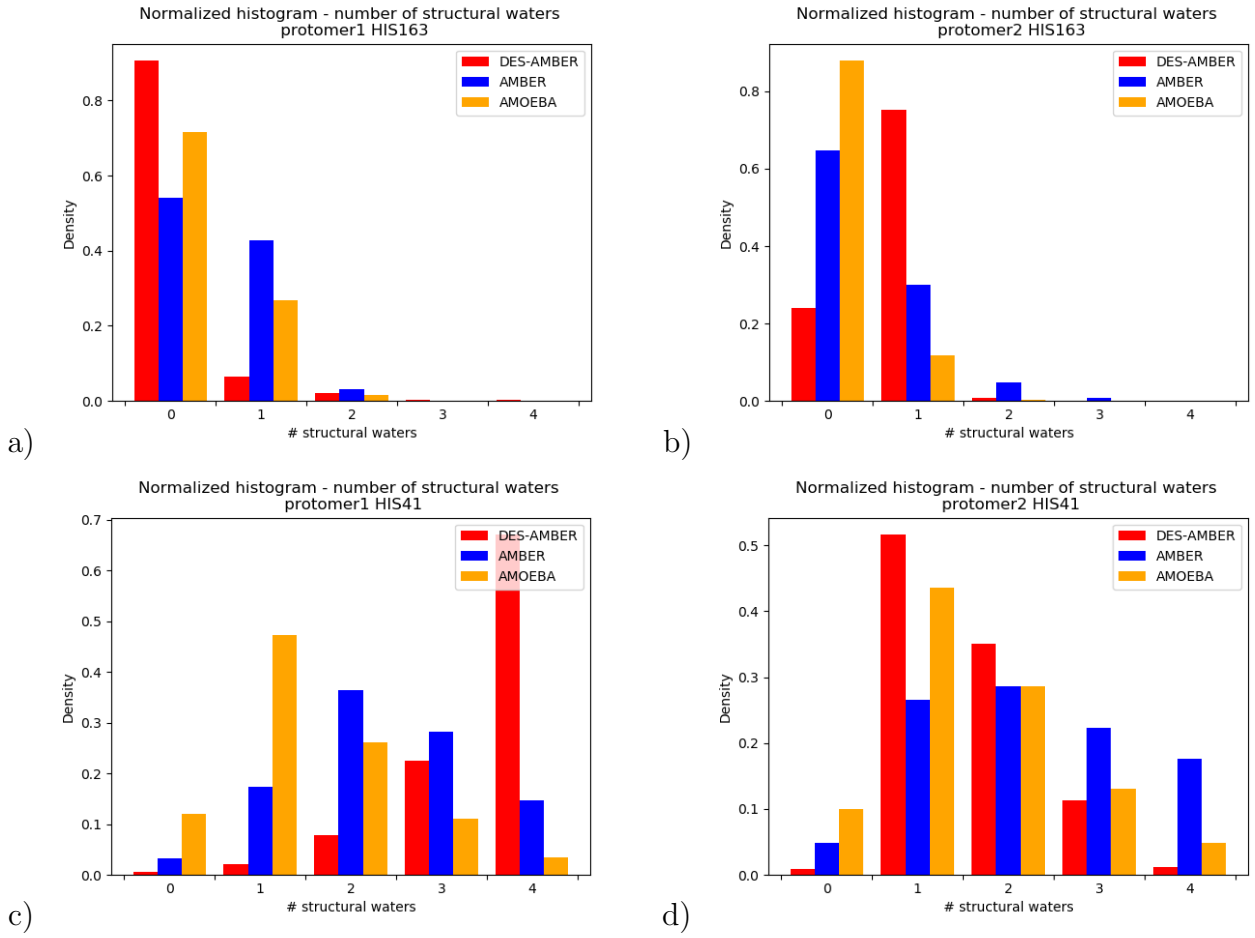


Figure 2.23: Nombre de molécules d'eau structurales autour de **His163 a)**, **b)** et **His41 c), pour les protomères 1 et 2**, **d)**.

concernées et avons calculé le nombre de molécules d'eau à l'intérieur du site actif de chaque protomère au cours du temps ainsi que leur moment dipolaire. La Figure 2.22 montre la distribution des dipôles des eaux structurelles pour les promoteurs 1 et 2 de **His163** a), b) et **His41**, c), d). Les résultats d'AMOEBA PFF sont frappants. Ils montrent que : i) les molécules d'eau dans chacun des deux sites actifs des protomères sont fortement polarisées ; et que ii) la distribution AMOEBA des molécules d'eau étant significativement différente de celles observées dans les trajectoires DESRES TIP4-D (DES–AMBER) et RIKEN TIP3P (AMBER). Des études antérieures ont montré qu'une polarisation élevée est une caractéristique commune des molécules d'eau structurelles qui présentent des moments dipolaires élevés [54]. En pratique, le moment dipolaire moyen ayant la densité la plus élevée avec le champ de force AMOEBA est situé autour de 2.9 D tandis que pour DES-AMBER et AMBER n–PFF , les dipôles d'eau sont fixés respectivement à 2.403 D et 2.347 D. Comme les moments dipolaires AMOEBA ne sont pas fixes, nous observons de fortes fluctuations de polarisation dues au trafic d'eau à l'intérieur de la région catalytique. La Figure 2.23 représente le nombre de molécules d'eau structurelles pour les protomères 1 et 2 de **His163** a), b) et **His41** c), d). Toutes les

trajectoires montrent une densité la plus élevée pour aucune molécule d'eau à une distance de 4 Å du protomère 1 de **His163**. Cependant cette observation est différente pour le protomère 1 de **His41** où les trajectoires Tinker-HP ont trouvé la densité la plus élevée pour la présence d'une seule molécule d'eau alors que RIKEN et DESRES les trouvent pour 2 et 4 molécules respectivement. Une distribution non symétrique des molécules d'eau par rapport au protomère 1 est trouvée pour le protomère 2. Les trajectoires Tinker-HP et RIKEN ne prédisent pas de présence fréquente de molécules d'eau à la distance choisie de **His163**, tandis que DESRES présente une densité plus élevée pour 1 molécule. Concernant **His41** du protomère 2, les trajectoires de Tinker-HP et de DESRES montrent une densité la plus fréquente pour une seule molécule d'eau, tandis que la densité la plus élevée de RIKEN est trouvée pour 2 molécules d'eau. Ces observations démontrent que la polarisation de l'eau fluctue intensément à l'intérieur du site actif confiné, suggérant un rôle dynamique de la polarisation sur le trafic de l'eau qui influence fortement les interactions des molécules d'eau avec **His163** et **His41** de chacun des deux protomères. Cependant ces interactions ne sont pas distribuées symétriquement entre les protomères. Alors est-ce compatible avec les données expérimentales ? Encore une fois, des données de rayons X relativement détaillées existent pour d'autres coronavirus, y compris le SARS-CoV-1 où le rôle des histidines a été largement discuté [210, 231]. La présence d'une eau structurale autour de **His41** est toujours confirmée. Pour le SARS-CoV-2, les articles décrivant la structure de la protéase M^{pro} dans son état *apo* [243, 109] dans des conditions de pH physiologique discutent également de la présence d'une telle molécule trouvée près de la dyade catalytique. Cependant, la molécule d'eau structurelle interagissant avec **His163** n'apparaît que proposée dans les données de Zhou *et coll.* [243].

Concernant la prédiction du nombre exact de molécules d'eau autour de **His41**, AMBER et DES–AMBER ont en moyenne un nombre plus élevé d'eaux structurales avec au plus 2.38 à 4.01 par rapport à AMOEBA qui prédit la présence de 1.5 molécule d'eau plus conforme aux données expérimentales accumulées. La Figure 2.23 montre que les simulations non polarisables capturent des configurations fréquentes ayant jusqu'à 4 molécules d'eau, ce qui pourraient être une conséquence de la non-inclusion de l'effet de polarisation conduisant à un moment dipolaire plus faible et constant des molécules d'eau ce qui pourrait générer plus de trafic d'eau. Comparant à l'**His41**, toutes les analyses AMOEBA, AMBER et DES-AMBER ont trouvé significativement moins de molécules d'eau autour de **His163**. En pratique, AMOEBA a trouvé le plus faible nombre d'eau de toutes les méthodes avec une moyenne de 0.13–0.31 molécules autour de **His163**, tandis que les tendances plus élevées observées pour **His41** sont toujours présentes pour tous les n–PFF à l'exception d'un protomère de DES–AMBER 0.77 molécules (voir Tableau 2.2).

En clair, la présence d'une molécule d'eau structurale autour de l'**His163** semble moins probable pour toutes les simulations (aux conditions de pH actuelles) et en concurrence avec le trafic d'eau entrant dans la sphère de mesure. La distribution dipolaire des molécules d'eau offre une analyse plus approfondie car elle s'avère légèrement plus grande pour l'**His163** et associée à une densité plus

	His163		His41	
	Protomer 1	Protomer 2	Protomer 1	Protomer 2
DES-AMBER	0.14 σ=0.48	0.77 σ=0.44	4.01 σ=1.17	1.61 σ=0.75
AMBER	0.49 σ=0.57	0.44 σ=0.41	2.38 σ=1.11	2.25 σ=1.23
AMOEBA	0.31 σ=0.51	0.13 σ=0.34	1.48 σ=0.99	1.62 σ=1.06
Expériences [230, 241, 243, 109]	0 ou 1		1	

Table 2.2: Moyenne et écart type du nombre de molécules d'eau autour des résidus **His163** et **His41** dans les simulations de champs de force DES-AMBER, AMBER et AMOEBA (pH 7.4).

	His163		His41	
	Protomer 1	Protomer 2	Protomer 1	Protomer 2
AMOEBA pH 6	0.37 σ=0.65	0.27 σ=0.57	1.95 σ=1.04	1.42 σ=0.97
AMOEBA pH 7.4	0.31 σ=0.51	0.13 σ=0.34	1.48 σ=0.99	1.62 σ=1.06
Experiments	0 ou 1		1	

Table 2.3: Moyenne et écart type du nombre de molécules d'eau autour des résidus **His163** et **His41** en utilisant AMOEBA pour les simulations à pH 7.4 et 6.

faible de moments dipolaires totaux hautement polarisés. Dans tous les cas, la présence d'eau dans le site actif apparaît donc cohérente avec le besoin d'une molécule d'eau pour modéliser le mécanisme de réaction enzymatique [247, 109].

2.4 Étude de la région de dimérisation

2.4.1 Stabilité de l'interface de dimérisation : études sur les liaisons hydrogènes

Pour commencer notre analyse de la dynamique structurelle de la M^{pro} au niveau de l'interface de dimérisation, nous avons déterminé le nombre d'interactions de liaison hydrogène (liaison H) afin d'évaluer la robustesse des interactions non covalentes entre les deux protomères. Partant du pH physiologique, nous avons analysé les trajectoires DES-AMBER (DESRES), AMBER (RIKEN) et AMOEBA (Tinker-HP) fournies au sein des ensembles de conformation disponibles. Nous avons trouvé des fonctions de densité de probabilité d'interaction de liaison H relativement similaires entre les trois profils (voir Figure 2.24, a)) qui présentent tous une forte stabilité de l'interface de dimérisation. En comparant la distribution physiologique des liaisons H à des simulations AMOEBA à pH inférieur (voir Figure 2.24, b)), nous avons trouvé une transition d'une distribution gaussienne nette centrée

à 14 liaisons H (pH 7.4) à une distribution plus diffuse pour les pH 6 et au-dessous, présentant les implications d'interactions plus faibles et désorganisées.

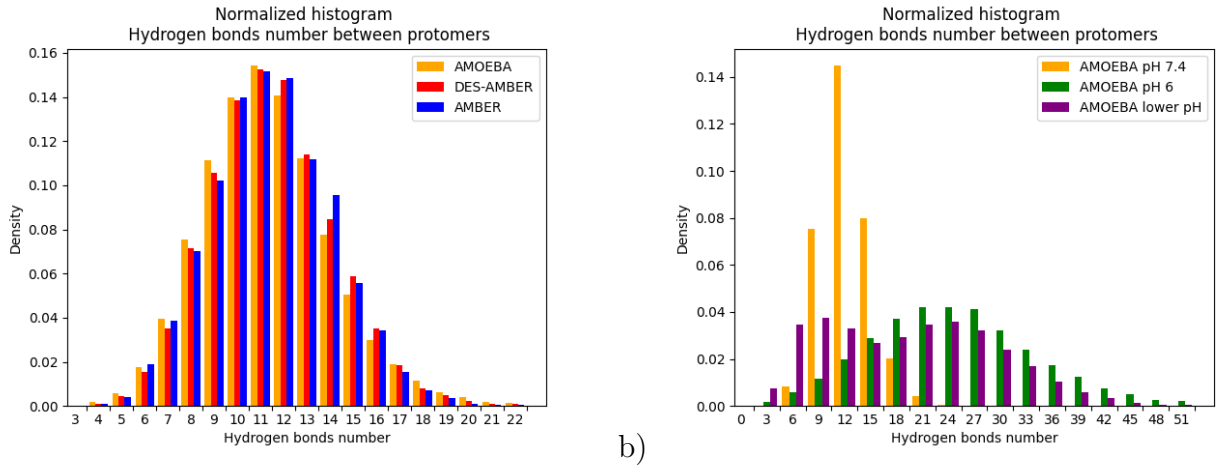


Figure 2.24: Représentation par histogramme de la densité de probabilité des liaison H pour les champs de force **a)** DES-AMBER, AMBER et AMOEBA à pH = 7.4 et **b)** pour les trajectoires AMOEBA à pH 7.4, 6 et au-dessous.

Clairement, nos résultats montrent un effondrement de l'interface du dimère à des valeurs de pH inférieures à celles physiologiques suite aux protonations successives des résidus d'histidine (**His172** puis **His163**) [210, 93, 217]. Parmi les interactions observées dans le Tableau 2.4, les interactions **Arg4–Glu290** et **Gly11–Glu14** ont la densité de probabilité la plus élevée de tout le trajectoires DES-AMBER, AMBER et AMOEBA à pH physiologique. Pourtant, ces interactions ne sont pas détectées à un pH inférieur, ce qui est cohérent avec les études expérimentales rapportant qu'un pH faible est responsable de la perte de l'interface du dimère [72, 44]. Il est important de noter ici que la protonation de **His172** à un pH plus bas a récemment été démontrée [217, 109, 93] comme étant également la source d'un effondrement partiel au niveau du site catalytique. Étant donné que l'interface du dimère est connue pour être pleinement fonctionnelle à pH physiologique, nos résultats à plusieurs pH renforcent le rôle critique de l'état de protonation de l'**His172** et sont cohérents avec les découvertes de Verma *et coll.* [217] sur l'**His172** non protoné à pH physiologique. Un examen détaillé du profil d'interaction de la liaison H dans le Tableau 2.4 met en évidence le rôle clé de **Arg4** dans le maintien de la dimérisation à travers plusieurs interactions, principalement avec **Glu290** mais aussi avec **Lys137**, **Ser139**, **Glu288** et **Asp289** à pH physiologique. Ceci est cohérent avec la description des résidus clés pour le maintien de la dimérisation de la M^{pro} du SARS-CoV-2 dans la littérature expérimentale [123] : **Arg4, Ser10, Gly11, Glu14, Asn28 , Ser139, Phe140, Ser147, Glu166, Glu290, Arg298**, Figure 2.26.

Ces résidus apparaissent tous le long de notre analyse, à l'exception de la **Ser147**. Néanmoins, nous avons pu ici élargir la liste de ces résidus après une analyse détaillée des simulations DES-AMBER, AMBER et AMOEBA. Comme le montre le Tableau 2.4, AMOEBA prédit une liste plus

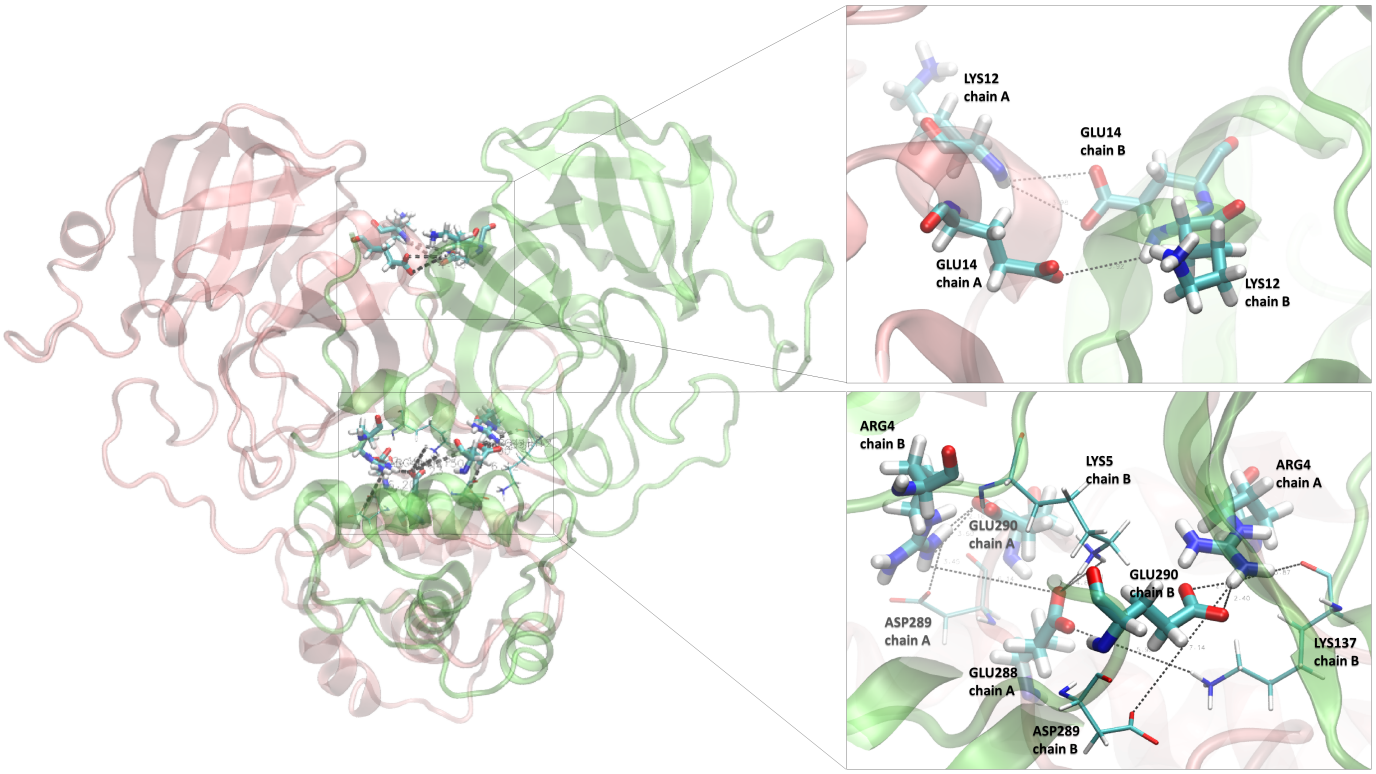


Figure 2.25: Représentation des interactions H les plus fréquentes à l'interface de dimérisation de la M^{pro} du SARS-CoV-2, les chaînes **A** et **B** sont présentées respectivement en rose et en citron vert.

riche et plus exhaustive de résidus impliqués par la dimérisation par rapport à AMBER et DES-AMBER. Les formes spéciales détectées de liaison H et d'autres interactions, à pH physiologique, sont mises en évidence dans la Figure 2.24, c). Il est important de noter que lorsque des protonations successives d'histidines se produisent, **His172** et **His163** passent d'histidines neutres à pH=7,4 à chargées positivement à pH=6 et moins, modifiant la nature de certaines de leurs interactions avec d'autres résidus et de l'eau (par exemple le passage des liaisons H aux ponts salins dans certains cas [124, 93]). Bien que l'abaissement du pH affecte également d'autres résidus qui ne sont pas tous pris en compte dans nos calculs [217], ce changement physico-chimique dans la nature des interactions histidines est central dans l'affaiblissement de la stabilité de l'interface, l'obligeant à redistribuer son réseau de liaison H selon une configuration différente et moins structurée. Enfin, le Tableau 2.4 révèle également que les interactions **Arg4–Glu290** et **Gly11–Glu14** sont les liaisons H les plus importantes responsables de la stabilisation de l'interface de dimérisation car elles présentent les densités les plus élevées au pH physiologique et sont absentes de la simulations aux pH plus bas. Globalement, ces résultats mettent en évidence le fait que le réseau complexe de liaisons H est celui de la force motrice stabilisant l'interface de dimérisation.

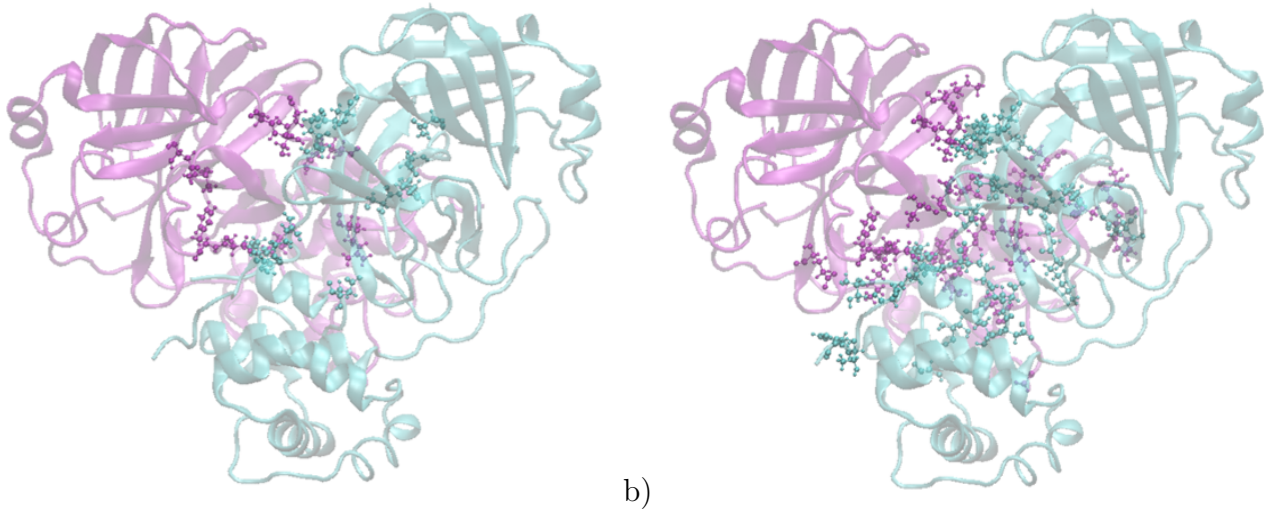
DES-AMBER		
Donor	Acceptor	Density
ARG4-Side	**GLU290-Side**	1.57
GLY11-Main	**GLU14-Side**	1.38
VAL125-Main	ALA7-Main	1.01
ALA7-Main	VAL125-Main	0.97
ARG4-Side	LYS137-Main	0.86
SER10-Main	**GLU14-Side**	0.29
SER1-Side	**GLU166-Side**	0.28
SER1-Main	**GLU166-Side**	0.27
SER139-Side	**ARG4-Main**	0.20
SER1-Main	**PHE140-Main**	0.18
ARG4-Main	SER139-Side	0.18
SER139-Side	GLN299-Side	0.13
SER301-Side	SER139-Side	0.12
SER10-Main	**SER10-Side**	0.10
THR304-Side	**GLU166-Side**	0.10

AMBER		
Donor	Acceptor	Density
ARG4-Side	**GLU290-Side**	1.52
GLY11-Main	**GLU14-Side**	1.41
ALA7-Main	VAL125-Main	1.02
VAL125-Main	ALA7-Main	0.99
ARG4-Side	LYS137-Main	0.73
SER1-Main	**GLU166-Side**	0.50
SER1-Main	**PHE140-Main**	0.43
SER1-Side	**GLU166-Side**	0.41
SER10-Main	**GLU14-Side**	0.38
SER139-Side	ARG4-Main	0.38
ARG4-Main	SER139-Side	0.34
GLY2-Main	**GLU166-Side**	0.32

AMOEBA, pH 7.4		
Donor	Acceptor	Density
ARG4-Side	**GLU290-Side**	2.07
GLY11-Main	**GLU14-Side**	1.39
ALA7-Main	VAL125-Main	1.02
VAL125-Main	ALA7-Main	1.00
ARG4-Side	LYS137-Main	0.82
SER1-Side	GLU166-Side	0.37
SER139-Side	**ARG4-Main**	0.37
LYS137-Side	SER284-Side	0.26
GLY278-Main	ASN238-Side	0.24
ARG4-Side	GLU288-Side	0.21
ARG298-Side	SER123-Main	0.17
ARG4-Main	**SER139-Side**	0.16
ARG4-Side	ASP289-Side	0.15
ASN277-Main	LYS236-Main	0.13
SER139-Side	GLN299-Side	0.13
THR304-Main	**GLU166-Side**	0.12
THR304-Side	**GLU166-Side**	0.12

AMOEBA, pH 6		
Donor	Acceptor	Density
VAL125-Side	ALA7-Main	0.35
MET6-Side	TYR126-Side	0.29
GLY11-Main	**GLY11-Main**	0.22
MET6-Side	VAL125-Main	0.17
PRO9-Side	**GLU14-Side**	0.16
ARG4-Side	**GLU290-Side**	0.14
VAL125-Side	PHE8-Main	0.12
ARG4-Side	TYR126-Side	0.11

Table 2.4: Liste des résidus impliqués dans les interactions de liaisons hydrogène avec une occupation supérieure à 10% tout au long des trajectoires, entre les deux protomères SARS-CoV-2 M^{pro}. Les résidus en gras ont déjà été mentionnés dans la littérature comme des résidus importants pour la dimérisation M^{pro}.



a) b)

Figure 2.26: Représentation des résidus impliqués dans la dimérisation due la M interactions H les plus fréquentes à l'interface de dimérisation de la M^{pro} du SARS-CoV-2, a) dans la littérature (**Arg4, Ser10, Gly11, Glu14, Asn28, Ser139, Phe140, Ser147, Glu290, Arg298**) et b) d'après nos simulations moléculaires (**Arg4, Ser10, Gly11, Glu14, Asn28, Ser139, Phe140, Ser147, Glu290, Arg298**).

2.4.2 Interactions allostériques entre l'interface de dimérisation avec le site catalytique

Pour approfondir la compréhension de la complexité de l'interface de dimérisation, nous avons décidé d'examiner ses interactions allostériques potentielles au sein de la M^{pro}. En effet, l'allostérie correspond aux changements conformationnels se produisant sur un site d'une protéine et provoquant des changements structurels ou dynamiques sur un site distant et topologiquement indépendant. De tels changements conduisent à une réduction ou à une augmentation de l'activité catalytique parmi d'autres réarrangements structurels. La prédiction basée sur la structure des sites allostériques, des modulateurs et des voies de communication est importante pour une meilleure compréhension du comportement moléculaire et est importante en vue de découvrir des petites molécules à effets thérapeutiques afin de réguler la fonction des protéines [149, 73]. Étant donné que les liaisons H jouent un rôle très important dans la région de dimérisation, elles peuvent être en mesure d'influencer son volume, ce qui pourrait également avoir des effets structurels sur d'autres poches de surfaces protéiques via des corrélations allostériques [149]. La "druggabilité" de l'interface de dimérisation a été discutée dans la littérature [72, 93] mais peu de travaux se sont penchés sur les interactions allostériques potentielles. En effet, l'importance de la connectivité allostérique entre les sites allostériques et fonctionnels a été de plus en plus observée ces dernières années [174, 47]. Plusieurs sites allostériques potentiels ont récemment été discutés afin d'offrir des stratégies de ciblage de médicaments à effets allostériques [75, 206, 41] au niveau de la M^{pro} du SARS-CoV-2. Par exemple, Stromich *et coll.* [206] ont étudié le score de sites allostériques putatifs et ont souligné une zone située dans le site de dimérisation

montrant une connectivité élevée vers le site catalytique actif. Ils ont définit un site potentiel de dimérisation allostérique formé par les six résidus suivants de l'interface de dimérisation : **Arg131**, **Asp197**, **Thr199**, **Asp289** et **Glu290** de la chaîne **A** et **Arg4** de la chaîne **B**. Étant donné que plusieurs de ces résidus ont été montrés par nos simulations comme déterminants pour la stabilisation de l'interface (Tableau 2.4 et discussion précédente), nous avons décidé d'étudier ce site. Afin d'évaluer une connectivité allostérique potentielle du site de dimérisation allostérique vers le site actif catalytique des deux chaînes et d'analyser sa dynamique structurelle, nous avons eu recours à une analyse approfondie de la propension des liaison à liaison [13]. En utilisant cette approche, nous mesurons les fluctuations d'ensembles donnés d'interactions atome-atome et analysons comment elles affectent tout autre ensemble d'interactions situées ailleurs dans la protéine, permettant ainsi de mesurer leur connectivité instantanée à chaque moment de la dynamique. Nous avons calculé dans un premier temps l'évolution des distances situées à l'intérieur du site de dimérisation allostérique avec d'autres distances caractéristiques impliquées dans les résidus formant la dyade catalytique. Ainsi, grâce à des atomes ou résidus de référence bien choisis, cette étude nous renseigne indirectement sur la co-évolution des deux volumes de cavité. En effet, comparer leurs fluctuations de volume le long des trajectoires peut nous renseigner sur une possible connectivité allostérique entre eux. [189, 206].

Nous montrons sur la Figure 2.27, a) et b) un graphique en 2D des distances séparant les résidus de la dyade catalytique pour les deux chaînes **A** et **B** *vs* les distances entre les résidus du site de dimérisation allostérique : **Arg4** de la chaîne **B** et **Glu290** de la chaîne**A** présentant une interaction robuste. Les trajectoires AMOEBA montrent une forte densité de structures ayant à la fois des sites de dimérisation catalytique et allostérique étroits, respectivement autour de 4Å et 3Å comme le montre la figure2.27, a) et b). Cependant nous sommes également capables de détecter une organisation différente de structures qui se caractérisent par un site de dimérisation allostérique étroit et un site catalytique relaxé et inversement, ce qui propose une connectivité allostérique possible entre les tailles des sites allostérique de dimérisation et catalytique. Cette connectivité supplémentaire trouvée dans les simulations AMOEBA n'est pas observée dans DES-AMBER ni dans les simulations AMBER, Figure 2.27, c) et d).

Dans notre plan d'échantillonnage adaptatif, le score est défini comme le rapport entre les probabilités d'obtenir une structure donnée q_i dans la simulation biaisée et dans une simulation débiaisée. Ici, nous nous limitons aux structures avec un score de repondération supérieur à 1 car ils sont plus susceptibles d'être visités lors d'une simulation de MD conventionnelle. En revanche, les structures avec des scores inférieurs à 1 ont été favorisées par l'algorithme adaptatif pour maximiser l'exploration et sont donc moins pertinentes physiquement pour la statistique du système (plus d'informations peuvent être trouvées dans la référence [93]). Ainsi, les structures présentées en orange sur la Figure 2.27, sont plus représentatives des vraies statistiques AMOEBA. Dans ce cas, nous détectons principalement des structures ayant un site catalytique relaxé et un site de dimérisation allostérique étroit. Ceci suggère que cette dépendance spécifique est détectée grâce à l'utilisation de l'AMOEBA FF polar-

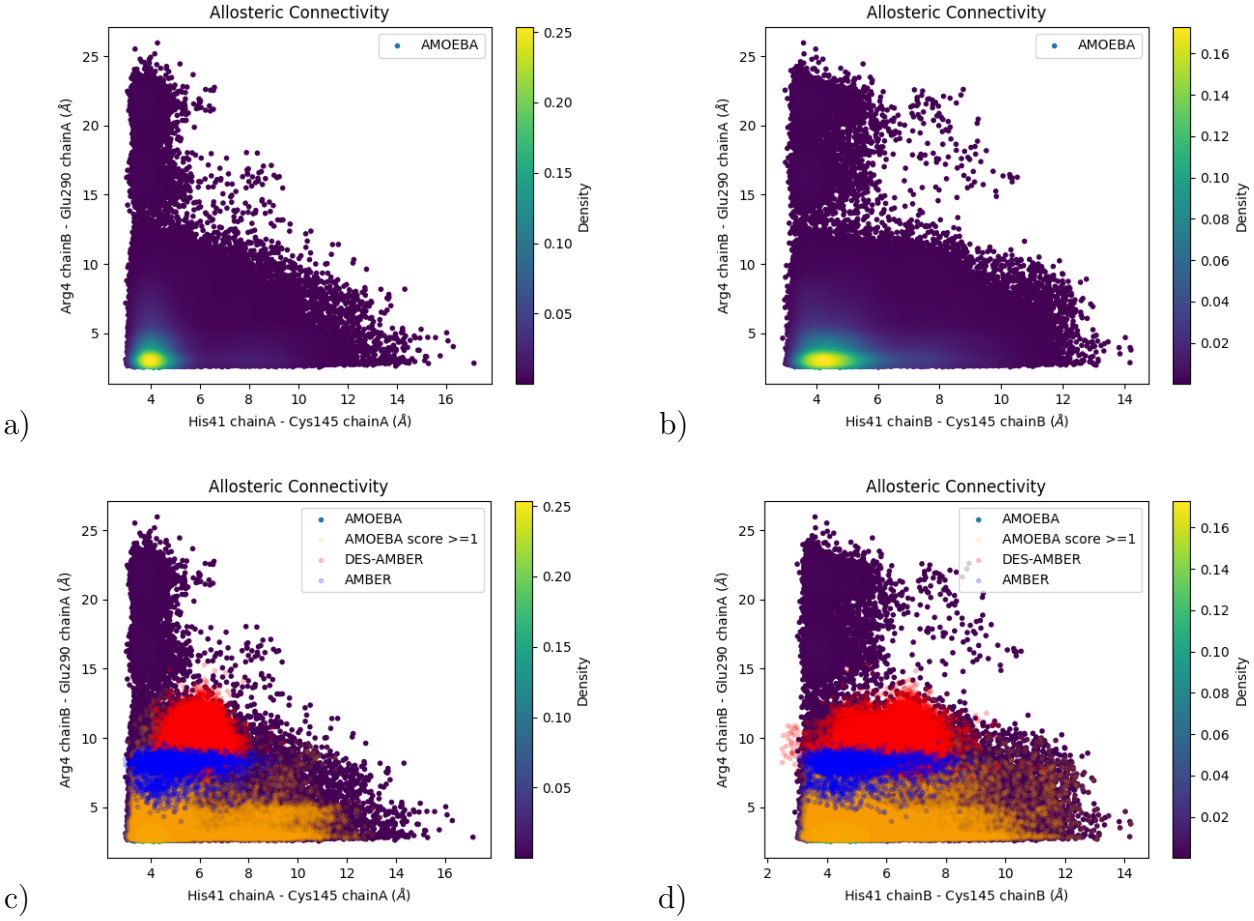


Figure 2.27: Représentation graphique 2D de la distance ***Arg4*** *chaîne* ***B*** – ***Glu290*** *chaîne* ***A*** vs la distance ***His41*** *chaîne* ***A*** – ***Cys145*** *chaîne* ***A*** **(a), c)**) et vs **His41** chaîne **B** – **Cys145** chaîne **B** (**b), d)**). En **c)** et **d)** nous avons projeté sur les cadres AMOEBA 15.14μs DES-AMBER 100μs, AMBER 10μs et AMOEBA avec un score de repondération supérieur à 1.

isable, alors que l'échantillonnage de l'algorithme adaptatif est celui responsable de la détection des structures associées à la fois à un site catalytique étroit et à un site de dimérisation allostérique relaxé. Des conclusions similaires peuvent être tirées en considérant **Arg131**, **Asp197** et **Thr199** au lieu de **Glu290**, comme le montre la Figure 2.28. Ces observations démontrent l'importance du couplage de l'algorithme d'échantillonnage adaptatif au PFF AMOEBA pour faire ressortir des conformations qui ont échappé aux simulations MD standard non polarisables.

Puisqu'une certaine connexion allostérique a été trouvée entre la région de dimérisation et les sites actifs, nous avons décidé de fournir une autre vue des différences de simulations observées avec les différents champs de force. Ainsi, nous avons effectué une analyse de matrice de corrélation croisée dynamique (DCCM) [139, 88] pour les trois trajectoires. La DCCM permet d'étudier les changements dynamiques du système au cours du temps et de quantifier les coefficients de corrélation des mouvements entre les atomes.

Le premier résultat à signaler est que comme précédemment, les données AMOEBA diffèrent de celles AMBER/DES-AMBER. DCCM montre plus de valeurs positives/négatives que celles obtenues à partir des n–PFF, indiquant des mouvements d'atomes corrélés/anti-corrélés plus forts dans les simulations PFF (voir la Figure 2.29). Il est à noter que de forts mouvements d'anti-corrélation sont observés au niveau de la région α-hélicoïdale de chaque protomère de la M^{pro} (une région participant fortement à la dimérisation, c'est-à-dire plage de résidus : 220-280 et 470-570) dans les trajectoires AMOEBA. En revanche, les régions correspondantes ont une (anti-)corrélation beaucoup plus faible dans les trajectoires DES-AMBER et AMBER. Le Tableau 2.5 propose une analyse plus fine des régions d'intérêt pour les interactions allostériques (ie le site de dimérisation allostérique) et révèle un mouvement anti-corrélé plus global entre les résidus du site de dimérisation allostérique et la dyade catalytique de la chaîne **A** qu'en AMBER/DES-AMBER. Pour la chaîne **B**, cette corrélation doit dépendre des résidus de la dyade catalytique. Dans tous les cas, la valeur DCCM de corrélation la plus forte est trouvée dans la simulation AMOEBA. La corrélation la plus positive est trouvée pour **Cys145** (chaîne B) et Arg4 (chaîne B) alors que la corrélation la plus négative est trouvée pour **Cys145** (chaîne A) et **Glu290** (chaîne A). Ceci confirme encore la présence d'une corrélation allostérique entre les sites et soutient également l'hypothèse d'une forte asymétrie entre les protomères [93].

2.4.3 Importance des modèles de solvatation et des effets de polarisation dans les interactions allostériques entre les sites

Comme notre analyse précédente a confirmé les différences entre les simulations des différents champs de force, entraînant des prédictions différentes des connexions allostériques et des mouvements corrélés entre les sites, nous avons tenté de retracer les divergences en étudiant la dynamique structurelle globale de l'interface. Comme nous l'avons déjà expliqué, la stabilité globale de l'interface de dimérisation est liée à un réseau complexe de liaisons H qui est exposé au solvant aqueux. Au sein de la M^{pro}, les

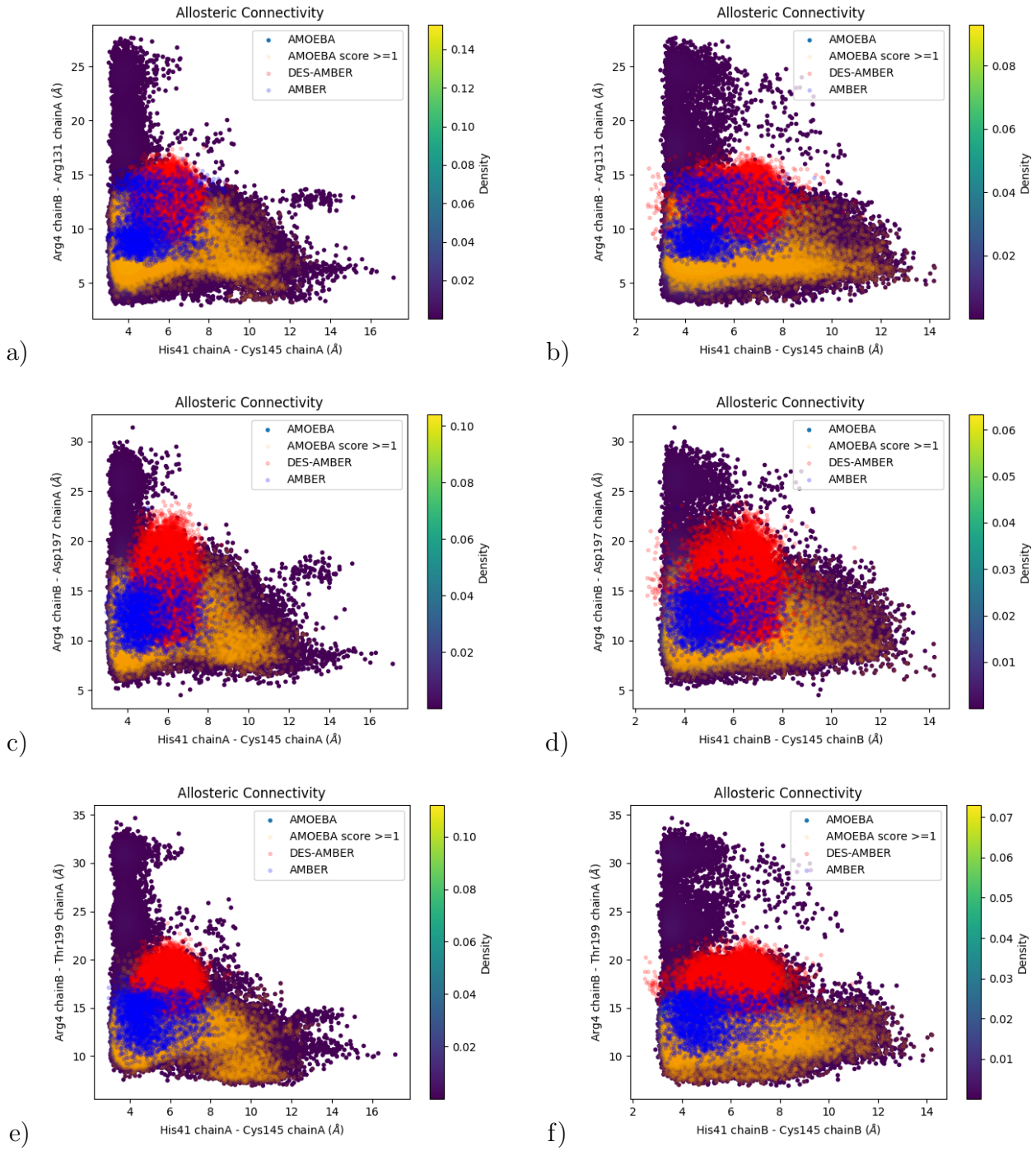
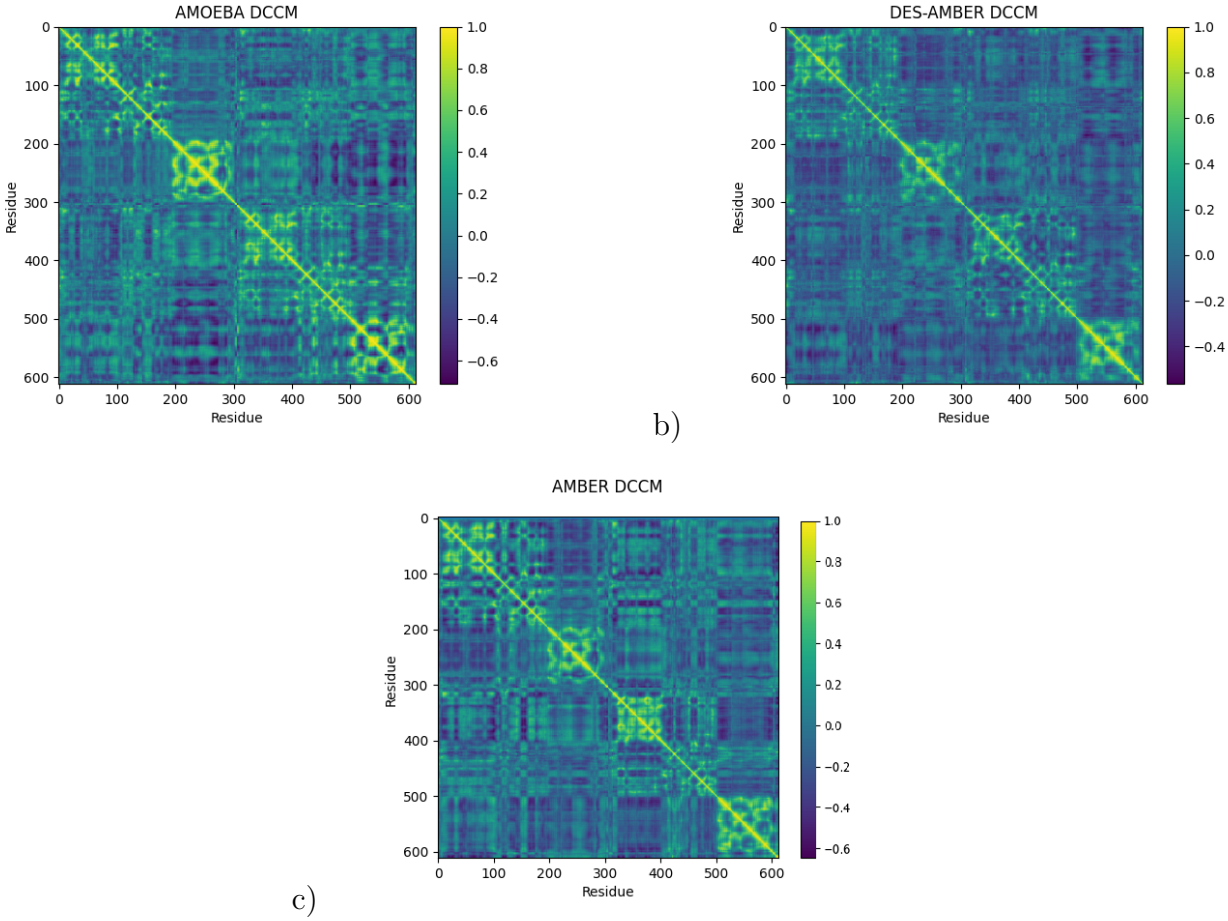


Figure 2.28: Représentation graphique 2D des distances **His41–Cys145** des chaînes A et B par rapport aux distances **Arg4**$_{\mathrm{B}}$–**Arg131**$_{\mathrm{A}}$, **Arg4**$_{\mathrm{B}}$–**Asp197**$_{\mathrm{A}}$ et **Arg4**$_{\mathrm{B}}$–**Thr199**$_{\mathrm{B}}$



a) b) c)

Figure 2.29: Cartes d'intercorrélation dynamique utilisant l'atome C_α de chaque résidu pour a) les trajectoires AMOEBA, b) DES–AMBER et c) AMBER.

fluctuations du volume des cavités et des poches conduisent à un trafic de molécules d'eau qui est essentiel pour maintenir la structure des protéines. En un sens, la connexion allostérique s'effectue "à travers l'eau" et l'analyse qui en résulte de sa présence est donc impactée par la qualité du modèle de l'eau utilisé. En effet, les molécules d'eau se trouvent couramment dans les sites enzymatiques et peuvent former des ponts hydriques entre les résidus et ainsi maintenir les structures secondaires des protéines via des interactions de liaison H (voir la référence [142] et les références qui y sont contenues). En utilisant des champs de force polarisables, il a été démontré que certaines molécules d'eau structurelles présentent des moments dipolaires améliorés, dans les sites actifs de kinase par exemple [54]. Nos travaux sur M^{pro} ont aussi clairement démontré un comportement très différent des molécules d'eau lorsqu'elles sont modélisées avec AMOEBA PFF, qui prend en compte les effets à plusieurs corps [93]. Étant donné que l'eau joue un rôle important dans les activités structurelles et fonctionnelles, nous avons recherché les molécules d'eau présentes autour de certains résidus clés au niveau de l'interface de dimérisation à pH physiologique. Ainsi, nous avons considéré une sphère de rayon 3.5 Å centrée au niveau de l'atome capable d'être engagée dans une liaison hydrogène avec l'eau pour les résidus les plus importants et impliqués dans les interactions non covalentes entre protomères, à savoir : **Arg4**, **Glu290**, **Gly11** et **Glu14**.

Le nombre de molécules d'eau détectées (voir Figure 2.30), présente notamment des profils de distribution différents selon les simulations : eau polarisable AMOEBA, DES-AMBER(TIP4D) et AMBER (TIP3P). En effet, le nombre de molécules d'eau détectées dépend fortement du type de résidu, de la chaîne de la M^{pro} considérée et du champ de force lui-même. **Arg4** de la chaîne **A** par exemple, interagit principalement avec une molécule d'eau pour AMBER, 1-2 molécules pour DES-AMBER et 2-3 molécules pour AMOEBA. Cependant, **Arg4** de la chaîne **B** interagit principalement avec 3 molécules d'eau pour AMBER et DES-AMBER et avec 2 molécules pour AMOEBA conformément à l'asymétrie prédite entre les protomères trouvés dans la M^{pro} [93]. Bien que le trafic d'eau soit détecté pour tous les champs de force, les modèles de solvatation et les différences entre les champs de force semblent dépendre des résidus. Les molécules d'eau extraites des trajectoires AMOEBA autour des résidus concernés sont polarisables (et le modèle de l'eau est flexible [180]) et donc leur distribution est principalement contrôlée par la nature physico-chimique des résidus (polaire, apolaire, chargées positivement/négativement etc...) générant des champs polarisants spécifiques. En pratique, le moment dipolaire moyen de l'eau en vrac AMOEBA s'élève à 2.78 D, en bon accord avec l'expérience, alors que les modèles non PFF présentent des moments dipolaires fixes plus petits de 2.40 D et 2.35 D pour TIP4P-D et TIP3P respectivement. La Figure 2.31 montre les valeurs dipolaires moyennes pour les molécules d'eau à proximité des résidus ciblés. Leurs valeurs moyennes (environ 2.6 D en moyenne) sont inférieures à la valeur de référence globale AMOEBA. Ce résultat est cohérent avec l'idée que l'environnement d'interface dense génère un effet dépolarisant global à plusieurs corps (par rapport à l'eau en vrac) influençant les dipôles induits par les molécules d'eau. Dans l'ensemble, le réseau de liaisons H de l'interface se connecte au propre motif de liaisons H du solvant, formant un

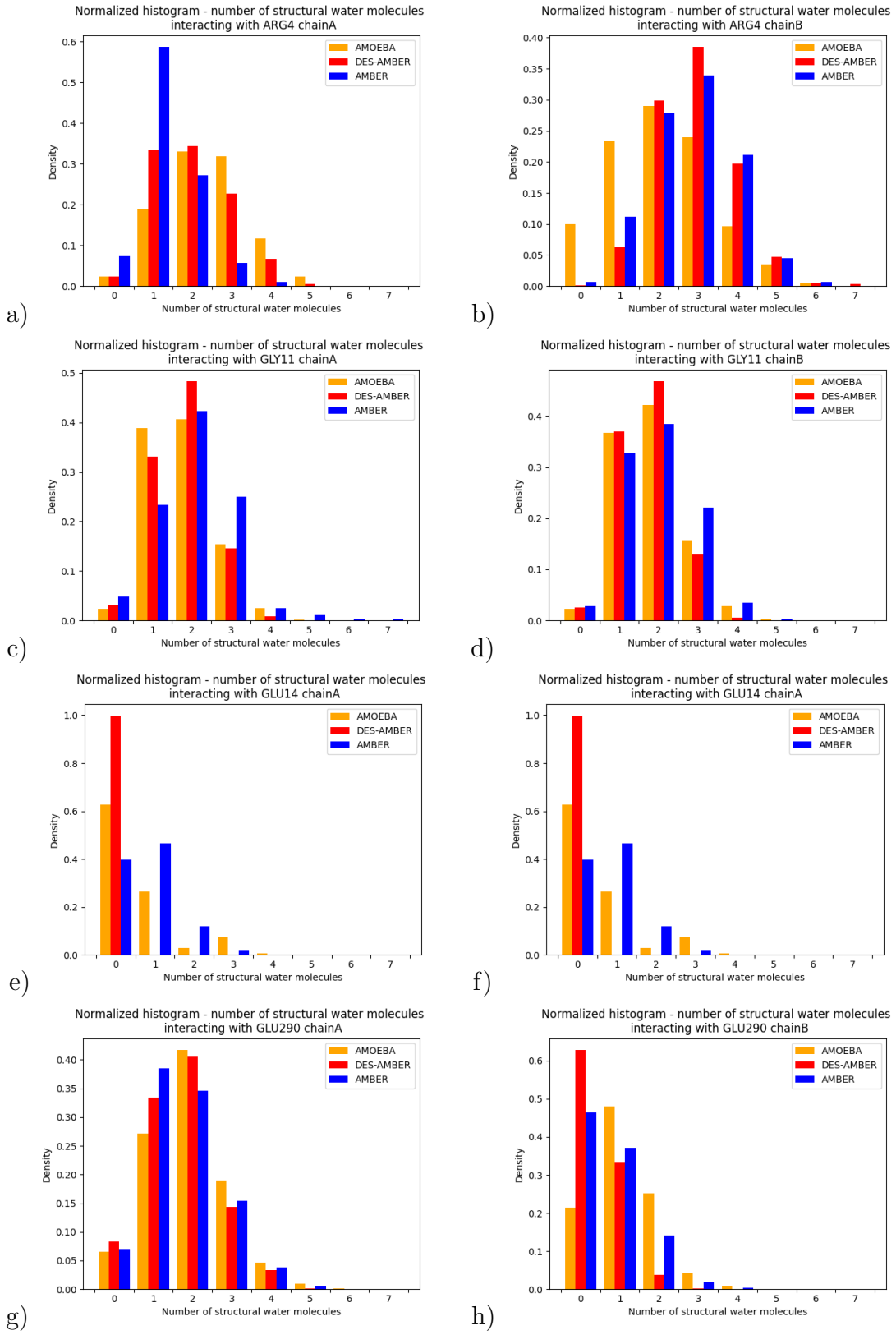


Figure 2.30: Nombre de molécules d'eau détectées dans un rayon de 3.5 Åcentré sur les résidus **Arg4**$_{\mathrm{A}}$, **Arg4**$_{\mathrm{B}}$, **GLY11**$_{\mathrm{A}}$, **Gly**$_{\mathrm{B}}$, **Glu4**$_{\mathrm{A}}$, **Glu4**$_{\mathrm{B}}$, **Glu290**$_{\mathrm{A}}$ et **Glu290**$_{\mathrm{B}}$

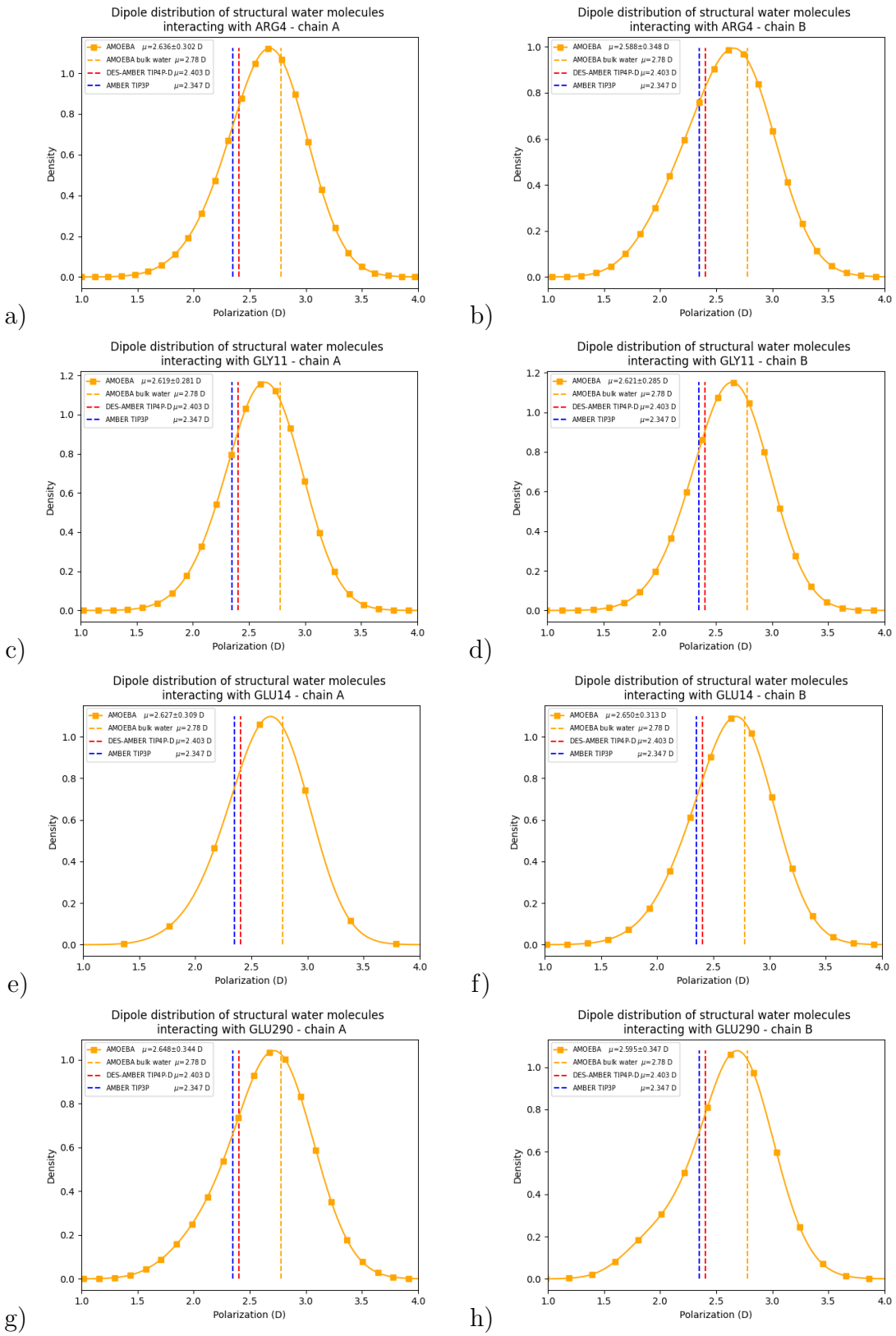
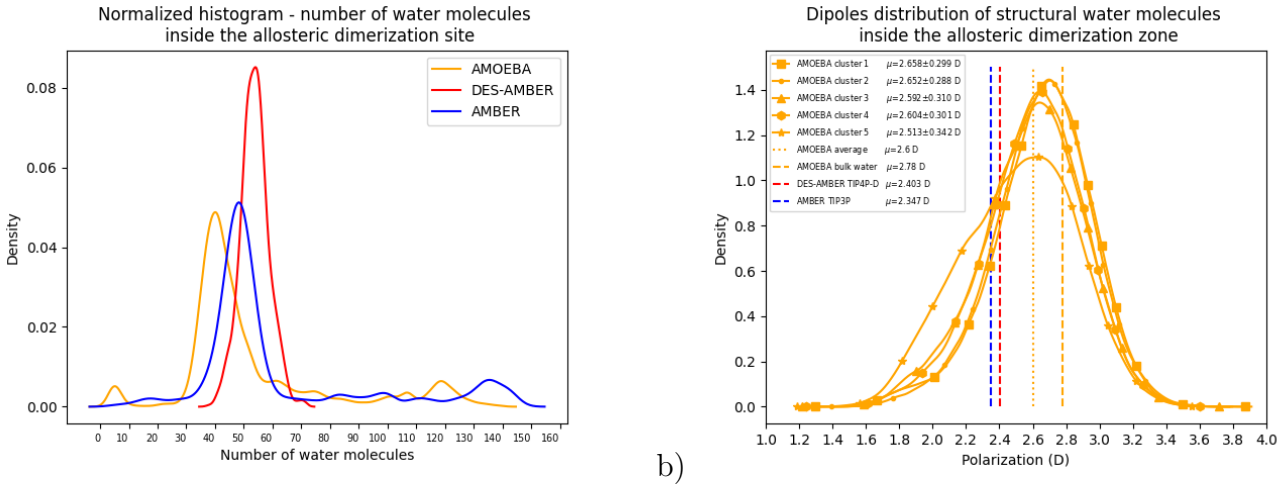


Figure 2.31: Distribution dipolaire des molécules d'eau structurelles interagissant avec $\mathbf{Arg4}_{\mathrm{A}}$, $\mathbf{Arg4}_{\mathrm{B}}$, $\mathbf{GLY11}_{\mathrm{A}}$, $\mathbf{Gly11}_{\mathrm{B}}$, $\mathbf{Glu4}_{\mathrm{A}}$, $\mathbf{Glu4}_{\mathrm{B}}$, $\mathbf{Glu290}_{\mathrm{A}}$ et $\mathbf{Glu290}_{\mathrm{B}}$



a) b)

Figure 2.32: Représentation de **a)** la probabilité du nombre de molécules d'eau structurales à l'intérieur du site de dimérisation allostérique et **b)** leur distribution de dipôles.

niveau de complexité plus élevé. De toute évidence, le comportement des molécules d'eau est fortement influencé par la nature des résidus d'interface par le biais des effets à plusieurs corps, générant divers modèles de microsolvatation en fonction de l'environnement local. Ces modèles sont eux-mêmes affectés par leurs interactions avec le solvant de manière auto-cohérente.

Afin d'évaluer davantage la différence dans les modèles de solvatation, nous nous sommes concentrés sur le site de dimérisation allostérique précédemment introduit, un emplacement spécifique au sein de l'interface qui permet aux molécules d'eau de circuler entre les résidus d'interface. Pour mieux comprendre ce qui se passe, nous devons évaluer le nombre de molécules d'eau présentes et leurs durées de vie au sein de ce site. Il est important de mentionner ici que les six résidus formant le site allostérique à l'interface de dimérisation sont soit ioniques soit polaires. **Asp** et **Glu** sont chargés négativement alors que **His** est chargé positivement. Les chaînes latérales telles que **Thr** peuvent retenir les molécules d'eau à l'intérieur de la cavité. Les flèches noires de la Figure 2.35 montrent le flux de molécules d'eau dans le site enfoui. Puisque la plus grande distance séparant **Arg4** chaîne **B** et **Glu290** chaîne **A** est d'environ 24 Å, nous avons défini une sphère avec un rayon (coupure) de 10Å, centré au centre géométrique des six résidus formant la poche au site de dimérisation allostérique, et calculé le nombre de molécules d'eau présentes au sein de cette sphère. La Figure 2.32, a) montre une différence frappante entre les simulations AMOEBA et non PFF. Les simulations PFF donnent beaucoup moins de molécules d'eau à l'intérieur du site de dimérisation allostérique et une densité de probabilité de présence la plus élevée centrée à 40, à comparer à 50 pour AMBER et 55 pour DES-AMBER.

Nous avons ensuite mesuré les durées de vie de l'eau dans la sphère 10 Å en utilisant les simulations cMD de 400 ns produites avec les champs de force AMBER et AMOEBA. Comme montré sur la Figure 2.33, nous observons une durée de vie moyenne de l'eau de 0.171 ns pour AMBER et une durée de vie plus longue ; de 0.516 ns pour AMOEBA. Cela montre clairement que les effets de

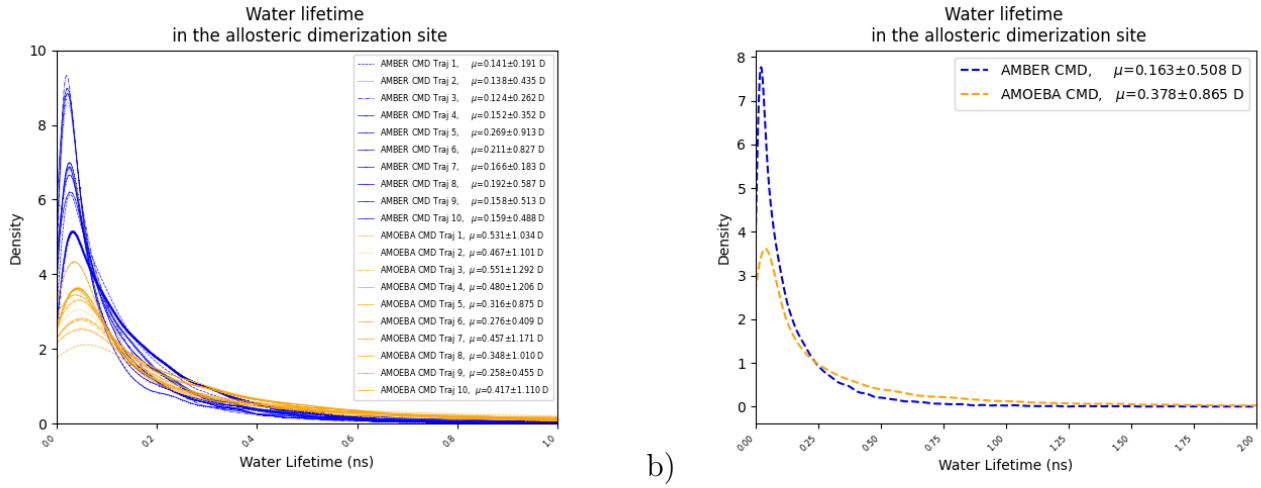


Figure 2.33: Répartition de la durée de vie de l'eau à l'intérieur du site de dimérisation allostérique pour les trajectoires AMBER et AMOEBA : 20 cMD (400 ns x 2), a). La distribution de la durée de vie de l'eau pour toutes les trajectoires AMBER et AMOEBA cMD, b).

polarisation à plusieurs corps ont tendance à agir comme une colle entre l'interface de dimérisation et les molécules d'eau, en particulier sur le site de dimérisation allostérique, les retenant plus longtemps à la surface des résidus du site de dimérisation. La combinaison de ces deux résultats nous permet de mieux comprendre pourquoi la dynamique de l'eau à l'extérieur de l'interface est si différente de la dynamique (plus lente) trouvée dans la partie la plus confinée du site allostérique de dimérisation. Le plus petit nombre de molécules d'eau à l'intérieur du site de dimérisation allostérique reflète donc un trafic d'eau plus lent, car ces molécules d'eau polarisées ont tendance à se déplacer lentement, étant engagées dans beaucoup plus de liaisons H. En effet, la constante de diffusion AMOEBA est plus conforme à l'expérience que les modèles TIP3P et TIP4-D. Cependant, comme nous l'avons vu, les valeurs du moment dipolaire de l'eau AMOEBA peuvent présenter de fortes variations locales en raison des modèles de microsolvatation locaux qui ne peuvent pas être capturés par l'approximation du champ moyen qui est à la base des non-PFF classiques [142]. Comme pour la situation précédente, la Figure 2.34 montre une situation globale plutôt sous-polarisée pour l'eau qui présente un moment dipolaire moyen inférieur au volume. Néanmoins, la Figure 2.34 met également en évidence la collection de multiples situations différentes où les modèles de microsolvatation ont tendance à générer simultanément des distributions partielles de molécules d'eau hautement polarisées et sous-polarisées dans le site de dimérisation allostérique puisque cette distribution est principalement contrôlée par la nature physico-chimique des résidus.

Comme le montre la Figure 2.32, c) et la Figure 2.34, la plupart des molécules d'eau sous-polarisées se trouvent dans la section la plus enfouie du site de dimérisation allostérique où le confinement génère plus d'effets dépolarisants. Ceux-ci sont bien connus pour diminuer les valeurs moyennes des moments dipolaires des eaux confinées et sont observés ici. Encore une fois, AMOEBA présente une densité de probabilité plus élevée inférieure à la masse à 2.60 D alors que les dipôles d'eau DES-AMBER et

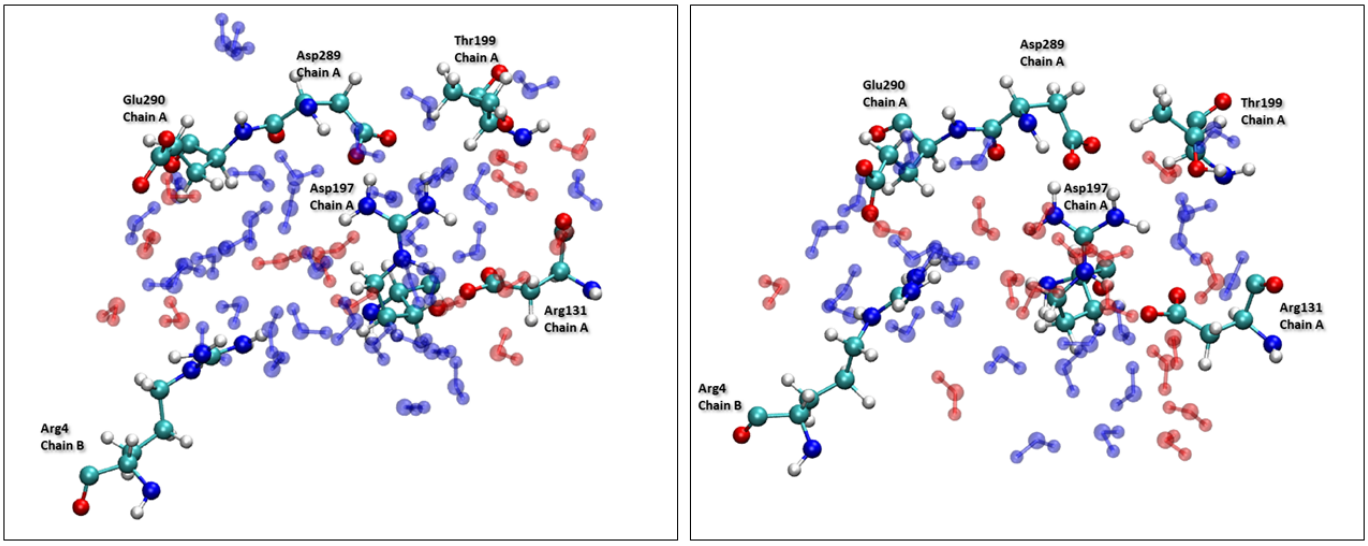


Figure 2.34: Représentation de la distribution des dipôles d'eau à l'intérieur du site de dimérisation allostérique, avec 5.29Å entre **Arg4** et **Glu290** (gauche) et 8.7Å (droite). Les molécules d'eau recouvertes de rouge ont un moment dipolaire $\geq$ 2.78 D, celles recouvertes de bleu ont un moment dipolaire $\leq$ 2.78 D. Asp et Glu ont des chaînes latérales chargées électriquement (acide). Arg ont des chaînes latérales chargées électriquement (de base). Thr a une chaîne latérale polaire.

AMBER restent fixés respectivement à 2.40 D et 2.35 D respectivement (voir Figure 2.32, b)). Les Figures 3.34b) donnent également une vue des moments dipolaires moyens trouvés après clusterisation des trajectoires AMOEBA (voir [93] pour plus d'informations sur les 5 différents clusters). Le site maintient une valeur de solvant dipolaire moyenne relativement stable grâce à la fluctuation à la fois des volumes (c'est-à-dire différents dans les différents clusters) et du nombre de molécules d'eau (voir Figure 2.36), mettant en évidence l'interconnexion de l'interface H-bond network et le solvant. Cela suggère qu'il existe une interaction complexe entre la distribution des dipôles des molécules d'eau polarisables et les résidus (et les volumes associés) du site allostérique de dimérisation. Ce réseau d'interaction contribue à réguler les effets allostériques avec le site catalytique des deux protomères. La modélisation de telles connexions entre les cavités nécessite de capturer l'équilibre subtil entre la dynamique des protéines et des solvants. Les fluctuations dipolaires du trafic d'eau ont tendance à être extrêmement complexes, conduisant à un comportement radicalement différent dans différentes parties de l'interface où la dynamique locale de l'eau peut être très différente (c'est-à-dire, pour AMOEBA, un ralentissement dynamique prédit dans le site de dimérisation allostérique enfoui , etc...). Un tel trafic d'eau façonne l'interface et participe à la modulation de la "respiration" structurelle du site de dimérisation allostérique qui est impliquée dans les effets allostériques globaux avec le site catalytique principal. Une telle implication critique des molécules d'eau "polarisables" dans les sites de reconnaissance ou de régulation des protéines avait été montrée antérieurement [54] et il est clair que le nombre de molécules d'eau dans un site de liaison est important. En effet, les molécules d'eau interagissant avec leur environnement proche via des modes de liaison à travers l'eau ("through

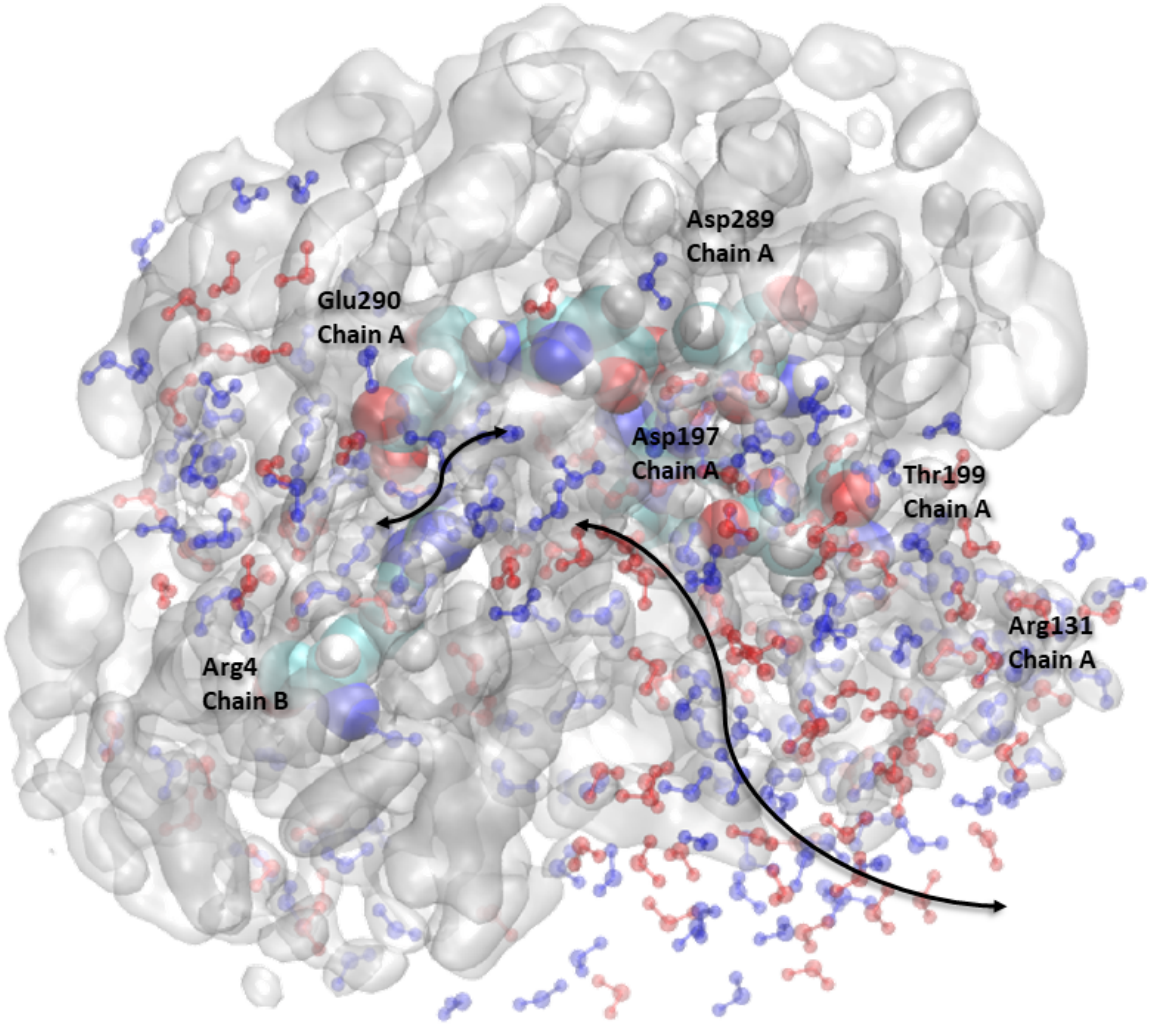
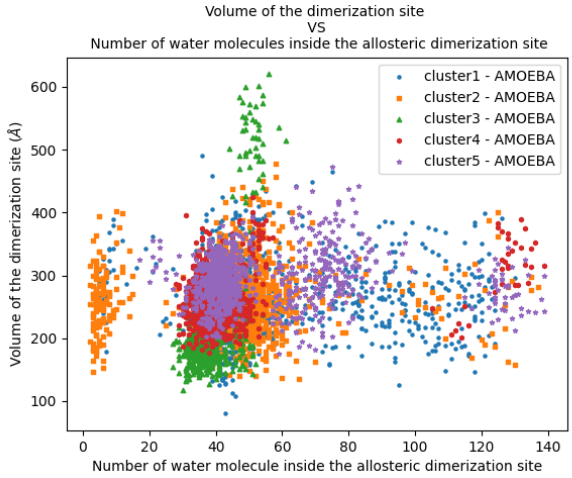


Figure 2.35: Représentation de la distribution des dipôles d'eau à l'intérieur du site de dimérisation allostérique. Les molécules d'eau recouvertes de rouge ont un moment dipolaire $\leq$ 2.78 D, celles recouvertes de bleu ont un moment dipolaire $\geq$ 2.78 D. Asp et Glu ont des chaînes latérales chargées négativement (acide). Arg ont des chaînes latérales chargées positivement (base). Thr a une chaîne latérale polaire. La distance entre **Arg4** et **Glu290** est de 5.29Å. Les résidus à moins de 10Å du site de dimérisation allostérique sont présentés en mode quicksurf en blanc. Les flèches noires montrent le flux des molécules d'eau dans ce site.

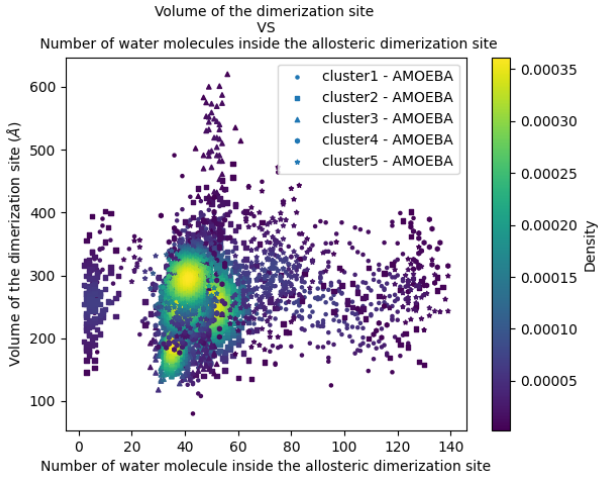
water") sont courantes et capables d'influencer fortement les propriétés électroniques locales [53]. Les configurations à travers l'eau peuvent médier les interactions entre un inhibiteur (voir par exemple [54, 154]) et les résidus liés indirectement du site de reconnaissance. Dans de telles situations, également considérées dans le contexte des PFF, un décompte précis des molécules d'eau peut être critique car les effets à plusieurs corps (en particulier l'énergie de polarisation) pourraient faire basculer l'équilibre énergétique (libre) entre les inhibiteurs concurrents. L'absence de cet aspect dans la modélisation entraîne certainement une perte dans la prédiction du signal dans la communication allostérique. Il est également important de mentionner qu'au-delà de cette vision énergétique du phénomène, le lien entre les molécules d'eau inter-faciales et la dynamique/flexibilité des protéines a été largement discuté dans la littérature expérimentale (voir les références [48, 22, 45] et les références qui y figurent): la dynamique des protéines et la dynamique des couches de solvatation ont été caractérisées au niveau régional. Plus précisément, il a été observé que les régions flexibles des protéines englobent généralement les eaux à mouvement rapide, tandis que les régions stables sont intégrées dans les molécules d'eau de la couche d'hydratation plus lente. C'est exactement ce que nous voyons ici et ce qui est nouveau dans nos résultats, c'est qu'il est démontré qu'une telle modélisation de la dynamique régionale est fortement affectée par les effets de polarisation. En effet, ils influencent fortement la dynamique des molécules d'eau inter-faciales agissant sur leur "viscosité" locale et donc la dynamique locale. Comme les poches de liaison et les sites allostériques doivent être raisonnablement stables dans le temps pour être ciblés par les médicaments, dans certaines situations, les simulations sans PFF peuvent avoir tendance à prédire les modèles de solvatation associés à un trafic d'eau excessif et à des molécules inter-faciales se déplaçant trop rapidement. Cela pourrait malheureusement conduire à la déstabilisation des régions ciblés par les molécules à effets thérapeutiques potentiels ou à surpasser des régions stables qui resteraient inconnus par les modélisateurs moléculaires.

2.5 Conclusions

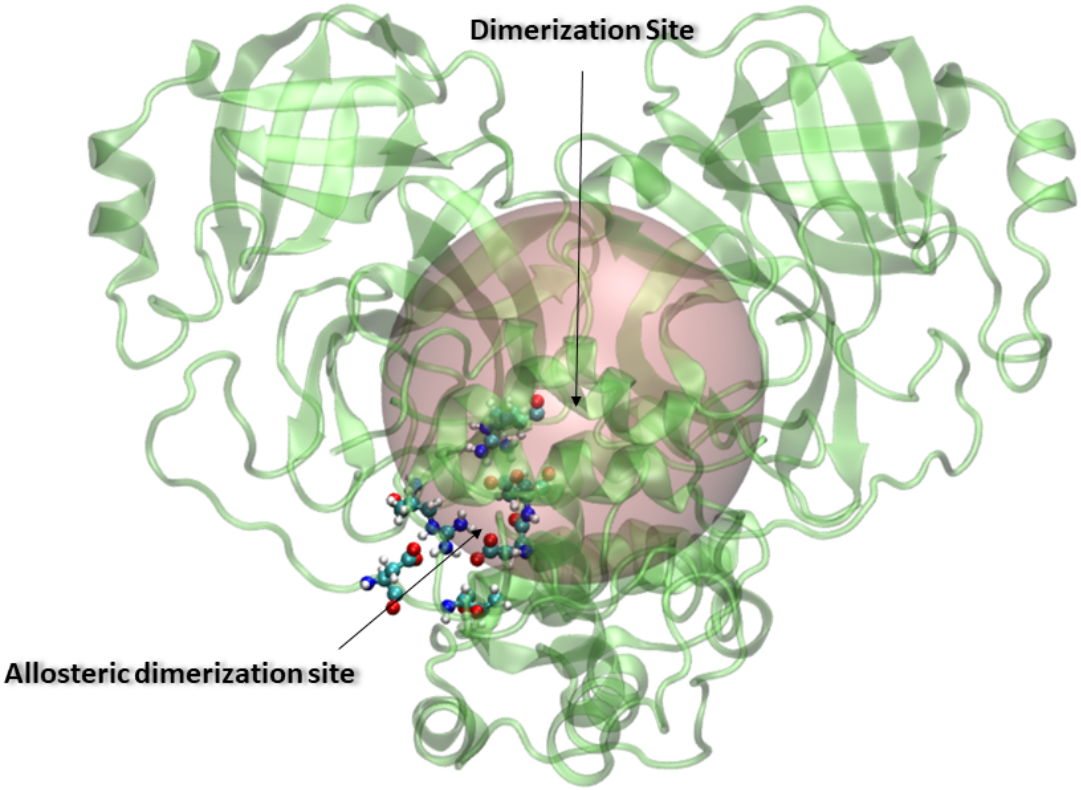
Pour conclure, afin de proposer un modèle de haute qualité de l'interface de dimérisation de la M^{pro} du SARS-CoV-2 qui pourrait être utilisé pour la conception ultérieure de médicaments, il est important de bien comprendre et modéliser son réseau complexe de liaisons H qui est intégré dans un réseau dipolaire dynamique du solvant. L'eau semble être un acteur clé dans la dynamique structurelle globale de l'interface de dimérisation, étant un élément constitutif des effets allostériques globaux entre les sites de la protéine via des interactions de polarisation à plusieurs corps avec les résidus de l'interface de dimérisation. Comme nous l'avons souligné, la protéase principale ; M^{pro} du SARS-CoV-2 est un système moléculaire difficile et complexe qui nécessite la capacité simultanée: i) de décrire avec précision tous les types d'interactions non covalentes au sein de la protéine et du solvant nécessitant donc un champ de force précis capable de décrire les effets locaux de polarisation ; et ii) effectuer un échantillonnage étendu allant au-delà de l'échelle de temps des μs. Bien entendu, nous n'avons analysé



a)



b)



c)

Figure 2.36: Représentation graphique 2D du volume du site de dimérisation en fonction du nombre de molécules d'eau à l'intérieur du site de dimérisation allostérique (pas de repondération), a). La densité de chacun des clusters est indiquée dans **b)**. **c)** Représentation schématique du dimère de la M^{pro} du SARS-CoV-2. Le site de dimérisation est présenté en rose et les résidus du site de dimérisation allostérique sont présentés en CPK.

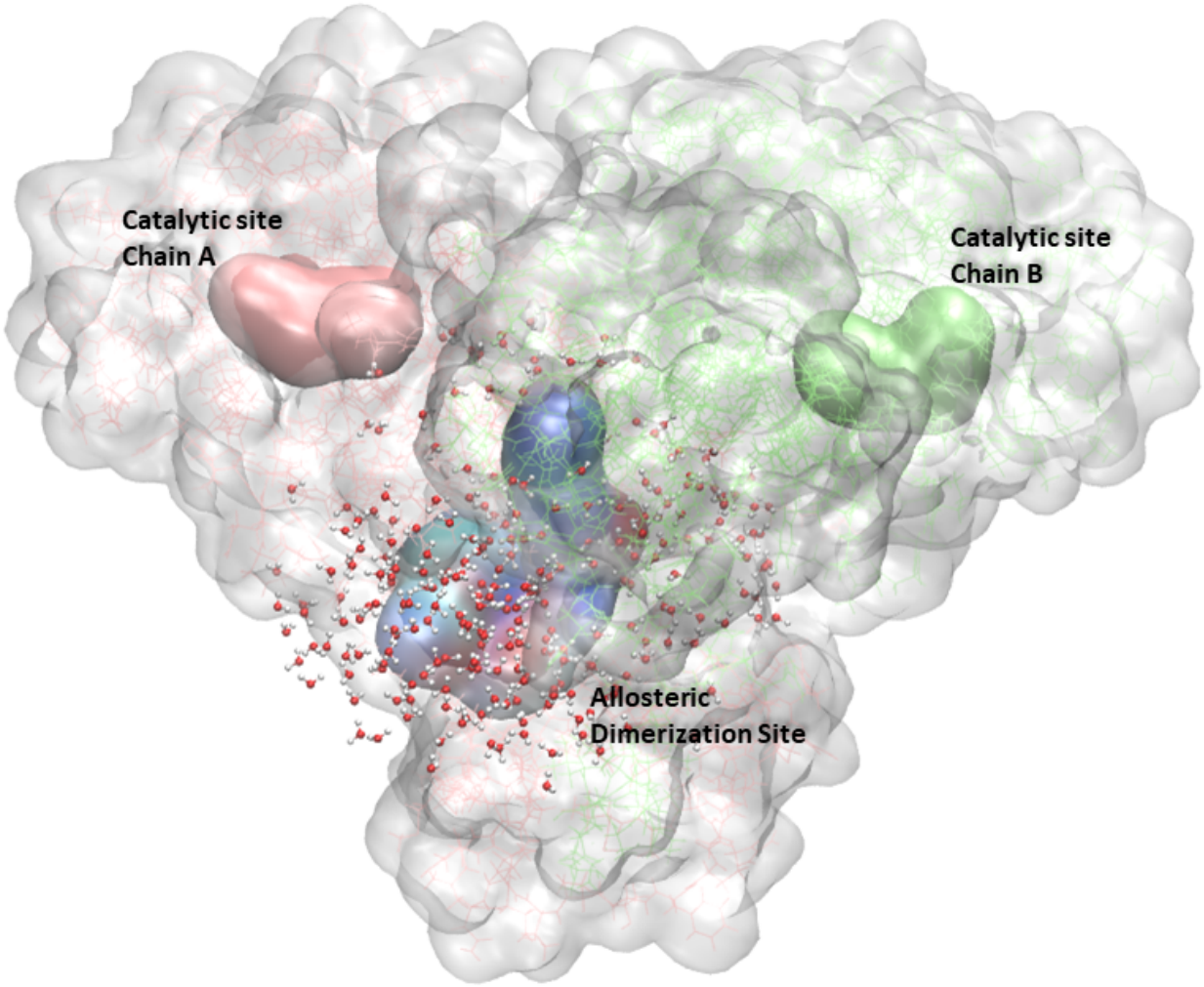


Figure 2.37: Vue globale de la M^{pro} du SARS-CoV-2, montrant le site catalytique des deux chaînes **A** et B et le site de dimérisation allostérique. Les molécules d'eau à moins de 10Å du site de dimérisation allostérique sont présentées en mode CPK.

ici qu'un exemple des interactions allostériques au sein de la M^{pro} et bien d'autres resteront peut-être à découvrir : nous espérons que ces analyses et trajectoires de dynamique moléculaire (disponibles via le référentiel BioExcel/MolSSI) aideront les chasseurs de molécules à effets thérapeutique ciblant l'interface de dimérisation de la M^{pro}. Un projet qui porte sur la conception de ligands inhibiteurs qui ciblent la M^{pro} a été débuté. Par ailleurs, des premières simulations ont été faites également sur la protéine S "Spike" sous sa forme trimérique.

Cross-Correlation		AMOEBA	DES-AMBER	AMBER
Residues of the allosteric dimerization site	Arg4 ChainB - Glu290 ChainA	-0.016	0.124	-0.227
	Arg4 ChainB - Arg131 ChainA	-0.037	0.146	-0.345
	Arg4 chainB - Asp197 chainA	0.009	0.043	-0.229
	Arg4 ChainB - Thr199 ChainA	0.048	0.082	-0.255
	Arg4 ChainB - Asp289 ChainA	-0.018	0.155	-0.168
Arg4 ChainB vs Catalytic Dyad	Arg4 ChainB - His41 ChainA	-0.224	-0.190	-0.184
	Arg4 ChainB - Cys145 ChainA	0.013	0.034	-0.071
	Arg4 ChainB - His41 ChainB	-0.259	-0.081	-0.170
	Arg4 ChainB - Cys145 ChainB	0.439	0.323	0.275
Glu290 ChainA vs Catalytic Dyad	Glu290 ChainA - His41 ChainA	-0.169	-0.196	-0.095
	Glu190 ChainA - Cys145 ChainA	-0.355	-0.035	-0.142
	Glu290 ChainA - His41 ChainB	-0.004	-0.044	-0.189
	Glu190 ChainA - Cys145 ChainB	-0.044	0.172	0.074
Arg141 ChainA vs Catalytic Dyad	Arg131 ChainA - His41 ChainA	0.049	-0.126	0.090
	Arg131 ChainA - Cys145 ChainA	0.129	-0.029	0.045
	Arg131 ChainA - His41 ChainB	-0.012	0.089	-0.032
	Arg131 ChainA - Cys145 ChainB	-0.066	0.072	0.016
Asp197 ChainA vs Catalytic Dyad	Asp197 ChainA - His41 ChainA	0.010	-0.092	0.035
	Asp197 ChainA - Cys145 ChainA	0.000	-0.246	0.104
	Asp197 ChainA - His41 ChainB	0.139	0.097	0.125
	Asp197 ChainA - Cys145 ChainB	0.058	-0.067	0.010
Thr199 ChainA vs Catalytic Dyad	Thr199 ChainA - His41 ChainA	-0.151	-0.230	-0.111
	Thr199 ChainA - Cys145 ChainA	-0.211	-0.286	-0.187
	Thr199 ChainA - His41 ChainB	-0.010	0.018	-0.024
	Thr199 ChainA - Cys145 ChainB	0.075	0.032	0.045
Asp289 ChainA vs Catalytic Dyad	Asp289 ChainA - His41 ChainA	-0.181	-0.139	-0.116
	Asp289 ChainA - Cys145 ChainA	-0.299	-0.084	-0.192
	Asp289 ChainA - His41 ChainB	0.050	0.014	-0.178
	Asp289 ChainA - Cys145 ChainB	-0.051	0.138	0.077

Table 2.5: Valeurs extraites des cartes de corrélation croisée dynamiques. La valeur maximale (0.429) correspond à la corrélation entre **Arg4** chaîne B et **Cys145** chaîne B et la valeur minimale (-0.355) correspond à la correlation entre **Glu290** chaîne A et **Cys145** chaîne A

.

2.6 Publications

2.6.1 Publication 1

Chemical Science

ROYAL SOCIETY OF CHEMISTRY

EDGE ARTICLE

View Article Online
View Journal | View Issue

Check for updates

Cite this: *Chem. Sci.*, 2021, **12**, 4889

All publication charges for this article have been paid for by the Royal Society of Chemistry

High-resolution mining of the SARS-CoV-2 main protease conformational space: supercomputer-driven unsupervised adaptive sampling†

Théo Jaffrelot Inizan, ‡[a] Frédéric Célerse, ‡[ab] Olivier Adjoua,[a] Dina El Ahdab, [ac] Luc-Henri Jolly,[d] Chengwen Liu,[e] Pengyu Ren,[e] Matthieu Montes,[f] Nathalie Lagarde,[f] Louis Lagardère,*[ad] Pierre Monmarché*[ag] and Jean-Philip Piquemal *[aeh]

We provide an unsupervised adaptive sampling strategy capable of producing μs-timescale molecular dynamics (MD) simulations of large biosystems using many-body polarizable force fields (PFFs). The global exploration problem is decomposed into a set of separate MD trajectories that can be restarted within a selective process to achieve sufficient phase-space sampling. Accurate statistical properties can be obtained through reweighting. Within this highly parallel setup, the Tinker-HP package can be powered by an arbitrary large number of GPUs on supercomputers, reducing exploration time from years to days. This approach is used to tackle the urgent modeling problem of the SARS-CoV-2 Main Protease (M^{pro}) producing more than 38 μs of all-atom simulations of its apo (ligand-free) dimer using the high-resolution AMOEBA PFF. The first 15.14 μs simulation (physiological pH) is compared to available non-PFF long-timescale simulation data. A detailed clustering analysis exhibits striking differences between FFs, with AMOEBA showing a richer conformational space. Focusing on key structural markers related to the oxyanion hole stability, we observe an asymmetry between protomers. One of them appears less structured resembling the experimentally inactive monomer for which a 6 μs simulation was performed as a basis for comparison. Results highlight the plasticity of the M^{pro} active site. The C-terminal end of its less structured protomer is shown to oscillate between several states, being able to interact with the other protomer, potentially modulating its activity. Active and distal site volumes are found to be larger in the most active protomer within our AMOEBA simulations compared to non-PFFs as additional cryptic pockets are uncovered. A second 17 μs AMOEBA simulation is performed with protonated His172 residues mimicking lower pH. Data show the protonation impact on the destructuring of the oxyanion loop. We finally analyze the solvation patterns around key histidine residues. The confined AMOEBA polarizable water molecules are able to explore a wide range of dipole moments, going beyond bulk values, leading to a water molecule count consistent with experimental data. Results suggest that the use of PFFs could be critical in drug discovery to accurately model the complexity of the molecular interactions structuring M^{pro}.

Received 10th January 2021
Accepted 27th January 2021

DOI: 10.1039/d1sc00145k

rsc.li/chemical-science

1 Introduction

At the end of December 2019, a novel coronavirus (CoV) that induces severe acute respiratory disease (SARS) was discovered and labeled SARS-CoV-2.[1] It causes the disease named COVID-19, which led to a global pandemic in 2020 and finally to an urgent global issue.

Great effort has been made to gain insights into the action of the virus on the human body. As the genome of the virus has been rapidly determined,[2] a similarity between the SARS-CoV-2 virus and the older SARS-CoV (2003) and Middle East respiratory syndrome coronavirus (MERS-CoV in 2012) was observed. Besides vaccines, researchers started the hunt for small molecules to treat the disease. Rapidly,[2] different classes of proteins have been experimentally characterized that could be useful

[a]Sorbonne Université, LCT, UMR 7616 CNRS, Paris, France. E-mail: louis.lagardere@sorbonne-universite.fr; pierre.monmarche@sorbonne-universite.fr; jean-philip.piquemal@sorbonne-universite.fr
[b]Sorbonne Université, IPCM, UMR 8232 CNRS, Paris, France
[c]Université Saint-Joseph de Beyrouth, UR-EGP Faculté des Sciences, Lebanon
[d]Sorbonne Université, IP2CT, FR 2622 CNRS, Paris, France
[e]University of Texas at Austin, Department of Biomedical Engineering, Texas, USA
[f]Laboratoire GBCM, EA 7528, CNAM, Hésam Université, Paris, France
[g]Sorbonne Université, LJLL, UMR 7598 CNRS, Paris, France
[h]Institut Universitaire de France, Paris, France

† Electronic supplementary information (ESI) available. See DOI: 10.1039/d1sc00145k

‡ These authors contributed equally to this work.

targets for drugs. Among the different classes of proteins that have been experimentally characterized, the main protease[3] is essential for processing the precursor polyprotein for the replication of the virus. Indeed, proteases are responsible for activating viral proteins for particle assembly. Due to their importance within the replication cycle of the virus, they have been proven to be successful targets for antiviral agents and are used to treat many diseases including HIV and hepatitis.[4] In the case of SARS-CoV-2, the main protease is called M^{pro} or $3CL^{pro}$. Many efforts have been made to refine the crystallographic structure of M^{pro} as the number of experimental structures available in the Protein Data Bank is increasing. While more than one hundred M^{pro} structures exist and massive efforts to discover a successful inhibitor are underway, computational approaches involving virtual screening and Molecular Dynamics (MD) simulations are needed to help experimentalists to *in silico* optimize their millions of test molecules.[5–8]

Molecular Dynamics is a powerful tool for understanding the structural and dynamical details of complex biological systems. It also enhances the ability to identify promising protein inhibitors. Two main research groups, DE Shaw Research (DESRES) and RIKEN Center for Biosystems Dynamics Research, recently released multi-microsecond MD simulations of the M^{pro} dimer.[5,6] These MD conformational ensembles both used non-polarizable force fields (n-PFFs) including DES-AMBER[9] and AMBER14ff.[10] Although the simulations are of great help for the scientific community, conventional MD (cMD) simulation results are limited by the daunting complexity of M^{pro}'s conformational space, which requires very large computational resources. In practice, both DESRES and RIKEN results were obtained on special-purpose petascale supercomputers designed for MD (Anton[11] and MD-GRAPE-4A[12] for DESRES and RIKEN, respectively). So, what can be done next? Besides these large scale MD simulations, the question of accuracy still remains open. Indeed, conformational space sampling depends by definition also on the force field used for the simulations. Our group has been involved for many years in the demonstration of the importance of considering explicit many-body effects in classical MD and free energy methods through the use of polarizable force fields (PFFs).[13–17] Indeed, electronic polarization affects solvation and modifies the stability of secondary and quaternary structures of proteins, playing therefore a crucial role in defining the conformational space of a protein. Applying such methods to COVID-19 research could provide additional insights for drug modelers and experimental teams. When our project started (end of March 2020) in response to the international High-Performance Computing (HPC) global effort to mitigate the impact of the COVID-19 pandemic,[18–20] performing long timescale MD simulations using new generations of PFFs on SARS-CoV-2 proteins encompassing hundreds of thousands of atoms (or more), such as M^{pro}, was out of reach of generalist supercomputers. Such simulations would have required years of computation.

To overcome these limitations we introduce a density-driven unsupervised adaptive sampling method based on statistical models and principal component analysis (PCA). It has been deployed on a generalist supercomputer. Since the global exploration problem is decomposed into a set of separate MD trajectories, the process can be restarted using an iterative selection method, and various computations can take place on a large number of Graphics Processing Units (GPUs) that are now available in generalist supercomputers. Such a strategy enables the Tinker-HP package,[21] which recently proposed a GPU-accelerated implementation,[22] to perform multi-microsecond MD simulations within a few days, where years would have been required with single GPU card or CPU-based conventional MD simulations. We additionally provide the capability to re-weight our simulations, which enables full exploitation of the total amount of MD trajectories to compute statistical properties that can therefore benefit from the long simulations. After describing our sampling strategy, we will detail our conformational space exploration results that notably expand over those obtained by other groups. We will unveil critical structural behavior not fully captured with n-PFFs. We particularly investigated the differences in clustering results, active site volumes, cryptic pockets, key structural activation markers linked to the oxyanion hole structuring, interactions between the C-terminal chain and the active site, and solvation patterns of some key residues. The effect of pH is also discussed.

2 Unsupervised adaptive sampling strategy for exploration: exploiting pre-exascale machines and GPUs

Adaptive sampling has been used for many years and has proven to be a powerful exploration tool to study protein folding and dynamics, ligand binding and a variety of rare molecular events.[23–26] For this family of approaches, multiple iterations of independent molecular dynamics simulations are performed, basing the initial conditions at each iteration on the results of previous iteration steps. We propose here a new unsupervised (*i.e.* fully automated) adaptive sampling strategy dedicated to our specific use of PFFs within large supercomputer systems allowing for the simultaneous use of hundreds or thousands of GPU cards. This characteristic is important as it allows us to benefit from the full potential of pre-exascale supercomputers, and will naturally transfer to future exascale machines. The results presented here benefit from a GPU acceleration in the newly developed Tinker-HP GPU code[22] that was first used here for COVID-19 simulations. However the procedure is completely general and can be applied to any homogeneous or heterogeneous computational platforms compatible with Tinker-HP[21,27] or any MD software. Therefore, in view of the particular distribution of available numerical resources, the simulations are organized by iterations as follows. At the beginning of each iteration, some initial structures are selected among the configurations sampled in the past iterations, from which independent MD simulations are run, generating new configurations. The selection of the initial structures at each iteration follows an adaptive procedure designed to enhance the exploration of a low-dimensional space of slow variables.

More precisely, M_k denotes the number of configurations available at the beginning of iteration $k \geq 0$, and $(q_i)_{1 \leq i \leq M_k}$ the configurations. Here, a configuration means the positions $q \in \mathbb{R}^{3N}$ of all the atoms of the system. In particular, at the very beginning of the algorithm, we suppose that we start with $M_0 \geq 1$ configurations, obtained from an initial conventional MD simulation (which is in practice non-polarizable), or previously available studies. At the beginning of iteration k, first, the protein is aligned in all configurations, using the backbone atoms of the 6LU7 crystal structure from the Protein Data Bank.[3] A principal component analysis (PCA)[28] is then performed, using the scikit-learn[29] and MDTraj[30] packages, on the protein atoms $(q_i)_{1 \leq i \leq M_k}$, from which the $n = 4$ principal modes are considered. This choice was made after a global analysis of the first 20 PCA modes of the first AMOEBA 0.14 μs which showed that $n > 4$ modes had variance contributions below 4% (Fig. 1, ESI†). This has also been corroborated by an analysis of RIKEN and DESRES trajectories, for which, respectively, 3 and 4 PCA modes are above 4% (Fig. 2, ESI†). We denote by $\xi_k : \mathbb{R}^{3N} \to \mathbb{R}^n$ the orthogonal projection on these n principal modes and we write $x_i = \xi_k(q_i)$. At the beginning of iteration k, this represents the current guess of slow variables of the system, and in order to enhance the sampling, we would like to explore all the values of these slow variables. In other words, ideally, we would like the values of x sampled to be uniformly distributed over some compact set of $\mathbb{R}^n$. The selection procedure is designed to push the exploration in the direction of this ideal target.

The density ρ_k of the collective variables is approximated by a Gaussian kernel, *i.e.* for $x \in \mathbb{R}^n$

$$\rho_k(x) = \frac{1}{(2\pi\sigma^2)^{n/2} M_k} \sum_{i=1}^{M_k} \exp\left(-\frac{|x - x_i|^2}{2\sigma^2} \right),$$

for some $\sigma > 0$. In practice we used the D.W. Scott method, implemented in Scipy,[31] to estimate a suitable bandwidth σ. Denoted by s_k is the number of MD trajectories that are going to be run during iteration k. In order to select the initial structures $(q_{I_1}, \ldots, q_{I_{s_k}})$ of these simulations, the indexes $I_1, \ldots, I_{s_k}$ are generated as independent random variables in $\{1, \ldots, M_k\}$ distributed according to

$$\mathbb{P}(I = i) = \frac{\rho_k^{-1}(x_i)}{\sum_{j=1}^{M_k} \rho_k^{-1}(x_j)}.$$

In other words, among all the structures currently available, q_i is selected to be the initial structure of a new simulation with a probability inversely proportional to its density (in the low-dimensional space given by the first four PCA components). The effect of this selection can intuitively be illustrated as follows: if two domains of similar size (in the sense of the Lebesgue measure on $\mathbb{R}^n$) have been visited, with one that concentrates most of the past trajectories while the other contains only a few points, then approximately half of the new initial structures will be selected in each domain; in contrast, a uniform selection among the past configurations would have put much more weight on the dense domain.

From the initial structures $(q_{I_1}, \ldots, q_{I_{s_k}})$, s_k independent MD simulations are sampled, and the state of each simulation is recorded every 0.1 ns (the initial structure is not recorded, since it has already been recorded in one of the past iterations). Here, independent means that the initial velocities (sampled according to the equilibrium Gaussian density) and the white noises of the Langevin thermostats are independent (and, of course, independent from previous iterations, so that a trajectory starting at some configuration q_i will be different from the trajectory that initially produced this q_i). At the end of this kth iteration, structures $(q_j)_{M_k < j \leq M_{k+1}}$ have been added, and iteration $k + 1$ starts.

The procedure penalizes areas that have already been extensively visited, and is in a way reminiscent of the metadynamics[32] method except that the statistical biasing is done through a selection step between each iteration rather than a biasing force updated along the trajectory. By comparison with metadynamics, this unsupervised selection step has the advantage of overcoming the critical choice of initial collective variable at the beginning of the simulation reinforcing automation of the sampling scheme.

This strategy belongs to the family of counts based adaptive sampling algorithms, where one only exploits the number of passages in the different states (micro or macro) visited in the previous iterations to choose which state to restart trajectories from. These are known to be efficient for pure exploration purposes (as is the case here), even though more refined algorithms exist when some information is available as to where the sampling should be guided.[24] However, in contrast to what is usually done in the context of Markov State Models (MSMs),[23] the states are not defined by applying a clustering algorithm to the already explored structures, but are the projection on the n principal components generated by PCA (here, $n = 4$ as we discussed) of all the previous data. This has the advantage of providing an unsupervised sampling strategy that does not rely on a particular clustering algorithm (and therefore its associated parameters) and treating every point of this 4-dimensional representation differently.

At the end of the simulation, M_K configurations have been sampled with K, the total number of iterations. For a large K, the distribution of these configurations does not converge to the canonical distribution because of the statistical bias induced by the selection. To compute thermodynamic quantities, this bias should be taken into account. In that case, we interpret the previous selection as an importance sampling scheme. Thus, we have to compute a score $\omega_i > 0$ for each $i \in \{1, \ldots, M_K\}$ so that the canonical average of an observable φ is estimated by

$$\langle \varphi \rangle \simeq \frac{\sum_{i=1}^{M_K} \omega_i \varphi(q_i)}{\sum_{i=1}^{M_K} \omega_i}.$$

The score ω_i is the ratio between the probabilities to obtain q_i in the biased simulation and in an unbiased simulation (where, between each iteration, the next initial conditions are uniformly

chosen among all currently available configurations, *i.e.* all with probability $1/M_k$). As a consequence, it is computed as follows: for all $i \leq M_0$, $\omega_i = 1$. Suppose by induction that ω_i has been computed for all $i \leq M_{k-1}$ for some k. Let $(i_1, \ldots, i_{s_k})$ be the indexes that have been randomly selected for the initial conditions at the beginning of iteration k. For each $h \in \{i_1, \ldots, i_{s_k}\}$, α_h is computed:

$$\alpha_h = \frac{1}{M_k \mathbb{P}(I = h)} = \frac{\rho_k(x_h)}{M_k} \sum_{j=1}^{M_k} \rho_k^{-1}(x_j).$$

Then, the score of all the configurations that are generated during iteration k from the initial condition q_h is $\alpha_h \omega_h$. That way, ω_i is computed for all $i \leq M_k$.

This latest point is important since it means that the total simulation time can be used to compute average statistical properties that are unbiased and therefore exploitable. For example, it is possible to compare them to those obtained upon performing conventional MD runs.

Finally, it should be noticed that, instead of the PCA, this adaptive sampling strategy may be used with any other collective variables and/or dimensionality reduction algorithm. Overall the procedure is fully unsupervised, fast and can be used within Tinker-HP in a fully automated way.

3 Large scale unsupervised adaptive simulation using polarizable force fields (PFFs) and GPUs

3.1 Preparation of systems and choice of initial structures

In order to perform a large scale unsupervised adaptive sampling simulation, starting structures have to be selected from a conventional MD simulation (using either n-PFF or PFF approaches). We chose the RIKEN dataset as the starting point. From their 10 μs conventional MD simulation (PDB: 6LU7, pH = 8)[3] using the n-PFF AMBER14ff[10] approach and using PCA as a guiding thread, we carefully extracted 14 relevant structures that represent our starting point for the study. It is worth noting that the 6LU7 crystal structure is a holo structure including a covalently bound inhibitor. The inhibitor-unbound apo structure was initially obtained by RIKEN removing the inhibitor and relaxed over 10 μs of simulation (https://data.mendeley.com/datasets/vpps4vhryg/1). Each Amber14ff structure was then minimized with the AMOEBA PFF[33–36] and an L-BFGS algorithm until a Root Mean Square (RMS) of 1 kcal mol^{-1} on the gradient was reached. It is important to note that not all histidine residues are protonated in the RIKEN structure similarly to the DESRES one. Since it has been recently demonstrated that the highest pK_a for possible protonation of histidine sites was lower in the SARS-CoV-2 M^{pro} than in the SARS-CoV-1 M^{pro}, being about 6.6,[37] the present simulation is therefore consistent with physiological pH conditions (pH = 7.4).[38]

3.2 Simulation protocol

The presented all-atom simulation was performed using the newly developed GPU module[22] within the Tinker-HP package,[21] which is part of the Tinker 8 platform.[39] This newly developed module is able to efficiently exploit mixed precision[22] offering a strong acceleration of simulations using GPUs. The 98 694 atom initial structure of the fully solvated M^{pro} dimer was extracted from the Protein Data Bank (PDB: 6LU7) and the AMOEBA PFF[33,34,36] was used to describe all atoms (protein and water). Periodic boundary conditions using a cubic box with side lengths of 100 Å were used. Langevin molecular dynamics simulations were performed using the BAOAB-RESPA1 integrator[40] using a 10 fs outer timestep, a preconditioned conjugate gradient polarization solver (with a 10^{-5} convergence threshold), hydrogen-mass repartitioning (HMR) and random initial velocities. Periodic boundary conditions (PBCs) were employed using the Smooth Particle Mesh Ewald (SPME) method with a grid of dimensions 128 Å × 128 Å × 128 Å. The Ewald-cutoff was taken to be 7 Å and the van der Waals cutoff to be 9 Å. As we explained, we started the simulation by running a 10 ns cMD for each of RIKEN's 14 representative structures (as mentioned in Section 3.1). A first adaptive sampling selection was then conducted on those 140 ns initial structures. We chose to use the first four PCA components (see the method section) as conformational space for the adaptive sampling method. At each iteration, the adaptive sampling procedure is then used on these newly computed first four PCA components in order to select 100 structures. Then, 100 independent molecular simulations of 10 ns were performed in the NVT ensemble at 300 K on single NVIDIA V100 GPU cards. Each trajectory belonging to the same adaptive sampling iteration was run simultaneously on the HPE Jean Zay Supercomputer (IDRIS, GENCI, France). A single adaptive sampling iteration took less than 18 hours to complete, allowing a production rate of 15.14 μs in two weeks. Overall, the simulations ran over 12 working days in line with computer center resources availability.

The complete 15.14 μs trajectories with and without water are freely accessible through the Swiss National Supercomputing Center (CSCS)[41] and have been linked to the BioExcel/Molssi COVID-19 community portal. A movie depicting the progress of the exploration can be found in the ESI.†

3.3 Performance of the adaptive sampling exploration: comparisons with other available simulations

As we mentioned in the method section, we use the PCA[28] as an intermediate quantity to orient the consecutive sampling iteration. However, it is also a good quantity to quickly assess the performance of the adaptive sampling scheme for the exploration of the conformational space. Indeed, the analysis of MD trajectories with PCA is a well-known strategy known in the community as the "essential dynamics".[42–44] PCA, being a dimensionality reduction algorithm that evaluates directions maximizing the variance of the dataset, is thus a revealer of a system conformational diversity. Therefore, it can be seen as a way to assess the amount of sampling and can also detect explicit "essential motions" otherwise not discernible using predefined collective variables. Thus, it is interesting to compare the amount of sampling on the space of these reduced variables. This is why we projected the RIKEN, the DESRES and

the first 2 μs Tinker-HP data set on the first two PCA components of the first 2 μs of the Tinker-HP data set (Fig. 1a and b). One can see that, in this space, the Tinker-HP adaptive scheme already captured the RIKEN and DESRES major main PCA features. It also appears that the RIKEN trajectory sampled a portion of conformational space close to the Tinker-HP data set while the DESRES trajectory seems to explore only the area that is most sampled by Tinker-HP. The same procedure was applied for the PCA components and associated data of the entire Tinker-HP data set (Fig. 1c and d) and it is striking that a much larger portion of conformational space has been sampled by our adaptive scheme. Additionally, we also projected the same data sets on the first two principal components of the RIKEN trajectory which gives the same justification of the larger sampling obtained by our method (see Fig. 4 in the ESI†).

As a preliminary conclusion, we can say that our adaptive sampling strategy allowed us to generate a multi-microsecond polarizable MD simulation that sampled a vast area of the free energy landscape. In addition, we analyzed the Root Mean Square Deviation (RMSD) on protein backbones *versus* the radius of gyration (see Fig. 5 in the ESI†) for the AMOEBA 15.14 μs. It revealed large conformational changes. Variations for the radius of gyration are about 2 Å, while the variation is 1 Å for non-polarizable conventional MD. Such plots are very useful to understand one key question: what makes the AMOEBA results different? Is it the choice of PFF (*vs.* n-PFF) or is it the choice of adaptive sampling strategy. In order to provide a fair (and somewhat quantitative) comparison between the FFs and to decorrelate the effects of the FFs themselves from the gains due to adaptive sampling, we limit ourselves to structures with a reweighting score (see the section above) greater than 1 as it is the score of the frames visited during a conventional MD simulation and as frames with scores lower than 1 are the ones that have been favored by the adaptive algorithm to maximize exploration. 3/4 of the points are therefore removed using this criterion offering a view of the performance of the adaptive sampling. The plot representing the remaining point is presented in Fig. 3 (ESI)† for AMOEBA and it can be directly compared to the RIKEN plot for example. Clearly differences exist between AMBER and AMOEBA results, and they also come from the choice of FF. In addition, important changes are also observed in different important areas of the protease such as the dimerization site. The RMSD of the protein backbone *versus* the RMSD of the chain A dimerization site (see Fig. 6 in the ESI†) depicts large fluctuations between 6 and 7 Å. DESRES and RIKEN trajectories exhibited only 2 Å, which is in the order of the size of the observed PCA features. Overall, these first observations of the differences between the non-polarizable and the polarizable simulations motivate a further analysis of the different simulations.

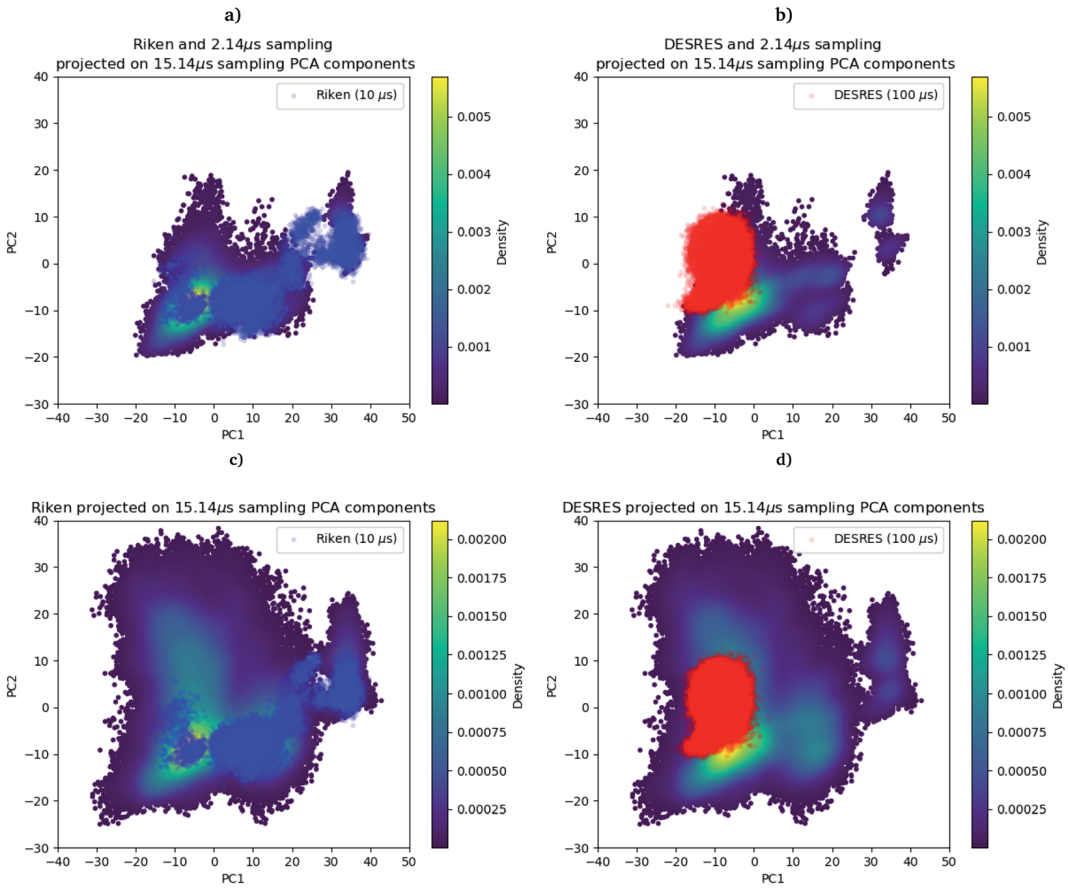


Fig. 1 RIKEN and DESRES datasets superposed on the 6LU7 protein backbone and projected on the first two PCA components fitted to, respectively, the 2 μs (a and b) and 15.14 μs (c and d) simulations.

3.4 Unsupervised clustering and extraction of the unbiased relative free energy between representative domains

First, if the PCA analysis reveals useful information, a proper clustering of the produced ensembles is a more precise and quantitative framework to discuss differences between simulations and possible new features captured by the AMOEBA force field. Therefore, we applied to all trajectories the density-based spatial clustering of applications with the noise (DBSCAN) method.[45] DBSCAN is an unsupervised machine learning algorithm that groups together data in clusters according to their density. It has the particularity to label points as noise if they are not in a dense region and are then not assigned to any cluster. DBSCAN is particularly well suited in our case as it is especially designed to target arbitrary shape clusters. To evaluate the density, DBSCAN uses two parameters, ε the distance at which two points are considered to be neighbors and MinPts the minimum number of points needed to define a cluster. ε was chosen using the nearest neighbor graph procedure, *i.e.* by plotting the distance to the nearest n-neighbor for each point, ordered from the largest to the smallest value, and evaluating ε for which the graph starts forming an elbow. For a given ε we then scanned different values of MinPts until relatively large clusters covering a wide range of the space are found. In practice we evaluated the distance to the 4th nearest neighbor on the 4 dimensions composed of the first four 15.14 μs principal components generated by PCA (see Fig. 7 in the ESI†). For DESRES and RIKEN, after being aligned to their respective PDB, the structures were projected on this 4D space.

Our choice of using the AMOEBA 15.14 μs PCA components as the starting point of the clustering is driven by the conformational diversity brought about by the coupling of the PFF and the adaptive sampling scheme. For visualization, clusters are then projected on the first two principal components (Fig. 2). To evaluate the quality of the clustering we used three scoring methods for unknown labeled data:[46] Silhouette coefficient, Calinski–Harabasz and Davies–Bouldin indices. These indices confirmed our parameter optimization procedure and the high quality of the clustering. Our new adaptive sampling scheme has the main advantage of offering access to true statistical properties such as free energies. To understand the cluster stability, the free energies for each cluster are computed (Fig. 3c and d) through the evaluation of the probability distribution over the total number of structures. Notice that, since not all the structures are part of a cluster, the cluster probabilities do not add up to one. The unbiased probability distribution (Fig. 3a and b) is estimated with the de-biasing procedure explained in the previous section. The de-biasing step preserves the trend

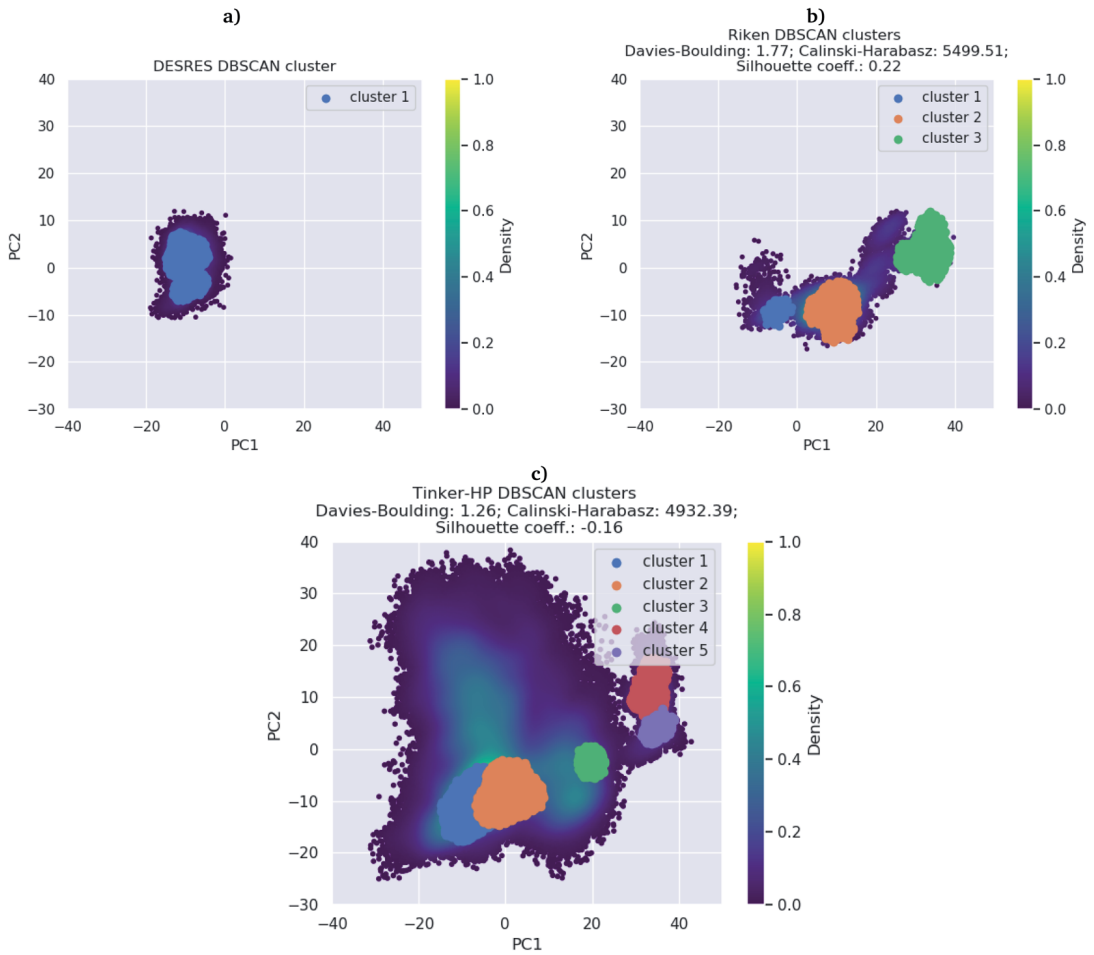


Fig. 2 DBSCAN clustering of (a) DESRES (100 μs) and (b) RIKEN (10 μs) datasets and (c) the Tinker-HP 15 μs simulation.

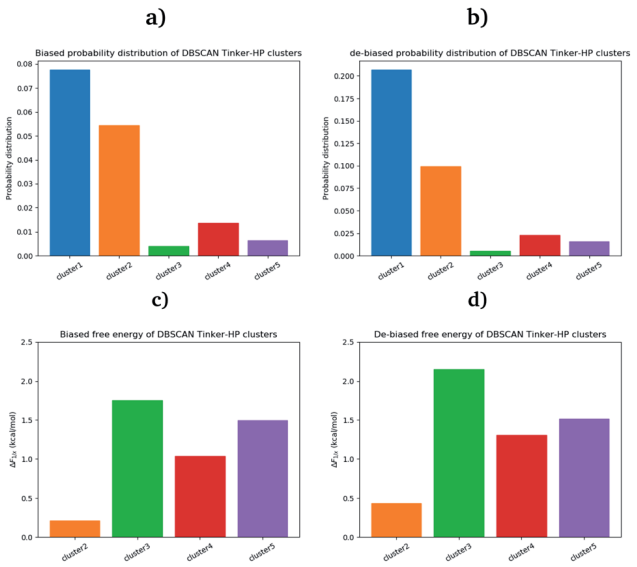


Fig. 3 Biased (a) and unbiased (b) probability distribution of DBSCAN Tinker-HP clusters. Biased (c) and unbiased (d) relative free energies of the DBSCAN Tinker-HP 15.14 μs clusters, with respect to cluster 1.

between clusters but increases the probabilities. It means that the five clusters were disadvantaged by the adaptive sampling. For example, the biased simulation assessed an 8% probability for the presence of cluster 1, which should have contained, in an unbiased simulation, 20% of the configurations. Besides, cluster 1 is indeed the most explored region by both DESRES and RIKEN. Hence, the algorithm managed to disadvantage this part of the conformational space which is what we could have expected as it favored intermediate transition areas to the detriment of dense regions in order to discover new regions. The effect of the polarizability on structural properties such as volumes and RMSF is further depicted in the next section. Overall, our approach demonstrated our capability to reach high-resolution conformational space exploration using a PFF. We identified 5 different clusters using AMOEBA (see Fig. 2). While some of these states were already identified in previous n-PFF simulations (RIKEN and DESRES), we found two new non-negligible conformations (according to Fig. 3) that can be critical, *e.g.*, for the computation of thermodynamic properties and finally guide further ensemble docking simulations and/or to help to interpret experimental results.

4 Correlation with experimental data: structural markers for protomer activity and new features

4.1 Markers of the structuring of the oxyanion hole

To ensure the validity of our AMOEBA simulations, we compared our computed properties with available experimental data. Since the beginning of the COVID-19 pandemic various X-ray structures have been released (PDB: 6Y84, 6LU7, 6Y2G, …).[3,47,48] They provided important insight on specific interactions between residues as well as structural information about the active site. To be consistent with RIKEN simulations we used as reference the same PDB: 6LU7.[3] Note that DESRES used another PDB, 6Y84,[47] which we used as a reference in the computation of its properties. Crystal structures have been projected on the first two PCA components of the Tinker-HP simulations (see Fig. 8 in the ESI†).

Recently, Zhou *et al.* published an experimental study of the apo structure (PDB 1UJ1)[49] at physiological pH. They found several features allowing for the characterization of the presence of the oxyanion hole structure which is a key structural element of the activity of each protomer. In particular, they proposed to monitor the distance between Glu166 and His172 and the π–π stacking between Phe140 and His163. The definitions of these structural markers are not new and were initially also discussed for the SARS-CoV-1 M^{pro}.[50,51] The oxyanion hole is responsible for the stabilization of the substrate in the active site and is of crucial importance for the enzyme's kinetics and activity. Indeed, the substrate binding site is composed of 4 pockets labelled S1 to S4 with the S1 pocket involving very conserved residues such as Glu166, His172, His163 and Phe140. The oxyanion hole of the cysteine protease encompasses backbone amides (Gly143, Ser144, and Cys145) while residues 138 to 145 form the so-called oxyanion-binding loop.[48,51,52] The existence of this latter is responsible in part for the structuring of the S1 pocket.[51] When the stacking and the Glu166–His172 interaction are broken, a rearrangement occurs leading eventually to the collapse of the oxyanion hole. In this case, Glu166 potentially interacts with His163 instead of His172. In other words, strong interactions of Glu166 with His172 associated with a Phe140–His163 stacking are consistent with a structured oxyanion hole, and can be used as a marker of the activation of the enzyme protomer. Inversely, a strong interaction of Glu166 with His163 would rather be a marker of the protomer inactivation linked with a collapse of the S1 substrate-binding pocket. Of course, such analysis is only interpretative, the oxyanion hole structuring being far more complex. However, it has been shown to be useful since the initial studies on the SARS-CoV-1 main protease.[51] In practice, the absence of a well-structured oxyanion hole leads to the inhibition of the enzyme's activity. Experimentally, it is known that the M^{pro} monomeric form is inactive while the active form is a homodimer containing two protomers.[53] In the holo state of SARS-CoV-1, the first protomer is active while the second one is found inactive.[54] For SARS-CoV-2, a pH = 6 crystal structure (PDB: 1UJ1)[53] predicted a strong asymmetry of the protomers with an inactive conformation for one of the protomers linked to a broken Glu166 and His172 interaction. However, the inactivity of one of the protomers is still a hypothesis as crystallographic studies of the dimer in the space group *C*2 encounter difficulties in capturing the details of each individual protomer. Indeed, data are only available on one of the protomers in the asymmetric unit which always leads to the more ordered conformation and therefore to the most active one. Concerning the apo state, recent experimental results lead to a potential low activity of the apo dimer linked with an observed destructured oxyanion hole.[49] It is important to point out that distances/markers exhibit a distribution of

different values centered around a maximum of frequency due to the liquid conditions that differ from the crystal ones (Fig. 4).

Then we investigated these markers. To study the Phe140–His163 stacking interaction, we use a stacking-index developed by Branduardi and Parrinello[55] who described it as a product of 2 Fermi functions, one considering the radial dependence, and the other the angular dependence of the interaction. The model provides an index ranging from 0 for a non-stacked interaction to 0.6 for a perfect one. The Glu166 interactions and π–π stacking were thus calculated for both chains of all RIKEN, DESRES and Tinker-HP structures and then classified into histograms. Finally, each histogram has been unbiased (*i.e.* reweighted) and extrapolated using a univariate kernel density estimator. Final results are given in Fig. 9 of the ESI.† Furthermore a 6 μs adaptive sampling simulation was performed (on the Irene Joliot Curie Machine (TGCC, GENCI, France)) on the monomer species (PDB: 6LU7) and the same features as discussed below (π–π stacking between Phe140 and His163, and Glu166 interactions with both His172 and His163) were calculated. Since the monomer is known to be in an inactive conformation, it helps us to rationalize the behavior observed in our simulations. Results are depicted in Fig. 10 in the ESI.† The preparation and simulation protocols are similar to what we did for the dimer. Therefore, since His172 and His163 are also unprotonated, we minimized the structure up to a RMS of the gradient of 1 kcal mol^{-1} and generated an initial cMD of 200 ns. We then selected 100 random initial structures according to the Adaptive Sampling protocol of structure selection using the PCA, and we performed 6 iterations of 1 μs for a total simulation time of 6 μs.

For the interaction formed by Glu166, in the case of Tinker-HP, we observed an asymmetry between the two protomers. In one protomer the Glu166–His172 interaction is significantly weaker than in the other exhibiting a well-defined marker of a smaller activity of the protomer. This relative non-interaction is in accordance with the results obtained on the monomer which appears to be similar (see ESI Fig. 10†). The situation is more complex in the other protomer where we observe an oscillation between two states, presenting either a formed Glu166–His172 interaction or its absence leading to only some partial activity markers. However, the "interacting" state clearly dominates the statistics. These results demonstrate that the oxyanion hole is only partially organized in the other protomer. This is consistent with experimental data on the apo state[49] and also with the data on the active protomer of the holo state which shows distances of around 5 Å (see ref. 37 and references therein for a discussion of the different available crystal structures). It is, of course, only one single marker but it could already corroborate the asymmetry observed in the holo state where only one protomer is found to be active,[48] a similar feature to what was previously observed in SARS-CoV-1.[54] Based on the analysis of this single marker, we tend to have an inactive first protomer coupled to a second protomer that exhibits some partial but clear activity features (two states) when compared to its inactive counterpart and to the monomer. Similar interpretations can be deduced from the DESRES and RIKEN simulations despite a less clear picture of the His172–Glu166 interactions which appear extremely flexible with more mixed states, especially for AMBER. This is not surprising as Glu–His interactions can be classified as H-bonds, a class of directional weak interactions that are known to be difficult to model using n-PFFs[56,57] as polarizability contributes significantly to the accuracy of simulations of structures with hydrogen bonds.[15,58] However, a single distance is not enough to reach a conclusion and should be combined with other markers such as the Glu166–His163 distance. We note here a stronger asymmetry of such distances in protomers for DESRES while in the case of RIKEN and Tinker-HP we could again observe a mixture between interacting/non-interacting states. However, this second marker should be carefully considered as a direct comparison with our monomer simulation (see ESI Fig. 10†) shows that this distance criterion is less well-defined for discussing the protomer "activity" than the Glu166–His172 distance. Since our monomer is known to be inactive, it could be deduced that this marker should always be associated with the evaluation of the Glu166–His172 distance. In practice, one should look at the relative strength of these interactions and the Glu166–His163 distance here appears to be clearly longer than the Glu166–His172 ones. Glu166–His163 distances appear consistent with data on the active protomer of the holo state which shows distances going beyond 6–8 Å (see ref. 37 and references therein for a discussion of the different available crystal structures). In that connection, a better conservation of the catalytic dial is observed in the RIKEN and Tinker-HP simulations with a smaller Cys145–His41 distance compared to DESRES (see ESI Fig. 9†). The active site of the M^{pro} protease comprises a catalytic dyad composed of residues Cys145 and His41. X-ray crystal structures of SARS-CoV-1 (ref. 51 and 52) found a Cys145–His41 distance between 3 and 3.9 Å. In comparison, our simulations revealed distances of around 4 Å while AMBER and DES-AMBER distances are, respectively, around 4.5 and 6–7 Å. Regarding the relatively small differences between the SARS-CoV-1 and SARS-CoV-2 main proteases, AMOEBA results appear closer to experimental data.

Finally, a last marker is studied to confirm our observations: the π–π stacking between Phe140 and His163. Results are

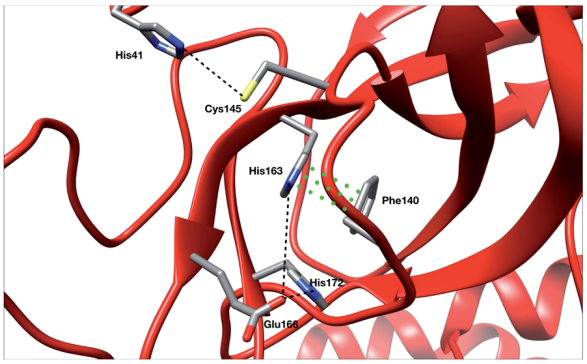


Fig. 4 Representation of the π–π stacking interaction between His163 and Phe140 residues (green points) and of several distances of interest which are responsible for the stability of the active site (black dashed lines).

depicted in Fig. 9 in the ESI.† Tinker-HP does not capture this stacking in one protomer while again two mixed-states (stacked and un-stacked) are observed in the other protomer. The same observations can be made for DESRES and RIKEN although the states are less well defined in connection with the well-known difficulty of capturing π–π stacking with n-PFFs.[59] Despite these differences, the 3 simulations appear consistent. Overall, our initial conclusion stands: we describe an asymmetric situation where one protomer is fully inactive and the other shows some partial activity features. It is important to point out that these results are not artificial and linked to our starting structure. Fig. 11 of the ESI† shows the convergence of the stacking

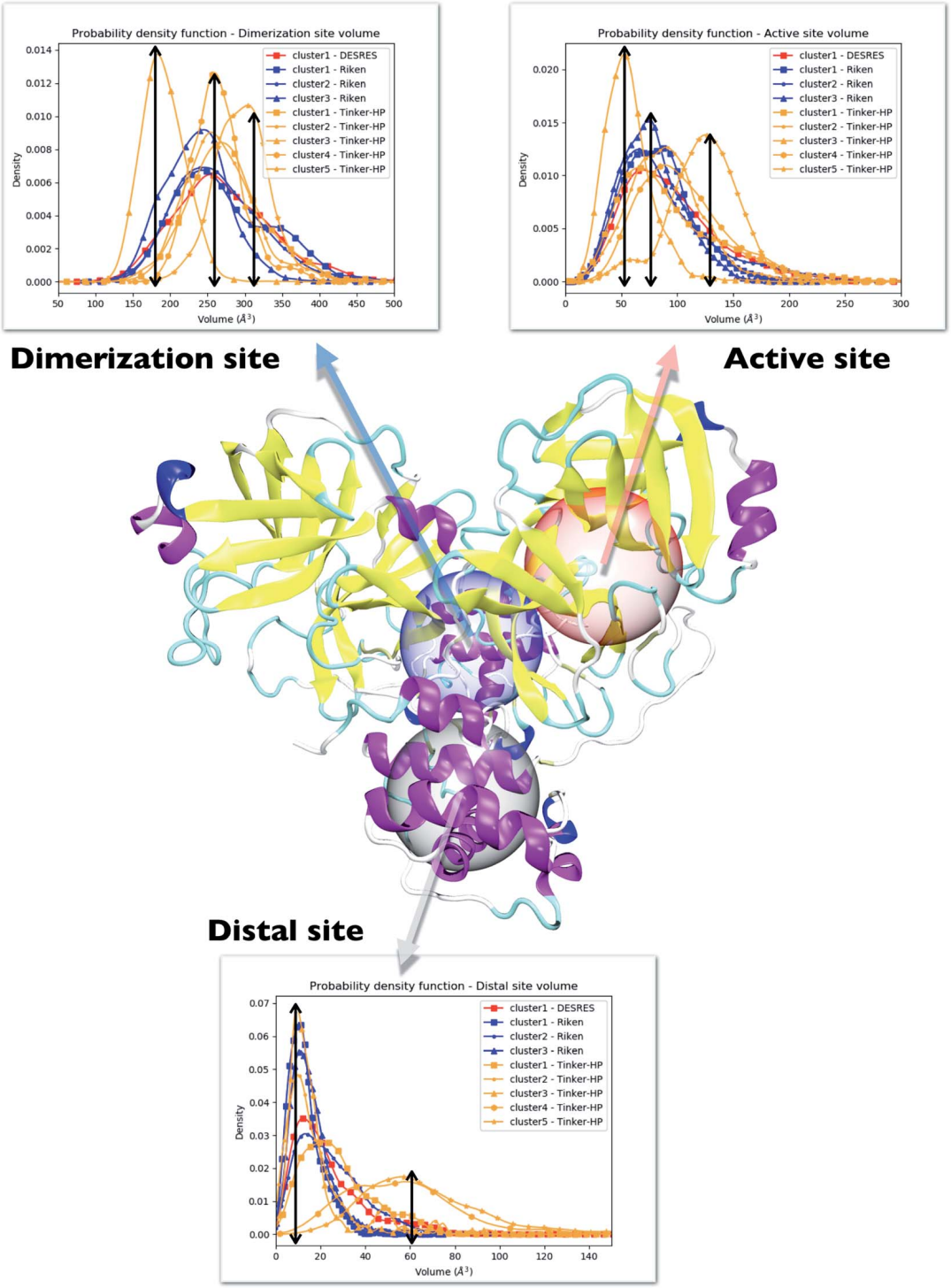


Fig. 5 Representation of the 3 cavities considered in this study: the dimerization site, active site and distal site. For each cavity, trends inferred from each cluster are depicted and superposed on three different graphs. Each curve has been unbiased according to the reweighting approach described in this work. Cavity volumes are the sum of volumes found in both protomers. The black arrows link the maxima of frequency to the volume axis to highlight the difference between clusters.

marker over the 15.14 μs simulation. If protomer 1 is clearly not evolving over the simulation, protomer 2 evolves slowly towards the discussed 2 state organization. Overall, our results are compatible with the description of the apo crystal structure by Zhou *et al.*[49] who observed an incomplete structured oxyanion hole exhibiting several mixed states of structuring. This highlights the large flexibility of the enzyme discussed in the experimental literature at room temperature.[38] Our data also support the possible strong asymmetry between protomers discussed in the holo state.[53]

4.2 Evaluation of the volumes of the enzyme cavities

One way to measure some potential global differences between the different simulations is to measure the active site volume in each cluster and to depict the observed trend similarly to the π–π stacking previously. Besides the main active site cavity, the main protease exhibits 2 other cavities: the distal site and the dimerization site. Represented in Fig. 5, these cavities are considered as potential targets for drug inhibition.[60,61] An accurate description of each of these cavities is essential to the estimation of efficient inhibitors. For each cluster of each dataset, we thus estimated those 3 cavity volumes. Volumes were calculated for each isolated cluster using POVME 3.0 software.[62] For each cavity, a 1.0 Å grid spacing was chosen. Residues 7–198 and 198–306 and all residues within 3.5 Å from the other protomer were selected for the active, distal and dimerization sites with, respectively, 12 Å, 10 Å and 10 Å. 1000 structures were randomly chosen per cluster for the analysis. When a cluster had less than 1000 structures, we chose all the structures. Detailed information is given in the ESI† on the size of each cluster as well as their relative size (see Table 1 in the ESI†). Similarly to the π–π stacking and the Glu166 distances, we used the univariate kernel density estimator on the volumes. The final volumes are depicted in Fig. 5. Additionally, each cluster has a normal distribution supporting the quality of DBSCAN clusters. Different trends appear, represented by black arrows. For the 3 cavities, we observed a similarity between the single DESRES cluster, clusters 1 and 2 from RIKEN and Tinker-HP's clusters 1 and 2. Agreement is also found with volumes obtained by Sztain *et al.* using a Gaussian accelerated MD (GaMD) enhanced sampling strategy coupled with AMBER ff14SB[8] which also match these results confirming the importance of simulating long enough in conventional MD. Overall, while Tinker-HP clusters 1 and 2 are in good agreement with RIKEN and DESRES clusters, our clusters 3, 4 and 5 appear to be different and specifically highlight the importance of the PFF choice, *i.e.* these data are not obtained using enhanced sampling coupled with non-PFFs.[8] As we pointed out earlier, differences indeed occur between clusters and between different datasets, going in the same direction of the previous analysis of the π–π stacking between residues Phe140 and His163 in chains A and B. For Tinker-HP, we observed a contraction for the three cavities in cluster 3 while in cluster 4 and especially cluster 5, we observed a strong difference with a non-negligible increase of the cavity volumes. Cavities from clusters 4/5 depict stronger volume fluctuations when using the AMOEBA PFF. While cavity volumes obtained from AMBER/DES-AMBER simulations and from clusters 1 and 2 from AMOEBA simulations are in agreement, the AMOEBA results clearly capture an additional feature not captured by the DES-AMBER and AMBER simulations. This information could be important for designing potential new inhibitors.

Consequently, since strong differences between methods are observed in the volume evaluations of the different clusters, it is interesting to estimate the global protomer volumes if one wants to try to capture further the discussed asymmetry. Protomer volumes can be found in Fig. 6. Protomer 1 (predicted to be non-active) depicts a strong gaussian behavior while protomer 2 (predicted to be oscillating between an active and a non-active state) is characterized by a spread gaussian with more important associated volume compared to protomer 1. This increase of volume is therefore concomitant with the previous asymmetry related to the various discussed structural markers. It is worth noting that this asymmetry is also found for the DESRES simulation but to a lesser extent compared to that for the AMOEBA Tinker-HP simulations. Concerning the RIKEN dataset, this feature is not found as both protomers depict a similar gaussian trend with very similar values.

4.3 Analysis of the local fluctuations: high flexibility of the C-terminal region

Finally, it is also possible to study local fluctuations in the structural dynamics of the M^{pro} dimer system to uncover other types of difference between datasets. We calculated the fluctuation of residues in each cluster on the same 1000 previously randomly chosen structures per cluster using the Root Mean Square Fluctuation (RMSF). These were calculated on the 5 clusters from Tinker-HP (AMOEBA), the 3 clusters from RIKEN (AMBER) and the single cluster from DESRES (DES-AMBER). Results are depicted in Fig. 7. The most interesting fluctuation as well as the main differences between clusters originates from a different spatial rearrangement of the C-terminal region

Table 1 Average and standard deviation of the number of water molecules around His163 and His41 residues in DES-AMBER, AMBER and AMOEBA force field simulations (pH 7.4)

	His163		His41	
	Protomer 1	Protomer 2	Protomer 1	Protomer 2
DES-AMBER	0.14, $\sigma = 0.48$	0.77, $\sigma = 0.44$	4.01, $\sigma = 1.17$	1.61, $\sigma = 0.75$
AMBER	0.49, $\sigma = 0.57$	0.44, $\sigma = 0.41$	2.38, $\sigma = 1.11$	2.25, $\sigma = 1.23$
AMOEBA	0.31, $\sigma = 0.51$	0.13, $\sigma = 0.34$	1.48, $\sigma = 0.99$	1.62, $\sigma = 1.06$
Experiments[38,49,50,53]	0 or 1		1	

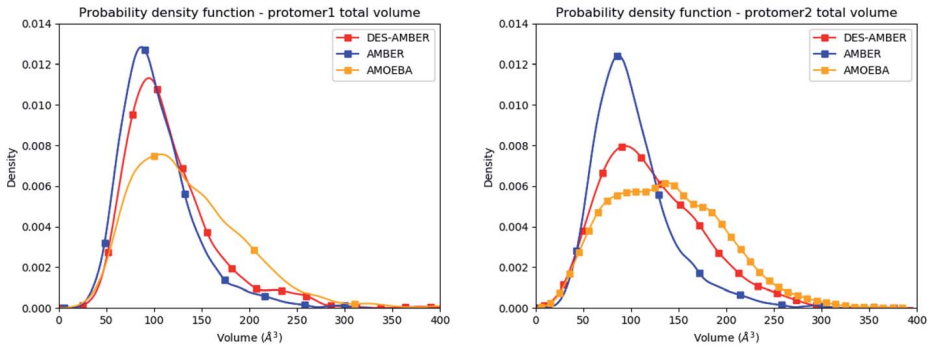


Fig. 6 Graphical representation of the distal + active sites for protomer 1 (on the left) and protomer 2 (on the right) for the DESRES, RIKEN and Tinker-HP simulations.

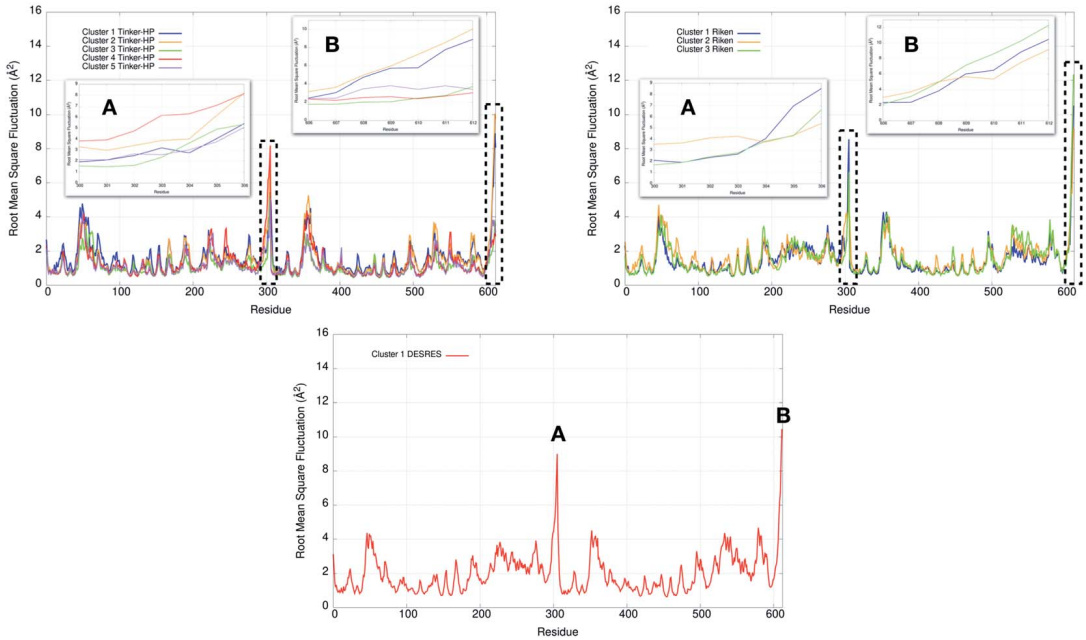


Fig. 7 Representation of the RMSF for each cluster of each simulation (Tinker-HP, RIKEN and DESRES). Zoomed-in images of both chains (A and B) are represented in subgraphics and correspond to the C-terminal end where the most important fluctuations are found (residues 300 to 306 for chains A and B).

of the protein (*e.g.* residues 300 to 306 on chains A and B of the dimer). In fact, this region is highly dynamical, which is in accordance with experimental X-ray observations where the electron density of the C-terminal domain was insufficient for backbone tracing, suggesting the flexibility of this region.[49] Visual enlargements of this region are provided in the subgraphics of Fig. 7 for chains A and B that do not differ significantly. Cluster 1 from the DESRES simulation depicts the same fluctuation as cluster 1 from the RIKEN simulation. This behaviour of the C-terminal region in these two clusters is characterized by a π–π interaction between Phe305 and His41, eventually blocking the access of any ligand to the active site. When the C terminal region does not interact with His41, it adopts an unfolded configuration which shows the high flexibility of these terminal amino acids. Structural representations can be found in Fig. 8. As this event is observed on the active site of only one chain and not both of them, it could be another marker of the previously mentioned protomer inactivation. We also observed such fluctuations in clusters 1 and 2 extracted from our Tinker-HP/AMOEBA simulations. However, in cluster 1, while the Phe305–His41 π–π interaction is indeed observed, we measure a lower fluctuation of chain A for cluster 1. It corresponds to a weaker interaction between Phe305 and His41 as configurations where the C-terminal branch is less structured are preferred. A similar feature is observed for cluster 2 of RIKEN, but with an inversion of fluctuation peaks between A and B. Overall, clusters 1 and 2 obtained from the Tinker-HP and RIKEN simulations appear relatively similar in the PCA space. They correspond to clusters where the C terminal region can oscillate between two states: one with a π–π stacking interaction between Phe305 and His41, and another with a less structured C-terminal branch with higher flexibility. Clusters 4 and 5 from our Tinker-HP simulations and to a lesser extent RIKEN's cluster 3 correspond to another configuration of the C-

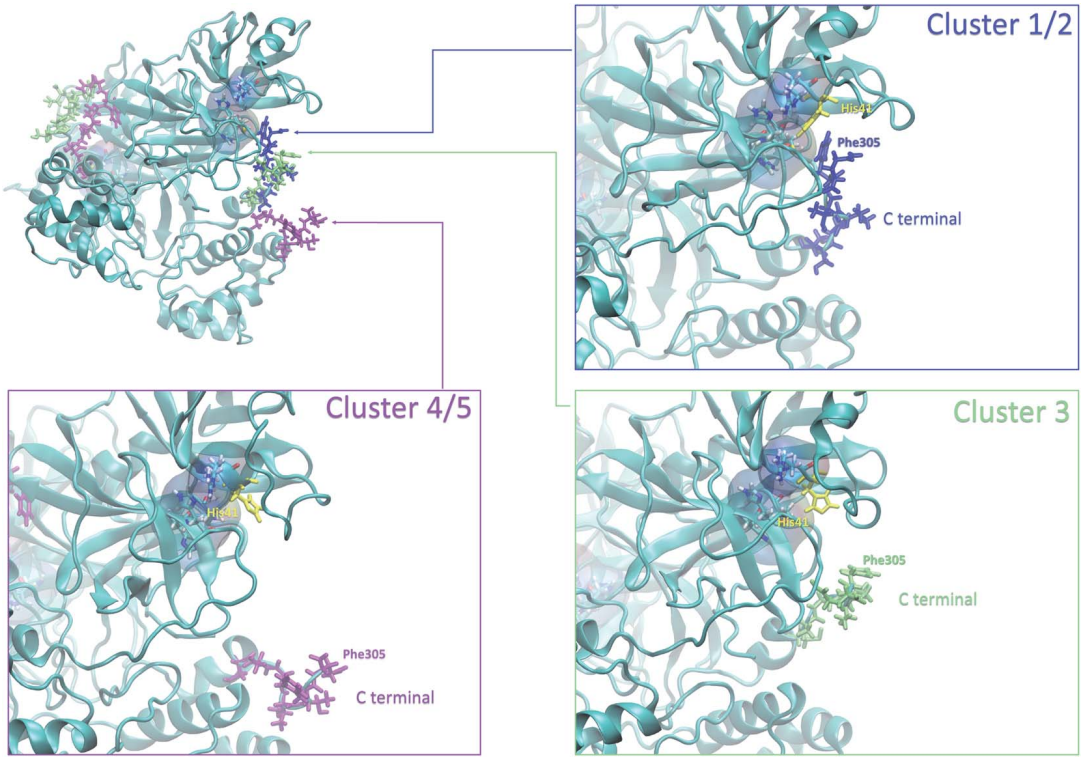


Fig. 8 Representation of the 3 possible states of the C terminal end. The whole protein is presented in ice blue. The C-terminal end presented in sapphire blue depicts most of the states in clusters 1 and 2, where the Phe305 residue of the C-terminal region is stacked with His41 of the catalytic site. The C-terminal end presented in lime depicts most of the states in cluster 3, and the one presented in purple depicts most of the states in clusters 4 and 5.

terminal region. Representative pictures are provided in Fig. 8 for each cluster C-terminal conformations. In these clusters, the C-terminal region appears more preserved/organized as it is localized further from the active site. To summarize the discussion concerning this specific feature, the high C-terminal flexibility observed in the X-ray experiments can be traced back to a modulated access to the active site linked to the absence of π–π stacking between Phe305 and His41. In other words, the C-terminal region of the fully inactive protomer is shown to oscillate between several states and one of them directly interacts with the other protomer active site. Such interaction tends to block the active site access, therefore modulating down the activity of the potentially most active site. This high flexibility is captured by both RIKEN and Tinker-HP, exemplifying the importance of the local conformational sampling and supporting the experimental analysis of a full inactivation of the apo state.[49]

5 Comparative ligandability analysis: searching for cryptic pockets

In order to check if all the previous features could affect the ligandability of the M^{pro} dimer system, we decided to search if new cryptic pockets are detected in each cluster. By taking into account the same sets as for the cavity volume analysis, cryptic pockets were searched using DoGSite Scorer software,[63] an automated tool for pocket detection and pocket descriptor calculation. DoGSite Scorer detected 18 pockets located on chain A or at the interface of chains A and B of the SARS-CoV-2 protease 6LU7 crystal structure. Among these pockets, 6 are already described in the literature:[8,64] pockets 'P_1_1', 'P_3' and 'P_15' corresponding to the dimerization site; the 'P_2' pocket corresponding to the active site and the 'P_6' and 'P_11' pockets located in the distal region. These 18 pockets were used as a reference and all pockets detected on the DESRES, RIKEN and Tinker-HP selected structures were assigned to these reference pockets by comparing the list of residues of the different pockets and selecting the reference pocket with the maximum number of common residues. When the maximum number of common residues was lower than 5, and the ratio between the maximum number of common residues and the number of residues in the predicted pocket was below 0.25, the pocket was not assigned to any reference pocket and was defined as a new cryptic pocket. New cryptic pockets were named after the first structure in which they were detected and added to the set of reference pockets. For example, the 'R_c1_s1_P14' mentioned in Fig. 11 is the pocket P_14 detected by DoGSite Scorer in structure 1 (s1) of cluster 1 (c1) of the RIKEN (R) simulations. The results of pocket assignation and new cryptic pocket identification are presented in Fig. 10. We observed that the reference pockets previously highlighted as 'active site', 'dimerization site' and 'distal site', except 'P_6', are particularly conserved and detected in a large majority of analyzed structures. However, a consequent number of other pockets were also detected: (1) in a few structures such as 'R_c1_s2_P21',

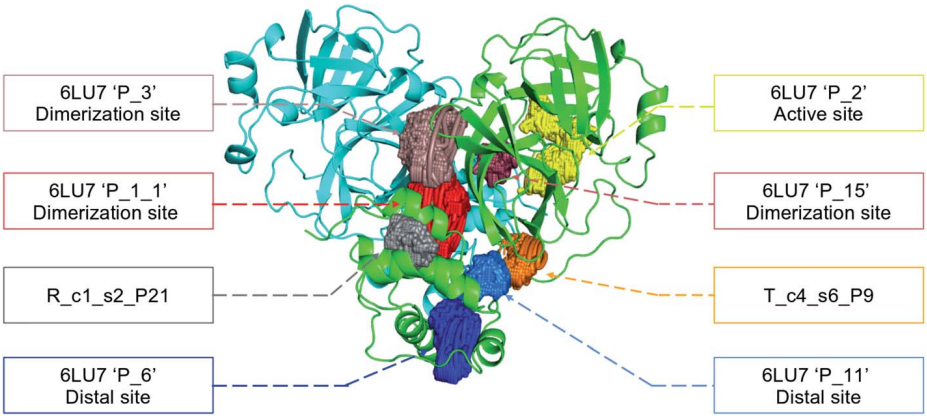


Fig. 9 Representation of the pocket locations on the 6LU7 SARS-CoV-2 main protease structure.

'R_c1_s18_P14' or 'T_c4_s19_P3' or (2) in many structures, such as 'R_c1_s2_P20', 'R_c1_s2_P25' or 'R_c1_s4_P7'. Interestingly, only 3 pockets were retrieved in clusters 4 and 5 of the Tinker-HP simulations: 'T_c4_s2_P8', 'T_c4_s5_P5' and 'T_c4_s6_P9'. The last one, 'T_c4_s6_P9' is of particular interest since its volume is equal to 199 Å^3 and its druggability score, DrugScore,[65] reaches 0.62. We repeated the pocket detection and analysis procedure on 100 randomly selected structures (20 for each of the 5 clusters) identified within the Tinker-HP simulations (see Fig. 10 in the ESI†). We observed that the 3 previously identified pockets 'T_c4_s2_P8', 'T_c4_s5_P5' and 'T_c4_s6_P9' were also detected on the structures randomly selected in clusters 4 and 5 of the Tinker HP simulations but also partially in cluster 3. We then evaluated if all the pockets assigned to the 'T_c4_s6_P9' pocket displayed similar properties. We observed that the mean volume of these pockets was 215 Å^3 but few structures presented extreme values far superior to this mean volume (Fig. 12 in the ESI†). Similarly, the DrugScore mean value was 0.37 but with large variations among the structures and the clusters (see Fig. 13 in the ESI†). For comparison, we also computed the DrugScore value distribution for each newly identified pocket, *i.e.* pockets that were not detected in the 6LU7 structure (Fig. 14 in the ESI†). One pocket, 'R_c1_s2_P21', displays peculiar properties with a mean druggability value of 0.6 and a mean volume value of 150 Å^3 which seems to indicate that this pocket may only accommodate very small compounds. The discovery of the 'T_c4_s6_P9' pocket is thus a very promising result, but one that underlines the necessity of carefully selecting one or several structure(s) in which the pocket properties are optimal for further *in silico* investigations to identify small molecules able to modulate the SARS-CoV-2 protease activity. All the pockets discussed herein are represented within the 6LU7 structure in Fig. 9.

6 Solvation analysis: the importance of including explicit polarization effects in water

Water molecules play critical roles in enzyme and protein functioning. In fact water can be a product or a reactant in condensation and hydrolysis reactions, a transition state intermediate in chemical reactions and a structural element at the molecular level. In the lattermost case, water interconnects the protein through hydrogen bonds in order to maintain and stabilize the positions of the residues and the fold.[66] Previous experimental studies on SARS-CoV-1 and SARS-CoV-2 have shown that one structural water molecule was conserved within the main protease of the two viruses and interacts with the cyclic nitrogen of His41.[38,51,52] A recent crystallographic study on SARS-CoV-2 suggests that another water molecule could be observed around His163.[49] In order to calculate the number of water molecules inside the active site and in proximity of His41 and His163 of both protomers, we have created a virtual sphere of 4 Å, centered on the nitrogen of each of the two concerned histidines and have calculated the number of water molecules inside the active site of each protomer over time. Fig. 11 shows the dipole distribution of structural water molecules for protomers 1 and 2 of His163 (a and b) and His41 (c and d). The AMOEBA results are striking. They show that (i) the water molecules in each of the two protomers' active sites are highly polarized, and (ii) the AMOEBA distribution of the water molecules is significantly different from the ones observed in

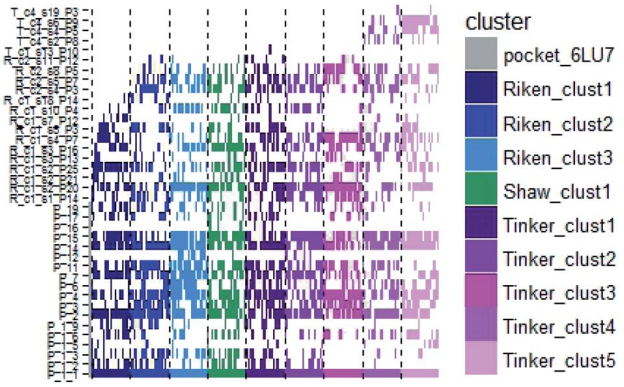


Fig. 10 Schematic representation of the detected DoGSite Score pockets within the 6LU7 structure (first column on the left, represented in grey) and 20 structures extracted from each cluster identified within RIKEN (blue gradient), DESRES (green) and Tinker-HP (magenta gradient) simulations.

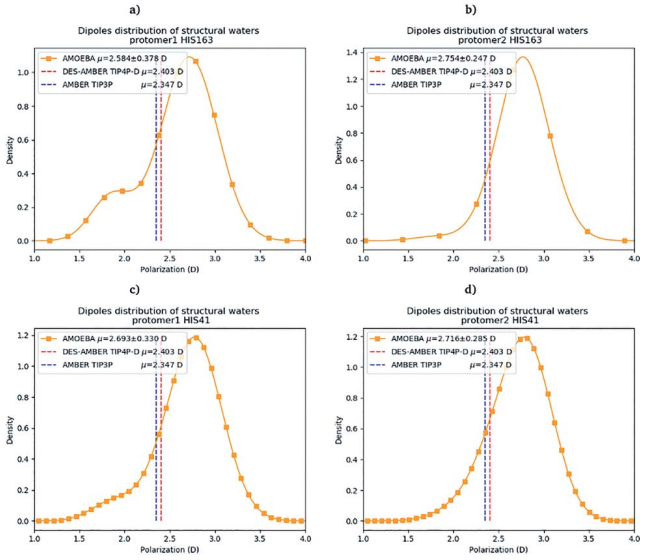


Fig. 11 Dipole distribution of water molecules for protomers 1 and 2 around His163 (a and b) and around His41 (c and d).

the DESRES TIP4-D (DES-AMBER) and RIKEN TIP3P (AMBER) trajectories. High polarization has been shown in past studies to be a common feature of structural water molecules that exhibit high dipole moments.[67] In practice, the average dipole moment having the highest density with the AMOEBA force field is located around 2.9 D while for the DES-AMBER and AMBER n-PFFs, the water dipoles are fixed at 2.403 D and 2.347 D, respectively (see Fig. 5). Since AMOEBA dipole moments are not fixed, we observe strong polarization fluctuations due to water traffic inside the catalytic region. Fig. 15 in the ESI† presents the number of structural water molecules for protomers 1 and 2 of His163 (a and b) and His41 (c and d). All trajectories show a highest density for no water molecules within a distance of 4 Å from protomer 1 of His163. However, this observation is different for protomer 1 of His41 where Tinker-HP trajectories found a highest density for the presence of one water molecule while it was 2 molecules for RIKEN's and 4 molecules for DESRES's trajectories. A non-symmetric distribution of water molecules compared to protomer 1 is found for protomer 2. Tinker-HP and RIKEN trajectories do not predict the frequent presence of water molecules within the chosen distance from His163, while DESRES's trajectories exhibit a higher density for 1 molecule. Concerning His41 of protomer 2, Tinker-HP's and DESRES's trajectories show a most frequent density of one water molecule, while RIKEN's highest density goes to 2 water molecules, and slightly less for 1 molecule. These observations demonstrate that water polarization intensively fluctuates inside the confined active site, suggesting a dynamic role of polarization on water traffic that strongly influences water molecule interactions with His163 and His41 of each of the two protomers. However these interactions are not distributed symmetrically between protomers. So is it compatible with experimental data? Again, relatively detailed X-ray data exist for other coronaviruses including SARS-CoV-1 where the role of histidines has been extensively discussed.[51,52] The presence of a structural water molecule around His41 is always confirmed. For SARS-CoV-2, papers describing the M^{pro} protease structure in its apo state[38,49] under physiological pH conditions also discuss the presence of such molecule found near the catalytic dyad (His41). However, the interaction of the structural water molecule with His163 appears to only be proposed in Zhou *et al.*'s report.[49]

Concerning the precise predicted water count around His41, AMBER and DES-AMBER have on average a higher number of structural water molecules (2.38 to 4.01 at the most) compared to AMOEBA which predicts the presence of 1.5 water molecules, more in line with accumulated experimental data. Fig. 15 in the ESI† shows that the non-polarizable simulations capture frequent configurations with up to 4 water molecules which could be a consequence of the non-inclusion of the polarization effect leading to a weaker and constant dipole moment of the water molecules that could generate more water traffic. Compared to His41, all AMOEBA, AMBER, and DES-AMBER analyses found significantly fewer water molecules around His163. In practice AMOEBA found the lowest water count of all methods with an average of 0.13–0.31 molecules around His163, while the higher trends observed for His41 are still present for all n-PFFs except for one protomer of DES-AMBER that exhibits 0.77 molecules (see Table 2). Clearly, the presence of a structural water molecule around His163 seems less probable for all simulations (under the present pH conditions) and in competition with the water traffic entering the measurement sphere. The dipole distribution of water molecules offers further analysis as it is found to be slightly larger for His163 and associated with a smaller density of highly polarized total dipole moments confirming the trends. In any case, the presence of water in the active site thus appears consistent with the need for a water molecule to model the enzyme reaction mechanism.[38,68]

Table 2 Average and standard deviation of the number of water molecules around His163 and His41 residues using AMOEBA for simulations at pH 7.4 and 6

	His163		His41	
	Protomer 1	Protomer 2	Protomer 1	Protomer 2
AMOEBA pH 6	0.37, $\sigma = 0.65$	0.27, $\sigma = 0.57$	1.95, $\sigma = 1.04$	1.42, $\sigma = 0.97$
AMOEBA pH 7.4	0.31, $\sigma = 0.51$	0.13, $\sigma = 0.34$	1.48, $\sigma = 0.99$	1.62, $\sigma = 1.06$
Experiments	0 or 1		1	

7 Further simulation at lower pH: impact of His172 protonation

From the past studies on SARS-CoV-1 (see ref. 51 and references therein) we know that the activity of the main protease system is pH dependent. While its activity is lower at low pH and high pH, it is higher at pH close to the physiological human pH (*i.e.* 7.4). Studies performed on the M^{pro} of SARS-CoV-1 show a bell-shaped pH–activity curve[51] for the enzyme. All proposed simulations (*i.e.* ours and the one from DESRES and RIKEN) were performed using neutral histidine residues. Indeed, one key element of the impact of lowering the pH is the protonation of His172 and His163.[51] Initially, based on SARS-CoV-1 knowledge, it was thought that if His172 and His163 were not protonated at pH = 8, His172 would be in a protonated state in both protomers at physiological pH (pH = 7.4) since its pK_a was found to be close to 7.6.[69] However, differences exist with the SARS-CoV-2 M^{pro}, and Verma *et al.* recently showed[37] that the pK_a of His172 would be actually lower than anticipated, being about 6.6. Such prediction appears consistent with recent experimental results.[38] Our proposed simulation setup using neutral histidines is therefore likely to be consistent with physiological pH conditions. In that connection, Verma *et al.* described the critical role of the protonation of His172 on the holo state that would happen at pH = 6 and they showed that it would lead to a partial collapse of the S1 pocket, linked with a strong destructuring of the oxyanion hole.[37] Thus, it appears critical to investigate the influence of pH on our apo results by performing an additional simulation compatible with pH = 6 conditions. So, in order to propose a starting point for this second simulation, we followed a protocol found in the literature for SARS-CoV-1.[51] We then selected 15 new structures from our pH = 7.4 simulation (3 structures per cluster). For each structure we then protonated the His172 on both protomers, which initiates the structural transformation from pH = 7.4 to pH = 6. The same simulation protocol (see Section 3.2) was followed and a total of 17 μs of simulation was thus generated using the Jean Zay Supercomputer (IDRIS, GENCI, France). In practice, with enough sampling, the structures should be able to relax. Of course, as pointed out by Verma *et al.*,[37] other residues could be impacted by lowering the pH but such simulation has strong interpretative interest. We therefore looked again at all the structural markers described for the previous simulation. We first studied the convergence of some of the properties. Fig. 11 in the ESI† shows that the simulation tends to converge more slowly than at physiological pH and starts to do so beyond 14 μs. Clearly, comparisons of both pH situations would not have been possible using nanosecond simulations even if initial local relaxation of the histidine residues appears to have happened at this timescale. Of course, we cannot state that the simulation is fully converged. However, we stopped the computation when the observed structural changes strongly diminished over time within the ensemble, leaving us with enough confidence in the computed properties. The key result obtained from this second long simulation is the strong variation of the activation features present in the previously described inactive protomer. Indeed, while a significant asymmetry between protomers was found at pH = 6 with protomer 1 exhibiting a poor structure oxyanion hole, the situation evolves with the protonation of His172. Indeed protomer 1 now exhibits a mix of several states with different structural markers (see Fig. 16, ESI†). Compared to pH = 7.4, the interaction of His172/163 with Glu166 changed from a H-bond type interaction (neutral His172/163 at pH = 7.4) to a salt-bridge (positively charged His172 at pH = 6).[70] The stacking index shows that the stacking interaction appears to be weaker than at physiological pH and therefore easier to break and to form (see ESI Fig. 17†). As a result of the protonation, protomer 1 now shows two relatively short maxima for the Glu166–His172 distance (see ESI Fig. 16†) associated with a continuum of values of distances going beyond 6 Å. The protomer 1 Glu166–His172 distance appears to explore a variety of situations including a favorable stacking second minimum which is a sign of a more structured state. However, while some ordered states are found, the absence of stacking is statistically dominant and associated with a striking set of Glu166–His163 interactions. Clearly some really short hydrogen-bonds are found between these residues, a sign of a strong destructuring of the oxyanion hole. These results are in line with the findings of Verma *et al.*[37] that associated the protonation of His172 with the collapse of the oxyanion loop toward the S1 pocket. However, for the other protomer, our apo results differ a bit from Verma *et al.*'s holo data. Indeed, the situation appears more contrasted. Despite a net destructuring effect, protomer 2 tends also to exhibit a mix of states after protonation. The protomer encompasses longer Glu166–His172 interactions than previously noted at physiological pH and the noticeable appearance of some states with short Glu166–His163 distances is observed. However, in the case of protomer 2, the stacking still statistically partially holds despite the existence of a second peak describing a non-negligible absence of stacking in some configurations. Overall, our computations show that the protomers tend to be both affected by the destructuring effect of the His172 protonation, leading to a more symmetrical situation between destructured protomers. Protonation of His172 definitively increases the dynamical aspect of the protease structure and favors the exploration of different states of the activation markers highlighting the instability of the oxyanion hole leading to the partial collapse of the S1 pocket. The impact of the increased flexibility can be further examined through the comparative RMSF of the two simulated pH states where the mobility of the C-terminal end appears further enhanced (see ESI Fig. 17†). This clearly correlates with our initial remark concerning the sampling, that such lower pH structure is far more complex to simulate than the situation at physiological pH as several states resonate due to the low structuring of the oxyanion loop. Finally, Table 2 shows the evolution of the solvation around His163 and His41. The number of water molecules found in the AMOEBA simulation tends to increase on both histidine sites compared to pH = 7.4 with more configurations including one and two water molecules for His163 and His141, respectively. If the presence of a structural water molecule is confirmed around His41, a similar presence around His163 tends to be statistically reinforced under these

protonation conditions. Clearly these findings have potentially an important impact in drug discovery as the presence of structural water molecules around His141 and potentially His163 would make rational drug design more difficult since the substrate or inhibitors would suffer from steric hindrance.[49] The use of PFFs could be critical in the evaluation of the free energies of binding of possible drug candidates. Indeed, our data confirm the high plasticity of the active site observed in X-ray structures[38] at room temperature. Modeling such plasticity including the structuring of the S1 pocket clearly requires the simultaneous capability to accurately evaluate various types of weak interaction including hydrogen bonds, salt bridges and π–π stacking while high-resolution modeling of solvation appears to also be mandatory. Of course, we also showed that extensive sampling beyond the μs-timescale was crucial to deal with such difficult flexible systems.

8 Conclusion and perspectives

In this work, designed in response to the urgent need for COVID-19 research, we demonstrated that it is now possible to perform long μs-timescale MD simulations of large biosystems using polarizable force fields such as AMOEBA that are able to account for physical many-body effects. Due to the inherent complexity of the SARS-CoV-2 proteins, performing such higher-resolution simulations is important as they could provide additional information about the structural dynamics of virus constituents to the COVID-19 experimental and computational research communities. To do so, we proposed a fully unsupervised adaptive sampling strategy that can be used on any type of computational resources. This automated framework allows for production simulations that benefit from advances in supercomputing and from our recent Tinker-HP HPC massively parallel software enhancements, that can now efficiently handle GPU-accelerated large petascale computers using lower precision arithmetic and MPI. In order to extract new information from this type of simulation, we also provided the necessary steps to remove the bias from (re-weight) the obtained data to collect useful and accurate structural dynamics features. More than 38 μs of all-atom MD simulation of the M^{pro} enzyme in its apo (ligand-free) state was produced using the AMOEBA polarizable force field.

Results were then compared to available state-of-the-art large scale simulation data. The results from the new generation PFF were shown to capture most of the structural dynamics features discussed in the experimental literature, confirming that M^{pro} is probably in a poorly active conformation in its apo state under physiological pH conditions. However, simulations detected some partial activity features in one of the protomers linked to a more structured oxyanion hole. This is consistent with the protomeric asymmetric activity observed in the holo state where only one protomer is found to be active,[48] a similar feature that was also observed in SARS-CoV-1.[54] This asymmetry can be related to several structural markers as well as to the total protomer volumes. The active site is found to be highly flexible at room temperature in agreement with recent experimental findings.[38] Overall, the apo state of M^{pro} clearly appears less organized than the holo state in agreement with experimental results discussed by Zhou *et al.*[49] A second simulation, including the protonation of the His172 residue to simulate the system under pH = 6 conditions, was performed and tends to confirm the role of the protonation in the collapse of the S1 pocket at lower pH. Under these conditions, the protomeric AMOEBA asymmetry remains although the protomers tend to be notably destructured. The AMOEBA simulations also captured the C-terminal high flexibility feature discussed in the literature.[49] Flexibility increases at lower pH and tends to further modulate down the activity of the apo state linked with the collapse of the S1 pocket. Striking differences were observed concerning the solvation patterns around the key His41 and His163 residues between AMOEBA and n-PFFs. Overall, the smaller AMOEBA water count around histidines is more in line with experimental data. If the presence of a structural water molecule around His41 is probable at all pH, the existence of a water molecule around His163 tends to be more statistically possible at pH = 6. These results can be explained by the capability of AMOEBA structural water molecules to exhibit an average dipole moment higher than that of bulk water and to explore a wider range of dipoles compared to n-PFFs. Structural water molecules around histidines will clearly affect rational drug design. The use of polarizable force fields could be critical in the evaluation of the free energies of binding of possible drug candidates competing with water to interact with the enzyme. In practice, the M^{pro} enzyme tends to be difficult for molecular mechanics approaches. Indeed, it encompasses all sorts of weak interactions. Therefore, it is not surprising that all the experimentally described features found within the AMOEBA simulations were not necessarily found with the non-polarizable simulations. Such systems tend to require both an accurate force field and an extensive sampling strategy as it is obvious that a few ns of PFF MD alone would not provide insights into a system where the statistical convergence is challenging due to its plasticity. These results provide a first direct validation of the stability of the AMOEBA polarizable force field and clearly demonstrate its applicability at long timescales. Besides correlating with experimental data, our results also show that our adaptive sampling approach coupled with AMOEBA led to enhanced volumes for the active site and to additional potential cryptic pockets as well. As the apo (ligand-free) state has been shown to be a relevant structure at room temperature to perform docking studies,[38] the new information provided could be useful for drug design. Our simulation data are fully available to the general public. They can therefore be used for further structural analysis and/or as an additional basis for ensemble docking studies.[71] Indeed, concentrating the GPU computing power on an apo state is useful to "mine" the conformations to obtain an accurate and more statistically converged set of MD binding site conformations that could be selected by a ligand. The new structural information provided here could help to design new drugs or to repurpose existing ones. These data could also be important to understand chemical reactivity at an atomic level *via* hybrid QM/MM simulations.[68,72] Finally, thanks to the presented divide and conquer strategy, our AMOEBA adaptive MD simulations were

shown to be simultaneously computationally competitive and in line with the available experimental data. Using 100 GPU cards, we show that an acceptable and competitive time to solution could be achieved as our "microsecond" results were obtained in a few days on an academic (and multipurpose) supercomputer. It is worth noting that each simulation could have run on full nodes or using more efficient A100 cards. In practice, a similar exploration of the available community data was already achieved in only 2.5 days (Fig. 1). It is also important to note that Tinker-HP can also produce an order of magnitude faster simulation using n-PFFs using GPUs. Since n-PFF simulations are also of great interest, capturing many experimental aspects, our dual-level (n-PFF + PFF) strategy is confirmed. Indeed, an optimal setup consists in first producing a long adaptive non-polarizable simulation that can be further refined with polarizable potentials within additional adaptive iterations. That way, our approach could also use Folding@-home COVID-19 community results[73] as an input (or any available data shared on the BioExcel/Molssi repository) in order to deliver a maximum of potentially new/useful information into COVID-19 research. Indeed, it is important to recall the importance of proposing accurate (and as much as possible converged) simulations of the COVID-19 targets. As a final perspective, we can mention that the present strategy is platform independent and not limited to supercomputers. Therefore, it can also be used at a smaller scale on "cheaper" laboratory GPU clusters which can benefit from the computational power of low arithmetic to obtain local supercomputing capabilities. On the other side of the spectrum, with the coming of the exascale era and the HPC–Artificial Intelligence (AI) convergence, the "big iron" supercomputer systems, and their cloud-computing counterparts, will considerably extend the high accuracy conformational mining capabilities leading to extended possibilities for the *in silico* modeling of complex biological systems.

Author contributions

T. J. I., F. C., D. El A., and N. L. performed simulations; O. A., T. J. I., and L.-H. J. contributed new code. P. M., T. J. I., J.-P. P, P. R., and L. L. contributed new methodology. N. L., M. M., L. L., F. C., and P. M. contributed analytical tools. F. C., T. J. I., D. El A., N. L., M. M., P. R., and J.-P. P. analyzed data. J.-P. P., P. M., L. L., N. L., T. J. I., F. C. and P. R. wrote the paper; J.-P. P. designed the research.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

This work has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement no. 810367), project EMC2 (JPP). FC acknowledges funding from the French state funds managed by the CalSimLab LABEX and the ANR within the Investissements d'Avenir program (reference ANR11-IDEX-0004-02) and support from the Direction Génerale de l'Armement (DGA) Maîtrise NRBC of the French Ministry of Defense. DEA acknowledges funding from the Lebanese National Council for Scientific Research, CNRS-L. Adaptive sampling computations were performed at GENCI thanks to a COVID19 emergency allocation on the Jean Zay machine (IDRIS, Orsay, France) under grant no. A0070707671 and on the Irene Joliot Curie machine thanks to a PRACE COVID-19 emergency grant (project COVID-HP). The authors thank the Swiss National Supercomputing Center (CSCS) for hosting our data through the FENIX infrastructure. JPP acknowledges a special COVID-19 funding from Sorbonne Université. PR is grateful for support by the Robert A. Welch Foundation (F-1691) and National Institutes of Health (R01GM106137 and R01GM114237).

References

1 J. Guarner, *Am. J. Clin. Pathol.*, 2020, **153**, 420–421.
2 F. Wu, S. Zhao, B. Yu, Y.-M. Chen, W. Wang, Z.-G. Song, Y. Hu, Z.-W. Tao, J.-H. Tian, Y.-Y. Pei, *et al.*, *Nature*, 2020, **579**, 265–269.
3 Z. Jin, X. Du, Y. Xu, Y. Deng, M. Liu, Y. Zhao, B. Zhang, X. Li, L. Zhang, C. Peng, *et al.*, *Nature*, 2020, 1–5.
4 D. Leung, G. Abbenante and D. P. Fairlie, *J. Med. Chem.*, 2000, **43**, 305–341.
5 T. S. Komatsu, Y. Koyama, N. Okimoto, G. Morimoto, Y. Ohno and M. Taiji, Mendeley Data, 2020, DOI: 10.17632/vpps4vhryg.2.
6 *DESRES: Molecular Dynamics Simulations Related to SARS-CoV-2*, 2020, DESRES-ANTON-10880334.
7 M. M. Ghahremanpour, J. Tirado-Rives, M. Deshmukh, J. A. Ippolito, C.-H. Zhang, I. C. de Vaca, M.-E. Liosi, K. S. Anderson and W. L. Jorgensen, *ACS Med. Chem. Lett.*, 2020, **11**(12), 2526–2533.
8 T. Sztain, R. Amaro and J. A. McCammon, *bioRxiv*, 2020, DOI: 10.1101/2020.07.23.218784.
9 S. Piana, P. Robustelli, D. Tan, S. Chen and D. E. Shaw, *J. Chem. Theory Comput.*, 2020, **16**, 2494–2507.
10 J. A. Maier, C. Martinez, K. Kasavajhala, L. Wickstrom, K. E. Hauser and C. Simmerling, *J. Chem. Theory Comput.*, 2015, **11**, 3696–3713.
11 D. E. Shaw, J. P. Grossman, J. A. Bank, B. Batson, J. A. Butts, J. C. Chao, M. M. Deneroff, R. O. Dror, A. Even, C. H. Fenton, A. Forte, J. Gagliardo, G. Gill, B. Greskamp, C. R. Ho, D. J. Ierardi, L. Iserovich, J. S. Kuskin, R. H. Larson, T. Layman, L. Lee, A. K. Lerer, C. Li, D. Killebrew, K. M. Mackenzie, S. Y. Mok, M. A. Moraes, R. Mueller, L. J. Nociolo, J. L. Peticolas, T. Quan, D. Ramot, J. K. Salmon, D. P. Scarpazza, U. B. Schafer, N. Siddique, C. W. Snyder, J. Spengler, P. T. P. Tang, M. Theobald, H. Toma, B. Towles, B. Vitale, S. C. Wang and C. Young, *SC'14: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, 2014, pp. 41–53.

12 I. Ohmura, G. Morimoto, Y. Ohno, A. Hasegawa and M. Taiji, *Philos. Trans. R. Soc., A*, 2004, **372**, 20130387.
13 Y. Shi, P. Ren, M. Schnieders and J.-P. Piquemal, Polarizable force fields for biomolecular modeling, in *Reviews in Computational Chemistry*, ed. A. L. Parrill and K. B. Lipkowitz, John Wiley and Sons, Inc., Hoboken, NJ, 2015, vol. 28, pp. 51–86, DOI: 10.1002/9781118889886.ch2.
14 Z. Jing, C. Liu, S. Y. Cheng, R. Qi, B. D. Walker, J.-P. Piquemal and P. Ren, *Annu. Rev. Biophys.*, 2019, **48**, 371–394.
15 J. Melcr and J.-P. Piquemal, *Front. Mol. Biosci.*, 2019, **6**, 143.
16 F. Célerse, L. Lagardère, E. Derat and J.-P. Piquemal, *J. Chem. Theory Comput.*, 2019, **15**, 3694–3709.
17 L. El Khoury, F. Célerse, L. Lagardère, L.-H. Jolly, E. Derat, Z. Hobaika, R. G. Maroun, P. Ren, S. Bouaziz, N. Gresh, *et al.*, *J. Chem. Theory Comput.*, 2020, **16**, 2013–2020.
18 *GENCI: lutte contre le COVID-19*, online https://www.genci.fr/fr/content/projets-contre-le-covid-19, 2020.
19 *European PRACE Support to Mitigate Impact of COVID-19 Pandemic*, https://prace-ri.eu/prace-support-to-mitigate-impact-of-covid-19-pandemic/, 2020.
20 *United States COVID-19 High Performance Computing Consortium*, https://covid19-hpc-consortium.org/, 2020.
21 L. Lagardère, L.-H. Jolly, F. Lipparini, F. Aviat, B. Stamm, Z. F. Jing, M. Harger, H. Torabifard, G. A. Cisneros, M. J. Schnieders, N. Gresh, Y. Maday, P. Y. Ren, J. W. Ponder and J.-P. Piquemal, *Chem. Sci.*, 2018, **9**, 956–972.
22 O. Adjoua, L. Lagardère, L.-H. Jolly, A. Durocher, Z. Wang, T. Very, I. Dupays, F. Célerse, J. Ponder, P. Ren and J.-P. Piquemal, *J. Chem. Theory Comput.*, 2021, arXiv: 2011.01207.
23 G. R. Bowman, D. L. Ensign and V. S. Pande, *J. Chem. Theory Comput.*, 2010, **6**, 787–794.
24 M. I. Zimmerman, J. R. Porter, X. Sun, R. R. Silva and G. R. Bowman, *J. Chem. Theory Comput.*, 2018, **14**, 5459–5475.
25 R. M. Betz and R. O. Dror, *J. Chem. Theory Comput.*, 2019, **15**, 2053–2063.
26 E. Hruska, J. R. Abella, F. Nüske, L. E. Kavraki and C. Clementi, *J. Chem. Phys.*, 2018, **149**, 244119.
27 L.-H. Jolly, A. Duran, L. Lagardère, J. W. Ponder, P. Ren and J.-P. Piquemal, *LiveCoMS*, 2019, **1**, 10409.
28 H. Abdi and L. J. Williams, *Wiley Interdiscip. Rev. Comput. Stat.*, 2010, **2**, 433–459.
29 F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and E. Duchesnay, *J. Mach. Learn. Res.*, 2011, **12**, 2825–2830.
30 R. T. McGibbon, K. A. Beauchamp, M. P. Harrigan, C. Klein, J. M. Swails, C. X. Hernández, C. R. Schwantes, L.-P. Wang, T. J. Lane and V. S. Pande, *Biophys. J.*, 2015, **109**, 1528–1532.
31 P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. Jarrod Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. Carey, İ. Polat, Y. Feng, E. W. Moore, J. Vand erPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt and SciPy 1.0 Contributors, *Nat. Methods*, 2020, **17**, 261–272.
32 A. Laio and M. Parrinello, *Proc. Natl. Acad. Sci. U. S. A.*, 2002, **99**, 12562–12566.
33 P. Y. Ren and J. W. Ponder, *J. Phys. Chem.*, 2003, **107**, 5933–5947.
34 Y. Shi, Z. Xia, J. Zhang, R. Best, C. Wu, J. W. Ponder and P. Ren, *J. Chem. Theory Comput.*, 2013, **9**, 4046–4063.
35 J. W. Ponder, C. Wu, P. Ren, V. S. Pande, J. D. Chodera, M. J. Schnieders, I. Haque, D. L. Mobley, D. S. Lambrecht, R. A. DiStasio, M. Head-Gordon, G. N. I. Clark, M. E. Johnson and T. Head-Gordon, *J. Phys. Chem. B*, 2010, **114**, 2549–2564.
36 C. Zhang, C. Lu, Z. Jing, C. Wu, J.-P. Piquemal, J. W. Ponder and P. Ren, *J. Chem. Theory Comput.*, 2018, **14**, 2084–2108.
37 N. Verma, J. A. Henderson and J. Shen, *J. Am. Chem. Soc.*, 2020, **142**, 21883–21890.
38 D. Kneller, G. Phillips, H. O'Neill, R. Jedrzejczak, L. Stols, P. Langan, A. Joachimiak, L. Coates and A. Kovalevsky, Structural plasticity of SARS-CoV-2 3CL M^{pro} active site cavity revealed by room temperature X-ray crystallography, *Nat. Commun.*, 2020, **11**, 3202.
39 J. A. Rackers, Z. Wang, C. Lu, M. L. Laury, L. Lagardère, M. J. Schnieders, J.-P. Piquemal, P. Ren and J. W. Ponder, *J. Chem. Theory Comput.*, 2018, **14**, 5273–5289.
40 L. Lagardère, F. Aviat and J.-P. Piquemal, *J. Phys. Chem. Lett.*, 2019, **10**, 2593–2599.
41 *Data Tinker-HP*, SARS–CoV–2 Main Protease, deposited at CSCS, 2020.
42 A. Amadei, A. B. Linssen and H. J. Berendsen, *Proteins*, 1993, **17**, 412–425.
43 A. Amadei, A. Linssen, B. De Groot, D. Van Aalten and H. Berendsen, *J. Biomol. Struct. Dyn.*, 1996, **13**, 615–625.
44 H. J. Berendsen and S. Hayward, *Curr. Opin. Struct. Biol.*, 2000, **10**, 165–169.
45 M. Ester, H.-P. Kriegel, J. Sander and X. Xu, *et al.*, *Kdd*, 1996, pp. 226–231.
46 Y. Liu, Z. Li, H. Xiong, X. Gao and J. Wu, *2010 IEEE International Conference on Data Mining*, 2010, pp. 911–916.
47 C. D. Owen, P. Lukacik, C. M. Strain-Damerell, A. Douangamath, A. J. Powell, D. Fearon, J. Brandao-Neto, A. D. Crawshaw, D. Aragao, M. Williams, R. Flaig, D. Hall, K. McAauley, D. I. F. Stuartvon Delft and M. A. Walsh, PDB 6Y84: Structure COVID-19 main protease with unliganded active site, 2020, https://www.wwpdb.org/.
48 L. Zhang, D. Lin, X. Sun, U. Curth, C. Drosten, L. Sauerhering, S. Becker, K. Rox and R. Hilgenfeld, *Science*, 2020, **368**, 409–412.
49 X. Zhou, F. Zhong, C. Lin, X. Hu, Y. Zhang, B. Xiong, X. Yin, J. Fu, W. He, J. Duan, *et al.*, *Sci. China: Life Sci.*, 2020, 1–4.
50 H. Yang, M. Yang, Y. Ding, Y. Liu, Z. Lou, Z. Zhou, L. Sun, L. Mo, S. Ye, H. Pang, *et al.*, *Proc. Natl. Acad. Sci. U. S. A.*, 2003, **100**, 13190–13195.

51 J. Tan, K. H. Verschueren, K. Anand, J. Shen, M. Yang, Y. Xu, Z. Rao, J. Bigalke, B. Heisen, J. R. Mesters, K. Chen, X. Shen, H. Jiang and R. Hilgenfeld, *J. Mol. Biol.*, 2005, **354**, 25–40.
52 H. Yang, M. Yang, Y. Ding, Y. Liu, Z. Lou, Z. Zhou, L. Sun, L. Mo, S. Ye, H. Pang, G. F. Gao, K. Anand, M. Bartlam, R. Hilgenfeld and Z. Rao, *Proc. Natl. Acad. Sci. U. S. A.*, 2003, **100**, 13190–13195.
53 L. Zhang, D. Lin, X. Sun, U. Curth, C. Drosten, L. Sauerhering, S. Becker, K. Rox and R. Hilgenfeld, *Science*, 2020, **368**, 409–412.
54 H. Chen, P. Wei, C. Huang, L. Tan, Y. Liu and L. Lai, *J. Biol. Chem.*, 2006, **281**, 13894–13898.
55 D. Branduardi, F. L. Gervasio, A. Cavalli, M. Recanatini and M. Parrinello, *J. Am. Chem. Soc.*, 2005, **127**, 9147–9155.
56 J. Hermans, in *Peptide Solvation and HBonds*, Academic Press, 2005, vol. 72, Advances in Protein Chemistry, pp. 105–119.
57 R. S. Paton and J. M. Goodman, *J. Chem. Inf. Model.*, 2009, **49**, 944–955.
58 J. A. Lemkul, J. Huang, B. Roux and A. D. MacKerell, *Chem. Rev.*, 2016, **116**, 4983–5013.
59 S. Cardamone, T. J. Hughes and P. L. A. Popelier, *Phys. Chem. Chem. Phys.*, 2014, **16**, 10367–10387.
60 B. Goyal and D. Goyal, *ACS Comb. Sci.*, 2020, **22**, 297–305.
61 J. Liang, C. Karagiannis, E. Pitsillou, K. K. Darmawan, K. Ng, A. Hung and T. C. Karagiannis, *Comput. Biol. Chem.*, 2020, 107372.
62 J. R. Wagner, J. Sørensen, N. Hensley, C. Wong, C. Zhu, T. Perison and R. E. Amaro, *J. Chem. Theory Comput.*, 2017, **13**, 4584–4592.
63 A. Volkamer, D. Kuhn, F. Rippmann and M. Rarey, *Bioinformatics*, 2012, **28**, 2074–2075.
64 B. Goyal and D. Goyal, *ACS Comb. Sci.*, 2020, **22**, 297–305.
65 P. Schmidtke and X. Barril, *J. Med. Chem.*, 2010, **53**, 5858–5867.
66 Y. Levy and J. N. Onuchic, *Annu. Rev. Biophys. Biomol. Struct.*, 2006, **35**, 389–415.
67 B. de Courcy, J.-P. Piquemal, C. Garbay and N. Gresh, *J. Am. Chem. Soc.*, 2010, **132**, 3312–3320.
68 K. Świderek and V. Moliner, *Chem. Sci.*, 2020, **11**, 10626–10630.
69 J. Yang, M. Yu, Y. N. Jan and L. Y. Jan, *Proc. Natl. Acad. Sci. U. S. A.*, 1997, **94**, 1568–1572.
70 S.-M. Liao, Q.-S. Du, J.-Z. Meng, Z.-W. Pang and R.-B. Huang, *Chem. Cent. J.*, 2013, **7**, 44.
71 R. E. Amaro, J. Baudry, J. Chodera, Ö. Demir, J. A. McCammon, Y. Miao and J. C. Smith, *Biophys. J.*, 2018, **114**, 2271–2278.
72 D. Loco, L. Lagardère, G. A. Cisneros, G. Scalmani, M. Frisch, F. Lipparini, B. Mennucci and J.-P. Piquemal, *Chem. Sci.*, 2019, **10**, 7200–7211.
73 M. I. Zimmerman, J. R. Porter, M. D. Ward, S. Singh, N. Vithani, A. Meller, U. L. Mallimadugula, C. E. Kuhn, J. H. Borowsky, R. P. Wiewiora, M. F. D. Hurley, A. M. Harbison, C. A. Fogarty, J. E. Coffland, E. Fadda, V. A. Voelz, J. D. Chodera and G. R. Bowman, *bioRxiv*, 2020, DOI: 10.1101/2020.06.27.175430.

2.6.2 Publication 2

THE JOURNAL OF PHYSICAL CHEMISTRY LETTERS
A JOURNAL OF THE AMERICAN CHEMICAL SOCIETY

pubs.acs.org/JPCL Letter

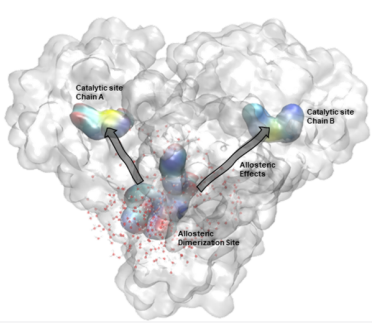
Interfacial Water Many-Body Effects Drive Structural Dynamics and Allosteric Interactions in SARS-CoV-2 Main Protease Dimerization Interface

Dina El Ahdab, Louis Lagardère, Théo Jaffrelot Inizan, Fréderic Célerse, Chengwen Liu, Olivier Adjoua, Luc-Henri Jolly, Nohad Gresh, Zeina Hobaika, Pengyu Ren, Richard G. Maroun, and Jean-Philip Piquemal*

Cite This: *J. Phys. Chem. Lett.* 2021, 12, 6218−6226 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Following our previous work (*Chem. Sci.* 2021, *12*, 4889−4907), we study the structural dynamics of the SARS-CoV-2 Main Protease dimerization interface (apo dimer) by means of microsecond adaptive sampling molecular dynamics simulations (50 μs) using the AMOEBA polarizable force field (PFF). This interface is structured by a complex H-bond network that is stable only at physiological pH. Structural correlations analysis between its residues and the catalytic site confirms the presence of a buried allosteric site. However, noticeable differences in allosteric connectivity are observed between PFFs and non-PFFs. Interfacial polarizable water molecules are shown to appear at the heart of this discrepancy because they are connected to the global interface H-bond network and able to adapt their dipole moment (and dynamics) to their diverse local physicochemical microenvironments. The water−interface many-body interactions appear to drive the interface volume fluctuations and to therefore mediate the allosteric interactions with the catalytic cavity.



In the context of COVID-19 drug discovery, both structural and nonstructural proteins are considered as promising targets for the development of antiviral agents against the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2).[1] Specifically, SARS-CoV-2 M^{pro} plays a pivotal role in controlling viral replication and transcription through proteolytic processing of viral poly proteins.[2] Many studies on inhibitor ligands are based on active site pocket targeting. However, advancing a drug toward clinical trials remains a daunting task[3] (as was the case for SARS-Cov1[4,5]). In practice, because of the dimeric nature of M^{pro}, another strategy can be employed to inhibit its activity through the development of dimerization inhibitors.[2,6] Indeed, dimerization inhibitor design was previously reported for many viral enzymes such as the HIV reverse transcriptase, integrase, herpes simplex virus ribonucleotide reductase, and DNA polymerase.[6,7] In fact, targeting dimerization could potentially affect the substrate pocket and thus inhibit the M^{pro} activity because of allosteric connectivity between the dimerization site and the catalytic site.[2,8] Recently, we provided extensive simulations on M^{pro}[9] using the AMOEBA polarizable force field (PFF)[10−12] and a new highly parallel GPUs-accelerated[13,14] unsupervised adaptive sampling strategy.[9] These multimicrosecond simulations and their associated conformational spaces were compared to available non-PFF long-time scale simulation data from D. E. Shaw Research (DESRES)[15] and RIKEN Center for Biosystems Dynamics Research.[16] It was found that AMOEBA results were closely correlated with experimental data, highlighting the observed strong flexibility of M^{pro}.[17] However, important differences in structural dynamics were observed compared to non-PFFs in key areas of the protease. For example, the overall richer conformational space led to enhanced volume cavities and to different solvation patterns within the active site. In order to drive further our high-resolution M^{pro} analysis, we present here a study of the factors structuring the dimerization interface as a function of different pH and solvation patterns. We particularly focus on the study of the role of many-body effects in the modeling of interfacial water and on their impact in allosteric interactions of the dimerization interface with other cavities/sites. To do so, we analyze more than 50 μs (including more than 12 μs of new simulations produced for the study) of AMOEBA molecular dynamics simulations and more than 110 μs of additional non-PFF simulations from other available data sets. All simulation details can be found in Theoretical Methods at the end of this Letter.

Received: May 6, 2021
Accepted: June 10, 2021

https://doi.org/10.1021/acs.jpclett.1c01460
J. Phys. Chem. Lett. 2021, 12, 6218−6226

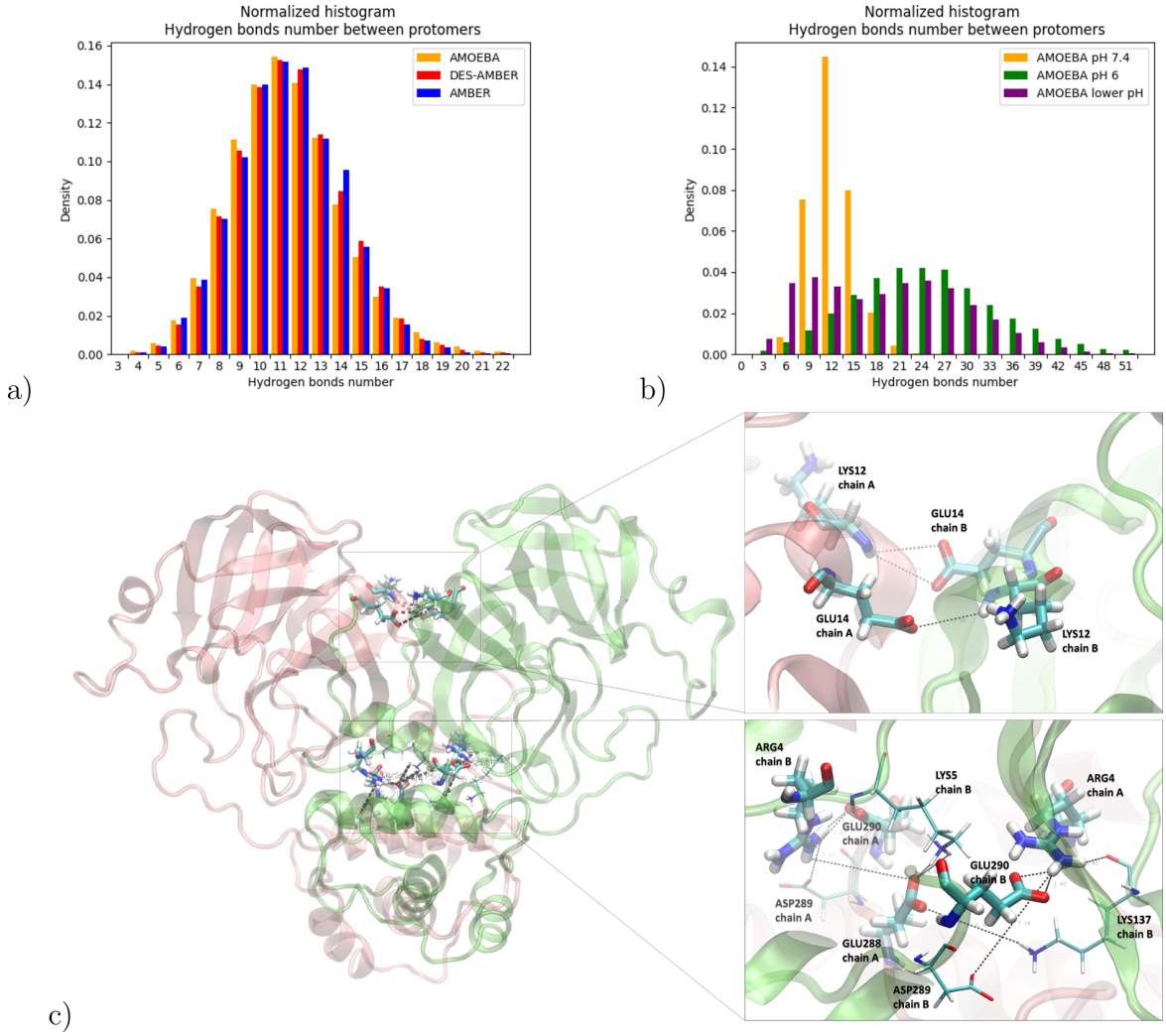


Figure 1. Histogram representation of H-bond probability density for (a) DES-AMBER, AMBER, and AMOEBA force fields at pH 7.4 and for AMOEBA trajectories at pH 7.4, 6, and lower. (b) Representation of the most frequent H-Bond interactions at the dimerization interface. Chains **A** and **B** are presented in pink and lime, respectively, (c).

To start our analysis of the M^{pro} structural dynamics at the dimerization interface, we determined the number of hydrogen bond (H-bond) interactions in order to evaluate the robustness of noncovalent interactions between the two protomers. Starting at physiological pH, we analyzed the DES-AMBER (DESRES), AMBER (RIKEN), and AMOEBA (Tinker-HP) trajectories (see Theoretical Methods for details) provided within the available conformation ensembles. We found relatively similar H-bond interaction probability density functions between the three profiles (see Figure 1a) that all present strong stability of the dimerization interface. Comparing the physiological H-bond distribution to lower pH AMOEBA simulations (see Figure 1b), we found a transition from a sharp Gaussian distribution centered at 14 H-bonds (pH 7.4) to a more diffuse one at pH 6 and below, exhibiting the involvements of weaker, disorganized, interactions. Clearly, our results show a collapse of the dimer interface at pH values lower than physiological as a consequence of the successive protonations of histidine residues (**His172** then **His163**).[9,18,19] Among the observed interactions (see Table 1 in the Supporting Information), **Arg4−Glu290** and **Gly11−Glu14** H-bond interactions have the highest probability density of all over DES-AMBER, AMBER, and AMOEBA trajectories at physiological pH. However, these interactions are not detected at lower pH, which is consistent with experimental studies reporting that low pH is responsible for the loss of the dimer interface.[20,21] It is important to note here that protonation of **His172** at lower pH has recently been shown[9,17,19] to be the source of a partial collapse in the catalytic site as well. Because the dimer interface is known to be fully functional at physiological pH, our multi-pH results reinforce the critical role of the **His172** protonation state and are consistent with Verma et al. findings[19] of a nonprotonated **His172** at physiological pH. A detailed look at the H-bond interaction profile in Table 1 of the Supporting Information highlights the key role of **Arg4** in maintaining the dimerization through several interactions, mainly with **Glu290** but also with **Lys137**, **Ser139**, **Glu288**, and **Asp289** at physiological pH. This is consistent with the description of key residues for the maintenance of SARS-CoV-2 M^{pro} dimerization in the experimental literature:[22] **Arg4**, **Ser10**, **Gly11**, **Glu14**, **Asn28**, **Ser139**, **Phe140**, **Ser147**, **Glu166**, **Glu290**, and **Arg298**. These residues all appear along our analysis, except for **Ser147**. Nevertheless, we were capable here of expanding

https://doi.org/10.1021/acs.jpclett.1c01460
J. Phys. Chem. Lett. 2021, 12, 6218−6226

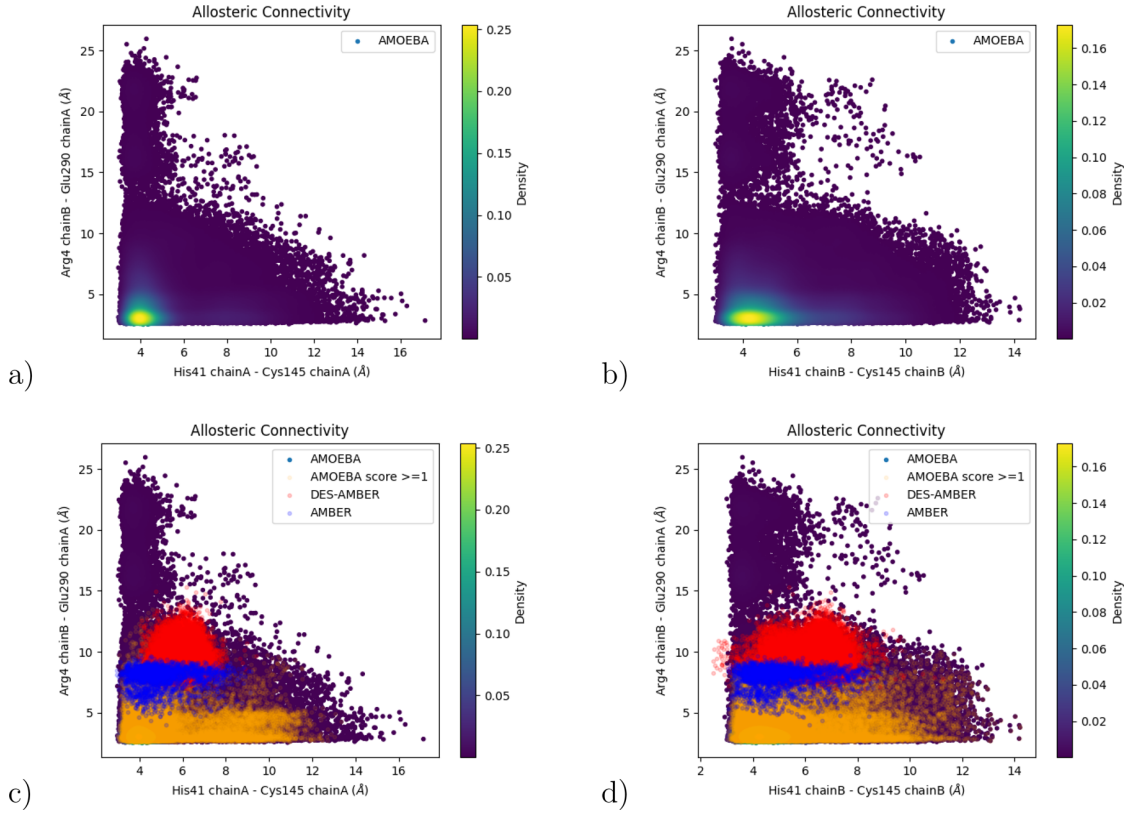


Figure 2. 2D plot representation of ***Arg4*** *chain* ***B−Glu290*** *chain* ***A*** distances vs ***His41*** *chain* ***A−Cys145*** *chain* ***A*** distances (a and c) and vs (b and d) His41 chain **B−Cys145** chain **B**. In panels c and d we have projected on the AMOEBA 15.14 μs, DES-AMBER 100 μs, AMBER 10 μs, and AMOEBA frames with a reweighting score greater than 1.

the list of these residues after a detailed analysis of DES-AMBER, AMBER, and AMOEBA simulations. As shown in Table 1 (Supporting Information), AMOEBA predicts a richer, more exhaustive, list of dimerization-implied residues compared to AMBER and DES-AMBER. The detected special forms of H-bond and other interactions, at physiological pH, are highlighted in Figure 1c. It is important to note that when successive histidine protonations occur, **His172** and **His163** switch from neutral histidines at pH 7.4 to positively charged at pH 6 and below, changing the nature of some of their interactions with other residues and water (for example, moving from H-bonds to salt-bridges in some cases[9,23]). Although pH lowering will affect also other residues that are not all considered in our computations,[19] this physicochemical change in the nature of the histidines interactions is central to the weakening of the interface stability, forcing it to redistribute its H-bond network into a different and less structured configuration. Finally, Table 1 (Supporting Information) also reveals that the Arg4−Glu290 and Gly11−Glu14 interactions are the most important H-bonds responsible for the stabilization of the dimerization interface because they exhibit the highest densities at physiological pH and are absent in the lower pH simulations. Overall, these results highlight the fact that the complex H-bond network is the one driving force stabilizing the interface.

To probe deeper into the complexity of the dimerization interface, we decided to look at its potential allosteric interactions within M^{pro}. Allostery occurs when conformational changes happening at one site of a protein and causing structural or dynamical changes at a topologically independent distant site. Such changes lead to a reduction or an increase in catalytic activity among other structural rearrangements. Structure-based prediction of allosteric sites, modulators, and communication pathway is important for a basic understanding of proteins and can lead drug discovery in order to regulate protein function.[24,25] Because H-bonds play a very important role in the dimerization region, they may be able to influence its volume, which could also have structural effects on other protein surface pockets via allosteric correlations.[24] The druggability of the dimerization interface has been discussed in the literature,[9,20] but fewer contributions looked at the potential allosteric interactions. Indeed, the importance of allosteric connectivity between allosteric and functional sites has been increasingly witnessed during recent years.[26,27] Several potential allosteric sites were recently discussed in order to offer allosteric drug target strategies[28−30] inside SARS-CoV-2 M^{pro}. For example, Stromich et al.[29] studied the scoring of putative allosteric sites and underlined a zone located in the dimerization site showing a high connectivity toward the catalytic active site. They proposed the definition of a potential allosteric dimerization site formed by the six following residues of the interface: **Arg131**, **Asp197**, **Thr199**, **Asp289**, and **Glu290** from chain **A** and **Arg4** from chain **B**. Because several of these residues were shown by our simulations to be instrumental to the interface stabilization (see Table 1, Supporting Information and previous discussion), we decided to study this site. In order to assess for a potential allosteric connectivity of the allosteric dimerization site toward both chains of the catalytic active site and to analyze its structural dynamics, we resorted to extensive bond-to-bond propensity

https://doi.org/10.1021/acs.jpclett.1c01460
J. Phys. Chem. Lett. 2021, 12, 6218−6226

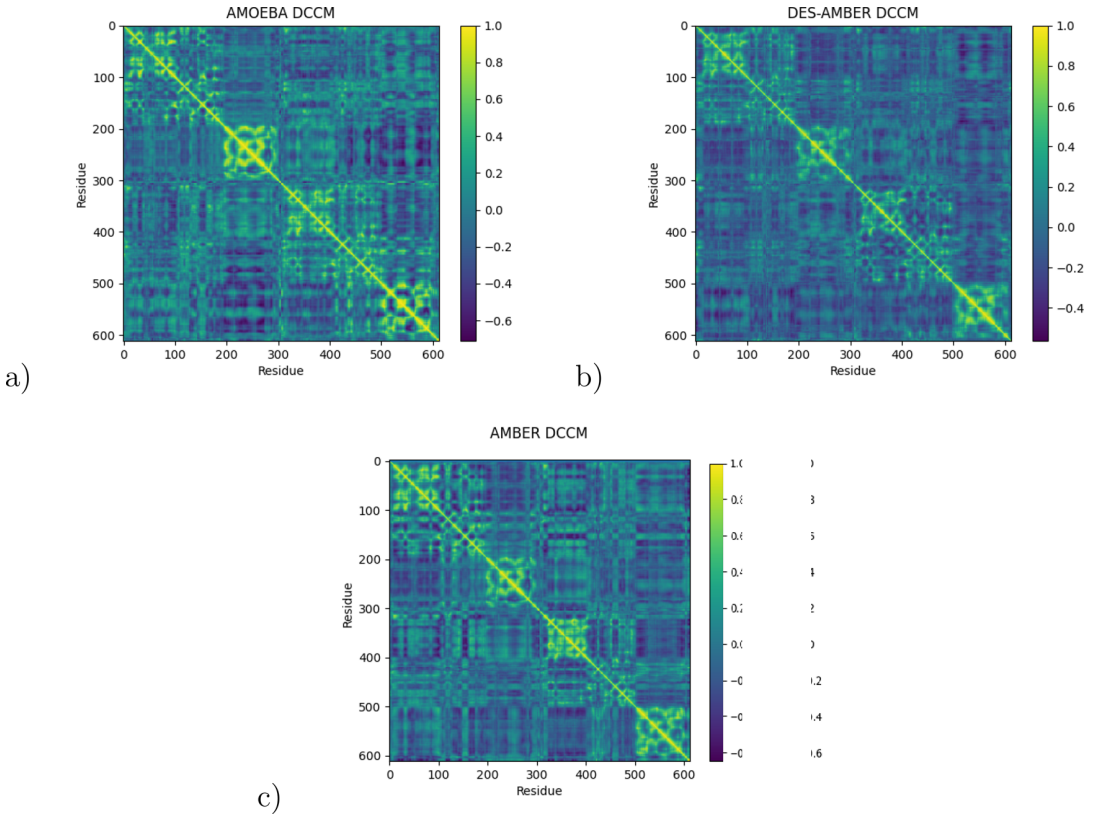


Figure 3. Dynamic cross-correlation maps using the C_α atom of each residue for (a) AMOEBA, (b) DES-AMBER, and (c) AMBER trajectories.

analysis.[31] Using this approach, we measure the fluctuations of given sets of atom−atom interactions and analyze how they affect any other set of interactions located elsewhere within the protein, allowing therefore to measure their instantaneous connectivity at each moment of the dynamics. We calculated first the evolution of distances located inside the allosteric dimerization site with other characteristic distances implicated in the residues forming the catalytic dyad. That way, thanks to well-chosen reference atoms or residues, this study informs us indirectly of the coevolution of the two cavity volumes. Indeed, comparing their volume fluctuations along trajectories can tell us about a possible allosteric connectivity between them.[29,32] We show in Figure 2a,b a 2D plot graphic of the distances separating the residues of the catalytic dyad for both chains **A** and **B** versus the distances between residues from the allosteric dimerization site: **Arg4** chain **B** and **Glu290** chain **A** because they present a robust interaction. AMOEBA trajectories show a high density of structures having both narrow catalytic and allosteric dimerization sites, respectively, around 4 and 3 Å, as shown in Figure 2a,b. However, we are also able to detect a different organization of the structures that are characterized by a narrow allosteric dimerization site and a relaxed catalytic site and, conversely, proposing possible allosteric connectivity between the sizes of the catalytic and allosteric dimerization sites. This additional connectivity found in the AMOEBA simulations is not observed in DES-AMBER nor in AMBER simulations (Figure 2c,d). Within our adaptive sampling scheme, the score is defined as the ratio between the probabilities to obtain the structure q_i in the biased simulation and in an unbiased simulation. Here, we limit ourselves to structures with a reweighting score greater than 1 as they are more likely to be visited during a conventional MD simulation. In contrast, frames with scores less than 1 have been favored by the adaptive algorithm to maximize exploration and are thus less physically relevant to the system statistic (more information can be found in ref 9). Thus, structures presented in orange in Figure 2 are more representative of the true AMOEBA statistics. In this case, we detect mostly structures having a relaxed catalytic site and a narrow allosteric dimerization site. This suggests that this specific dependency is detected thanks to the use of the polarizable AMOEBA FF, whereas the adaptive algorithm sampling is the one responsible for detecting structures associated with both a narrow catalytic site and a relaxed allosteric dimerization site. Similar conclusions can be reached upon considering **Arg131**, **Asp197**, and **Thr199** instead of **Glu290**, as shown in Figure 1 of the Supporting Information. These observations demonstrate the importance of the coupling of the adaptive sampling algorithm to the AMOEBA PFF for bringing out conformations that have escaped nonpolarizable standard MD simulations.

Because some allosteric connection was found between the dimerization and the active sites, we decided to provide another view of the simulation differences observed with the different force fields. To do so, we performed dynamic cross-correlation map (DCCM) analysis[33,34] for the three trajectories. DCCM allows us to investigate the dynamical changes of the system over time and to quantify the correlation coefficients of motions between atoms. The first result to point out is that as seen previously, AMOEBA data differ from the AMBER/DES-AMBER data. DCCM shows more positive/negative values than those obtained from non-PFFs, indicating a stronger correlated/anticorrelated atom motion in PFF simulations (see Figure 3). It is worth mentioning that strong

https://doi.org/10.1021/acs.jpclett.1c01460
J. Phys. Chem. Lett. 2021, 12, 6218−6226

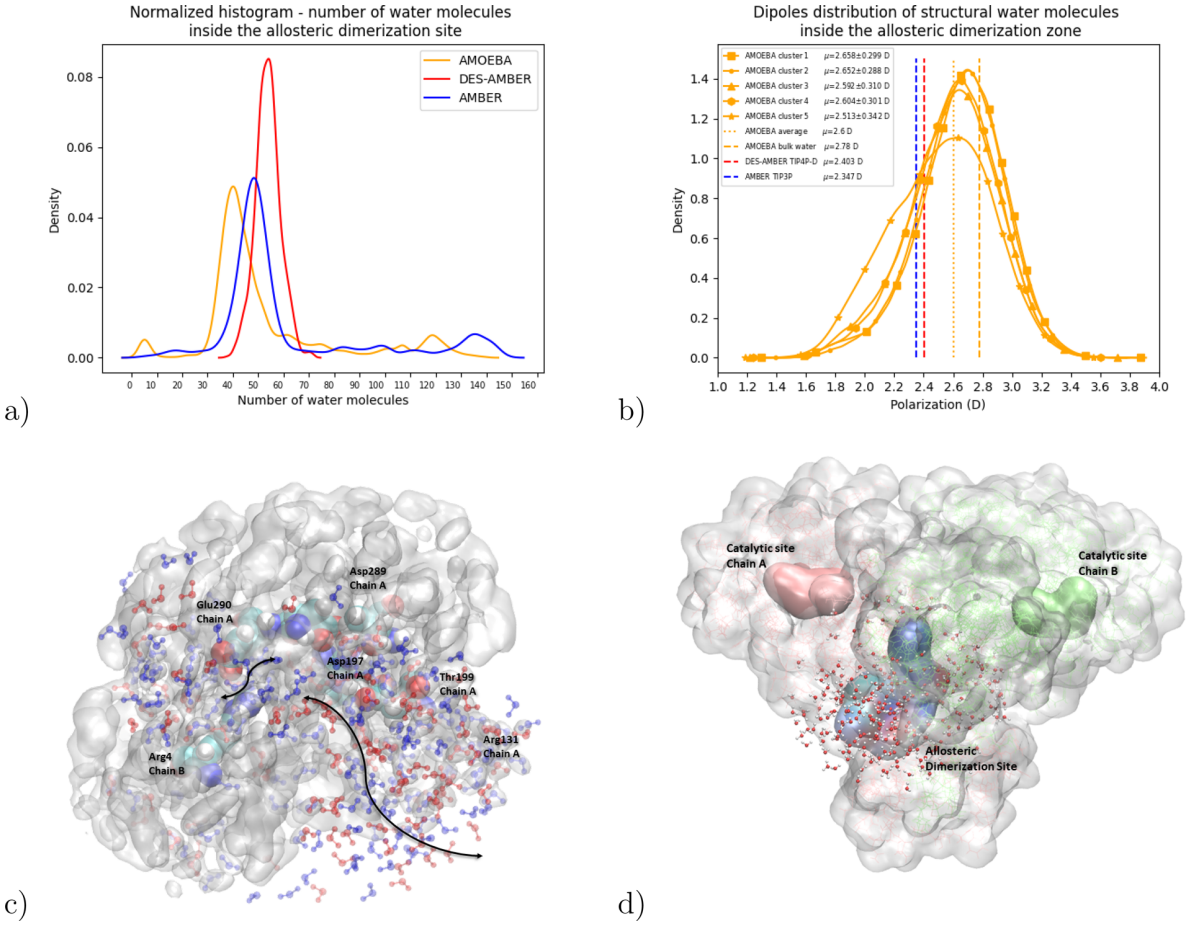


Figure 4. Representation of (a) the probability of structural water molecules number inside the allosteric dimerization site and (b) their dipoles distribution. (c) Representation of the water dipole distribution inside the allosteric dimerization site. Water molecules layered with red have dipole moment ≤2.78 D; those layered with blue have dipole moment ≥2.78 D. Asp and Glu have electrically charged side chains (acidic). Arg have electrically charged side chains (basic). Thr has polar side chain. The distance between **Arg4** and **Glu290** is 5.29 Å. Residues within 10 Å of the allosteric dimerization site are presented in quicksurf mode in white. Black arrows show the flow of water molecules in this site. (d) Global view of the M^{pro}, showing the catalytic site of both chain **A** and **B** and the allosteric dimerization site. Water molecules within 10 Å of the allosteric dimerization site are presented in cpk mode.

anticorrelation motions are observed between the α-helical region of each protomer of M^{pro} (a region strongly participating to the dimerization, i.e., residue range of 220−280 and 470−570) in AMOEBA trajectories. By contrast, the corresponding regions have much weaker (anti)correlation in both DES-AMBER and AMBER trajectories. Figure 2 in the Supporting Information proposes a closer analysis of the regions of interest for the allosteric interactions (i.e., the allosteric dimerization site) and reveals a more global anticorrelated motion between the residues of the allosteric dimerization site and the catalytic dyad of chain **A** than in AMBER/DES-AMBER. For chain **B**, this anti-correlation of the dimerization site with the catalytic dyad residues is also found. In all cases, the stronger correlation DDCM values are found within the AMOEBA simulation. The most positive correlation is found for Cys145 (chain B) and Arg4 (chain B) as the most negative correlation is found for Cy145 (chain A) and Glu290 (chain A). This further confirms the presence of an allosteric correlation between the sites and also supports the hypothesis of a strong asymmetry between protomers.[9]

As our previous analysis confirmed the differences between FF simulations, resulting in different predictions of allosteric connections and correlated motions between sites, we attempted to trace back the discrepancies studying the overall structural dynamics of the interface. As we explained in the first section, the dimerization interface overall stability is linked to a complex H-bond network that is exposed to the water solvent. Within M^{pro}, cavities and pocket volume fluctuations lead to water molecule traffic which is essential to maintain the protein structure. In a sense, the allosteric connection is performed "through water" and the resulting analysis of its presence is therefore impacted by the quality of water modeling. In practice, water molecules are commonly found within enzymatic sites, can form water bridges between the residues, and thus maintain protein secondary structures via H-bond interactions (see ref 35 and references therein). Using polarizable force fields, it has been demonstrated that some structural water molecules exhibit enhanced dipole moments, in kinase active sites for example.[36] Our previous work on M^{pro} clearly also demonstrated a very different behavior of water molecules when they are modeled with the AMOEBA PFF, which takes into account many-body effects.[9] Because water plays an important role in structural and functional activities, we looked for the water molecules present around some key

https://doi.org/10.1021/acs.jpclett.1c01460
J. Phys. Chem. Lett. 2021, 12, 6218−6226

interface residues at physiological pH. To do so, we considered a 3.5 Å radius sphere centered at the atom capable of being engaged in hydrogen bonds with water for the most important residues involved in noncovalent interactions between protomers, namely: **Arg4**, **Glu290**, **Gly11**, and **Glu14**. The number of detected water molecules (see Figure 3 in the Supporting Information), presents notably different distribution profiles depending on the simulations: AMOEBA polarizable water, DES-AMBER(TIP4D), and AMBER (TIP3P). In fact, the number of water molecules detected strongly depends on the type of residue, on the considered M^{pro} chain, and on the force field itself. **Arg4** of chain **A**, for example, is found to be mostly interacting with one water molecule for AMBER, 1−2 molecules for DES-AMBER, and 2−3 molecules for AMOEBA. However, **Arg4** of chain **B** is found to interact mostly with 3 water molecules for AMBER and DES-AMBER and with 2 molecules for AMOEBA in line with the predicted asymmetry between protomers found in M^{pro}.[9] Although water traffic is detected for all force fields, the solvation patterns and differences between force fields appear to be residue-dependent. Water molecules extracted from AMOEBA trajectories around the concerned residues are polarizable (and the water model is flexible[10]), and therefore, their distribution is mainly controlled by the physicochemical nature of the residues (polar, apolar, positively/negatively charged, etc.) generating specific polarizing fields. In practice, the AMOEBA bulk water average dipole moment amounts to 2.78 D, in nice agreement with experiment, whereas non-PFF models exhibit smaller fixed dipole moments of 2.40 and 2.35 D for TIP4P-D and TIP3P, respectively. Figure 4 in the Supporting Information shows the average dipole values for the water molecules in the vicinity of the targeted residues. Their mean values (around 2.6 D on average) is below the bulk AMOEBA reference value. This result is consistent with the idea that the dense interface environment generates a global many-body depolarizing effect (compared to bulk water) influencing the water molecule-induced dipoles. Overall, the interface H-bond network connects to the solvent's own H-bond pattern forming a higher level of complexity. Clearly, the water molecule behavior is strongly influenced by the nature of the interface residues through many-body effects, generating various microsolvation patterns according to the local environment. These patterns are themselves affected by their interactions with the solvent in a self-consistent fashion.

In order to further evaluate the difference in solvation patterns, we focused on the previously introduced allosteric dimerization site, a specific location within the interface that allows for water molecules to circulate between the interface residues. To get a better understanding of what is happening, we have to evaluate the number of water molecules present and their lifetimes within this site. It is important to mention here that the six residues forming the allosteric site at the dimerization interface are either ionic or polar. Asp and Glu are negatively charged, whereas His is positively charged. Side-chains such as Thr can retain water molecules inside the cavity. Black arrows in Figure 4 display the flow of water molecules in the buried site. Because the greatest distance separating **Arg4** chain **B** and **Glu290** chain **A** is around 24 Å, we defined a sphere with a (cutoff) radius of 10 Å, centered at the geometrical center of the six residues forming the pocket at the allosteric dimerization site, and calculated the number of water molecules present within this sphere. Figure 4a shows a striking difference between AMOEBA and non-PFF simulations. PFF simulations give far fewer water molecules inside the allosteric dimerization site and a highest probability density of presence centered at 40, to be compared with 50 for AMBER and 55 for DES-AMBER.

We then measured the water lifetimes in the 10 Å sphere using the 400 ns CMD simulations produced with both the AMBER and AMOEBA force fields. We observed an average water lifetime of 0.171 ns for AMBER and a longer lifetime of 0.516 ns for AMOEBA. This clearly shows that many-body polarization effects tend to act as glue between the dimerization interface and the water molecules, specifically at the allosteric dimerization site, retaining them longer at the surface of the residues of the dimerization site (Figure 5 in the Supporting Information). Putting these two findings together allows us to better understand why the water dynamics outside the interface is so different from the (slower) dynamics found in the most confined part of the dimerization allosteric site. The smaller number of water molecules inside the allosteric dimerization site reflects therefore a slower water traffic, because these polarized water molecules tend to move slowly, being engaged into many more H-bonds. Indeed, the AMOEBA diffusion constant is more in line with experiment than the TIP3P and TIP4-D models. However, as we discussed, the AMOEBA water dipole moment values can present strong local variations because of the local micro-solvation patterns that cannot be captured by the mean-field approximation, which is the basis of classical non-PFFs.[35] As for the previous situation, Figure 4 displays a rather underpolarized global situation for water that exhibits an average dipole moment lower than that of the bulk. Nevertheless, Figure 4 also highlights the collection of multiple different situations where the microsolvation patterns tend to generate simultaneously partial distributions of highly polarized and underpolarized water molecules in the allosteric dimerization site because this distribution is mainly controlled by the physicochemical nature of the residues. As shown in Figure 4c and in Figure 6 in the Supporting Information, mostly underpolarized water molecules are found in the most buried section of the allosteric dimerization site where confinement generates more depolarizing effects. These are well-known to decrease the average dipole moment values of confined waters and are observed here. Again, AMOEBA exhibits a higher probability density lower than bulk at 2.6 D, whereas DES-AMBER and AMBER water dipoles remain fixed at 2.403 and 2.347 D, respectively (see Figure 4b). Figures 4b also provides a view of the average dipole moments found after clusterization of the AMOEBA trajectories (see ref 9 for more information about the five different clusters). The site maintains a relatively stable average dipole solvent value because of the fluctuation of both the volumes (i.e., different in the different clusters) and the number of water molecules (see Figure 7 in the Supporting Information), highlighting the interconnection of the interface H-bond network and the solvent. This suggests that there is a complex interplay between the distribution of dipoles of polarizable water molecules and the residues (and associated volumes) of the dimerization allosteric site. This interaction network contributes to regulating the allosteric effects with the catalytic site of both protomers. Modeling such connections between cavities requires capturing the subtle equilibrium between the protein and solvent dynamics. The dipolar fluctuations of the water traffic tend to be extremely complex, leading to dramatically different behavior in different parts of the interface where the

https://doi.org/10.1021/acs.jpclett.1c01460
J. Phys. Chem. Lett. 2021, 12, 6218−6226

local water dynamics can be quite different (i.e., for the AMOEBA-predicted dynamic slowdown within the buried allosteric dimerization site, etc.). Such water traffic shapes the interface and participates in modulating the allosteric dimerization site structural "breathing" that is involved in the overall allosteric effects with the main catalytic site. Such critical involvement of the "polarizable" water molecule within recognition or regulatory sites of proteins had been postulated before,[36] and it is clear that the number of water molecules within a binding site matters. Indeed, waters interacting with their close environment via through-water binding modes are common and able to strongly influence local electronic properties.[37] Through-water configurations can mediate interactions between an inhibitor (see for example refs 36 and 38) and indirectly bound residues of the recognition site. In such situations, also considered in the context of pFFs, an accurate count of water molecules can be critical because many-body effects (particularly the polarization energy) could tip the (free) energy balance between competing inhibitors. Missing this aspect within the modeling certainly results in a loss in the prediction of signal in the allosteric communication. It is also important to mention that beyond this energetic view of the phenomenon, the connection between interfacial water molecules and protein dynamics/flexibility has been extensively discussed in the experimental literature (see references 39−41 and references therein): protein dynamics and solvation shell dynamics have been characterized regionally. More precisely, it has been observed that flexible regions of proteins generally encompass fast-moving waters, while stable regions are embedded into slower hydration layer water molecules. This is exactly what we see here, and what is new in our results is that such regional dynamics modeling is shown to be strongly affected by many-body effects. Indeed, they strongly influence the dynamics of interfacial water molecules acting on their local "viscosity" and therefore local dynamics. As binding pockets and allosteric sites require being reasonably stable over time to be targeted by drugs, in some situations, non-PFF simulations may tend to predict solvation patterns associated with an excessive water traffic and to too fast-moving interfacial molecules. This could unfortunately lead to the destabilization of druggable hotspots that therefore would potentially remain unknown to molecular modelers.

To conclude, in order to propose a high-quality model of the dimerization interface of SARS-CoV2 M^{pro} that could be used for further drug design, it is important to understand well and model its complex H-bonds network that is embedded within a dynamic dipolar water solvent network. Water appears to be a key player in the overall structural dynamics of the dimerization interface, being one building block of the global allosteric effects between sites through many-body polarization interactions with the interface residues. As we stressed before,[9] M^{pro} is a difficult and complex molecular system that requires the simultaneous ability to (i) accurately describe all types of noncovalent interactions within the protein and solvent requiring therefore an accurate force field able to describe local many-body polarization effects and (ii) perform extensive sampling going beyond the microsecond time scale. Of course, we analyzed here only one example of allosteric interactions within M^{pro} and many other ones may remain to be discovered; we hope that these analyses and molecular dynamics trajectories (available via the BioExcel/MolSSI repository) will help drug hunters targeting the M^{pro} dimerization interface.

THEORETICAL METHODS

To study the dimerization interface we extensively analyzed the all-atom conformation space produced previously[9] using the AMOEBA polarizable force field (AMOEBA protein force field[11,12] and AMOEBA03 flexible water model[10]) as well as the one provided by the RIKEN[16] (using the AMBER ff14SB force field[42] and the TIP3P water model[43]) and DESRES[15] (using the DES-AMBER[44] and TIP4P-D water model[45]) groups. Following the same simulation protocol (reference PDB structure 6LU7[46]) proposed in our previous work,[9] we performed separate additional runs of adaptive simulations for a total of 12 μs with AMOEBA to simulate low pH values. In this case, additional histidine residue protonation occurs. Therefore, to produce additional data to the pH 7.4 and pH 6 simulations proposed in our previous data set,[9] we also successively protonated (2 × 6 μs runs) the two **His163** residues to simulate further pH lowering (see discussion and Table 2 in ref 18). Further 800 ns AMOEBA and AMBER99SB conventional molecular dynamics simulations (400 ns × 2) were produced at physiological pH and restarting from starting points from our previous data set, taking a snapshot every 10 ps to enable an in-depth analysis of the role of the water solvent. All additional all-atom simulations were performed using the newly developed GPUs module[14] within the Tinker−HP package,[13] which is part of the Tinker 8 platform.[47] This recently developed module is able to efficiently leverage mixed precision,[14] offering a strong acceleration of simulations using GPUs. Periodic boundary conditions using a cubic box of side length 100 Å were used. Langevin molecular dynamics simulations were performed using the BAOAB−RESPA1 integrator[48] using a 10 fs outer time step, a preconditioned conjugate gradient polarization solver (with a 10^{-5} convergence threshold), hydrogen−mass repartitioning (HMR), and random initial velocities. Periodic boundary conditions (PBC) were employed using the smooth particle mesh Ewald (SPME) method with a grid of dimension 128 Å × 128 Å × 128 Å. The Ewald-cutoff was taken to 7 Å, and the van der Waals cutoff was taken to be 9 Å. Post processing analysis was done using the MDTraj,[49] Scikit-Learn,[50] and Scipy packages.[51] Dynamical cross-correlation matrices (DCCMs) were generated based on the C_α atom of each residue by using the functionality provided in the MD-TASK package.[52]

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jpclett.1c01460.

All the residues implicated in H-bond interactions (Table 1); 2D plot representation of distances **His41**$_X$**−Cys145**$_X$ (X = chain **A** or chain **B**) versus distances **Arg4**$_B$**−Arg131**$_A$, **Arg4**$_B$**−Asp197**$_A$ and **Arg4**$_B$**−Thr199**$_A$ showing that allosteric connectivity is present (Figure 1); extracted values from dynamic cross-correlation maps revealing the cross-correlation between residues implicated in allosteric connectivity (Figure 2); number of water molecules detected in a 3.5 Å radius from **Arg4**$_X$, **Gly11**$_X$, **Glu14**$_X$, or **Glu290**$_X$ (X = chain **A** or chain **B**) (Figure 3); dipole distribution of structural water molecules interacting with **Arg4**$_X$, **Gly11**$_X$, **Glu14**$_X$, or **Glu290**$_X$ (X = chain **A** or chain **B**) (Figure 4); water lifetime distribution inside the allosteric

https://doi.org/10.1021/acs.jpclett.1c01460
J. Phys. Chem. Lett. 2021, 12, 6218−6226

dimerization site (Figure 5); representation of the water dipole distribution inside the allosteric dimerization site, for 5.29 and 8.7 Å between **Arg4** and **Glu290** (Figure 6); 2D plot representation of the volume of the dimerization site vs the number of water molecules inside the allosteric dimerization site and schematic representation of the SARS-CoV-2 M^{pro} dimer showing the dimerization site and the allosteric dimerization site residues (Figure 7) (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Jean-Philip Piquemal − *Sorbonne Université, 75005 Paris, France; Institut Universitaire de France, 75005 Paris, France; Department of Biomedical Engineering, University of Texas at Austin, Austin, Texas 78712, United States;* orcid.org/0000-0001-6615-9426; Email: jean-philip.piquemal@sorbonne-universite.fr

Authors

Dina El Ahdab − *Sorbonne Université, 75005 Paris, France; Université Saint-Joseph de Beyrouth, 1104 2020 Beirut, Lebanon*

Louis Lagardère − *Sorbonne Université, 75005 Paris, France; Sorbonne Université, 75005 Paris, France*

Théo Jaffrelot Inizan − *Sorbonne Université, 75005 Paris, France*

Fréderic Célerse − *Sorbonne Université, 75005 Paris, France; Sorbonne Université, 75005 Paris, France;* Present Address: F.C.: EPFL, LCMD, Lausanne, 1015, Switzerland.; orcid.org/0000-0001-8584-6547

Chengwen Liu − *Department of Biomedical Engineering, University of Texas at Austin, Austin, Texas 78712, United States;* orcid.org/0000-0002-3930-7793

Olivier Adjoua − *Sorbonne Université, 75005 Paris, France*

Luc-Henri Jolly − *Sorbonne Université, 75005 Paris, France*

Nohad Gresh − *Sorbonne Université, 75005 Paris, France;* orcid.org/0000-0001-7174-2907

Zeina Hobaika − *Université Saint-Joseph de Beyrouth, 1104 2020 Beirut, Lebanon*

Pengyu Ren − *Department of Biomedical Engineering, University of Texas at Austin, Austin, Texas 78712, United States;* orcid.org/0000-0002-5613-1910

Richard G. Maroun − *Université Saint-Joseph de Beyrouth, 1104 2020 Beirut, Lebanon*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jpclett.1c01460

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (Grant Agreement No 810367), project EMC2 (J.-P.P.). D.E.A. acknowledges funding from the Lebanese National Council for Scientific Research, CNRS-L. F.C. acknowledges funding from the French state funds managed by the CalSimLab LABEX and the ANR within the Investissements d'Avenir program (reference ANR11-IDEX-0004-02) and support from the Direction Génerale de l'Armement (DGA) Maîtrise NRBC of the French Ministry of Defense. Adaptive sampling computations have been performed at GENCI thanks to a COVID19 emergency allocation on the Jean Zay machine (IDRIS, Orsay, France) on Grant No. A0070707671 and on the Irene Joliot Curie machine thanks to a PRACE COVID-19 emergency grant (Project COVID-HP). Additional conventional AMOEBA and AMBER MD simulations have been performed on the Amazon cloud platform thanks to an AWS COVID-19 special grant. The authors thank the Swiss National Supercomputing Center (CSCS) for hosting our data through the FENIX infrastructure. J.-P.P. acknowledges a special COVID-19 funding from Sorbonne Université. P.R. is grateful for support by National Science Foundation (CHE-1856173) and National Institutes of Health (R01GM106137 and R01GM114237).

■ REFERENCES

(1) Zumla, A.; Chan, J. F.; Azhar, E. I.; Hui, D. S.; Yuen, K.-Y. Coronaviruses-Drug Discovery and Therapeutic Options. *Nat. Rev. Drug Discovery* **2016**, *15*, 327−347.

(2) Ding, L.; Zhang, X.-X.; Wei, P.; Fan, K.; Lai, L. The interaction between severe acute respiratory syndrome coronavirus 3C-like proteinase and a dimeric inhibitor by capillary electrophoresis. *Anal. Biochem.* **2005**, *343*, 159−165.

(3) Cui, W.; Yang, K.; Yang, H. Recent Progress in the Drug Development Targeting SARS-CoV-2 Main Protease as Treatment for COVID-19. *Front. Biosci.* **2020**, *7*, 398.

(4) Pillaiyar, T.; Manickam, M.; Namasivayam, V.; Hayashi, Y.; Jung, S.-H. An Overview of Severe Acute Respiratory Syndrome−Coronavirus (SARS-CoV) 3CL Protease Inhibitors: Peptidomimetics and Small Molecule Chemotherapy. *J. Med. Chem.* **2016**, *59*, 6595−6628.

(5) Kuo, C.; Liang, P. Characterization and Inhibition of the Main Protease of Severe Acute Respiratory Syndrome Coronavirus. *ChemBioEng Rev.* **2015**, *2*, 118−132.

(6) Boggetto, N.; Reboud-Ravaux, M. Dimerization Inhibitors of HIV-1 Protease. *Biol. Chem.* **2002**, *383*, 1321−1324.

(7) Zutshi, R.; Brickner, M.; Chmielewski, J. Inhibiting the assembly of protein-protein interfaces. *Curr. Opin. Chem. Biol.* **1998**, *2*, 62−66.

(8) Wei, P.; Fan, K.; Chen, H.; Ma, L.; Huang, C.; Tan, L.; Xi, D.; Li, C.; Liu, Y.; Cao, A.; Lai, L. The N-terminal octapeptide acts as a dimerization inhibitor of SARS coronavirus 3C-like proteinase. *Biochem. Biophys. Res. Commun.* **2006**, *339*, 865−872.

(9) Jaffrelot Inizan, T.; Célerse, F.; Adjoua, O.; El Ahdab, D.; Jolly, L.-H.; Liu, C.; Ren, P.; Montes, M.; Lagarde, N.; Lagardère, L.; Monmarché, P.; Piquemal, J.-P. High-resolution mining of the SARS-CoV-2 main protease conformational space: supercomputer-driven unsupervised adaptive sampling. *Chem. Sci.* **2021**, *12*, 4889−4907.

(10) Ren, P. Y.; Ponder, J. W. Polarizable Atomic Multipole Water Model for Molecular Mechanics Simulation. *J. Phys. Chem. B* **2003**, *107*, 5933−5947.

(11) Shi, Y.; Xia, Z.; Zhang, J.; Best, R.; Wu, C.; Ponder, J. W.; Ren, P. Polarizable Atomic Multipole-Based AMOEBA Force Field for Proteins. *J. Chem. Theory Comput.* **2013**, *9*, 4046−4063.

(12) Zhang, C.; Lu, C.; Jing, Z.; Wu, C.; Piquemal, J.-P.; Ponder, J. W.; Ren, P. AMOEBA Polarizable Atomic Multipole Force Field for Nucleic Acids. *J. Chem. Theory Comput.* **2018**, *14*, 2084−2108.

(13) Lagardère, L.; Jolly, L.-H.; Lipparini, F.; Aviat, F.; Stamm, B.; Jing, Z. F.; Harger, M.; Torabifard, H.; Cisneros, G. A.; Schnieders, M. J.; Gresh, N.; Maday, Y.; Ren, P. Y.; Ponder, J. W.; Piquemal, J.-P. Tinker-HP: a massively parallel molecular dynamics package for multiscale simulations of large complex systems with advanced point dipole polarizable force fields. *Chem. Sci.* **2018**, *9*, 956−972.

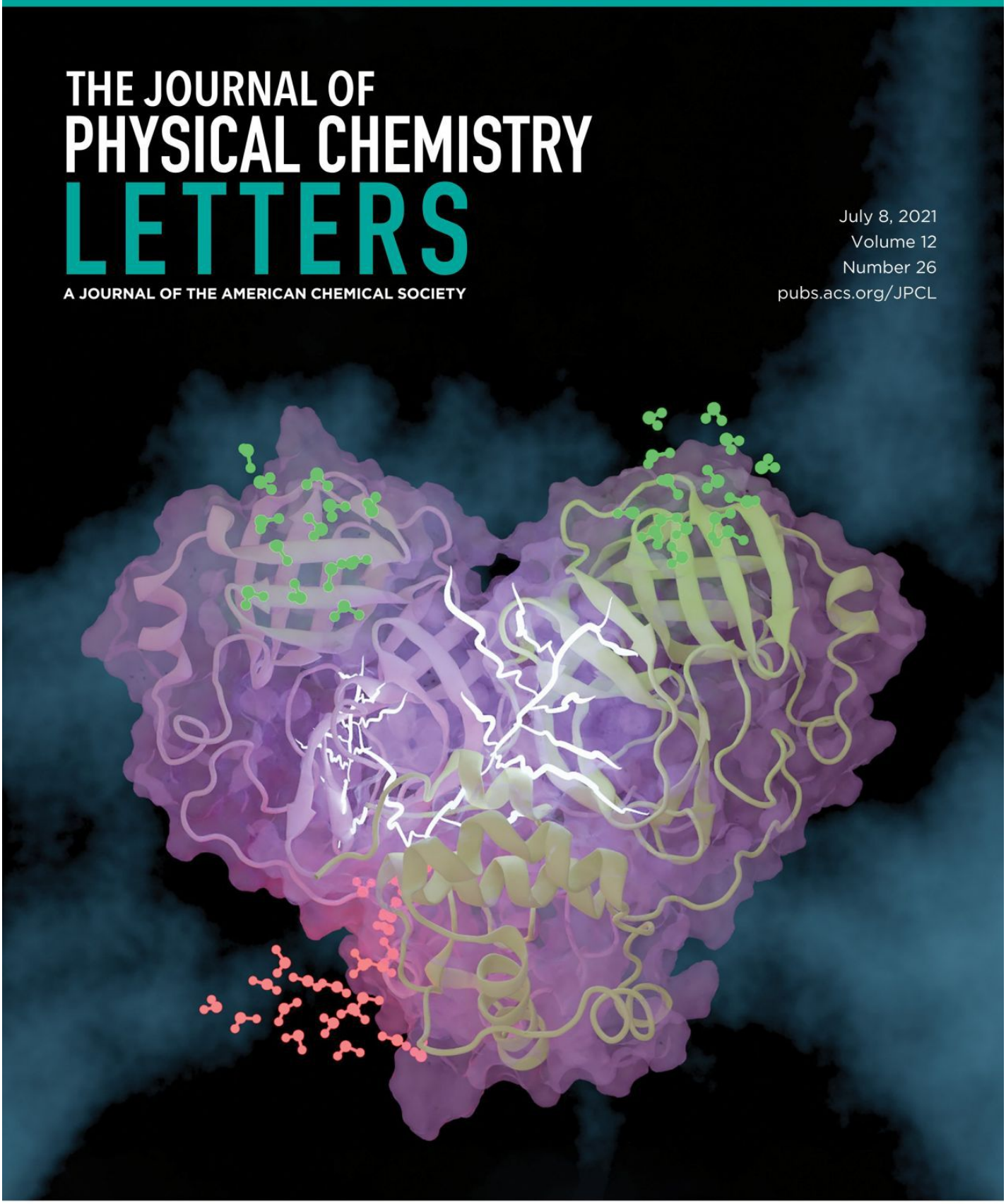
(14) Adjoua, O.; Lagardère, L.; Jolly, L.-H.; Durocher, A.; Very, T.; Dupays, I.; Wang, Z.; Inizan, T. J.; Célerse, F.; Ren, P.; Ponder, J. W.; Piquemal, J.-P. Tinker-HP: Accelerating Molecular Dynamics Simulations of Large Complex Systems with Advanced Point Dipole

Polarizable Force Fields Using GPUs and Multi-GPU Systems. *J. Chem. Theory Comput.* **2021**, *17*, 2034−2053.
(15) D. E. Shaw Research. *Molecular Dynamics Simulations Related to SARS-CoV-2*,. D. E. Shaw Research Technical Data, 2020, http://www.deshawresearch.com/resources_sarscov2.html.
(16) Komatsu, T. S.; Koyama, Y.; Okimoto, N.; Morimoto, G.; Ohno, Y.; Taiji, M. COVID-19 related trajectory data of 10 microseconds all atom molecular dynamics simulation of SARS-CoV-2 dimeric main protease, V2. *Mendeley Data* **2020**, *10*, 17632.
(17) Kneller, D. W.; Phillips, G.; O'Neill, H. M.; Jedrzejczak, R.; Stols, L.; Langan, P.; Joachimiak, A.; Coates, L.; Kovalevsky, A. Structural plasticity of SARS-CoV-2 3CL M pro active site cavity revealed by room temperature X-ray crystallography. *Nat. Commun.* **2020**, *11*, 3202.
(18) Tan, J.; Verschueren, K. H.; Anand, K.; Shen, J.; Yang, M.; Xu, Y.; Rao, Z.; Bigalke, J.; Heisen, B.; Mesters, J. R.; Chen, K.; Shen, X.; Jiang, H.; Hilgenfeld, R. pH-dependent Conformational Flexibility of the SARS-CoV Main Proteinase (Mpro) Dimer: Molecular Dynamics Simulations and Multiple X-ray Structure Analyses. *J. Mol. Biol.* **2005**, *354*, 25−40.
(19) Verma, N.; Henderson, J. A.; Shen, J. Proton-Coupled Conformational Activation of SARS Coronavirus Main Proteases and Opportunity for Designing Small-Molecule Broad-Spectrum Targeted Covalent Inhibitors. *J. Am. Chem. Soc.* **2020**, *142*, 21883−21890.
(20) Goyal, B.; Goyal, B. Targeting the Dimerization of the Main Protease of Coronaviruses: A Potential Broad-Spectrum Therapeutic Strategy. *ACS Comb. Sci.* **2020**, *22*, 297−305.
(21) Chou, C.-Y.; Chang, H.-C.; Hsu, W.-C.; Lin, T.-Z.; Lin, C.-H.; Chang, G.-G. Quaternary Structure of the Severe Acute Respiratory Syndrome (SARS) Coronavirus Main Protease. *Biochemistry* **2004**, *43*, 14958−14970.
(22) Liang, J.; Karagiannis, C.; Pitsillou, E.; Darmawan, K. K.; Ng, K.; Hung, A.; Karagiannis, T. C. Site mapping and small molecule blind docking reveal a possible target site on the SARS-CoV-2 main protease dimer interface. *Comput. Biol. Chem.* **2020**, *89*, 107372.
(23) Liao, S.-M.; Du, Q.-S.; Meng, J.-Z.; Pang, Z.-W.; Huang, R.-B. The multiple roles of histidine in protein interactions. *Chem. Cent. J.* **2013**, *7*, 44.
(24) Monod, J.; Wyman, J.; Changeux, J.-P. On the nature of allosteric transitions: a plausible model. *J. Mol. Biol.* **1965**, *12*, 88−118.
(25) Greener, J. G.; Sternberg, M. J. Structure-based prediction of protein allostery. *Curr. Opin. Struct. Biol.* **2018**, *50*, 1−8.
(26) Laskowski, R. A.; Gerick, F.; Thornton, J. M. The structural basis of allosteric regulation in proteins. *FEBS Lett.* **2009**, *583*, 1692.
(27) Suplatov, D. A.; Švedas, V. Study of Functional and Allosteric Sites in Protein Superfamilies. *Acta. Naturae* **2015**, *7*, 34−45.
(28) Günther, S.; et al. X-ray screening identifies active site and allosteric inhibitors of SARS-CoV-2 main protease. *Science* **2021**, *372*, 642−646.
(29) Strömich, L.; Wu, N.; Barahona, M.; Yaliraki, S. N. Allosteric Hotspots in the Main Protease of SARS-CoV-2. *BioRxiv* 2020, DOI: 10.1101/2020.11.06.369439.
(30) Carli, M.; Sormani, G.; Rodriguez, A.; Laio, A. Candidate Binding Sites for Allosteric Inhibition of the SARS-CoV-2 Main Protease from the Analysis of Large-Scale Molecular Dynamics Simulations. *J. Phys. Chem. Lett.* **2021**, *12*, 65−72.
(31) Amor, B. R. C.; Schaub, M. T.; Yaliraki, S. N.; Barahona, M. Prediction of allosteric sites and mediating interactions through bond-to-bond propensities. *Nat. Commun.* **2016**, *7*, 12477.
(32) La Sala, G.; Decherchi, S.; De Vivo, M.; Rocchia, W. Allosteric Communication Networks in Proteins Revealed through Pocket Crosstalk Analysis. *ACS Cent. Sci.* **2017**, *3*, 949−960.
(33) McCammon, J. A.; Harvey, S. C. *Dynamics of proteins and nucleic acids*; Cambridge University Press, 1988.
(34) Ichiye, T.; Karplus, M. Collective motions in proteins: a covariance analysis of atomic fluctuations in molecular dynamics and normal mode simulations. *Proteins: Struct., Funct., Genet.* **1991**, *11*, 205−217.
(35) Melcr, J.; Piquemal, J.-P. Accurate biomolecular simulations account for electronic polarization. *Front. Mol. Biosci.* **2019**, *6*, 143.
(36) de Courcy, B.; Piquemal, J.-P.; Garbay, C.; Gresh, N. Polarizable Water Molecules in Ligand-Macromolecule Recognition. Impact on the Relative Affinities of Competing Pyrrolopyrimidine Inhibitors for FAK Kinase. *J. Am. Chem. Soc.* **2010**, *132*, 3312−3320.
(37) de Courcy, B.; Pedersen, L. G.; Parisel, O.; Gresh, N.; Silvi, B.; Pilme, J.; Piquemal, J.-P. Understanding Selectivity of Hard and Soft Metal Cations within Biological Systems Using the Subvalence Concept. 1. Application to Blood Coagulation:Direct Cation−Protein Electronic Effects versus Indirect Interactions through Water Networks. *J. Chem. Theory Comput.* **2010**, *6*, 1048−1063.
(38) Gresh, N.; de Courcy, B.; Piquemal, J.-P.; Foret, J.; Courtiol-Legourd, S.; Salmon, L. Polarizable Water Networks in Ligand−Metalloprotein Recognition. Impact on the Relative Complexation Energies of Zn-Dependent Phosphomannose Isomerase with d-Mannose 6-Phosphate Surrogates. *J. Phys. Chem. B* **2011**, *115*, 8304−8316.
(39) Dahanayake, J. N.; Mitchell-Koch, K. R. How Does Solvation Layer Mobility Affect Protein Structural Dynamics? *Front. Biosci.* **2018**, *5*, 65.
(40) Bellissent-Funel, M.-C.; Hassanali, A.; Havenith, M.; Henchman, R.; Pohl, P.; Sterpone, F.; van der Spoel, D.; Xu, Y.; Garcia, A. E. Water Determines the Structure and Dynamics of Proteins. *Chem. Rev.* **2016**, *116*, 7673−7697.
(41) Combet, S.; Zanotti, J.-M. Further evidence that interfacial water is the main "driving force" of protein dynamics: a neutron scattering study on perdeuterated C-phycocyanin. *Phys. Chem. Chem. Phys.* **2012**, *14*, 4927−4934.
(42) Maier, J. A.; Martinez, C.; Kasavajhala, K.; Wickstrom, L.; Hauser, K. E.; Simmerling, C. ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *J. Chem. Theory Comput.* **2015**, *11*, 3696−3713.
(43) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **1983**, *79*, 926−935.
(44) Piana, S.; Robustelli, P.; Tan, D.; Chen, S.; Shaw, D. E. Development of a Force Field for the Simulation of Single-Chain Proteins and Protein−Protein Complexes. *J. Chem. Theory Comput.* **2020**, *16*, 2494−2507.
(45) Piana, S.; Donchev, A. G.; Robustelli, P.; Shaw, D. E. Water Dispersion Interactions Strongly Influence Simulated Structural Properties of Disordered Protein States. *J. Phys. Chem. B* **2015**, *119*, 5113−5123.
(46) Jin, Z.; Du, X.; Xu, Y.; Deng, Y.; Liu, M.; Zhao, Y.; Zhang, B.; Li, X.; Zhang, L.; Peng, C. Structure of M pro from SARS-CoV-2 and discovery of its inhibitors. *Nature* **2020**, *582*, 289.
(47) Rackers, J. A.; Wang, Z.; Lu, C.; Laury, M. L.; Lagardère, L.; Schnieders, M. J.; Piquemal, J.-P.; Ren, P.; Ponder, J. W. Tinker 8: Software Tools for Molecular Design. *J. Chem. Theory Comput.* **2018**, *14*, 5273−5289.
(48) Lagardère, L.; Aviat, F.; Piquemal, J.-P. Pushing the Limits of Multiple-Time-Step Strategies for Polarizable Point Dipole Molecular Dynamics. *J. Phys. Chem. Lett.* **2019**, *10*, 2593−2599.
(49) McGibbon, R. T.; Beauchamp, K. A.; Harrigan, M. P.; Klein, C.; Swails, J. M.; Hernández, C. X.; Schwantes, C. R.; Wang, L.-P.; Lane, T. J.; Pande, V. S. MDTraj: A Modern Open Library for the Analysis of Molecular Dynamics Trajectories. *Biophys. J.* **2015**, *109*, 1528−1532.
(50) Pedregosa, F.; et al. Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825−2830.
(51) Virtanen, P.; et al. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nat. Methods* **2020**, *17*, 261−272.
(52) Brown, D. K.; Penkler, D. L.; Sheik Amamuddy, O.; Ross, C.; Atilgan, A. R.; Atilgan, C.; Tastan Bishop, O. MD−TASK: a software suite for analyzing molecular dynamics trajectories. *Bioinformatics* **2017**, *33*, 2768−2771.

2.6.3 Publication 2 - "Back Cover"

THE JOURNAL OF
PHYSICAL CHEMISTRY
LETTERS
A JOURNAL OF THE AMERICAN CHEMICAL SOCIETY

July 8, 2021
Volume 12
Number 26
pubs.acs.org/JPCL



ACS Publications
Most Trusted. Most Cited. Most Read.

www.acs.org

Chapitre 3

Étude par dynamique moléculaire des structures des quadruplexes de guanine au niveau du promoteur de l'oncogène c-kit et du LTR-III du VIH-1

3.1 Détails computationnels et protocole de simulation

3.1.1 Choix des structures de départ et construction des systèmes

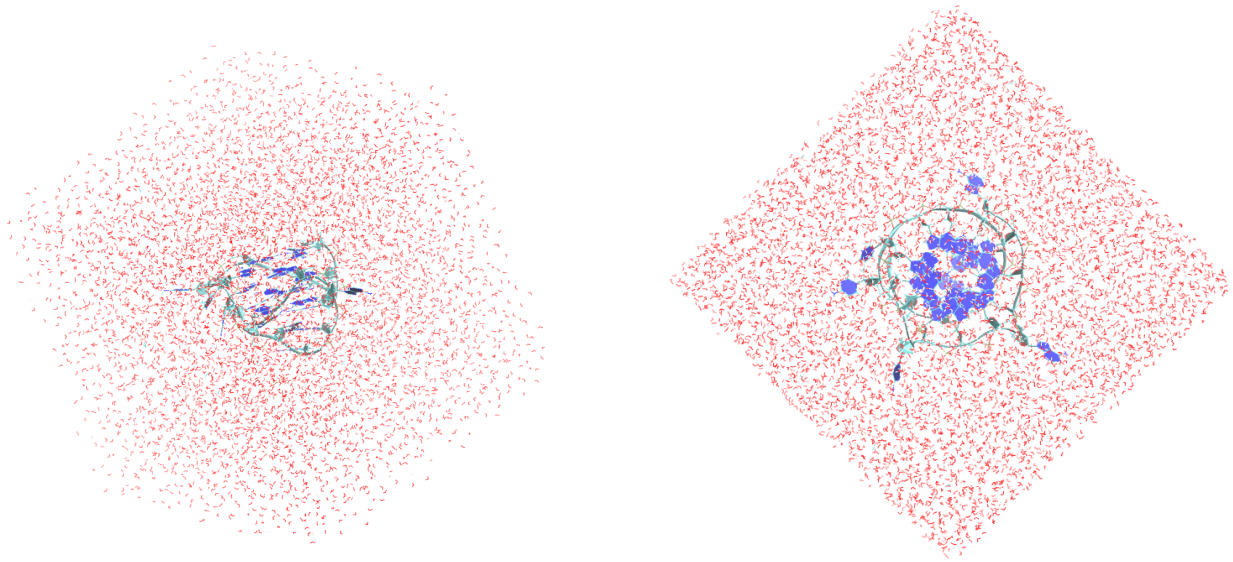
Des structures de haute résolution du GQ du c-kit 1 en présence de K^+ sont disponibles sur le site du PDB à partir de la spectroscopie RMN-3D (code PDB *2O3M*) et de la cristallographie de rayons X (codes PDB : *3QXR, 4WO2 et 4WO3*)[167, 225, 227]. La liste de ces structures est présentée dans la Tableau 3.1.

Code PDB	Nom de la structure	Méthode d'obtention de la structure	Résolution
3qxr	Structure cristalline de l'ADN quadruplex du promoteur du proto-oncogène CKIT-1 bromé	Cristallographie	1.62 Å
4wo2	Structure cristalline de l'ADN quadruplex du promoteur du proto-oncogène ckit natif humain	Cristallographie	1.82 Å
4wo3	La deuxième structure cristalline du quadruplex de l'ADN c-kit 1	Cristallographie	2.73 Å
2o3m	La deuxième structure cristalline du quadruplex de l'ADN du c-kit 1	RMN-3D	-

Table 3.1: Liste des structures PDB à haute résolution du GQ de c-kit 1

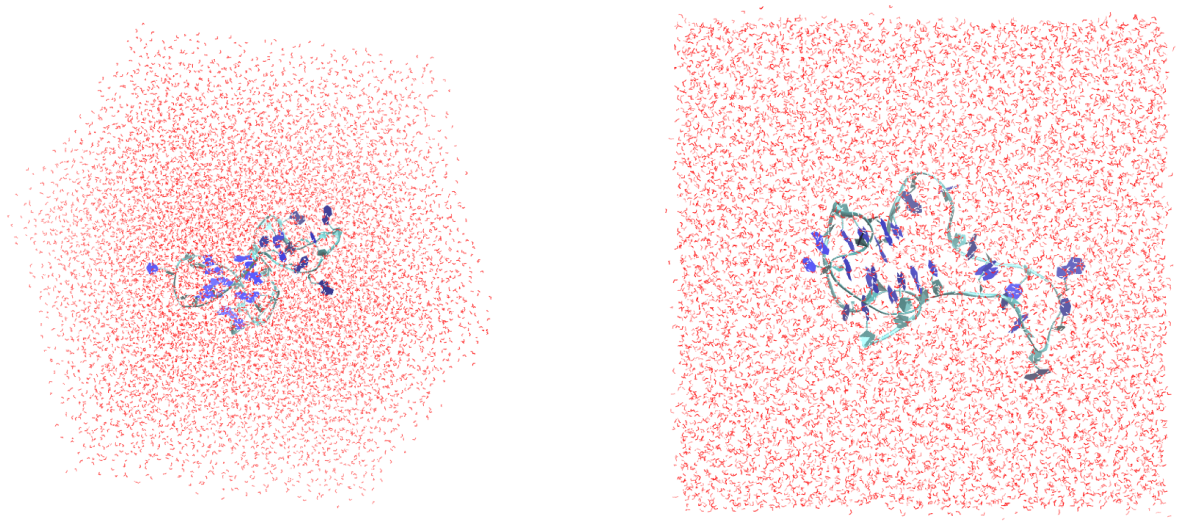
Bien que la structure *3QXR* ait la meilleure résolution cristallographique (1.62 Å), un résidu bromouridine (BRU) a été incorporé dans la séquence GQ du c-kit 1 afin de faciliter le phasage cristallographique. Afin d'éviter les effets de toute mutation de la structure cristallographique au niveau de la séquence native du promoteur du proto-oncogène natif humain c-kit 1 et puisque *4WO2* a la meilleure résolution pour la structure 22-mer (1.82 Å) non mutée nous avons sélectionné comme structure de départ pour notre étude la structure cristalline *4WO2* ayant une résolution de 1.82 Å.

Les structures de départ *4WO2* [227] et *6h1k* [37] ont été retirées de la PDB et solvatées dans une boîte cubique d'eau de dimensions 60 Å et de 70 Å respectivement de chaque coté, voire Figures 3.1 et 3.2 respectivement. Les deux systèmes ont été neutralisés avec des ions K^+ et 150 mM de sel [KCl] (K^+, Cl^-) y ont été ajoutés. La présence des ions dans le système est importante afin de mimer les conditions physiologiques. C'est aussi crucial pour le maintien de la structure de l'ADN lors des simulations de dynamique moléculaire. Le logiciel VMD (Visual Molecular Dynamic) [82] a été utilisé afin de concevoir la boîte de l'eau et d'ioniser le système et ultérieurement afin de pouvoir visualiser les structures et les dynamiques moléculaires. Tinker [176] et Tinker-HP [114, 7] ont été utilisés lors de la préparation du système en utilisant les paramètres de champ de force AMOEBA-BIO18 [235] et afin de simuler les trajectoires de dynamique moléculaire.



a) b)

Figure 3.1: Représentation de la boîte d'eau cubique de dimensions 60 Å contenant le GQ au niveau du promoteur du proto-oncogène c-kit 1 et les ions K^+ et Cl^-. a) Vue de coté et b) vue de haut.



a) b)

Figure 3.2: Représentation de la boîte d'eau cubique de dimensions 70 Å contenant le GQ de la région LTR-III du VIH-1 et les ions K^+ et Cl^-. a) Vue de coté et b) vue de haut.

3.1.2 Dynamique moléculaire conventionnelle

Les paramètres du champ de force multipolaire atomique polarisable AMOEBA pour les acides nucléiques ont déjà été affinés et validés pour des simulations de dynamique moléculaire de molécules d'ADN et d'ARN en solution aqueuse et en réseau cristallin replié en conformations classiques et non classiques [235]. Ce champ de force polarisable est implementé dans le logiciel Tinker-HP [114]. Toutes

les simulations d'atomes présentées ici ont été réalisées à l'aide du module GPU (processeur graphique) récemment développé [7] dans le logiciel Tinker-HP [114]. Ce module est capable d'exploiter efficacement la précision mixte [7] offrant une forte accélération des simulations utilisant des GPU. Des conditions aux limites périodiques utilisant une boîte cubique d'une longueur de côté de 60 Åont été utilisées. Le système a été équilibré pendant 1 ns à chacune des températures suivantes successivement : 50, 100, 150, 200, 250 et 300 Kelvins. Le seuil de cutoff d'Ewald a été porté à 7 Å et celui de van der Waals à 9 Å. De longues simulations ont été ainsi réalisées à une température de 300 K. Un temps de simulation long permettant un échantillonage exhaustif est nécessaire afin de permettre une meilleure découverte des états de transition et une analyse détaillée du comportement moléculaire du GQ de c-kit 1.

3.1.3 Algorithme adaptatif: longues dynamiques moléculaires polarisables

Une nouvelle technique d'échantillonnage adaptatif non supervisé basée sur des modèles statistiques et une analyse en composantes principales (PCA) [94] a été récemment développée afin de surmonter la limitation de temps imposée par les supercalculateurs généralistes comme déjà discuté dans le chapitre précédent. Cette méthode d'échantillonnage a permis à la plateforme Tinker-HP [114] d'effectuer des simulations de dynamique moléculaire de plusieurs microsecondes en un temps record grâce aux GPUs, rendant ainsi possible un échantillonage massif du comportement de GQ en présence et en l'absence des ions K^+ au sein du canal ionique formé par les guanine qui s'arrangent en 3 tétrades. Ceci est rendu possible du fait que plusieurs ensembles de trajectoires séparées peuvent être simulées en parallèle sur un grand nombre de GPUs disponibles sur les supercalculateurs généralistes [94] [7]. Afin d'effectuer une simulation d'échantillonnage adaptatif non supervisé à grande échelle, des structures de départ ont été sélectionnées à partir de notre simulation cMD initiale de départ (800 ns). En pratique, à partir de la simulation cMD de 800 ns et en utilisant la PCA, nous avons extrait 50 structures pertinentes qui représentent nos points de départ pour l'échantillonnage adaptatif.

3.1.4 Analyse de composantes indépendantes du temps et de la structure (tICA)

Actuellement, la tICA est largement utilisée et est préférée à la PCA dans les études des systèmes lents puisqu'elle utilise en plus que la PCA des informations temporelles sur la trajectoire d'entrée, ainsi, tICA peut être très sensible à de très petits changements. En effet, il a été démontré également [173] que tICA résout une approche variationnelle [158, 159] par laquelle elle trouve une approximation optimale des valeurs et fonctions propres de l'opérateur de Markov sous-jacent aux données de dynamique moléculaire. Par conséquent, tICA est capable de détecter et est parfaitement adaptée à la clusterisation et à la préparation des données pour la construction du modèle de Markov.

3.2 Résultats et Discussion

3.2.1 Évaluation de la stabilité conformationnelle et étude des états de transition

Évolution de la stabilité des différents nucléotides du GQ c-kit 1

L'écart quadratique moyen (RMSD) a été calculé pour l'ensemble des 7.5 μs d'échantillonnage adaptatif afin de mesurer la distance moyenne entre les atomes du squelette des structures superposées. Comme le montre la Figure 3.3, a), le RMSD moyen des structures GQ entières est de 3.5 Å, ce qui signifie que le système est relativement stable. Cependant, afin de suivre la dynamique de fluctuation des différentes composantes du GQ, le RMSD pour le noyau GQ, le LPloop et les boucles simples ont été calculés séparément. Le noyau du GQ et LPLoop semblent être très stables et bien structurés car leur RMSD est inférieur à 1 Å. En ce qui concerne les boucles nucléotidiques simples, le RMSD s'avère plus fluctuant, suggérant que la dynamique des boucles est responsable de la dynamique de la structure des trous. Le RSMF a été également calculé pour rechercher les résidus les plus dynamiques dans le GQ de c-kit 1. Les nucléotides formant les boucles simples, à nommer A1, A5, C9, C11 et T12, sont les résidus ayant le RMSF qui fluctuent le plus au sein de la structure du GQ. Le RMSF dépasse les 8 Å pour les résidus A1 et C11 et les 7 Å pour le résidu T12. A5 et C9 fluctuent également ; ils dépassent respectivement 4 et 6 Å. Prenant compte de ces fluctuations par rapport à leur structure moyenne, comme le montre la Figure 3.3, b), nous pensons que c'est la flexibilité de ces boucles simples qui affecte la dynamique de l'ensemble de la structure du GQ.

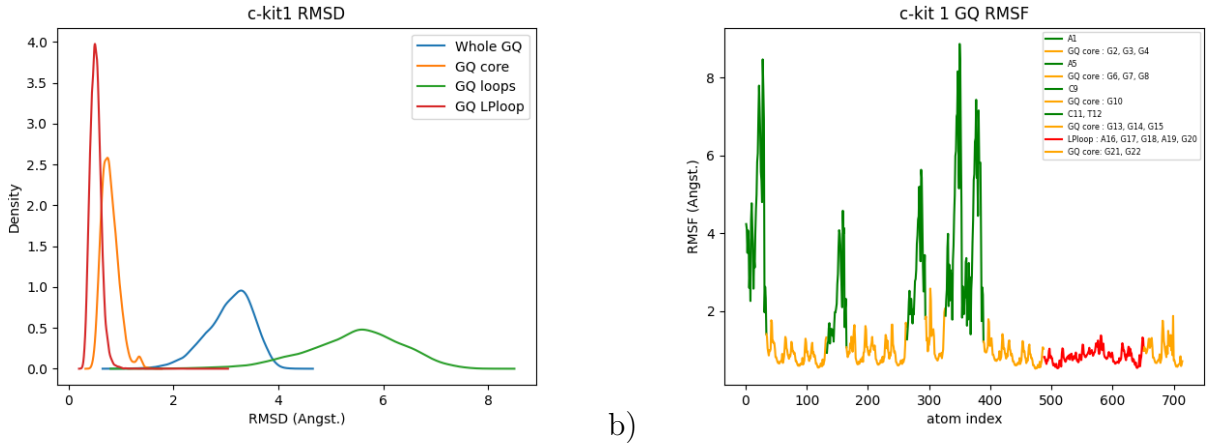


Figure 3.3: RMSD et RMSF du c-kit 1. Pour RMSF, les index atomiques des boucles nucléotidiques simples A1, A5, C9, C11 et T12 sont respectivement : 0 à 33, 133 à 164, 264 à 293, 327 à 356 et 357 à 388.

Angles	Atomes
α	O3'(i-1)-P-O5'-C5'
β	P-O5'-C5'-C4'
γ	O5'-C5'-C4'-C3'
δ	C5'-C4'-C3'-O3'
ϵ	C4'-C3'-O3'-P(i+1)
ζ	C3'-O3'-P(i+1)-O5'(i+1)
χ pyrimidines (Y)	O4'-C1'-N1-C2
χ purines (R)	O4'-C1'-N9-C4

Table 3.2: Liste des sept angles de torsion qui définissent la conformation d'un nucléotide et les atomes qui forment l'angle.

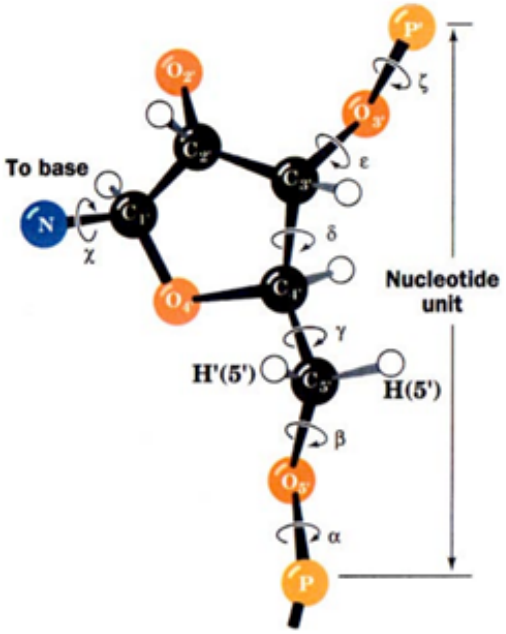
Études des angles de torsion et évolution de l'état conformationnel des boucles

Pour remplir sa fonction biologique, une protéine ou un acide nucléique doit adopter une multiplicité de conformations. L'équilibrage et le contrôle des différents états conformationnels sont au cœur des processus biologiques, notamment le repliement, l'allostérie et la signalisation des protéines, ainsi que l'assemblage avec d'autres molécules dans le cadre de complexe moléculaire. La conformation des nucléotides au sein de l'ADN et de l'ARN est définie particulièrement par les 7 angles de torsions [160] qui sont précisés comme présenté dans le Tableau 3.2 et la Figure 3.4. Plus précisément, l'angle de torsion χ caractérise l'orientation relative base/sucre. Dans le cas du duplex d'ADN des formes A et B), χ se situe dans les plages de +90° à +180° ; –90° à –180°, correspondant à une conformation dite *anti* (Figure 3.5, en haut). Parfois, χ adopte des valeurs comprises entre –90° et +90°, se référant à la conformation *syn* (Figure 3.5, en bas). C'est important de noter que dans l'ADN-Z gaucher ayant une séquence répétitive CG, la purine G adopte plutôt une conformation *syn* tandis que la pyrimidine C adopte la conformation *anti*. Vraisemblablement, la conformation *anti/syn* liée à χ est un concept géométrique simple. Néanmoins, la liaison N-glycosidique et l'angle de torsion χ correspondant illustrent que la base et le sucre sont deux entités distinctes, c'est-à-dire qu'il existe un degré de liberté interne entre eux [130]. Le retournement de base (commutation de conformation *anti/syn*) est l'un des facteurs associés aux deux orientations relatives possibles des bases azotés.

Un deuxième angle de torsion qui caractérise également la conformation des nucléotides de l'ADN est l'angle γ qui est formé par les atomes O5'-C5'-C4'-C3'. Cet angle précise l'orientation relative du sucre par rapport au squelette. En effet, le squelette joue un rôle important dans la restriction de l'espace conformationnel disponible pour l'empilement des bases [161] et dans le couplage des propriétés conformationnelles entre les dinucléotides voisins dans une séquence [25].

Vu les propriétés qui peuvent être déduites des angles χ et γ, nous avons suivi les évolutions de l'angle de torsion glycosidique χ, avec l'apparition possible de retournement de base *anti/syn* et l'angle γ spécifiant l'orientation du sucre par rapport au squelette de l'ADN.

La Figure 3.7 représente les graphiques 2D des angles χ et γ des nucléotides A1, A5, C9, C11 et



a)

Figure 3.4: Représentation des 7 angles de torsion qui définissent la conformation des nucléotides [160].

T12.

Une première vue de la Figure 3.7 nous apprend le comportement différent de chacun de ces nucléotides vu le profil différent de la répartition des points qui correspondent aux angles de chaque structure au sein de la trajectoire.

Plus précisément, l'angle γ semble être très libre pour A1, C9, C11 et T12, pouvant adopter une infinité de degrés de liberté, alors qu'il est plus limité pour A5 ayant peu ou pas de points entre 25° et 75°. C9 et C11 affichent de grandes amplitudes de variations χ et γ par rapport à A1, A5 et T12. A1 et T12 ont des variations d'angle χ limitées (-50° – +50° et +25° – +50° respectivement) mais ont de grandes fluctuations de l'angle γ. De telles fluctuations permettent la formation d'une paire de bases Hoogsteeen transitoire au-dessus du premier tétramère du GQ impliquant G2, G6, G10 et G13. Cette paire est stabilisée par deux liaisons H, entre HN6(A1) et O4(T12), et entre N7(A1) et HN3(T12).

L'implication de N7 de A1 au lieu de N1 comme dans une paire de bases W-C standard était imprévue et semble n'avoir aucun précédent dans d'autres simulations des GQ. Ceci permet de localiser N7 de A1 dans le même sillon que les trois bases ; G2, G3 et G4 sur le même brin qui forment les trois tétramères. En revanche, c'est la formation d'une paire de bases W-C qui a été précédemment rapportée dans les simulations AMBER MD [90].

Bien que les structures apparues le long des 7.5 μs de MD puissent adopter un large éventail de degrés de liberté correspondant aux angles χ et γ, certains angles sont plus favorisés que d'autres.

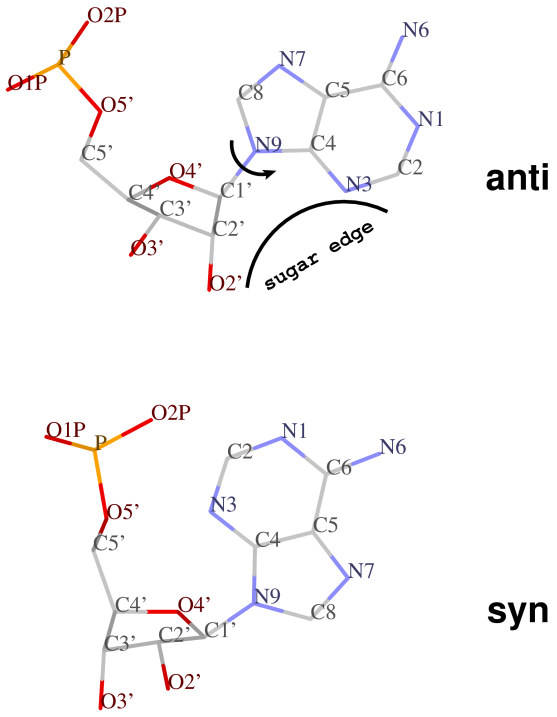


Figure 3.5: Représentation des conformations *syn* et *anti* que peuvent adoptées les bases azotées.

Nucléotides	γ	χ
A1	-25°	25°
A5	-110°	-160°
C9	75°	-40°
C11	-30°	-10°
T12	-50°	35°

Table 3.3: Les angles γ et χ adoptés par les nucléotides A1, A5, C9, C11 et T12 en se basant sur la densité maximale des points sur les graphes 2D de la Figure 3.7

En effet, les bassins présentés en jaune sur les graphiques de la Figure 3.7 représentent la densité maximale de structures pouvant adopter des angles χ et γ particuliers de chaque nucléotide. Ces angles ayant une forte probabilité d'existence sont présentés dans le Tableau 3.3.

Les angles χ et γ correspondant aux structures cristallographiques *4WO2* [167] et *4WO3* [225] et celle résolue par RMN *2O3M* [227] ont été également calculés, voir Figure 3.7. Clairement, nos simulations ont permis de détecter des conformations qui n'apparaissaient ni en cristallographie ni en RMN. Comme χ et γ des structures extraites de la PDB [167, 225, 227] ne correspondent pas forcément aux bassins présentant une forte densité de structures échantillonnées par AMOEBA le long des 7.5 μs, ceci propose que les structures obtenus par cristallographie aux rayons X ou par RMN ne correspondent pas obligatoirement aux structures ayant les conformations avec la plus forte chance d'existence, d'où l'importance de la dynamique moléculaire afin d'évaluer la stabilité et les conformations les plus stables d'un système donné.

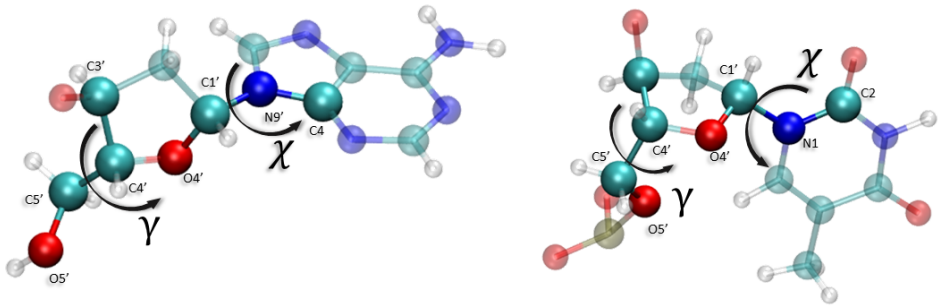


Figure 3.6: Représentation des angles γ et χ dans les nucléotides de type purine (ex. Adénine) et de type pyrimidine (ex. Thymine) respectivement à gauche et à droite.

Évaluation des distances entre nucléotides

Afin de comprendre les altérations de la conformation du GQ c-kit 1 en fonction de leur temps d'apparition, nous avons considéré d'abord la trajectoire MD classique initiale (800 ns), puis les trajectoires résultant de l'échantillonnage adaptatif (7,5 μs).

Nos simulations ont donné lieu à la formation de deux paires de bases : A1–T12 et une concurrente de plus longue durée entre A1–C11.

Les évolutions temporelles de ces deux paires de bases ont été suivies en calculant la distance HN6–O4 de A1–T12 et HN6–N3 de A1–C11 puisque ces 3 nucléotides ont été observés capables de s'empiler au-dessus de la première tétrade de guanine après retournement. Comme le montre la Figure 3.8, a), b) et c), le premier raccourcissement de la première distance à environ 2 Å est activé par le mouvement de retournement de T12 avec des fluctuations γ de grande amplitude, et son empilement au-dessus de la tétrade à face ouverte, en particulier au-dessus de G13 et G10. Ce phénomène se produit d'abord après 80 ns du début de la simulation mais ne dure que 1 ns. Il se reproduit ensuite entre 180 et 220 ns. À 260 ns, la formation de la paire de bases concurrente A1–C11 a lieu et cette paire de bases dure jusqu'à la fin du cMD à 800 ns.

Une telle interaction non classique n'avait pas été signalée auparavant dans les études PFF [190] ni dans les simulations d'études n–PFF [190, 90]. Cependant, AMBER MD [90] a enregistré un mouvement de C11 et une interaction de courte durée avec T12 ou le squelette phosphate de G10.

A5 et C9 restent très dynamique et n'interagissent avec aucun composant particulier des GQ en raison de leur position en tant que boucle latérale à nucléotide unique.

Par ailleurs, la Figure 3.9, d), montre la probabilité des distances A1–T12 et A1–C11 le long des 7.5 μs MD issues de l'échantillonnage adaptatif. Bien que la plus grande densité de structures échantillonnées partagent une distance qui varie entre 10 Å et 20 Å pour A1–C11 et A1–T12, il est intéressant de mentionner que pour une distance de 2-3 Å on peut détecter plus de structures qui partagent l'interaction A1–C11 que de structures partageant l'interaction A1–T12 ce qui vient confirmer une préférence en faveur de l'interaction A1–C11 plutôt que de A1–T12.

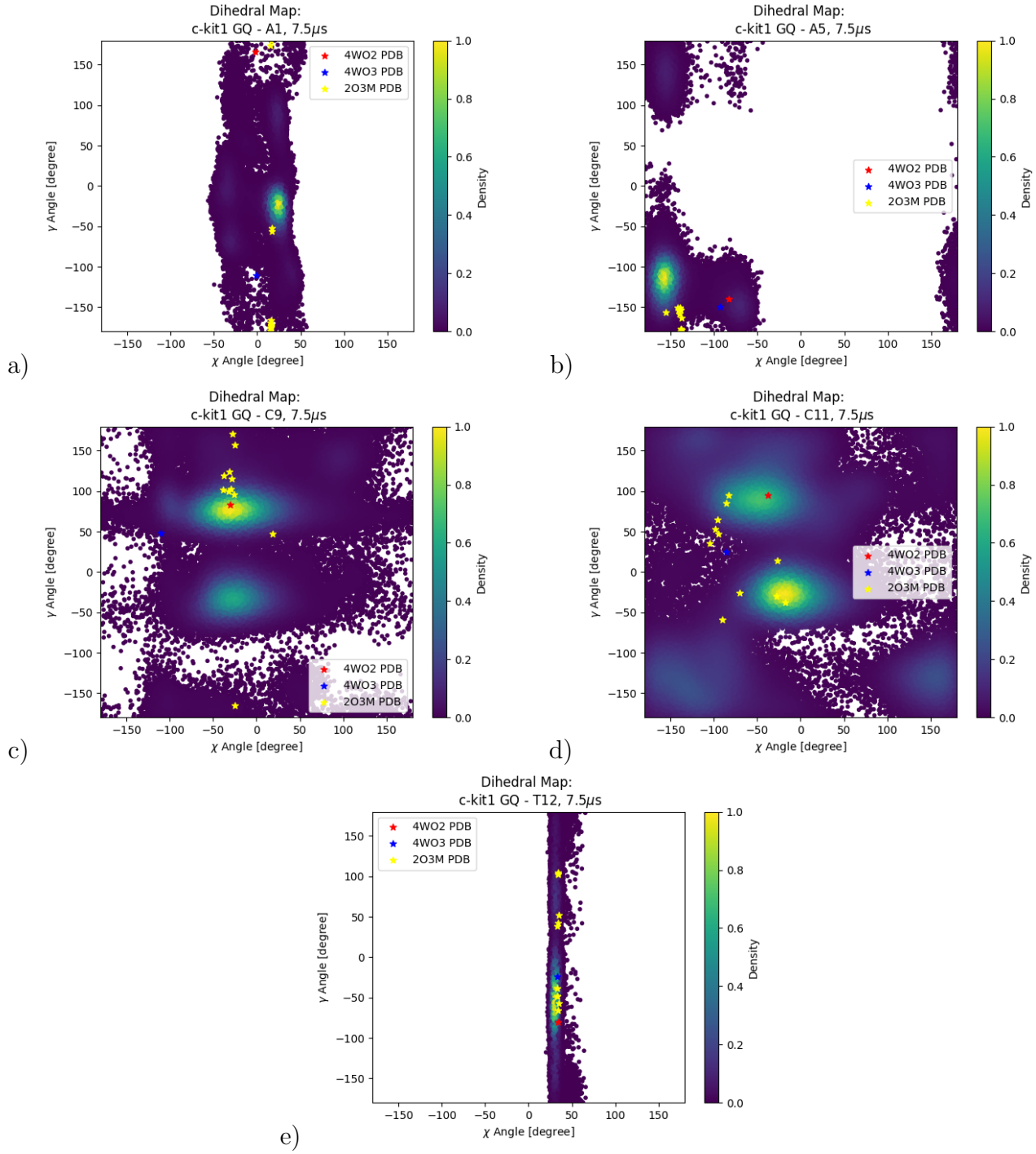


Figure 3.7: Représentation graphique 2D des angles γ et χ correspondant aux nucléotides A1, A5, C9, C11 et T12, respectivement en a), b), c), d) et e). Les angles correspondant aux structures cristallines de GQ de c-kit 1 déposées sur le site du PDB sont présentés en étoiles.

Au niveau de la boucle LP démontrée stable par RMSD, A16 peut interagir avec G20 et G17 peut interagir avec A19. Afin d'évaluer la stabilité de ces interactions au sein de la boucle LP, il convient de calculer les distances suivantes A16–G20 et G17–A19. Une interaction bien définie est observée entre les résidus A16(N1)–G20(H1) et G17(H22)–A19(N1). Comme le montre la Figure 3.9, a) et b),

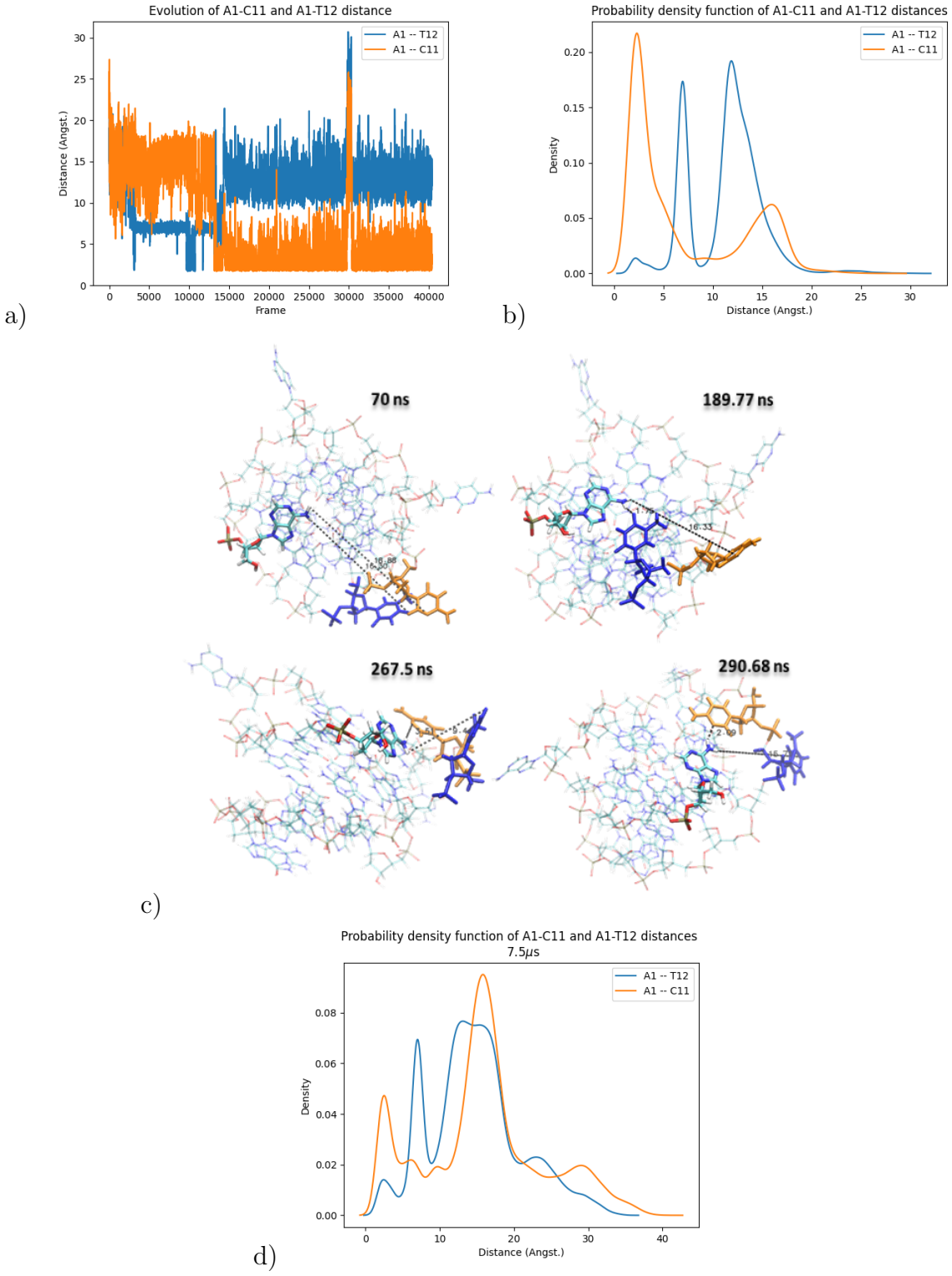


Figure 3.8: Évolution de la distance entre A1(H) et T12(O) présentée en bleu et entre A1(H) et C11(O) présentée en orange. Chaque 5000 "Frames" représente 100 ns. c) Représentation de 4 structures à quatre moments différents de la trajectoire initiale du cMD montrant l'évolution des distances A1–T12 et A1–C11.

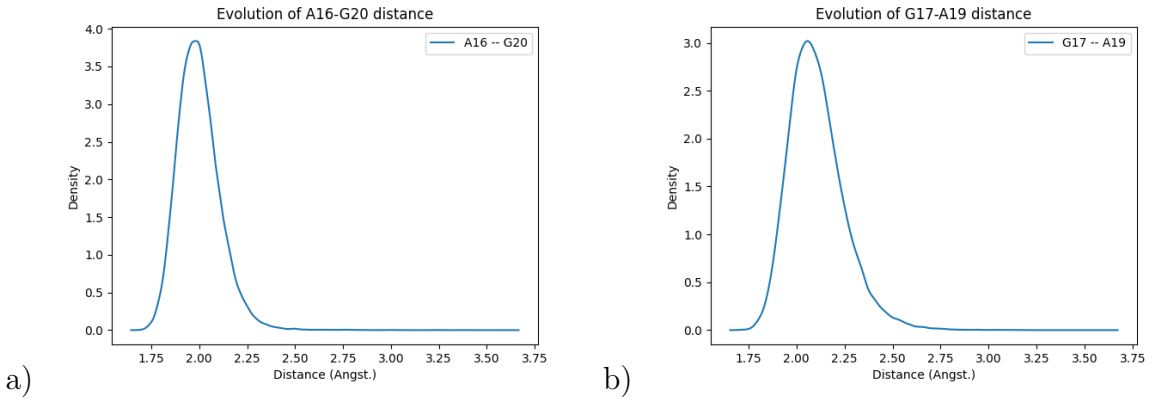


Figure 3.9: Distances entre A16–G20, G17–A19 calculées le long de la trajectoire cMD initiale (800ns).

la distance entre ces nucléotides varie autour de 2 Å pour A16–G20 et de 2.15 Å pour G17–A19, donc la boucle LP est bien structurée. Ces distances ont été enregistrées pour varier entre 2 Å et 3 Å en champ de force non polarisable [91]. L'appariement des bases Watson-Crick entre A16 et G20 est similaire à celui des structures cristallines et RMN [167, 225, 227].

Ces résultats suggèrent qu'AMOEBA est capable de détecter et de maintenir des interactions non classiques entre bases nucléiques. Leur apparition commence après 10 ns. En revanche, un mouvement de retournement A1 n'a été détecté qu'après 2 μs dans une simulation MD dans les simulations AMBER [90].

Études des empilements π-π

Comme nous avons observé des mouvements de torsion et de retournement des nucléotides A1, C11 et T12 qui s'empilent au-dessus de la première tétrade de guanine, il convient de quantifier cet empilement afin d'évaluer la stabilité structurelle des résidus en question. En effet, l'empilement π-π fait référence à des interactions attractives et non-covalentes entre les cycles aromatiques, car ils contiennent des liaisons de type π. L'empilement π-π est donc, par définition, une interaction faible intermoléculaire ou intramoléculaire non spécifique. Une entité aromatique peut pratiquement interagir avec n'importe quelle autre entité aromatique. La présence d'un moment quadripôle électrique seul est requise pour former un tel empilement. Effectivement, l'interaction se formant entre les deux moments quadripolaires électriques peut piloter les deux molécules à occuper différentes configurations possibles.

L'indice d'empilement π-π a été calculé pour les structures de la trajectoire cMD initiale (800 ns) et pour celles des simulations issues de l'échantillonnage adaptatif (7.5μs) (figure 3.10) afin d'obtenir une mesure quantitatif de cet empilement. En théorie, l'index d'un empilement π-π varie entre 0 (absence d'empilement) et 0.6 (empilement parfait). En effet, A1 partage des interactions π-π avec G2 et avec G13 simultanément, cependant, il est mieux empilé au-dessus de G13, puisqu'il partage un indice de 0.2 avec G13 contre un indice de 0.15 dans le cas de G2. Partant de la même logique, C11 et T12

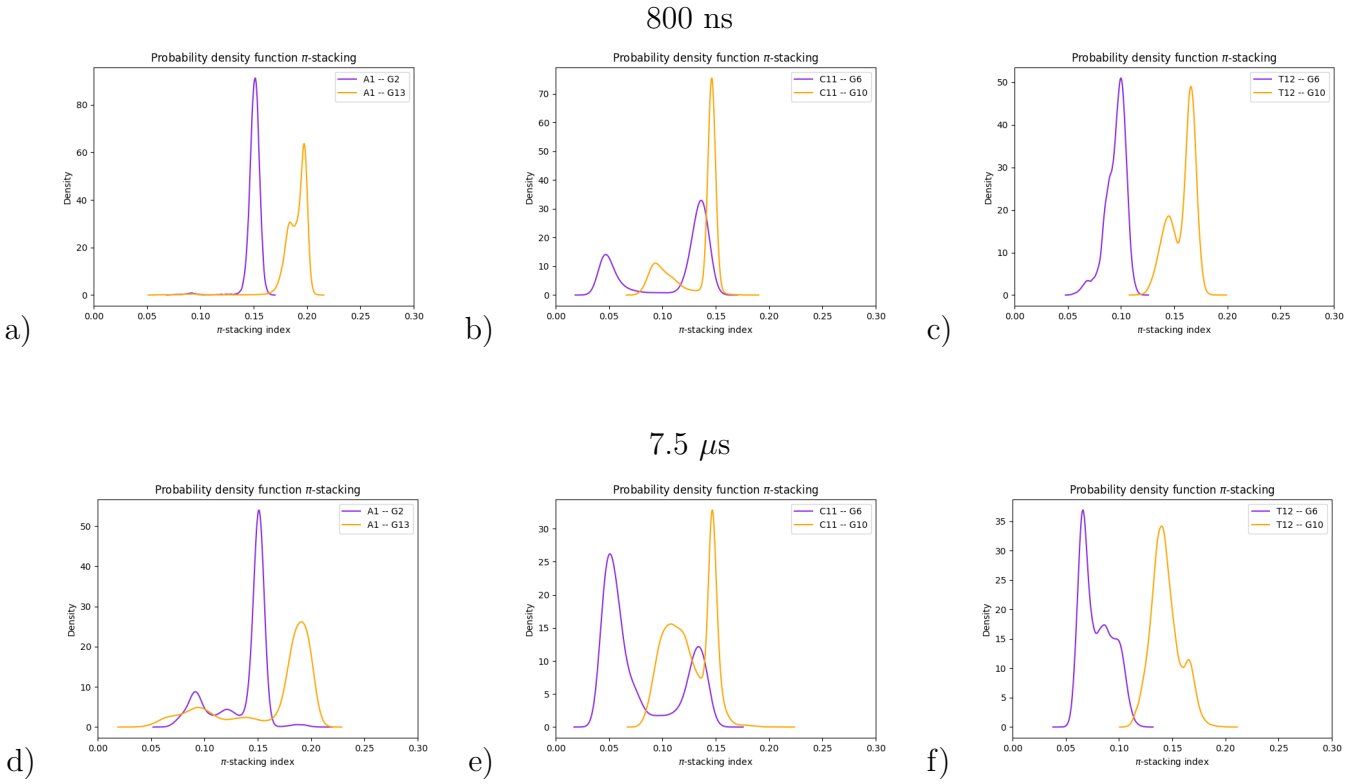


Figure 3.10: Indice d'empilement π-π pour A1–G2, A1–G12, C11–G6, C11–G10, T12–G6 et T12–G10 calculé pour les trajectoires cMD initiale et 7.5 μs.

partagent des interactions de type π-π avec C11 et avec T12, cependant les deux nucléotides de la boucle dinucléotidique semble mieux s'empiler au-dessus de G10.

3.2.2 Clusterisation et analyse conformationnelle des clusters

Clusters obtenus par tICA

Nous avons profité de l'algorithme d'échantillonnage adaptatif non-supervisé développé précédemment dans le cadre de l'accélération de la recherche sur la SARS-CoV-2 afin de l'appliquer sur le GQ de c-kit 1. Ceci a permis de produire 7.5 μs de simulation en AMOEBA PFF. Ensuite DBSCAN a été appliqué sur l'ensemble des trajectoires produites en utilisant tICA pour réduire la dimensionnalité des données. Cela a permis d'identifier les micro-états caractéristiques des trajectoires de MD issues de la stratégie d'échantillonage adaptative non supervisée, qui ont ensuite été regroupés en des macro-états représentatifs de chaque cluster.

De nombreux algorithmes exigent que les données d'entrée soient des vecteurs dans un espace vectoriel (euclidien). Cela inclut la clusterisation selon tICA. Puisqu'il n'y a généralement pas de cadre de référence de rotation ou de translation spécial dans une simulation MD, il est souvent souhaitable de supprimer le mouvement de rotation et de translation via une fonctionnalité qui est insensible aux rotations et aux translations.

Une alternative à la caractérisation "Featurization" : Bien que de nombreux algorithmes nécessitent des données vectorisables, d'autres algorithmes ne nécessitent qu'une métrique de distances par paires, par ex. RMSD entre deux conformations protéiques.

Dans notre cas, le modèle le moins concluant est celui qui est utilisé par le groupe de Pande, spécialiste des méthodes "MSM" (Markov State Models) [86], basées sur l'identification de distances caractéristiques du modèle comme données d'entrée ("featurization"). En général, cette approche consiste à calculer toutes les distances entre les paires de carbones alpha au sein d'une protéine ou entre les paires de phosphates dans le cas de l'ADN/ARN.
Par contre, l'utilisation des coordonnées cartésiennes comme données d'entrée du modèle parait plus concluante dans notre cas.

En effet, nous avons testé les deux procédures pour tICA en prenant en compte les coordonnées cartésiennes ainsi que les distances entre les pairs de phosphate de l'ADN [86]. Ceci a l'intérêt de rendre plus interprétable le modèle car on passe de 714 x 3 dimensions à 210.

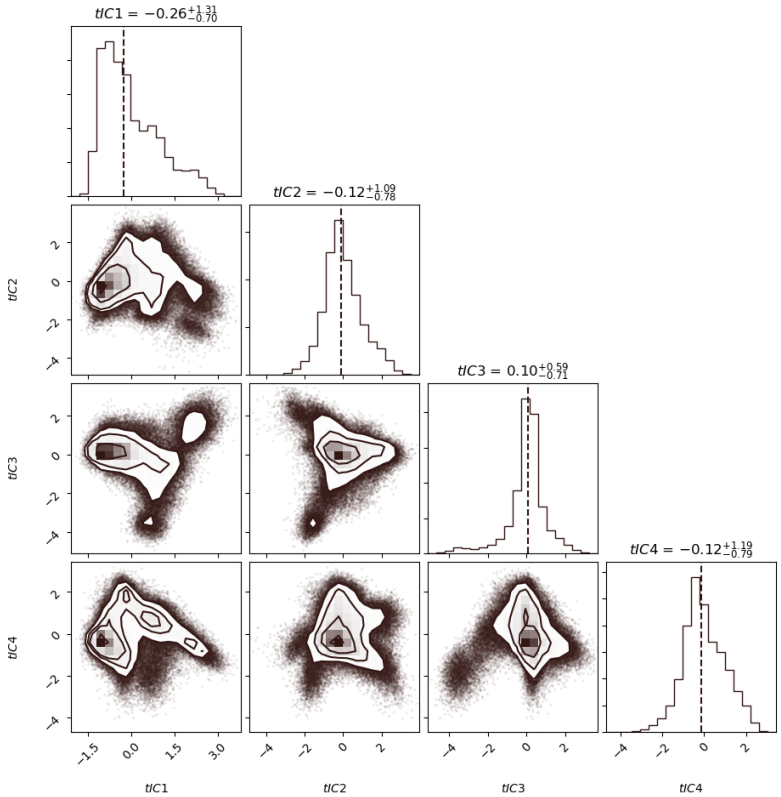
La Figure 3.11 représente les histogrammes/plot 2D sur les 4 premières tICA. Nous observons clairement une plus forte apparition de clusters avec la méthode basée sur les coordonnées cartésiennes. En effet, en réduisant à partir des distances entre les phosphates, il semble que nous perdons l'information sur les faibles fluctuations. Comme dans le cas de l'ADN il n'existe pas des changements conformationnels importants, la méthode basée sur la distance entre les phosphates est moins adaptée. Ainsi l'étape de "clusterisation" à été lancée en se basant sur la méthode basée sur les coordonnées cartésiennes.

Comme nous l'avons expliqué dans le chapitre précédent, DBSCAN est bien adapté dans notre cas car il est spécialement conçu pour cibler des clusters de forme arbitraire. En utilisant les deux paramètres ϵ (la distance à laquelle deux points sont considérés comme voisins) et $MinPts$ (le nombre minimum de points nécessaires pour définir un cluster), DBSCAN propose l'existence de 7 clusters indépendants, Figure 3.12. Nous proposons par la suite d'évaluer l'énergie libre de chaque cluster et de calculer les propriétés caractéristiques de chacun d'eux.

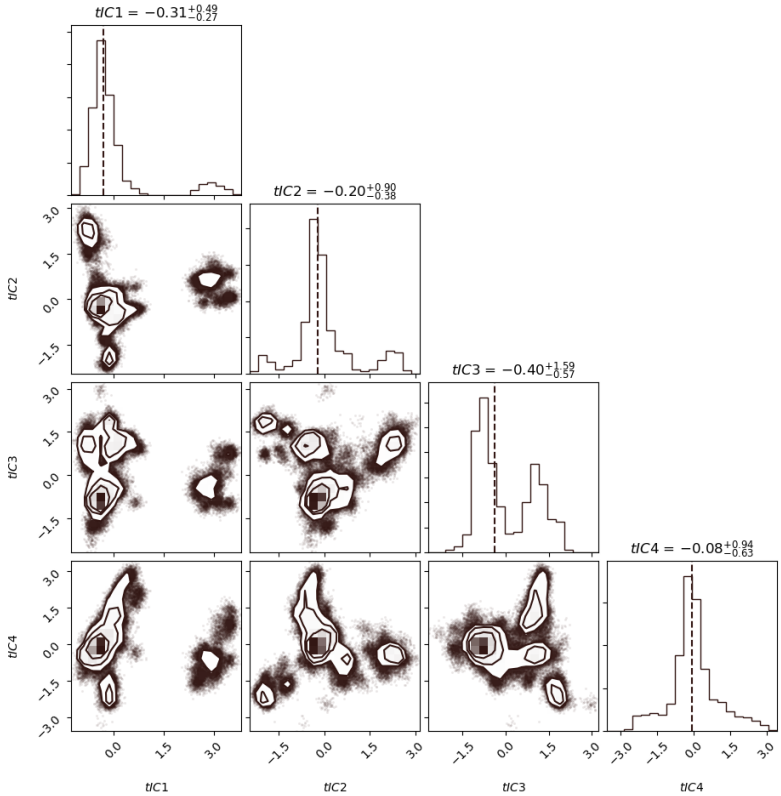
Extraction de l'énergie libre relative débiaisée entre domaines représentatifs

Le calcul d'énergie libre est très important dans les calculs de dynamique moléculaire, afin d'obtenir une description réduite d'un système complexe de grande dimension. En effet, les valeurs d'énergie libre donnent un aperçu des ensembles conformationnels de biomolécules de chaque cluster dans le but de trouver les conformations les plus probables d'un système moléculaire. Afin d'élucider les mécanismes sous-jacents aux changements de conformation, il était nécessaire d'extraire des états métastables avec un bruit limité, ceci est rendue possible grâce à la stratégie d'utilisation de l'approche tICA. Ainsi, dans le but d'évaluer la stabilité des clusters, l'énergie libre de chaque groupement est calculée à travers l'évaluation de la distribution de probabilité sur le nombre total de structures. Le Tableau 3.4 montre les valeurs de l'énergie libre de chaque cluster.

Cependant les énergies libres ne peuvent pas être totalement déterminées à partir d'un échantillon



a)



b)

Figure 3.11: Représentation des histogrammes/plot 2D sur les 4 premières tIC.

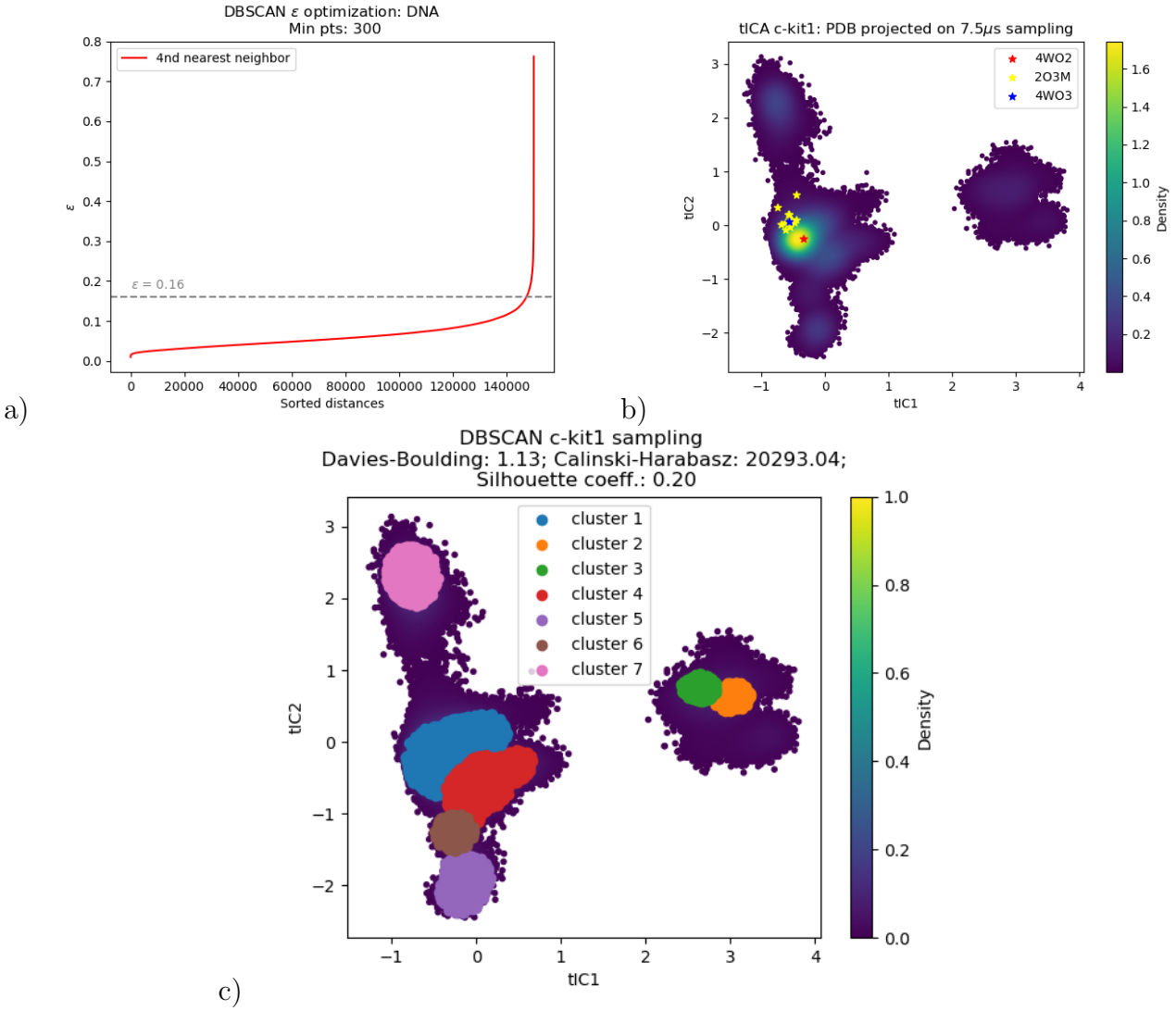


Figure 3.12: a) Optimisation des paramètres ϵ et $MinPts$ du DBSCAN. b) Représentation 2D des structures échantillonées pendant les 7.5 μs de simulation selon tIC1 et tIC2. Les différentes structures PDB disponibles dans la littérature sont y projetées. c) Identification des différents clusters par DBSCAN après optimisation des paramètres ϵ et $MinPts$.

(Kcal/mol)	Énergie libre	ΔG par rapport au cluster 1
Cluster 1	0.57	0
Cluster 2	2.36	1.79
Cluster 3	2.98	2.41
Cluster 4	2.55	1.98
Cluster 5	1.46	0.89
Cluster 6	2.58	2.01
Cluster 7	2.51	1.93

Table 3.4: Valeurs de l'énergie libre propre de chaque cluster identifié et de la valeur ΔG par rapport au cluster 1.

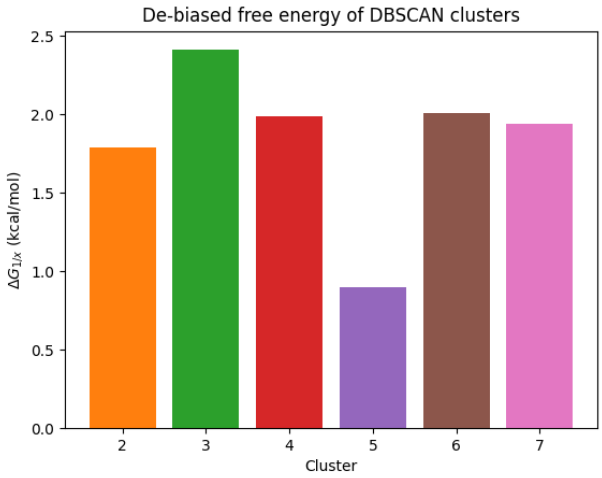


Figure 3.13: Représentation des énergies libres dé-biaisées relatives des clusters tICA DBSCAN, par rapport au cluster 1.

statistique car ces grandeurs dépendent de l'ensemble des configurations du système. Néanmoins, il est possible d'atteindre les différences d'énergies libres entre deux clusters, ce qui permet de quantifier la probabilité relative de différents états macroscopiques du système. Cette différence d'énergie libre est notée par la suite par ΔG.

La Figure 3.13 présente la différence de l'énergie libre entre chaque cluster et le cluster 1, puisque ce dernier représente le macro-état le plus stable ayant ce dernier étant le plus exploré. Un ΔG faible (<1 kcal/mol) est observé entre le cluster 1 et le cluster 5, ceci propose que ce dernier a une forte probabilité d'existence. Toutefois, la différence entre le cluster 1 et le cluster ayant la plus grande valeur de l'énergie libre ($E_{\text{cluster 3}}$ = 2.98 kcal/mol) n'est que de 2.41 kcal/mol.

RMSF

Nous avons calculé le RMSF pour les 7 clusters identifiés par DBSCAN en se basant sur tICA. Comme dans le cas l'analyse déjà faite sur l'ensemble des trajectoires issues de la stratégie d'échantillonage

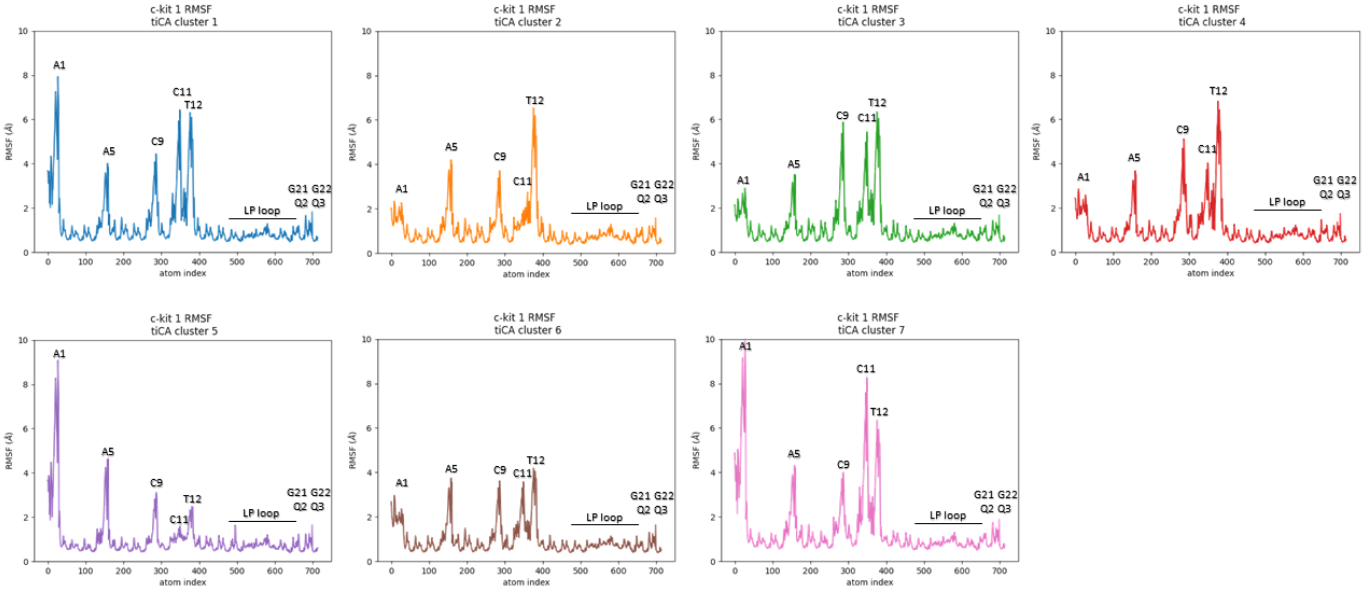


Figure 3.14: RMSF des clusters identifiés par tICA.

adaptatif non supervisé, les résidus les plus dynamiques sont ceux appartenant aux boucles simples (A1, A5, C9, C11 et T12). De façon intéressante, le RMSF de chacun de ces résidus hautement dynamiques est différent selon les clusters, Figure 3.14

En effet A1 semble être le plus mobile dans les clusters 7, 5 et 1 puisqu'il atteint respectivement 10, 9 et 8 ÅĊ11, bien qu'il effectue des fluctuations légèrement moins intenses que A1, il atteint 8 Å dans le cluster 7, 6 Å dans les clusters 3 et 4 Å dans les clusters 1, 4 et 6. T12 semble fluctuer autour de 6 Å dans les clusters 1, 2, 3, 4 et 7, et autour de 4 Å dans le cluster 6. A5 et C9 paraissent être plus homogènes au niveau de tous les clusters où ils fluctuent de l'ordre de 4 Å cependant, C9 peut atteindre les 6 Å dans les clusters 3 et 4. La boucle LP et le coeur du GQ semblent être très stables dans tous les clusters. Ces résultats viennent confirmer encore une fois la stabilité de notre système et valident l'aspect hautement dynamiques des nucléotides appartenant aux petites boucles qui relient les guanines formant le coeur du quadruplexe. Ces nucléotides semblent avoir un comportement différent selon les clusters.

Évaluation des angles de torsion χ et γ des nucléotides des boucles simples des différents clusters

Étant donné que A1, A5, C9, C11 et T12 sont les résidus les plus flexibles du GQ c-kit 1, ils semblent être responsables de la formation de 7 clusters différents. Il est donc important de comprendre comment ils fluctuent dans chaque cluster afin d'appréhender quelles conformations sont les plus intéressantes en vue de la conception d'une future molécule stabilisante. Une façon d'y parvenir est d'évaluer l'angle de torsion des résidus les plus dynamiques. Considérant les angles χ et γ comme le montre les Figures 3.15 et 3.16, le cluster 1 semble être largement dispersé, notamment au vu de

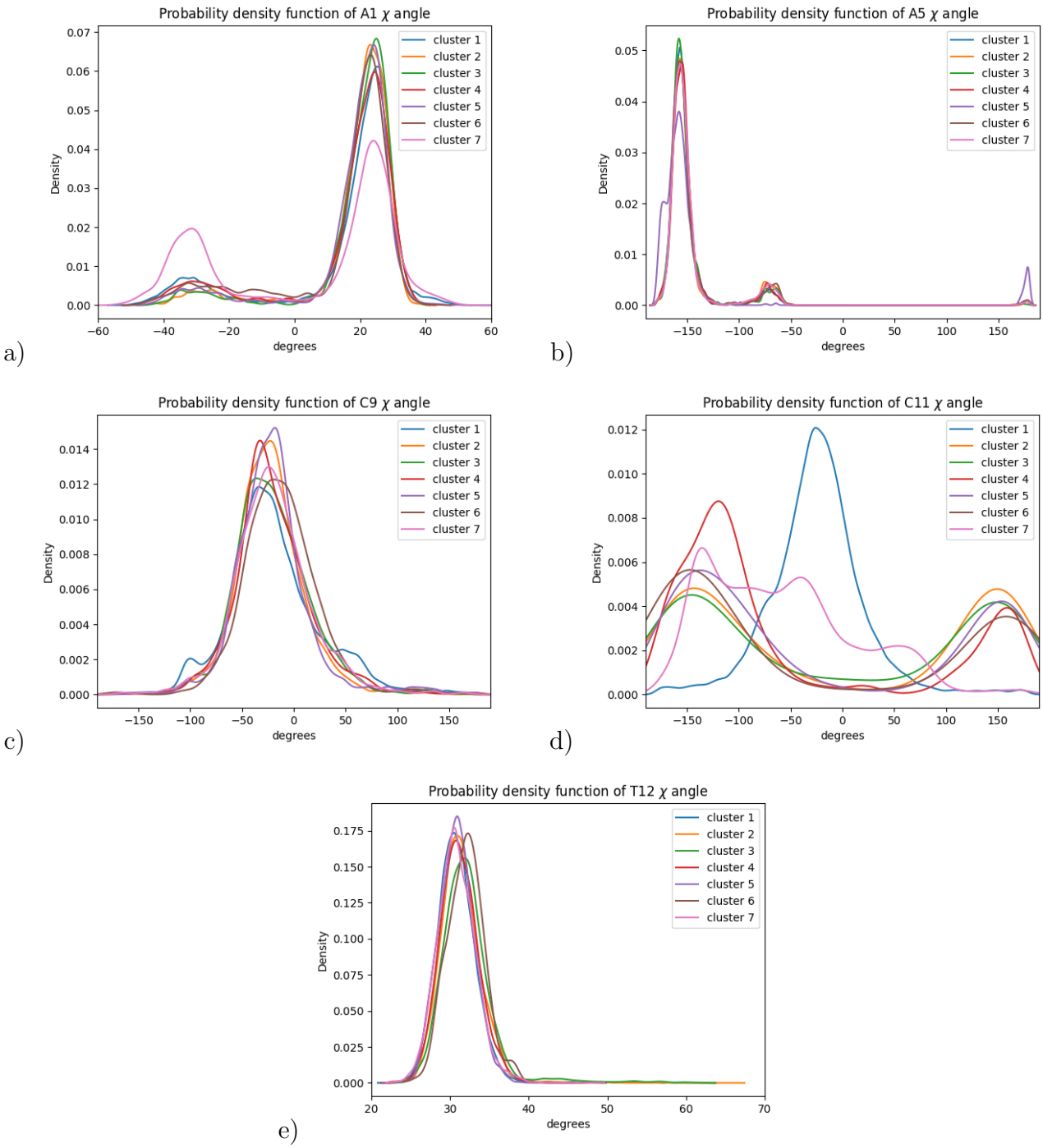


Figure 3.15: Représentation graphique de la fonction de la densité de probabilité de l'angle χ des nucléotides A1, A5, C9, C11 et T12 dans les sept clusters tICA, respectivement en a), b), c), d) et e).

A1, C9 et C11, ce qui signifie que ces nucléotides sont les plus fluctuants selon leurs angles de torsion (γ pour A1 et χ pour C9 et C11). Nous rappelons que l'angle χ réfère aux conformations *syn/anti* et que l'angle γ correspond à l'orientation du sucre par rapport au squelette. Spécifiquement, γ de A1 exprime plus de liberté dans le cluster 1, alors qu'il est plus limité dans les autres clusters. Bien

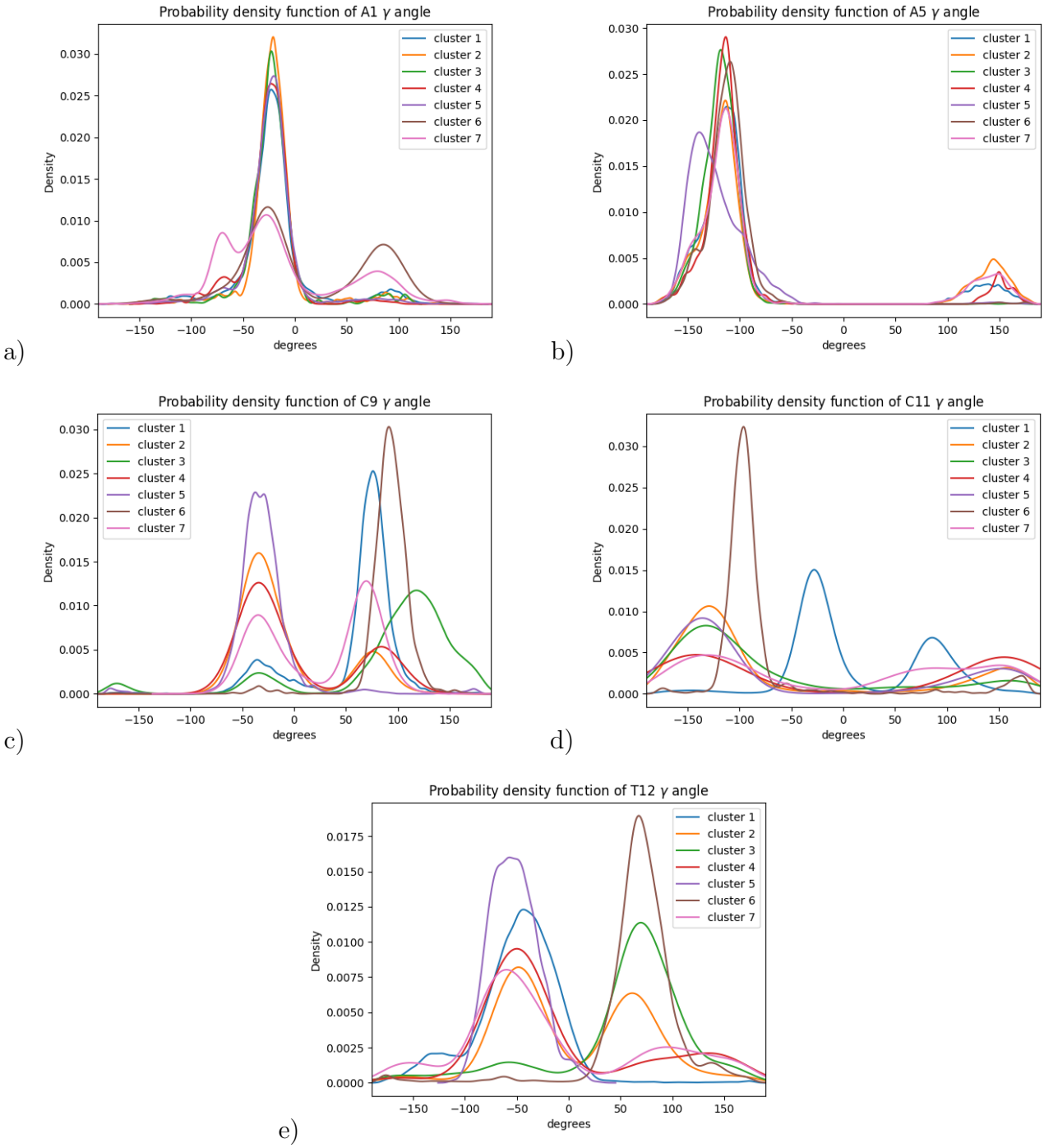


Figure 3.16: Représentation graphique de la fonction de la densité de probabilité de l'angle γ des nucléotides A1, A5, C9, C11 et T12 dans les sept clusters tICA, respectivement en a), b), c), d) et e).

que le cluster 3 révèle des angles γ particuliers pour C9 (entre 80° et 180°), les clusters 2, 4, 5, 6 et 7 expriment des angles similaires entre eux qui varient entre 0° et -70° et entre 50° et 90°. Cette similitude entre les clusters (à l'exception du cluster 1) est également observée pour le nucléotide C11. En fait, C11 peut largement adopter une très large gamme d'angles γ et χ dans le cluster 1.

Cependant, d'autres clusters représentent une gamme limitée de γ et χ entre 100° et 180° et -100° et -180°. Quant au résidu A5, ses angles χ et γ sont plus restreints dans tous les clusters. En fait, ils semblent être concentrés autour de -180° et -50° pour l'angle χ et entre -50° et -180° et 100° à 150° pour l'angle γ. Curieusement, T12 montre des degrés très restreints pour l'angle χ limité entre 30° et 45°, alors qu'il peut adopter des degrés illimités pour les angles γ.
Ces observations démontrent une grande flexibilité des nucléotides C11 et C9 capables d'adapter une large marge de degrés pour l'angle χ proposant un état d'équilibre entre les conformations *syn* et *anti*. Toutefois une conformation *syn* est favorisée pour C9 dans les clusters 1 jusqu'au 7, alors qu'une conformation *anti* est plutôt favorisée pour C11 à l'exception du cluster 1 qui peut alterner entre les deux conformation. A1 et T12 conserve une conformation *syn* mais exprime un degré de liberté qui est très vaste pour son angle γ. A5 semble être beaucoup moins flexible dans tous les clusters.

Comparaison des distances A1–C11 and A1–T12 entre les différents clusters

Nous avons détecté précédemment des interactions de types Watson-Crick qui peuvent exister transitoirement entre A1 et C11 et entre A1 et T12. Ceci à été déduit en évaluant la distance A1–C11 et A1–T12. Nous avons calculé ces mêmes distances pour les 7 clusters identifiés. Les résultats sont présentés dans la Figure 3.17. Dans le cluster 1 et 7, nous ne détectons pas de distances assez proches ni pour A1–C11, ni pour A1–T12. Par contre, une interaction entre A1 et C11 peut exister dans les clusters 2, 3, 4, 5 et 6. Curieusement T12 ne semble interagir avec A1 dans aucun des clusters. En effet, nous avons démontré précédemment après calcul des distances A1–C11 et A1–T12 que des interactions WC peuvent exister entre A1 et T12. Le fait de ne pas pouvoir détecter cette interaction dans les 7 clusters identifiés, nous mène à croire que les structures présentant une interaction A1–T12 au-dessus de la première tétrade, ne sont pas favorisées puisqu'elles n'apparaissant pas dans les clusters identifiés. Par contre, les structures partageant une interaction A1–C11 sont plus favorisées par la clusterisation.

Calcul des empilements π-π entre les différents clusters

L'index d'empilement π-π nous a servi précédemment, à évaluer la robustesse de l'empilement de A1, C11 et T12 au-dessus de la première tétrade de guanine. Nous avons conclu en analysant toute la trajectoire non "clusterisée" que A1 a tendance à s'empiler au-dessus de G13 et que C11 et T12 sont en compétition d'empilement au-dessus de G10 bien que T12 à plus de chance de s'empiler au-dessus de G10 que C11. Nous cherchons ici à évaluer ces indices d'empilement en prenant en considération les 7 clusters identifiés. En effet, selon la Figure 3.18, pour les clusters 1 à 6, A1 semble partager une interaction π avec G13 conformément à notre analyse précédente. Cependant la probabilité de la densité des structures du cluster 7 est diffuse et ne montre pas un pic intense pour un indice particulier, ce qui signale presque une absence d'interaction π de A1 avec G2 et G13. La Figure 3.19 montre que C11 semble s'empiler au-dessus de G10 plutôt que de G16 dans les clusters 1, 2, 3, 4

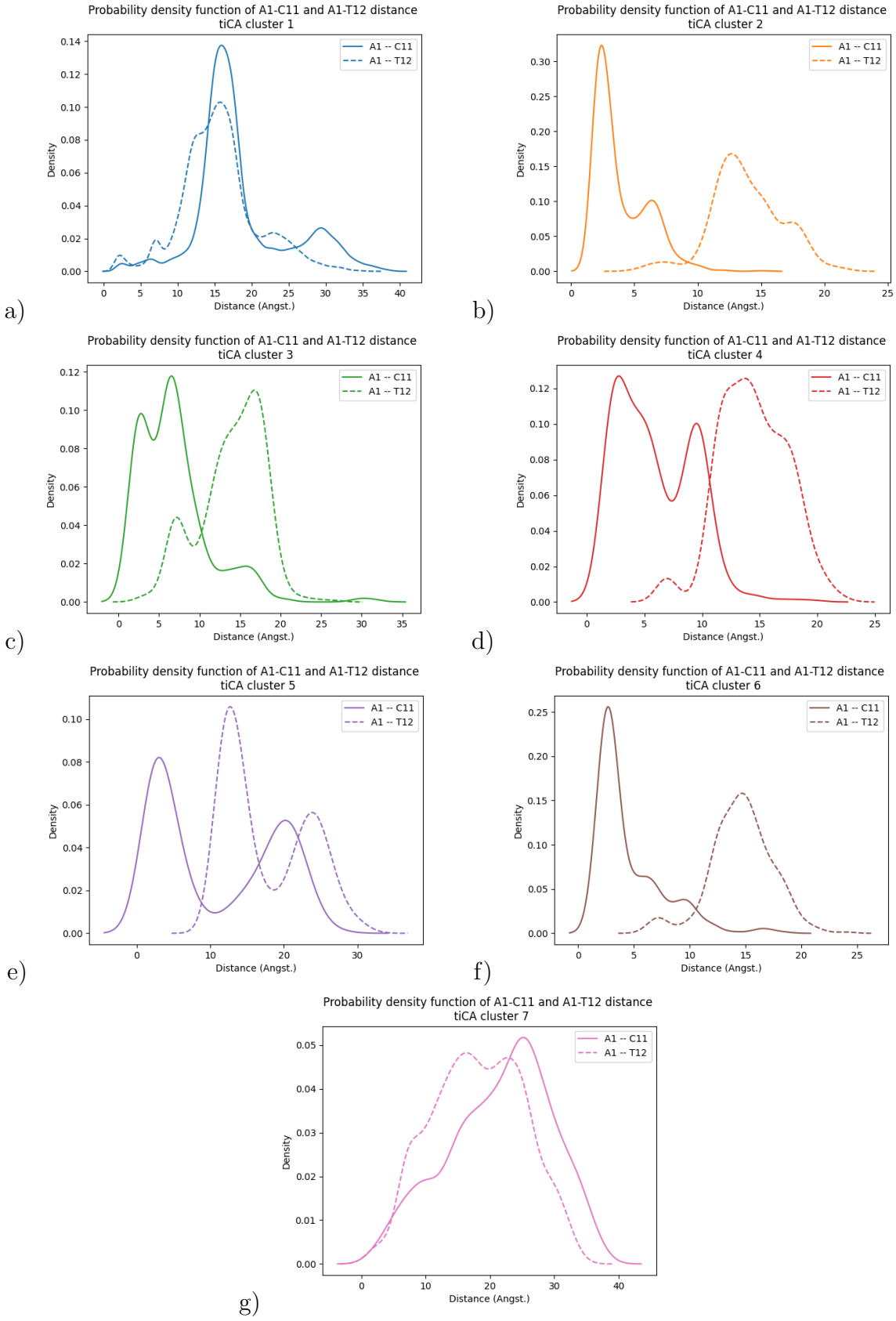


Figure 3.17: Représentation graphique des fonctions de la densité de probabilité des distances A1–C11 et A1–T12 selon les sept clusters.

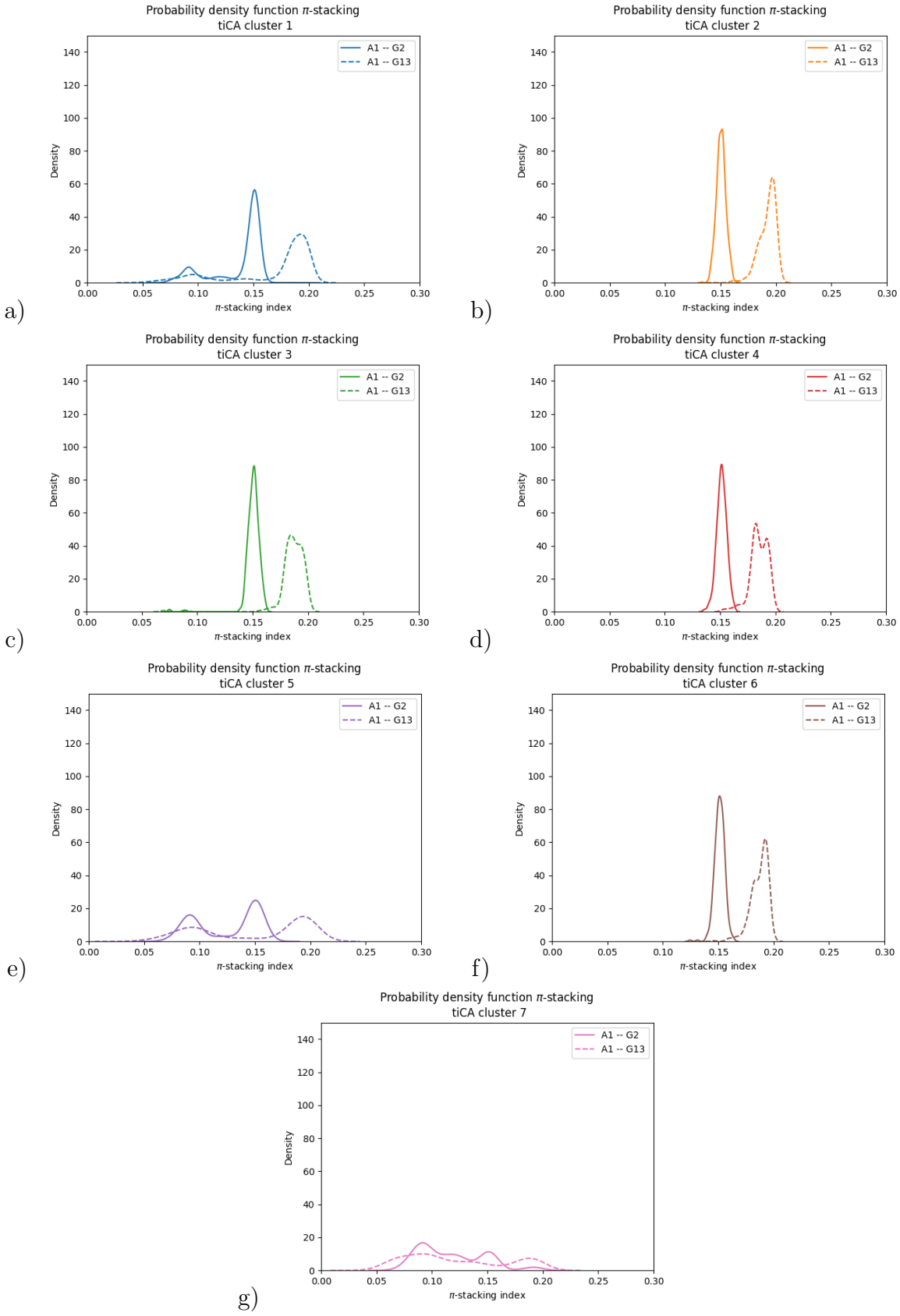


Figure 3.18: Représentation graphique des fonctions de la densité de probabilité de l'indice de l'empilement π-π entre A1-G2 et A1-G13 pour chacun des clusters tICA.

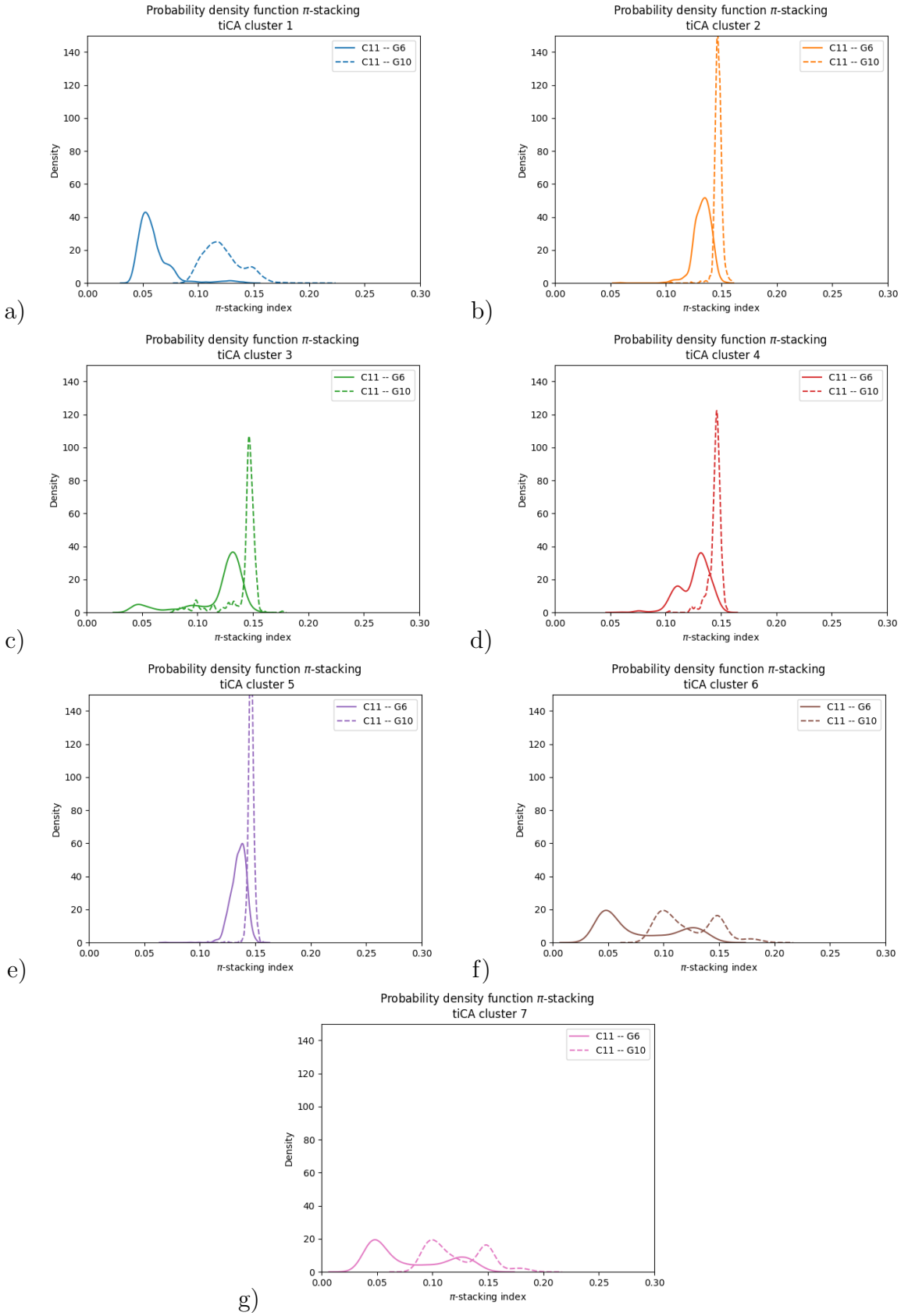


Figure 3.19: Représentation graphique des fonctions de la densité de probabilité de l'indice de l'empilement π-π entre C11-G6 et C11-G10 pour chacun des clusters tICA.

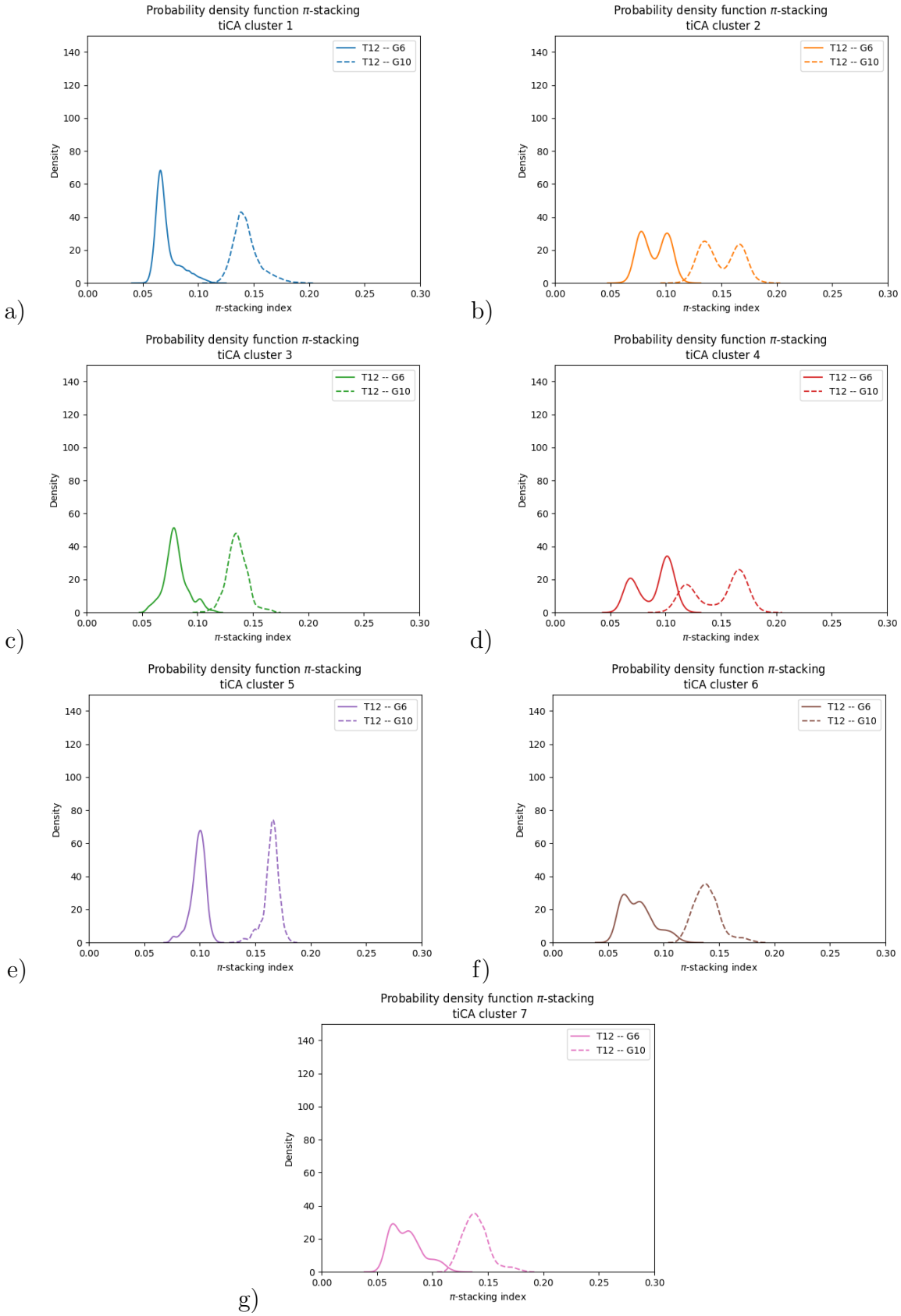


Figure 3.20: Représentation graphique des fonctions de la densité de probabilité de l'indice de l'empilement π-π entre T12-G6 et T12-G10 pour chacun des clusters tICA.

et 5. Dans les clusters 6 et 7, ces interactions ne semblent pas être d'une importance significative. En ce qui concerne T12, Figure 3.20, son empilement au-dessus de G10 est bien mis en évidence dans tous les clusters. L'analyse de ces résultats viennent dans le même esprit que celle faite sur la trajectoire complète non "clustérisée". Cependant l'intérêt de la clusterisation s'inscrit dans le but de pouvoir comprendre les différents changements structuraux qui peuvent exister au sein de chaque groupe de structures, permettant ainsi une meilleure caractérisation de conformations afin de faciliter ultérieurement la conception de futures molécules stabilisantes de GQ.

3.2.3 Stabilisation du GQ de c-kit 1 par les ions K^+ et les molécules d'eau polarisables

Sur la base des structures cristallographiques aux rayons X à haute résolution, il a été suggéré que les deux cations K^+ à l'intérieur du canal exercent un rôle stabilisateur essentiel sur le GQ du promoteur de c-kit 1.

Afin d'évaluer l'importance des cations, à l'intérieur du canal ionique formé par le noyau de guanine, sur l'intégrité du GQ, 10 simulations cMD indépendantes de 20 ns chacune ont été simulées en partant d'une structure de départ qui ne présente aucun ion à l'intérieur du canal ionique. Cette structure de départ à été obtenue en déplaçant les deux K^+ structuraux en dehors du GQ de c-kit 1, vers les molécules d'eau en vrac. Afin de détecter le nombre des ions K^+ et des molécules d'eau entrant dans le canal vide, nous avons considéré une sphère de rayon 5 Å centrée sur le centre de masse du noyau GQ. Comme le montre la Figure 3.22 et le Tableau 3.5, 3 scénarios se sont produits. Dans le premier, deux K^+ entraient dans le canal au niveau de la première tétrade et restaient en place, chacun à mi-chemin entre l'un ou l'autre des deux couples successifs de tétrades. Dans le second, un seul K^+ y est entré avec une ou deux molécules H_2O. Dans le troisième, il y avait trois molécules H_2O et aucun K^+ à l'intérieur du canal. Ces événements se produisent au tout début de la MD.

Un mécanisme représentatif d'entrée des K^+ peut être illustré par la première trajectoire (Figure 3.21). Une fois le canal ainsi rempli, toutes les boucles évoluent de la même manière que dans les simulations MD initiales avec deux cations K^+ présents dès le début de la MD sans changements structurels relatifs significatifs.

La Figure 3.22, c) montre les valeurs moyennes des moments dipolaires pour les molécules H_2O détectées à l'intérieur du canal GQ. Leurs valeurs moyennes (environ 2.72 D) sont légèrement inférieures à la valeur de référence globale AMOEBA (2.78 D). En fait, comme nous l'avons démontré dans le chapitre précédent, un environnement dense et compact génère un effet dépolarisant global (par rapport à l'eau en vrac) influençant les dipôles induits par les molécules d'eau.

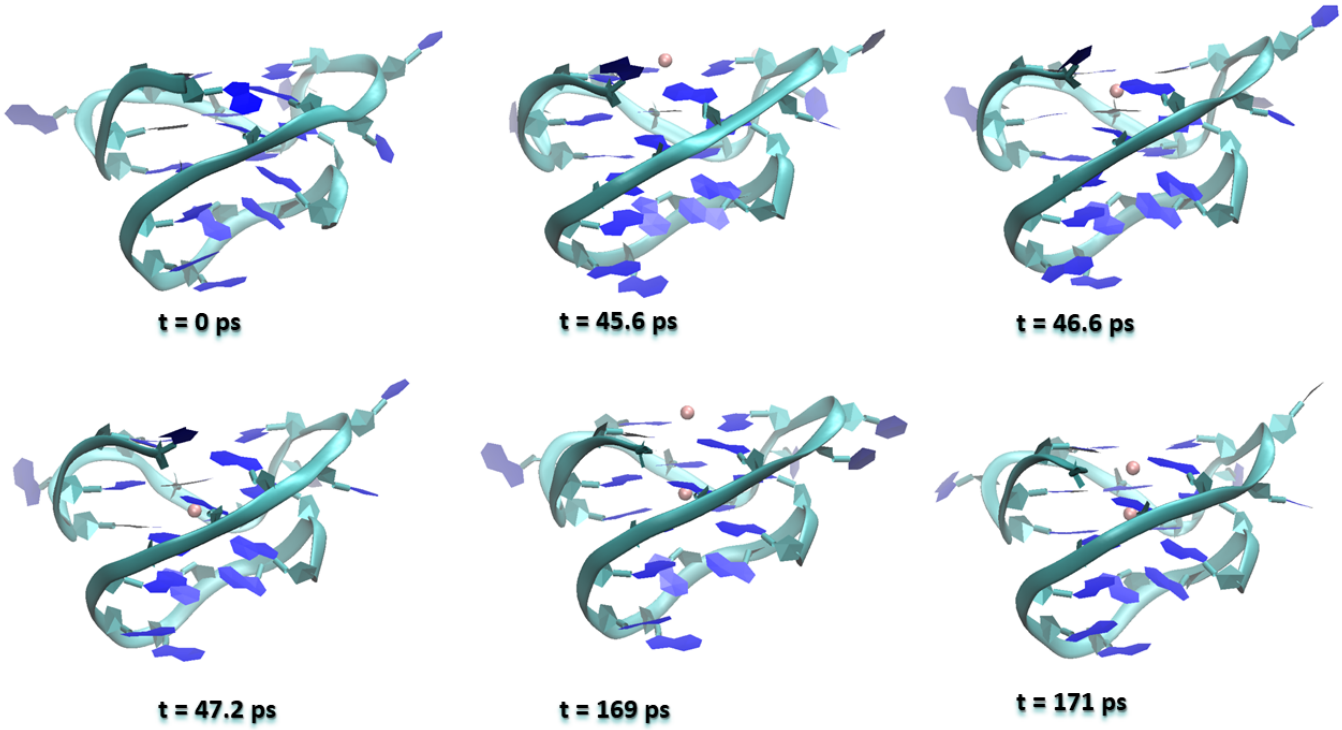
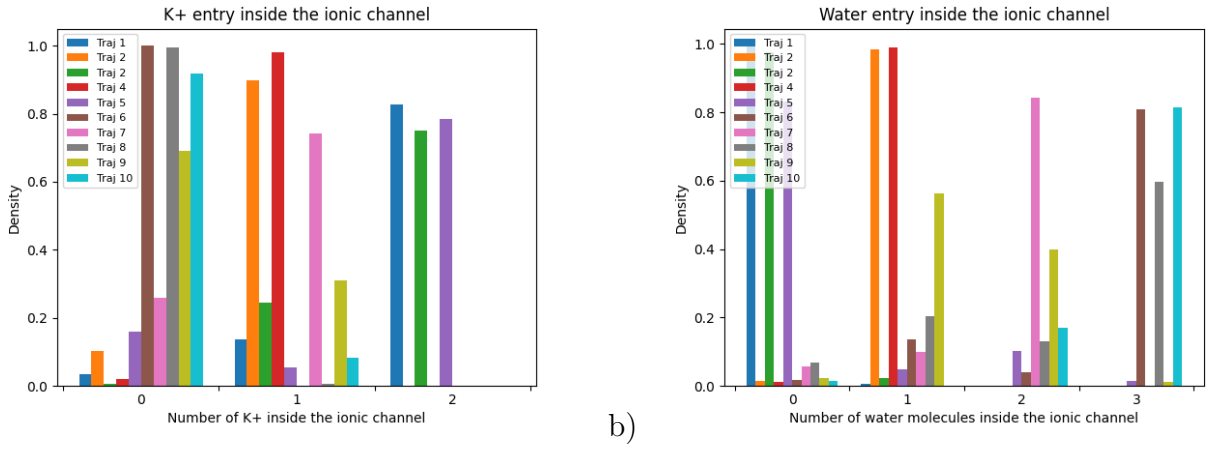


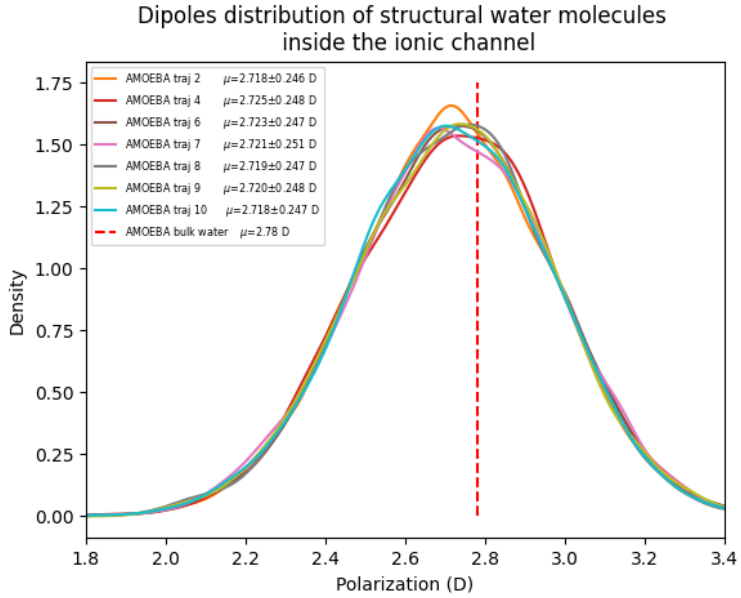
Figure 3.21: Représentation du mécanisme d'entrée des 2 K^+ à l'intérieur du canal ionique du GQ de c-kit 1 en prenant comme exemple la première trajectoire. À 45.6 ps, un ion K^+ se lie au-dessus de la première tétrade. À 46.6 ps, il se déplace le long de l'axe perpendiculaire et se lie par interaction non covalente aux guanines de la première et de la deuxième tétrade. À 47.2 ps l'interaction se traduit par une interaction du K^+ aux guanines de la deuxième et de la troisième tétrade. Il reste ensuite en place jusqu'à la fin de la simulation. À 169 ps, un deuxième K^+ se lie en bloc au-dessus de la première tétrade. Il se traduit alors par une interaction non covalente avec les guanines de la première et la deuxième tétrade.

3.2.4 Poches cryptiques au niveau des GQ de c-kit 1 et du VIH-1 LTR-III

Afin de détecter des cavités potentielles de liaison aux médicaments au sein des GQ, nous avons eu recours au logiciel DoGSite Scorer [220], un descripteur automatisé de poches cryptiques. Comme le montre la Figure 3.23, a) DoGSite Scorer a détecté 2 poches à l'intérieur de la structure cristalline du GQ c-kit 1. La première poche (P1) est formée par les résidus : A1, G2, G3, G13, G14, G15, A16. Tandis que la seconde poche (P2) est constituée par les résidus : G7, G8, C9, G10, C11, G21.
Dans le but d'évaluer la persistance de ces poches cryptiques tout au long de la MD adaptative, nous avons calculé certaines distances inter-bases clés au sein des poches cryptiques identifiées et nous avons comparé leur valeurs dans les sept clusters. Pour ce faire, nous avons considéré les paires de bases G3–G15, G2–G14 et A1–G13 à l'intérieur de P1 et les paires de bases G21–G8 et G10–G7 à l'intérieur de P2 et nous avons évalué la densité de probabilité de la distance entre chaque paire au sein de chaque cluster. Les distances intermoléculaires que nous avons suivies sont comprises entre N7 du premier G et HN1 du second. Dans la paire A–G, la distance correspondante se situe entre



a) b)



c)

Figure 3.22: a) et b) Représentation du nombre des ions K^+ et des molécules d'eau à l'intérieur du canal ionique après 20 ns de simulation cMD à partir d'une structure de départ ayant le canal ionique vide. c) Représentation des fonctions de la probabilité de densité de la distribution du moment dipolaire des molécules d'eau détectées à l'intérieur du canal ionique.

	Occupation en K^+	Occupation en H_2O
Trajectoire 1	2	0
Trajectoire 2	1	1
Trajectoire 3	2	0
Trajectoire 4	1	1
Trajectoire 5	2	0
Trajectoire 6	0	3
Trajectoire 7	1	2
Trajectoire 8	0	3
Trajectoire 9	0-1	1-2
Trajectoire 10	0	3

Table 3.5: Occupation du canal formé par les 16 guanines constituant le coeur du GQ. Chaque trajectoire est une cMD indépendante de 20 ns simulée en partant d'une structure de base ayant le canal ionique vide.

N7 de A1 et HN1 de G13 qui fait partie de la première tétrade. Ce choix des distances G–G devrait permettre d'évaluer si un dépilage partiel des quadruplexes ou une distorsion des liaisons-H au sein du GQ ont pu se produire au cours de la MD.

La Figure 3.24 montre que dans les sept clusters, pour chaque paire G–G, ces distances restent étroitement distribuées autour de pics aigus dans un intervalle de 4 à 4.5 Å. Ceci montre la stabilité du noyau du GQ dans tous les clusters. Elle doit favoriser la fixation d'un ligand entrant dans le sillon formé par les bases appartenant à P1 ou à P2. Par contre, la distance entre les bases de la paire A1–G13 subit de fortes fluctuations dans tous les clusters, notamment dans les clusters 5 et 7. Ceci est clairement causé par le mouvement de grande amplitude de A1, dont le RMSF dépasse 9 Å dans ces 2 clusters. En particulier, cette flexibilité de A1 au niveau de P1 pourrait faciliter une éventuelle interaction d'un ligand potentiel dans le sillon de l'ADN : lorsque A1 est libre, cela ouvre la voie à un ligand potentiel pour accéder au sillon et alors une adaptation conformationnelle de A1 peut être nécessaire pour interagir avec le ligand (sorte de porte qui se ferme après l'entrée du ligand). Cette hypothèse nécessite des simulations supplémentaires de GQ en présence de ligands potentiels et des calculs de l'énergie libre d'interaction entre ces deux molécules.

Nous avons effectué une extension préliminaire d'AMOEBA à la région LTR-III du VIH qui est également sujette à la formation de GQ. À ce stade, nous avons réalisé une cMD de 400 ns en utilisant une structure RMN à haute résolution comme point de départ (PDB ID : *6h1K*). Le logiciel DoGSite Scorer [220] a pu détecter deux poches cryptiques : P1, formées par les bases G2, A4, G5, G6, G8, T9, G11, C12, C13, T14, G15 et G19 et P2 formées par les bases G1, G2, G3, G15, G19, G20 et G21. P1 est principalement situé dans le domaine en épingle à cheveux tandis que P2 est situé dans la face supérieure du domaine GQ (Figure 3.26 a)). Similaire au GQ du c-kit 1, les calculs RMSD et RMSF (Figures 3.25 a) et b)) attestent d'un noyau GQ hautement structuré. En revanche, le domaine en épingle à cheveu et les boucles à trois nucléotides et à un seul nucléotide ne sont pas structurés.

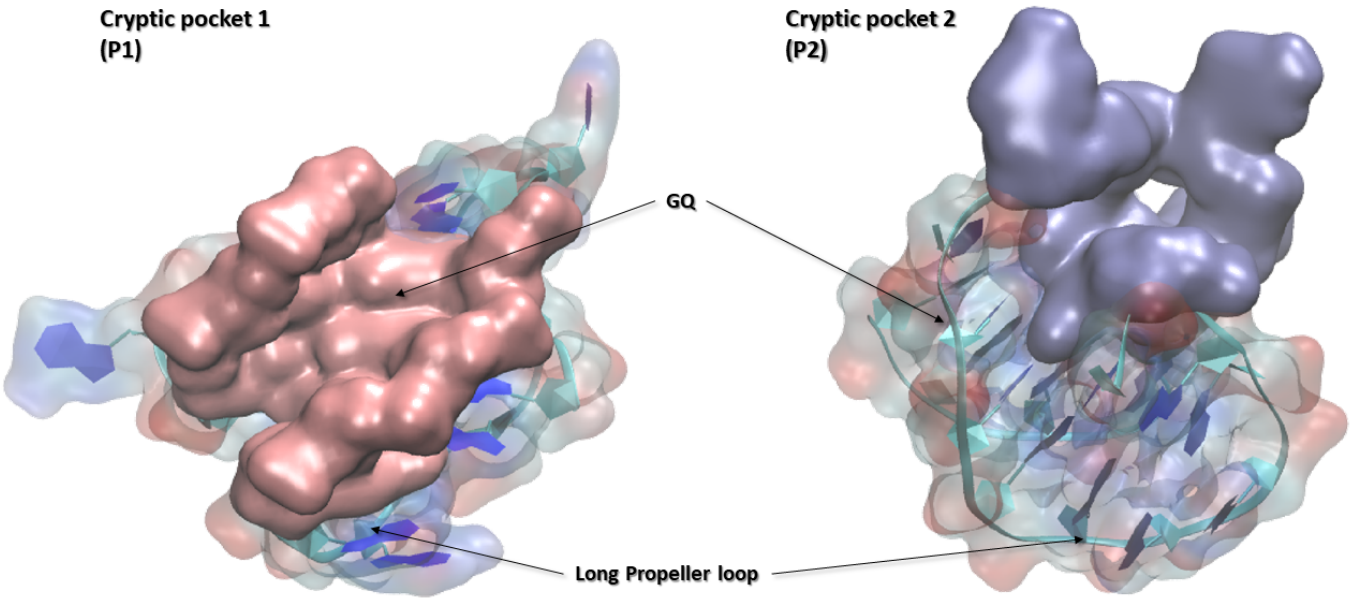


Figure 3.23: Représentation des deux poches cryptiques à l'intérieur de la structure GQ c-kit 1 : P1 (couleur rose) est formé par les bases A1, G2, G3, G13, G14, G15 et A16 et P2 (bleu glacier) est formé par les bases G7, G8, C9, G10, C11 et G21.

Comme indiqué ci-dessus pour le cas de c-kit 1, nous avons évalué les évolutions des distances inter-bases représentatives à l'intérieur de P1 et P2. En P1, nous avons considéré les trois paires de bases de type W-C A4–T14, G5–C13 et G6–C12. Dans P2, les deux bases sont G2–G19 et G1–G20. Pour la paire de bases A–T, les distances sont comprises entre H3N(T) et N1(A). Pour les deux paires de bases G–C, elles sont comprises entre H1N(G) et N3(C). Dans P2, pour G2–G19 et G1–G20, les distances ont été calculé entre les atomes O6 des deux bases. Ce choix, plutôt que celui fait pour le c-kit 1, a été motivé par les différents angles glycosidiques des bases consécutives : *anti* pour G2 et G20, et *syn* pour G1 et G19. Il en résulte les atomes N7 de deux bases consécutives situées sur des brins opposés. Les distances O6-O6 ont été considérées comme moins sensibles à l'alternance des angles glycosidiques. En revanche, toutes les guanines du noyau GQ dans c-kit sont *anti*.

Les fonctions de densité de probabilité de distance (Figure3.26 b), attestent d'une plus grande stabilité de P2 que de P1. Cela devrait avoir des implications claires concernant la conception des ligands spécifiques des GQ du LTR-III, qui devraient cibler P2 plutôt que P1. On observe ainsi que les deux distances O6-O6 dans P2 sont clairement centrées autour de deux pics dans un intervalle de 4.5-4.8 ÅËn P1, la distance entre la paire de bases A4–T14 a deux pics. Le premier correspond à une distance standard de liaison W-C, mais le second, à 4.2 Å correspond à une paire dissociée. Pour G6–C12, de même, il y a deux pics, le premier à 2.0 Å correspondant à une paire de bases W-C, et

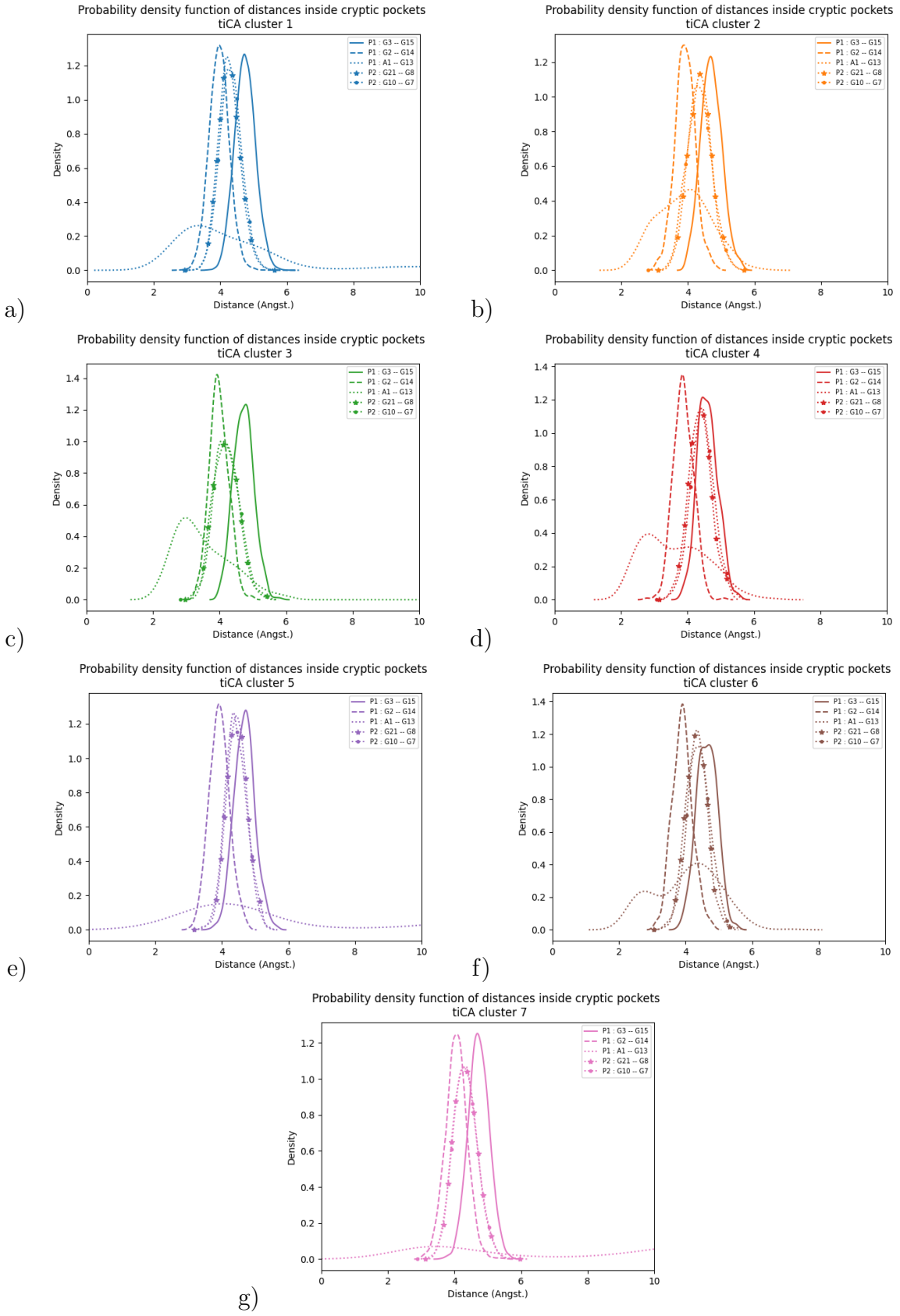


Figure 3.24: Représentation des fonctions de densité de probabilité des distances entre les nucléotides à l'intérieur des poches cryptiques identifiées.

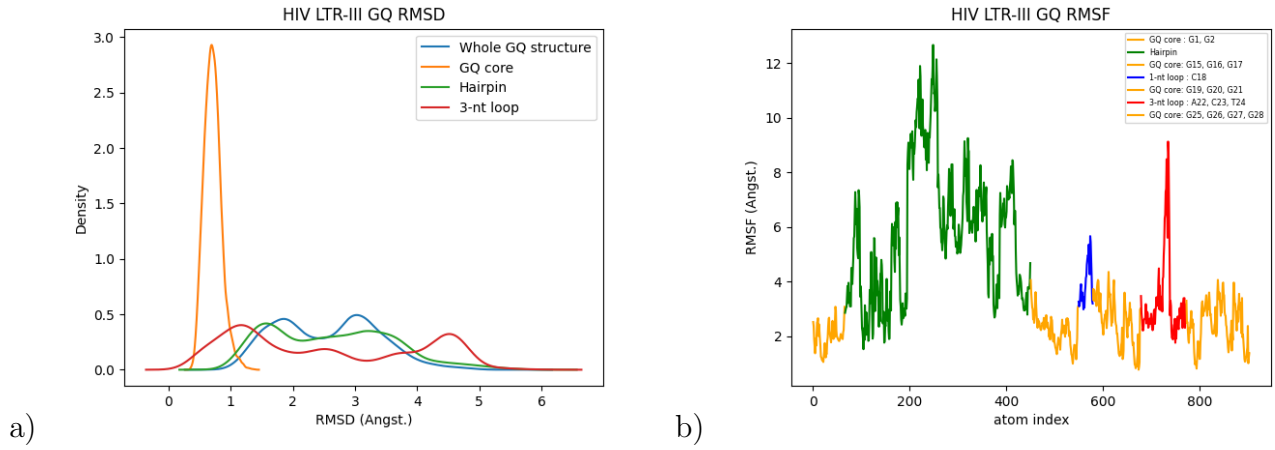


Figure 3.25: a) Représentation du RMSD (fonctions de probabilité de densité) and b) du RMSF du GQ au niveau de la région LTR-III du VIH-1.

le second à 4 Å correspondant à une paire de bases dissociée. Pour G5–C13, une faible évolution est trouvée dans la gamme 2 – 4.5 Å .

3.3 Conclusions

Dans le but de pouvoir proposer des structures GQ de haute qualité capables de servir à la conception de futures ligands thérapeutiques, nous avons effectué des simulations de MD polarisables AMOEBA de longue durée allant jusqu'à plusieurs μs sur deux GQ différents ; le premier est rencontré au niveau du promoteur du proto-oncogène c-kit et le deuxième est rencontré au niveau de la région LTR-III du rétrovirus VIH-1.

Nous nous sommes ainsi focalisés dans un premier temps sur un oligonucléotide 22-mer appartenant à l'oncogène c-kit 1, dont la surexpression peut avoir des conséquences néfastes conduisant au cancer, notamment le cancer gastro-intestinale. Cet oligonucléotide peut se replier en quadruplexes de guanine, un changement de conformation se produisant lors de la transcription de l'ADN en ARNm CD117. Il était donc essentiel de mieux comprendre la dynamique de cet oligomère de 22-mères dans sa conformation quadruplexe.

Notre étude a bénéficié de plusieurs atouts. Un élément essentiel est le recours à un champ de force précis, AMOEBA, pour les simulations d'ADN et d'ARN, récemment validé par les résultats *ab initio* [238]. Un autre est l'utilisation d'une extension GPU [7] du logiciel Tinker-HP [114] permettant une augmentation de l'ordre de grandeur de ses performances par rapport à la version CPU standard. De plus, la MD conventionnelle (cMD) a été complétée par des approches récemment adaptées en Tinker-HP, à savoir la MD adaptative non supervisée reposant sur des analyses PCA et tICA [93]. Nous avons suivi de près la stratégie adoptée pour étudier la dynamique conformationnelle du dimère de protéase SARS-CoV-2 [59]. Ceci nous a permis de couvrir un temps total de simulation de 7.5 μs.

Les analyses tICA ont permis d'identifier sept clusters distincts stabilisés transitoirement au cours de la MD.

Plusieurs résultats étaient tout à fait cohérents avec ceux des simulations précédentes utilisant les champs de force classiques : AMBER [90] et CHARMM [190] ou polarisable : Drude [190]. Cependant, quelques différences ont été trouvées. La stabilité au cours du temps du noyau GQ, constitué de trois quadruplexes de guanine empilés, et celle de la Long-Propeller Loop (LP loop), ont été confirmées. Ceci contraste avec la forte dynamique/mobilité des boucles mononucléotidiques A5, C9 et de la boucle dinucléotidique C11-T12, dont les mobilités impactent toute la dynamique du 22-mer dy GQ de c-kit 1.

De telles mobilités pourraient être évaluées dans les clusters individuels en termes de :

a) Distances de liaison H dans les paires de bases A1–T12 et A1–C11, A16–G20 et G17–G20.

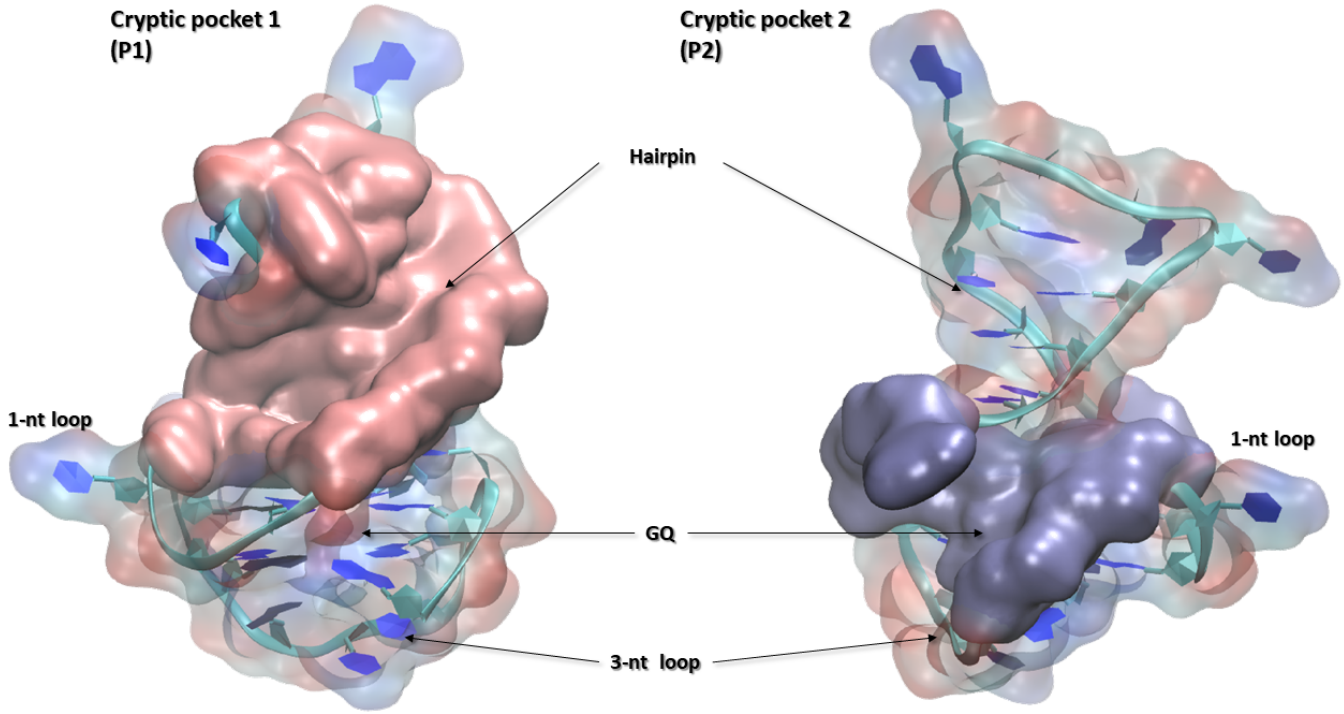
b) Deux angles de torsion : les angles glycosidiques χ et γ des bases A1, A5, C9, C11 et T12.

c) L'indice d'empilement π entre A1 et les deux bases, C11 et T12 des boucles dinucléotidiques, et les bases G de la première tétrade de guanine les plus proches, dont nous avons suivi l'évolution au cours de la MD adaptative pour les paires de bases A1–G2, A1–G13, C11–G6, C11–G10, T1–G6 et T12–G10.

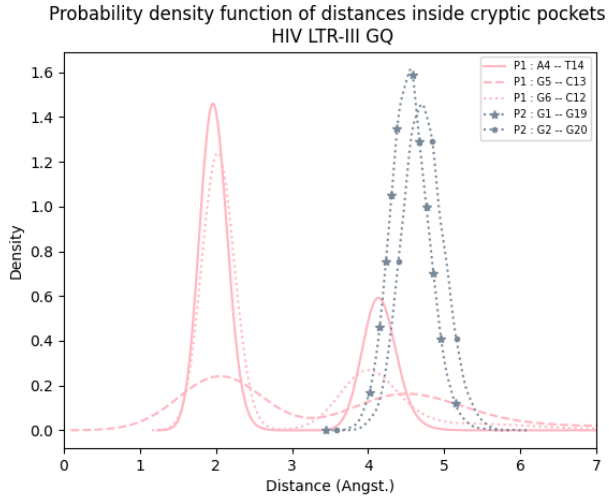
Les analyses tICA ont permis d'identifier sept clusters distincts, qui pourraient être transitoirement stabilisés au cours d'une MD de longue durée. Le logiciel DoGSite Scorer [220] a démêlé deux poches cryptiques, P1 et P2, qui pourraient servir de sites "droguables" pour des ligands pharmacologiques potentiels. Les Figures 3.23 et 3.26, b) donnent une représentation de la surface accessible de P1, révélant clairement l'existence d'un sillon pour leur fixation.

Dans ce sens, nous avons de façon préliminaire prolongé cette étude par une cMD de 400 ns sur un autre oligonucléotide formant un GQ, celui du LTR-III du VIH-1. DoGSite Scorer a ensuite identifié deux poches cryptiques, P1 et P2, dont seule P2, présentant des variations limitées de RMSD, pourrait éventuellement être utilisée comme site de liaison de ligand médicamenteux. La Figure 3.26, a) donne une représentation de la surface accessible de P2.

Sur la base de ces résultats, nous avons initié dans nos Laboratoires la conception de nouveaux ligands ciblant les sillons des poches cryptiques P1 et P2 de c-kit 1, et de P1 du LTR-III du BIH-1. Cela se fait en suivant de près les procédures utilisées dans le présent travail, et les résultats seront communiqués en temps voulu. Il est également prévu d'étendre ces études à d'autres quadruplexes formant G à partir d'autres séquences oncogènes et rétrovirales.



a)



b)

Figure 3.26: a) Représentation des deux poches cryptiques au sein de la structure GQ du LTR-III du VIH-1 : P1 (couleur rose) est formé par les bases localisées dans le domaine en épingle à cheveux (G2, A4, G5, G6, G8, T9, G11, C12, C13, T14, G15 et G19) et P2 (bleu glacier) est formé par les bases localisées au niveau du noyau du GQ (G1, G2, G3, G15, G19, G20 and G21). b) Représentation des fonctions de densité de probabilité des distances entre les bases des paires A4–T14, G5–C13, G6–C12 à l'intérieur de P1 et de G2–G19 et G1–G20 à l'intérieur de P2.

Chapitre 4

Conclusions générales et futures perspectives

4.1 Rappel de notre problématique

Afin d'aboutir aux objectifs de cette Thèse, il convenait d'appliquer des simulations de dynamique moléculaire polarisable massive sur des systèmes biologiques complexes d'importance thérapeutique dans le cadre de conception rationnelle de médicaments anti-cancéreux et anti-viraux (contre le SARS-CoV-2 et le VIH-1). Ainsi l'idée centrale était d'explorer au mieux les conformations possibles de chaque système étudié afin d'obtenir des modèles structuraux de haute qualité des qui pourraient servir à la conception de petites molécules inhibitrices ou stabilisantes des quadruplexes de guanine (GQ) et de la protéine M^{pro} du SARS-CoV-2. Nous avons choisi d'étudier les GQ formés au niveau du promoteur du proto-oncogène c-kit 1 et au niveau de la région LTR-III du VIH-1 et la protéase principale, M^{pro}, du SARS-CoV-2 comme complexes moléculaires à étudier du fait de leur implication dans des maladies graves : le cancer, le SIDA et le COVID-19.
En effet, le proto-oncogène c-kit 1 code pour le récepteur transmembranaire CD117/c-kit intervenant dans la croissance tumorale et la progression de divers types de cancer, notamment le cancer gastro-intestinal. La présence stabilisée des GQ au niveau du promoteur de l'oncogène c-kit 1 jouera le rôle d'un répresseur de la transcription. De même, au niveau de la région LTR-III du VIH-1, les GQ agissent comme des éléments répresseurs de l'initiation de la transcription virale.
Comme les quadruplexes de guanine sont des structures d'ADN non canonique à quatre brins qui sont sur-représentées dans les régions promotrices des oncogènes et dans le génome viral et proviral, ils sont considérés comme des cibles thérapeutiques émergentes, car la répression transcriptionnelle des gènes et des séquences d'ADN par la stabilisation de ces structures pourrait être une nouvelle stratégie thérapeutique.
Par ailleurs, la protéase principale, M^{pro}, joue un rôle essentiel dans le cycle de vie et la réplication du SARS-CoV-2. Alors que cette protéase n'a pas d'homologue humain, il s'agit d'une cible anti-virale idéale.

4.2 Spécificité et résultats de notre approche

Grâce aux avancées technologiques développées dans notre Laboratoire, notamment en réponse au besoin urgent de recherche sur le COVID-19, tant au niveau informatique qu'au niveau de l'échantillonnage, nous avons démontré la possibilité d'effectuer plusieurs multi-μs de simulations de MD en peu de temps de calcul sur de gros systèmes en solvant explicite en utilisant des champs de force polarisable, tels que AMOEBA, permettant une haute précision dans la prédiction des conformations possibles ayant une énergie stable des systèmes complexes. Nous avons pu étudier en détail la dynamique moléculaire des quadruplexes de guanine et de la protéase principale du SARS-CoV-2, M^{pro}.

En raison de la complexité inhérente des protéines du SARS-CoV-2, il est important d'effectuer de telles simulations à plus haute résolution car elles pourraient fournir des informations supplémentaires sur la dynamique structurelle des constituants du virus aux communautés de recherche COVID-19

expérimentales et informatiques. Au delà du contexte du COVID-19, il est intéressent de rappeler que la M^{pro} est un protéase à cystéine dont le site actif est formé par des résidus très conservés au sein de la famille des coronavirus en raison de leur importance fonctionnelle. Une bonne exploration des conformations possibles de la M^{pro} aidera non pas seulement à combattre le SARS-CoV-2 mais tous la virus de la même famille.

Ainsi, nous avons proposé une stratégie d'échantillonnage adaptatif entièrement non supervisée qui peut être utilisée sur tout type de ressources de calcul. Cette stratégie automatisée permet des simulations de production qui bénéficient des progrès de la super-informatique et de nos récentes améliorations au niveau de logiciel massivement parallèle Tinker-HP HPC, qui peuvent désormais gérer efficacement les grands ordinateurs pétascale accélérés par GPU. Afin d'extraire de nouvelles informations de ce type de simulation, nous avons également fourni les étapes nécessaires pour supprimer le biais (i.e. pour re-pondérer) les données obtenues afin de collecter des caractéristiques de dynamique structurelle utiles et précises.

La qualité de nos simulations nous a permis de bien décrire les comportements globaux et locaux de la M^{pro}. Une attention spéciale a été consacrée à l'élucidation des changements conformationnels au niveau du site catalytique et au niveau de l'interface de dimérisation de la protéine.

Plus de 38 μs de simulation MD de l'enzyme M^{pro} dans son état *apo* (sans ligand) ont été produites à l'aide du champ de force polarisable AMOEBA.

Les résultats ont ensuite été comparés aux données des simulations à grande échelle déjà disponibles dans la littérature récente. Il a été démontré que les résultats issus du PFF de nouvelle génération capturent la plupart des caractéristiques de la dynamique structurelle discutées dans la littérature expérimentale, confirmant que la M^{pro} est probablement dans une conformation peu active dans son état *apo* dans des conditions de pH physiologique. Cependant, les simulations ont détecté des caractéristiques d'activité partielle dans l'un des protomères liés à un trou d'oxyanion plus structuré. Ceci est cohérent avec l'activité asymétrique protomérique observée dans l'état *holo* où un seul protomère est actif, une caractéristique similaire qui a également été observée dans le SARS-CoV-1. Cette asymétrie peut être liée à plusieurs marqueurs structurels ainsi qu'aux volumes totaux des protomères. Le site actif s'avère très flexible à température ambiante, en accord avec des découvertes expérimentales récentes. Globalement, l'état *apo* de la M^{pro} apparaît clairement moins organisé que l'état *holo* en accord avec les résultats expérimentaux discutés par Zhou *et al* [243]. Une seconde simulation, incluant la protonation du résidu **His172** pour simuler le système dans des conditions à pH = 6, a été réalisée. Cette simulation tend à confirmer le rôle de la protonation dans l'effondrement de la poche S1 à plus bas pH. Dans ces conditions, l'asymétrie des protomères AMOEBA est conservée, bien que les protomères aient tendance à être notablement déstructurés. Les simulations AMOEBA ont également capturé la haute flexibilité de la région C-terminale discutée dans la littérature. Cette flexibilité augmente à un pH plus bas et tend à moduler davantage l'activité de l'état *apo* liée à l'effondrement de la poche S1. Des différences frappantes ont été observées concernant les modèles de

solvatation autour des résidus clés **His41** et **His163** entre AMOEBA et les n–PFF . Dans l'ensemble, le plus petit nombre de molécule d'eau en AMOEBA détecté autour des histidines est plus conforme aux données expérimentales. Si la présence d'une molécule d'eau structurelle autour de **His41** est probable à tous les pH, l'existence d'une molécule d'eau autour de **His163** a tendance à être statistiquement plus possible à pH = 6. Ces résultats peuvent s'expliquer par la capacité des molécules d'eau structurelles à présenter en AMOEBA un moment dipolaire moyen supérieur à celui de l'eau en vrac et d'explorer une plus large gamme de dipôles par rapport aux n–PFF . Les molécules d'eau structurelles autour des histidines affecteront clairement la conception rationnelle des médicaments.

Par ailleurs, le réseau complexe de liaisons hydrogène au niveau de l'interface de dimérisation du dimère de la M^{pro}, s'est avéré impliqué dans un réseau dipolaire dynamique du solvant, où l'eau semble être un acteur clé dans la dynamique structurelle globale de cette interface. En plus nous avons pu démontrer l'existence d'une interaction allostérique entre la région de dimérisation et le site catalytique de chaque monomère de la M^{pro}. Cette allostérie peut être modulée par le flux et la polarisation des molécules d'eau structurales qui existent au niveau de l'interface de dimérisation. Ces deux régions au niveau de la protéase font l'objet d'une attention spéciale puisqu'elles modulent l'activité fonctionnelle de la M^{pro} et le blocage de l'une d'elle, ou les deux, pourra affecter son activité catalytique. Inhiber l'activité de la M^{pro} aura comme résultat direct le blocage du cycle viral du SARS-CoV-2 au sein des cellules infectées et empêchera la production de nouveaux virions.

La partie de notre travail portant sur l'étude des quadruplexes de guanine nous a permis de mieux comprendre le comportement de ces structures non-canoniques qui peuvent exister au niveau de l'ADN et de l'ARN. Nous avons considéré deux structures de GQ intéressantes : la première au niveau du promoteur du proto-oncogène c-kit 1 et la deuxième au niveau de la région LTR-III du VIH-1. L'évaluation de la stabilité de ces deux structures présentant des arrangements en GQ est très important dans la cadre du développement des stratégies contre le cancer (notamment gastro-intestinal) et contre le SIDA. En effet la formation d'un GQ au niveau du promoteur du proto-oncogène c-kit 1 exerce un effet inhibiteur sur l'ARN polymérase responsable de la transcription du proto-oncogène c-kit 1 en ARN messager qui sera ensuite traduit en CD117, dont la sur-expression est responsable de l'activation des voies de signalisation impliquées dans la progression du cancer. Alors que la formation d'un GQ au niveau de la région LTR-III du VIH-1 empêchera l'ARN polymérase de transcrire l'ADN viral integré dans le génome cellulaire, étape nécessaire à la continuité du cycle viral à l'intérieur des cellules infectée. Bien que la présence des GQ dans l'ADN humain ou viral est transitoire et peut se former et se déformer par les protéines cellulaires, une stratégie thérapeutique consiste à stabiliser davantage ces structures non-canoniques de l'ADN via des petites molécules stabilisantes.

Nous avons pu démontrer une forte stabilité des tétramères de guanine et de la grande boucle formée de 5 nucléotides au sein de la structure GQ de l'ADN. Cependant, les nucléotides libres qui forment des boucles mono- et dinucléotidiques sont hautement dynamiques bien que certaines puissent

former des interactions de type Watson-Crick entre elles. Tout comme dans le cas de la protéase principale de SARS-CoV-2, nous avons pu démontrer l'importance de molécules d'eau polarisables dans la zone confinée de la structure pour le maintien de structures stables de tétramères de guanine en l'absence des ions K^+.

Nous avons détecté l'existence de deux poches cryptiques stables au sein de deux systèmes différents de GQ étudiés : le premier au niveau du promoteur du proto-oncogène c-kit 1 et le deuxième au niveau de la région LTR-III du VIH-1. Les poches identifiées au niveau du coeur du GQ ont été démontrées être bien stables en comparant avec la poche identifiée au niveau du domaine en épingle à cheveux du VIH-1. L'évaluation de la stabilité a été faite suite à la caractérisation de certaines distances inter-nucléotides au sein de chacune des poches détectées. Ces poches cryptiques offrent une opportunité thérapeutique intéressante.

L'utilisation de champs de force polarisables pourrait être critique dans l'évaluation des énergies libres de liaison d'éventuels médicaments candidats en compétition avec l'eau pour interagir avec la M^{pro} du SARS-CoV-2 et avec les GQ au niveau du promoteur de l'oncogène c-kit 1 et au niveau de la région LTR-III du VIH-1. En pratique, les systèmes biologiques considérés dans cette Thèse ont tendance à être difficile pour les approches de mécanique moléculaire. En effet, ils englobent toutes sortes d'interactions faibles. Par conséquent, il n'est pas surprenant que toutes les caractéristiques décrites expérimentalement trouvées dans les simulations AMOEBA n'aient pas nécessairement été trouvées avec les simulations non polarisables. De tels systèmes ont tendance à nécessiter à la fois un champ de force précis et une stratégie d'échantillonnage étendue car il est évident que quelques ns de PFF MD à eux seuls ne fourniraient pas d'informations sur un système où la convergence statistique est difficile en raison de sa plasticité. Ces résultats fournissent une première validation directe de la stabilité du champ de force polarisable AMOEBA et démontrent clairement son applicabilité à de longues échelles de temps. Outre la corrélation avec les données expérimentales, nos résultats montrent également que notre approche d'échantillonnage adaptatif couplée à AMOEBA a conduit à mieux échantillonner les conformations possibles du site actif et de la région de dimérisation de la M^{pro} du SARS-CoV-2, du coeur des GQ au niveau de l'ADN, notamment au niveau du promoteur du proto-oncogène c-kit 1 et au niveau de la région LTR-III du VIH-1, ainsi qu'au niveau des boucles formées par les bases nucléotidiques entre les guanines formant le coeur des structures GQ et finalement à des poches cryptiques potentielles supplémentaires dans chacun des systèmes étudiés. Comme les états *apo* (sans ligand) des systèmes biologiques étudiés se sont avérés être des structures pertinentes à température ambiante, les nouvelles informations fournies pourraient être utiles pour la conception de médicaments. Nos données de simulation sont entièrement accessibles au grand public. En effet, concentrer la puissance de calcul du GPU sur un état *apo* est utile pour "exploiter" les conformations afin d'obtenir un ensemble précis et plus convergé statistiquement. Des conformations du site de liaison pourraient être ainsi sélectionnées pour y docker un ligand. Les nouvelles informations structurelles fournies ici pourraient aider à concevoir de nouveaux médicaments ou à réutiliser des médicaments

existants. Ces données pourraient également être importantes pour comprendre la réactivité chimique au niveau atomique via des simulations hybrides QM/MM.

Enfin, grâce à notre approche, nos simulations de MD adaptative en AMOEBA se sont avérées simultanément compétitives sur le plan informatique et conformes aux données expérimentales disponibles. À l'aide de 100 cartes GPU, nous montrons qu'un délai de résolution acceptable et compétitif pourrait être atteint car nos résultats en "μs" ont été obtenus en quelques jours sur un super-calculateur académique (et polyvalent). Il est également important de noter que Tinker-HP peut également produire une simulation plus rapide d'un ordre de grandeur à l'aide de n–PFF à l'aide de GPU. Étant donné que les simulations n–PFF sont également d'un grand intérêt, capturant de nombreux aspects expérimentaux, notre stratégie à deux niveaux (n-PFF + PFF) est confirmée. En effet, une configuration optimale consiste à produire d'abord une longue simulation adaptative non polarisable qui peut être encore affinée avec des potentiels polarisables au sein d'itérations adaptatives supplémentaires.

4.3 Perspectives sur la modélisation des GQ

Ayant donné une attention particulière à l'analyse des trajectoires de MD des GQ au niveau du promoteur du proto-oncogène c-kit 1, les trajectoires cMD du GQ au niveau de la région du LTR-III du VIH-1 nous ont servi comme application directe pour la recherche des poches cryptiques après validation de la stabilité des tétramères de guanine par RMSD et RMSF. Il convient ainsi, dans un futur proche, d'étendre les simulations adaptative multi-μs utilisant AMOEBA. Il a été montré qu'il est intéressant ensuite de réaliser une clusterisation de toutes les structures générées en se basant sur tICA, et d'étudier le comportement locale de chacune des bases nucléotidiques en évaluant leurs angles de torsion χ et γ afin de détecter les changements quant au retournement des bases et des changements conformationnelles *syn*/*anti* et de mesurer certaines distances clés qui sont cruciales au maintien de la stabilité de la structure globale.
Les résultats de ces études pourront aider à mieux imaginer la structure d'un futur ligand à visée thérapeutique et à imaginer un potentiel mécanisme d'entrée dans le sillon d'ADN au niveau des poches cryptiques identifiées.

4.4 Conception des ligands à visée thérapeutique

En perspective, ces travaux laisse envisager de pouvoir concevoir de petites molécules thérapeutiques spécifiques de la M^{pro} du SARS-CoV-2 et des quadruplexes de guanine. La protéine M^{pro} peut être ciblée aussi bien au niveau de son site catalytique qu'au niveau de l'interface de dimérisation. Le ciblage du site catalytique bloquera directement l'activité de clivage de la protéase alors que le ciblage de l'interface de dimérisation empêchera la dimérisation de la M^{pro} qui perdra ses fonctions, car la M^{pro} n'agit que sous sa forme dimérique. Il est important ensuite de calculer l'énergie libre

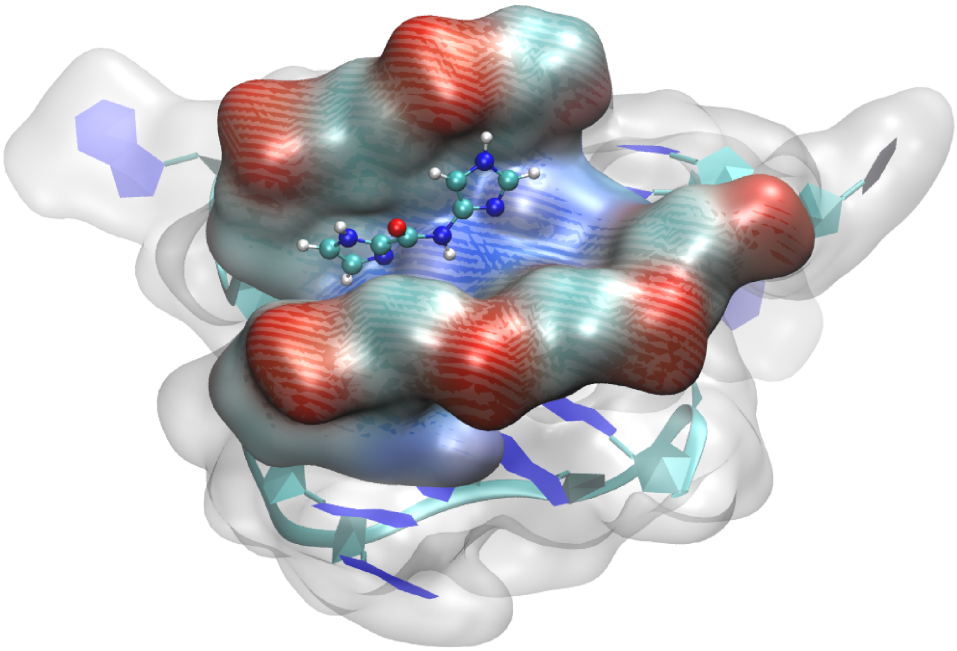
d'interaction entre la protéase principale et les molécules candidates afin de déterminer la robustesse des interactions.

Quant au quadruplexes de guanines, la modélisation de petites molécules stabilisantes de la structure de l'ADN en quadruplexe inhibera l'activité transcriptionnelle de l'ARN polymérase, étape cruciale avant la production des ARNm correspondants.

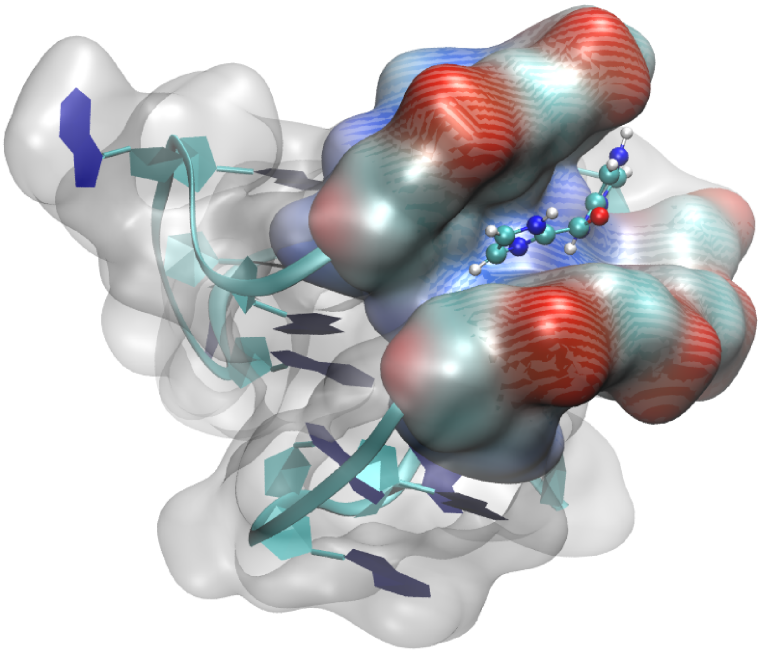
L'importance émergente des GQ en tant que cibles pour des molécules anti-cancéreuses et antivirales ouvre la possibilité d'explorer de nouveaux ligands qui leur sont spécifiques. Il convient dans une future perspective d'optimiser les petits ligands stabilisants et d'estimer la robustesse de leur interaction au niveau de P1 et de P2 du promoteur de c-kit 1 et au niveau de P2 du LTR-III du VIH-1. Considérant le teneur élevée en GQ dans les cellules humaines, l'un des principaux défis est d'atteindre la sélectivité envers les GQ ciblés. Étant donné que la majorité des ligands spécifiques des GQ testés jusqu'à présent présentent des fonctionnalités destinées à cibler les tétrades de guanine principalement par interaction de type empilement π [195, 16, 151, 99, 40, 100, 222, 111, 17, 104] et une sélectivité élevée, nos découvertes ouvrent de nouvelles perspectives sur la possibilité de discriminer entre les différentes conformations des GQ. Une future approche peut donc être orientée vers le développement de petites molécules avec des caractéristiques structurelles compatibles avec les séquences et les arrangements de boucles uniques. Cette stratégie envers les structures du promoteur de l'oncogène c-kit et la région LTR-III du VIH-1 présentées dans ce travail peut fournir de nouveaux agents anti-cancéreux et anti-VIH-1 sélectifs pour intervenir dans le mécanisme d'action intra-cellulaire en stabilisant les acides nucléiques présentant des motifs GQ.

Plusieurs petites molécules font l'objet de prototypes pour la conception des ligands spécifiques contre les GQ. Le Chelerythrine (CHE) peut être ainsi considérée comme une molécule prometteuse qui pourrait servir de base candidats-médicaments ciblant spécifiquement les GQ à des fins à la fois antitumorales et antivirales [9]. Des polyamides contenant du N-méthylimidazole et des acides aminés N-méthylpyrrole sont des ligands synthétiques qui complexent le petit sillon de l'ADN B et dont certains ont une affinité et une spécificité pour l'ADN comparables à celles de nombreuses protéines naturelles de liaison à l'ADN [185]. Elles pourront servir de base de départ pour la conception d'inhibiteurs destinés à interagir dans certains 'sillons' creusés dans les quadruplexes d'ADN et que la dynamique moléculaire a pu identifier (Figure 4.1).

L'importance d'élucider le comportement des tétrades de guanine empilées ne se limite pas à l'étude des GQ au niveau du proto-oncogène c-kit 1 ou au niveau de la région LTR-III du VIH. En effet, les GQ sont rencontrés au niveau des télomères des chromosomes et jouent un rôle clé contre le vieillissement de la cellule. Par ailleurs, les GQ peuvent être considérés dans les applications nano- et biotechnologiques ; par exemple nous citons le système de commutation pour la formation de nano-fils de guanine (G-wire) par des signaux externes [216]. De tels systèmes sont prometteurs pour réguler la formation de nano-fils pour diverses applications, y compris les circuits électroniques destinés aux utilisations de la nano-biotechnologie.



a)



b)

Figure 4.1: Représentation sur VMD d'une optimisation préliminaire du complexe du GQ de c-kit 1 et d'un premier ligand destiné à cibler la poche cryptique P1. a) Vue de face et b) vue de coté.

4.5 Perspectives techniques : raffinements des champs de force polarisables sur la base des calculs SAPT-DFT

Parmi les limitations de la MD, les simulations de très grands complexes moléculaires et / ou à très longue échelle de temps restent encore hors d'atteinte dans une diversité de domaines. Ceux-ci englobent la conception de médicaments, le repliement des protéines et la science des matériaux. Une alternative porte sur les approches de la Mécanique Quantique/Mécanique Moléculaire (QM/MM), dans lesquelles le noyau du site de reconnaissance moléculaire est traité de manière quantique *ab initio* tandis que la périphérie est traitée par la mécanique moléculaire classique [153]. Toutefois de telles approches peuvent poser des problèmes en traitant, par exemple, des sites de reconnaissance élargis, de sorte que les effets de polarisation à longue portée pourraient brouiller la distinction entre les zones de la QM et la MM. Cela peut se produire avec les complexes d'inhibiteurs des métalloenzymes à zinc, dans lesquels des réseaux de molécules d'eau polarisables peuvent relier l'inhibiteur, dans le site de reconnaissance, à des sites plus éloignés sur les sites N- et / ou C-terminaux [155]. En revanche, les potentiels de la mécanique moléculaire / dynamique moléculaire, MM / MD sont pleinement capables de gérer de très grands complexes moléculaires sur des échelles de temps de l'ordre de la μs. Afin d'améliorer de telles simulations, il convient d'optimiser un potentiel de mécanique moléculaire fondé sur la chimie quantique, formulé et calibré de manière à reproduire séparément chacune des contributions de la chimie quantique.

Avec les nouvelles approches de chimie quantique développées ces dernières années, des données de référence très précises pour de nombreuses propriétés atomiques et moléculaires deviennent disponibles pour la paramétrisation et la validation de ces méthodes. L'étude des interactions non covalentes est un domaine de recherche particulièrement important dans de nombreux domaines de la chimie, de la biologie et de la science des matériaux. Ces interactions sont fortement impactées par la corrélation électronique. Il est donc nécessaire d'utiliser des méthodes avancées pour en rendre compte, et de recourir a des bases étendues de Gaussiennes.

Un sous-ensemble de potentiels MM / MD "polarisables" a d'abord raffiné la représentation de la contribution électrostatique au premier ordre, en ayant recours à des multipôles distribués dérivés de calculs *ab initio* sur les fragments. C'est particulièrement le cas de la méthode SIBFA (Sum of Interactions Between Fragments Ab initio computed) [156]. La calibration des paramètres de SIBFA est réalisée en minimisant la différence des propriétés désirées des calculs de la MM par rapport aux calculs de la QC. Une possibilité très prometteuse consiste à simuler les interactions ligand-molécule à grande échelle en se basant sur la mécanique et la dynamique moléculaire anisotrope et polarisable (APMM / APMD) comme alternative aux calculs QM/MM. Le potentiel SIBFA dans son entièreté est en cours de portage sur Tinker-HP [114]. Comme ces simulations sont fondées sur la chimie quantique qui décrit le comportement électronique des atomes et les énergies d'interaction intermoléculaire ΔE, avec grande précision, plus les ΔE et ses contributions *ab initio* calculées au niveau des sites de

reconnaissance moléculaire sont proches de celles calculées par les simulations APMM / APMD, plus ces derniers sont fiableS.

Dans les sites d'interaction moléculaire, la liaison du (des) cation(s) se fait avec certains ligands tels que: l'imidazole, qui fait partie de la chaîne latérale de l'histidine ; le formate, des chaînes latérales de l'aspartate et du glutamate ; le methanethiolate de celle du cystéinate et le formamide, de celles de l'asparagine et de la glutamine, et qui est aussi partie intégrante de la chaîne principale. Une interaction hors du plan dans le cas des cations divalents, pourrait dans une certaine mesure améliorer les interactions simultanées avec tous les ligands qui ont une liberté limitée de torsion du fait de leur ancrage dans la chaîne principale de la protéine. Ainsi, il s'avère important de tenir compte des dépendances de ΔE non pas seulement dans le plan des ligands mais aussi hors du plan.

Dans un travail antérieur effectué par El Khoury *et coll.* [131] des raffinements de SIBFA au niveau HF non corrélé, ont porté sur les deux contributions énergétiques à courte portée, la répulsion et le transfert de charge. Ces raffinements consistaient à étaler les paires libres sp et sp2. Ceci a été inspiré de la fonction de localisation électronique (Electron Localization Function, ELF) qui mesure la possibilité de trouver un électron dans le voisinage d'un électron de référence à un point donné. D'un point de vue physique, elle augmente la localisation spatiale d'un électron dans un bassin de localisation en fournissant une cartographie pour la densité de présence d'une paire électronique pour un système multiélectronique. Dans SIBFA ce concept se traduit pour les ligands tels que le formamide et l'imidazole par la prise en compte d'un total de huit et de onze paires libres après étalement des doublets sp2 de l'oxygène du formamide et du doublet sp de l'azote de l'imidazole avec en plus les deux paires libres supplémentaires de type π, d'où un total de dix et de treize paires libres, respectivement pour ces deux ligands. Pour le formate et le méthanethiolate le total des paires libres sera de quatorze et de six, respectivement. Sans cette stratégie d'étalement des paires libres l'énergie de répulsion pour les variations hors du plan calculée par SIBFA diminue trop vite par rapport à la contribution d'échange-répulsion de la chimie quantique [131]. Nous avons cherché a raffiner la méthode SIBFA sur la base des calculs *ab initio* corrélés ; "Symmetry Adapted Perturbation Theory" (SAPT), et sur la décomposition de l'énergie d'interaction de SAPT, dans le cadre de la continuité des enrichissements et des perfectionnements de cette méthode.

4.5.1 Le champ de force polarisable: SIBFA (Sum of Interactions Between Fragments *Ab initio* computed)

Dans la procédure SIBFA, l'énergie d'interaction intermoléculaire totale entre molécules ou fragments moléculaires est calculée comme une somme de cinq contributions:

$$\Delta E_{tot} = E_{MTP} + E_{rep} + E_{pol} + E_{ct} + E_{disp} \tag{4.1}$$

On désigne par ordre, les contributions : multipolaire électrostatique, répulsion à courte portée, polarisation, transfert de charge et dispersion.

- E_{MTP} est calculé comme une somme d'interactions multipôles-multipôles, englobant six termes de monopole-monopole jusqu'à quadrupôle-quadrupôle. Les multipôles sont situés sur les atomes et les points médians des liaisons chimiques et sont dérivés des orbitales moléculaires *ab initio* QC du fragment et sont pré-calculés pour chaque molécule individuellement. Ils étaient à l'origine issus d'une procédure due à Vigné-Maeder et Claverie [218]. Les multipôles distribués ont été dérivés, pour de grands ensembles de base, en utilisant la méthode d'analyse multipolaire généralisée (GDMA) par Stone [205], et peuvent également être dérivés de la méthode d'atomes sphériques itérés "space-Iterated Stockholder Atom" (ISA). Les améliorations subséquentes à E_{MTP} ont consisté à inclure un terme explicite de "pénétration", Epen, qui dépend du recouvrement: il traduit le fait qu'à de courtes distances intermoléculaires, il y a un moindre blindage de la charge nucléaire d'un atome donné par sa propre densité électronique, en raison de sa "pénétration" par la densité correspondante de l'atome interagissant, et inversement. Ceci résulte en une augmentation réelle de l'attraction électron-noyau de la paire en interaction [97].

  La différence entre GDMA et ISA :
  GDMA : Le programme GDMA effectue une analyse multipolaire distribuée des fonctions d'onde calculées par le logiciel Gaussian G09. Le résultat est un ensemble de moments multipolaires à des sites définis par l'utilisateur (généralement aux positions des noyaux atomiques) qui, avec une fonction d'onde moléculaire, fournissent une description détaillée du potentiel électrostatique généré par les atomes de la molécule.

  ISA : La méthode ISA (atome sphérique itéré) est une procédure de séparation de la densité électronique moléculaire en composants atomiques [10]. Des travaux antérieurs ont montré que les moments multipolaires calculés en utilisant la méthode ISA seraient préférables à ceux obtenus à partir de l'analyse multipolaire distribuée de Stone [204] ; présentant de meilleures propriétés de convergence notamment dans le cas des molécules conjuguées. Chaque atome "a" est associé à une fonction de poids à symétrie sphérique, et la densité électronique totale est répartie de façon itérative entre les atomes proportionnellement à la fonction de poids de chaque point.

- L'autre contribution de nature principalement électrostatique est la polarisation. Elle traduit le gain en énergie lors du réarrangement de la distribution électronique d'un fragment moléculaire donné causé par le champ électrostatique généré sur lui par tous les autres fragments du complexe intermoléculaire. E_{pol} sur tout centre "polarisable" est ainsi fonction du carré du champ électrostatique qu'il subit et de sa polarisabilité. Le champ est calculé avec les mêmes multipôles distribués que E_{MTP} et est écranté par une fonction gaussienne, S, de la distance entre ce centre et chacun des centres des molécules avec lesquelles il interagit. Cette distance est divisée par

la moyenne des rayons effectifs de la paire de centres en interaction. Ainsi, pour tout fragment moléculaire, nous dérivons une fois pour toutes, et de manière consistante, à la fois les multipôles distribués et les polarisabilités. Ceux-ci sont stockés dans la bibliothèque SIBFA sous la forme d'un fichier, avec les informations sur la géométrie interne, les connectivités, et les types d'atomes. Chaque fichier spécifique au fragment peut être extrait et concaténé avec d'autres chaque fois que le fragment est nécessaire pour assembler une grande molécule flexible, ou une biomolécule, telle qu'une protéine ou un acide nucléique.

SIBFA comprend deux contributions qui dépendent du recouvrement :

- E_{rep} au premier ordre est formulé sous la forme d'une somme des interactions de liaison-liaison, de liaison-paire libre et de paire libre-paire libre, où la répulsion-échange est proportionnelle au carré du recouvrement intermoléculaire entre les orbitales localisées [150]. Pour chaque paire d'orbitales en interaction, elle peut être formulée comme S^2 / R^n, S indiquant leur recouvrement et R la distance entre leurs centroïdes, n prenant les valeurs de 1, 2 et éventuellement au-delà. Les orbitales moléculaires sont optimisées comme étant localisées sur les liaisons et les doublets. La présence explicite des doublets libres étalés confère la directionalité à E_{rep}.

- E_{ct} est fonction du recouvrement entre les représentations des orbitales localisées du fragment donneur d'électrons et des orbitales virtuelles du fragment accepteur d'électrons. La contribution E_{ct} du second ordre est formulée comme l'intégrale d'une densité de transition de recouvrement convoluée avec le potentiel électrostatique qu'elle subit.

- E_{disp} est calculée comme une somme de termes atome-atome avec les dépendances $1/R^6$, $1/R^8$ et $1/R^{10}$. Elle est augmentée d'un terme explicite "échange-dispersion" ; $E_{xch-disp}$, et de contributions des paires libres en leurs positions centroïdes.

Comme cela avait été souligné dans des articles de synthèse précédents [96], une méthodologie de mécanique moléculaire visant à reproduire les résultats CQ, devrait avoir les caractéristiques suivantes : séparabilité, anisotropie, non-additivité et transférabilité. Chacune est détaillée dans les articles de synthèse du potentiel SIBFA.

4.5.2 Construction d'une base de données pour les interactions intermoléculaires sur la base de calculs perturbationnels SAPT(DFT)

Les calculs perturbationnels SAPT(DFT) permettent un découpage de l'énergie d'interaction entre deux molécules en six contributions bien définies: au premier ordre, E1, Coulombienne, EC et répulsion d'échange EX, au second ordre, induction, Eind, comportant séparément la polarisation, Epol, et le transfert de charge, ECT, et dispersion, Edisp, comportant une contribution sans échange, Edi, et une contribution répulsive dite 'd'échange-dispersion', Xdi.

Ce découpage devrait permettre de calibrer chacune des six contributions correspondantes de la méthode SIBFA. Ainsi nous pourrions appliquer SAPT a l'étude de complexes des chaînes latérales et des bases nucléiques avec des 'sondes' en vue de cette calibration. Ceci constituerait une première étape dans la construction d'un nouveau 'champ de force' SIBFA sur la base de calculs SAPT avec des bases étendues de Gaussiennes: en l'occurrence, la calibration et la validation des interactions intermoléculaires des chaînes latérales des protéines et des bases de l'ADN. La calibration de leurs chaînes principales (N-methylformamide des protéines, groupements phosphodiester et deoxyribose de l'ADN) pourra être considérée par la suite.

Nous avons ainsi considéré les fragments suivants: formate, formamide, imidazole, méthanethiolate, acétamide, acétate, benzène, méthanol, méthylimidazole, méthylindole, méthylphénol et N-Méthylformamide, et les quatre bases nucléiques de l'ADN: guanine, adénine, thymine et cytosine.

Chaque molécule été 'sondée' par une molécule dipolaire, l'eau (Figure 6.1 dans les annexes), et par un dication à couche fermée, Zn^{2+} (Figure 6.2 dans les annexes) dans une grande diversité de positions autour de son volume. Nous avons ainsi considéré d'abord des complexes dans le plan en faisant varier la distance d'approche a chaque atome polaire, et ensuite, pour des distances a l'équilibre et proches de l'équilibre, l'angle Θ. Pour une distance et un angle Θ donnés, nous avons ensuite fait varier l'angle Φ (variations sur un cône). Nous avons, pour les molécules conjuguées, également sondé la verticale au plan a partir de chacun de ses atomes lourds.

Le choix de Zn^{2+} comme sonde dicationique plutôt que d'un autre cation tel que Mg^{2+} et Ca^{2+} permet 'd'exalter' la réponse de la molécule sondée en termes de transfert de charge et de dispersion, ces deux contributions ayant des poids beaucoup plus faibles avec ces autres cations. Le choix de l'eau comme sonde dipolaire est motivé par sa présence ubiquitaire.

Nous avons également considéré l'extension à une grande base de données des énergies d'interaction [95] pour 66 complexes moléculaires, dans leurs géométries d'équilibre de référence (Figure 6.13 dans les annexes). L'ensemble de ces données avait été conçu en vue de couvrir les modes les plus courants d'interactions non covalentes entre biomolécules, en conservant une représentation équilibrée des contributions individuelles. Nous avons de ce fait complété les résultats de référence CCSD(T) publiés par Rezac *et coll.* par les calculs SAPT(DFT) permettant de connaître les poids des six contributions au sein de DEtot. Celles-ci sont données en annexes.

Théorie des Perturbations Adaptée à la Symétrie (SAPT)

La Théorie des perturbations adaptée à la symétrie (SAPT: Symmetry Adapted Perturbation Theory) [30, 207] part de deux molécules non perturbées et traite leur énergie d'interaction comme une somme de contributions au premier et au deuxième ordre résultant de leurs perturbations mutuelles.

Toutefois dans sa formulation originale, SAPT ne sépare pas, au sein de la contribution d'induction au deuxième ordre,la contribution de transfert de charge de la contribution de polarisation.

Des méthodes ont été récemment proposées afin d'extraire cette contribution dans le cadre des

approches SAPT.

Il a d'abord été proposé que E_{ct} pourrait se calculer comme la différence entre d'une part l'énergie d'induction calculée avec une base centrée sur le dimère, (ie sur le supersystème), et d'autre part, celle calculée avec des bases centrées sur les monomères individuels [203, 202]. Cependant, avec cette définition E_{ct} disparaît lorsque la base des monomères devient complète. Ceci a conduit A. J. Misquitta à proposer un autre découpage [146] [148]. Il consiste en une séparation du potentiel Coulombien électron-noyau en deux parties: une partie singulière avec un terme en 1/R qui diverge lorsque R diminue, et une partie non singulière, régularisée', qui s'amortit lorsque R diminue. Un calcul de l'induction avec le seul potentiel régularisé permet d'avoir alors la valeur de la seule contribution de polarisation. E_{ct} s'obtient alors comme la différence des valeurs de l'induction calculée avec le potentiel Coulombien complet et ce potentiel régularisé.

La contribution E_{ct} au second ordre de SAPT est alors définie comme

$$E_{CT}{}^{(2)} = [E_{ind}{}^{(2)} - E_{ind}{}^{(2)}(Reg)] + [E_{ind-exch}{}^{(2)} - E_{ind-exch}{}^{(2)}(Reg)] \tag{4.2}$$

L'énergie E_{ct} totale du second ordre est la somme des contributions A $\rightarrow$ B et B $\rightarrow$ A.

Dans le cadre des calculs SAPT(DFT), un terme "δHF" est défini comme la différence entre l'énergie d'interaction Hartree-Fock corrigée de la BSSE et l'énergie SAPT de second ordre sans les termes de dispersion et d'échange-dispersion. Il est utilisé pour inclure l'effet d'induction d'ordre élevé manquant. Ce terme est rajouté à la contribution E_{ct} formulée ci-dessus.

Chaque calcul se fait en deux étapes: i) un calcul SAPT utilisant avec le logiciel Psi4 [181] et ii) un calcul DFT avec le logiciel CamCASP [147]. Psi4 permet de calculer les contributions E_{MTP}, E_{rep} et la valeur de δHF. CamCasp permet de calculer E_{pol} et E_{ct} et E_{disp}.

Il est également nécessaire dans ce cadre d'introduire la valeur du potentiel d'ionisation (IP) de chaque molécule ou ion. Cette valeur correspond a la valeur expérimentale si elle est connue, ou a défaut résulte d'un calcul préalable de chimie quantique, comme la différence entre l'énergie de la molécule considérée et celle résultant de la perte d'un électron:

$$IP_{\text{formamide}} = E_{\text{formamide}} - E_{\text{formamide}^+} \tag{4.3}$$

4.5.3 Raffinements supplémentaires du potentiel polarisable SIBFA sur la base des calculs SAPT(DFT)

Pour les complexes de la guanine, du formamide, de l'imidazole, du formate et du méthanethiolate avec Zn^{2+} et H_2O, chacune des six contributions de DEtot(SIBFA) a été recalibrée pour reproduire la contribution correspondante SAPT(DFT) pour l'ensemble des variations radiales et angulaires dans et hors du plan. Comme dans des travaux récents du Laboratoire [135, 74] cette calibration a été réalisée avec un minimiseur de type quasi-Newton intégré dans le code SIBFA (code 'VA10A':

http://www.hsl.rl.ac.uk [1]). Pour chacune des contributions, il minimise par moindres carrés l'erreur relative SIBFA-SAPT(DFT). Ceci est fait en optimisant les paramètres de cette contribution: ce sont essentiellement les rayons effectifs du type d'atomes, les variables internes définissant les positions des doublets d'électrons libres et les incréments/décréments des rayons effectifs des hétéroatomes le long des directions des doublets d'électrons libres.

Le graphes présentés dans les Figures 6.3 et 6.4 dans les annexes portent sur les complexes de la guanine avec l'eau et avec Zn^{2+}. Ils représentent les évolutions comparées des six contributions SIBFA, en rouge, et SAPT(DFT), en bleu. 129 complexes guanine-H_2O et 116 complexes guanine-Zn^{2+} ont été considérés. Les graphes correspondants de ces évolutions dans les complexes du formamide, de l'imidazole, du formate et du methanethiolate sont reportés en annexe.

Nous avons choisi de privilégier la guanine, en raison de son rôle central dans la formation de quadruples hélices de l'ADN, et dans une perspective prochaine de simulations des ses complexes avec des potentiels polarisables d'un degré augmenté de précision.

Les contributions individuelles SIBFA, pour les approches des deux sondes dans, et hors du plan de la guanine, reproduisent de très près leur contrepartie *ab initio*. Certains écarts au niveau de E2 peuvent apparaître dans les complexes de la guanine avec l'eau, ou E2(SIBFA) est sousestimé, mais ils se produisent pour des complexes ayant un DEtot positif, donc dans des régions qui ne seront pas peuplées dans un calcul de dynamique moléculaire.

Dans le cas de Zn^{2+}, DEtot(SIBFA) est sous-estimée de 6% par rapport à DEtot(SAPT) dans le complexe d'énergie la plus basse: la distance entre Zn^{2+} et N7 de la guanine est de 1.9 Å et l'angle $\theta = 105°$), permettant une interaction simultanée avec O6. Cette sous-estimation est due essentiellement à E_{pol}. Ceci devrait avoir néanmoins une portée limitée dans des complexes multimoléculaires: en effet d'une part, l'anticooperativité de E_{pol} et de E_{ct} dans ces complexes [131] devrait en réduire l'impact ; d'autre part, les distances Zn^{2+} -ligand sont plus longues que dans les complexes mono-ligand correspondants, et cela contribue également à réduire E_{pol} dans ces complexes.

4.5.4 Validation sur des oligomères

Dans quelle mesure les calibrations ainsi obtenues sur la base de complexes 'bimoléculaires' pourraient-elles être validées par des calculs sur des complexes multimoléculaires ? Et quelle précision attendre de la méthode SIBFA compte tenu de la non-additivité de E_{pol} et de E_{ct} en passant de n=2 a n=12-24 molécules, et du fait que les molécules interagissant entre elles ne sont pas celles qui auront servi à la calibration ?

Nous avons ainsi considéré trois types de complexes multimoléculaires : des complexes intermoléculaires de la guanine (n=2-8), des complexes polymorphes du formamide et des complexes oligo-ligands de Zn^{2+}.

Complexes intermoléculaires de la guanine (n=2-8)

Le premier est le complexe de tétramères cycliques de guanines, qui structurent les quadruples hélices dans de nombreux oligonucleotides riches en guanine, et dont l'étude de deux, c-kit 1 et le LTR-III du HIV-1, constitue le sujet central de cette Thèse. Deux complexes entre huit guanines ont été considérés (Figure 4.2). Le premier est celui de deux tétramères cycliques en empilement, comme dans les quadruples hélices. Dans le deuxième complexe, les deux tétramères sont dans le même plan, et interagissent mutuellement via deux liaisons H entre deux guanines: elles se font entre le proton du groupement N9H de la première et l'azote N2 de la seconde et réciproquement. Cet arrangement est issu d'une publication de Paragi et Fonseca Guerra qui avaient considéré un complexe cyclique de quatre tétramères de guanine [162].

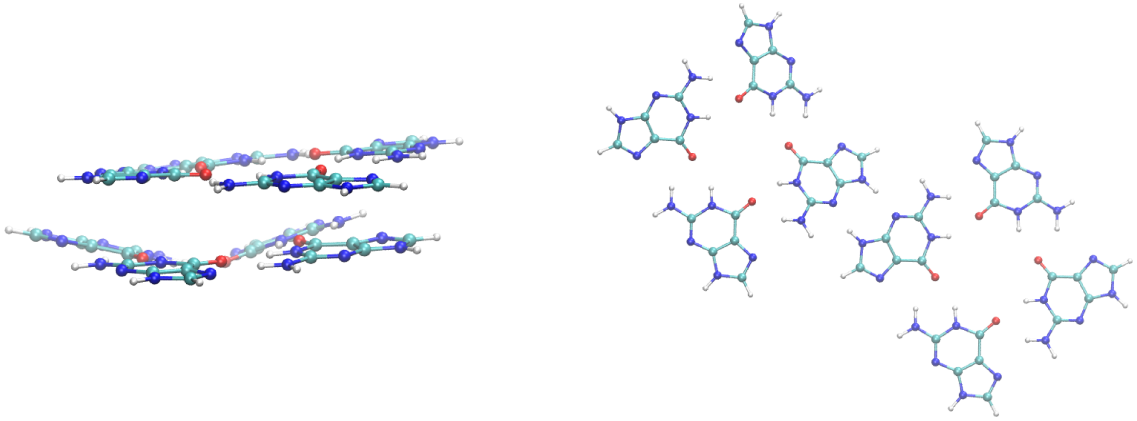
Nous avons minimisé par SIBFA ces deux structures. Nous en avons également extrait, à l'issue de la minimisation, des complexes avec n=2, 4, 6 et huit guanines afin de voir dans quelle mesure cette méthode rendait compte de l'évolution de DEtot en fonction de n. Nous avons effectué des calculs ponctuels QC DFT-D3 avec la fonctionnelle B97-D3 de Grimme *et coll.* [201] codée dans le logiciel G09 et W-B97D de Chai et Head-Gordon [42] codée dans le logiciel Q-Chem. Tous les calculs QC prennent en compte la correction de l'erreur d'extension de base (BSSE, 'Basis Set Superposition Error').

Les évolutions comparées de DEtot(SIBFA) en rouge et de DEtot(QC) avec ces deux fonctionnelles en orange et en bleu respectivement sont reportées sur la Figure 4.2, c). Les arrangements avec empilement des tétramères et les arrangements plans ont les indices 's' et 'p' respectivement. Les valeurs de DEtot(SIBFA) sont très proches des valeurs de DE(QC). Elles ont en fait des valeurs intermédiaires entre celles trouvées avec ces deux fonctionnelles. Le complexe par empilement (avant-dernier point de la courbe) s'avère plus stable que le complexe plan (dernier point) de 8.4, 8.4 et 11.4 kcal/mol par les calculs SIBFA, B97D3 et W97-D respectivement.

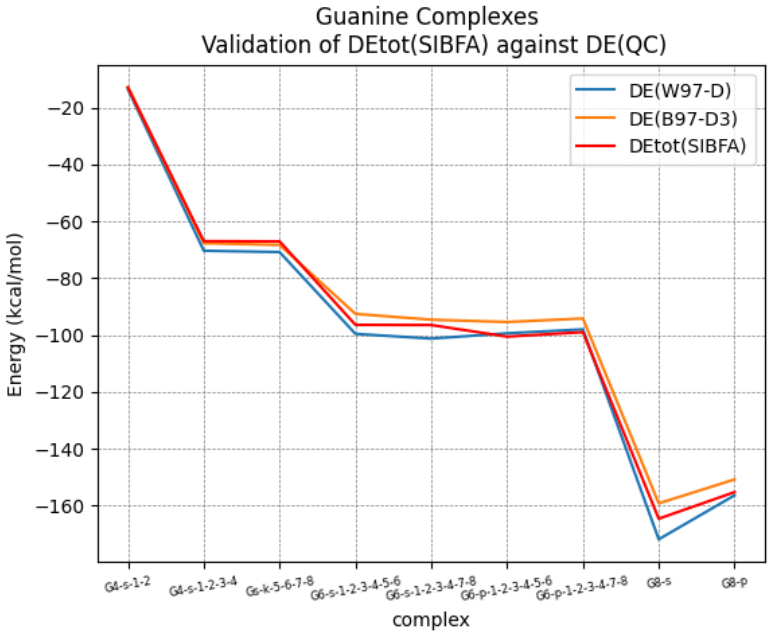
La consistance et la persistance des accords entre DEtot(SIBFA) et DE(DFT-D3) devraient dans un proche avenir motiver d'une part, des études de dynamique moléculaire qui pourraient porter sur des complexes organisés de guanine ou de guanosine en phase condensée en présence de cations métalliques et/ou d'autres entités ; et d'autre part, la construction d'un nouveau champ de force SIBFA pour les ADN et l'ARN intégrant la guanine ainsi calibrée, et les autres bases par la suite, pour le calcul de la contribution intermoléculaire.

Complexes polymorphes du formamide

Le deuxième type de complexe multimoléculaire est extrait d'un complexe 'polymorphe' du formamide [187]. La Figure 4.3 b) en donne une représentation dans le plan pour n=24 molécules. Ce complexe par liaisons H multiples est stabilisé par une grande diversité de complexes formamide–formamide: complexes bidentés ou le carbonyle d'un formamide accepte un proton du groupement NH de l'autre formamide et réciproquement donne un proton de son groupement NH au carbonyle de ce dernier, mais

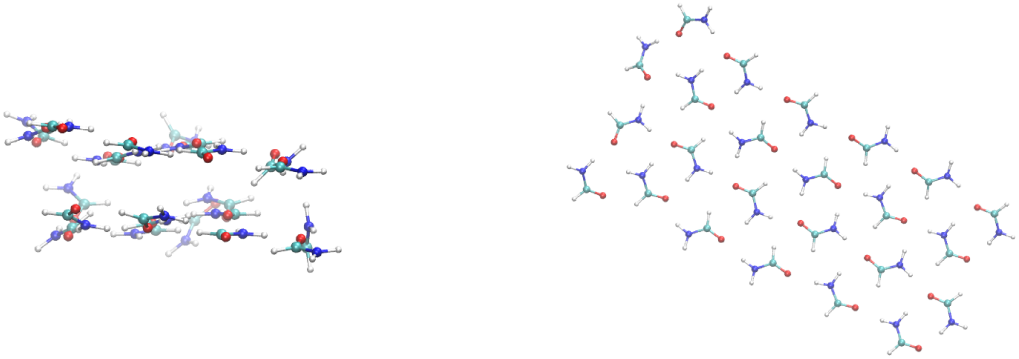


a) b)

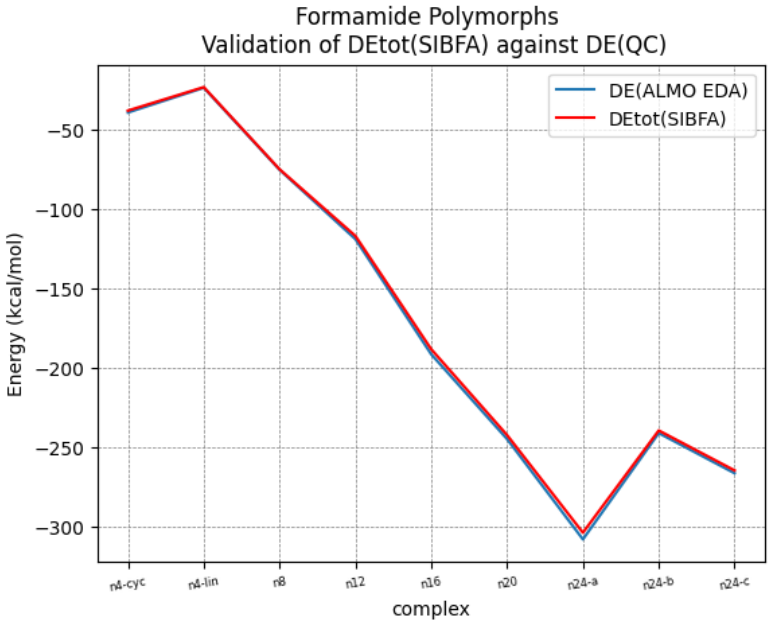


c)

Figure 4.2: Représentation des complexes intermoléculaires de la guanine a) empilés et b) interagissant dans le plan. c) Représentation graphique montrant l'énergie d'interaction pour chacun des complexes. Les complexes indiqués par 's' réfèrent aux tétramères empilés et ceux indiqués par 'p' réfèrent aux complexes interagissant dans le plan.



a) b)



c)

Figure 4.3: Représentation des complexes intermoléculaires des formamides a) empilés et b) interagissant dans le plan. c) Représentation graphique montrant l'énergie d'interaction pour chacun des complexes.

également complexes monodentés, et complexes où le proton partiellement 'acide' du groupement CH interagit avec le carbonyle d'un autre formamide, etc. Nous avons d'abord avec SIBFA et la nouvelle calibration du formamide effectué une minimisation de l'énergie intermoléculaire du complexe. La structure optimisée, dénotée A est celle représentée sur la Figure 4.3 b). Nous avons ensuite séparé les 24 formamides en deux complexes multimoléculaires de 12 formamides. Le deuxième dodécamère a été translaté sur un plan parallèle au plan du premier de façon à avoir un recouvrement optimal pour z=3.4 A. Une deuxième minimisation de l'énergie a été faite sur le second dodécamère, considéré comme un 'bloc' rigide, en ne relaxant que les six variables intermoléculaires qui définissent son orientation par rapport au premier dodécamère. À partir de la position ainsi relaxée (structure B), une troisième minimisation a été faite en relaxant la totalité des variables intermoléculaires pour chacun des formamides des deux dodécamères hormis le premier formamide du premier dodecamère, puisqu'il sert de repère (6 variables pour chacun des 23 autres formamides). La structure ainsi minimisée, dénotée C, est représentée sur la Figure 4.3, a). Un calcul ponctuel DFT-D a été ensuite effectué sur ces trois complexes avec la fonctionnelle W-B97D de Chai et Head-Gordon [42] et le logiciel Q-Chem [201].

Afin de voir dans quelle mesure SIBFA peut rendre compte de l'évolution de DEtot en fonction de n, nous avons également extrait du complexe plan avec n=24, des complexes comportant n=4, 8, 12, 16 et 20 formamides.

Sur la Figure 4.3, c), sont représentées les évolutions de DEtot(SIBFA) et de DE(DFT/W97-D) dans les différents complexes. Les trois derniers points correspondent aux structures A, B, et C pour n=24. Les valeurs de DEtot(SIBFA) sont extrêmement proches de celles de DE(DFT/W97-D) pour l'entièreté des complexes. Un point essentiel est que, pour n=24, les différences QC des énergies d'interaction entre le complexe A entièrement planaire et les complexes B et C stabilisés tant par liaisons H multiples que par empilement, sont reproduites de près par les calculs SIBFA. La préférence en faveur de l'arrangement entièrement planaire est à contraster avec la préférence inverse trouvée avec les deux tétramères de guanine, favorisant une structure par empilement plutôt qu'une structure planaire. Il est également essentiel de rendre compte du fait que l'évolution de DE(QC) en fonction de n est reproduite également. Ceci indique un contrôle correct des comportements non-additifs individuels de E_{pol} et de E_{ct}.

L'ensemble de ces résultats qui ne pouvaient pas être anticipés sur la base des seules calibrations dans les complexes bimoléculaires du formamide avec Zn^{2+} et H_2O est concluant: d'une part dans la perspective prochaine de simulations par dynamique moleculaire du formamide et de ses complexes en phase condensée, et d'autre part pour représenter la partie intermoléculaire d'un champ de force pour les protéines où le formamide constitue la partie terminale de la chaîne latérale de l'asparagine et de la glutamine.

Complexes oligo-ligands de Zn^{2+}

Le troisième type de complexes multimoléculaires est constitué de plusieurs complexes poly-ligands de Zn^{2+}. Nous avons ainsi considéré cinq complexes de Zn^{2+} avec les ligands imidazole (H), méthanethiolate (C), et formate (E), qui constituent les extrémités terminales des chaînes latérales de l'histidine, du cystéinate et du glutamate, respectivement. Les arrangements considérés se retrouvent dans les doigts à zinc des protéines rétrovirales et des facteurs de transcription, ainsi que dans des sites enzymatiques de métallo-enzymes à zinc. "LF" ('Lethal Factor') se retrouve dans le doigt à zinc d'une métalloprotéine de l'anthrax. La Figure 4.4 montre une représentation des cinq complexes oligo-ligands considérés. La charge totale autour du Zn^{2+} varie entre 0 (complexes HHHH) et -3 (complexes CCC et CCCH). Tout comme dans des études antérieures [131, 70], il était essentiel de vérifier si DEtot(SIBFA)) pouvait reproduire correctement DE(QC) compte tenu de la très forte anti-cooperativité de E_{pol} et E_{ct} pour l'ensemble de ces complexes. Nous avons ainsi comparé DEtot(SIBFA à DEtot(QC) calculé en MP2, ainsi qu'avec les fonctionnelles B97-D3 et WB97-D. Les courbes des évolutions de DEtot sont reportées sur la Figure 4.5 et le Tableau 4.1 et démontrent que les valeurs de DEtot(SIBFA) peuvent effectivement reproduire de près les valeurs de DEtot(QC) dans intervalle compris entre -400 et -700 kcal/mol. Les accords les plus proches sont ceux trouvés avec DE(MP2).

Ces résultats valident la perspective de l'intégration des fragments imidazole, formate et méthanethiolate comme extrémités des chaînes latérales de l'histidine, de l'aspartate, du glutamate, et du cystéinate, dans la construction d'un nouveau champ de force SIBFA pour les protéines, calibré sur la base des calculs SAPT(DFT).

	HHHH d = 2 Å	HHHH d = 2.1 Å	HHHH d = 2.2 Å	HHCC	CCCH	CCC	LF
ΔE(W97-D)	-431.9	-427.8	-415.8	-685.3	-700.8	-686.9	-553.8
ΔE(B97-D3)	-422.2	-419.7	-411.3	-688.8	-699.2	-690.7	-547.3
ΔE(MP2)	-418.7	-413.8	-401.8	-673.4	-688.8	-675.2	-542.8
ΔEtot(SIBFA)	-408.2	-417.6	-413.5	-683.3	-687.8	-662.7	-556.6

Table 4.1: Valeurs comparées de ΔEtot(SIBFA) et ΔE(QC) avec les fonctionnelles B97-D3, WB97-D et en MP2

Nous prévoyons d'étendre prochainement SAPT(DFT) à la calibration d'une série d'autres cations métalliques, mono- et divalents.

Nous projetons, sur la base de ces validations, d'entreprendre à court terme des simulations par dynamique moléculaire de complexes multimoléculaires impliquant la stabilisation d'arrangements par liaisons hydrogène multiples coexistant avec des complexes par empilement, et d'étudier aussi des complexes mixtes. Les dépendances à la température et à la pression seront étudiées dans ce cadre.

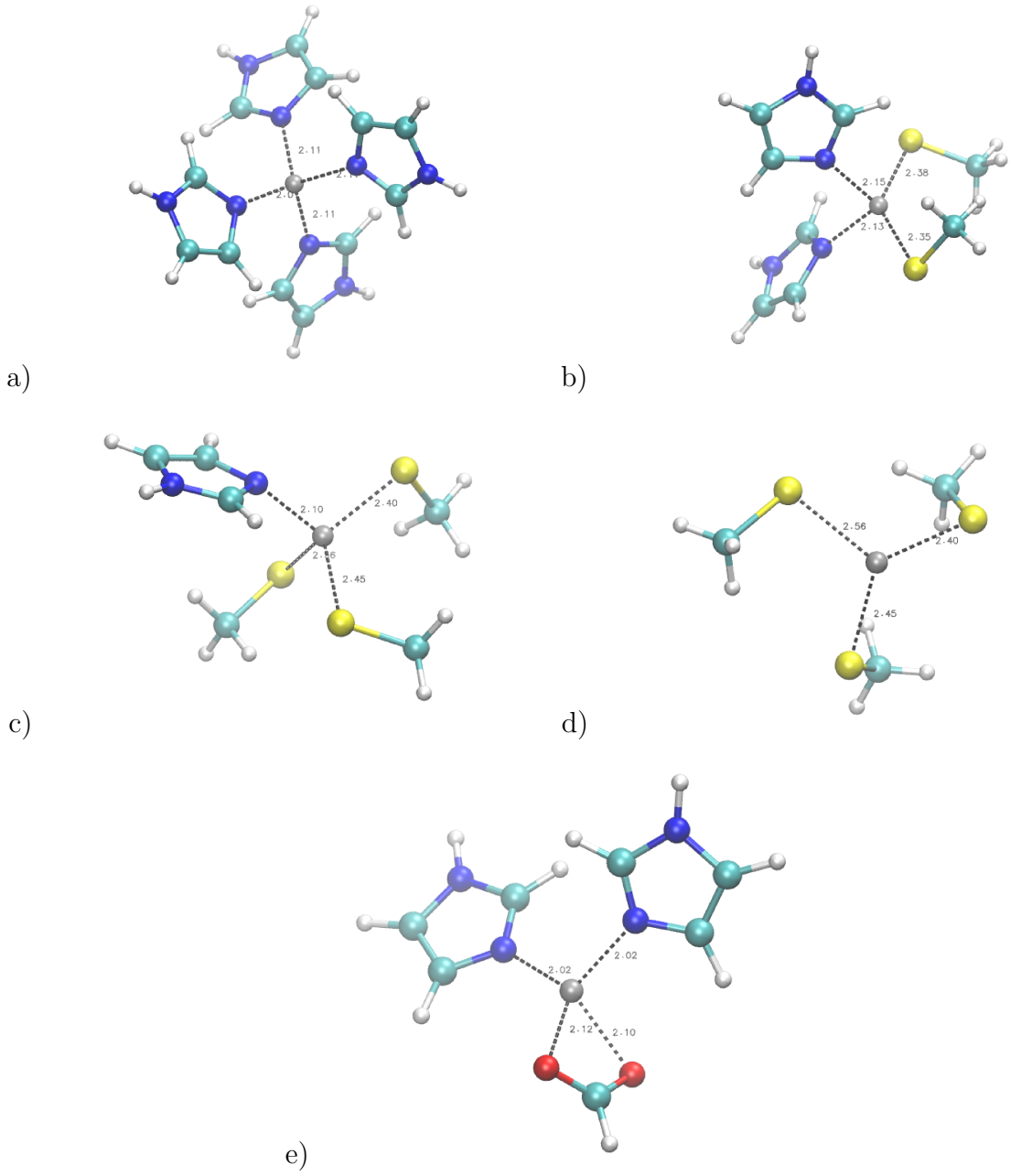


Figure 4.4: Représentation des cinq complexes poly-ligands de Zn^{2+}. a) HHHH: complexe avec quatre imidazoles b) HHCC: complexe avec deux imidazoles et deux méthanethiolates c) CCCH: complexe avec trois méthanethiolates et un imidazole d) CCC: complexe avec trois méthanethiolates et e) LF: complexe avec deux imidazoles et un formate

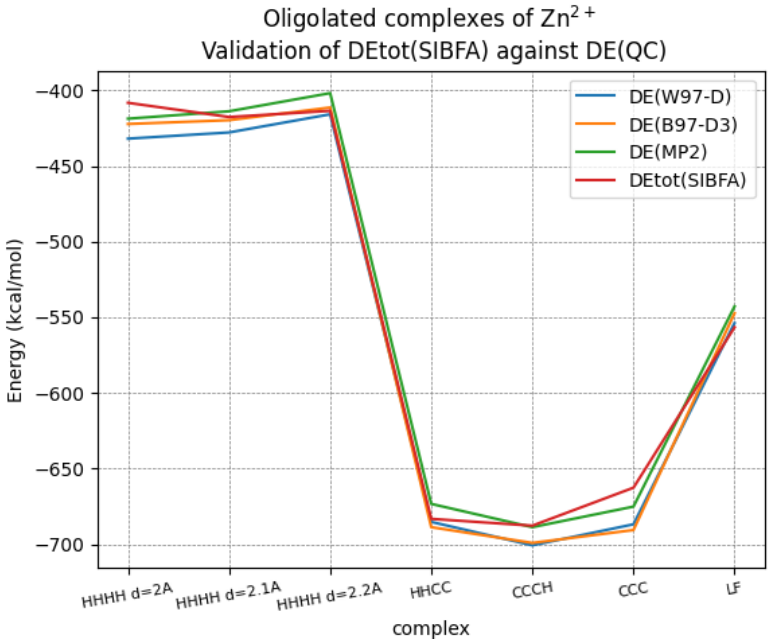


Figure 4.5: Représentation graphique des montrant l'énergie d'interaction intermoléculaire de chaque complexe oligo-ligands de Zn^{2+}

Chapitre 5

Références bibliographiques

Références bibliographiques

[1] The HSL Mathematical Software Library.

[2] ONU SIDA, Fiche d'information — Dernières statistiques sur l'état de l'épidémie de SIDA.

[3] WHO Coronavirus (COVID-19) Dashboard.

[4] Data Tinker–hp, SARS–CoV–2 Main Protease, deposited at CSCS, 2020.

[5] PDB 6Y84 : Structure COVID-19 main protease with unliganded active site., 2020.

[6] Hervé Abdi and Lynne J. Williams. Principal component analysis. *WIREs Computational Statistics*, 2(4):433–459, 2010.

[7] Olivier Adjoua, Louis Lagardère, Luc-Henri Jolly, Arnaud Durocher, Thibaut Very, Isabelle Dupays, Zhi Wang, Théo Jaffrelot Inizan, Frédéric Célerse, Pengyu Ren, Jay W. Ponder, and Jean-Philip Piquemal. Tinker-hp: Accelerating molecular dynamics simulations of large complex systems with advanced point dipole polarizable force fields using gpus and multi-gpu systems. *J. Chem. Theory. Comput.*, 17(4):2034–2053, 2021.

[8] Rodríguez-Morales AJ, MacGregor K, Kanagarajah S, Patel D, and Schlagenhauf P. Going global - travel and the 2019 novel coronavirus. *Travel. Med. Infect Dis.*, 33(101578), 2020.

[9] Angelo Spinello-Bernhard K Keppler Christophe Chipot François Dehez Giampaolo Barone Alessio Terenzi, Hugo Gattuso and Antonio Monari. Targeting g-quadruplexes with organic dyes: Chelerythrine-dna binding elucidated by combining molecular modeling and optical spectroscopy. 8:472, 2019.

[10] Anthony J. Stone Alston J. Misquitta and Farhang Fazeli. Distributed multipoles from a robust basis-space implementation of the iterated stockholder atoms procedure. *J. Chem. Theory Comput.*, 10(12):5405–5418, 2014.

[11] A Amadei, ABM Linssen, BL De Groot, DMF Van Aalten, and HJC Berendsen. An efficient method for sampling the essential subspace of proteins. *J. Biomol. Struct. Dyn.*, 13(4):615–625, 1996.

[12] Andrea Amadei, Antonius BM Linssen, and Herman JC Berendsen. Essential dynamics of proteins. *Proteins.*, 17(4):412–425, 1993.

[13] B. R. C. Amor, M. T. Schaub, S. N. Yaliraki, and M. Barahona. Prediction of allosteric sites and mediating interactions through bond-to-bond propensities. *Nat. Commun.*, 7:12477, 2016.

[14] Ramu Anandakrishnan, Aleksander Drozdetski, Ross C. Walker, and Alexey V. Onufriev. Speed of conformational change: Comparing explicit and implicit solvent molecular dynamics simulations. *Biophys J.*, 108(5):1153–1164, 2015.

[15] Hans C. Andersen. Molecular dynamics simulations at constant pressure and/or temperature. *J. Chem. Phys.*, 72(2384), 1980.

[16] Hai Yan Gaw Anh Tuân Phan, Vitaly Kuryavyi and Dinshaw J Patel. Small-molecule interaction with a five-guanine-tract g-quadruplex structure from the human myc promoter. *Nat Chem Biol.*, 1(3):167–73, 2005.

[17] Arun Shivalingam Jorge Gonzalez-Garcia Ramon Vilar Anita Kotar, Baifan Wang and Janez Plavec. Nmr structure of a triangulenium-based long-lived fluorescence probe bound to a g-quadruplex. *Angew Chem Int Ed Engl.*, 26(55):12508–11, 2016.

[18] Indwiani Astuti and Ysrafil. Severe acute respiratory syndrome coronavirus 2 (sars-cov-2): An overview of viral structure and host response. *Diabetes Metab Syndr.*

[19] Shankar Balasubramanian, Laurence H. Hurley, and Stephen Neidle. Targeting g-quadruplexes in gene promoters: a novel anticancer strategy? 10:261–275.

[20] Alessandro Barducci, Massimiliano Bonomi, and Michele Parrinello. Metadynamics. *WIREs Computational Molecular Science Metadynamics*, 1, 2011.

[21] B.Coutard, C.Valle, X.de Lamballeriea B.Canard, N.G.Seidah, and E.Decroly. The spike glycoprotein of the new coronavirus 2019-ncov contains a furin-like cleavage site absent in cov of the same clade. 2020.

[22] Marie-Claire Bellissent-Funel, Ali Hassanali, Martina Havenith, Richard Henchman, Peter Pohl, Fabio Sterpone, David van der Spoel, Yao Xu, and Angel E Garcia. Water determines the structure and dynamics of proteins. *Chem. Rev.*, 116(13):7673–7697, 2016.

[23] Sandrine Belouzard, Jean K Millet, Beth N Licitra, and Gary R Whittaker. Mechanisms of coronavirus cell entry mediated by the viral spike protein. *Viruses*, 2012.

[24] Herman JC Berendsen and Steven Hayward. Collective protein dynamics in relation to function. *Curr. Opin. Struct. Biol.*, 10(2):165–169, 2000.

[25] H Bertrand, T Ha-Duong, S Fermandjian, and B Hartmann. Flexibility of the b-dna backbone: effects of local and neighbouring sequences on pyrimidine-purine steps. *Nucleic Acids Res.*, 26(5):1261–7, 1998.

[26] Robin M. Betz and Ron O. Dror. How effectively can adaptive sampling methods capture spontaneous ligand binding? *J. Chem. Theory. Comput.*, 15(3):2053–2063, 2019.

[27] Giulia Biffi, David Tannahill, John McCafferty, and Shankar Balasubramanian. Quantitative visualization of dna g-quadruplex structures in human cells. *Nat Chem.*, 5(3):182–6, 2013.

[28] Giulia Biffi, David Tannahill, Jodi Miller, William J Howat, and Shankar Balasubramanian. Elevated levels of g-quadruplex formation in human stomach and liver cancer tissues. *PLoS One.*, 17(9):7, 2014.

[29] Nicole Boggetto and Michèle Reboud-Ravaux. Dimerization inhibitors of hiv-1 protease. *Biol. Chem.*, 383:1321 – 1324,, 2002.

[30] Robert Moszynski Bogumil Jeziorski and Krzysztof Szalewicz. Perturbation theory approach to intermolecular potential energy surfaces of van der waals complexes. *Chem. Rev.*, 94(7):1887–1930, 1994.

[31] Gregory R. Bowman, Daniel L. Ensign, and Vijay S. Pande. Enhanced modeling via network theory: Adaptive sampling of Markov state models. *J. Chem. Theory. Comput.*, 6(3):787–794, 2010.

[32] Davide Branduardi, Francesco Luigi Gervasio, Andrea Cavalli, Maurizio Recanatini, and Michele Parrinello. The role of the peripheral anionic site and cation- π interactions in the ligand penetration of the human AChE gorge. *J. Am. Chem. Soc.*, 127(25):9147–9155, 2005.

[33] Tracy A. Brooks and Laurence H. Hurley. The role of supercoiling in transcriptional control of myc and its importance in molecular therapeutics. 9:849–861.

[34] David K Brown, David L Penkler, Olivier Sheik Amamuddy, Caroline Ross, Ali Rana Atilgan, Canan Atilgan, and Ozlem Tastan Bishop. MD–TASK: a software suite for analyzing molecular dynamics trajectories. *J. Bioinform.*, 33(17):2768–2771, 05 2017.

[35] Sarah Burge, Gary N. Parkinson, Pascale Hazel, Alan K. Todd, and Stephen Neidle. Quadruplex DNA: sequence, topology and structure. 34(19):5402–5415.

[36] Elena Butovskaya, Brahim Heddi, Blaž Bakalar, Sara N. Richter, and Anh Tuân Phan. Major g-quadruplex form of HIV-1 LTR reveals a (3 + 1) folding topology containing a stem-loop. 140(42):13654–13662.

[37] Elena Butovskaya, Paola Soldà, Matteo Scalabrin, Matteo Nadai, and Sara N. Richter. HIV-1 nucleocapsid protein unfolds stable RNA g-quadruplexes in the viral genome and is inhibited by g-quadruplex ligands. 5(12):2127–2135.

[38] Li C, Qi Y, Teng X, Yang Z, Wei P, Zhang C, Tan L, Zhou L, Liu Y, and Lai L. Maturation mechanism of severe acute respiratory syndrome (sars) coronavirus 3c-like proteinase. *J Biol Chem*, 285:28134–28140, 2010.

[39] Salvatore Cardamone, Timothy J. Hughes, and Paul L. A. Popelier. Multipolar electrostatics. *Phys. Chem. Chem. Phys.*, 16:10367–10387, 2014.

[40] Anna Rita Bilia Francesca Scheggi Carla Bazzicalupi, Marta Ferraroni and Paola Gratteri. The crystal structure of human telomeric dna complexed with berberine: an interesting case of stacked ligand to g-tetrad ratio higher than 1:1. *Nucleic Acids Res.*, 41(1):632–638, 2013.

[41] Matteo Carli, Giulia Sormani, Alex Rodriguez, and Alessandro Laio. Candidate binding sites for allosteric inhibition of the sars-cov-2 main protease from the analysis of large-scale molecular dynamics simulations. *J. Phys. Chem. Lett.*, 12(1):65–72, 2021.

[42] Jeng-Da Chaia and Martin Head-Gordon. Long-range corrected hybrid density functionals with damped atom–atom dispersion corrections. *Phys. Chem. Chem. Phys.*, 10(44):6615, 2008.

[43] Hao Chen, Ping Wei, Changkang Huang, Lei Tan, Ying Liu, and Luhua Lai. Only one protomer is active in the dimer of SARS 3C-like proteinase. *J. Biol. Chem.*, 281(20):13894–13898, 2006.

[44] Chi-Yuan Chou, Hui-Chuan Chang, Wen-Chi Hsu, Tien-Zheng Lin, Chao-Hsiung Lin, , and Gu-Gang Chang. Quaternary structure of the severe acute respiratory syndrome (sars) coronavirus main protease. *Biochemistry.*, 43(47):14958–14970, 2004.

[45] Sophie Combet and Jean-Marc Zanotti. Further evidence that interfacial water is the main "driving force" of protein dynamics: a neutron scattering study on perdeuterated c-phycocyanin. *Phys. Chem. Chem. Phys.*, 14:4927–4934, 2012.

[46] Christopher J. Cramer and Donald G. Truhlar. Implicit solvation models: equilibria, structure, spectra, and dynamics. *Chem. Rev.*, 99(8):2161–2200, 1999.

[47] D Suplatov D and V Švedas. Study of functional and allosteric sites in protein superfamilies. *Acta. Naturae.*, 7(4):34–45, 2015.

[48] Jayangika N. Dahanayake and Katie R. Mitchell-Koch. How does solvation layer mobility affect protein structural dynamics? *Front. Biosci.*, 5:65, 2018.

[49] Rhodes Daniela and Lipps Hans J. G-quadruplexes and their regulatory roles in biology. 43(18):8627–8637.

[50] Martin Zacharias Danilo Roccatano 1, Andre Barthel. Structural flexibility of the nucleosome core particle at atomic resolution studied by molecular dynamics simulation. *Biopolymers.*, 85(5-6):407–21, 2007.

[51] Charles C David and Donald J Jacobs. Principal component analysis: a method for determining the essential dynamics of proteins. *Methods Mol Biol.*, 1084:193–226, 2014.

[52] Marie-Pierre de Béthune. Non-nucleoside reverse transcriptase inhibitors (nnrtis), their discovery, development, and use in the treatment of hiv-1 infection: a review of the last 20 years (1989-2009). *Antiviral Res.*, 85(1):75–90, 2009.

[53] B. de Courcy, L. G. Pedersen, O. Parisel, N. Gresh, B. Silvi, J. Pilme, and J.-P. Piquemal. Understanding selectivity of hard and soft metal cations within biological systems using the subvalence concept. 1. application to blood coagulation:direct cation–protein electronic effects versus indirect interactions through water networks. *J. Chem. Theory. Comput.*, 6(4):1048–1063, 2010.

[54] Benoit de Courcy, Jean-Philip Piquemal, Christiane Garbay, and Nohad Gresh. Polarizable water molecules in ligand-macromolecule recognition. impact on the relative affinities of competing pyrrolopyrimidine inhibitors for fak kinase. *J. Am. Chem. Soc.*, 132:3312–3320, 2006.

[55] B Delmas and H Laude. Assembly of coronavirus spike protein into trimers and its role in epitope expression. *J Virol.*, 1990.

[56] Philippe DELVENNE. Référentiel des examens. biologie clinique - génétique - anatomie et cytologie pathologique. `https://www.chu.ulg.ac.be/jcms/c_637276/fr/cd-117`.

[57] DESRES. Desres-anton-10880334. *DESRES : Molecular Dynamics Simulations Related to SARS-CoV-2*, pages DESRES–ANTON–10880334, 2020.

[58] Marie-Cécile Didiot, Zhaoxia Tian, Céline Schaeffer, Murugan Subramanian, Jean-Louis Mandel, and Hervé Moine. The g-quartet containing fmrp binding site in fmr1 mrna is a potent exonic splicing enhancer. *Nucleic Acids Res.*, 36(15):4902–12, 2008.

[59] Théo Jaffrelot Inizan Fréderic Célerse Chengwen Liu Olivier Adjoua Luc-Henri Jolly Nohad Gresh Zeina Hobaika Pengyu Ren Richard G. Maroun Dina El Ahdab, Louis Lagardère and Jean-Philip Piquemal. Interfacial water many-body effects drive structural dynamics and allosteric interactions in sars-cov-2 main protease dimerization interface. 12(26):6218–6226, 2021.

[60] Li Ding, Xin-Xiang Zhang, Ping Wei, Keqiang Fan, and Luhua Lai. The interaction between severe acute respiratory syndrome coronavirus 3c-like proteinase and a dimeric inhibitor by capillary electrophoresis. *Anal. Biochem.*, 434(1):159–165, 2005.

[61] Gaurav Sharma Dorota Piekna-Przybylska, Mark A Sullivan and Robert A Bambara. U3 region in the hiv-1 genome adopts a g-quadruplex structure in its rna and dna sequence. *biochemistry*, 53(16):2581–93, 2014.

[62] Johanna Eddy and Nancy Maizels. Conserved elements with potential to form polymorphic g-quadruplex structures in the first intron of human genes. *Nucleic Acids Res.*, 36(4):1321–33, 2008.

[63] Matteo Scalabrin Rosalba Perrone Elena Butovskaya Matteo Nadai Giorgio Palù Dan Fabris Elena Tosoni, Ilaria Frasson and Sara N Richter. Nucleolin stabilizes g-quadruplex structures folded by the ltr promoter and silences hiv-1 viral transcription. *Nucleic Acids Res.*, 43(18):8884–97, 2015.

[64] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd*, volume 96, pages 226–231, 1996.

[65] Peng Zhou et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature*, 2020.

[66] P. Ewald. Evaluation of optical and electrostatic lattice potentials. *Ann. Phys*, 64:253–287, 1921.

[67] Brittni M. Foster, Danish Zaidi, Tyler R. Young, Mary E. Mobley, , and Bethany A. Kerr. Cd117/c-kit in cancer stem cell-mediated progression and therapeutic resistance. *Biomedicines.*, 6(1):31, 2018.

[68] Andrey Frolov, Santiago Chahwan, Michael Ochs, Juan Pablo Arnoletti, Zhong-Zong Pan, Olga Favorova, Jonathan Fletcher, Margaret von Mehren, Burton Eisenberg, and Andrew K. Godwin. Response markers and the molecular mechanisms of action of gleevec in gastrointestinal stromal tumors. 2(8).

[69] Sharon R Lewin Gary Maartens, Connie Celum. Hiv infection: epidemiology, pathogenesis, treatment, and prevention. *Lancet*, 384(9939):258–71, 2014.

[70] Claude Giessner-Prettre Lee G. Pedersen Gilles Tiraboschi, Nohad Gresh and David W. Deerfield. Parallel ab initio and molecular mechanics investigation of polycoordinated zn(ii) complexes with model hard and soft ligands: Variations of binding energy and of its components with number and charges of ligands. *J. Comput. Chem.*, 21(12):1011–1039, 2000.

[71] Alexander E Gorbalenya, Luis Enjuanes, John Ziebuhr, and Eric J Snijder. Nidovirales: evolving the largest rna virus genome. *Virus Res.*, 2006.

[72] Bhupesh Goyal and Bhupesh Goyal. Targeting the dimerization of the main protease of coronaviruses: A potential broad-spectrum therapeutic strategy. *ACS. Comb. Sci.*, 22(6):297–305, 2020.

[73] Joe G Greener and Michael JE Sternberg. Structure-based prediction of protein allostery. *Curr. Opin. Struct. Biol.*, 50:1–8, 2018.

[74] Nohad Gresh and David Perahia. Multimolecular complexes of the phosphodiester anion with zn(ii) or mg(ii) and water molecules-preliminary validations of a polarizable potential by ab initio quantum chemistry. *J. Comput. Chem.*, 42(20):1430–1446, 2021.

[75] Sebastian Günther, Patrick Y. A. Reinke, Yaiza Fernández-García, Julia Lieske, Thomas J. Lane, Helen M. Ginn, Faisal H. M. Koua, Christiane Ehrt, Wiebke Ewert, Dominik Oberthuer, Oleksandr Yefanov, Susanne Meier, Kristina Lorenzen, Boris Krichel, Janine-Denise Kopicki, Luca Gelisio, Wolfgang Brehm, Ilona Dunkel, Brandon Seychell, Henry Gieseler, Brenna Norton-Baker, Beatriz Escudero-Pérez, Martin Domaracky, Sofiane Saouane, Alexandra Tolstikova, Thomas A. White, Anna Hänle, Michael Groessler, Holger Fleckenstein, Fabian Trost, Marina Galchenkova, Yaroslav Gevorkov, Chufeng Li, Salah Awel, Ariana Peck, Miriam Barthelmess, Frank Schluenzen, Paulraj Lourdu Xavier, Nadine Werner, Hina Andaleeb, Najeeb Ullah, Sven Falke, Vasundara Srinivasan, Bruno Alves França, Martin Schwinzer, Hévila Brognaro, Cromarte Rogers, Diogo Melo, Joanna J. Zaitseva-Doyle, Juraj Knoska, Gisel E. Peña-Murillo, Aida Rahmani Mashhour, Vincent Hennicke, Pontus Fischer, Johanna Hakanpää, Jan Meyer, Philip Gribbon, Bernhard Ellinger, Maria Kuzikov, Markus Wolf, Andrea R. Beccari, Gleb Bourenkov, David von Stetten, Guillaume Pompidor, Isabel Bento, Saravanan Panneerselvam, Ivars Karpics, Thomas R. Schneider, Maria Marta Garcia-Alai, Stephan Niebling, Christian Günther, Christina Schmidt, Robin Schubert, Huijong Han, Juliane Boger, Diana C. F. Monteiro, Linlin Zhang, Xinyuanyuan Sun, Jonathan Pletzer-Zelgert, Jan Wollenhaupt, Christian G. Feiler, Manfred S. Weiss, Eike-Christian Schulz, Pedram Mehrabi, Katarina Karničar, Aleksandra Usenik, Jure Loboda, Henning Tidow, Ashwin Chari, Rolf Hilgenfeld, Charlotte Uetrecht, Russell Cox, Andrea Zaliani, Tobias Beck, Matthias Rarey, Stephan Günther, Dusan Turk, Winfried Hinrichs, Henry N. Chapman, Arwen R. Pearson, Christian Betzel, and Alke Meents. X-ray screening identifies active site and allosteric inhibitors of sars-cov-2 main protease. *Science*, 2021.

[76] Charles C. Hardin, Thomas Watson, Matthew Corregan, and Charles Bailey. Cation-dependent transition between the quadruplex and watson-crick hairpin forms of d(CGCG3GCG). 31(3):833–841.

[77] Michael C Heinrich, Christopher L Corless, George D Demetri, Charles D Blanke, Margaret von Mehren, Heikki Joensuu, Laura S McGreevey, Chang-Jie Chen, Annick D Van den Abbeele, Brian J Druker, Beate Kiese, Burton Eisenberg, Peter J Roberts, Samuel Singer, Christopher D M Fletcher, Sandra Silberman, Sasa Dimitrijevic, and Jonathan A Fletcher. Kinase mutations and imatinib response in patients with metastatic gastrointestinal stromal tumor. 21(23):4342–9.

[78] Jan Hermans. Hydrogen bonds in molecular mechanics force fields. In *Peptide Solvation and HBonds*, volume 72 of *Advances in Protein Chemistry*, pages 105 – 119. Academic Press, 2005.

[79] Rolf Hilgenfeld. From sars to mers: crystallographic studies on coronaviral proteases enable antiviral drug design. *The FEBS Journal*, 281(18):4085–4096, 2014.

[80] Eugen Hruska, Jayvee R. Abella, Feliks Nüske, Lydia E. Kavraki, and Cecilia Clementi. Quantitative comparison of adaptive sampling methods for protein dynamics. *J. Chem. Phys.*, 149(24):244119, 2018.

[81] R.J.G. Hulswit, C.A.M. de Haan, and B.-J. Bosch. Coronavirus spike protein and tropism changes. *Adv Virus Res*, 2016.

[82] William Humphrey, Andrew Dalke, and Klaus Schulten. VMD – Visual Molecular Dynamics. *Journal of Molecular Graphics*, 14:33–38, 1996.

[83] Julian L Huppert and Shankar Balasubramanian. Prevalence of quadruplexes in the human genome. 33:2908–16, 2005.

[84] Julian L Huppert and Shankar Balasubramanian. G-quadruplexes in promoters throughout the human genome. 35:406–13, 2007.

[85] Julian Leon Huppert, Anthony Bugaut, Sunita Kumari, and Shankar Balasubramanian. G-quadruplexes: the beginning and end of utrs. 36(19).

[86] Brooke E. Husic and Vijay S. Pande. Markov state models: From an art to a science. *J. Am. Chem. Soc.*, 140, 7, 2386–2396(7):2386–2396, 2018.

[87] Clarisse Berlioz-Torrent Hélène Bauby, Sandra L Lopez-Vergès. Tip47, un cofacteur cellulaire essentiel à l'incorporation des glycoprotéines d'enveloppe du vih-1 dans la particule virale. 12(3):201–13.

[88] Toshiko Ichiye and Martin Karplus. Collective motions in proteins: a covariance analysis of atomic fluctuations in molecular dynamics and normal mode simulations. *Proteins.*, 11(3):205–217, 1991.

[89] Pfizer Inc. Pfizer initiates phase 1 study of novel oral antiviral therapeutic agent against sars-cov-2. https://www.pfizer.com/news/press-release/press-release-detail/pfizer-initiates-phase-1-study-novel-oral-antiviral.

[90] Barira Islam, Petr Stadlbauer, Miroslav Krepl, Jaroslav Koca, Stephen Neidle, Shozeb Haider, and Jiri Sponer. Extended molecular dynamics of a c-kit promoter quadruplex. 43(18):8673–8693.

[91] Barira Islam, Petr Stadlbauer, Miroslav Krepl, Jaroslav Koca, Stephen Neidle, Shozeb Haider, and Jiri Sponer. Extended molecular dynamics of a c-kit promoter quadruplex. 43(18):8673–8693.

[92] B R Gelin J A McCammon and M Karplus. Dynamics of folded proteins. *Nature.*, 267(5612):585–90, 1977.

[93] Theo Jaffrelot-Inizan, Frédéric Célerse, Olivier Adjoua, Dina El Ahdab, Luc-Henri Jolly, Chengwen Liu, Pengyu Ren, Matthieu Montes, Nathalie Lagarde, Louis Lagardère, Pierre Monmarché, and Jean-Philip Piquemal. High-resolution mining of sars-cov-2 main protease conformational space: Supercomputer-driven unsupervised adaptive sampling. *Chem. Sci.*, 2021.

[94] Théo Jaffrelot-Inizan, Fréderic Célerse, Olivier Adjoua, Dina El Ahdab, Luc-Henri Jolly, Matthieu Montes, Nathalie Lagarde, Pengyu Ren, Louis Lagardère, Pierre Monmarché, and Jean-Philip Piquemal. High-resolution mining of the sars-cov-2 main protease conformational space: supercomputer-driven unsupervised adaptive sampling. 2020.

[95] Kevin E. Riley Jan Řezáč and Pavel Hobza. Erratum to "s66: A well-balanced database of benchmark interaction energies relevant to biomolecular structures". *J. Chem. Theory Comput.*, 10(3):1359–1360, 2014.

[96] Manuel Ruiz-Lopez Jean-Louis Rivail and Xavier Assfeld. Quantum modeling of complex molecular systems. *Theoretical and Computational Chemistry*, 2015.

[97] Nohad Gresh Jean-Philip Piquemal and Claude Giessner-Prettre. Improved formulas for the calculation of the electrostatic contribution to the intermolecular interaction energy from multipolar expansion of the electronic distribution. *J. Phys. Chem. A*, 107:10353–10359, 2003.

[98] Zhenming Jin, Xiaoyu Du, Yechun Xu, Yongqiang Deng, Meiqin Liu, Yao Zhao, Bing Zhang, Xiaofeng Li, Leike Zhang, Chao Peng, et al. Structure of m pro from sars-cov-2 and discovery of its inhibitors. *Nature.*, pages 1–5, 2020.

[99] Laurence H Hurley Jixun Dai, Megan Carver and Danzhou Yang. Solution structure of a 2:1 quindoline-c-myc g-quadruplex: insights into g-quadruplex-interactive small molecule drug design. *J Am Chem Soc.*, 133(44):17673–80, 2011.

[100] Philip D Jeffrey Steven P Barrett Paul R Rablen Thomas J Lawton John M Nicoludis, Stephen T Miller and Liliya A Yatsunyk. Optimized end-stacking provides specificity of n-methyl mesoporphyrin ix for human telomeric g-quadruplex dna. *J Am Chem Soc.*, 134(50):20446–56, 2012.

[101] Luc-Henri Jolly, Alejandro Duran, Louis Lagardère, Jay W. Ponder, Pengyu Ren, and Jean-Philip Piquemal. Raising the performance of the tinker-HP molecular modeling package [article v1.0]. *LiveCoMS.*, 1(2):10409, 2019.

[102] William L. Jorgensen, Jayaraman Chandrasekhar, Jeffry D. Madura, Roger W. Impey, and Michael L. Klein. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.*, 79(2):926–935, 1983.

[103] Ignacio Tinoco Jr and Jin-Der Wen. Simulation and analysis of single-ribosome translation. *Phys Biol.*, 6(2):025006, 2009.

[104] Hendrik R A Jonker J Tassilo Grün Francesco Papi Carla Bazzicalupi Luigi Messori Paola Gratteri Julia Wirmer-Bartoschek, Lars Erik Bendel and Harald Schwalbe. Solution nmr structure of a ligand/hybrid-2-g-quadruplex complex reveals rearrangements that affect ligand binding. *Angew Chem Int Ed Engl.*, 56(25):7102–7106, 2017.

[105] Jakub Kaminsky and Frank Jensen. Force field modeling of amino acid conformational energies. *J. Chem. Theory Comput*, 3(5):1774–1788, 2007.

[106] Hengning Ke, Julhash U Kazi, Hui Zhao, and Jianmin Sun. Germline mutations of kit in gastrointestinal stromal tumor (gist) and mastocytosis. *Cell Biosci/*, 18(6):55, 2016.

[107] Dongwan Kim, Joo-Yeon Lee, Jeong-Sun Yang, Jun Won Kim, V. Narry Kim, and Hyeshik Chang. The architecture of sars-cov-2 transcriptome. *Cell.*

[108] Jeong-Min Kim, Yoon-Seok Chung adn Hye Jun Jo, Nam-Joo Lee, Mi Seon Kim, Sang Hee Woo, Sehee Park, Jee Woong Kim, Heui Man Kim, and Myung-Guk Hana. Identification of coronavirus isolated from a patient in korea with covid-19. 2020.

[109] Daniel W Kneller, Gwyndalyn Phillips, Hugh M O'Neill, Robert Jedrzejczak, Lucy Stols, Paul Langan, Andrzej Joachimiak, Leighton Coates, and Andrey Kovalevsky. Structural plasticity of sars-cov-2 3cl m pro active site cavity revealed by room temperature x-ray crystallography. *Nat. Commun.*, 11(1):1–6, 2020.

[110] Teruhisa S. Komatsu, Yohei Koyama, Noriaki Okimoto, Gentaro Morimoto, Yousuke Ohno, and Makoto Taiji. Covid-19 related trajectory data of 10 microseconds all atom molecular dynamics simulation of sars-cov-2 dimeric main protease, v2. *Mendeley Data*, 10, 17632, 2020.

[111] Anita Kotar, Riccardo Rigo, Claudia Sissi, and Janez Plavec. Two-quartet kit* g-quadruplex is formed via double-stranded pre-folded structure. *Nucleic Acids Research*, 47(5):2641–2653, 2019.

[112] Chih-Jung Kuo and Po-Huang Liang. Characterization and inhibition of the main protease of severe acute respiratory syndrome coronavirus. *ChemBioEng. Rev.*, 2(2):118–132, 2015.

[113] A Peeters M J Churchill L A Pereira, K Bentley and N J Deacon. A compilation of cellular transcription factor interactions with the hiv-1 ltr promoter. *Nucleic Acids Res.*, 28(3):663–8, 2000.

[114] Louis Lagardère, Luc-Henri Jolly, Filippo Lipparini, Félix Aviat, Benjamin Stamm, Zhifeng F. Jing, Matthew Harger, Hedieh Torabifard, G. Andrés Cisneros, Michael J. Schnieders, Nohad Gresh, Yvon Maday, Pengyu Y. Ren, Jay W. Ponder, and Jean-Philip Piquemal. Tinker-hp: a massively parallel molecular dynamics package for multiscale simulations of large complex systems with advanced point dipole polarizable force fields. *Chem. Sci.*, 9:956–972, 2018.

[115] Louis Lagardère, Félix Aviat, and Jean-Philip Piquemal. Pushing the limits of multiple-time-step strategies for polarizable point dipole molecular dynamics. *J. Phys. Chem. Lett.*, 10(10):2593–2599, 2019.

[116] M M Lai and S A Stohlman. Comparative analysis of rna genomes of mouse hepatitis viruses. *J Virol.*, 1981.

[117] Alessandro Laio and Michele Parrinello. Escaping free-energy minima. *Proc Natl Acad Sci U S A.*, 99(20):12562–6, 2002.

[118] Alessandro Laio and Michele Parrinello. Escaping free-energy minima. *Proc. Nat. Acad. Sci. USA.*, 99(20):12562–12566, 2002.

[119] Justin A. Lemkul, Jing Huang, Benoît Roux, and Alexander D. MacKerell. An empirical polarizable force field based on the classical drude oscillator model: Development history and recent applications. *Chemical Reviews*, 116(9):4983–5013, 2016. PMID: 26815602.

[120] Yaakov Levy and José N Onuchic. Water mediation in protein folding and molecular recognition. *Annu Rev Biophys Biomol Struct*, 2006.

[121] Wei Li, Peng Wu, Tastuo Ohmichi, and Naoki Sugimoto. Characterization and thermodynamic properties of quadruplex/duplex competition. 526(1-3):77–81.

[122] Julia Liang, Chris Karagiannis, Eleni Pitsillou, Kevion K Darmawan, Ken Ng, Andrew Hung, and Tom C Karagiannis. Site mapping and small molecule blind docking reveal a possible target

site on the SARS-CoV-2 main protease dimer interface. *Comput. Biol. Chem.*, page 107372, 2020.

[123] Julia Liang, Chris Karagiannis, Eleni Pitsillou, Kevion K. Darmawan, Ken Ng, Andrew Hung, and Tom C. Karagiannis. Site mapping and small molecule blind docking reveal a possible target site on the sars-cov-2 main protease dimer interface. *Comput. Biol. Chem.*, 89(107372):1476–9271, 2020.

[124] Si-Ming Liao, Qi-Shi Du, Jian-Zong Meng, Zong-Wen Pang, and Ri-Bo Huang. The multiple roles of histidine in protein interactions. *Chem. Cent. J.*, 7(1):44, 2013.

[125] Cynthia Liu, Qiongqiong Zhou, Yingzhu Li, Linda V. Garner, Steve P. Watkins, Linda J. Carter, Jeffrey Smoot, Anne C. Gregg, Angela D. Daniels, Susan Jervey, and Dana Albaiu. Research and development on therapeutic agents and vaccines for covid-19 and related human coronavirus diseases. 2020.

[126] Lin Liu, Susan M Bailey, Maja Okuka, Purificación Muñoz, Chao Li, Lingjun Zhou, Chao Wu, Eva Czerwiec, Laurel Sandler, Andreas Seyfang, Maria A Blasco, and David L Keefe. Telomere lengthening early in development. *Nat Cell Biol.*, 9(12):1436–41, 2007.

[127] Xin Liu and Xiu-Jie Wang. Potential inhibitors for 2019-ncov coronavirus m protease from clinically approved medicines. 2020.

[128] Y. Liu, Z. Li, H. Xiong, X. Gao, and J. Wu. Understanding of internal clustering validation measures. In *2010 IEEE International Conference on Data Mining*, pages 911–916, 2010.

[129] B J Longley, M J Reguera, and Y Ma. Classes of c-kit activating mutations: proposed mechanisms of action and implications for disease classification and therapy. *Leuk Res.*, 25(7):571–6, 2001.

[130] Xiang-Jun Lu. The chi () torsion angle characterizes base/sugar relative orientation. `https://x3dna.org/highlights/the-chi-x-torsion-angle-characterizes-base-sugar-relative-orientation`.

[131] Karolina Kwapien Zeina Hobaika Richard G Maroun Jean-Philip Piquemal Léa El Khoury, Sehr Naseem-Khan and Nohad Gresh. Importance of explicit smeared lone-pairs in anisotropic polarizable molecular mechanics. torture track angular tests for exchange-repulsion and charge transfer contributions. *J Comput Chem.*, 38(22):1897–1920, 2017.

[132] G. J. Martyna M. Tuckermar and B. J. Berne. Reversible multiple time scale molecular dynamics. 97(3), 1992.

[133] D. H. J. Mackay, A. J. Cross, and A. T. Hagler. The role of energy minimization in simulation strategies of biomolecular systems. *Prediction of Protein Structure and the Principles of Protein Conformation*, pages 317–358, 1989.

[134] James A. Maier, Carmenza Martinez, Koushik Kasavajhala, Lauren Wickstrom, Kevin E. Hauser, and Carlos Simmerling. ff14sb: Improving the accuracy of protein side chain and backbone parameters from ff99sb. *J. Chem. Theory. Comput.*, 11(8):3696–3713, 2015.

[135] Laurent Salmon Marion Devillers, Jean-Philip Piquemal and Nohad Gresh. Calibration of the dianionic phosphate group: Validation on the recognition site of the homodimeric enzyme phosphoglucose isomerase. *J. Comput. Chem.*, 41(8):839–854, 2020.

[136] Paul S Masters. The molecular biology of coronaviruse. *Virus research*, 2006.

[137] Emanuela Ruggiero Rosalba Perrone Elena Tosoni Sara Lago Martina Tassinari-Giorgio Palù Matteo Scalabrin, Ilaria Frasson and Sara N Richter. The cellular protein hnrnp a2/b1 enhances hiv-1 transcription by unfolding ltr promoter g-quadruplexes. *Sci Rep.*, 24(7):45244, 2017.

[138] Zhi Wang Kevin Dalby Louis Lagardère Jean-Philip Piquemal Jay Ponder Matthew Harger, Daniel Li and Pengyu Ren. Tinker-openmm : Absolute and relative alchemical free energies using amoeba on gpus. *J Comput Chem.*, 38(23):2047–2055, 2017.

[139] J Andrew McCammon and Stephen C Harvey. *Dynamics of proteins and nucleic acids.* Cambridge University Press, 1988.

[140] Robert T. McGibbon, Kyle A. Beauchamp, Matthew P. Harrigan, Christoph Klein, Jason M. Swails, Carlos X. Hernández, Christian R. Schwantes, Lee-Ping Wang, Thomas J. Lane, and Vijay S. Pande. MDTraj: A modern open library for the analysis of molecular dynamics trajectories. *Biophys. J.*, 109(8):1528 – 1532, 2015.

[141] Michael Medinger, Manuela Kleinschmidt, Klaus Mross, Barbara Wehmeyer, Clemens Unger, Hans-Eckart Schaefer, Renate Weber, and Marc Azemar. c-kit (cd117) expression in human tumors and its prognostic value: an immunohistochemical analysis. *Pathol Oncol Res.*, 16(3):295–301, 2010.

[142] Josef Melcr and Jean-Philip Piquemal. Accurate biomolecular simulations account for electronic polarization. *Front. Mol. Biosci.*, 6:143, 2019.

[143] Vineet D Menachery, Rachel L Graham, and Ralph S Baric. Jumping species—a mechanism for coronavirus persistence and survival. *Preventive and therapeutic vaccines*, 2017.

[144] Yinglong Miao and J Andrew McCammonn. Gaussian accelerated molecular dynamics: Theory, implementation and applications. *Annual Reports in Computational Chemistry*, 13:231–278, 2017.

[145] Ze-Yi Yu Lu-Ni Hu Tian-Miao Ou Shuo-Bin Chen Zhi-Shu Huang Ming-Hao Hu, Yu-Qing Wang and Jia-Heng Tan. Discovery of a new four-leaf clover-like ligand as a potent c-myc transcription inhibitor specifically targeting the promoter g-quadruplex. *J Med Chem.*, 61(6):2447–2459, 2018.

[146] Alston J. Misquitta. Charge transfer from regularized symmetry-adapted perturbation theory. *. Chem. Theory Comput. 2013, 9, 12, 5313–5326*, 9(12):5313–5326, 2013.

[147] Alston J. Misquitta and Anthony Stone. CamCASP: a program for studying intermolecular interactions and for the calculation of molecular properties in distributed form. *University of Cambridge*, 2018.

[148] Alston J. Misquitta and Krzysztof Szalewicza. Symmetry-adapted perturbation-theory calculations of intermolecular forces employing density-functional description of monomers. *J. Chem. Phys.*, 112:214109, 2005.

[149] J Monod, J Wyman, and J-P Changeux. On the nature of allosteric transitions: a plausible model. *J. Mol. Biol.*, 12:88–118, 1965.

[150] J.N. Murrell and J.J.C. Teixeira-Dias. The dependence of exchange energy on orbital overlap. *Mol. Phys.*, 521:521–531, 1970.

[151] Anthony P Reszka Nancy H Campbell, Gary N Parkinson and Stephen Neidle. Structural basis of dna quadruplex recognition by an acridine drug. *J Am Chem Doc*, 130(21):6722–4, 2008.

[152] Yusuke Naritomi and Sotaro Fuchigami. Slow dynamics of a protein backbone in molecular dynamics simulation revealed by time-structure based independent component analysis. *J. Chem. Phys.*, 139:215102, 2013.

[153] Elodie Goldwaser Benoit De Courcy-Robin Chaudret David Perahia Christophe Narth-Louis Lagardere Filippo Lipparini Nohad Gresh, Krystel El Hage and Jean-Philip Piquemal. Addressing the issues of non-isotropy and non-additivity in the development of quantum chemistry-grounded polarizable molecular mechanics in quantum modeling of complex molecular systems. *Quantum Modeling of Complex Molecular Systems*, 21:1–49, 2015.

[154] Jean-Philip Piquemal Johanna Foret-Stéphanie Courtiol-Legourd Nohad Gresh, Benoit de Courcy and Laurent Salmon. Polarizable water networks in ligand–metalloprotein recognition. impact on the relative complexation energies of zn-dependent phosphomannose isomerase with d-mannose 6-phosphate surrogates. *J. Phys. Chem. B.*, 115(25):8304–8316, 2011.

[155] Jean-Philip Piquemal Johanna Foret-Stéphanie Courtiol-Legourd Nohad Gresh, Benoit de Courcy† and Laurent Salmon. Polarizable water networks in ligand–metalloprotein recognition. impact on the relative complexation energies of zn-dependent phosphomannose isomerase with d-mannose 6-phosphate surrogates. *J. Phys. Chem. B*, 115(25):8304–8316, 2011.

[156] Pierre Claverie Nohad Gresh and Alberte Pullman. Theoretical studies of molecular conformation. derivation of an additive procedure for the computation of intramolecular interaction energies. comparison with ab initio scf computations. *Theoret. Chim.*, 66:1–20, 1984.

[157] Young H. Yuh Norman L. Allinger and Jenn Huei Lii. Molecular mechanics. the mm3 force field for hydrocarbons. 1. *J. Am. Chem. Soc.*, 111(23):8551–8566, 1989.

[158] Frank Noé and Feliks Nüske. A variational approach to modeling slow processes in stochastic dynamical systems. 11:635–655, 2013.

[159] Feliks Nüske, Bettina G Keller, Guillermo Pérez-Hernández, Antonia S J S Mey, and Frank Noé. Variational approach to molecular kinetics. 8:1739–52, 2014.

[160] Wilma K. Olson. Nucleic acid structural principles. `https://casegroup.rutgers.edu/lnotes/BioPhysChem_week5.pdf`.

[161] M J Packer and C A Hunter. Sequence-dependent dna structure: the role of the sugar-phosphate backbone. *J Mol Biol.*, 17(280):407–20, 1998.

[162] Gábor Paragi and Célia FonsecaGuerra. Cooperativity in the self-assembly of the guanine nucleobase into quartet and ribbon structures on surfaces. *Chem. Eur. J.*, 23(13):3042–3050, 2017.

[163] Robert S. Paton and Jonathan M. Goodman. Hydrogen bonding and -stacking: How reliable are force fields? a critical evaluation of force field descriptions of nonbonded interactions. *J. Chem. Inf. Model.*, 49(4):944–955, 2009. PMID: 19309094.

[164] Fabian Pedregosa, Gael Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Edouard Duchesnay. Scikit-Learn: Machine learning in python. *J. Mach. Learn. Res.*, 12(null):2825–2830, November 2011.

[165] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-Learn: Machine learning in python. *J. Mach. Learn. Res.*, 12(null):2825–2830, 2011.

[166] Stanley Perlman and Jason Netland. Coronaviruses post-sars: update on replication and pathogenesis. *Nat Rev Microbiol.*, 2009.

[167] Anh Tuân Phan, Vitaly Kuryavyi, Sarah Burge, Stephen Neidle, and Dinshaw J. Patel. Structure of an unprecedented g-quadruplex scaffold in the human c-kit promoter. 129(14):4386–4392.

[168] Stefano Piana, Alexander G. Donchev, Paul Robustelli, and David E. Shaw. Water dispersion interactions strongly influence simulated structural properties of disordered protein states. *J. Phys. Chem. B.*, 119(16):5113–5123, 2015.

[169] Stefano Piana, Paul Robustelli, Dazhi Tan, Songela Chen, and David E. Shaw. Development of a force field for the simulation of single-chain proteins and protein–protein complexes. *J. Chem. Theory. Comput.*, 16(4):2494–2507, 2020.

[170] Thanigaimalai Pillaiyar, Manoj Manickam, Vigneshwaran Namasivayam, Yoshio Hayashi, and Sang-Hun Jung. An overview of severe acute respiratory syndrome–coronavirus (sars-cov) 3cl protease inhibitors: Peptidomimetics and small molecule chemotherapy. *J. Med. Chem.*, 59(14):6595–6628, 2016.

[171] Krzysztof Pluta and Magdalena Marta Kacprzak. Use of hiv as a gene transfer vector. *Acta Biochim Pol.*, 56(4):531–95, 2009.

[172] Jay W. Ponder, Chuanjie Wu, Pengyu Ren, Vijay S. Pande, John D. Chodera, Michael J. Schnieders, Imran Haque, David L. Mobley, Daniel S. Lambrecht, Robert A. DiStasio, Martin Head-Gordon, Gary N. I. Clark, Margaret E. Johnson, and Teresa Head-Gordon. Current status of the amoeba polarizable force field. *The Journal of Physical Chemistry B*, 114(8):2549–2564, 2010. PMID: 20136072.

[173] Guillermo Pérez-Hernández, Fabian Paul, Toni Giorgino, Gianni De Fabritiis, and Frank Noé. Identification of slow molecular order parameters for markov model construction. 139, 2013.

[174] RA Laskowski RA, F Gerick, and JM Thornton. The structural basis of allosteric regulation in proteins. *FEBS. Lett.*, 583(11):1682–8, 2009.

[175] M J Raamsman, J K Locker, A de Hooge, A A de Vries, G Griffiths, H Vennema, and P J Rottier. Characterization of the coronavirus mouse hepatitis virus strain a59 small membrane protein e. *J Virol.*, 2000.

[176] Joshua A. Rackers, Zhi Wang, Chao Lu, Marie L. Laury, Louis Lagardère, Michael J. Schnieders, Jean-Philip Piquemal, Pengyu Ren, and Jay W. Ponder. Tinker 8: Software tools for molecular design. *J. Chem. Theory and Comput.*, 14(10):5273–5289, 2018.

[177] Sarah Rankin, Anthony P Reszka, Julian Huppert, Mire Zloh, Gary N Parkinson, Alan K Todd, Sylvain Ladame, Shankar Balasubramanian, and Stephen Neidle. Putative dna quadruplex formation within the human c-kit oncogene. *J Am Chem Soc.*, 127(30):10584–9, 2005.

[178] THOMAS D. RASMUSSEN, PENGYU REN, JAY W. PONDER, and FRANK JENSEN. Force field modeling of conformational energies: Importance of multipole moments and intramolecular polarization. *International Journal of Quantum Chemistry*, 107:1390 –1395, 2007.

[179] Pengyu Ren and Jay W. Ponder. Polarizable atomic multipole water model for molecular mechanics simulation. *J. Phys. Chem. B.*, 107(24):5933–5947, 2003.

[180] Pengyu Y. Ren and Jay W. Ponder. "polarizable atomic multipole water model for molecular mechanics simulation". *J. Phys. Chem.*, 107(24):5933–5947, 2003.

[181] Daniel G A Smith-Andrew C Simmonett A Eugene DePrince 3rd Edward G Hohenstein Uğur Bozkaya Alexander Yu Sokolov Roberto Di Remigio Ryan M Richard Jérôme F Gonthier Andrew M James Harley R McAlexander Ashutosh Kumar Masaaki Saitow Xiao Wang Benjamin P Pritchard Prakash Verma Henry F Schaefer 3rd Konrad Patkowski Rollin A King Edward F Valeev Francesco A Evangelista Justin M Turney T Daniel Crawford Robert M Parrish, Lori A Burns and C David Sherrill. Psi4 1.1: An open-source electronic structure program emphasizing automation, advanced libraries, and interoperability. *J Chem Theory Comput.*, 13(7):3185–3197, 2017.

[182] Dirk Daelemans Giorgio Palù Christophe Pannecouque Rosalba Perrone, Elena Butovskaya and Sara N Richter. Anti-hiv-1 activity of the g-quadruplex ligand braco-19. *J Antimicrob Chemother.*, 69(12):3248–58, 2014.

[183] Ilaria Frasson Jerrod A Poe Elena Butovskaya Thomas E Smithgall Manlio Palumbo Giorgio Palù Rosalba Perrone, Matteo Nadai and Sara N Richter. A dynamic g-quadruplex region regulates the hiv-1 long terminal repeat promoter. *J Med Chem.*, 56(16):6521–30, 2013.

[184] Gustave Roussy. Soutenez les projets de recherche covid-19 de gustave roussy.

[185] Eldon E Baird Ryan E Bremer and Peter B Dervan. Research paper 119 inhibition of major-groove-binding proteins by pyrrole-imidazole polyamides with anhrg-pro-arg positive patch. *Chemistry Biology*, 5:119–133, 1998.

[186] Chen S, Jonas F, Shen C, and Hilgenfeld R. Liberation of sars-cov main protease from the viral polyprotein: N-terminal autocleavage does not depend on the mature dimerization mode. *Protein Cell*, 1:59–74, 2010.

[187] W. G. Marshall S. A. Moggach and S. Parsons. High-pressure neutron diffraction study of l-serine-i and l-serine-ii, and the structure of l-serine-iii at 8.1 gpa. *Acta Crystallographica Section B*, 62:815–825, 2006.

[188] F Xue X Kang X Ren J Chen C Jin Z Lou S Zhang, N Zhong and B Xia. Three-dimensional domain swapping as a mechanism to lock the active conformation in a super-active octamer of sars-cov main protease. *Protein Cell*, 1:371–383, 2010.

[189] Giuseppina La Sala, Sergio Decherchi, and Marco De Vivo andand Walter Rocchia. Allosteric communication networks in proteins revealed through pocket crosstalk analysis. *ACS. Cent. Sci.*, 3:949–960, 2017.

[190] Alexa M. Salsbury and Justin A. Lemkul. Molecular dynamics simulations of the c-kit1 promoter g-quadruplex: Importance of electronic polarization on stability and cooperative ion binding. 123(1):148–159.

[191] Amina Bedrat Brune Vialet Marie-Line Andreola Samir Amrane, Abdelaziz Kerkour and Jean-Louis Mergny. Topology of a dna g-quadruplex structure formed in the hiv-1 promoter: a potential target for anti-hiv drug development. *J Am Chem Soc.*, 136(14):5249–52, 2014.

[192] Peter Schmidtke and Xavier Barril. Understanding and predicting druggability. a high-throughput method for detection of drug binding sites. *J. Med. Chem.*, 53(15):5858–5867, 2010.

[193] Christian R. Schwantes and Vijay S. Pande. Improvements in markov state model construction reveal many non-native interactions in the folding of ntl9. 9:2000–2009, 2013.

[194] Yue Shi, Zhen Xia, Jiajing Zhang, Robert Best, Chuanjie Wu, Jay W. Ponder, and Pengyu Ren. Polarizable atomic multipole-based amoeba force field for proteins. *J. Chem. Theory. Comput.*, 9(9):4046–4063, 2013. PMID: 24163642.

[195] Gary N Parkinson Shozeb M Haider and Stephen Neidle. Structure of a g-quadruplex-ligand complex. *J Mol Biol*, 326(1):117–25, 2003.

[196] T. Simonsson. G-quadruplex DNA structures–variations on a theme. 382(4):621–628.

[197] Sona Sivakova and Stuart J Rowan. Nucleobases as supramolecular motifs. *Chem Soc Rev.*, 34(1):9–21, 2005.

[198] E J Snijder, E Decroly, and J Ziebuhr. The nonstructural proteins directing coronavirus rna synthesis and processing. *Adv Virus Res*, 2016.

[199] Isabel Sola, Fernando Almazán, Sonia Zúñiga, and Luis Enjuanes. Continuous and discontinuous rna synthesis in coronaviruses. *Annu Rev Virol*, 2015.

[200] Karmen Stankov, Stevan Popovic, and Momir Mikov. C-kit signaling in cancer treatment. *Curr Pharm Des.*, 20(17):2849–40, 2014.

[201] Stephan Ehrlich Stefan Grimme and Lars Goerigk. Effect of the damping function in dispersion corrected density functional theory. *J. Comput. Chem.*, 32(7):1456–1465, 2011.

[202] A. J. Stone and A. J. Misquitta. Charge-transfer in symmetry adapted perturbation theory. *Chem. Phys. Lett.*, 473:201–205, 2009.

[203] A.J. Stone. Computation of charge-transfer energies by perturbation theory. *Chem. Phys. Lett.*, 211(1):101–109, 1993.

[204] Anthony J. Stone. Distributed multipole analysis: Stability for large basis sets. *J. Chem. Theory Comput.*, 1(6):1128–1132, 2005.

[205] Anthony J Stone. Electrostatic damping functions and the penetration energy. *J. Phys. Chem.*, 115(25):7017–27, 2011.

[206] Léonie Strömich, Nan Wu, Mauricio Barahona, and Sophia N. Yaliraki. Allosteric hotspots in the main protease of sars-cov-2. *bioRxiv*, 2020.

[207] Krzysztof Szalewicz. Symmetry-adapted perturbation theory of intermolecular forces. *Comput. Mol. Sci.*, 2:254–272, 2012.

[208] Terra Sztain, Rommie Amaro, and J. Andrew McCammon. Elucidation of cryptic and allosteric pockets within the sars-cov-2 protease. *bioRxiv*, page 2020.07.23.218784, 2020.

[209] Muramatsu T, Kim YT, Nishii W, Terada T, Shirouzu M, and Yokoyama S. Autoprocessing mechanism of severe acute respiratory syndrome coronavirus 3c-like protease (sars-cov 3clpro) from its polyproteins. *FEBS J*, 280:2002–2013, 2013.

[210] Jinzhi Tan, Koen H.G. Verschueren, Kanchan Anand, Jianhua Shen, Maojun Yang, Yechun Xu, Zihe Rao, Janna Bigalke, Burkhard Heisen, Jeroen R. Mesters, Kaixian Chen, Xu Shen, Hualiang Jiang, and Rolf Hilgenfeld. ph-dependent conformational flexibility of the sars-cov main proteinase (mpro) dimer: Molecular dynamics simulations and multiple x-ray structure analyses. *J. Mol. Biol.*, 354(1):25 – 40, 2005.

[211] Clifton Ming Tay, Chee Wee Ong, Victor Kwan Min Lee, and Brendan Pang. Kit gene mutation analysis in solid tumours: biology, clincial applications and trends in diagnostic reporting. *Pathology.*, 16(3):295–301, 2010.

[212] Alan K Todd, Matthew Johnston, and Stephen Neidle. Highly prevalent putative quadruplex sequence motifs in human dna. 33:2901–7, 2005.

[213] Randall Tressler and Catherine Godfrey. Nrti backbone in hiv treatment: will it remain relevant? *Drugs.*, 72(16):2051–62, 2012.

[214] Makoto Ujike and Fumihiro Taguchi. Incorporation of spike and membrane glycoproteins into coronavirus virions. *Viruses.*, 2015.

[215] Sven Ullrich and Christoph Nitsche. The sars-cov-2 main protease as drug target. *Bioorg. Med. Chem. Lett.*, 2020.

[216] Kenji Usui, Arisa Okada, Shungo Sakashita, Masayuki Shimooka, Takaaki Tsuruoka, Shu ichi Nakano, Daisuke Miyoshi, Tsukasa Mashima, Masato Katahira, and Yoshio Hamada. Dna g-wire formation using an artificial peptide is controlled by protease activity. *Molecules*, 22(1991), 2017.

[217] Neha Verma, Jack A. Henderson, and Jana Shen. Proton-coupled conformational activation of sars coronavirus main proteases and opportunity for designing small-molecule broad-spectrum targeted covalent inhibitors. *J. Am. Chem. Soc.*, 142(52):21883–21890, 2020.

[218] Fabienne Vigné-Maeder and Pierre Claverie. The exact multicenter multipolar part of a molecular charge distribution and its simplified representations. *J. Chem. Phys.*, 88(8):4934, 1988.

[219] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, CJ Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake Vand erPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nat. Methods.*, 17:261–272, 2020.

[220] Andrea Volkamer, Daniel Kuhn, Friedrich Rippmann, and Matthias Rarey. Dogsitescorer: a web server for automatic binding site prediction, analysis and druggability assessment. *Bioinformatics.*, 28(15):2074–2075, 2012.

[221] Jeffrey R. Wagner, Jesper Sørensen, Nathan Hensley, Celia Wong, Clare Zhu, Taylor Perison, and Rommie E. Amaro. POVME 3.0: Software for mapping binding pocket flexibility. *J. Chem. Theory. Comput.*, 13(9):4584–4592, 2017.

[222] Masayuki Tera Keisuke Iida-Kazuo Nagasawa Wan Jun Chung, Brahim Heddi and Anh Tuân Phan. Solution structure of an intramolecular (3 + 1) human telomeric g-quadruplex bound to a telomestatin derivative. *J. Am. Chem. Soc.*, 135(36):13495–13501, 2013.

[223] A Warshel and M Levitt. Theoretical studies of enzymic reactions: dielectric, electrostatic and steric stabilization of the carbonium ion in the reaction of lysozyme. *J Mol Biol.*, 103(2):227–49, 1976.

[224] J D WATSON and F H CRICK. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature*, 171(4356):737–8, 1953.

[225] Dengguo Wei, Jarmila Husby, and Stephen Neidle. Flexibility and structural conservation in a c-kit g-quadruplex. 43(1):629–644.

[226] Dengguo Wei, Jarmila Husby, and Stephen Neidle. Flexibility and structural conservation in a c-kit g-quadruplex. *Nucleic Acids Res.*, 43(1):629–44, 2015.

[227] Dengguo Wei, Gary N. Parkinson, Anthony P. Reszka, and Stephen Neidle. Crystal structure of a c-kit promoter quadruplex reveals the structural role of metal ions and water molecules in maintaining loop conformation. 40(10):4691–4700.

[228] Ping Wei, Keqiang Fan, Hao Chen, Liang Ma, Changkang Huang, Lei Tan, Dong Xi, Chunmei Li, Ying Liu, Aoneng Cao, and Luhua Lai. The n-terminal octapeptide acts as a dimerization inhibitor of sars coronavirus 3c-like proteinase. *Biochem. Biophys. Res. Commun.*, 20;339(3):865-72, 2006.

[229] Xue X, Yang H, Shen W, Zhao Q, Li J, Yang K, Chen C, Jin Y, Bartlam M, and Rao Z. Production of authentic sars-cov mpro with enhanced activity: application as a novel tag-cleavage endopeptidase for protein overproduction. *J Mol Biol*, 266:965–975, 2007.

[230] Haitao Yang, Maojun Yang, Yi Ding, Yiwei Liu, Zhiyong Lou, Zhe Zhou, Lei Sun, Lijuan Mo, Sheng Ye, Hai Pang, et al. The crystal structures of severe acute respiratory syndrome virus main protease and its complex with an inhibitor. *Proc. Nat. Acad. Sci. USA.*, 100(23):13190–13195, 2003.

[231] Haitao Yang, Maojun Yang, Yi Ding, Yiwei Liu, Zhiyong Lou, Zhe Zhou, Lei Sun, Lijuan Mo, Sheng Ye, Hai Pang, George F. Gao, Kanchan Anand, Mark Bartlam, Rolf Hilgenfeld, and Zihe Rao. The crystal structures of severe acute respiratory syndrome virus main protease and its complex with an inhibitor. *Proceedings of the National Academy of Sciences*, 100(23):13190–13195, 2003.

[232] Klaus Schulten Yi Wang, Christopher B. Harrison and J. Andrew McCammon. Implementation of accelerated molecular dynamics in namd. *Comput Sci Discov.*, 4(1), 2011.

[233] Y Yogo, N Hirano, S Hino, H Shibuta, and M Matumoto. Polyadenylate in the virion rna of mouse hepatitis virus. *J. Biochem.*

[234] Ting Zhang Fang Li Wensheng Yang Rafal Kaminski Philip Regis Fagan Raj Putatunda Won-Bin Young Kamel Khalili Yonggang Zhang, Chaoran Yin and Wenhui Hu. Crispr/grna-directed synergistic activation mediator (sam) induces specific, persistent and robust reactivation of the hiv-1 latent reservoirs. *Sci Rep.*, 5:16277, 2015.

[235] Changsheng Zhang, Chao Lu, Zhifeng Jing, Chuanjie Wu, Jean-Philip Piquemal, Jay W. Ponder, and Pengyu Ren. Amoeba polarizable atomic multipole force field for nucleic acids. 14(4):2084–2108.

[236] Changsheng Zhang, Chao Lu, Zhifeng Jing, Chuanjie Wu, Jean-Philip Piquemal, Jay W. Ponder, and Pengyu Ren. AMOEBA polarizable atomic multipole force field for nucleic acids. 14(4):2084–2108.

[237] Changsheng Zhang, Chao Lu, Zhifeng Jing, Chuanjie Wu, Jean-Philip Piquemal, Jay W. Ponder, and Pengyu Ren. AMOEBA polarizable atomic multipole force field for nucleic acids. *J. Chem. Theory. Comput.*, 14(4):2084–2108, 2018.

[238] Changsheng Zhang, Chao Lu, Zhifeng Jing, Chuanjie Wu, Jean-Philip Piquemal, Jay W. Ponder, and Pengyu Ren. Amoeba polarizable atomic multipole force field for nucleic acids. *J. Chem. Theory. Comput.*, 14(4):2084–2108, 2018.

[239] Linlin Zhang, Daizong Lin, Xinyuanyuan Sun, Ute Curth, Christian Drosten, Lucie Sauerhering, Stephan Becker, Katharina Rox, and Rolf Hilgenfeld. Crystal structure of sars-cov-2 main protease provides a basis for design of improved -ketoamide inhibitors. *Science.*, 368(6489):409–412, 2020.

[240] Linlin Zhang, Daizong Lin, Xinyuanyuan Sun, Ute Curth, Christian Drosten, Lucie Sauerhering, Stephan Becker, Katharina Rox, and Rolf Hilgenfeld. Crystal structure of SARS-CoV-2 main protease provides a basis for design of improved α-ketoamide inhibitors. *Science*, 368(6489):409–412, 2020.

[241] Linlin Zhang, Daizong Lin, Xinyuanyuan Sun, Ute Curth, Christian Drosten, Lucie Sauerhering, Stephan Becker, Katharina Rox, and Rolf Hilgenfeld. Crystal structure of SARS-CoV-2 main protease provides a basis for design of improved α-ketoamide inhibitors. *Science.*, 368(6489):409–412, 2020.

[242] Yuan Chu Zhengtong Lv and Yong Wang. Hiv protease inhibitors: a review of molecular selectivity and toxicity. *HIV AIDS (Auckl).*, 8(7):95–104, 2015.

[243] Xuelan Zhou, Fanglin Zhong, Cheng Lin, Xiaohui Hu, Yan Zhang, Bing Xiong, Xiushan Yin, Jinheng Fu, Wei He, Jingjing Duan, et al. Structure of SARS-CoV-2 main protease in the apo state. *Sci. China Life. Sci.*, pages 1–4, 2020.

[244] Maxwell I. Zimmerman, Justin R. Porter, Xianqiang Sun, Roseane R. Silva, and Gregory R. Bowman. Choice of adaptive sampling strategy impacts state discovery, transition probabilities, and the apparent mechanism of conformational changes. *J. Chem. Theory. Comput.*, 14(11):5459–5475, 2018.

[245] A. Zumla, J. F. Chan, E. I. Azhar, D. S. Hui, and K.-Y. Yuen. Coronaviruses-drug discovery and therapeutic options. *Nat. Rev. Drug Discovery*, 15:328–347, 2016.

[246] Reena Zutshi, Michelle Brickner, and Jean Chmielewski. Inhibiting the assembly of protein-protein interfaces. *Curr. Opin. Chem. Biol.*, 2(1):62–66, 1998.

[247] Katarzyna Świderek and Vicent Moliner. Revealing the molecular mechanisms of proteolysis of sars-cov-2 mpro by QM/MM computational methods. *Chem. Sci.*, 11(39):10541–10852, 2020.

Chapitre 6

Annexes

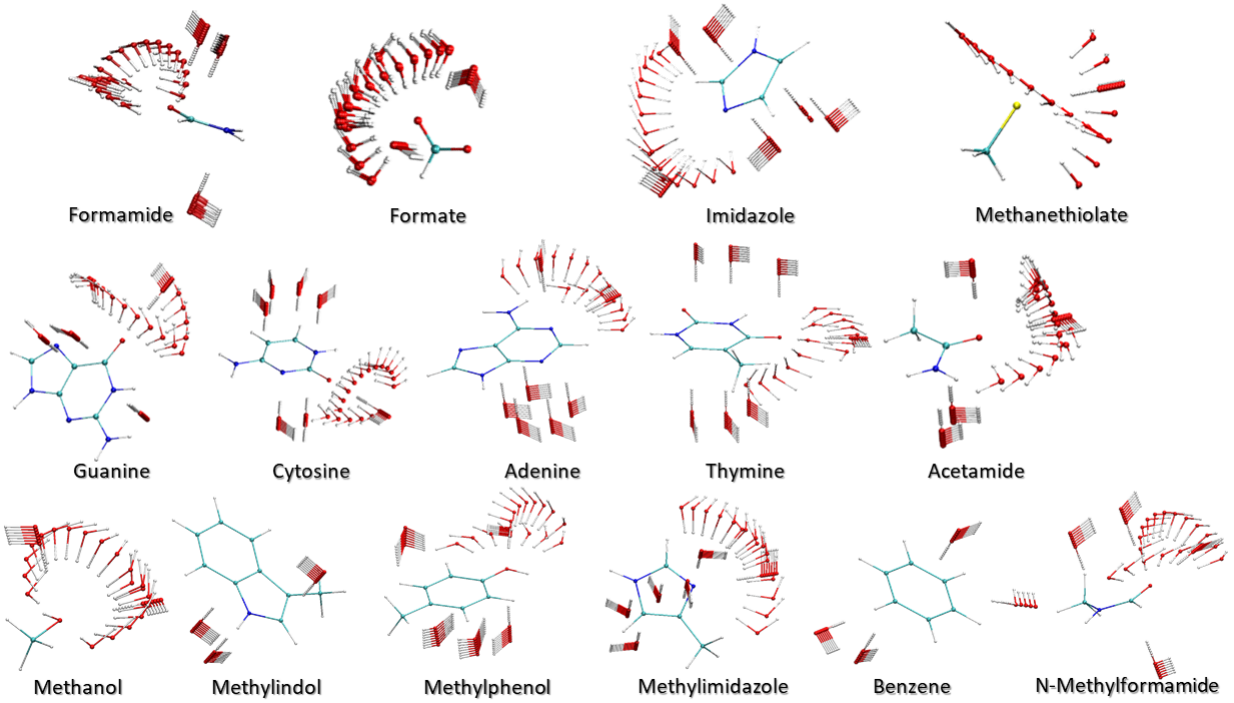


Figure 6.1: Représentation des 16 fragments protéiques et nucleiques. Chaque fragment est 'sondé' par une sonde dicationique, Zn^{2+}, et une sonde dipolaire, H_2O dans et hors plan.

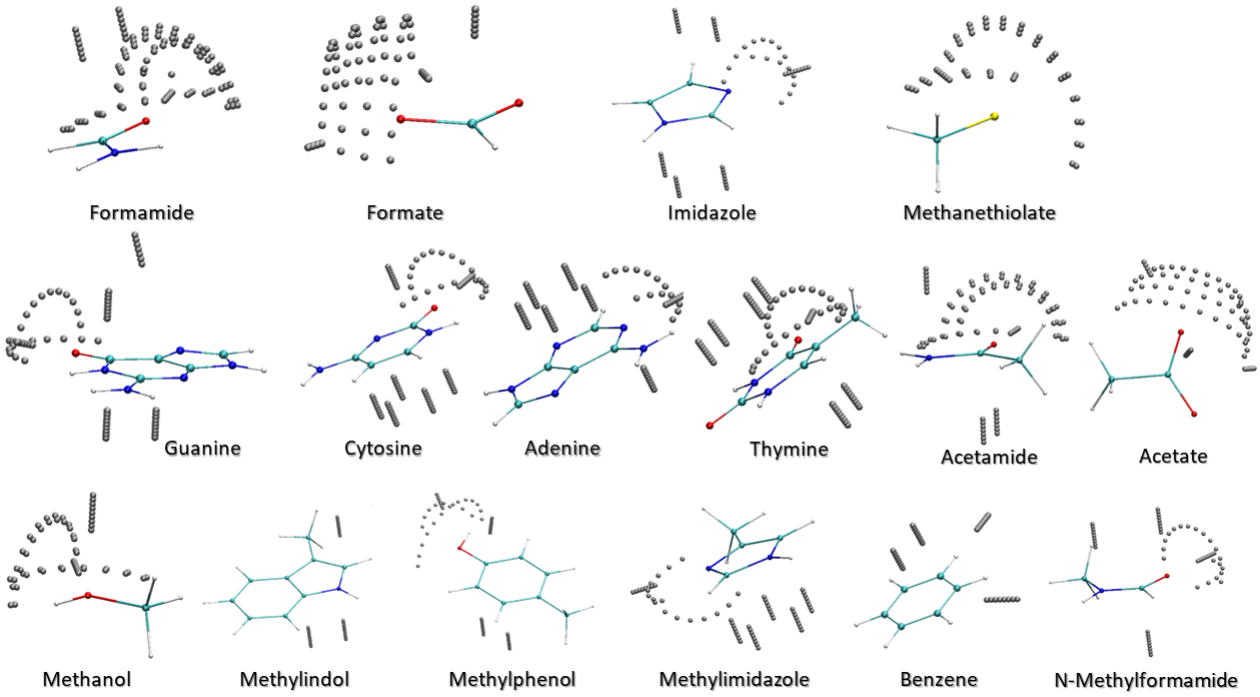


Figure 6.2: Représentation des 16 fragments appartenant aux protéine et à l'ADN. Chaque fragment est sondé par un ion cationique, Zn^{Zn2+} dans et hors du plan.

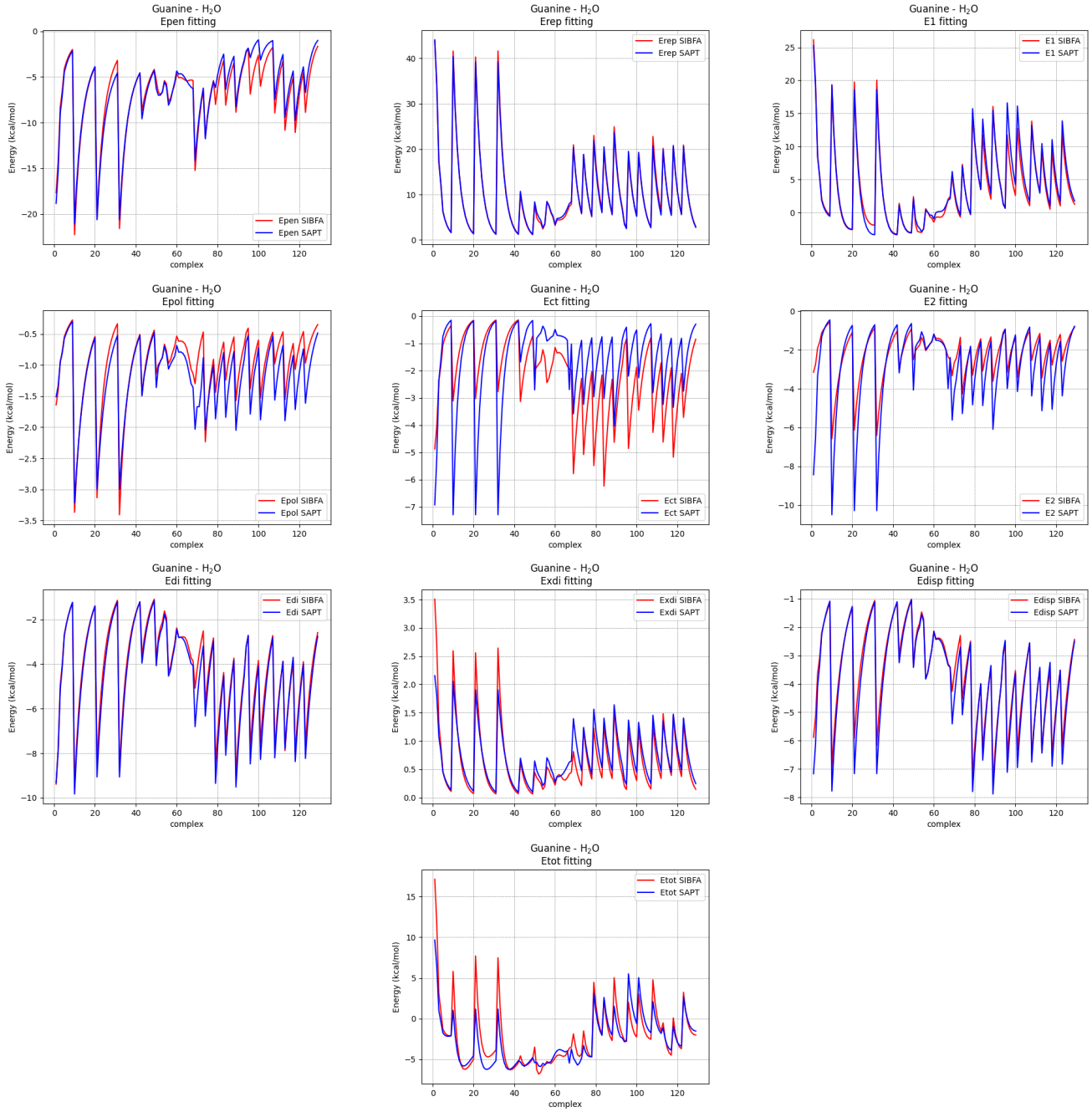


Figure 6.3: Guanine - H_2O SIBFA fitting on SAPT(DFT)

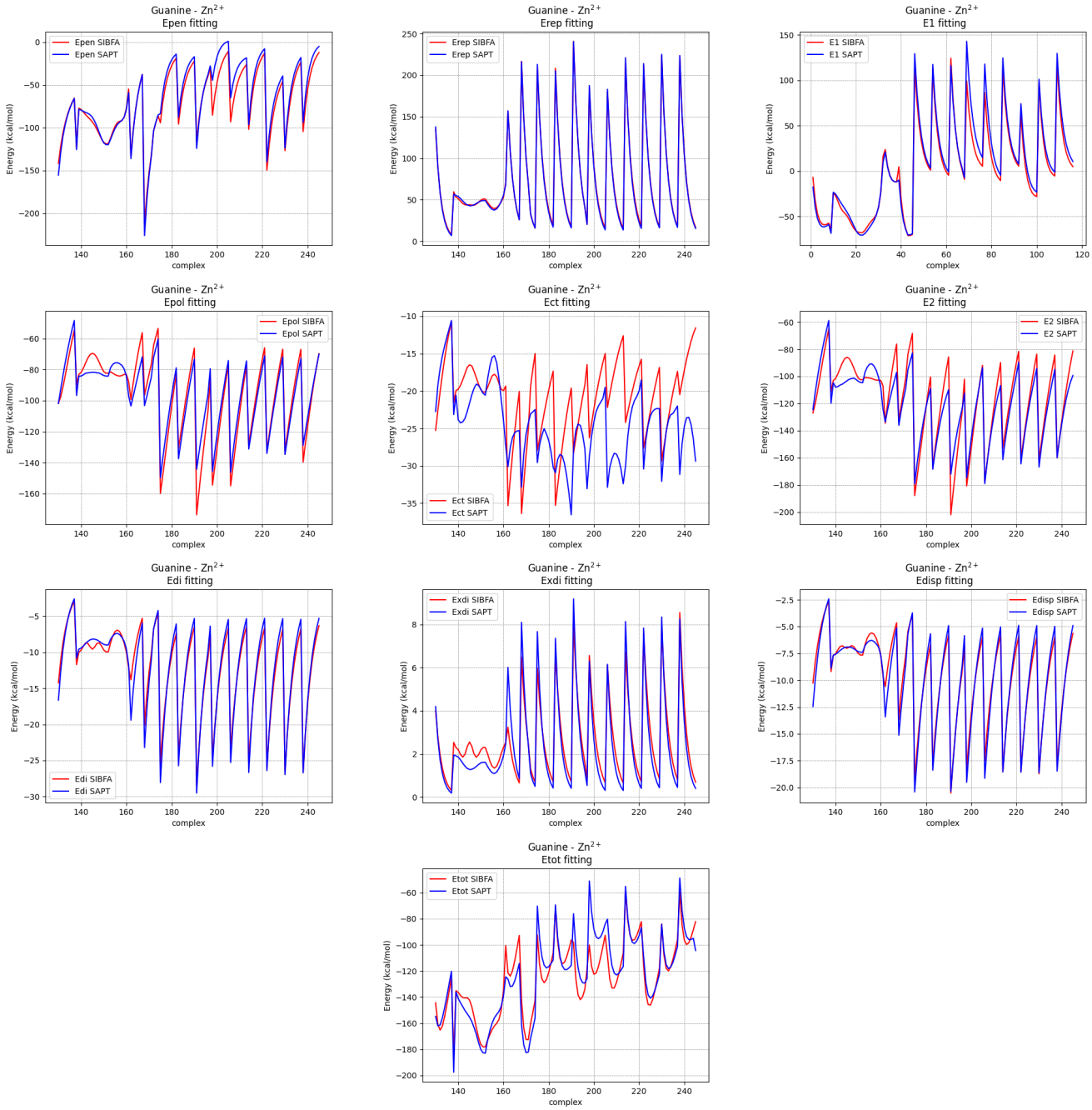


Figure 6.4: Guanine - Zn^{2+} SIBFA fitting on SAPT(DFT)

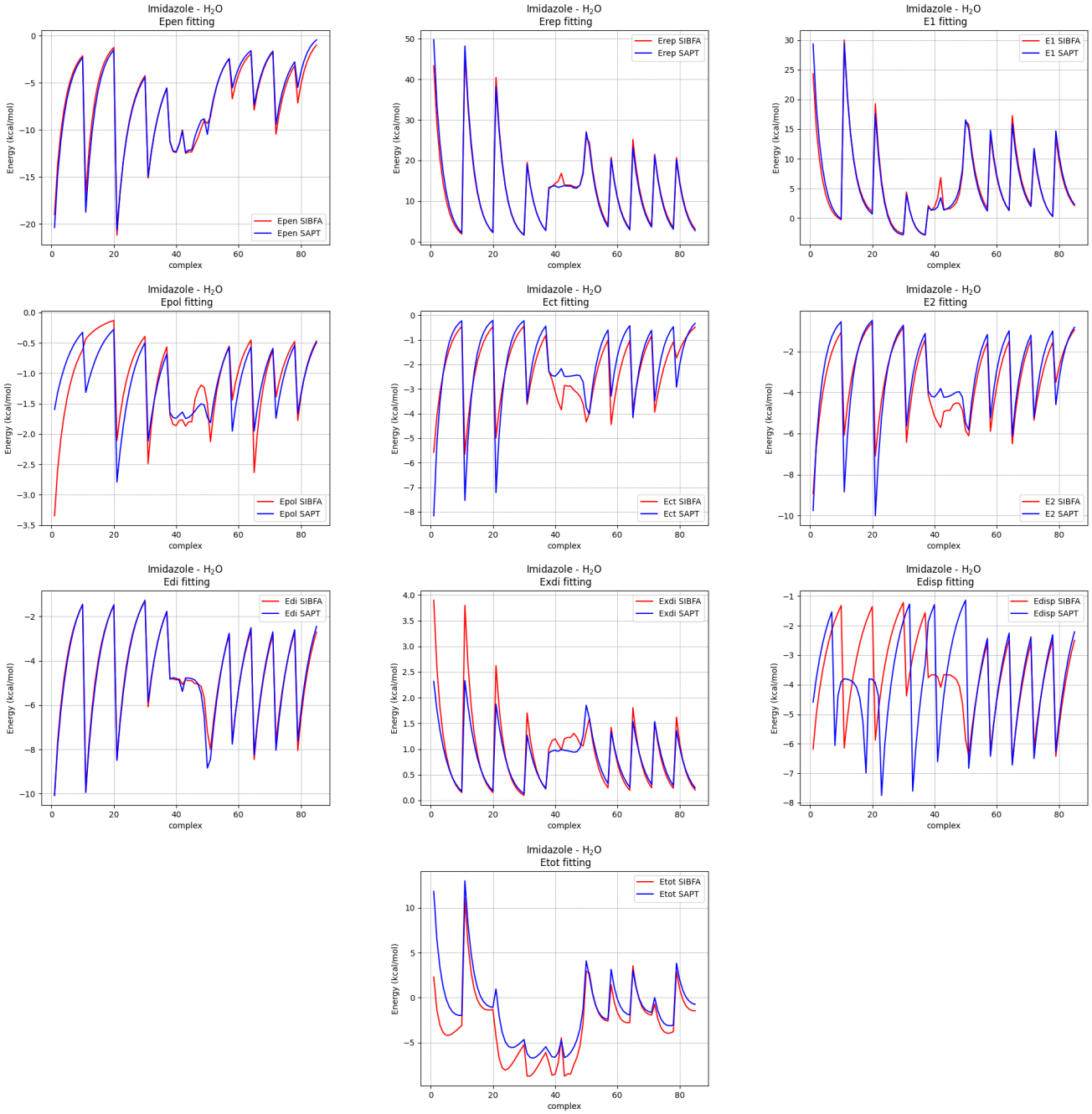


Figure 6.5: Imidazole - H_2O SIBFA fitting on SAPT(DFT)

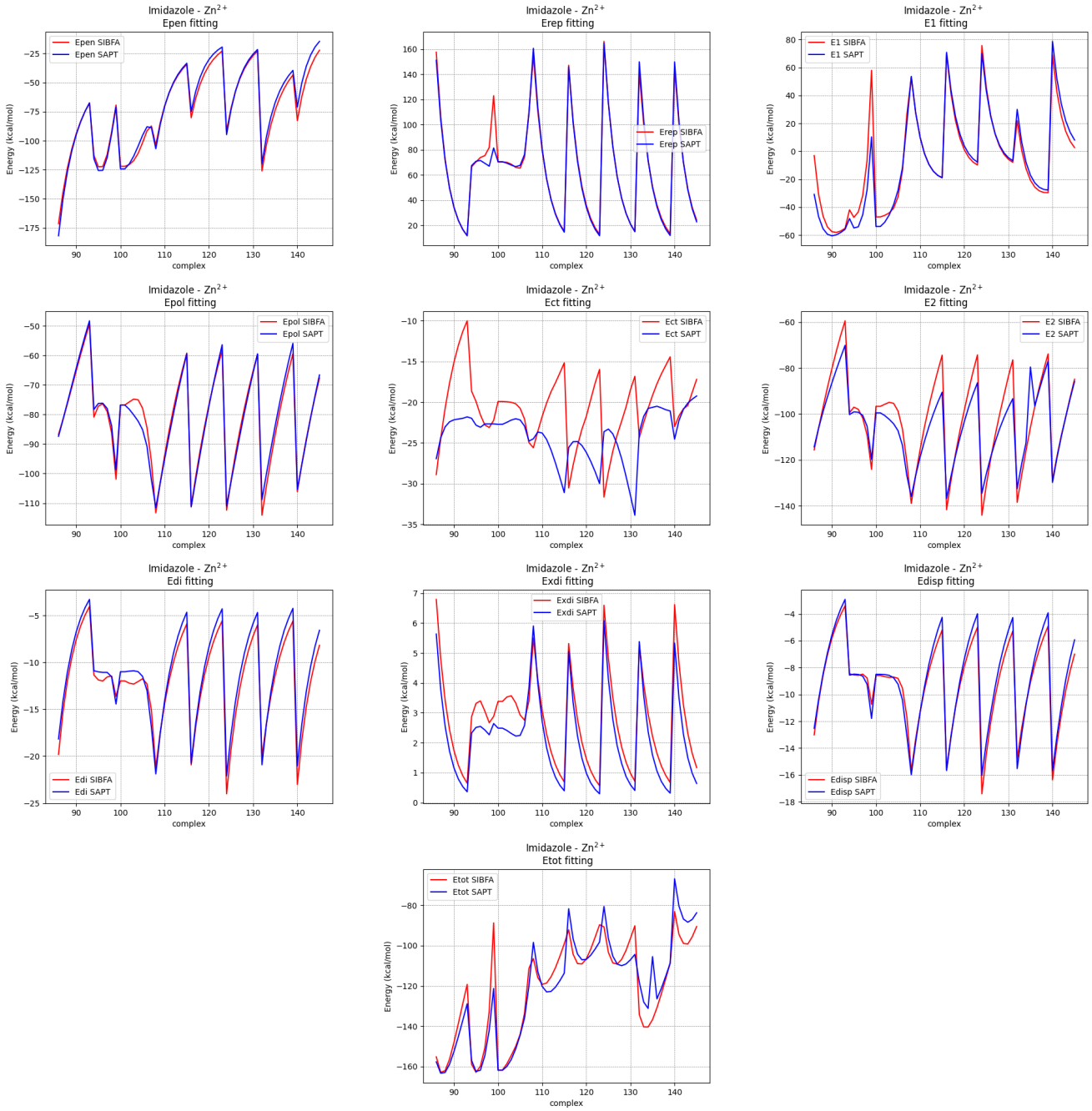


Figure 6.6: Imidazole - Zn^{2+} SIBFA fitting on SAPT(DFT)

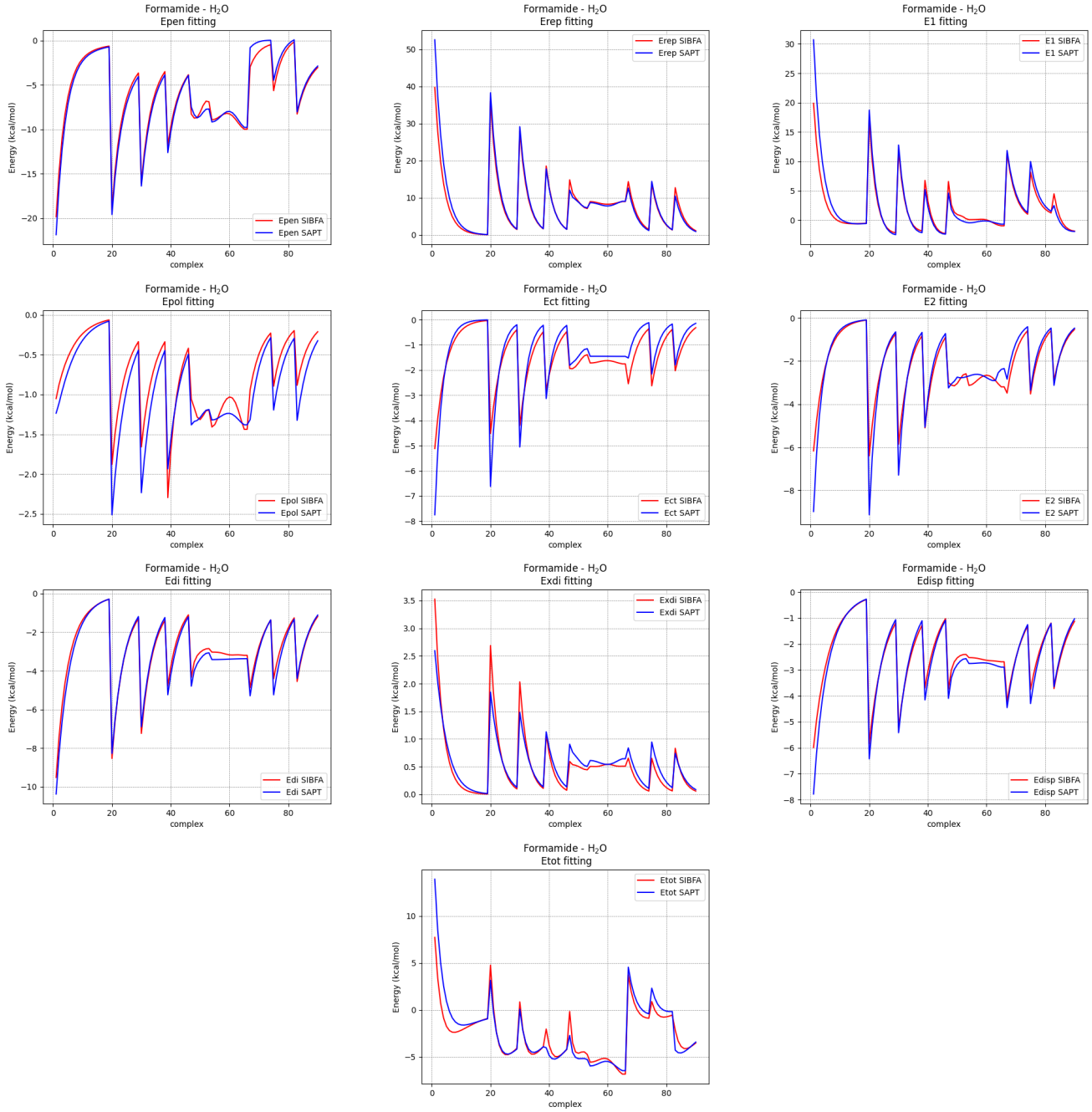


Figure 6.7: Formamide - H_2O SIBFA fitting on SAPT(DFT)

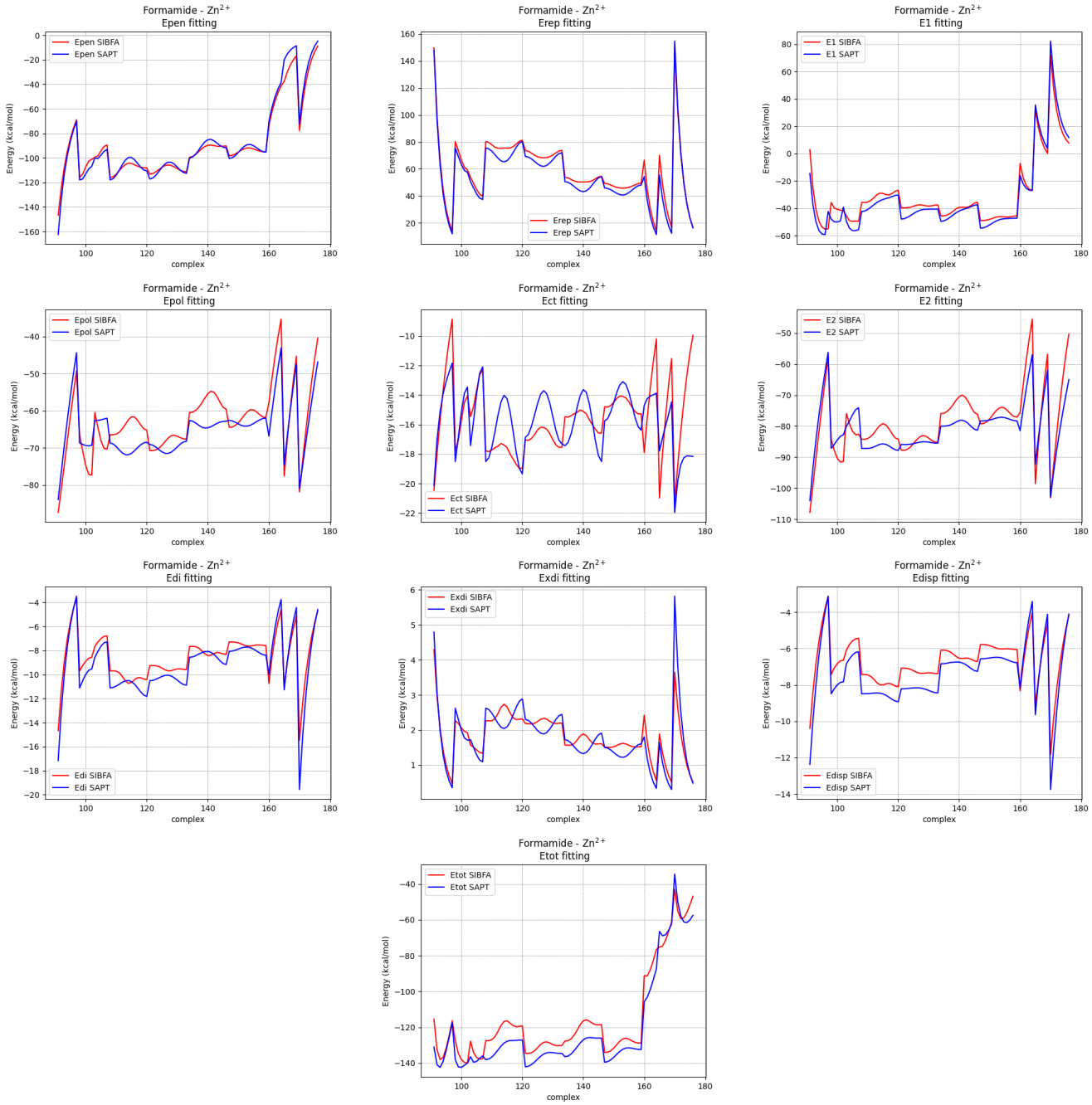


Figure 6.8: Formamide - Zn^{2+} SIBFA fitting on SAPT(DFT)

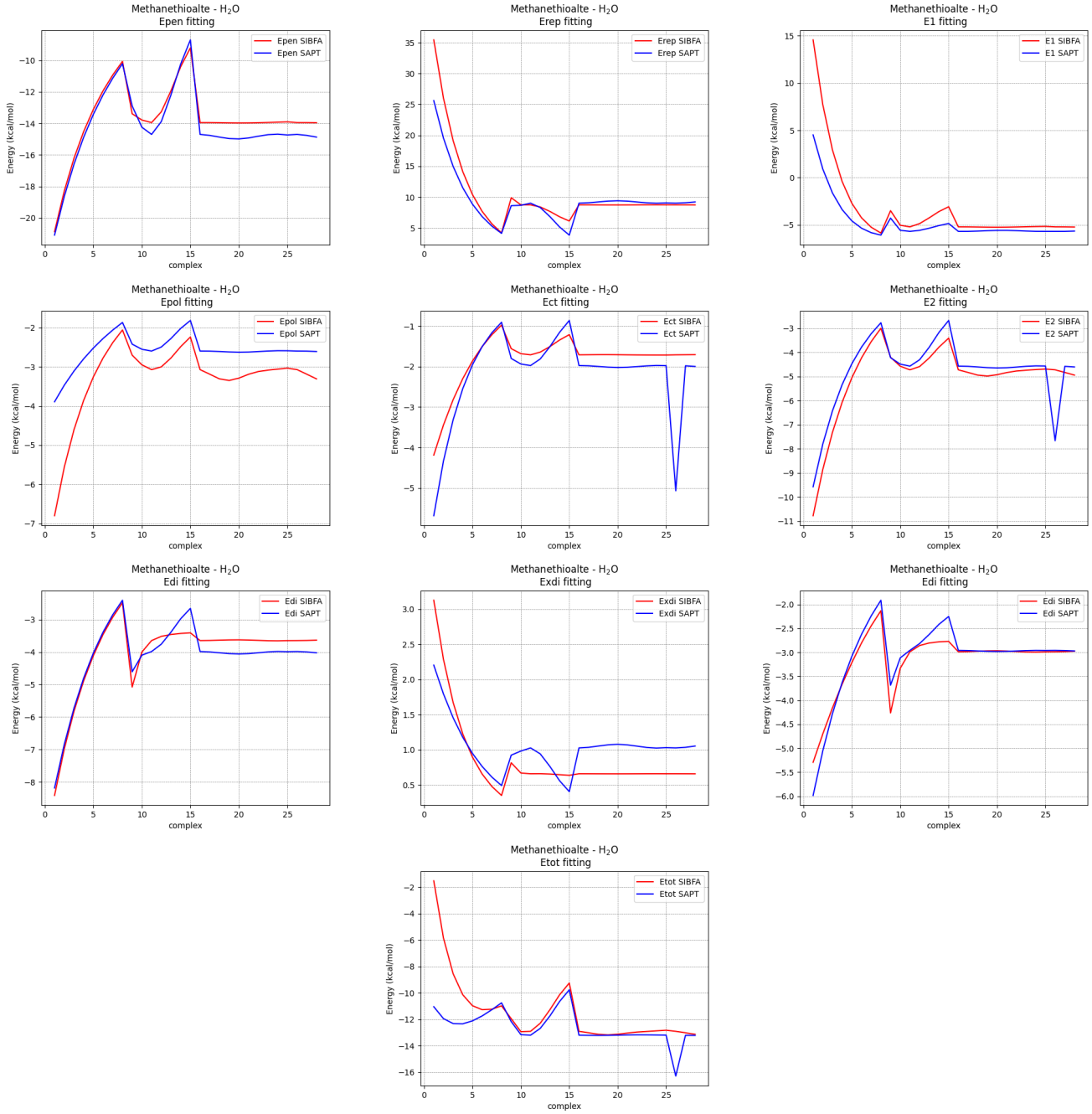


Figure 6.9: Methanethiolate - H_2O SIBFA fitting on SAPT(DFT)

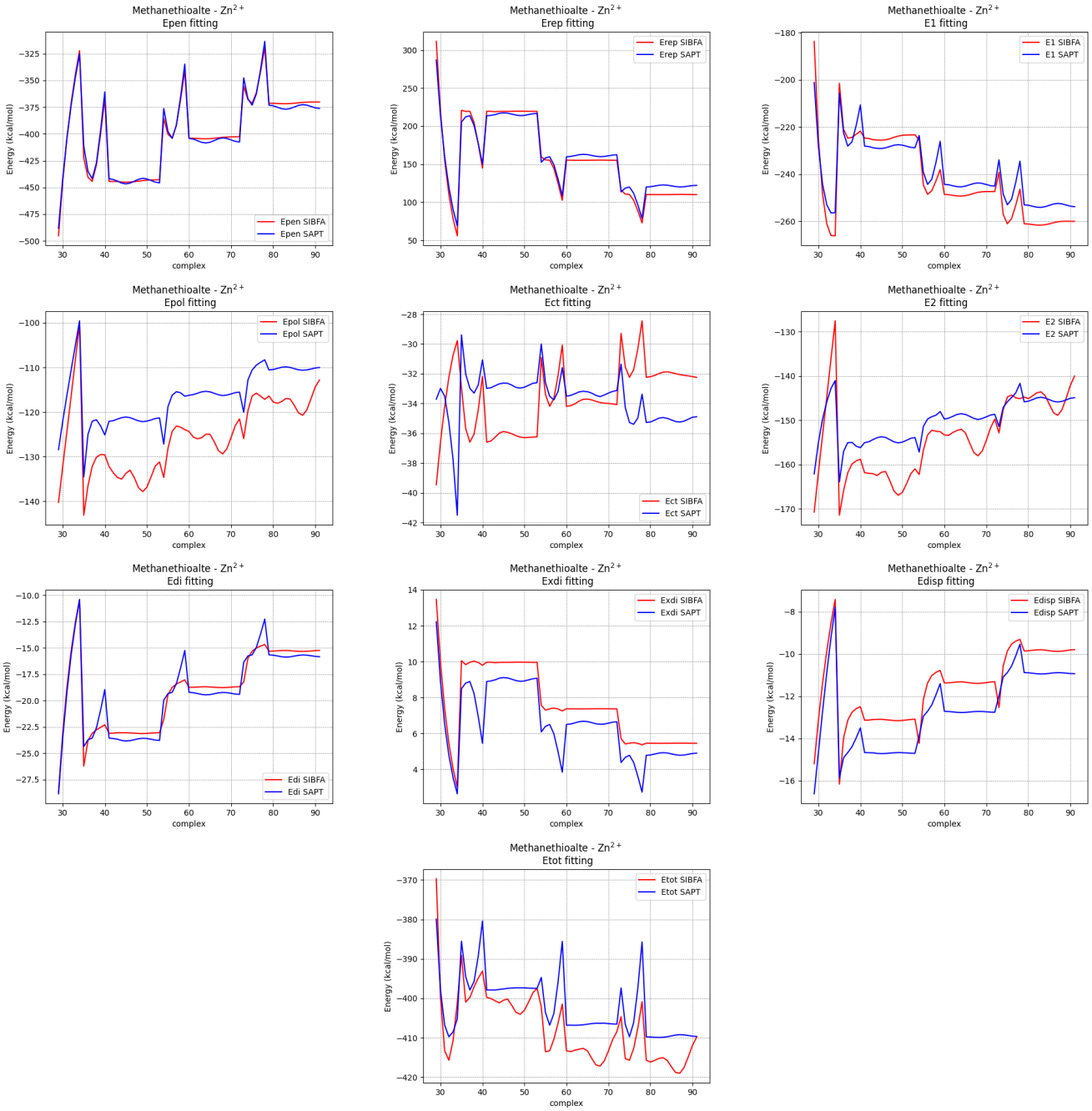


Figure 6.10: Methanethiolate - Zn^{2+} SIBFA fitting on SAPT(DFT)

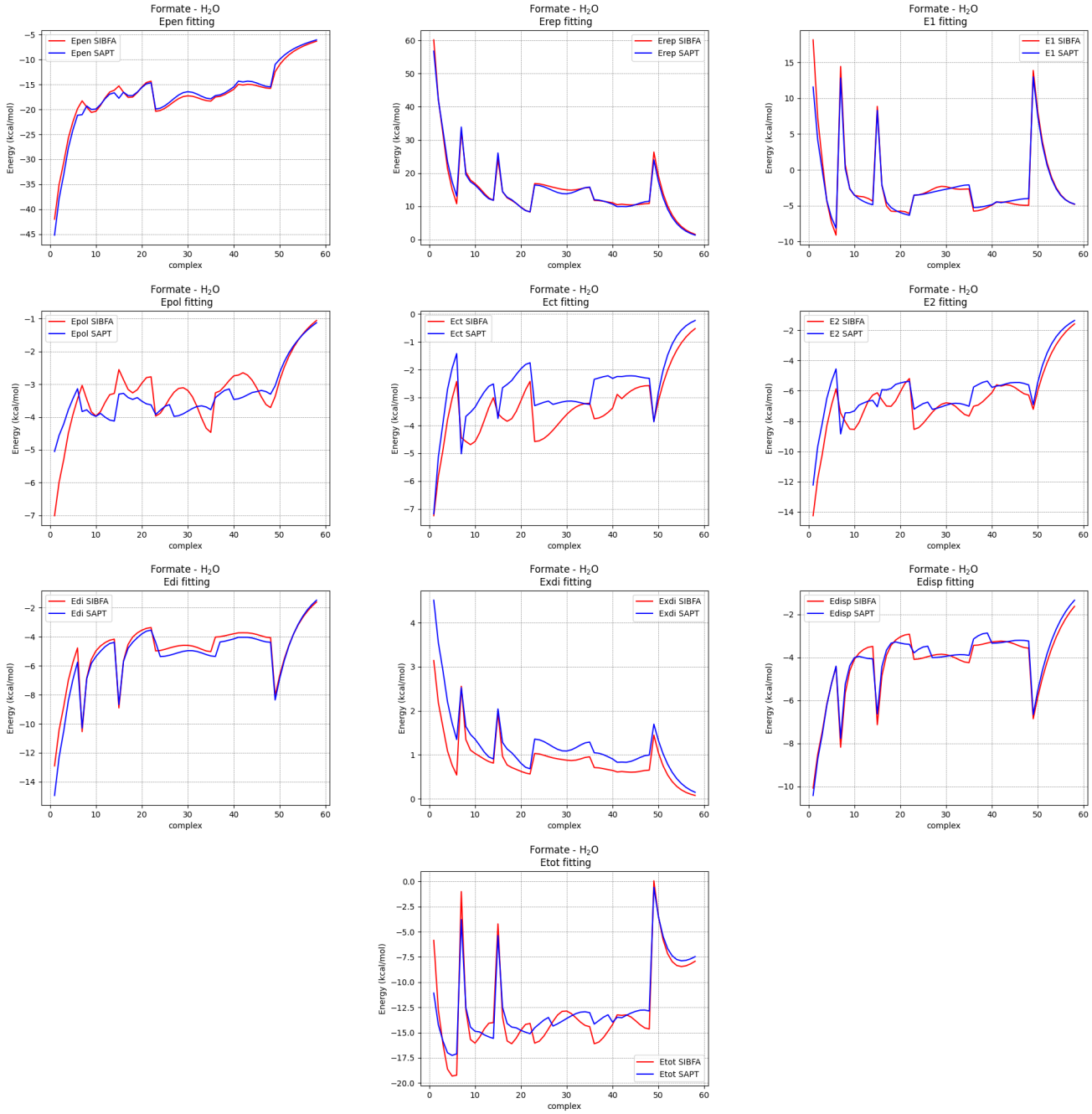


Figure 6.11: Formate - H_2O SIBFA fitting on SAPT(DFT)

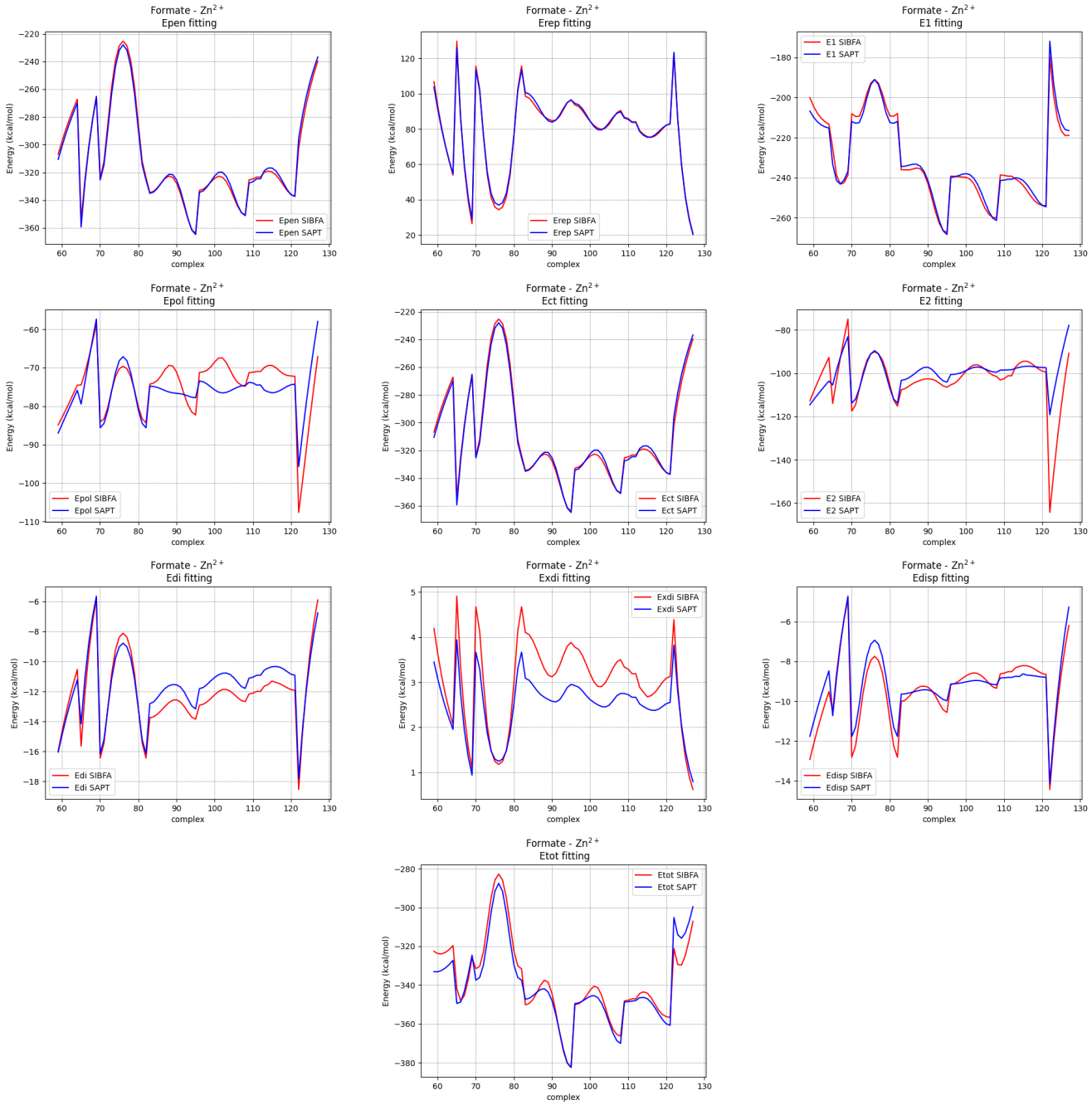


Figure 6.12: Formate - Zn^{2+} SIBFA fitting on SAPT(DFT)

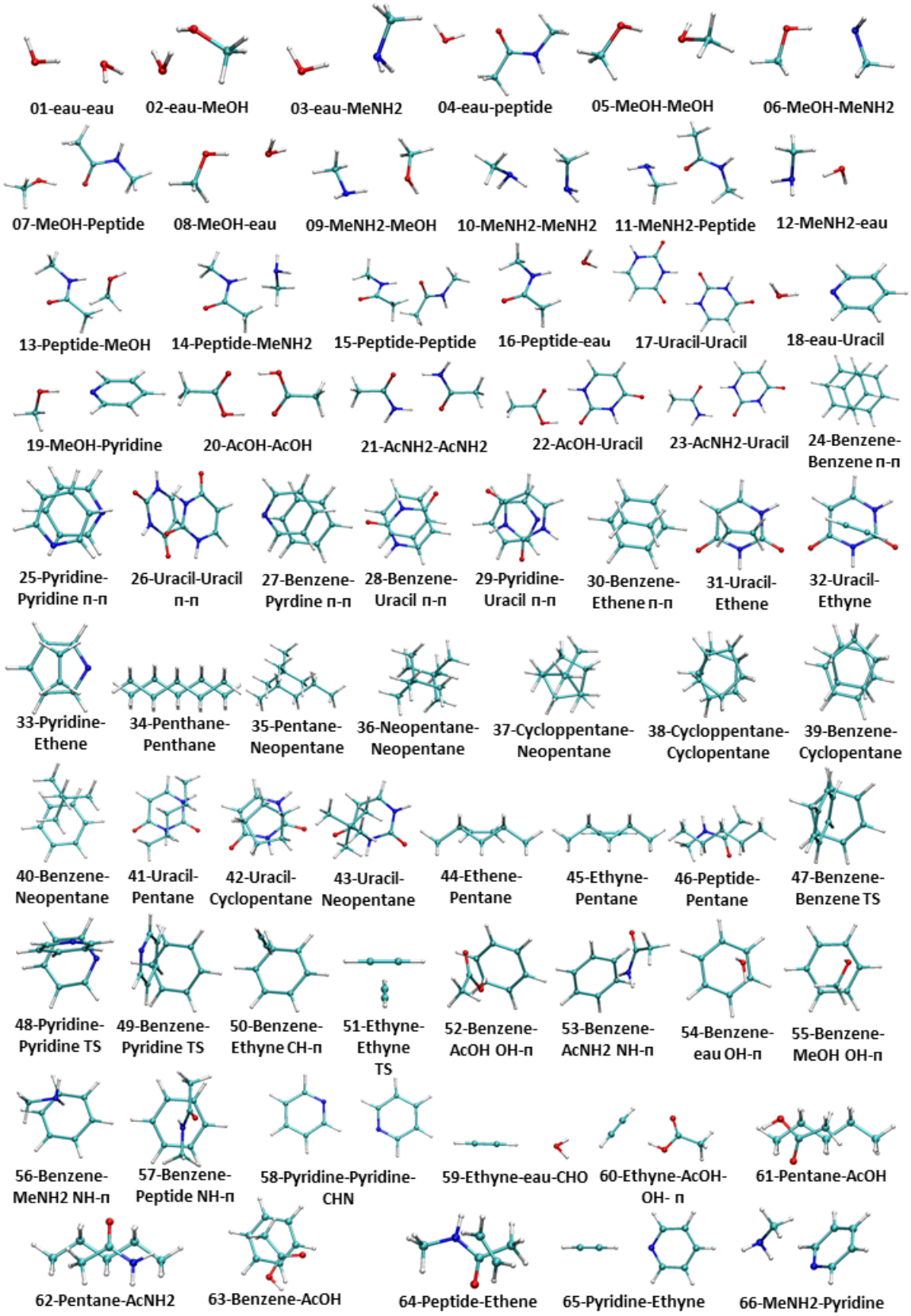


Figure 6.13: Représentation des 66 dimères (S66) des structures des biomolécules modèles.

kcal/mol	elst	exch	E1	POL	CT	IND	dHF	DISP	disp	exdisp	Etot
water-water	-7.77304	7.49013	-0.28291	-0.95504	-0.29317	-1.24821	-0.87953	-2.30113	-2.81539	0.51426	-4.71178
water-MeOH	-8.64903	8.86081	0.21178	-1.14535	-0.37706	-1.52241	-1.09877	-3.00153	-3.62034	0.61880	-5.41093
water-MeNH2	-11.88477	12.69063	0.80586	-1.54104	-0.56532	-2.10636	-1.84180	-3.46230	-4.40802	0.94572	-6.6046
water-Peptide	-12.25121	12.83336	0.58215	-1.78642	-0.49055	-2.27697	-1.63589	-4.34237	-5.28829	0.94592	-7.67308
MeOH-MeOH	-8.67242	9.09681	0.42439	-1.16663	-0.37693	-1.54356	-1.20854	-3.37388	-4.03639	0.66251	-5.70159
MeOH-MeNH2	-12.42719	13.92449	1.4973	-1.66346	-0.61308	-2.27654	-2.09916	-4.52661	-5.61470	1.08809	-7.40501
MeOH-Peptide	-12.38739	13.52555	1.13816	-1.80809	-0.48930	-2.29739	-1.81876	-5.02466	-6.09405	1.06939	-8.00265
MeOH-water	-7.68163	7.56115	-0.12048	-0.98430	-0.29201	-1.27631	-0.95206	-2.57951	-3.12088	0.54137	-4.92836
MeNH2-MeOH	-4.07937	4.69737	0.618	-0.44969	-0.10141	-0.5511	-0.45423	-2.70708	-3.11003	0.40295	-3.09441
MeNH2A-MeNH2B	-5.78222	7.15458	1.37236	-0.65666	-0.17219	-0.82885	-0.69917	-3.85822	-4.49206	0.63384	-4.01388
MeNH2-Peptide	-6.86658	8.46179	1.59521	-0.86283	-0.13798	-1.00081	-0.74651	-5.07968	-5.84944	0.76976	-5.23179
MeNH2-water	-11.99578	13.08544	1.08966	-1.62520	-0.60094	-2.22614	-1.86800	-3.94311	-4.88599	0.94288	-6.94759
Peptide-MeOH	-7.83513	8.11430	0.27917	-1.03950	-0.21170	-1.2512	-0.95985	-4.14067	-4.77074	0.63007	-6.07255
Peptide-MeNH2	-10.74519	12.16931	1.42412	-1.36911	-0.37587	-1.74498	-1.70103	-5.02865	-6.04127	1.01262	-7.05054
Peptide-Peptide	-11.21167	12.36526	1.15359	-1.62105	-0.29246	-1.91351	-1.52016	-5.78709	-6.79708	1.00999	-8.06717
Peptide-water	-6.73092	6.21612	-0.5148	-0.87462	-0.15832	-1.03294	-0.68662	-2.73158	-3.19562	0.46405	-4.96594
UracilA-UracilB	-26.23983	30.39790	4.15807	-4.81636	-1.10504	-5.9214	-5.12682	-8.61382	-10.49666	1.88285	-15.504
water-Pyridine	-11.03232	11.85002	0.8177	-1.52805	-0.54005	-2.0681	-1.64246	-3.60295	-4.47154	0.86859	-6.49581
MeOH-Pyridine	-11.57209	12.87253	1.30044	-1.62380	-0.58135	-2.20515	-1.88397	-4.40847	-5.39989	0.99143	-7.19715
AcOH-AcOH	-32.71212	41.03285	8.32073	-6.33951	-1.84037	-8.17988	-7.61450	-9.26622	-11.82361	2.55739	-16.7399
AcNH2-AcNH2	-25.56870	28.26343	2.69473	-4.28755	-0.89425	-5.1818	-4.58015	-7.63463	-9.53250	1.89787	-14.7019
AcOH-Uracil	-30.69663	35.99256	5.29593	-5.80753	-1.48785	-7.29538	-6.48418	-8.93241	-11.16176	2.22935	-17.416
AcNH2-Uracil	-28.85562	31.46185	2.60623	-5.09761	-1.15732	-6.25493	-5.51894	-8.39806	-10.39298	1.99492	-17.5657
Benzene-Benzene	-1.39002	5.64436	4.25434	-0.18593	0.03533	-0.1506	-0.46995	-6.12422	-6.88826	0.76404	-2.49043
Pyridine-Pyridine	-2.79177	6.60670	3.81493	-0.13435	0.00428	-0.13007	-0.63562	-7.05417	-7.92206	0.86788	-4.00493
Uracil-Uracil	-9.08358	12.64568	3.5621	-0.85972	-0.03466	-0.89438	-0.61901	-11.25050	-12.49915	1.24865	-9.20179
Benzene-Pyridine	-2.22307	6.29047	4.0674	-0.14884	0.00539	-0.14345	-0.56704	-6.90158	-7.76722	0.86564	-3.54467
Benzene-Uracil	-4.61795	9.67642	5.05847	-0.36273	-0.00072	-0.36345	-0.65931	-9.35407	-10.52366	1.16959	-5.31836
Pyridine-Uracil	-5.69497	9.67602	3.98105	-0.41265	-0.00902	-0.42167	-0.68579	-9.32136	-10.42250	1.10113	-6.44777
Benzene-Ethene	-0.74319	3.78007	3.03688	-0.08389	0.00249	-0.0814	-0.37861	-4.15503	-4.74235	0.58732	-1.57816
Uracil-Ethene	-3.12045	5.57997	2.45952	-0.17039	-0.01205	-0.18244	-0.35571	-5.19362	-5.90178	0.70815	-3.27225
Uracil-Ethyne	-3.90967	5.55059	1.64092	-0.23535	-0.01378	-0.24913	-0.32003	-4.77174	-5.43078	0.65904	-3.69998
Pyridine-Ethene	-2.22711	6.67553	4.44842	-0.08181	-0.00794	-0.08975	-0.52440	-5.39969	-6.25768	0.85799	-1.56542
Pentane-Pentane	-1.95553	6.68149	4.72596	-0.01214	-0.07981	-0.09195	-0.38655	-7.66019	-8.36314	0.70295	-3.41273
Neopentane-Pentane	-1.43359	4.74666	3.31307	-0.01255	-0.05993	-0.07248	-0.29496	-5.33487	-5.84135	0.50649	-2.38924
Neopentane-Neopentane	-0.83760	3.28510	2.4475	-0.01750	-0.05236	-0.06986	-0.20326	-3.79730	-4.12259	0.32530	-1.62292
Cyclopentane-Neopentane	-1.41633	4.69641	3.28008	-0.00985	-0.06491	-0.07476	-0.30098	-5.11911	-5.61393	0.49482	-2.21477
Cyclopentane-Cyclopentane	-1.57535	5.22177	3.64642	-0.01178	-0.07496	-0.08674	-0.32592	-5.97143	-6.51221	0.54078	-2.73767
Benzene-Cyclopentane	-2.29422	6.09156	3.79734	-0.07941	-0.03990	-0.11931	-0.48572	-6.68821	-7.46644	0.77824	-3.4959
Benzene-Neopentane	-2.30312	6.11267	3.80955	-0.07916	-0.03974	-0.1189	-0.48572	-6.68753	-7.46804	0.78051	-3.4826
Uracil-Cyclopentane	-2.32791	6.80825	4.48034	-0.19775	-0.04278	-0.24053	-0.30902	-7.49729	-8.25537	0.75809	-3.5665
Uracil-Neopentane	-2.57818	5.87409	3.29591	-0.16085	-0.03063	-0.19148	-0.26036	-6.18369	-6.84261	0.65893	-3.33962
Ethene-Pentane	-1.10417	3.56683	2.46266	-0.04818	-0.03585	-0.08403	-0.25052	-4.05474	-4.44296	0.38822	-1.92663
Ethyne-Pentane	-1.16856	3.05607	1.88751	-0.06738	-0.02059	-0.08797	-0.24723	-3.36655	-3.72910	0.36255	-1.81424
Peptide-Pentane	-2.70331	7.35587	4.65256	-0.37968	-0.05459	-0.43427	-0.42991	-7.61096	-8.39695	0.78599	-3.82258
Benzene-Benzene-TS	-1.86987	4.23743	2.36756	-0.14322	-0.03162	-0.17484	-0.37530	-4.61332	-5.12448	0.51115	-2.7959
Pyridine-Pyridine-TS	-2.78880	4.93766	2.14886	-0.24212	-0.04349	-0.28561	-0.41603	-4.82069	-5.33879	0.51810	-3.37347
Benzene-Pyridine-TS	-2.34632	4.46489	2.11857	-0.20640	-0.03654	-0.24294	-0.44024	-4.67816	-5.18984	0.51167	-3.24277
Benzene-Ethyne-CHpi	-2.16693	3.23667	1.06974	-0.32341	-0.04364	-0.36705	-0.44483	-3.17104	-3.47944	0.30839	-2.91318
Ethyne-Ethyne-TS	-1.85403	2.11405	0.26002	-0.19336	-0.04235	-0.23571	-0.30396	-1.38504	-1.55229	0.16726	-1.66469
Benzene-AcOH-OHpi	-4.14716	6.20656	2.0594	-0.93934	-0.13142	-1.07076	-0.79680	-4.75512	-5.38320	0.62808	-4.56328
Benzene-AcNH2-NHpi	-4.87246	6.50674	1.63428	-0.74454	-0.06340	-0.80794	-0.72428	-4.39317	-5.05570	0.66252	-4.29111
Benzene-Water-OHpi	-2.86043	3.55802	0.69759	-0.51502	-0.08594	-0.60096	-0.39007	-2.97648	-3.33885	0.36236	-3.26992
Benzene-MeOH-OHpi	-3.19062	4.76987	1.57925	-0.54354	-0.09614	-0.63968	-0.55207	-4.63074	-5.12770	0.49696	-4.24324
Benzene-MeNH2-NHpi	-2.40679	4.48763	2.08084	-0.22264	-0.04626	-0.2689	-0.40819	-4.60232	-5.13659	0.53427	-3.19857
Benzene-Peptide-NHpi	-3.93120	6.43770	2.5065	-0.60163	-0.07598	-0.67761	-0.67052	-6.26267	-6.96302	0.70035	-5.1043
Pyridine-Pyridine-CHN	-4.77276	6.17730	1.40454	-0.68478	-0.10688	-0.79166	-0.53153	-3.92059	-4.46751	0.54692	-3.83924
Ethyne-Water-CHO	-3.96126	3.51878	-0.44248	-0.47371	-0.08364	-0.55735	-0.39362	-1.54799	-1.78529	0.23731	-2.94144
Ethyne-AcOH-OH	-7.22030	8.63675	1.41645	-1.14469	-0.22044	-1.36513	-1.14511	-3.77561	-4.47169	0.69608	-4.8694
Pentane-AcOH	-1.81091	5.05043	3.23952	-0.19051	-0.04189	-0.2324	-0.27940	-5.33991	-5.84489	0.50498	-2.61219
Pentane-AcNH2	-2.42457	6.29084	3.86627	-0.47145	-0.05107	-0.52252	-0.39859	-6.17494	-6.84104	0.66609	-3.22978
Benzene-AcOH	-3.08071	5.75329	2.67258	-0.26615	-0.02088	-0.28703	-0.49499	-5.57601	-6.23300	0.65699	-3.68545
Peptide-Ethene	-2.71947	4.64743	1.92796	-0.36166	-0.04117	-0.40283	-0.34037	-4.11904	-4.59671	0.47767	-2.93428
Pyridine-Ethyne	-5.88313	6.24515	0.36202	-0.70035	-0.15557	-0.85592	-0.86778	-2.56140	-2.99942	0.43802	-3.92308
MeNH2-Pyridine	-4.28347	6.34851	2.06504	-0.38112	-0.09256	-0.47368	-0.57238	-4.85840	-5.48040	0.62200	-3.83942

Table 6.1: Les valeurs des énergies d'interaction calculées en SAPT(DFT) pour les 66 structures biomoléculaire bien équilibrées de référence (S66).