Etude des ADN glycosylases de la superfamille structurale Fpg/Nei par modélisation moléculaire, de nouvelles cibles thérapeutiques potentielles dans les stratégies anti-cancer

Charlotte Rieux

UNIVERSITÉ D'ORLÉANS

ÉCOLE DOCTORALE SANTÉ, SCIENCES BIOLOGIQUES ET CHIMIE DU VIVANT

Centre de Biophysique Moléculaire, CNRS Orléans

THÈSE présentée par :

Charlotte RIEUX

soutenue le : **20 décembre 2017**

pour obtenir le grade de : **Docteur de l'université d'Orléans**

Discipline/Spécialité : **Biologie structurale et fonctionnelle**

Étude des ADN glycosylases de la superfamille structurale Fpg/Nei par modélisation moléculaire, de nouvelles cibles thérapeutiques potentielles dans les stratégies anti-cancer

THÈSE dirigée par :

Norbert GARNIER	Maître de Conférences, Université d'Orléans

et co-encadrée par :

Stéphane BOURG	Ingénieur de Recherche, Institut de Chimie Organique et Analytique

RAPPORTEURS :

Manuel DAUCHEZ	Professeur des Universités, Université de Reims Champagne-Ardenne
Laurence SERRE	Chargée de Recherche, Grenoble Institut de Neurosciences

---

JURY :

Hélène BÉNÉDETTI	Directrice de Recherche, Centre de Biophysique Moléculaire
Pascal BONNET	Professeur des Universités, Université d'Orléans, Président du Jury
Stéphane BOURG	Ingénieur de Recherche, Institut de Chimie Organique et Analytique
Manuel DAUCHEZ	Professeur des Universités, Université de Reims Champagne-Ardenne
Catherine ETCHEBEST	Professeur des Universités, Université de Paris Diderot - Paris 7
Norbert GARNIER	Maître de Conférences, Université d'Orléans
Chantal PRÉVOST	Chargée de Recherche, Laboratoire de Biochimie Théorique
Laurence SERRE	Chargée de Recherche, Grenoble Institut de Neurosciences

Remerciements

Ce travail de thèse a été réalisé au Centre de Biophysique Moléculaire (CBM), au CNRS d'Orléans, et a été financé par le ministère de l'enseignement supérieur et de la recherche. Je remercie le Dr. Éva Jakab Toth de m'avoir acceuilli au sein du laboratoire.

Je remercie le Dr. Laurence Serre et le Pr. Manuel Dauchez d'avoir accepté d'examiner mon travail de thèse. Je remercie également le Dr. Hélène Bénédetti, le Pr. Pascal Bonnet, le Pr. Catherine Etchebest et le Dr. Chantal Prévost d'avoir accepté de faire partie de mon jury de thèse.

Je tiens à remercier tous ceux qui ont participé de prêt ou de loin à ce travail :

Norbert Garnier d'avoir été mon directeur de thèse et Stéphane Bourg pour l'encadrement de mes travaux. Je remercie également Bertrand Castaing de m'avoir accueilli dans son équipe. Pour leurs enseignements et leurs bons conseils.

Stéphane Goffinont, pour les tests biochimiques des molécules sur les activités des ADN glycosylases et pour m'en avoir enseigné les principes.

Franck Coste, pour avoir partagé ses connaissances des structures des protéines *Ll*Fpg et hNEIL1 ainsi que des modes d'interactions des $2TX_n$ sur ces cibles.

Nos collaborateurs du groupe « Chimie des nucléosides et hétérocycles – Recherche en infectiologie » (Institut de Chimie Organique et Analytique, ICOA d'Orléans) du Pr. Luigi Agrofoglio ainsi que le Dr. Vincent Roy et les Dr. Magali Lorion et Zahira Tber pour la synthèse des molécules $2TX_n$ et des P_n, ainsi que la conception des molécules VR.

Nos collaborateurs du groupe « Bioinformatique Structurale et Chémoinformatique » (ICOA d'Orléans) du Pr. Pascal Bonnet, pour le criblage virtuel de hNEIL1, ainsi que le Centre de Calcul Scientifique de la région Centre pour la puissance de calcul alloué à ce projet. Merci également à Samia Aci-Sèche pour ses conseils sur le logiciel AMBER, les paramètres du THF et son aide pour la TMD^{-1}. Je

remercie Jose Manuel Gally d'avoir répondu à mes questions et m'avoir permis de mieux comprendre le fonctionnement de VSPrep. Je remercie également Fabrice Carles de m'avoir éclairé sur les principes des fingerprints et la classification des molécules par critère de similarité. Merci également à Sonia Zaida pour son aide concernant les simulations de TMD^{-1}.

Notre collaborateur de Greepharma (Orléans), le Dr. Quoc-Tuan Do pour son travail de docking aveugle des monomères et polymères de 2TX et 2TX3. Je le remercie également pour sa participation en tant qu'orateur au 15ème colloque de l'ADOC.

Je remercie l'équipe « Matrice Extracellulaire et Dynamique Cellulaire » (Université de Reims Champagne-Ardenne) du Pr. Manuel Dauchez ainsi que le Dr. Stéphanie Baud et le Dr. Nicolas Belloy de m'avoir accueillie dans l'équipe et encadrée pendant une semaine à Reims, de m'avoir appris à utiliser AMIDE et m'avoir permis d'accéder au Centre de Calcul Régionnal de Champagne-Adenne ROMEO que je remercie également.

Je souhaite également remercier Alain Boyer et Yannick Berteaux pour leur aide concernant les logiciels et la maintenance du cluster Lucrece au CBM. Je voudrais aussi remercier le Centre de Calcul Scientique en région Centre pour m'avoir permis d'utiliser le cluster Artemis.

Je tiens à remercier tous les membres, présents ou anciens, de mon groupe thématique d'accueil, « Réparation de l'ADN : Structure, Fonction et Dynamique » qui m'ont accompagnée au cours de ma thèse : Serge Boiteux, Abdennour Braka, Bertrand Castaing, Franck Coste, Julien Cros, Françoise Culard, Norbert Garnier, Martine Guérin, Stéphane Goffinont, Rémy Le Meur, Virginie Nadan, Antonin Nourisson et Alan Waquiez.

Julien, Antonin et Alan avec qui j'ai partagé le bureau D208 pendant quelques temps, pour les petites blagues et énigmes qu'il fallait résoudre pour retrouver mon clavier et se mettre au travail de bon matin, la vraie question était « dans quel état j'ère… ».

Je remercie également les doctorants et postdocs du laboratoire, du CEMHTI, de l'ICARE, de l'ICOA et de l'ISTO pour les afterworks, sorties et événèments organisées par l'ADOC et l'ADSO : Abdennour, Alan, Alba, Anna, Baptiste, Cédric, Célia, Chloé, Clément, Fabrice, Francesco, Franck,

Geoffrey, Jade, José, Justine, Julien, Kelly, Kévin, Mamar, Martin, Maxime, Patrick, Rémy, Sandra et Sophie. Un special thanks et bisous de daim à Dounia et Willfried pour les bons souvenirs. Merci à tous pour les bons moments passés ensemble.

Je souhaite beaucoup de réussite à ceux qui termine leurs thèses en cette fin d'année : Abdennour, Alba, Baptiste, Dounia, José, Mamar, Patrick et Wilfried. Je souhaite aussi du courage à Julien qui commence sa thèse prochainement. Bon courage aussi à mes élèves Édifice qui commencent leurs études supérieures : Éléna, Laëtitia, Léonie et Sawfane.

Je souhaite enfin remercier ma famille et bons amis pour leur affection et leur soutien. Vous étiez tous très loin, mais bientôt nous serons à nouveau réunis.

Table des matières

Abréviations

2TX : 2-thioxantine

5-mC : 5-méthylcytosine

5-hmC : 5-hydroxyméthylcytosine

5-fC : 5-formylcytosine

5-caC : 5-hydroxyméthylcytosine

(6-4)PP : Pyrimidine(6-4)pyrimidone

8-oxoG : 8-oxoguanine

8-oxoA : 8-oxoadénine

A : Adénine

aa : Acide Aminé

AAG/ANPG : Alkyladénine DNA Glycosylase

AD : AutoDock

AD4 : AutoDock 4

ADN : Acide Désoxyribose Nucléique

ADV : AutoDock Vina

AGOG : « Archaeal 8-OxoGuanine DNA Glycosylase »

Alt-NHEJ : « Alternative Non-Homologous End Joining »

AMIDE : « Automated Molecular Inverse Docking Engine »

AN : Acide Nucléique

ARN : Acide Ribo Nucléique

ARN_m : Acide Ribo Nucléique messager

ARNnc : Acide Ribo Nucléique non-codant

ARNt : Acide Ribo Nucléique de transfert

ARNr : Acide Ribo Nucléique ribosomal

BA : « Base Activation »

BER : « Base Excision Repair »

BET : Bromure d'Éthidium

bzFaPyG : benzyl-FaPyG

BRCA : « Breast Cancer gene »

BFGS : algorithme Broyden-Fletcher-Goldfarb-Shanno

C : Cytosine

C-ter : extrémité C-terminale d'une protéine

c8-oxoG : carba-8-oxoguanine

cFaPyG : carba-FaPyG

CG : Centre Géométrique

CHA : Classification Hiérarchique Ascendante

CPD : « Cyclobutane Pytimidine Dimer »

D-loop : « Displacement loop »

Da : Dalton

DB : Double Brin

DCC : « Dynamics Cross-Correlation »

DHU : 5,6-dihydrouracil

DHT : 1,3-dihydrothymine

DM : Dynamique Moléculaire

DSB : « Double Strand Break »

DSBR : « Double Strand Break Repair »

ERO : Espèces Réactives de l'Oxygène

FaPyG, FaPyA : 2,6-diamino-4-hydroxy-5-formamidopyrimidine

FDA : « Food and Drug Administration »

FEN1 : Flap Endonucléase 1

Fpg : Formamidopyrimidine DNA glycosylase

G : Guanine

GA : Algorithme Génétique

Gh : Guanidinohydantoine

GPCR : « G protein-coupled receptors »

H_2O_2 : Peroxyde d'hydrogène

HA : « Hydrogen Acceptor »

HD : « Hydrogen Donnor »

HEAT repeat : « **H**untingtin elongation factor 3 » (**E**F3), « protein phosphatase 2A » (PP2**A**) et kinase **T**OR1 de la levure

HTT/HD : « Huntingtin gene »

hX : Hypoxanthine

Hyd : Hydantoine ; glycolylurée ; 2,4-imidazolidinedione

KRAS : « Kirsten rat sarcoma viral oncogene homolog gene»

Ia : Iminoallantoine

IC_{50} : Concentration inhibitrice médiane

IR : Infra Rouge

K_i : Constante d'inhibition

KO : « Knock Out »

LigI : ADN Ligase I

LigIII : ADN Ligase III

LSCX : « Latent Structural Cluster n° X »

MBD4 : « Methyl Binding Domain IV »

MC : algorithme Monte-Carlo

MCM : algorithme Monte-Carlo/Metropolis

MRX : complexe des protéines Mre11, Rad50, Xrs2

MSH2 : « MutS Homolog 2 gene »

MW : « Molecular Weigh »

N-ter : extrémité N-terminale d'une protéine

Nei : Endoclucléase VIII

NEIL1 : Endonucléase VIII like-1

O_2^- : Anion superoxide

$^1O_2$: Oxygène singulet

OGG1 : « 8-oxoguanine DNA glycosylase 1 »

OHU : 2-dioxy-5-hydroxyuridine 5'-(dihydrogenphosphate)

PAINS : « Pan-Assay Interference Compounds »

PARP1 : Protéine poly(ADP)-Ribose Polymérase 1

pb : Paire de bases

PBI : Propanediol

PDB : « Protein Data Bank »

Polβ : ADN polymérase beta

Polε : ADN polymérase sigma

Polδ : ADN polymérase delta

PSA : « Polar Surface Area »

PUA : sites 3'-phospho-aldéhyde α,β-insaturés

QM/MM : « Quantum Mechanics/Molecular Mechanics »

RA : « Ribose Activation »

RI : Radiations Ionisantes

RMN : Résonnance Magnétique Nucléaire

RMSD : « Root Mean Square Deviation »

RMSF : « Root Mean Square Fluctuation »

RNA_{si} : petit ARN interférent

RO5 : « Linpinki's rule of five »

RPA : « Replication Protein A »

RX : Rayons X

SB : Simple Brin

SDSA : « Synthesis-Dependent Strand Annealing »

Sp : Spiroiminodihydantoine

SSB : « Single Strand Break »

SSBR : « Single Strand Break Repair »

T : Thymine

TET : « Ten-Eleven Translocation »

TDG : « Thymine DNA Glycosylase »

Tg : Thymine glycol

THF : Tetrahydrofuranose

TIP3P : « Transferable Intermolecular Potential with 3 Points »

UV : Ultra Violet

UDG/UNG : « Uracile DNA Glycosylase »

vdW : van der Waals

WF : WorkFlow

VSPrep : « Virtual Screening Preparator »

WT : « Wild Type »

X : Xanthine

XRCC1 : « X-Ray Cross-Complementation group 1 »

ZnF : « Zinc Finger »

ZnLF : « Zinc Less Finger »

Index des Tableaux, Équations et Figures

Tableaux :

Équations :

Figures :

I. **I**ntroduction

Chapitre 1 : Contexte biologique

Chapitre 2 : Réparation de l'ADN par excision de base (BER)

Chapitre 3 : Les ADN glycosylases de la superfamille Fpg/Nei

I.1 **Chapitre 1 : C**ontexte biologique

I.1.1 L'acide désoxyribonucléique

I.1.1.1 Découverte de l'ADN, support de l'information génétique

Hérédité, un terme qui, sous l'ancien régime, était utilisé par les juristes pour signifier l'appartenance de biens à une personne et en médecine où l'on désignait par le terme « maladies héréditaires » les pathologies qui se transmettent des parents à l'enfant, et dont le vecteur n'est pas un organisme tiers. Les savants et philosophes de l'époque associent l'hérédité à un phénomène biologique dont les secrets restent encore à découvrir. Au XIX^{eme} siècle, le naturaliste anglais Charles Darwin publie son manuscrit « l'Origine des espèces » [1] et fait de l'hérédité le pilier de sa théorie de l'évolution. Dans la même période, le moine tchèque Gregor Mendel étudie l'hybridation des végétaux (pois) et pose les bases théoriques de la génétique et de l'hérédité moderne [2]. Mendel postule sur l'existence de facteurs responsables de la transmission des caractères entre les générations. Ces facteurs seront renommés « gènes » après la redécouverte de ses travaux au XX^{eme} siècle par trois botanistes Hugo de Vries, Carl Correns et Erich von Tscermak, et une nouvelle science nommée génétique est fondée. Les biologistes Walter Sutton, Theodor Boveri et Frans Alfons Janssens mettront ces connaissances en relation avec celles qu'ils obtiendront sur les chromosomes et affirment que ces objets sont les porteurs des gènes et sont donc le support de l'hérédité. Les gènes bien que localisés sur les chromosomes ne sont encore que des entités théoriques et leur caractère physique est encore inconnu. C'est en 1944 que la molécule d'acide désoxyribonucléique (ADN) est isolée par Oswald Theodore Avery, Colin Mc Leod et Maclyn Mc Carty. Ils identifient cette molécule comme le composant primaire des chromosomes. Un modèle tridimensionnelle hélicoïdal double-brin est proposé en 1953 par les chercheurs James Watson et Francis Crick (prix Nobel de médecine 1962) [3] à partir des clichés de diffraction des rayons X par des fibres d'ADN (obtenus par Rosalind Franklin et communiqués à son insu à Watson par Wilkins) [4]. Ce modèle sera confirmé par la suite par cristallographie des rayons X.

La molécule d'ADN est donc le support de l'information génétique de tous les organismes vivants et des virus. Elle est contenue dans un noyau chez les organismes eucaryotes et baigne dans le cytoplasme chez les organismes procaryotes ainsi que dans certaines organelles telles que les mitochondries et les chloroplastes (exclusifs aux cellules végétales). Elle contient toutes les instructions et informations nécessaires à la croissance, au fonctionnement et à la reproduction de l'organisme. L'ADN est une macromolécule formée d'une grande succession de 4 briques

élémentaires, les nucléotides, et leur enchainement définit une séquence précise. La longueur de cette séquence dépasse plusieurs millions ou milliards de nucléotides selon l'organisme. Par exemple depuis le séquençage complet de l'ADN humain en 2003 lors du projet international génome humain, on sait que l'ADN humain est constitué de 2,9 milliards de nucléotides portés en doublon sur 23 paires de chromosomes. Chez l'Homme, seulement 2% de l'ADN contient des parties codantes pour les fonctions vitales des cellules, le reste n'est pas exprimé, mais probablement transcrit sous forme d'ARN_{nc} (ARN non-codant) tels que les ARN_t (ARN de transfert) et ARN_r (ARN ribosomaux) entre autres. Dans la partie codante, nous retrouvons les gènes qui sont exprimés et permettent la synthèse d'Acides RiboNucléiques messagers (ARN_m) lors de la transcription. Ces ARN_m sont ensuite traduits en protéines, autres macromolécules essentielles au vivant décrites dans la section **I.1.2** p. **33**. L'ensemble des gènes d'une cellule forme ce que l'on appelle le génome. Le génome humain est composé d'environ 20 000 à 25 000 gènes. Il existe plusieurs versions de ces gènes, appelées allèles, qui sont responsables du phénotype, c'est-à-dire des traits et des caractéristiques morphologiques observables, de l'organisme comme par exemple la couleur des cheveux ou des yeux chez l'Homme. Une modification permanente dans la séquence nucléotidique d'un gène est une mutation (insertion, délétion, substitution), et peut avoir des conséquences sur le phénotype plus ou moins grave ou peut devenir un avantage pour l'organisme porteur. Ce genre d'évènement peut arriver spontanément dans la vie de l'organisme et peut se transmettre à sa descendance si cela n'est pas létal pour le parent et ne l'empêche pas de se reproduire. Le maintien de la structure et de la séquence de l'ADN est donc important pour la pérennité d'un organisme. Les molécules d'ARN et d'ADN sont aussi appelés Acides Nucléiques (AN).

L'ADN est sujet à des interactions avec d'autres molécules, notamment des protéines. Par exemple, l'ADN interagit avec des protéines appelées histones, qui ont pour but de structurer et condenser l'ADN en chromatine et rendre certaines parties du génome inaccessibles (enfouies) à la machinerie cellulaire. Les gènes contenus dans ces régions hyper compactées de l'ADN (hétérochromatine) ne sont donc pas exprimés. Seule l'euchromatine, formée essentiellement de la fibre nucléosomique (dite en collier de perle) est suffisamment accessible pour être transcrite et contribuera au phénotype cellulaire de l'organisme. L'ADN interagit donc avec des protéines telles que les histones mais aussi avec beaucoup d'autres facteurs impliqués dans son métabolisme (réplication, transcription, transposition, réparation et recombinaison). Les interactions ADN-protéine permettent à celui-ci d'adopter sa structure, d'être modifié si nécessaire (modulation de l'expression ou réparation) et d'être exprimé. Ces mécanismes sont essentiels à la vie.

I.1.1.2 Introduction à la structure de l'ADN

L'ADN est une molécule formée de 2 chaines antiparallèles généralement structurée en double-hélice décalée droite (**Figure 1** et **Figure 3**) [3]. Cette structure bicaténaire, aussi dite en Double Brin (DB), est polarisée 5' → 3' / 3' → 5'. Chaque chaine polymérique est appelée Simple Brin (SB) et est constituée d'une succession d'éléments constitutifs appelés nucléotides liés entre eux par des liaisons phosphodiesters. Chaque nucléotide résulte de l'estérification d'un nucléoside avec l'acide phosphorique. Le nucléoside résulte de la condensation d'une base azotée (purine ou pyrimidine) et d'un désoxyribose (2-désoxy-β-D-furanose) reliés par une liaison N-glycosidique (**Figure 1**). Seules quatre bases azotées différentes sont retrouvées dans l'ADN : deux « purines », l'Adénine (A) et la Guanine (G) et deux « pyrimidines » telles que la Thymine (T) et la Cytosine (C). L'Uracile (U) est la cinquième base azotée et remplace le T dans la molécule d'ARN. Les groupements phosphates sont exposés à l'extérieur de la double hélice d'ADN et forment avec le désoxyribose le squelette désoxyribose phosphate. Les bases azotées s'empilent les unes sur les autres à l'intérieur de la double-hélice par recouvrement de leurs orbitales π intervenant dans des interactions de type « stacking ». En établissant des liaisons hydrogène, chaque base d'un brin interagit avec la base qui lui est opposée sur l'autre brin en respectant les règles d'appariement des bases dites de Watson-Crick (**Figure 1**).

À l'intérieur de la double hélice, les bases azotées peuvent adopter deux orientations différentes appelées *anti* et *syn* en tournant autour de la liaison *N*-glycosidique (angle χ), changeant profondément la nature des interactions des appariements (**Figure 2**). L'orientation préférentielle des bases azotées dans la double hélice des types A et B (**Figure 3**) est l'orientation *anti*, et correspond aux appariements de type Watson et Crick A(*anti*) - T(*anti*) et C(*anti*) - G(*anti*) [3]. Lorsque deux bases aux orientations différentes sont appariées, ces paires sont dites de type « Hoogsteen » A(*syn*) - T(*anti*) et G(*syn*) - C^+(*anti*) [5]. Ces derniers appariements sont typiques de la double hélice gauche retrouvée dans l'ADN en conformation Z (**Figure 3**). Deux bases appariées quelles que soient leurs natures et le type d'appariement forment ce que l'on appelle une paire de bases (pb). Une pb Watson-Crick est également susceptible d'établir des liaisons hydrogène supplémentaires de type Hoogsteen avec une troisième base, ce qui permet la formation de structures d'ADN à trois brins antiparallèles, aussi appelées triple hélices [6].

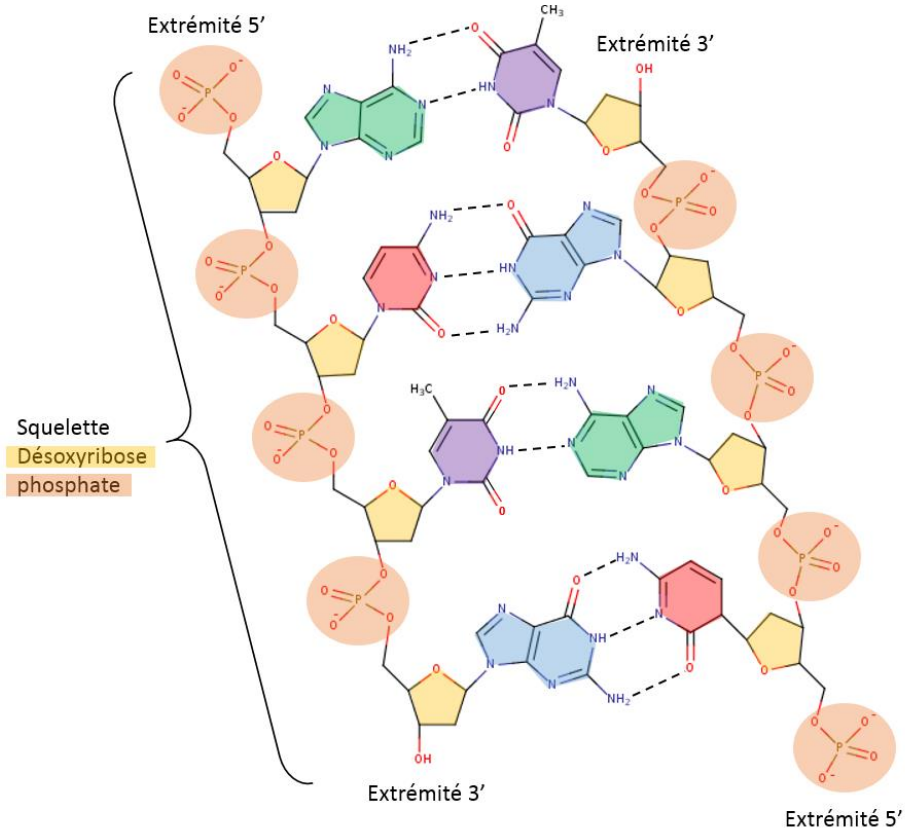


Figure 1 : Structure bicaténaire de l'ADN
Les bases azotées sont l'adénine (vert), la thymine (mauve), la cytosine (rouge) et la guanine (bleu). Le squelette de la structure de l'ADN bicaténaire est formé par l'alternance de phosphodiester (orange) et les sucres riboses (jaune). Les bases azotées interagissent entre elles à l'intérieur de la double hélice *via* des liaisons hydrogène qui forment les paires Watson-Crick [3].

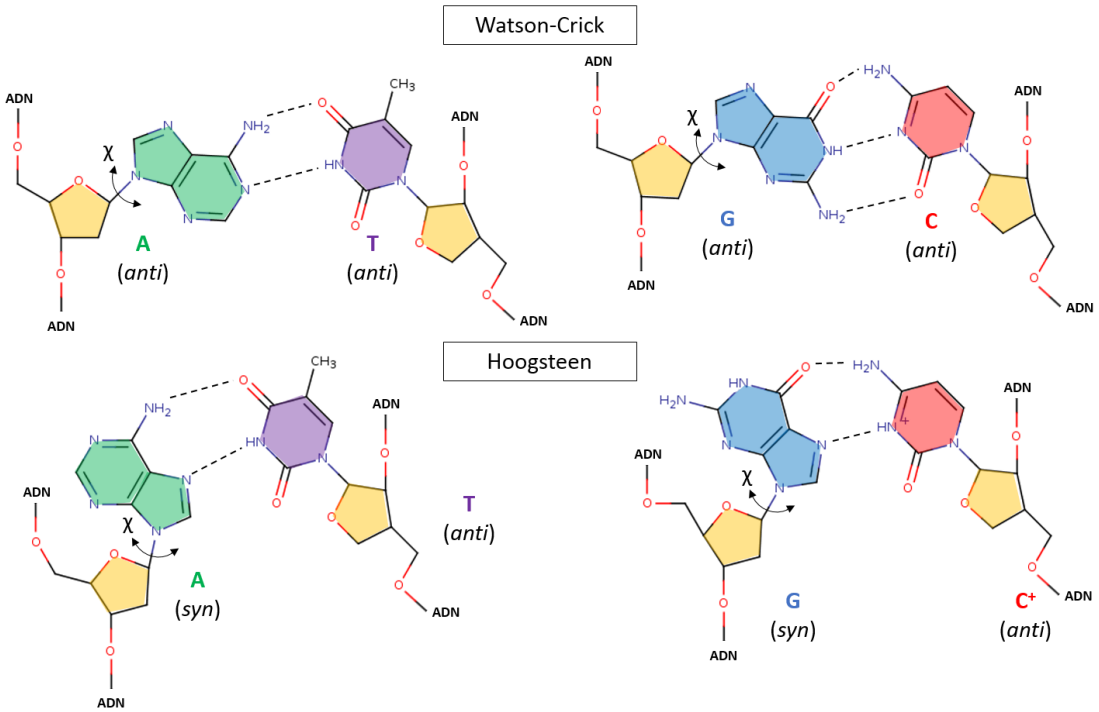


Figure 2 : Orientations *syn* et *anti* des bases azotées
Pour les bases puriques, c'est la distance du noyau pyrimidine par rapport au sucre (éloignée = *anti* ou proche = *syn*) qui permet de les distinguer [3, 5].

La double hélice est donc formée par deux chaines polymériques SB de nucléotides ayant le même sens de rotation (gauche = horaire ou droite = antihoraire) mais décalées entre elles ce qui aboutit à la formation de deux sillons adjacents aux paires de bases de tailles inégales appelés petit et grand sillon plus ou moins larges et profonds. Ces sillons fournissent des sites d'interaction pour diverses molécules (les facteurs de transcription par exemple). Il existe plusieurs sortes de doubles hélices en fonction du sens de rotation, du pas de l'hélice, du nombre de paires de bases par tour etc. La double hélice peut prendre 3 formes selon son hydratation, sa séquence, le taux de surenroulement, les modifications chimiques des bases, de la nature et de la concentration des ions métalliques en solutions. Elles sont appelées les formes A, B et Z (pour Zig zag) (**Figure 3**) [7]. La forme B étant la plus courante, car le taux d'hydratation dans les cellules est élevé. Seules les formes B et Z ont été observées *in vivo*. La forme A est la plus compactée et apparait exclusivement dans un environnement faible en eau. La forme Z semble surtout dépendre de la séquence et notamment de régions riches en paires G-C et contrairement aux formes A et B, a un sens de rotation horaire.

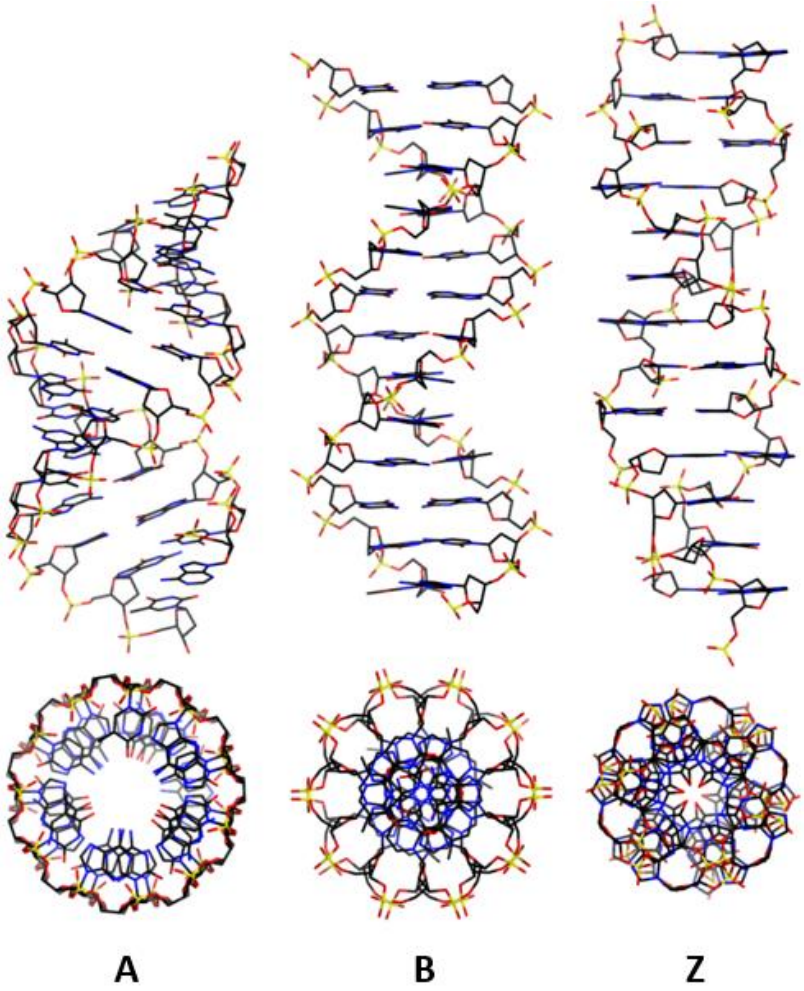


Figure 3 : Les trois formes d'hélices d'ADN, A, B et Z

Les caractéristiques de ces trois types d'hélices sont présentées dans le **Tableau 1** [7].

Tableau 1 : Caractéristiques des trois types de double hélice d'ADN A, B et Z

		A	B	Z	
Sens de l'hélice		Droite	Droite	Gauche	
Pas de l'hélice (Å)		28	34	44	
Torsion (°)		32-33	36	CG : -15	GC : -45
Sucre		C3' endo	C2' endo	Purine C3' endo	Pyrimidine C2' endo
Dist. P-P intrabrin (Å)		5,9	7,0	-	
Diamètre (Å)		23	20	18	
Rotation base-sucre		*Anti*	*Anti*	Purine *syn*	Pyrimidine *anti*
Grand sillon	Largeur (Å)	2,7	11,7	2,0	
	Profondeur (Å)	13,5	8,5	13,8	
Petit sillon	Largeur (Å)	11,0	5,7	8,8	
	Profondeur (Å)	2,8	7,5	3,7	
Pb par tour		11	10	12	

À l'état relâché, les 2,9 milliards de paires de bases de l'ADN des 23 chromosomes humains mis bout à bout atteignent la longueur de 1,9 m. Pour que la molécule d'ADN puisse rentrer dans le noyau d'une cellule d'environ 10 à 100 µm de diamètre, elle doit être compactée précisément pour rester active et adopte une structure « entortillée » sur elle-même. Elle forme ce que l'on appelle des supertours positifs ou négatifs. *In vivo*, l'ADN présente généralement un surenroulement négatif sous l'effet d'enzymes appelées les ADN topoisomérases. Ces enzymes sont également indispensables pour relâcher les contraintes lors des processus tels que la réplication ou la transcription de l'ADN. Seules les cellules eucaryotes possèdent un noyau contenant l'ADN. Chez les cellules procaryotes, l'ADN nage dans le cytoplasme et peut se retrouver également sous forme d'ADN circulaire appelé plasmide (chez l'Homme, les chromosomes sont constitués de molécules d'ADN DB linéaires). Dans les cellules, l'information génétique est le plus souvent portée par un ADN DB linéaire ou circulaire, mais elle existe aussi sous forme d'ADN SB et ARN DB et ARN SB chez les virus.

Les bases azotées sont des molécules très similaires entre elles à quelques groupements fonctionnels près et il suffit de quelques modifications comme des oxydations ou des méthylations de

ces dernières pour passer de l'une à l'autre ce qui peut être associé à un effet mutagène. Ces modifications peuvent aussi engendrer des dérivés de bases qui ne sont pas reconnus par la machinerie cellulaire ou qui sont encombrants dans la double hélice. La cellule est donc dans l'incapacité d'exploiter son ADN altéré, ce qui a des conséquences délétères [8]. Les modifications que peut subir l'ADN sont abordées plus en détails dans la section **I.1.3** p. **39**.

I.1.2 Les protéines

I.1.2.1 Fonctions

Les protéines sont des macromolécules essentielles au vivant. Elles sont formées par les acides aminés, décrits dans la section **I.1.2.2** p. **34**. Leur structure définit leurs fonctions qui sont très variées au sein de la cellule. Les protéines interviennent dans la formation du cytosquelette, qui donne à la membrane cellulaire sa forme globale ainsi qu'une certaine flexibilité. Elles interviennent également dans le transport de métabolites, d'ions, et d'autres molécules à travers les membranes plasmique et nucléaire, dans le métabolisme qui est la formation des composés nécessaires à la cellule. Elle joue aussi un rôle dans l'expression du code génétique contenu dans l'ADN. Pour résumer, elles interviennent dans tous les phénomènes d'assimilation, de croissance, et de division des organismes uni ou pluricellulaires, ces trois grandes propriétés définissant le vivant. Ce sont les machines microscopiques du vivant, servant à la structure, au transport et au métabolisme. L'ensemble des protéines d'un organisme est appelé le protéome.

Pour se donner une idée, le protéome humain est composé d'environ 19 000 protéines putatives [9]. « Putatives » signifie qu'elles sont prédites par la bio-informatique à partir du séquençage du génome humain, mais pas encore identifiées. L'identification du protéome humain est à ce jour en cours, et 93% des protéines ont été assignées à une fonction. Parmi les 19 000 protéines identifiées, 2 300 protéines font partie du « house keeping proteome », elles sont ubiquistes, c'est-à-dire qu'elles sont exprimées dans tous les tissus, leurs fonctions sont le contrôle général et la maintenance de la cellule. Elles composent 75% de la masse des protéines. Les autres sont donc spécifiques à un type cellulaire. De plus, les protéines peuvent subir des modifications post-traductionnelles et être décorées de groupements fonctionnels (acétylation, alkylation, phosphorylation entre autres…), lipides et sucres supplémentaires. Ces modifications permettent l'adressage des protéines dans les bons compartiments cellulaires et influencent leurs fonctions et temps de demi-vie. Ainsi, un gène produit une protéine qui sera ensuite modifiée, et ces modifications

peuvent produire plusieurs versions d'une protéine adaptant sa fonction selon le contexte (type cellulaire, environnement etc…).

I.1.2.2 Les acides aminés

Les acides aminés (aa) sont les briques fondamentales formant les protéines. Ce sont des molécules comportant toutes la même base formée des groupements carboxylate (COOH) et amine (NH_2) formant le squelette composé des atomes N-Cα-C-O. Les aa sont exclusivement composés des atomes de carbones (C), d'oxygènes (O), d'azotes (N), d'hydrogènes (H) et deux seulement comportent des atomes de soufre (S). Le Cα est porteur d'une chaîne latérale *R* pouvant avoir différentes structures qui déterminent la nature l'aa ainsi que ses propriétés physico-chimiques. Une vingtaine d'aa composent les protéines, cependant il en existe plusieurs centaines dans la nature et tous ne rentrent pas dans la synthèse de protéines. Ils peuvent être classés en fonction de leurs propriétés physico-chimiques telles que la polarité, leur nature aliphatique, aromatique ou cyclique (**Figure 5**). Ils sont tous reliés les uns aux autres *via* des liaisons peptidiques O-N (**Figure 4**) et forment une chaine polymérique aussi appelé chaine « polypeptidique ». La chaine polypeptidique peut être décrite par les angles dièdres ϕ (C_i-$C\alpha_i$-N_i-C_{i+1}) et ψ (N_{i-1}-C_i-$C\alpha_i$-N_i). La synthèse de la chaine polypeptidique est réalisée *in vivo* lors de la lecture d'un ARN_m et consiste en l'assemblage des aa, apportés par les ARN_t, correspondant aux codons (séquence de 3 nucléotides sur l'ARN_m) lus par le ribosome lors de la traduction. La chaîne polypeptidique possède deux extrémités nommées N-terminale (premier aa, aussi appelé extrémité N-ter) et C-terminale (dernier aa, aussi appelée extrémité C-ter), et se replie sous la force des interactions qu'entretiennent les aa les uns avec les autres pour former différentes structures. Le repliement de la protéine est dépendant de sa séquence en aa.

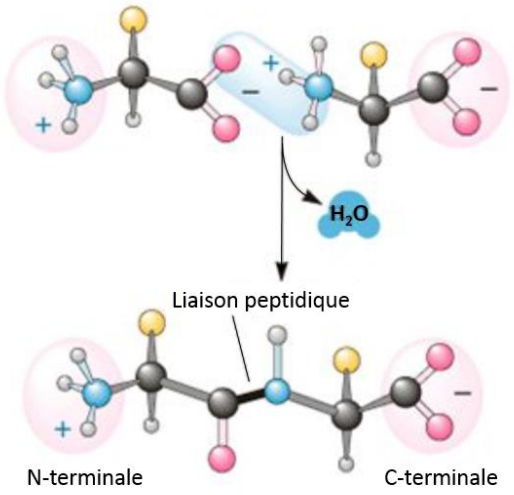


Figure 4 : Formation d'une liaison peptidique
C'est une liaison covalente entre le groupement carboxylate chargé négativement et amine chargée positivement de deux aa. La réaction engendre la production d'une molécule d'eau. Sur cette figure, les carbones sont en noir, les azotes en bleu, les oxygènes en rouge, les hydrogènes en blanc et les sphères jaunes correspondent aux chaines latérales des aa décrits dans la **Figure 5**.

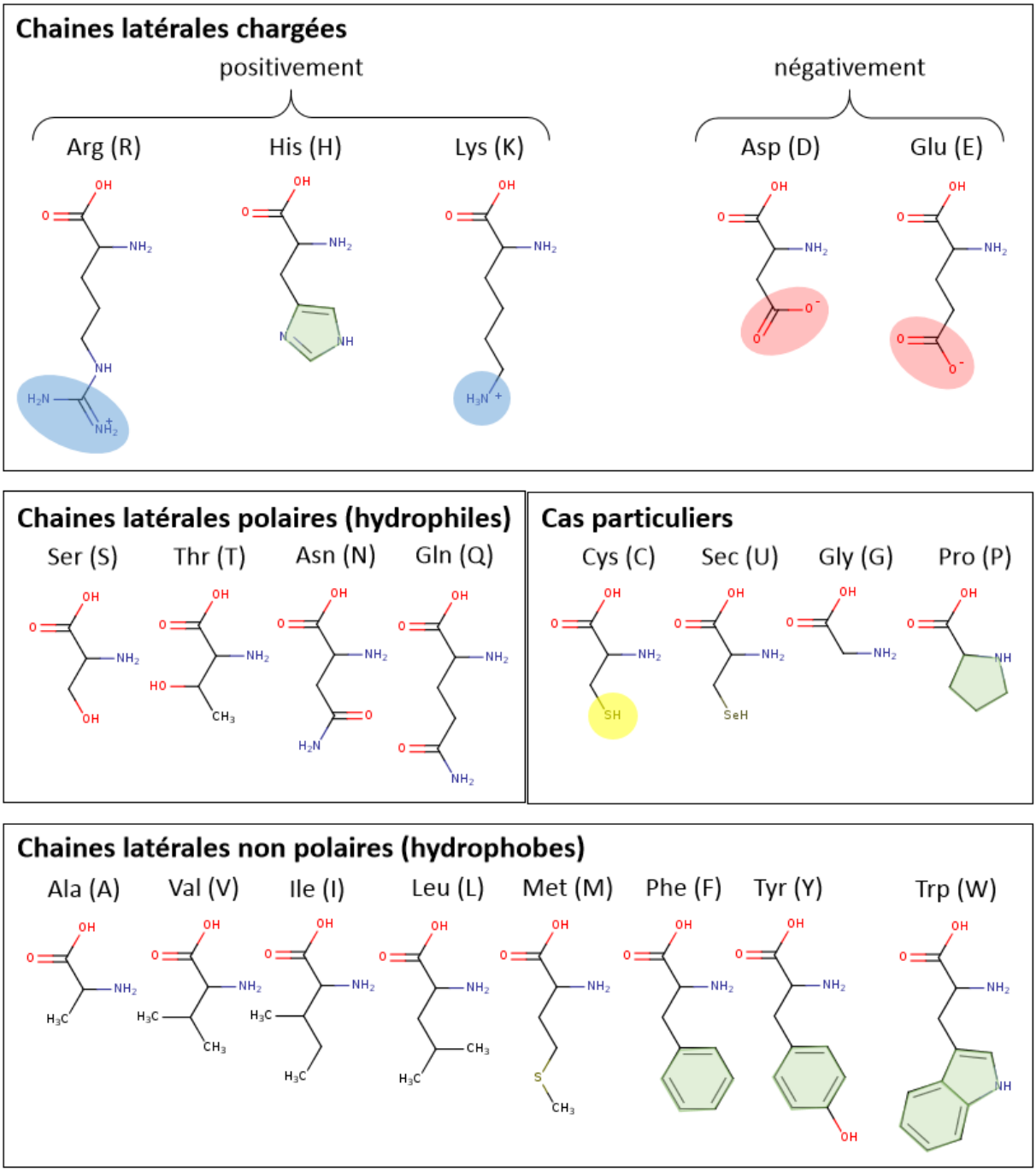


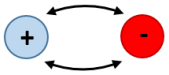
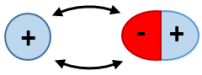
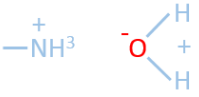
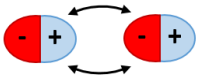
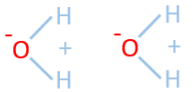
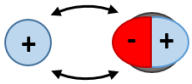
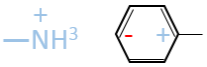
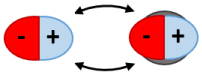
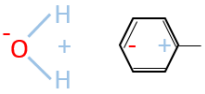
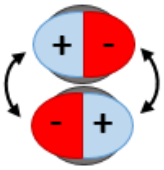
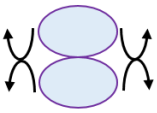
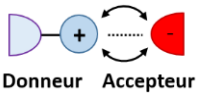
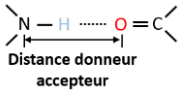
Figure 5 : Les acides aminés entrant dans la composition des protéines eucaryotes

Ceux-ci sont classés en 4 groupes selon la nature de leurs chaines latérales. Parmi les cas particuliers, la cystéine est capable de former une liaison covalente nommée pont disulfure avec une autre cystéine ou une molécule portant un groupement thiol (SH). La sélénocystéine (Sec, U) est plus rare et son incorporation est effectuée lors de la traduction grâce au « recodage » d'un codon Stop (UGA) de l'ARNm de la sélénoprotéine par l'anticodon d'un ARN_t spécifique pour la Sec [10]. La glycine et la proline sont des acides aminés particuliers dont la chaîne latérale est inexistante pour l'une ou incluse dans le squelette de ce dernier pour l'autre, respectivement, donnant une flexibilité ou une rigidité importante à la séquence protéique où elles sont intégrées.

I.1.2.3 Structures

Il existe plusieurs niveaux de complexité dans la structure des protéines. La structure primaire d'une protéine correspond à l'enchaînement des aa qui décrit une séquence. Sa composition déterminera sa(es) structure(s) tridimensionnelle(s) (**Figure 6 A**). Les autres niveaux de structuration sont en 3 dimensions et font intervenir des interactions de différents types entre les squelettes et les chaines latérales des aa. La protéine se replie sur elle-même et se combine parfois avec d'autres protéines pour adopter une conformation plus stable, et ces différents « niveaux » de repliement (secondaire, tertiaire, quaternaire) sont maintenus par différentes interactions décrits dans le **Tableau 2** [11]. Ces interactions peuvent être de nature covalente ou non covalente et impliquent ou non, respectivement, le partage d'électrons entre deux atomes. Des liaisons covalentes peuvent se créer spontanément au sein d'une protéine entre deux cystéines *via* leurs atomes de soufre (pont disulfure). Les interactions non covalentes aussi appelées interactions « à distance » sont des interactions électrostatiques. Les interactions électrostatiques impliquent deux objets porteurs de charges positive ou négative (ions). Selon les charges de ces objets, deux effets sont possibles : la répulsion (mêmes signes : -/-, +/+) et l'attraction (signes opposés : +/-) (**Tableau 2 A** et **B**). Les forces de van der Waals (vdW) sont des interactions électrostatiques faibles qui impliquent des molécules possédant un moment dipolaire permanent et/ou induit. Elles impliquent deux dipôles permanents (forces de Keesom, **Tableau 2 C**) ou un dipôle permanent et un dipôle induit (forces de Debye, **Tableau 2 E**) ou deux dipôles induits (forces de dispersion de London, **Tableau 2 F**). Lorsque ces phénomènes impliquent des cycles aromatiques, ils sont aussi appelés « effets π » font intervenir deux objets dont au moins un groupe aromatique d'un aa (Phe, Tyr et Trp), et créer un phénomène d'attraction dans les interactions dites « *stacking* π-π » (empilement de deux cycles aromatiques, forces de London **Tableau 2 F**), « cation-π et anion-π » (entre une charge positive et négative, respectivement, et un cycle aromatique, **Tableau 2 D**), « dipôle-π » (entre un dipôle des aa Ser, Thr, Asn, Gln, Cys, Sec et Tyr par exemple, et un benzène, forces de Debye **Tableau 2 E**) et « alkyl-π » (entre un groupe alkyle des aa Thr, Ala, Val, Ile, Leu et Met par exemple et un cycle aromatique). Il en résulte des phénomènes d'attraction ou de répulsion entre les atomes ou groupes d'atomes selon les charges portées par ces derniers. Les liaisons hydrogène sont des interactions fortes entre un atome accepteur d'hydrogène, porteur d'une charge partielle négative et un donneur d'hydrogène porteur d'une charge partielle positive (**Tableau 2 H**). Les interactions hydrophobes, moins bien définies à ce jour, forcent certains groupements fonctionnels aussi dits « apolaires » à « fuir » les molécules d'eau et donc à s'enfouir à l'intérieur de l'édifice tridimensionnel que définit la protéine. Dans les interactions engendrant une attraction, les forces de répulsion de vdW empêchent la collision des orbitales atomiques (**Tableau 2 G**).

Tableau 2 : Interactions de type non covalente ou « à distance » responsables du repliement des protéines

Type d'interaction	Modèle	Exemple	Dépendance énergie/distance
A Ion-ion Longue portée			$1/r$
B Ion-dipôle Dépend de l'orientation du dipôle			$1/r^2$
C Dipôle-dipôle (vdW, force de Keesom) Dépend de l'orientation des deux dipôles			$1/r^3$
D Effet π, cation-π et anion-π Ion-dipôle induit Dépend de la polarisabilité de la molécule dans laquelle le dipôle est induit			$1/r^4$
E Effet π, dipôle-π Dipôle-dipôle induit (vdW, force de Debye) Dépend de la polarisabilité de la molécule dans laquelle le dipôle est induit			$1/r^4$
F *Stacking* π-π Dispersion (vdW, force de London) Entre deux dipôles induits, implique une synchronisation des fluctuations des charges			$1/r^6$
G Répulsion de Van der Waals Survient lorsque deux orbitales électroniques se superposent			$1/r^{12}$
H Liaison hydrogène Attraction + liaison covalente partielle			-

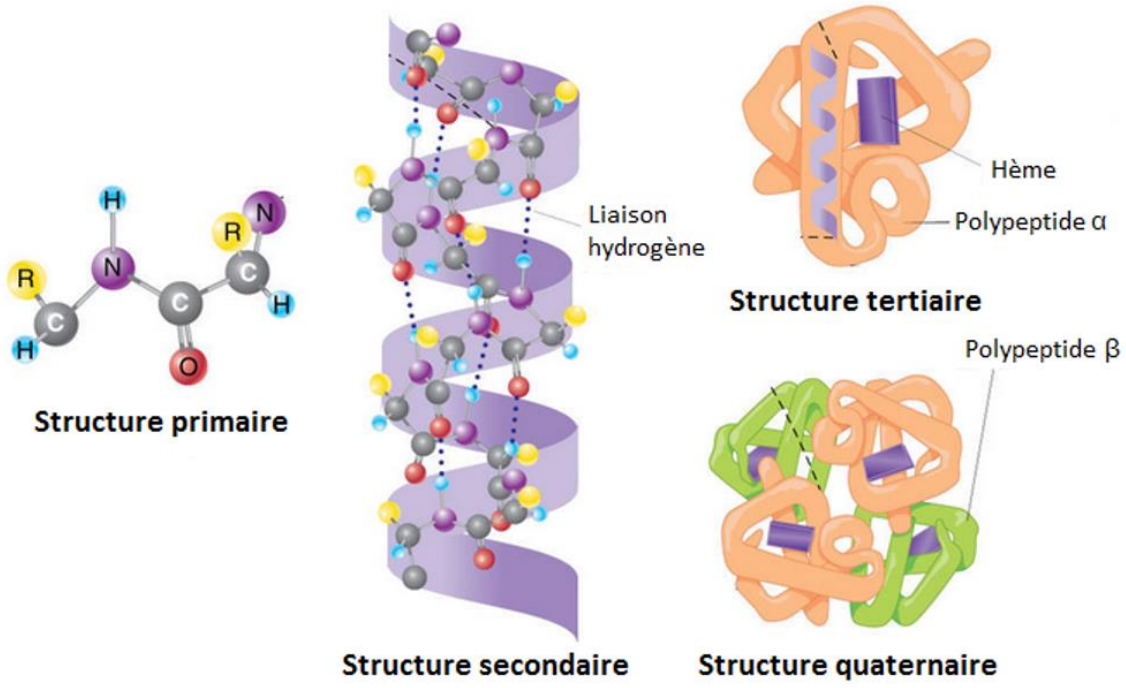


Figure 6 : Différents niveaux des structures des protéines
Les protéines peuvent être décrites avec différents niveau de structure, de la structure primaire (séquence en acides aminés) à la structure quaternaire (oligomérisation de plusieurs protéines en complexes multiprotéiques), en passant les structures secondaires et tertiaires (repliements de la chaîne polypeptidique et organisation en motifs de ce dernier).

On appelle structure secondaire le repliement local de la chaine peptidique à l'intérieur d'une protéine pour former des motifs tels que les hélices α et les brins β, repliements les plus courants, ainsi que des hélices 3_{10}, des hélices π, des ponts et des tours. Les aa ne respectant aucun des motifs cités précédemment sont donc dépourvus d'éléments de structure secondaire canoniques et adoptent une structure dite en boucle. Tous ces éléments forment le dictionnaire DSSP (« The Dictionary of Protein Secondary Structure ») qui est utilisé pour décrire les structures secondaires des protéines par des lettres en se basant sur les liaisons hydrogène formées par la chaine polypeptidique repliée et d'autres caractéristiques géométriques (**Tableau 3**) [12]. Les hélices, quel que soit leur type, sont caractérisées par des liaisons hydrogène formées entre le carbonyle -CO de l'aa i et l'amide -NH de l'aa $i + x$, x étant le nombre d'aa nécessaires pour former un tour de l'hélice (3, 4 ou 5). Ces motifs de structures secondaires peuvent interagir entre eux et former des éléments de structure dits tertiaire comme par exemple des tonneaux β, qui correspond à la succession de brins β antiparallèles organisés en un seul feuillet β s'enroulant pour former une structure quasi-cylindrique. Les protéines peuvent adopter une ou plusieurs structures tertiaires. La structure quaternaire résulte de l'interaction de plusieurs chaines peptidiques appelées sous-unités conduisant à la formation de protéines dites multimériques (dimère, trimère, tétramère, …). Par exemple les capsides protégeant l'ADN viral sont des structures composées

de globules eux même composés de plusieurs dizaines de protéines. Bien évidemment les fonctions des protéines sont établies par leur structure, qu'elles soient intrinsèquement désordonnées ou non.

Tableau 3 : Caractéristiques permettant de distinguer les éléments du dictionnaire des structures secondaires des protéines

Type d'élément	Éléments DSSP	Alphabet DSSP	Caractéristiques		Aa favorables
			Aa/tour	Translation/aa (Å)	
Hélices	Hélice α	H	3,6	1,5	Met, Ala, Leu, Glu, Lys
	Hélice 3_{10}	G	3,0	2,0	
	Hélice π	I	4,4	1,2	
Brins étendus	Brin β	E	Ap : 2,0 ; p : 2,0	Ap : 3,4 ; p : 3,2	Thr, Ile, Val, Phe, Tyr, Trp
	Pont β	B	Aa isolé formant une lh avec un brin β		
Boucles	Tour	T	Coude fermé par une lh		
	Coude	S	Pas de lh		
	Pelote	C	Aa ne rentrant dans aucune des catégories ci-dessus		Gly, Ser, Pro, Asp

Abréviations : DSSP, « Dictionary of Protein Secondary Structure » ; Aa, acide aminé; Ap, antiparallèle; p, parallèle; lh, liaison hydrogène.

I.1.3 Vue globale sur les altérations de l'ADN

Les cellules subissent constamment des agressions qui sont sources d'altérations de la nature chimique de l'ADN et donc de sa structure. Elles sont estimées à 20 000 lésions par cellule et par jour [8]. Ces phénomènes sont d'origines endogènes et exogènes. En effet la mitochondrie, fournisseur d'énergie de la cellule, produit des déchets toxiques tels que les espèces réactives de l'oxygène (ERO). La cellule subit également des agressions provenant de son environnement tels que les produits chimiques issus de l'activité humaine ou de la nature, le rayonnement solaire *via* les rayons ultraviolets (UV-A et UV-B), les radiations ionisantes (IR), comme les rayons γ et rayons X issus de la radioactivité naturelle des roches de l'écorce terrestre (formation de gaz radon) ou produite par l'Homme pour des applications médicales (thérapies anti-cancer, imagerie, stérilisation) et pour la production d'énergie (**Figure 7**). Dans cette partie, je présenterai succinctement les différentes altérations physiques et

chimiques de l'ADN (**Figure 7** et **Figure 8**) et je détaillerai les lésions de type oxydations qui sont les substrats des protéines que j'ai étudiées dans ces travaux.

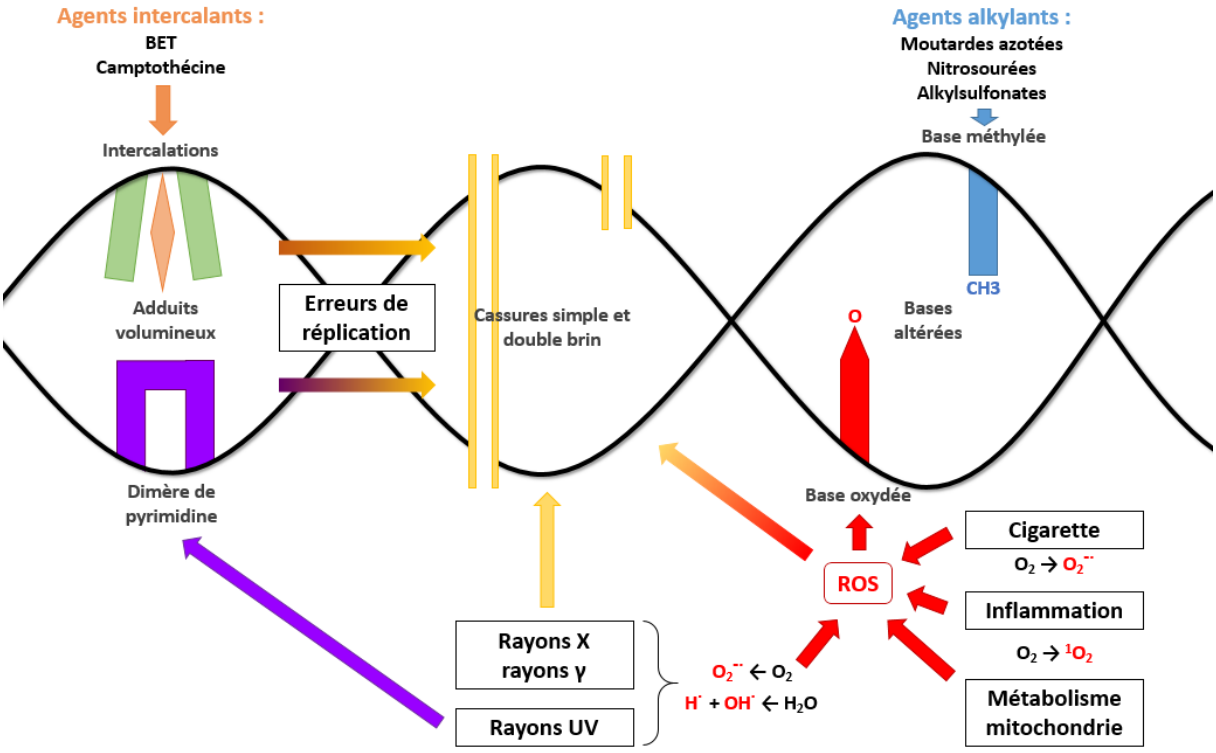


Figure 7 : Les sources d'altérations de l'ADN
Elles sont diverses et peuvent provenir de l'extérieur de la cellule (radiations ionisantes, pollution [13], agents chimiques) ou de sa propre activité métabolique (mitochondrie). L'ADN peut être endommagé directement ou indirectement par le même facteur comme par exemple les rayons UV qui produisent des ERO (responsables de l'apparition de bases oxydées) tout en induisant des dimères de pyrimidines [14]. Ces dimères déforment la double hélice de l'ADN et sont souvent à l'origine d'un blocage de la machinerie de réplication associé à l'apparition de cassures simple- et double-brin de l'ADN elles-mêmes à l'origine d'instabilité génétique [15]. Les ERO sont également impliqués dans la production de bases oxydées dans l'ADN qui induisent souvent des erreurs de réplications et donc contribuent à l'apparition de mutations (c'est par exemple le cas de la 8-oxoguanine qui induit les transversions GC → TA, voir plus loin) [16, 17].
Abréviations : $OH^{\cdot}$, radical hydroxyle ; $H^{\cdot}$, radical hydrogène ; $O_2^{\cdot-}$, radical superoxide; ${}^1O_2$, oxygène singulet.

I.1.3.1 Sources et types des dommages de l'ADN

Tous les éléments constitutifs de la molécule d'ADN (bases, sucre et lien phosphodiester) sont susceptibles de voir leurs structures natives endommagées par des agents physiques ou chimiques (**Figure 8**). Les dommages les plus couramment rencontrés dans l'ADN seront présentés dans les sections suivantes.

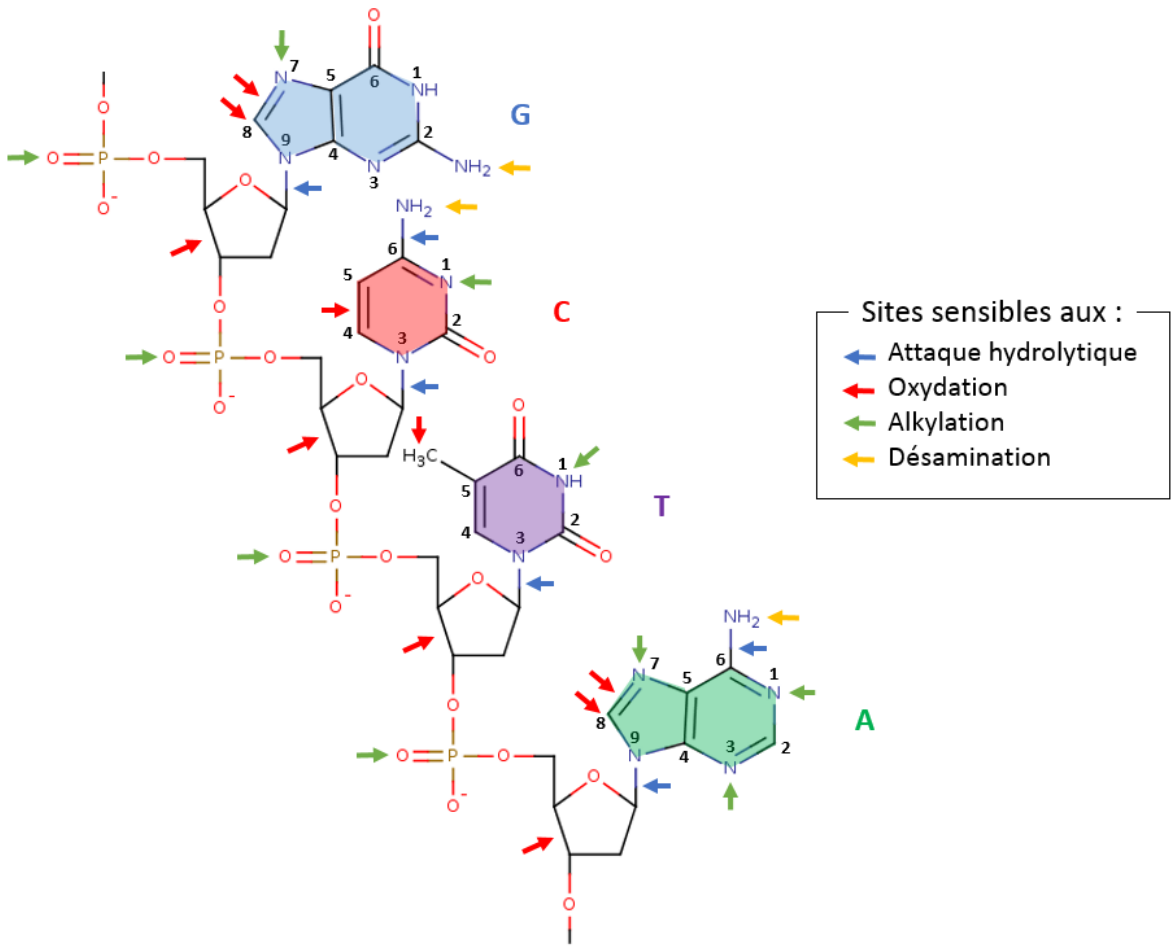


Figure 8 : Sites de l'ADN sensibles aux agressions physiques et chimiques
Ces positions des nucléotides d'un ADN sont susceptibles de subir une modification *via* des agents chimiques endogènes et exogènes [8]. Ces modifications peuvent altérer la nature, la lisibilité et la transmission de l'information génétique (hydrolyse de liaison covalente : flèche bleue ; oxydation : flèche rouge ; alkylation : flèche verte ; désamination : flèche jaune).

I.1.3.1.1 Cassures simple et double brin

Une cassure simple ou double brin est la rupture d'une liaison covalente dans le squelette phosphodiester de l'ADN sur une position dans la séquence soit sur un brin (cassure simple brin) soit sur les deux brins (cassure double brin). Ces cassures se produisent naturellement lors de la division cellulaire, notamment pendant la duplication de l'information génétique par la cellule mère (méiose/mitose) et a en même temps de lourdes conséquences pour les cellules filles si elles restent telles quelles.

En parallèle, l'interaction directe entre les photons des rayons X et rayons γ, aussi appelés radiations ionisantes (RI), avec l'ADN peuvent entrainer la rupture de liaisons covalentes dans les molécules touchées. Ces RI invisibles sont caractérisées par des photons fortement énergétiques dont

la longueur d'onde se situe entre 10^{-8} et 10^{-11} m pour les rayons X et au-delà pour les rayons γ. S'ils viennent à rencontrer un atome sur leur route, ils seront absorbés par celui-ci tout en transférant son énergie aux électrons de la couche superficielle de cet atome. Cet électron excité sera alors arraché, perturbant l'organisation des couches électroniques de l'atome et forçant les électrons à se réarranger pour atteindre un état plus stable. Les radiations ionisantes peuvent donc endommager directement l'ADN causant les cassures d'un ou des deux brins de la double hélice donnant respectivement des cassures simple brin (« Single Strand Break », SSB) ou double brin (« Double Strand Break », DSB) [15]. Les rayons X, γ et UV peuvent indirectement engendrer des DSB dans l'ADN en créant des radicaux libres qui peuvent attaquer la double hélice au niveau de son squelette : sur le sucre et/ou le groupement phosphate [18]. Plus spécifiquement aux rayons UV, la création de dimères entre bases voisines notamment les thymines sont possibles. Lorsque ces dimères se retrouvent dans la fourche de réplication, ils causent un arrêt de la réplication, ce qui induit des cassures SB [15]. Ces altérations sont extrêmement mutagènes car il peut en résulter des réarrangements et pontages intra- et inter-brins indésirables. De plus, le système de réparation des cassures SB et DB est fortement sujet aux erreurs [19].

I.1.3.1.2 Pontages intra-brin, inter-brin, ADN/protéine

L'exposition de l'ADN à des agents chimiques et aux rayonnements UV, peuvent mener à des pontages intra-brin (à l'intérieur d'un même brin), inter-brin (entre deux brins) ainsi qu'à des pontages ADN/protéine. Si ces dommages encombrants persistent, ils conduisent au blocage de la réplication et de la transcription et pour finir à la mort cellulaire. Ainsi, le cisplatine (complexe II du platine cis [$Pt(NH_3)2Cl_2$]), un agent chimique génotoxique, est utilisé en chimiothérapie contre un grand nombre de cancers. C'est une molécule très réactive et capable de former des pontages chimiques covalents deux purines d'un même brin ou de brins différents de l'ADN et entre une purine et une amine réaction d'une protéine. Dans le cas des pontages dans l'ADN, on parle d'adduits de l'ADN bifonctionnels intra- et inter-brin (**Figure 9**) [20, 21]. Ces pontages lorsqu'ils sont trop nombreux dans la molécule d'ADN sont responsables de l'initiation de l'apoptose [22]. Certains cancers étant résistants au cisplatine classique [23], d'autres études ont permis la conception de nouvelles molécules plus efficaces telles que la carboplatine, l'oxaliplatine et la nédaplatine (**Figure 10**) [24, 25].

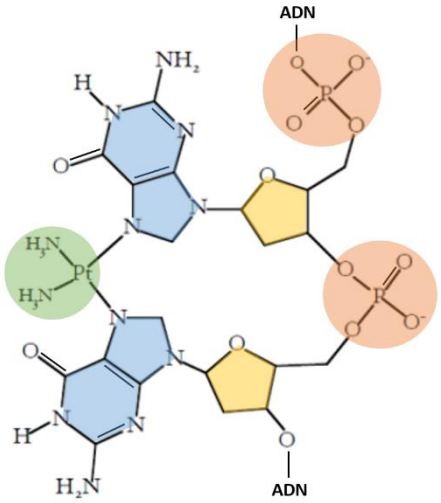


Figure 9 : Structure d'un adduits bifonctionnel intra-brin
Les guanines sont en bleues et le cisplatine en vert

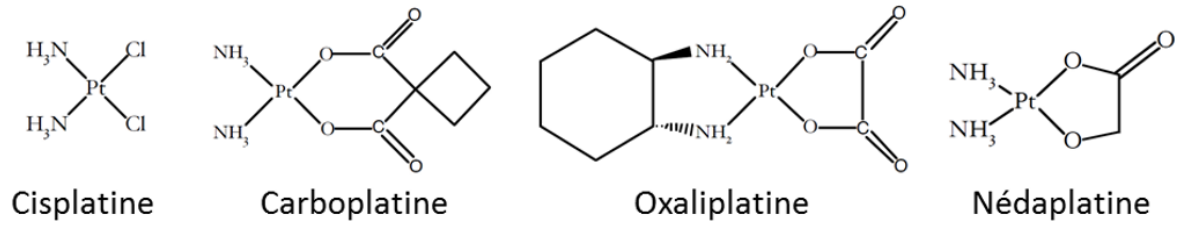


Figure 10 : Le cisplatine et ses analogues utilisés en chimiothérapie anticancéreuse

Les rayonnements UV, notamment les UV-A (315-400 nm) qui représentent 95% des UV atteignant la surface terrestre, peuvent également induire dans l'ADN des produits de pontage inter-brin tels que les dimères de thymine (engageant deux thymines adjacentes d'un même brin) sous deux formes différentes, les cyclobutanes de pyrimidines (CPDs) et les 6,4-pyrimidine-pyrimidones (6-4PP) (**Figure 11**). Ces altérations sont mises en cause dans les cancers de la peau (mélanomes) [26, 27]. Outre le cisplatine et ces analogues et les UV, il existe une multitude d'agents génotoxiques naturels ou issus de l'activité humaine susceptibles de former des pontages protéine-ADN et ADN-ADN dont nous ne parlerons pas ici [28].

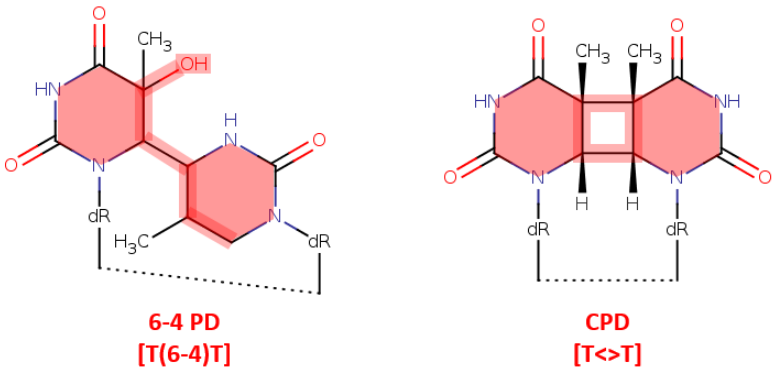


Figure 11 : Dimères de thymines générés par les rayonnements UV

I.1.3.1.3 Intercalations d'éléments étrangers dans la double-hélice de l'ADN

Les intercalants sont généralement des molécules planes, aromatiques et polycycliques capables de se glisser dans la double hélice d'ADN de façon parfois séquence-spécifique. Ils se lient de manière covalente et/ou s'empilent (interaction dite de « stacking ») entre deux plateaux de bases azotées à l'intérieur de la double hélice. Les molécules dont la base est de type acridine (hétérocycle azoté) possèdent de par leur nature aromatique de grandes capacités à s'intercaler dans l'ADN [29, 30]. Par exemple, le Camptothécine est utilisée en chimiothérapie pour tuer des cellules cancéreuses (pièges moléculaires de la topoisomérase I) [31] et le Bromure d'Éthidium (BET) en biochimie comme outil de fluorescence pour révéler l'ADN sous rayons UV [32]. L'intercalation dans l'ADN de ces molécules induit une déformation de la double hélice et ainsi un blocage de la réplication et/ou de la transcription, faisant de ces molécules des agents génotoxiques [33].

I.1.3.1.4 Alkylations des bases azotées

La méthylation de l'ADN est un processus physiologique participant à la régulation de l'expression des gènes. On parle de marquage épigénétique de l'ADN. En effet, une portion d'ADN très méthylée sera moins accessible par la machinerie de transcription, et sera donc moins exprimée qu'une portion d'ADN peu ou pas méthylée. Ce phénomène est impliqué dans l'inactivation de certains gènes et du chromosome X, le vieillissement et la carcinogénèse [34]. En temps normal, seules deux nucléotides peuvent subir une alkylation dite « physiologique » : la cytosine (C) en 5-méthyl-cytosine (5-mC) chez les eucaryotes et l'adénine (A) en N^6-méthyladénine (N^6-mA, **Figure 12**) chez les bactéries. Ces modifications spécifiques de l'ADN jouent un rôle important dans la compaction de l'ADN, la régulation de la réplication, de la transcription, de la transposition et de la réparation de l'ADN (système de réparation des mésappariements chez *Escherichia coli*) et également dans les processus de défense des procaryotes (le système de restriction/modification chez les bactéries) [35, 36]. Ces méthylations enzymatiques spécifiques sont assurées par les ADN méthyltransférases (« DNA Methyl Transferase », DNMT). Ces alkylations ne concernent en rien les propriétés d'appariement de type Watson-Crick des bases alkylées. Ainsi, la 5-mC et 6-mA ne sont pas mutagènes et ne sont pas identifiées par les systèmes catalytiques de réparation de l'ADN comme des bases endommagées. Cependant les agents chimiques alkylants (non-enzymatiques) tel que le S-AdénosylMéthionine (SAM) utilisé comme cofacteur par la cellule (facteur endogène) et les molécules telles que les moutardes azotées utilisées dans les thérapies anti-cancéreuses (molécules issues des gaz moutardes utilisés comme une arme chimique lors des première et seconde guerres mondiales) sont susceptibles d'alkyler les bases (et les phosphates) de l'ADN en positions N1 et N7 des purines et en N3 de l'adénine (**Figure 12**). Les phosphoramides ou moutardes azotées utilisés désormais en chimiothérapie ciblent de manière spécifique la position N7 des guanines de l'ADN [37] et induisent la formation de guanines aux cycles imidazoles ouverts (FaPyG), de sites abasiques et de pontages intra- (dimère composé de deux N7-mG opposées) et inter-brin [38, 39]. Ce type d'agents chimiques engendre la création d'un grand nombre de bases modifiées *via* les alkylations en O et N aussi bien sur les pyrimidines que sur les purines (**Figure 12**). Ces lésions forcent la cellule cancéreuse à rentrer dans le processus de mort programmée ou apoptose.

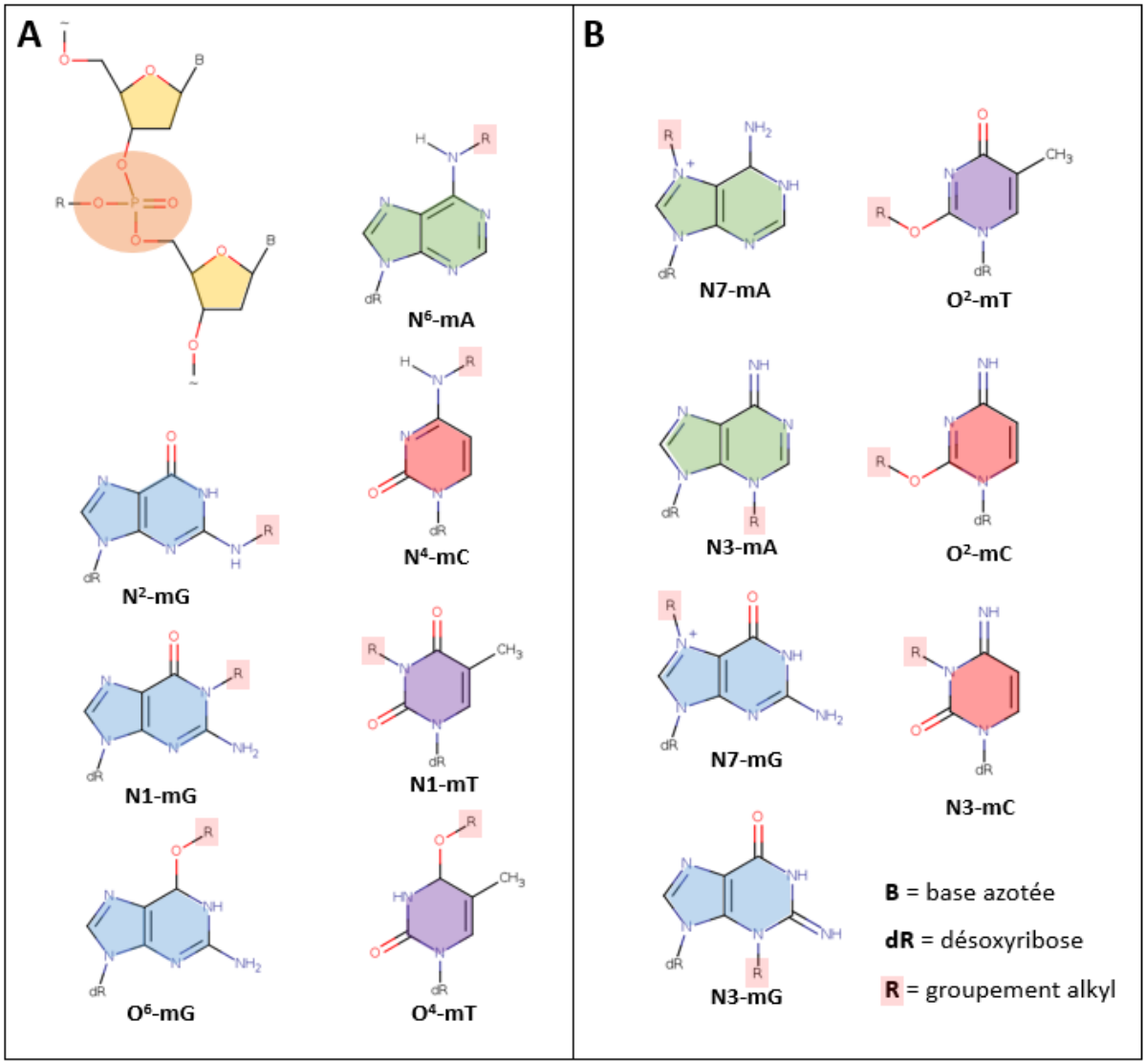


Figure 12 : Produits d'alkylation de l'ADN

A) Produits stables et **B)** instables [40]. Les produits instables conduisent à la dégradation des bases et/ou à leur perte. Par exemple, l'alkylation en N7 des purines déstabilise le cycle imidazole ce qui peut provoquer son ouverture (formation de lésions appelées formamidopyrimidines). Cette alkylation peut aussi déstabiliser la liaison N-glycosidique et aboutir à une dépurination (formation de site abasique). Cette dernière réaction d'alkylation (utilisant le diméthylsulfate) a été exploitée par Maxam et Gilbert pour séquencer l'ADN [41].

Abréviations : N1-mG, N1-méthylguanine ; N1-mT, N1-méthyltymine ; N^2-mT, N^2-méthylthymine ; N3-mA, N3-méthyladénine ; N3-mC, N3-méthylcytosine ; N3-mG, N3-méthylguanine ; N^4-mC, N^4-méthylcytosine ; N^6-mA, N^6-méthyladénine ; N7-mA, N7-méthyladénine ; N7-mG, N7-méthylguanine ; O^2-mC, O^2-méthylcytosine ; O^2-mT, O^2-méthylthymine ; O4-mT, O4-méthylthymine ; O6-mG, O6-méthylgunanine.

I.1.3.1.5 Désaminations des bases azotées

La désamination des bases azotées de l'ADN peut être contrôlée ou spontanée. Lorsqu'elle est contrôlée, elle est réalisée par les enzymes appelées désaminases et est nécessaire pour la création volontaire de mutation dans certains gènes, notamment ceux codant les protéines formant les chaines légères des anticorps. C'est l'hyper-mutagénèse somatique impliquée dans la maturation des immunoglobulines. Par opposition avec ces mécanismes enzymatiques hyper-régulés, on appelle désamination spontanée ou aussi désamination hydrolytique, le retrait d'un groupement amine d'une base azotée sous l'effet de l'eau et du pH du milieu (**Figure 13**).

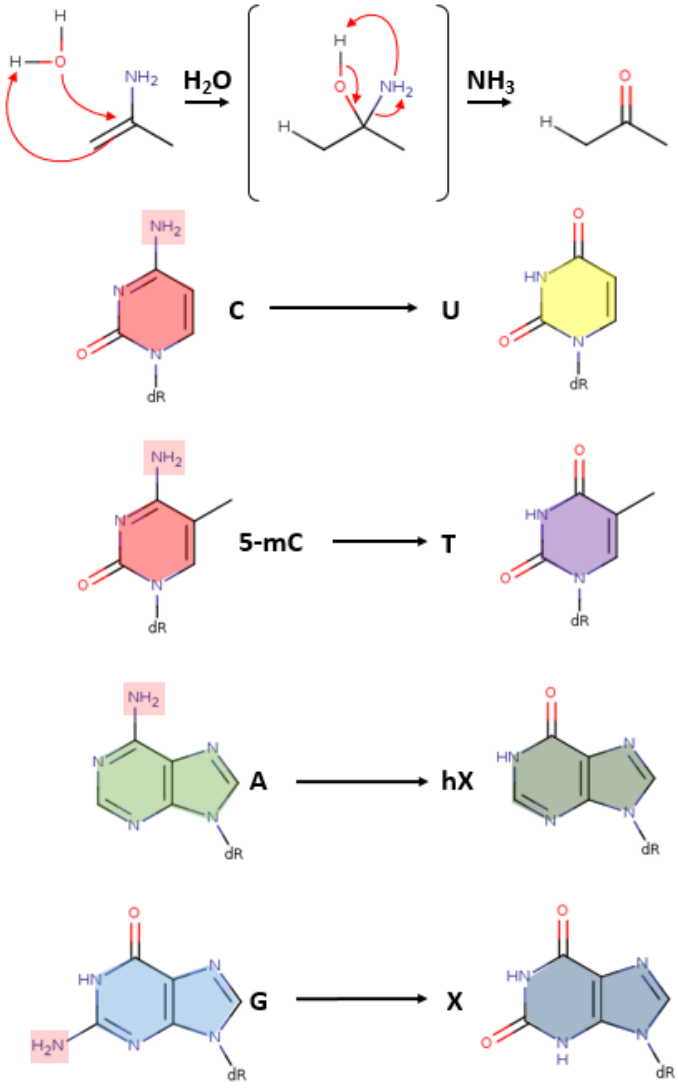


Figure 13 : Produits des désaminations contrôlées ou accidentelles des bases azotées de l'ADN
Les positions impactées par cette altération sont surlignées en rouge pâle.
Abréviations : 5-mC, 5-méthylcytosine ; hX, hypoXanthine ; X, Xanthine.

Ce phénomène ne dépend pas de facteurs extérieurs à la cellule, et se produit rarement mais suffisamment pour justifier la présence de systèmes de réparation de ces lésions car elles peuvent conduire à des mutations si elles ne sont pas réparées. En effet, la désamination de la cytosine en C4 par hydrolyse spontanée engendre l'apparition d'un uracile (U), et du mésappariement U.G, pouvant induire la mutation de type transition C:G → T:A. La désamination de la 5-méthylcytosine (5-mC) en thymine, aboutit au mésappariement T.G puis à la mutation C:G → T:A. La désamination de la guanine en C2 et de l'adénine en C6 les convertit respectivement en xanthine (X) et hypoxanthine (hX). Leurs propriétés d'appariements (X.T et hX.C) conduisent respectivement aux mutations de type transitions G:C → A:T et A:T → G:C.

I.1.3.1.6 Oxydation des bases azotées et du squelette phosphodiester

Les espèces réactives de l'oxygène (ERO ou « Reactive Oxygen Species », ROS en anglais) sont créées sous l'effet d'une intense excitation des électrons situés sur la couche extérieure d'un atome d'oxygène. Les ERO ne cherchent qu'à se combiner avec une autre molécule pour combler le vide dans leurs couches de valence et atteindre un niveau d'énergie plus bas. Ce sont des espèces chimiques possédant sur leur couche électronique externe un ou plusieurs électrons non appariés. Ces électrons sont libres, et notés par un point « $^{\cdot}$ » pour les espèces radicalaires comme le radical hydroxyle $OH^{\cdot}$ et l'ion superoxyde $O_2^{-\cdot}$ [42].

Une source importante de radicaux hydroxyles provient de la réduction des métaux de transition par le peroxyde d'hydrogène issue de la réaction de Fenton (**1**) [43]. Cette réduction peut être catalysée par le fer, le cuivre ou le chrome. De plus, cette réaction peut se dérouler au sein d'une cellule en présence de protéines portant des ions métalliques et de peroxyde d'hydrogène [13].

Équation 1 : Réaction de Fenton

$$(1)\ Fe^{2+} + H_2O_2 + H^+ \leftrightarrow Fe^{3+} + OH^{\cdot} + H_2O$$

Les rayons X utilisés en radiothérapie, les rayons γ issus de la radioactivité naturelle ou non peuvent endommager indirectement l'ADN *via* la création d'espèces chimiques nommées ERO lors de la radiolyse de l'eau et du dioxygène présents dans la cellule. La photolyse de l'atmosphère (notamment de la couche d'ozone) et de l'eau par les rayons UV-A et UV-B produits par le soleil sont également responsables de la création d'ERO. Les molécules d'eau et de dioxygène dissoutes sont cassées lors du phénomène d'ionisation sous l'effet d'un rayonnement énergétique intense. Ainsi, la

radiolyse de l'eau (2) et du dioxygène (3) produit des radicaux libres de l'oxygène tel que le superoxide ($O_2^{-\cdot}$), le radical hydroxyle ($OH^{\cdot}$), et d'autres ERO non radicalaires tel que l'oxygène singulet ($^1O_2$) ainsi que des radicaux libres d'hydrogène ($H^{\cdot}$).

Équation 2 : Radiolyse de l'eau

$$(2)\ H_2O \leftrightarrow OH^{\cdot} + H^{\cdot}$$

Équation 3 : Radiolyse du dioxygène

$$(3)\ O_2 + e^- \leftrightarrow O_2^{-\cdot}$$

Les réponses inflammatoires par exemples liées à une infection sont aussi génératrices d'ERO. En effet, les neutrophiles produisent le radical superoxyde et du peroxyde d'hydrogène qui peuvent réagir ensemble dans la réaction d'Haber-Weiss (4) [44, 45].

Équation 4 : Réaction d'Haber-Weiss

$$(4)\ O_2^{-\cdot} + H_2O_2 \leftrightarrow OH^{\cdot} + OH^-$$

Par ailleurs, la cellule produit elle aussi des ERO *via* son métabolisme notamment lors de la respiration cellulaire (5) siégeant dans les mitochondries et permettant la synthèse d'adénosine triphosphate (ATP), une molécule riche en énergie essentielle au métabolisme cellulaire.

Équation 5 : Respiration cellulaire

$$(5)\ O_2 + e^- \rightarrow O_2^{-\cdot} + e^- + 2H^+ \rightarrow H_2O_2 + e^- + H^+ - H_2O \rightarrow OH^{\cdot} \rightarrow H_2O$$

Les ERO sont captés par les peroxysomes (lors de la neutralisation par les catalases et la superoxyde dismutase), organites cellulaires qui les réduisent et les décomposent en molécules d'eau et dioxygène, molécules inoffensives. Lorsque les peroxysomes ne sont plus capables de contenir et/ou de traiter les ERO, ceux-ci peuvent être relâchés entrainant un stress oxydatif important pour la cellule, ce phénomène est appelé pexophagie. Les ERO sont une des causes de la modification de la structure chimique des bases de l'ADN, créant notamment des bases oxydées [46-49].

Il existe de nombreux produits d'oxydation des bases que nous ne passerons pas tous en revue mais dont les plus connus sont présentés dans la **Figure 14**.

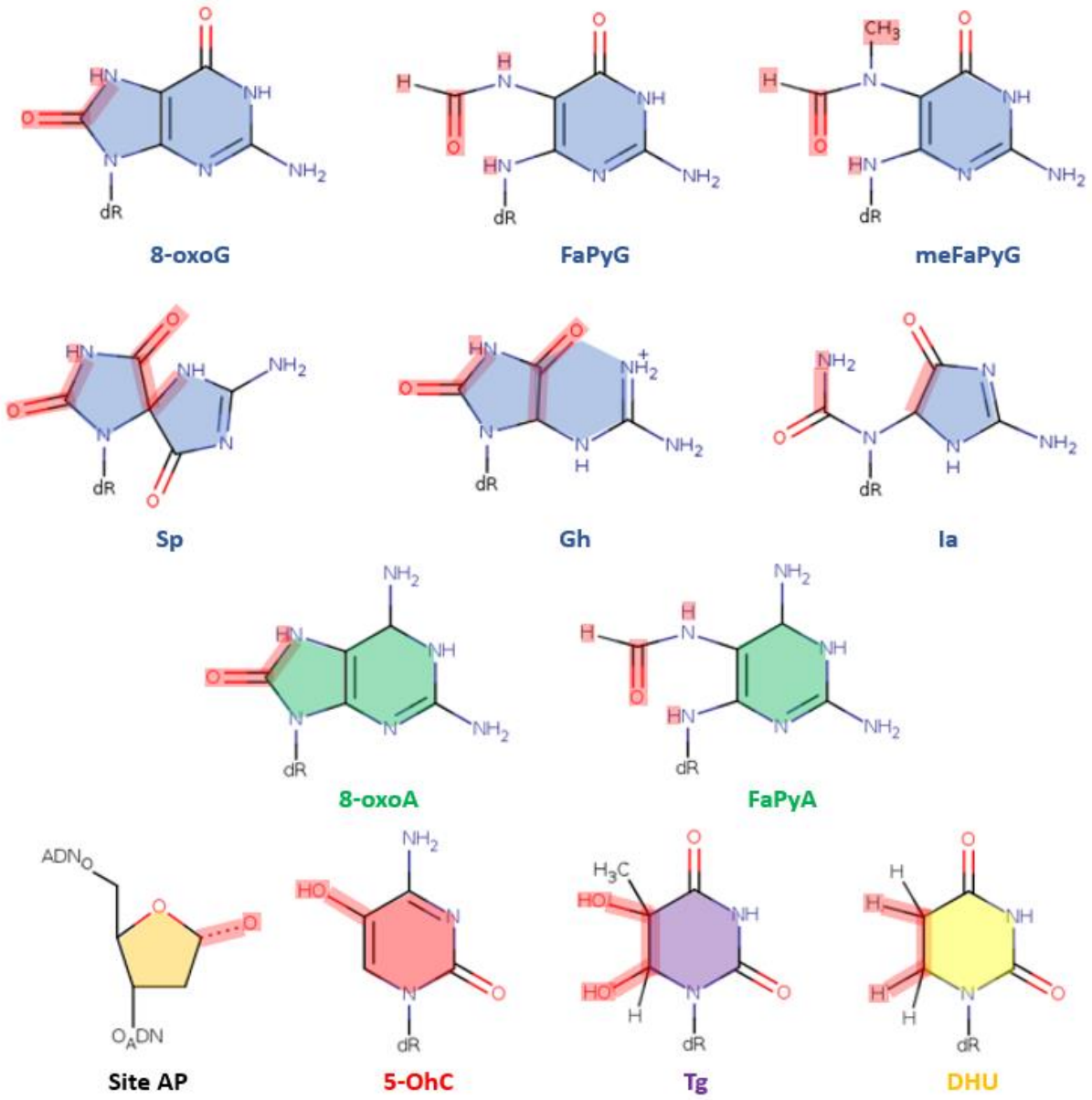


Figure 14 : Liste non-exhaustive des bases oxydées de l'ADN
Oxydations possibles des bases azotées de l'ADN (8-oxoG, FaPyG, me-FaPyG, Sp, Gh, Ia, 8-oxoA, FaPyA, 5-ohC, Tg, DHU) [50]. Les modifications de structure sont surlignées en rouge.
Abréviations : 8-oxoG, 8-oxoguanine, FaPyG/A, formamidopyrimidine ; me-FaPyG, méthyl-FaPyG ; Sp, Spiroiminodihydantoine ; Gh, Guanidinohydantoine ; Ia, Iminoallantoine ; 8-oxoA, 8-oxoadénine ; Site AP, site abasique ; 5-ohC, 5-hydroxycytosine ; Tg, Thymine glycol ; DHU, 5,6-dihydrouracile ; εA, éthénoadénine ; 5-foU, 5-formyluracile ; 6-4PD, 6-4 Dimère de Pyrimidines ; CPD, Dimère de Pyrimidines Cyclobutane.

La guanine est particulièrement sensible au stress oxydant car elle possède le potentiel redox le plus bas des bases de l'ADN. L'un des produits majeurs d'oxydation de la guanine dans l'ADN est la 8-oxo-7,8-dihydroguanine (8-oxoG) (**Figure 15 A**). Cette lésion est créée par l'attaque en C8 de la guanine par d'un radical hydroxyle ou par l'oxygène singulet. Du fait de ses propriétés particulières d'appariement, la 8-oxoG est une lésion de l'ADN extrêmement mutagène (**Figure 15 B**). En effet elle peut s'apparier indifféremment avec une C pour former une paire Watson-Crick classique ou avec une

A pour former une paire de type Hoogsteen. En raison de ces propriétés particulières d'appariement, les ADN polymérases insèrent généralement une C ou une A en face d'une 8-oxoG contenue dans le brin matrice avec une fréquence de 50 %. À la suite d'un deuxième cycle de réplication, il en résulte une transversion G:C → T:A [16, 17]. L'oxydation d'une guanine par les radicaux hydroxyles peut aussi aboutir à la formation de 2,6-diamino-4-oxo-6-formamidopyrimidine (FaPyG), une guanine dont le cycle imidazole est ouvert. Le FaPyG peut être mutagène exactement pour les mêmes raisons que la 8-oxoG (transversion G:C → T:A) ou il peut bloquer la réplication une base avant la lésion lorsqu'il est alkylé en N7 [51, 52]. D'autres altérations découlent aussi de la 8-oxoG, comme la spiroiminodihydantoine (Sp), la guanidinohydantoine (Gh) ainsi que son isomère, l'iminoallantoine (Ia) (**Figure 14**) décrites comme facteurs de mutations G:C → T:A et C:G [53].

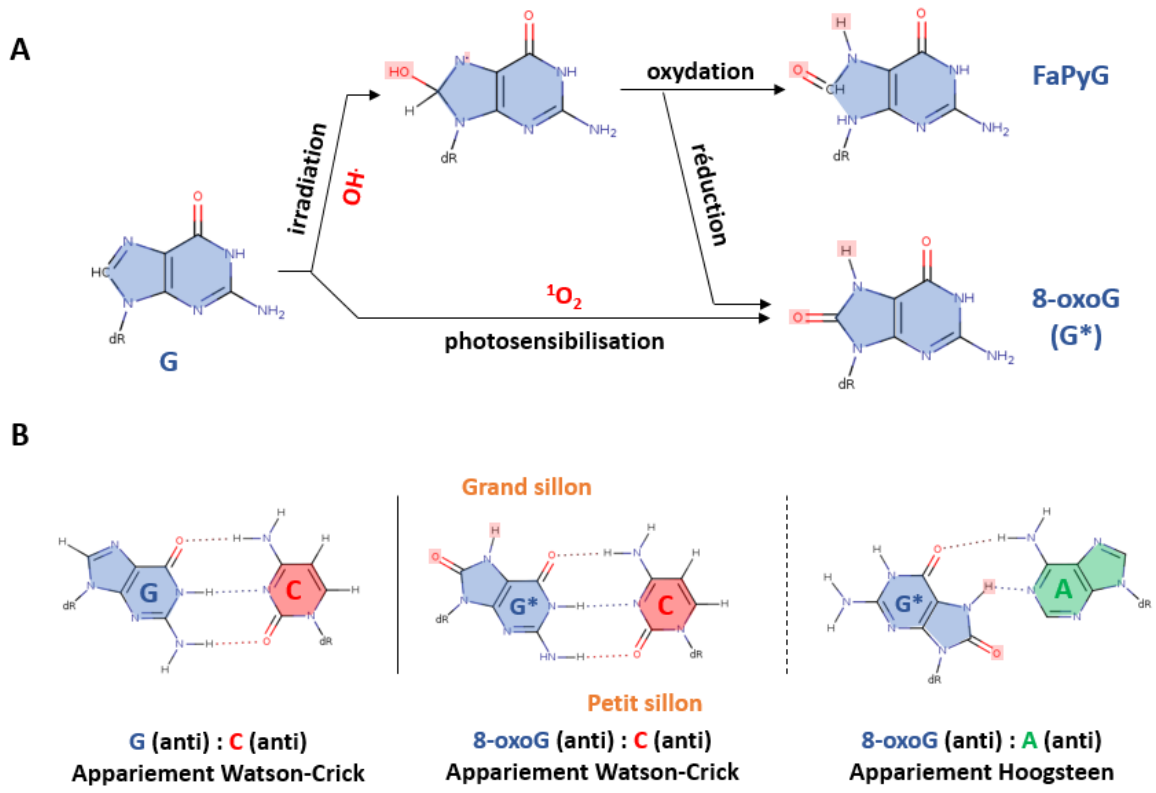


Figure 15 : Induction de la 8-oxoguanine dans l'ADN, élément mutagène
A) Deux mécanismes pour l'oxydation de la guanine en 8-oxoguanine (8-oxoG ou G*) dans l'ADN et ensuite **B)** Propriétés d'appariement de la 8-oxoguanine. La plupart des 8-oxoguanine ADN glycosylases, enzymes participant à la réparation de l'ADN, sont capables d'éliminer efficacement la 8-oxoG uniquement lorsque celle-ci est appariée à une cytosine [54].

L'adénine peut également être la cible d'oxydations ce qui conduit à la création de 4,6-diamino-5-formamidopyrimidine (FaPyA) et de 7,8-dihydro-8-oxo-adenine (8-oxoA). Dans de moindres mesures que la 8-oxoG et le FaPyG elles sont tout de même mutagènes et peuvent générer des transversions correspondantes A:T → G:C et A:T → C:G [17].

Les pyrimidines ne sont pas épargnées par les phénomènes d'oxydation, notamment sur les positions 5 et 6 de la thymine générant les bases oxydées 5,6-dihydroxy-5,6-dihydrothimine ou Thymine glycol (Tg), dihydrothymine (DHT), dihyrouracile (DHU), 5-hydroxyuracile (5-OHU), 5-hydroxycytosine (5-OHC), 5-hydroxyméthyluracile (5-hmC) et 5-formyluracile (5-fU) (**Figure 14**) [55].

I.1.4 Les systèmes de réparation de l'ADN

Au cours de l'évolution seuls les organismes comportant des stratégies de défense pour lutter contre les altérations de leur ADN ont persisté. Il existe plusieurs mécanismes de réparation pouvant prendre en charge un grand nombre de modifications de l'ADN (**Tableau 4**) [50]. Les protéines participant à la réparation sont aussi impliquées dans d'autres voies cellulaires telles que la réplication et la transcription de l'ADN, notamment les ADN polymérases β/ε/δ intervenant dans le système de réparation par excision de base et les facteurs de transcription XPA/C/G participant également au système de réparation par excision de nucléotide. Il existe deux grandes stratégies de réparation qui dans leurs principes de base ont été conservées de la bactérie jusqu'à l'Homme : (i) la réparation par réversion directe (DRR) n'impliquant pas de remaniement important de l'ADN et qui consiste simplement en la catalyse de la réaction inverse à celle qui a conduit à la lésion et (ii) la réparation par excision-resynthèse qui consiste en l'excision du dommage conduisant de façon transitoire à une cassure/lacune de une à plusieurs nucléotides puis à son comblement par une ADN polymérase et à la restauration d'une molécule d'ADN continue par une ADN ligase. La réparation par excision-resynthèse exploite la nature bicaténaire de l'ADN qui suppose de recopier une séquence d'ADN par complémentation tant que l'un des deux brins d'ADN porte l'information correcte. À l'exception du système de réparation par excision de base (BER) qui sera décrit en détails dans le chapitre suivant **I.2** (p. **65**), les autres mécanismes de réparation seront présentés succinctement dans les sections suivantes.

Tableau 4 : Liste des sources et des lésions communes de l'ADN ainsi que les mécanismes de réparation et le nom des enzymes d'*E. coli*/*H. sapiens* impliqués correspondants

Agents provocant les modifications	Altérations de l'ADN	Mécanismes de réparation majeurs	Noms des enzymes réparatrices (*E. coli*/*H. sapiens*)
Agents alkylants	O^6-mG, O^4-mT	DRR	Transférases : Ada et Ogt/Agt (MGMT)
	1-mA, 3-mC	DRR	Oxydoréductases : AlkB/Abh2
	3-mA, 3-mG, 7-mA, 7-mG	BER	Glycosylases : AlkA/Aag
Hydrolyse	Sites abasiques	BER	Endonucléases : EndoIV/APE1
	Désamination (de C donnant U)	BER	Glycosylases : Ung
	Désamination (de C donnant U et de G donnant une hypoxanthine)	NIR	Endonucléase : EndoV
ERO	8-oxoG, FaPyG, faPyA, Tg, 5-ohC, DHU, DHT	BER	Glycosylases : Fpg/NEIL1, Nth/OGG1, Nth1
	5-ohC, DHU, DHT	NIR	Endonucléases : EndoIV/Ape1
Erreurs de réplication	Mésappariements de base Insertion/délétion	MMR	Protéines de mésappariements : MutS, MutL, MutH/MutSα/β, MutLα
Rayons UV	Adduits volumineux	NER	XPA-XPF et autres*
	CPDs, 6-4PDs	DRR	Photolyases CpD et (6-4)*

Abréviations : ● **Agents provocant les modifications :** ERO, Espèce Réactive de l'Oxygène ; UV, Ultra-Violet ● **Altérations de l'ADN :** O^6-mG, O^6-méthylguanine ; O^4-mT, O^4-méthylthymine ; 1-mA, 1-méthyladénine ; 3-mA, 3-méthyladénine ; 3-mG, 3-méthylguanine ; 7-mA, 7-méthyladénine ; 7-mG, 7-méthylguanine ; 8-oxoG, 8-oxoguanine ; 8-oxoA, 8-oxoadénine ; FaPyG/A, formamidopyridine guanine/adénine ; Tg, Thymine glycol ; 5-ohC, 5-hydroxycytosine ; DHU, 5,6-dihydrouracil ; DHT, 5,6-dihydrothymine ; CPDs, Dimères de Pyrimidines Cyclobutane ; 6-4PDs, 6-4 Dimères de Pyrimidines.
● **Mécanismes de réparation majeurs** : BER, Réparation par excision de base ; DRR, Réparation par réversion directe ; NER, réparation par excision de nucléotide ; NIR, réparation par incision de nucléotide.

● **Noms des enzymes réparatrices :** Ada, O^6 Alkylguanine transférase I (O^6 AGT I) ; Agt, O6-Alkylguanine transférase ; AlkB/Abh2, dioxygénases Fe(ii) et acide α-cétoglutarique dépendentes ; AlkA, 3-méthyladénine ADN glycosylase ; EndoIV, Endonucléase IV ; Ape1, AP Endonucléase 1 ; Ung, Uracile ADN glycosylase ; EndoV, Endonucléase V ; Fpg, Formamidopyrimidine ADN glycosylase ; MGMT, O^6-méthylguanine ADN méthyltransférase ; MutS/L/H/Sα/Sβ/Lα, protéines Mut ; Nth/1, Endonucléase III /1 ; OGG1, 8-oxoguanine ADN glycosylase 1 ; Ogt, O^6-alkylguanine ADN alkyltransférase II (O^6 AGT II) ; XPA-XPF, protéines XPA à XPF.
* Protéines de réparation chez la bactérie uniquement.

I.1.4.1 Réparation par réversion directe (DRR)

La réparation par réversion directe (« Direct Reversal Repair », DRR) est la façon la plus directe de réparer les bases altérées de l'ADN (**Figure 16**). Elle ne concerne que les bases dont la structure chimique est modifiée ou autres altérations du squelette phosphodiester (mais pas les cassures SB et DB) de l'ADN. Ce mécanisme ne concerne que deux types d'altération de l'ADN : les dimères de pyrimidine et certaines bases méthylées [56]. Il permet la réparation sans ouverture ni retrait de base ou nucléotide ou resynthèse de brin d'ADN dans la double hélice. Il fait intervenir différentes protéines telles que les CPD et (6-4)TT photolyases [57], les protéines Ada, Ogt et Agt [58] (aussi appelée MGMT pour Methyl Guanine Methyl Transferase) ou les oxydodéméthylases de la famille AlkB(procaryote)/Abh(homologue humain) (attention, ces protéines n'ont rien à voir avec les ADN glycosylases AlkA, AlkC ou AlkD) selon le type de dommage [59].

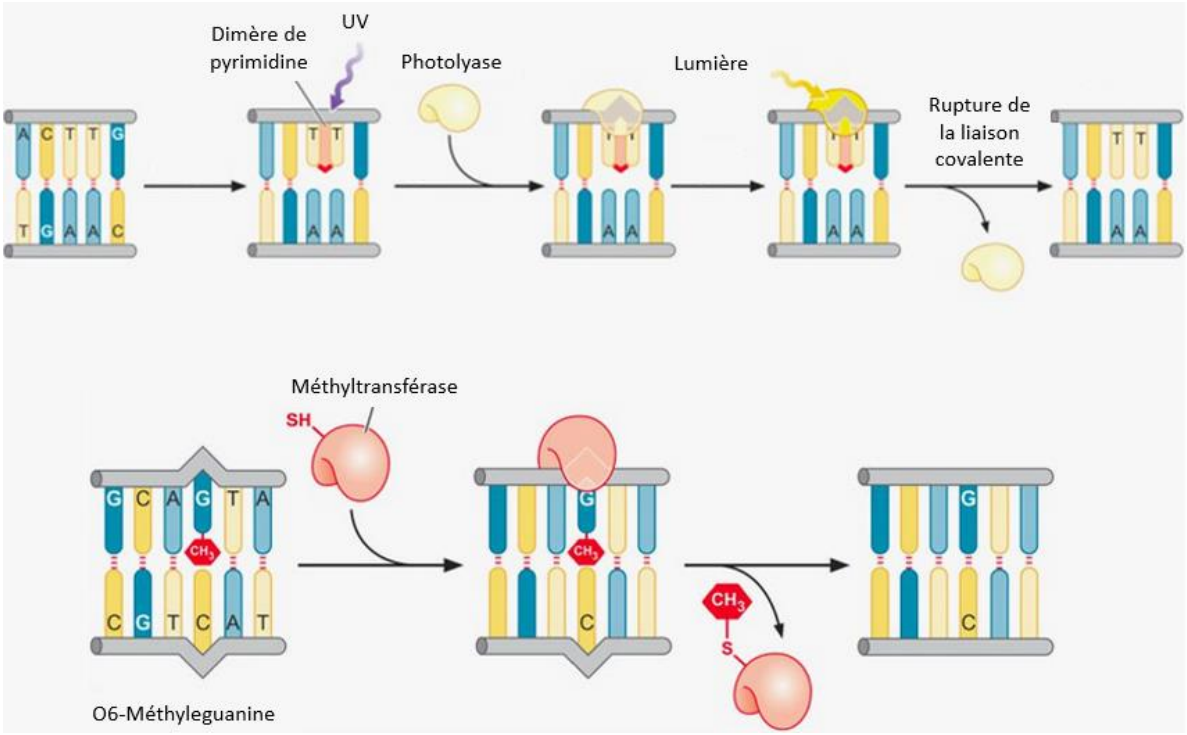


Figure 16 : Système de réparation par réversion directe (DRR)

Ce système de réparation porte bien son nom car il ne nécessite pas d'ouvrir la double hélice pour l'excision des nucléotides ou bases endommagées. Il fait intervenir des photolyases (CPPD et (6-4)TT photolyases) et des méthyltransférases (Ada, Ogt, Agt) [56].

I.1.4.2 Systèmes de réparation de cassures simple brin (SSBR) et double brin (DSBR)

Les altérations physiques de l'ADN telles que les cassures simple et double brin (« Single Strand Break », SSB et « Double Strand Break », DSB) peuvent être causées par les rayons γ et X (radiations de fortes énergies) ou par blocage de la réplication au niveau d'un dommage. Elles doivent être réparées car elles représentent un problème majeur dans l'expression de l'information génétique et la réplication de l'ADN.

Le système de réparation des cassures simple brin (« Single Strand Break Repair », SSBR) est le plus efficace car il ne commet pas d'erreur, et fait intervenir des facteurs de recrutements de protéine tels que la Poly(ADN-Ribose) Polymérase 1 (PARP1) et la « X-ray Repair Cross-Complementing protein 1 » (XRCC1) ainsi que l'ADN polymérase β (Polβ) aussi impliquées dans les systèmes de réparation par excision de base ou excision de nucléotide (mécanismes présentés plus loin dans le manuscrit), et le système de réparation des mésappariements. Si en plus d'une coupure, une ou plusieurs nucléotides sont manquantes dans la séquence de l'ADN elles seront resynthétisées et réintégrées en se basant sur les informations contenues sur le brin complémentaire.

La réparation des cassures double brin (« Double Strand Break Repair », DSBR) est plus hasardeuse et peut parfois entrainer des erreurs et des réarrangements dans le génome. La DSBR regroupe trois mécanismes distincts, (i) la recombinaison homologue (« Homologous Recombinaison », HR) (**Figure 17 A**), (ii) la jonction d'extrémités homologues (« Microhomology-Mediated End Joining », MMEJ) (**Figure 17 B**) et (iii) la jonction d'extrémités non homologues (« Non-Homologous End Joining », NHEJ) (**Figure 17 C**) [60]. Le choix de l'un ou l'autre de ces mécanismes se fait en fonction de la position de la cellule dans son cycle cellulaire. En effet, le NHEJ qui ne se produit qu'en phase G_0/G_1 et en début de phase S, le MMEJ intervient exclusivement lors de la phase G_1, tandis que le HR est mis en place en phase S et G_2, juste avant que la cellule entre en mitose, c'est-à-dire pendant la phase M [61, 62].

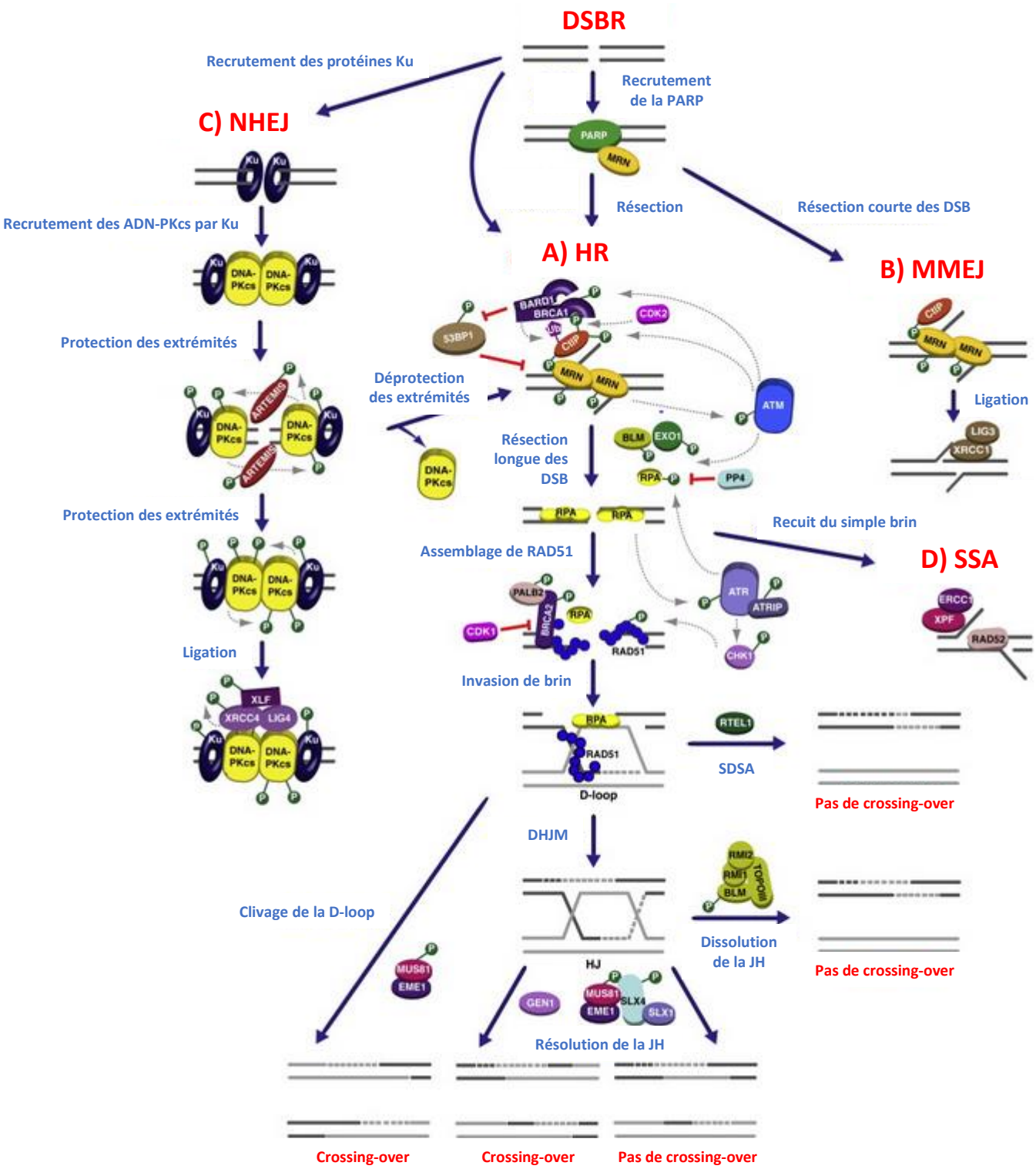


Figure 17 : Systèmes de réparation des cassures doubles brins (DSBR)

La réparation des cassures doubles brins dépend du positionnement de la cellule par rapport à son cycle cellulaire. Ces altérations de l'ADN bloquent la transcription et la réplication de l'ADN et sont donc un problème majeur pour la cellule [62].

Abréviations : DHJM, « Double Holliday Junctions Model » ; DSBR, « Double Strand Break Repair » ; HR, « Homologous Recombinaison » ; MMEJ, « Microhomology-Mediated End Joining » ; NHEJ, « Non-Homologous End Joining » ; JH (ou « HJ »), Jonction de Holliday ; SSA, « Single-Strand Annealing » ; SDSA, « Synthesis-Dependant Strand Annealing ».

I.1.4.2.1 La recombinaison homologue (HR)

La recombinaison homologue (« Homologous Recombinaison », HR) est un mécanisme essentiel chez les eucaryotes d'échange de deux séquences de nucléotides entre deux molécules d'ADN similaires ou identiques. Outre son activité dans la réparation efficace des DSB, il intervient dans le phénomène appelé enjambement (ou « crossing-over ») entre deux chromosomes homologues qui se met en place lors de la méiose pendant la création des gamètes eucaryotes permettant la diversité génétique en créant de nouvelles combinaisons de gènes [63]. La recombinaison homologue intervient aussi dans le transfert d'information génétique entre bactéries (lors des phénomènes de conjugaison ou de transformation) et lors de l'insertion d'ADN viral dans une bactérie (transduction) ou d'un autre type d'organisme (infection). C'est donc un mécanisme utilisé par tous les organismes vivants et entités, variant quelque peu selon l'organisme et le type cellulaire mais dont les principales étapes sont semblables (**Figure 17 A**). Lorsqu'une DSB se produit, la HR est initiée par la dégradation du brin d'ADN d'au moins un kilobase dans le sens 5' → 3' de part et d'autre de la cassure lors d'une étape appelée la résection. Il en résulte une extrémité 3' plus longue et ballottante sur le brin complémentaire. Chez l'Homme, ce sont le complexe MRX (Mre11, Rad50 et Nbs1) et la protéine Sae2 qui permettent la résection, en recouvrant les extrémités 5' de chaque côté de la cassure créant une superposition courte des deux extrémités 3'. Puis, les activités hélicase de Sgs1 et nucléase d'Exo1 et Dna2 permettent de terminer la résection [64]. Cette extrémité sera liée à la protéine RPA (« Replication Protein A » homologue eucaryote de la protéine SSB pour « Single-Strand Binding protein ») qui sera décalée pour permettre la fixation d'une recombinase (RecA chez les procaryotes, Rad51 chez les eucaryotes aussi couplée à BRCA2 chez l'Homme) pour assurer la stabilité et passer à l'étape d'invasion. L'étape suivante est appelée « invasion de brin », la longue extrémité 3' stabilisée peut ensuite rechercher un brin d'ADN similaire ou identique intact à envahir pour former ce que l'on appelle la D-loop (« Displacement loop ») formée d'un ADN DB hétéroduplex (composé du brin chercheur hybridé au brin homologue trouvé) et d'un ADN SB déplacé (le brin complémentaire au brin trouvé). Une ADN polymérase permet de copier l'information perdue au site de cassure lors d'une étape de synthèse réparatrice grâce au brin trouvé servant de matrice [65]. Dans certains cas on n'observe pas d'invasion et donc pas la formation d'une D-loop, les deux brins comportent eux même des séquences homologues et s'apparient entre eux, c'est le mécanisme du recuit simple brin (« Single-Strand Annealing », SSA) (**Figure 17 D**). Une fois la synthèse terminée plusieurs scénarii sont possibles, la D-loop peut être démantelée ou résolue selon deux modèles différents : (i) le modèle de réparation de cassure double brin aussi appelé le modèle de la double jonctions Holliday (« Double Holliday Junctions Model », DHJM), (ii) ou le modèle de recuit de brin synthèse dépendant (« Synthesis-Dependant Strand Annealing », SDSA) [66, 67]. La particularité du DHJM est que l'extrémité 3' non

impliqué dans l'invasion d'un brin similaire est recrutée pour former une jonction de Holliday avec le chromosome homologue. La double jonction de Holliday permet la recombinaison de portion d'ADN, correspondant à un crossing-over, et dans ce cas, la D-loop est excisée, donc résolue. Dans le modèle SDSA, la D-loop est démantelée causant la migration des branches, la nouvelle extrémité 3' rallongée à partir du brin envahi est relâchée et s'hybride avec l'autre extrémité 3' sur le chromosome endommagé grâce à la complémentarité des bases créant maintenant deux cassures simple brin, qui seront ensuite réparées. Ce processus restaure la séquence d'ADN originale et est dit conservatif [68].

I.1.4.2.2 Par jonction d'extrémités micro-homologues (MMEJ ou Alt-NHEJ)

Après le mécanisme de résection décrit ci-dessus, le MMEJ, aussi connu sous le nom de système de jonction d'extrémités non homologue alternatif (« Alternative Non-Homologous End-Joining », Alt-NHEJ), réalise la jonction entre deux brins possédant des micro-séquences homologues allant de 5 à 25 pb, ce qui fait la particularité de ce mécanisme de réparation (**Figure 17 B**) [69]. Une fois les séquences homologues identifiées sur les deux brins coupés, ils sont hybridés et une endonucléase vient retirer les surplus du chaque brin. Pour finir une ADN ligase permet de réparer les coupures de part et d'autre de la zone d'homologie. Ce système peut cependant engendrer des délétions notamment directement autour du site de coupure, et des translocations et inversions à proximité, et est donc qualifié de système non conservatif car il ne restaure pas la séquence d'ADN originale, contrairement à la HR décrite dans la section précédente [62].

I.1.4.2.3 Par jonction d'extrémités non homologues (NHEJ)

Le NHEJ est la voie de réparation la plus classiquement suivie par la cellule pour réparer les DSB car ce mécanisme n'est pas exigeant pour être mené à terme (**Figure 17 C**). En effet, et comme son nom l'indique, une homologie de séquence importante sur les deux brins de l'ADN cassés n'est pas requise. Cette voie de réparation des DSB est dépendante des protéines Ku70 et Ku80 ainsi que du facteur protéine kinase ADN-dépendent (« DNA-dependant Protein Kinase catalytic subunit », DNA-PKcs) qui permettent la détection des zones d'homologie dans l'ADN. Les DNA-PKcs couplées aux protéines Ku de part et d'autre de la cassure DB protègent les deux extrémités. Elles sont phosphorylées sur plusieurs aa par la protéine Artemis. C'est la ligase LigIV couplée à d'autres protéines qui réalise la synthèse permettant de résorber la cassure [62, 69].

I.1.4.3 Réparation de mésappariements (MMR)

Le mésappariement de bases se produit lorsque les appariements de Watson-Crick ne sont pas respectés, c'est-à-dire lorsqu'une adénine n'est pas incorporée en face d'une thymine ou encore lorsqu'une guanine n'est pas insérée en face d'une cytosine. Ce type de situation est généré par des erreurs de l'ADN polymérase lors de la réplication. Ces erreurs peuvent résulter d'un changement de tautomérie des bases (équilibre céto-énol et amino et imino) ou d'une modification chimique des bases directement dans la double hélice aboutissant dans tous les cas à un changement des propriétés d'appariement des bases (ex. : désamination d'une cytosine donnant ainsi un uracile). Les ADN polymérases se trompent fréquemment, environ 1 fois toutes les 10^{-4}-10^{-5} bases incorporées. Si l'ADN polymérase est associée à une activité 3'→ 5' exonucléase (dite de « correction »), son taux d'erreur chute à 10^{-7}-10^{-8}. Si l'on prend maintenant en compte les systèmes de réparation et notamment le MMR, le taux d'erreurs de la réplication observée *in vivo* est de 10^{-9}-10^{-10} donc extrêmement fidèle. Si l'erreur n'est pas corrigée, elle peut se transformer en mutation avantageuse, ou non après un deuxième cycle de réplication [70].

Le système de réparation des mésappariements ou « DNA mismatch repair » (MMR) reconnait et élimine les mésappariements mais aussi les erreurs d'insertion et de délétion (**Figure 18**). Ce système est brin spécifique, et est couplé à la réplication chez la bactérie. Lors de la réplication, c'est-à-dire la synthèse d'un nouveau brin d'ADN à partir d'une matrice servant de modèle, le MMR différencie le brin modèle du brin néo-synthétisé *via* un système de marquage, appelé hémi-méthylation. Chez les bactéries gram négatif, le brin « parent » est méthylé par Dam sur les A (O^6-méthyladénine) aux sites GATC tandis que le brin fils ne l'est pas encore (chez les autres procaryotes et les eucaryotes le mécanisme de reconnaissance est moins clair). Chez *Escherichia coli*, le MMR est orchestré par les protéines MutS, MutH et MutL formant le complexe MutSHL, immédiatement après la réplication. Ces protéines permettent respectivement la reconnaissance, la stabilisation et l'activation de la coupure de la base à retirer. L'excision de la base est effectuée par une exonucléase [71]. En dehors de la réplication, d'autres protéines telles que MutY et MutM opèrent pour réparer certains mésappariements spécifiquement causés par la 8-oxoG (**Figure 15 B**). La 8-oxoG est capable de s'apparier avec une adénine (8-oxoG.A). Ce mésappariement est reversé en 8-oxoG.C grâce à la protéine MutY, capable de retirer l'adénine en face d'une guanine pourtant non classique. C'est MutM (aussi appelée Fpg) qui permet le retrait de la 8-oxoG, prévenant ainsi la formation d'une mutation G:C → T:A [72]. Récemment, le MMR eucaryote a pu être reconstitué *in vitro* à partir de protéines humaines purifiées MutLα, MutSα, de l'exonucléase I et de l'ADN polymérase δ ainsi que du facteur de réplication C (RFC). Dans cette reconstitution, le couple MutLα et MutSα ayant détecté un

mésappariement dans la fourche de réplication produit une coupure à proximité, à partir de laquelle l'exonucléase digère, la polymérase δ resynthétise le brin digéré et l'ADN ligase I reconstitue une double hélice continue [73].

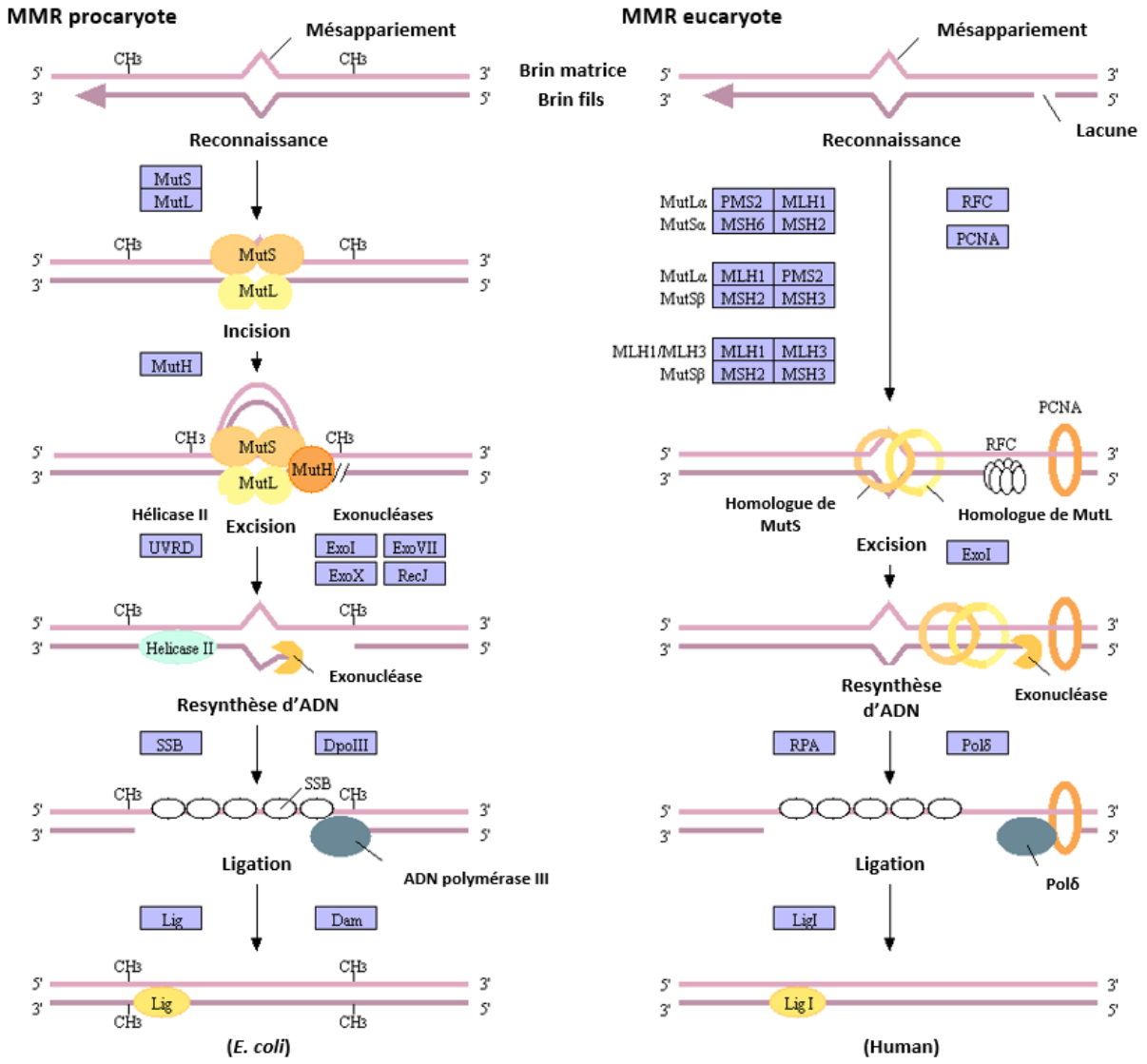


Figure 18 : Système de réparation des mésappariements (MMR) procaryote et eucaryote
Schéma issu de la base de données KEGG (« Kyoto Encyclopedia of Genes and Genomes ») représentant les acteurs intervenant dans la réparation des mésappariements chez *E. coli* et chez l'Homme [74, 75].
Abréviations : MMR, « MisMatch Repair ».

I.1.4.4 Réparation par excision de nucléotides (NER)

Le système de réparation par excision de nucléotides (« Nucleotide Excision Repair », NER) est impliqué dans la réparation de réarrangement de base en dommages volumineux tels que les adduits du cysplatine et les dimères de pyrimidine induits par les UV (**Figure 19**). D'une façon générale, tout dommage induisant une forte torsion de la double-hélice et un blocage de la réplication sera pris en charge par le NER. Le NER est différent chez les procaryotes et les eucaryotes, mais suit les grandes lignes : (i) identification, (ii) excision de nucléotides autour de l'aberration et (iii) resynthèse et incorporation d'une nouvelle portion d'ADN. Chez les organismes procaryotes, ce sont les endonucléases UvrA/B/C qui interviennent dans le NER, et les protéines XPA/B/C/D/E/F/G chez les eucaryotes. Le NER eucaryote se divise en deux voies, le NER génomique global (GG-NER) et le NER couplé à la transcription (TC-NER) impliquant chacun différents groupes de protéines. Le GG-NER est capable d'opérer sur les parties du génome silencieuse ou mise en veille, contrairement au TC-NER qui comme son nom l'indique, agit en parallèle de la transcription. Plusieurs déficiences enzymatiques participant au NER (résultantes de mutations dans les gènes du NER) sont associées à des symptômes comme le retard mental et la photosensibilité (Xeroderma Pigmentosum et syndrome de Cockayne) [76, 77].

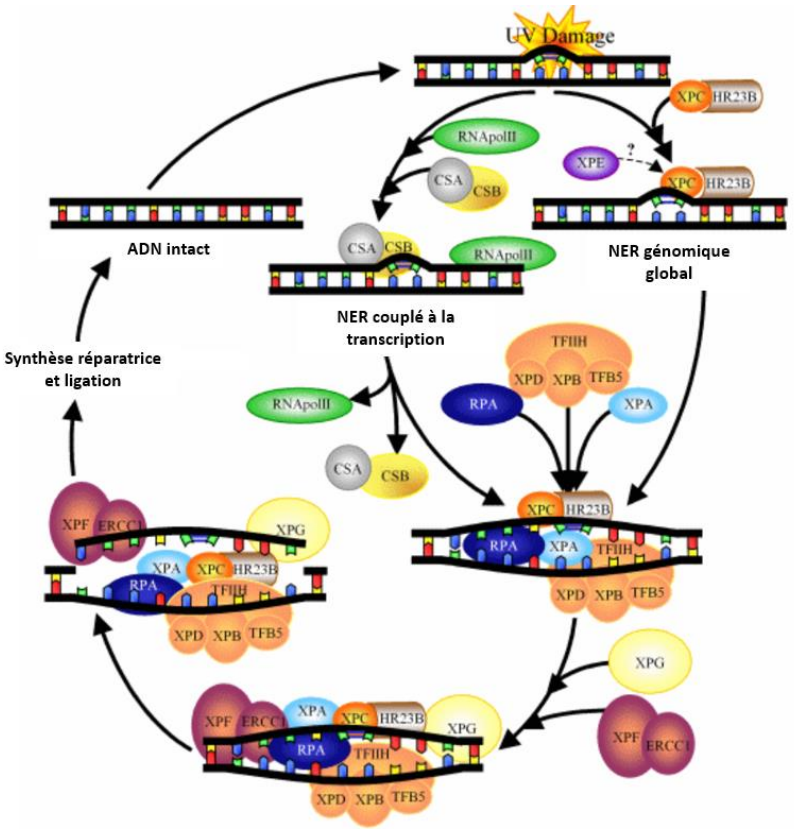


Figure 19 : Système de réparation par excision de nucléotide (NER) eucaryote
Au moment de la reconnaissance du dommage, ce mécanisme de réparation se divise en deux branches, le système de réparation NER génomique global et le système de réparation NER couplé à la

transcription. Quelle que soit la voix choisie, l'excision de nucléotide se fait de la même manière, en enlevant une portion de l'ADN et en resynthétisant un néo brin dans la lacune ainsi formée [78].

I.1.4.5 Réparation par incision de nucléotides (NIR)

Le système de réparation par incision de nucléotides (« Nucleotide Incision Repair », NIR) recrute deux classes fonctionnelles d'enzymes dans deux scénarii différents (**Figure 20**). La première classe est représentée par l'endonucléase IV (EndoIV) chez les procaryotes et l'AP endonucléase 1 (APE1) chez les eucaryotes [79], et la deuxième classe est composée de l'endonucléase V (EndoV) [80]. Les membres de la première classe peuvent reconnaitre et inciser quelques lésions résultantes d'oxydations (telles que 5-ohC, DHU, DHT), et possèdent également un rôle majeur dans un autre système de réparation par excision de base en résolvant les sites abasiques. En parallèle, l'EndoV participe aussi à la réparation de certains mésappariements T.G, U.G, Hx.C (Hx : hypoxanthine) provoqués respectivement par la désamination de la 5-méthylcytosine, de la cytosine et de la guanine [81, 82].

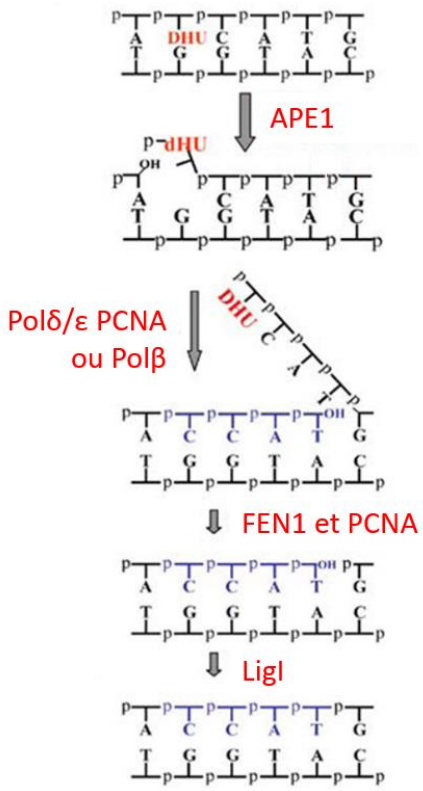


Figure 20 : Système de réparation par incision de nucléotide (NIR)
L'APE1 ouvre la double hélice au niveau de la base endommagée en clivant la liaison phosphodiester induisant deux extrémités 3'-OH et 5'-phosphate. Les ADN polymérases Polδ/ε couplées à la protéine « Proliferating Cell Nuclear Antigen » (PCNA) ou la Polβ seule synthétise un nouveau brin soulevant la section de l'ADN portant le dommage. Cette partie ballotant est appelée « Flap » et est exicée par la « Flap Endonucléase I » (FEN1). La protéine Ligase LigI reforme la liaison phosphodiester entre les deux extrémités [83].

I.2 **Chapitre 2 : R**éparation de l'ADN par excision de base (BER)

I.2.1 Mécanismes et acteurs du BER

Le système de réparation par excision de base (« Base Excision Repair », BER) est un système multienzymatique dédié à la réparation des bases endommagées et des cassures simple brin (« Single-Strand Break Repair », SSBR survenant à la suite d'irradiation ou consécutive à un blocage de la réplication). On peut considérer que le SSBR est partiellement inclus dans le BER et diffère simplement par les enzymes qui initient le processus en reconnaissant la cassure [84]. Le BER proprement dit se réalise en deux phases : (i) une phase d'excision consistant en l'élimination du dommage au travers de l'excision de 1 à 10 nucléotides contenant le dommage et (ii) une phase de resynthèse réparatrice conduisant à la reconstitution d'une molécule d'ADN normale. En fonction de la nature des enzymes qui initient le système BER, on distingue : (i) la voie à brèche courte (« Short Patch », SP) initiée par la reconnaissance et l'excision de la base endommagée par une ADN glycosylase monofonctionnelle ou bifonctionnelle et (ii) la voie à brèche longue (« Long Patch », LP) initiée soit par une ADN glycosylase bifonctionnelle, soit par la reconnaissance d'une cassure simple brin par la Poly(ADP-ribose) polymérase 1 (PARP1) assistée de la protéine XRCC1 qui participe à la stabilisation de la cassure simple-brin et au recrutement d'autres enzymes et par conséquent, au choix du SP- ou LP-BER. Il existe 12 ADN glycosylases humaines différentes contre seulement 6 chez *E. coli*. Le paragraphe suivant sera dédié à la présentation détaillée de ces enzymes car c'est sur celles-ci que mon travail de thèse a porté. Le site abasique résultant du clivage du lien *N*-glycosidique entre la base lésée et le désoxyribose par une ADN glycosylase peut être pris en charge soit par une AP endonucléase (hydrolyse du lien phosphodiester en 3' du site AP, APE1 chez l'Homme) soit par une AP lyase (activité associée aux ADN glycosylases bifonctionnelles consistant au clivage du lien phosphodiester en 3' ou en 3' et en 5' du site AP par β ou β,δ-élimination). Le choix de la voie SP-ou LP-BER peut dépendre d'abord de la nature de la base endommagée, substrat pour une ADN glycosylase monofonctionnelle ou bifonctionnelle donnée, mais peu aussi dépendre de l'abondance relative des enzymes pouvant agir en aval de l'ADN glycosylase recrutée (cette abondance relative peut varier d'un type cellulaire à un autre). La phase d'excision de la base lésée conduit immanquablement et de façon transitoire à une cassure simple brin dont les extrémités doivent être nettoyées afin de constituer une matrice correcte (avec une extrémité 5'-phosphate et 3'-OH) pour la phase de resynthèse (les protéines PNK, APE1, XRCC1 et l'ADN polymérase β contribuent à cela). Dans la voie SP-BER, c'est la polymérase β non-processive (Polβ) qui remplit la lacune (« gap ») résultat de l'excision du nucléotide endommagé (activité « gap-filling »). Les

ADN polymérases réplicatives δ et Ɛ sont recrutées dans la voie LP-BER. Finalement, la continuité de la molécule d'ADN est rétablie par une ADN ligase [85] (**Figure 21**).

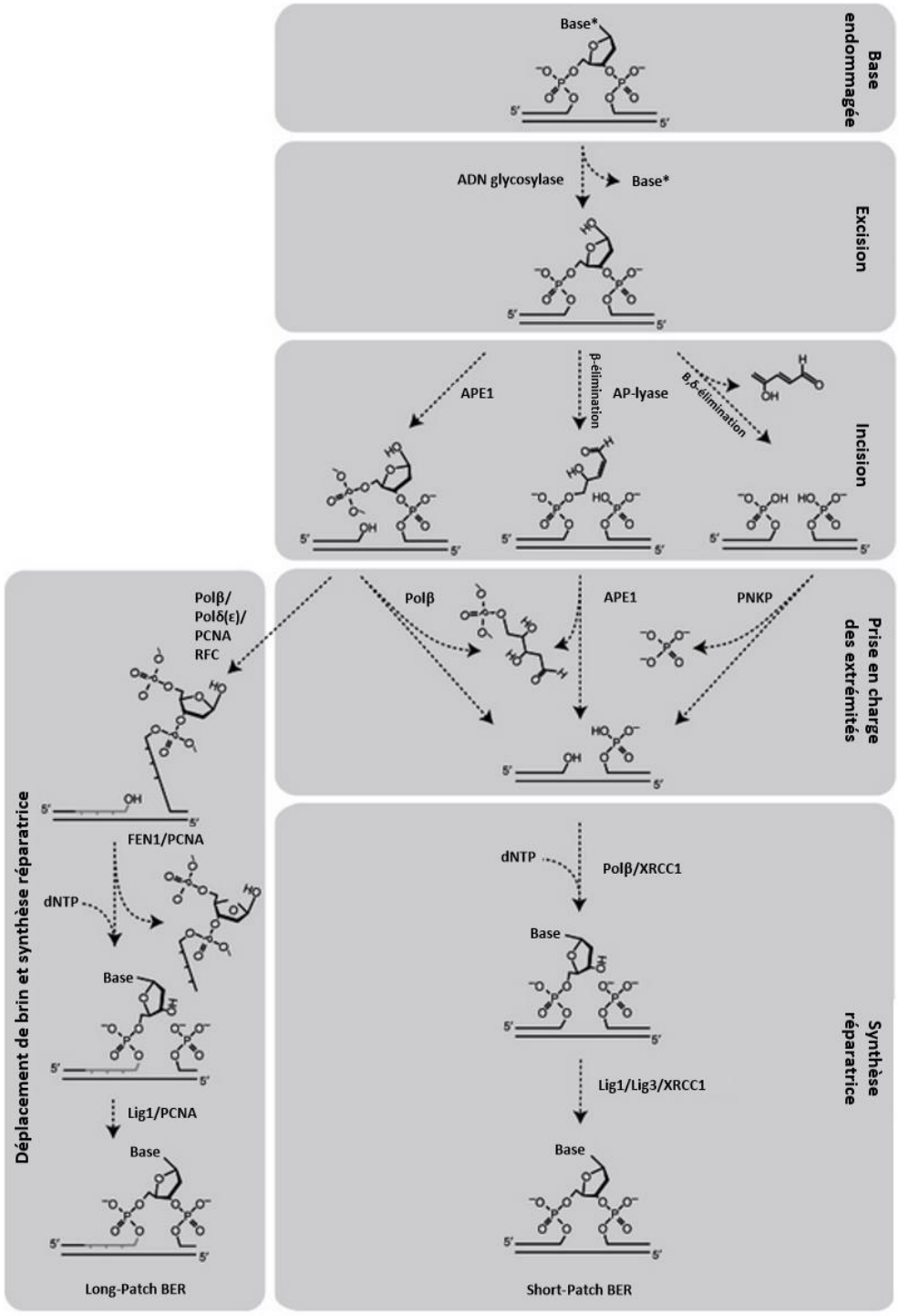


Figure 21 : Le système de réparation par excision de base (BER) chez les mammifères
Beaucoup d'enzymes différentes participent aux deux mécanismes mais les éléments clés sont les ADN glycosylases bifonctionnelles et monofonctionnelles, l'AP endonucléase 1 (APE1), la polymérase β (Polβ), la Flap endonucléase 1 (FEN1) et les ligases LigI, LigIII [84].
Abréviations : APE1, AP Endonucléase 1 ; PNKP, PolyNucleotide Kinase Protein ; Polβ/δ/ε, polymérase β/δ/ε ; LigI/III, ligase I/III ; XRCC1, X-Ray Cross-Complementation group 1 ; PCNA, Proliferating Cell Nuclear Antigen; FEN1, Flap Endonucléase 1.

Les protéines PARP1 et XRCC1 sont des protéines impliquées respectivement dans le signalement cellulaire et dans la stabilisation des SSB [86, 87]. La protéine PARP1 favoriserait également la réparation de la 8-oxoG dans des cellules Polβ déficientes [88], et serait surexprimée dans des cellules normales soumises à des agents alkylants et radiations ionisantes [89], permettant de faire un lien entre PARP1 et le recrutement du BER et de la protéine APE1 intervenant dans l'étape suivante [90]. XCRR1 pour sa part stimulerait la polymérase β intervenant une étape après [91] et régulerait l'aiguillage entre les voies SP- et LP-BER [92].

La protéine APE1 clive le lien phosphodiester en 5' (i) des sites abasiques (AP) qui résultent de la perte spontanée (ou induite) de bases ou de l'action d'une ADN glycosylase monofonctionnelle, (ii) des sites AP pré-incisés en 3' (sites 3'-phospho-aldéhyde α,β-insaturé, PUA) par une AP lyase (telles qu'une ADN glycosylase bifonctionnelle et Polβ) et (iii) d'une extrémité 3'-phosphate (activité 3'phosphodiestérase). Les produits de réaction peuvent être une cassure simple brin délimitée par une extrémité 5'-désoxyribose phosphate (5'-dRP) et 3'-OH (clivage d'un site AP) ou d'une lacune de 1 nucléotide délimitée par une extrémité 5'phosphate et 3'-OH (clivage d'un site AP pré-incisé par une ADN glycosylase bifonctionnelle). Il existe 4 classes d'AP endonucléases, les classes I et II coupent la liaison phosphodiester entre le désoxyribose et le groupement phosphate laissant deux extrémités adjacentes 3'-OH et 5'-phosphate et les classes III et IV qui coupent entre le groupement phosphate et le sucre pour laisser deux extrémités 3'-phosphate et 5'-OH. Chez l'Homme, c'est la protéine APE1 (HAP1 ou APEX) de la classe II qui est majoritairement exprimée dans 95% des cellules. C'est une hydrolase possédant un ion Mg^{2+} dans son site actif essentiel à la catalyse enzymatique. La protéine APE2 est aussi une endonucléase de classe II dont l'activité AP endonucléase est plus faible que APE1, mais qui possède également des activités 3'→ 5' exonucléase et 3'-phosphodiestérase [93, 94]. Cela signifie qu'elles sont plus efficaces sur les extrémités 3' et 5' de l'ADN. Leurs homologues chez la levure *S. cerevisiæ* sont les protéines APN1 et APN2 et Xth (Exonucléase III, ExoIII) et Nfo (Endonucléase IV, EndoIV) chez *E. coli* [95].

La lacune résultant de l'action concertée des ADN glycosylases et d'APE1 est comblée par une ADN polymérase. Il existe un grand nombre d'ADN polymérases, de I à V chez les procaryotes et de α à ν chez les eucaryotes. Chez l'Homme, c'est l'ADN polymérase non processive Polβ qui comble les petites lacunes dans la voie SP-BER [91]. La Polβ possède également une activité 5'dRP lyase capable d'enlever le groupement 5'-dRP selon un mécanisme de β-élimination après l'action de l'APE1, donc généralement lorsque le BER est initié par une ADN glycosylase monofonctionnelle [96]. Dans le cas du LP-BER ce sont les ADN polymérases réplicatives processives Polε et Polδ qui assurent l'étape de resynthèse réparatrice de l'ADN par un processus de déplacement de brin générant des structures

transitoires dites « Flap » [97]. Ces structures « Flap » sont excisées jusqu'à l'arrêt de la polymérisation par la Flap Endonucléase 1 (FEN1). Cette protéine est également impliquée dans l'élimination des structures « Flap » générées lors de la synthèse des fragments d'Okazaki et dans la maintenance des télomères. Cette protéine détient un rôle clé dans la maintenance de l'intégrité du génome, non seulement en participant à la réplication de l'ADN mais aussi en participant à la réparation de l'ADN *via* le BER [98].

L'action combinée de la Polβ et/ou Polε/δ et FEN1 conduit à la formation d'une SSB franche délimitée par des extrémités 5'-phosphate et 3'-OH, substrat pour l'ADN ligase I (LigI). Cette enzyme interagit avec « Proliferating Cell Nuclear Antigen » (PCNA, stimulant la processivité des ADN polymérases réplicatives Polε et Polδ) et est donc impliquée dans le LP-BER. L'ADN ligase III (LigIII) est recrutée par XRCC1 et est donc impliquée dans la voie SP-BER. LigIII est également impliquée dans le système de réparation des SSB et des DSB [99].

Cette description générale du BER est valable dans le noyau de la cellule. Néanmoins, il existe une variante du BER (SP et LP) dans les organites cellulaires contenant de l'ADN comme les mitochondries. C'est l'ADN polymérase γ (Polγ) propre à la mitochondrie qui assure l'étape de resynthèse du BER. Il semble que la LigIII que l'on retrouve aussi dans le noyau soit essentielle au BER mitochondrial [100].

I.2.2 Le BER et ses implications dans les fonctions cellulaires

Le BER est bien plus qu'un système de réparation car il intervient dans de nombreuses fonctions cellulaires telles que la régulation de la transcription, diversification des anticorps, le remodelage de la chromatine et des télomères. De ce fait, et lorsqu'il est déficient, il est par conséquent responsable de plusieurs maladies comme le cancer, les maladies neurodégénératives et le vieillissement prématuré (**Figure 22**).

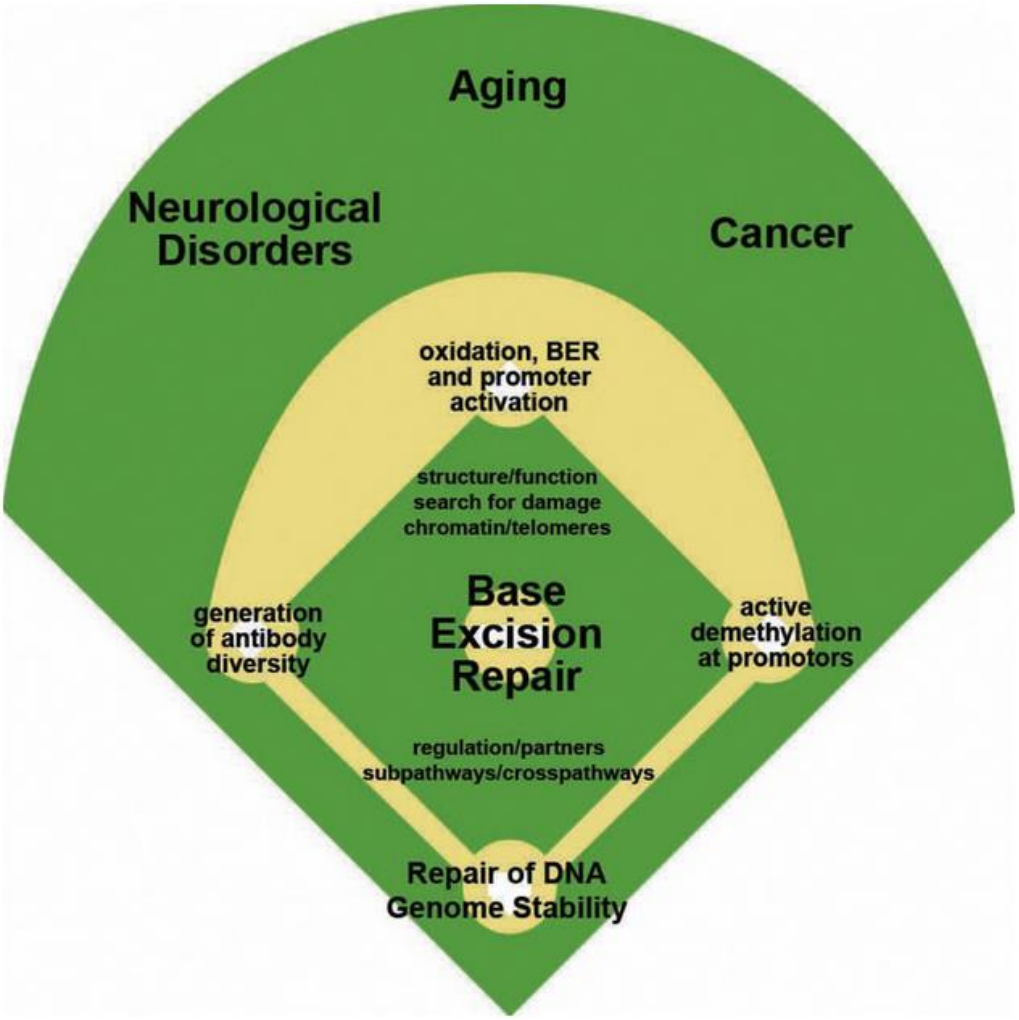


Figure 22 : Le BER, un acteur important dans le métabolisme cellulaire
Cette image est inspirée par la présentation de Failla en 1996 à la « Radiation Research Society » discutant des implications du BER dans la réparation des altérations de l'ADN induites par des radicaux libres produits par des radiations ionisantes. Elle représente un terrain de baseball avec la base principale en bas (aussi appelé quatrième base), et de droite à gauche sur la partie jaune, la première, la deuxième et la troisième base (rôles supplémentaires du BER). Pour finir dans les champs extérieurs droit, centre et gauche, se trouvent les pathologies liées à des défauts de fonctionnement du BER. Elle a été utilisée pour décrire les nouvelles activités connues des quatre plus grands acteurs du BER, les ADN glycosylases en première base, les AP endonucléases en deuxième base, les ADN polymérases en troisièmes bases et les ADN ligases en quatrième base. Le BER est donc un élément essentiel jouant sur plusieurs tableaux, « A critical player in many games » [101].

I.2.2.1 Modulation de l'expression génétique

Le BER est également impliqué dans des mécanismes épigénétiques (modulation de l'expression) en influençant la méthylation de l'ADN. Ce mécanisme est lié au remodelage de la chromatine permettant l'accès ou non de l'ADN à la machinerie cellulaire de transcription et donc à l'activation et l'inactivation de certains gènes. Ce mécanisme opère sur les gènes, les rétrotransposons (séquences d'ADN capables de se déplacer et de se multiplier donnant des séquences répétées et

dispersées) ou sur des chromosomes entiers (notamment un des doubles du chromosome X chez les mammifères femelles). La méthylation de l'ADN à la position 5 des cytosines (5-mC) est une marque épigénétique physiologique chez les eucaryotes qui joue un rôle fondamental dans le développement embryonnaire, dans la régulation de la transcription des gènes et dans bien d'autres processus cellulaires physiologiques ou pathologiques [102]. La formation des 5-mC dans l'ADN est régulée par plusieurs ADN méthyltransférases (« DNA Methyl Transferases », DNMTs), (i) la DNMT1 rétablit le profil de méthylation du brin néo-synthétisé après réplication et (ii) la DNMT3 méthyle l'ADN au niveau des îlots de dinucléotides CpG (longue succession de C terminée par un G dans le sens 5' → 3' de la séquence d'ADN) [103]. Contrairement à de nombreux produits d'alkylation des bases de l'ADN (voir la section **I.1.3.1.4** p. **44**), la 5-mC n'est pas identifiée par le système BER comme un dommage et forme une paire Watson-Crick stable à 3 liaisons hydrogène avec la guanine (5-mC.G). Bien que la 5-mdC soit essentielle, sa désamination en thymine induit dans l'ADN l'apparition du mésappariement T:G ce qui peut être associé à des sites chauds ou « hotspots » de mutation aux sites CpG (transition C → T), mutations souvent associées à des maladies génétiques et dans le génome de cellules cancéreuses chez l'Homme [104]. Récemment, des études ont montré que le BER participe à la déméthylation active de l'ADN via des produits spécifiques d'oxydation de la 5-mC (5-formyl-dC, 5-fC ; 5-carboxyl-dC, 5-caC) générées à partir de la 5-hydroxyméthyl-dC (5-hmC) (**Figure 23**) [105]. La formation de ces produits d'oxydation de la 5-mC est catalysée par les ADN dioxygénases de la famille « Ten-Eleven Translocation » (protéines TETs) [106-108]. La 5-fC et la 5-caC seraient reconnues et excisées par la thymine ADN glycosylase (« Thymine DNA glycosylase », TDG) de la superfamille des l'Uracile ADN glycosylases [106]. La 5-hmC pourrait être désaminée catalytiquement en 5-hydroxyméthyl-uracile (5-hmU) par des cytosines désaminases (« Activation-Induced cytidine Deaminase », AID ou « APOliporotein B $_{m}$RNA Editing Catalytic polypeptide », APOBECs) et opposée à une guanine et deviendrait ainsi un substrat des ADN glycosylases TDG, MBD4 (« Methyl CpG Binding Domain IV ») et SMUG1 (« Single-stranded specific Mono-fonctional Uracil DNA Glycosylase I »). Ainsi, l'action concertée et finement régulée des protéines TETs et des ADN glycosylases du BER contribue à une déméthylation active de l'ADN par un mécanisme d'oxydation contrôlée de la 5-mC [36]. L'ADN glycosylase hNEIL1 sur laquelle j'ai travaillé au cours de ma thèse pourrait être aussi impliquée dans la déméthylation active de l'ADN [109].

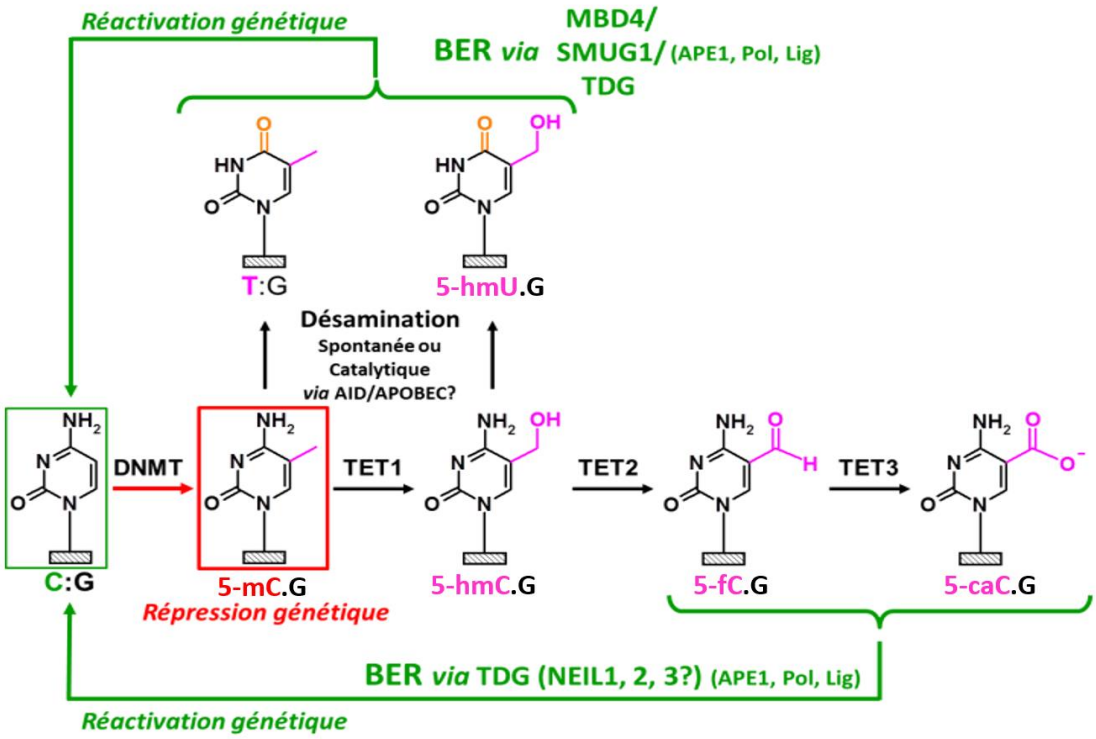


Figure 23 : Démétylation active de l'ADN via le système BER

I.2.2.2 Variabilité génétique et immunité

Les mécanismes d'hypermutagénèse somatique (« Somatic HyperMutation », SHM) et de commutation de classe (« Class Switching Recombination », CSR) dans les lymphocytes B participent tous les deux à la diversification et à la maturation des immunoglobulines du système immunitaire nécessaire pour détecter le plus grand nombre de corps étrangers pouvant être potentiellement des menaces [110]. Le SHM est le mécanisme cellulaire permettant au système immunitaire de s'adapter et de répondre en présence de corps étrangers, et le CSR consiste au remplacement d'une chaine constante de l'anticorps participant à la détection des corps étrangers. Les systèmes SHM et CSR contribuent de concert à la diversité du répertoire des immunoglobulines et à la variabilité antigénique en général. Outre leur rôle dans la réparation des altérations accidentelles de l'ADN, les ADN glycosylases UNG2 (« Uracile DNA glycosylase 2») et SMUG1 du BER des eucaryotes joueraient aussi un rôle clé dans ces deux mécanismes (**Figure 24**) [111].

Les premières protéines à intervenir dans le SHM sont des cytosines désaminases appelées AID et APOBEC mentionnées dans la section précédente **I.2.2.1** p. **69**. Dans les lymphocytes B, les protéines AID interviennent dans la désamination catalytique des cytosines des gènes des immunoglobulines (Ig). Leur action conduit à la formation d'uraciles mésappariées à des guanines (G.U). UNG2 reconnait

l'uracile dans le mésappariment et procède à son excision, ce qui a pour conséquence la formation de sites AP extrêmement mutagènes (s'ils ne sont pas éliminés par le BER). UNG2 initie également le SHM. L'initiation du CSR se fait en revanche après la formation SSB induites par APE1 au niveau des sites AP générés par UNG2 (**Figure 21**). S'en suit des réarrangements chromosomiques dépendants du système de recombinaison NHEJ. Ces remaniements chromosomiques au niveau des gènes codant pour les parties constantes des Ig contribuent, avec le SHM, à la variabilité du répertoire des anticorps dont disposent les mammifères. Dans ce contexte, le BER *via* AID/UNG2 favorise les mutations génétiques pour créer de la variabilité antigénique au sein des structures des anticorps. C'est exactement le contraire de l'activité antimutagène du BER dans le processus « canonique » de réparation de l'ADN (**Figure 21**) [112].

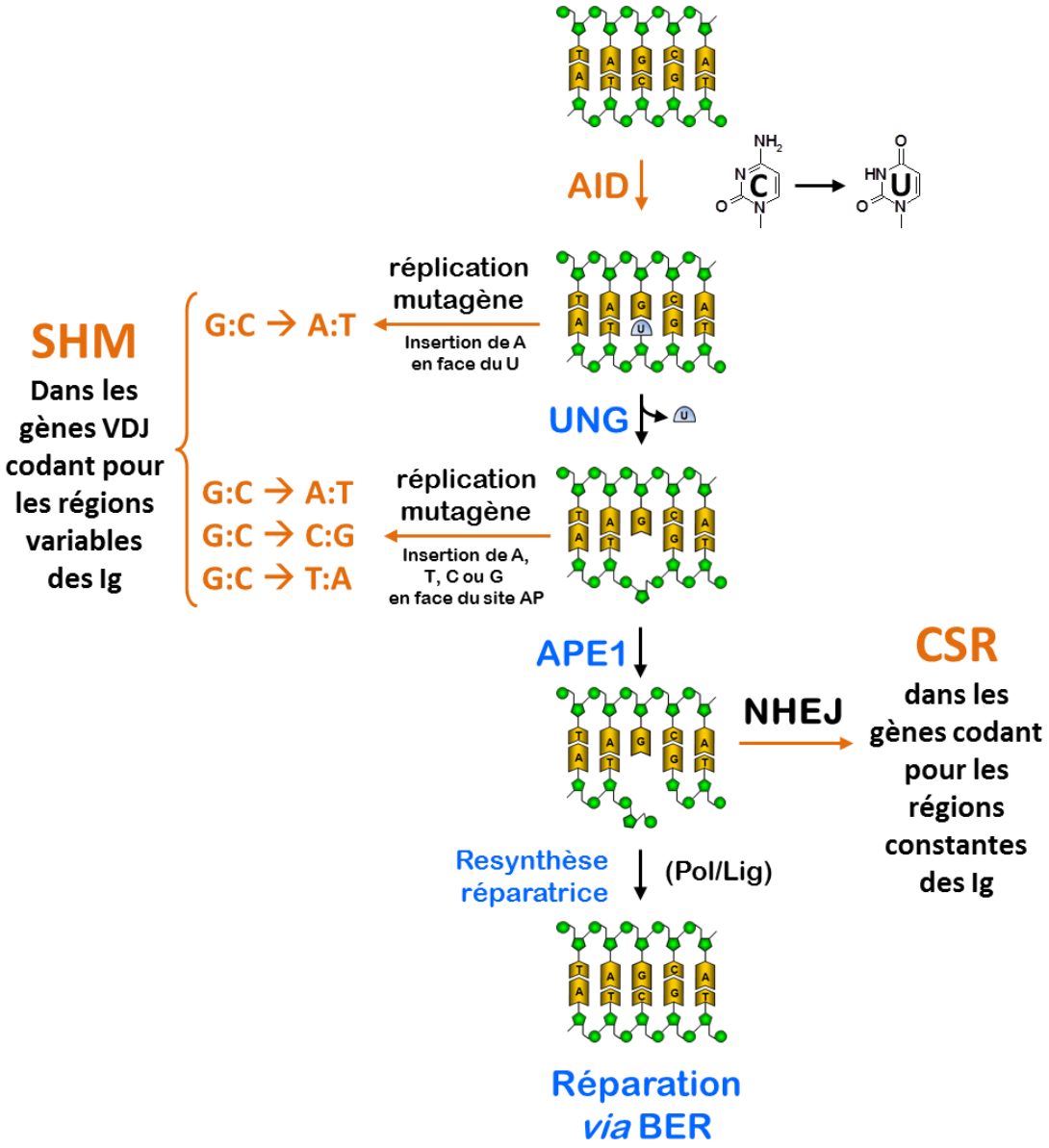


Figure 24 : Implication du BER dans les mécanismes de diversification des immunoglobulines
L'hypermutagenèse sommatique (SHM) et dans les réarrangements chromosomiques (CSR) des gènes codant pour les immunoglobulines dans les lignées lymphocytaires B.

En parallèle, UNG2 intervient lors de l'immunité innée après une infection par certains virus herpétiques et les poxvirus. Ces virus expriment leurs propres UNG virales, dont la structure et les propriétés biochimiques sont très proches des UNG de leur hôte. Les UNG virales interviennent dans le cycle de réplication de ces virus en éliminant les uraciles incorporées dans les brins d'ADN néoformés. Les deux types d'UNG étant présentes dans la cellule, elles interviennent toutes les deux sur l'ADN viral en formation, cependant les UNG de l'hôte ont tendance à générer dans l'ADN viral des sites AP qui sont ensuite clivés par APE1. Cela conduit finalement à la dégradation de l'ADN viral par l'activité 3'→5' exonucléase de l'APE1. Dans ce cadre, les protéines du BER UNG2 et APE1 participent également à la défense antivirale et d'une façon générale, contre tout ADN étranger [113].

I.2.3 Disfonctionnement du BER dans les pathologies humaines

Le BER contribue à la longévité et à la viabilité cellulaire par le maintien de l'intégrité du génome et participe à de nombreuses autres fonctions cellulaires. De ce fait, des défauts dans la voie du BER sont associés à un certain nombre de pathologies.

I.2.3.1 BER et vieillissement prématuré

En dehors du cerveau, lorsque des cellules sont BER déficientes, l'augmentation et ainsi l'accumulation des dommages oxydatifs dans leur ADN conduit à un vieillissement rapide de ces dernières. Ainsi, plusieurs études menées sur la souris ont démontré que des défauts dans le BER ou une diminution de la réparation des lésions issues d'oxydations ou encore une sensibilité accrue au ROS menait à des phénotypes dégénératifs et au vieillissement accéléré et prématuré des animaux [114].

En parallèle, les télomères (extrémités des chromosomes) sont extrêmement sensibles aux lésions oxydatives, notamment au niveau des répétitions TTAGGG. Les lésions uraciles et 8-oxoG sont donc très abondantes au niveau des télomères. Le mécanisme BER propre aux télomères n'est toujours pas bien compris, il a cependant été mis en évidence que les protéines garde du corps (« shelterin ») TRF1, TRF2 (« Telomeric Repeat Factor 1 et 2 ») et POT1 (« Protection Of Telomeres protein 1 ») interagissent physiquement avec les acteurs du BER tels que Polβ, FEN1, APE1 et NEIL3 [115, 116].

I.2.3.2 BER et cancers

Les disfonctionnements du BER et l'instabilité génomique dont il en résulte sont liés à de plus grandes prédispositions à certains cancers. Ainsi, une forme de Polβ délétée entre les positions 208 et 236 ou encore les formes mutées K289M et I260M ont été retrouvées dans certaines tumeurs humaines différentes, mais leur présence n'a pas pu être reliée directement à ces maladies. Cependant, il a été montré que le BER était beaucoup moins efficace en présence de ces formes modifiées de Polβ [117]. Chez la souris, certains mutants de Polβ ont été corrélés avec le lupus [118]. La déficience des protéines APE1, XRCC1, et LIG1 ont été reliées à la prédisposition à des cancers [117].

La déficience des ADN glycosylases est impliquée dans de nombreuses pathologies humaines, dont des cancers et des maladies neurodégénératives telle que la Chorée de Huntington. Des études ont montré que les ADN glycosylases MUTYH, MBD4, hNEIL1 et hOGG1 peu ou non fonctionnelles augmentaient individuellement le risque de différents types de cancers (**Tableau 5**). Certaines ADN glycosylases telle qu'UNG1 ou SMUG1 ne semblent pas induire de pathologie lorsqu'elles sont les seules inactives. Cependant, lorsque les deux gènes relatifs à ces enzymes sont supprimés chez la souris en même temps, on observe l'apparition de maladies telles que des cancers (**Tableau 5**). Il en est de même pour NEIL1 et NTHL1, MUTYL et OGG1 (**Tableau 5**). Cela signifie que, dans certains cas, lorsqu'une ADN glycosylase est non fonctionnelle, un système de secours existe pour pallier à l'enzyme défectueuse. Les ADN glycosylases prennent en charge des panels substrats spécifiques (**Tableau 7**, p. **85**), et partagent des substrats communs. Ainsi, le recouvrement des activités des ADN glycosylases permet de limiter les conséquences de l'inactivité de l'un des initiateurs du BER.

En parallèle, des altérations au niveau des gènes codants pour les protéines impliquées dans le BER ont récemment été reliées à certains cancers, notamment sur les gènes produisant les ADN-glycosylases MBD4 et NEIL1 (**Tableau 5**). MBD4 initie donc le BER en se liant à des régions de l'ADN hyperméthylées, sur des îlots CpG méthylés (voir la section précédente **I.2.2.1** p. **69**). Cette ADN glycosylase excise les bases mésappariées aux G, provenant généralement de la désamination de la cytosine en uracile et de la 5-méthylcytosine en thymine (potentiellement mutagènes car cela mène respectivement aux mésappariements G.U et G.T) [36]. Des études récentes ont montré un lien entre le « silencing » épigénétique du gène de *MBD4* et un certain type de cancer colorectal et ovarien [119]. Une déficience de l'ADN glycosylase NEIL1 a été corrélée à une augmentation en 8-oxoG dans l'ADN, et aux mésappariements de bases associés tels que 8-oxoG.A, menant à des transversions G:C → T:A [120]. De plus, chez des patients atteints de cancer de l'estomac, la production d'ARNm codant pour NEIL1 était diminuée de 46% en moyenne [121]. Dans une autre étude, 42% des patients ayant un

cancer des poumons présentent la région promotrice du gène *NEIL1* hyperméthylée [122]. Ces informations permettent de souligner l'implication d'ADN glycosylases telles que MBD4 et hNEIL1 dans la prévention de l'apparition de cancers chez l'homme.

I.2.3.3 BER et maladies neurodégénératives

Les neurones sont des cellules qui n'ont pas la capacité de se diviser, et l'accumulation de mutations dues à une déficience du BER favorise la décroissance du nombre de cellules neuronales. Dans des études récentes, l'hypersensibilité des neurones au stress oxydant suivie de l'apoptose a été reliée aux protéines UNG, AAG et APE1 défectueuses [117]. Une baisse de production de l'ADN glycosylase hOGG1 dans les neurones a également été corrélée à la maladie d'Alzheimer chez la souris [101, 123].

Pour finir, certaines ADN glycosylases (hNEIL1, hOGG1) ainsi que la PARP1 et FEN1 ont été reliées au Syndrome de Cockayne (« Cockayne Syndrome », CS), une maladie génétique où le sujet présente une croissance insuffisante, une dysmorphie spatiale, une photosensibilité cutanée, des troubles neurologiques progressifs atteignant la vision et l'audition ainsi qu'un retard mental. L'ADN des cellules des patients atteints du CS semblent accumuler les lésions oxydatives et UV-induites [101].

En parallèle, plusieurs protéines du BER interviennent dans l'augmentation de la pénétrance de la Chorée de Huntington (**Tableau 5**). Cette maladie est causée par l'augmentation du nombre de répétitions du triplet de nucléotides (TNR) « CAG » dans la séquence du gène *HTT* (ou *HD*) codant pour la protéine Huntingtine. *HTT* joue un rôle essentiel dans la régulation du trafic vésiculaire et la sécrétion de facteurs neurotrophiques. En dessous de 28 répétitions, le gène code pour une protéine normale, et à partir de 40 répétitions, la pénétrance de la maladie est complète. Le LP-BER intervient dans le mécanisme « TriNucleotide Repeat expansion » (TNR) par lequel le nombre de répétitions de ce triplet CAG augmente. Ce mécanisme fait intervenir les ADN glycosylases hOGG1, hNEIL1 et la FEN1 chez l'Homme. Les guanines et les cytosines des triplets CAG du gène *HTT* peuvent subir des oxydations ce qui induit ainsi la formation de 8-oxoG et de 5-OhC. Ces lésions sont réparées par hOGG1 et hNEIL1, un nouveau fragment est synthétisé ensuite les Polβ. L'incapacité de FEN1 à couper efficacement le brin déplacé lors de la phase de resynthèse du LP-BER conduit à l'expansion des triplets [124].

Tableau 5 : Liste non exhaustive des implications connues des ADN glycosylases dans les maladies humaines

Gènes	Localisation cellulaire	Phénotypes des souris Knockout	Association connue à une pathologie
UNG1	Noyau	Souris KO UNG1 et SMUG1, augmentation la prédisposition aux cancers associés au gène *MSH2* [125]	-
UNG2	Mitochondrie et noyau	Déficience partielle du système de commutation de classe et déficience totale du système d'hypermutagénèse somatique [126], lymphome diffus à grandes cellules B [127]	Déficience du système de commutation de classe, syndrome d'hyper-IgM [128], hyperplasie lymphoïde
SMUG1	Noyau	Viables et fertiles [125]	-
TDG	Noyau	Mortalité embryonnaire [129]	-
MBD4	Noyau	Viables et fertiles, fréquence des transitions C → T multipliée par un facteur 3 dans les îlots CpG [130]	Mutée chez les patients atteints de cancer colorectal et instabilité des microsatellites [131] ainsi que de cancer ovarien [119]
AAG (MPG)	Noyau	Viables et fertiles, accumulation des mutations A → T et G → T [132]	-
OGG1	Mitochondrie et noyau	Viables et fertiles, accumulation de 8-oxoG dans l'ADN [133, 134]	Associée avec l'expansion des répétitions CAG dans la maladie de Huntington [135], mutée dans le cas de cancer du poumon, de la prostate, la leucémie [131, 136] et la maladie d'Alzheimer [123]
MUTYH	Mitochondrie et noyau	Viables et fertiles [134]. Souris KO MUTYH et OGG1, cancer des poumons associé au gène *KRAS* [137]	Polymorphismes associés aux polypes colorectal [138, 139]
NTHL1	Mitochondrie et noyau	Viables et fertiles [140]. Souris KO NTHL1 et NEIL1, cancer des poumons associé au gène *KRAS* [141]	-
NEIL1	Noyau	Obésité sévère au bout de 7 mois [142]	Région promotrice hyperméthylée chez les

			patients atteints d'un cancer du poumon [122], production d'ARNm réduite dans le cas du cancer de l'estomac [121], associée avec l'expansion des répétitions CAG dans la maladie de Huntington [143]
NEIL2	Noyau	Inflammation [144]	-
NEIL3	Noyau	Maladies auto-immunes [116], neurogénèse déficiente chez l'embryon [145, 146]	-

Abréviations : • **Gènes** : UNG1/2, Uracile DNA glycosylase 1/2 ; SMUG1, Single-strand selective Monofunctional uracil DNA glycosylase 1 ; TDG, Thymine DNA Glycosylase ; MBD4, Methyl-CpG binding domain 4, DNA Glycosylase ; AAG (MPG), 3-Methyladenine-DNA glycosylase ; OGG1, 8-oxoguanine DNA glycosylase 1 ; MUTYH, MUTY DNA glycosylase ; NTHL1, Endonuclease III-like protein ; NEIL1/2/3, Endonuclease VIII like protein 1/2/3.
• **Phénotypes** : KO, knockout ; MSH2, MutS Homolog 2 ; IgM, Immunoglobulin M.

I.2.4 Les ADN glycosylases

Les ADN glycosylases ont fait l'objet de nombreuses études et une grande quantité de données structurales, biochimiques et fonctionnelles est disponible dans la littérature. Ces enzymes initient le BER en reconnaissant et en éliminant spécifiquement des bases altérées de l'ADN. Les bases oxydées, alkylées, dégradées ainsi que les bases mésappariées ou ne devant pas exister dans l'ADN (comme l'uracile et l'hypoxanthine par exemple) sont les substrats des ADN glycosylases. Le mode opératoire des ADN glycosylases consiste en (i) la diffusion sur l'ADN, (ii) reconnaissance des lésions puis (iii) leur extraction de la double hélice dans un mouvement appelé le « base extrusion ». Le nucléotide endommagé est stabilisé dans le site actif tandis que des aa remplissent le « gap » dans la double hélice en s'intercalant dans l'ADN. Comme décrit dans la section précédente **I.2.2** p. **68** sur le BER, certaines ADN glycosylases ont aussi une activité AP lyase qui leur permet de couper le squelette phosphodiester après l'élimination de la base altérée, ces enzymes sont dites bifonctionnelles.

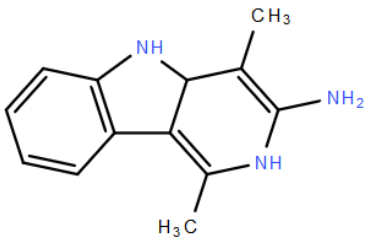
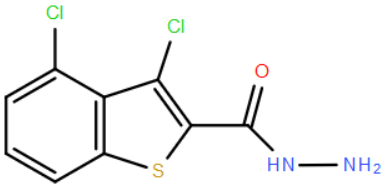
I.2.4.1 Les ADN glycosylases, de nouvelles cibles thérapeutiques

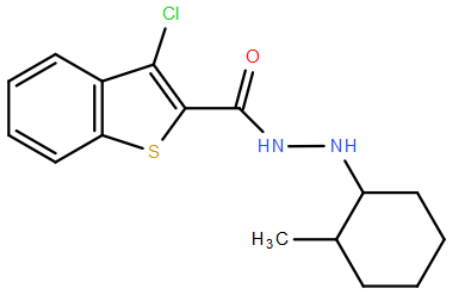
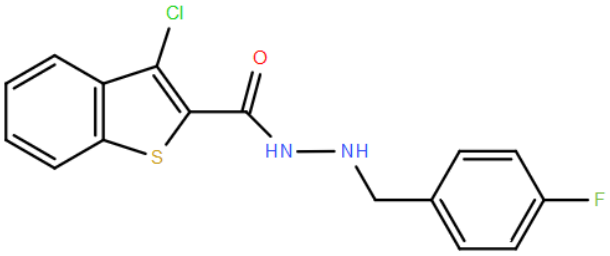
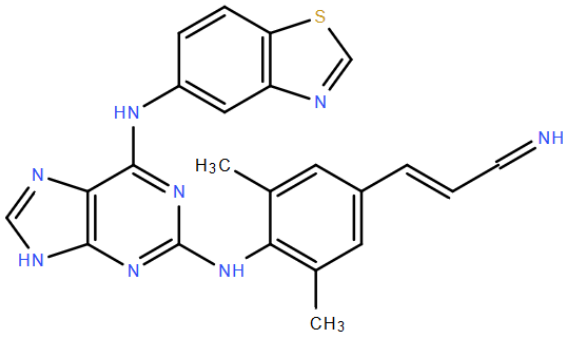
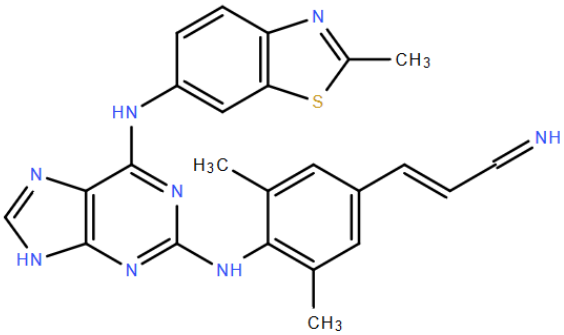
Les traitements tels que la chimio- et la radiothérapie engendrent des lésions dans l'ADN des cellules ce qui entraine leur entrée en apoptose. Cependant, le BER est un des systèmes qui permet aux cellules cancéreuses d'échapper au traitement, entraînant une résistance et une persistance de la maladie [147]. Par conséquent, les ADN glycosylases sont des cibles thérapeutiques d'intérêts dans les stratégies anti-cancers [148].

De plus, l'ablation isolée des ADN glycosylases hOGG1 ou hNEIL1 par $_{si}$RNA (petit ARN interférent) dans une population de cellules U2OS (ostéosarcome) traitées par méthotrexate ou raltitrexed (inhibiteur de la voie de thymidylate synthase, et par conséquent bloquant la production de TTP) induit l'incorporation d'UMP dans l'ADN, l'arrêt de la réplication puis l'apoptose des cellules [149]. Dans ce contexte, on parle de léthalité synthétique, cela signifie que l'inactivation d'une des deux voies n'a pas d'effet délétère sur la cellule, contrairement à la suppression simultanée des deux voies [150]. Le concept de léthalité synthétique a été identifié dans le cas d'inactivation de la protéine PARP1 dans les lignées cellulaires des cancers du sein et ovarien liés aux gènes *BRCA*. L'exploitation de ce mécanisme a initié la conception d'inhibiteurs de la PARP1 dont certains ont été approuvés par la « Food and Drug Administration » (FDA) en 2016 [151].

Dans ces deux contextes cités ci-dessus, les ADN glycosylases ont fait l'objet de recherches d'inhibiteurs dans le but de compléter les traitements anti-cancers déjà existants (**Tableau 6**) [152]. Cependant, les inhibiteurs de hNEIL1 présentés dans ce tableau sont, d'après les auteurs, toxiques car elles n'ont que très peu de spécificité pour leur cible, et les résultats décrits dans la publication de Jacobs *et al.* [153] n'ont pas pu être reproduits au CBM.

Tableau 6 : Liste non exhaustive des inhibiteurs des ADN glycosylases

Cible	Inhibiteur	IC_{50} (µM)
MPG		1 [154]
hOGG1		0,22 ± 0,08 [155]

		0,27 ± 0,1 [155]
		0,33 ± 0,02 [155]
hNEIL1		9 [153]
		28 [153]

Abréviations : hOGG1, 8-oxoguanine ADN glycosylase 1 humaine ; MPG, Méthylpurine ADN glycosylase humaine ; hNEIL1, Nei like protein 1 humaine ; IC_{50}, concentration inhibitrice médiane.

I.2.4.2 Les superfamilles structurales des ADN glycosylases

On distingue donc les ADN glycosylases monofonctionnelles et les ADN glycosylases bifonctionnelles (**Tableau 7** p. **85**). On peut classer les ADN glycosylases en fonction de leurs repliements tridimensionnels en six superfamilles structurales (**Figure 25**). Tous les organismes vivants possèdent un panel d'ADN glycosylases, ce qui leur permet de faire face à un large spectre d'altérations des bases de l'ADN. Même si les ADN glycosylases présentent des spécificités de substrat différentes, (**Tableau 7** p. **85**) certaines d'entre elles pouvant appartenir à des superfamilles structurales différentes présentent des spécificités croisées et chevauchantes (substrats communs) au sein d'un même organisme (ex. : la 8-oxoG est excisée par Fpg et la Ogg procaryote et la thymine glycol par Nei/ENdoIII et Nth/EndoVIII).

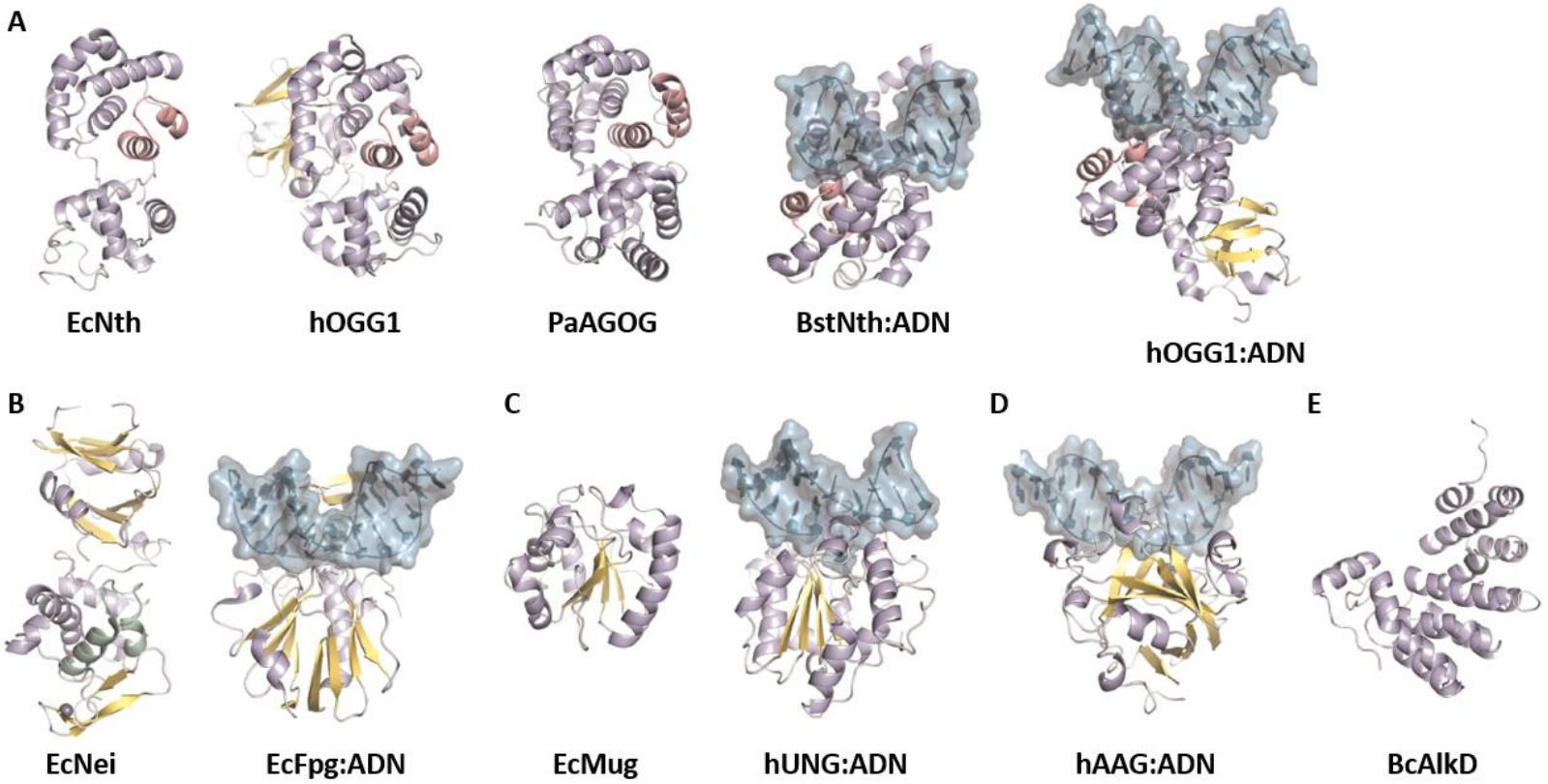


Figure 25 : Les cinq différents repliements structuraux des ADN glycosylases

A) Enzymes appartenant à la superfamille des HhH (« Helix hairpin helix ») motif en rose (PDBid : 2ABK [156], 1KO9 [157], 1XQO [158], 1P59 [159], 1EBM [160]) et **B**) protéines de la superfamilles des Fpg/Nei (H2TH, « Helix 2 Turns Helix ») motif en vert (PDBid : 1Q39 [161], 1K82 [162]). **C**) Les ADN glycosylases de la superfamille des UDG comportent un motif en sandwich αβα conservé (PDBid : 1MUG [163], 1SSP [164]). **D**) Les membres de la superfamilles des Aag sont composés d'un repliement compact d'hélice α et de feuillet β, et tordent l'ADN avec un angle de 22° (PDBid : 1BNK [165]). **E**) Modèle reconstruit par homologie de la protéine AlkD de *Bacillus cereus* libre, sans ADN, de la superfamille « HEAT-like repeat DNA glycosylases » [166]. Sur ces figures, la protéine est représentée en cartoon et colorée en fonction de la structure secondaire (les boucles en gris, les hélices α en bleu, les brins β en jaune), l'ADN est présenté en surface transparente cyan [50].

Abréviations : ● Organismes : *Ec*, *Escherichia coli* ; h, human ; *Pa*, *Pyrobaculum aerophilum* ; *Bst*, *Bacillus stearothermophilus* ; *Bc*, *Bacullis cereus*

● Protéines : Nth, Endonucléase III ; OGG1, 8-oxoguanine ADN glycosylase 1; AGOG : Archaeal 8-OxoGuanine DNA Glycosylase

I.2.4.2.1 Les ADN glycosylases de dimère de pyrimidines (PDG)

Les premiers représentants des ADN glycosylases forment une petite superfamille, car peu d'entre eux ont été identifiés, contrairement aux autres ADN glycosylases. Cette enzyme, anciennement connue sous le nom de T4-endonucléase V (EndoV) chez les procaryotes, est désormais nommée glycosylase de dimère de pyrimidine (T4 pyrimidine dimer glycosylase, T4-PDG), nommée ainsi pour son activité d'hydrolyse des liaisons *N*-glycosidiques en 5' des CPDs. Quelques représentants de cette famille ont d'abord été isolés chez le bactériophage T4, puis la bactérie *Neissera mucosa* (*Nmu*-Pdg I et II) ainsi que des séquences homologues du gène codant pour T4-PDG ont pu être identifiées chez les bactéries *Bordetella*, *Brucella*, *Haemophilus*, *Pasteurella* et *Prochlorococus*, mais l'existence de ces protéines reste putative [167]. La détection des CPDs par les PDGs semble passer par la reconnaissance du squelette phosphodiester de l'ADN déformé en présence de telles lésions. Grâce à une structure cristallographique, il a été possible de comprendre comment les PDG réalisaient leur catalyse [168]. Les PDGs sont des ADN glycosylases bifonctionnelles qui catalysent le clivage du lien *N*-glycosidique d'une des thymines du dimère de façon similaire à celle des ADN glycosylases de la superfamille structurale Fpg/Nei. Ainsi, l'enzyme forme un complexe covalent avec l'ADN sous la forme d'une base de Schiff (décrite plus loin dans ce manuscrit) impliquant l'amine de la thréonine N-terminale de l'enzyme. La formation de ce complexe est concomitante au clivage de la liaison *N*-glycosidique de la pyrimidine en 5' du dimère. Le site AP résultant est clivé en 3' par un mécanisme de β-élimination. L'activité glycosylase/lyase de la PDG résulte de la cassure SB délimitée par un site AP insaturé en 3' et un adduit CPD dont la thymine en 5' n'est plus associée à son sucre. Les structures ont révélé aussi que, contrairement aux autres ADN glycosylases, les PDG ne déplacent pas leurs substrats en position extra-hélicale mais la base faisant face au dommage. Elles accèdent ensuite à la lésion à l'intérieur de la double hélice et procèdent à la coupure. Cette originalité ne change pas le fait que quelques aa sont mobilisés pour l'intercalation dans l'ADN DB, pour stabiliser le système [168, 169].

I.2.4.2.2 Les ADN glycosylases « Heat-like » (HTL)

Les HTL sont des protéines dites « HEAT-like repeat » formant une superfamille récemment découverte initiant la réparation des bases alkylées de l'ADN, d'où le nom des protéines AlkC et AlkD. « HEAT repeat » est à l'origine un motif de structure secondaire identifié chez trois protéines sans réel point commun si ce n'est le repliement : la « **H**untingtin, **E**longation Factor 3 » (**E**F3), la « Protein Phosphatase 2**A »** (PP2A) et la kinase **T**OR1 issue de la levure [170]. La première structure était un modèle construit par homologie grâce à des structures d'autres enzymes dont la séquence primaire

était similaire. À partir de cette structure, il a tout d'abord été possible de distinguer cette protéine des autres ADN glycosylases car elle comportait un motif de structure secondaire formé de 6 répétitions de couples d'hélices α antiparallèles alors inédit, d'où le nom Heat-Like Repeat [166]. Puis les premières structures cristallographiques ont enfin été résolues, notamment la structure de la protéine AlkD [171]. Cette superfamille est représentée dans beaucoup d'organismes procaryotes mais n'a pas encore été observée chez des organismes eucaryotes unicellulaires.

I.2.4.2.3 Les Alkyladénine ADN glycosylases (AAG)

Cette superfamille est composée d'un seul représentant monofonctionnel, l'Alkyladénine ADN glycosylase (AAG ou N-méthyle purine ADN glycosylase MPG). La superfamille AAG comporte un motif structural composé de feuillets β antiparallèles entourés d'hélices α. Elle excise exclusivement les purines méthylées et désaminées de l'ADN, et est la seule enzyme à exciser les bases méthylées chez l'Homme [172].

I.2.4.2.4 Les Uracile ADN glycosylases (UDG ou UNG)

Les Uracile ADN glycosylases (UDG ou UNG) présentent toutes un cœur constitué par quatre feuillets β parallèles espacés par des hélices α. Les UNG sont réparties en 4 sous-familles suivant les substrats qu'elles prennent en charge. Les UDG-1 (Ung procaryote et UDG eucaryote) réparent les uraciles intégrées par erreur dans l'ADN, ou bien les cytosines désaminées qui peuvent mener à des transitions C:G → T:A. Les UDG-4 (UDG archées) réparent également les uraciles intégrés à de l'ADN SB, tandis que les UDG-2 (MUG procaryote et TDG eucaryote) et UDG-3 (SMUG eucaryote) réparent un panel plus important de bases oxydées issues de la 5-mC ou de la désamination de la cytosine et de la 5-mC en uracile et thymine respectivement résultant en les mésappariements U.G et T.G. La réparation de ces mésappariements est donc indépendante du système général de la réparation des mésappariements (MMR). Ces ADN glycosylases sont toutes monofonctionnelles et nécessitent donc l'activité d'une AP-endonucléase après leur action pour poursuivre la réparation [173].

I.2.4.2.5 Les ADN glycosylases de la superfamille structurale de l'endonucléase III (HhH)

La superfamille structurale HhH (Helix hairpin Helix) est composée de protéines comportant des domaines aux structures secondaires très conservées. La superfamille des HhH est la plus diverse et la plus représentée dans les règnes du vivant. Quelques protéines de ce groupe possèdent en plus des cofacteurs tels que des clusters Fe^4S^4 ne jouant pas de rôle critique dans la catalyse mais dans la

structure de ces dernières. Certaines ADN glycosylases HhH sont monofonctionnelles (AlkA, MBD4 et MutY) et bifonctionnelles (AGOG, EndoIII et OGG1) et reconnaissent un vaste ensemble de lésions provoquées par l'oxydation et l'alkylations des bases azotées [174].

I.2.4.2.6 Les ADN glycosylases de la superfamille structurale Fpg/Nei (ou H2TH)

Mes travaux de thèse ont porté sur l'étude des ADN glycosylases de la superfamille structure Fpg/Nei, c'est pourquoi j'ai dédié le chapitre suivant à leur caractérisation fonctionnelle et structurale.

Tableau 7 : Liste non exhaustive des substrats des ADN glycosylases

Dans la colonne « substrats », les mésappariements X.Y pris en charge sont précisés s'ils sont connus, la base excisée est située à gauche du « . », l'élément de droite étant une condition pour que le substrat soit excisé, « * » représentant toutes les bases azotées classiques A, T, C et G [50, 55].

Groupe ou superfamille	Substrat(s)	Enzyme			
		Archées	Procaryotes	Eucaryotes	Phages et virus
PDG	*Cis-syn* T<>T		DenV/*Nmu*-Pdg		T4-PDG
HTL	3-mA, 3-mG, 7-mG, 7-POB-G, O^2-POB-C		AlkC, AlkD		
AAG (MPG)	3-mA, 7-mG, εA, hX, A, G			AAG (MPG)	
UDG-1	U.G		Ung	UNG (UDG)	
UDG-2	T.G, U.G, U.A, 5-fC, 5-caC, 5-FU.G, 5-FU.G, 5-FU.A, 5-BrU.G, 5-Br.A, 5-hmU.G, 5-OHU.G, Tg.G, εV, εC:A, hX.G, 8-hmεC, εC, X		MUG	TDG	
UDG-3	U (ADN SB), U.G, U.G, 6-hmU, 5-OHU, 5-fU			SMUG1	
UDG-4	U (ADN SB), U.G	UDG			
HhH	8-oxoG : C, FaPyG, FaPyA	OGG	Ogg	OGG1	
	8-oxoG.*	OGG2			
	8-oxoG.* (ADN DB, ADN SB)	AGOG			
	3-mA, 3-mG, 7-mG, 7-CEG, 7-HEG, εA, hX, G			MAG, Mag1	

	3-mA, 7-mG, εA, 1-mA, 3-mC	*Af*AlkA			
	3-mA, 7-mG	MpgII			
	3-mA, 3-mG, 7-mG, 7-CEG, 7-HEG, 7-EG, O²-mdT, O²-mC, εA, hX, A, G, T, C, X		AlkA		
	3-mA, 7-mG.*		MagIII		
	3-mA, 3-mG		TAG		
	T.G, U.G, 5-FU.G, εC, 5-mC			MBD4	
HhH/FeS²	Tg, Ug, DHU, 5-OHU, 5-OHC, urée	EndoIII	Nth/EndoIII	NTH1	
	A.8-oxoG, A/G	MutY	MutY	MUYTH	
	T.G	MIG			
	5-mC, T.G			DME, ROS1, DML2, DML3	
Fpg/Nei (H2TH)	8-oxoG, FaPyG, 7-mFaPyG, Sp, Gh, Tg, Ug, DHT, DHU, 5-OHU, 5-OHC, FU, […]		MutM/Fpg		

	Tg, DHT, DHU, 5-OHU, 5-OHC, 5-fU, 5-hmU, FaPyG, FaPyA, 8-oxoA, Gh, Sp, Ia ; Nei : 8-oxoG, 7-mFaPyG, UG, 5,6-hC, 5-OHT		Nei/EndoVIII	NEIL1	NEIL1
	Gh.Ia, 5-OHU, FaPyG			NEIL2	
	Sp, Gh, FaPyG, FaPyA			NEIL3	

Abréviations :

● **Groupes ou superfamilles :** PDG, Pyrimidine Dimer DNA Glycosylase ; HTL, HeaT Like repeat ; AAG (MPG), 3-Methyladenine DNA Glycosylase ; UDG, Uracil DNA Glycosylase ; HhH, Helix hairpin Helix ; H2TH, Helix 2 Turns Helix.

● **Substrats :** AP, site abasique ; THF, Tetrahydrofuranose ; HPD, 1-hydroxypentane-3,4-diol ; PDI, 3-hydroxypropyle ; PED, pentane-3,4-diol ; 8-oxoG, 8-oxo-7,8-dihydroguanine ; FaPyG, 2,6-diamino-4-hydroxy-5-formamidopyrimidine ; FaPyA, 4,6-diamino-5-formamidopyrimidine ; 7-mFaPyG, N7-methylFaPyG ; Tg, Thymine glycol ; Ug, Uracile glycol ; DHT, dihydrothymine ; DHU, dihydrouracile ; 5-OHC, 5-hydroxycytosine ; 5-OHT, 5-hydroxythymine ; 5,6-dhC, 5,6-dihydroxycytosine ; 5-OHU, 5-hydroxyuracile ; 5mC, 5-methylcytosine ; 5-hmC, 5-dihydroxymethylcytosine ; 5-fC, 5-formylcytosine ; 5-caC, 5-carboxylcytosine ; 5-hmU, 5-hydroxymethyluracile ; 5-fU, 5-formyluracile ; 5-FU, 5-fluorouracile ; 5-BrU, 5-bromouracile ; Gh, guanidinohydantoine ; Ia, iminoallantione ; Sp, spiroiminodihydantoin ; 3-mA, N3-méthyladénine ; 3mG, N3-méthylguanine ; 7-mG, N7-méthylguanine ; 7-CEG, 7-(2(chloroéthyl)guanine ;7-HEG, 7-(2-hydroxyéthyl)guanine ; 7-POB-G, N7-pyridyloxobutylguanine ; O^2-POB-C, O^2-pyridyloxobultylcytosine ; εA, 1,N^6,éthénoadénine ; εC, 1,N^4-éthénocytosine ; 8-hmεC, 8-(hydroxyméthyl)-3,N^4-ethenocytosine ; hX, Hypoxanthine ; X, xanthine.

● **Enzymes :** OGG/Ogg/Ogg1, 8-Oxoguanine DNA glycosylase 1 ; AGOG, Archaeal 8-Oxoguanine DNA glycosylase ; *Af*AlkA, *Archaeoglobus fulgidus* Alkylpurine DNA glycosylase ; MpgII, Methylpurine DNA glycosylase ; EndoIII, Endonuclease III ; MutY, Adenine DNA glycosylase ; MIG, archaeal MIsmatch specific Glycosylase ; DenV, Endonuclease V ; T4/*Nmu*-Pdg, bactériophage T4/*Neisseria mucosa* pyrimidine glycosylase ; Alk A/C/D, Alkylpurine DNA glycosylases A/C/D ; Ung, Uracil DNA glycosylase ; MUG, Monofunctional Uracil DNA Glycosylase ; MagIII/MAG/Mag1, 3-methyladenine DNA glycosylase ; TAG, 3-methylAdenine Glycosylase I ; Nth/NTH1, Endonucléase III ; MutM/Fpg, Formamidopyrimidine DNA glycosylase ; Nei/EndoVIII, Endonucléase VIII ; TDG, Thymine DNA Glycosylase, SMUG, Single-stranded Monofunctional Uracil Glycosylase ; MBD4, Methyl-CpG Binding Domain protein 4 ; MUTYH, MUTY Homolog ; DME, 5-methylcytosine DNA glycosylase DEMETER ; ROS1, Repressor of Silencing 1 ; DML2/3, DME like 2/3 ; NEILl1/2/3, Nei like protein 1/2/3.

I.3 **Chapitre 3 : L**es ADN glycosylases de la superfamille Fpg/Nei

Je présenterai ici la superfamille structurale des ADN glycosylases Fpg/Nei en focalisant mes propos sur les éléments structure-fonctions à notre disposition au début de la thèse. Je détaillerai ces aspects en particulier sur les protéines *Ll*Fpg et hNEIL1, les objets d'étude de mon projet de thèse.

I.3.1 Fonction et structure des protéines Fpg/Nei

L'objectif de cette section est de donner un aperçu de l'état de l'art sur la compréhension que nous avons sur le fonctionnement de ces protéines mais également d'en appréhender toute la complexité. Inéluctablement, il s'agira cependant d'une introduction à l'étude de ces protéines survolant les connaissances et de fait elle restera non-exhaustive. À titre d'exemple, la modulation de l'activité de ces enzymes et de leurs interactions avec des partenaires cellulaires par des modifications post-traductionnelles ne sera pas du tout abordée comme bien d'autres aspects.

Le premier représentant connu de la superfamille structurale Fpg/Nei est la protéine Fpg ou MutM initialement identifiée chez *E. coli* comme une ADN glycosylase capable d'exciser les purines à cycle imidazole ouvert ou formamidopyrimidines d'où son nom Formamidopyrimidine-ADN glycosylase (Fpg) [175]. Plus tard, il a été découvert que Fpg était capable d'exciser de nombreuses bases oxydées dont en particulier le produit majeur d'oxydation des purines la 8-oxoG (**Figure 28** p. **95**) [176, 177]. Peu de temps après la découverte de Fpg, l'endonucléase VIII ou Nei est identifiée chez *E. coli* comme une ADN glycosylase très similaire structuralement à la protéine Fpg mais présentant des spécificités de substrats recouvrant seulement partiellement celles de Fpg (**Figure 28** p. **95**) [178]. Seules les protéobactéries comme *E. coli* possèdent en même temps Fpg et Nei (**Figure 26 A**) [179]. Suite au séquençage du génome humain, trois homologues structuraux ont été identifiés et appelés Nei-like protéine (NEIL) 1, 2 et 3 (NEIL1, NEIL2 et NEIL3) pour leurs spécificités de substrats similaires à celles des Nei (**Figure 26 A**) [180]. Plus récemment encore, une protéine Fpg/Nei proche de la protéine humaine NEIL1 a été identifiée et caractérisée chez les virus géants des protistes tel que Mimivirus (MvNei1) [181, 182]. L'accumulation des données génomiques révèle que ces protéines sont présentes dans tous les règnes du vivant (à l'exception des archées, pour l'instant) et permet d'établir un arbre phylogénétique de plus en plus précis (**Figure 26 A**). Néanmoins, la distinction entre sous-types 1, 2 ou 3 reste encore à éclaircir. Une analyse fine *in silico* des différentes structures 3D des protéines Fpg/Nei disponibles dans la Protein Data Bank (PDB) a permis d'identifier dans ces protéines 7 motifs structuraux (LSC pour « Latent Structural Cluster ») initialement non-apparents dans les analyses bioinformatiques classiques (voir le recensement de ces LSCs dans l'Annexe A **VI.1** p. **319**)

[178]. Une filiation protéique basée sur ces LSCs propose que les protéines Fpg/Nei dérivent toutes d'un ancêtre commun de type Fpg, à l'origine des Fpg actuelles et des protéines procaryotes Nei et eucaryotes NEILs (**Figure 26 A**). Chaque changement significatif dans ces LSCs est associé à un nouveau sous-type de protéine Fpg/Nei et en particulier à l'apparition des Nei et NEILs. Ce qui différencie clairement une Nei (NEIL) d'une Fpg est son incapacité à exciser la 8-oxoguanine (8-oxoG). Cette spécificité des Fpg pour la 8-oxoG est sans ambiguïté contenue dans un motif structural appelé « Lesion Capping Loop » (LCL) ou « Oxidized guanine Capping Loop » (OCL) (correspondant au motif LCS7) et qui a été perdu au cours de l'évolution par les protéines Nei et NEILs (Figure 26 **B**) [183-185].

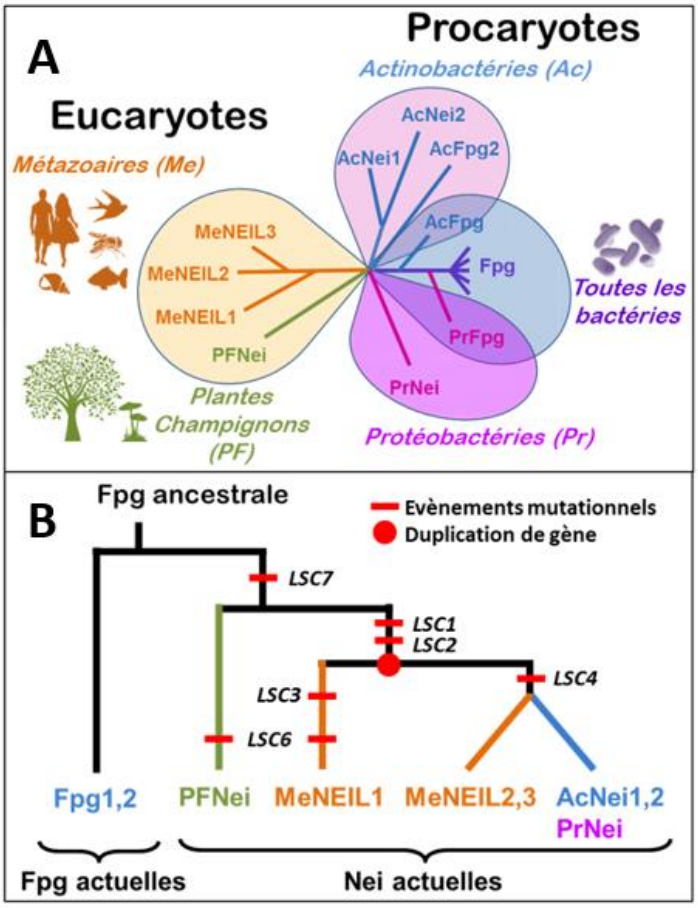


Figure 26 : Un ancêtre commun pour les ADN glycosylases Fpg/Nei
A) Arbre phylogénétique simplifié. Il manque sur cet arbre les archées (pas d'homologue identifié à ce jour) et les virus géants qui sont souvent proposés pour définir un 4ème règne par certains auteurs et pour lesquels il a été décrit et caractérisée une protéine de sous-type 1, nommée MvNei1. **B**) Filiation génétique et structurale entre les gènes Fpg/Nei basée sur l'identification des LSCs. À titre d'exemples : (i) les changements dans le motif LSC7 (« Latent Structural Cluster n° 7 ») aboutissent à la perte de la reconnaissance de la 8-oxoG c'est-à-dire au passage d'une protéine de type Fpg à une protéine de type Nei et (ii) et le passage d'un doigt à zinc (ZnF) à un doigt sans zinc (ZnLF) est associé à des modifications dans le motif LSC6 caractéristique des protéines hNEIL1 et MvNei1 [179].

Abréviations : Fpg, Formamidopyrimydine ADN glycosylase ; Nei, Endonucléase VIII ; NEIL, Nei-like.

Du point de vue mécanistique général, les protéines Fpg/Nei sont des ADN glycosylases/AP lyases (ADN glycosylases bifonctionnelles) dont l'activité AP lyase particulière résulte dans les clivages successifs en 3' et 5' du site AP par un mécanisme de β,δ-élimination (voir détails dans la section **I.3.5** p. **100**). Le repliement global 3D du domaine ADN glycosylase des protéines Fpg/Nei est très similaire (**Figure 27 A**). Ces protéines sont constituées de deux domaines globulaires reliés entre eux par un domaine charnière flexible. Le domaine N-terminal commence *quasi* toujours par une proline (P1) et

est riche en brins β tandis que le domaine C-terminal est riche en hélices α et contient deux motifs structuraux caractéristiques des protéines Fpg/Nei impliqués dans l'interaction avec l'ADN : (i) un motif hélice-2 tours-hélice (ou H2TH) (**Figure 27 A-B** et **Figure 31** motif en jaune) et (ii) un motif en épingle à cheveux formé de 2 brins β antiparallèles associé ou non à un atome de zinc (**Figure 27 A-C**, **Figure 31** motif en vert et **Figure 27 C**) [161, 186]. L'orientation relative des deux domaines est très différente entre les protéines bactériennes Fpg et Nei et similaire entre les NEILs eucaryotes et Fpg [161, 186, 187]. Lors de l'interaction avec l'ADN lésé, l'orientation relative entre ces domaines ne change pas (ou très peu) pour les Fpg et NEILs tandis qu'elle devient similaire à celles que l'on retrouve chez Fpg et NEIL1 pour Nei [188-190]. Outre le domaine ADN glycosylase et contrairement aux enzymes bactériens, les protéines humaines NEIL1 et NEIL3 présentent en C-terminal une extension prédite partiellement non-structurée pouvant représenter un tiers de la protéine et dont le(s) rôle(s) fonctionnel(s) reste(nt) encore à éclaircir. À ce jour, il semble que les régions additionnelles de NEIL1 et NEIL3 soient impliquées dans des interactions avec des partenaires cellulaires modulant leur activité dans le BER et d'autres fonctions telles que : (i) leur accessibilité à l'ADN endommagé et aux télomères, (ii) leurs localisations subcellulaires (noyau, mitochondries) et, plus généralement, (iii) leur régulation, leur « turnover » au cours du cycle cellulaire et leur abondance relative selon le type cellulaire [191, 192].

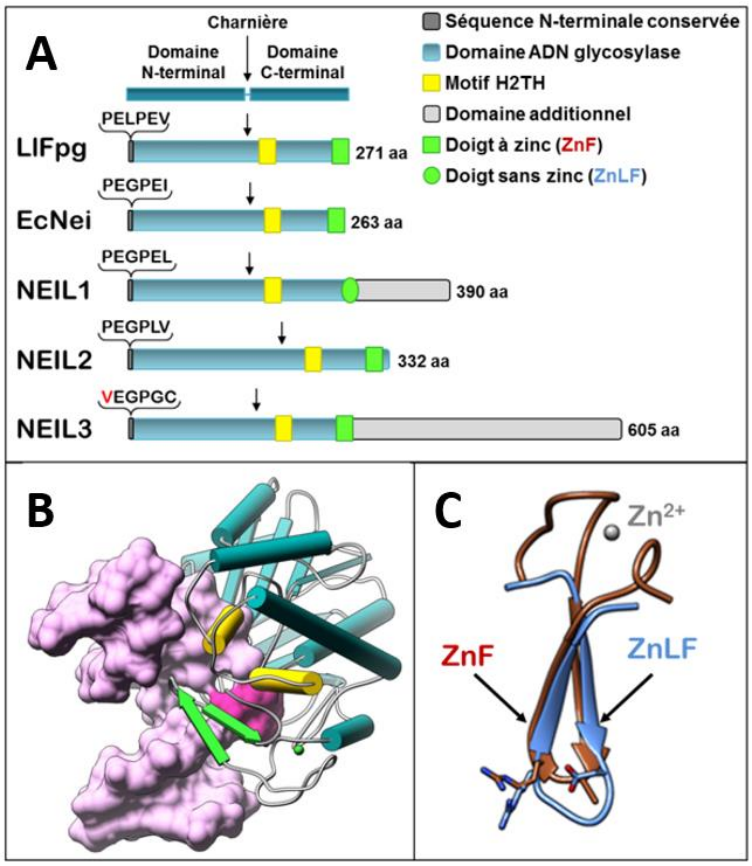


Figure 27 : Éléments structuraux généraux des ADN glycosylases Fpg/Nei

A) Alignements des structures primaires. **B**) Structure cristallographique du complexe *Ll*Fpg avec un duplexe d'ADN contenant un résidu FapyG (PDBid : 1XC8 [184]). En jaune, le motif H2TH et en vert le motif ZnF. **C**) Superposition du ZnF de *Ll*Fpg (en marron) avec le ZnLF de hNEIL1 (en bleu). La sphère en gris correspond à l'ion Zn^{2+} coordonné par 4 Cys chez *Ll*Fpg [54]. Abréviations : *Ll*, *Lactococcus lactis* ; *Ec*, *Escherichia coli* ; ZnF, zinc finger ; ZnLF, zincless finger; H2TH, hélice-2 tours-hélice; Zn^{2+}, ion zinc.

Le défi pour les protéines Fpg/Nei et pour toute ADN glycosylase en général est de (i) rechercher, (ii) identifier puis (iii) métaboliser une base endommagée parmi un million de bases normales. Ce processus est d'autant plus complexe que dans la cellule procaryote ou eucaryote, ces enzymes ne sont pas confrontées à la recherche d'une base endommagée dans un ADN « nu » mais dans une structure chromatinienne (protéines/ADN) en définitif peu accessible car l'ADN se retrouve largement masqué par des protéines de structure de l'ADN (dites histones ou de types histone ou non-histones) et toutes celles participant au métabolisme de l'ADN (réplication, transcription, recombinaison…). Le problème d'accessibilité des protéines à l'ADN peut être illustré par l'observation suivante : la vitesse de réparation est d'autant plus grande que la structure locale de la chromatine est plus ouverte. Ainsi, l'ADN correspondant à des gènes activement transcrits (chromatine peu condensée) est plus vite réparé que celui de gènes non-transcrits (structure de la chromatine beaucoup plus condensée) [193]. Par conséquent, on peut noter des couplages entre réparation et réplication/transcription en particulier chez les procaryotes pour lesquels l'ADN chromosomique baigne dans le cytoplasme. Pour des raisons de format du manuscrit de thèse et de complexité de cette question très récemment explorée chez les eucaryotes, la réparation des bases endommagées dans la chromatine ne sera pas abordée ici. Il faut donc garder à l'esprit que le recrutement des ADN glycosylases Fpg/Nei (ou de toute autre enzyme de réparation) à la chromatine pourrait être associé à des processus passifs et/ou actifs qui ne sont pas forcément dépendants de la présence d'une base endommagée au site de recrutement car ces enzymes, assurant une fonction de ménage, sont censées réparer l'ADN en permanence [191].

I.3.2 Substrats des protéines Fpg/Nei

La superfamille structurale Fpg/Nei est classiquement subdivisée en deux sous-familles : (i) les ADN glycosylases de type Fpg essentiellement bactériennes et plutôt spécifiques pour les purines oxydées et (ii) les ADN glycosylases de type Nei que l'on retrouve chez les procaryotes, les eucaryotes et certains virus comme MvNei1 (homologue de NEIL1 chez le virus géant Mimivirus) et plutôt spécifiques des pyrimidines oxydées (**Figure 28**). Même si elles ne sont pas reliées structuralement, les protéines Nei partagent les mêmes spécificités de substrats dans l'ADN double-brin que les protéines Nth (Endonucléase III), des ADN glycosylases procaryotes et eucaryotes de la superfamille structurale HhH spécialisée elles aussi dans la réparation des pyrimidines oxydées (voir plus haut). Les protéines Fpg/Nei peuvent aussi se différencier par leur capacité à fonctionner sur l'ADN simple ou double brin ou sur les deux. Les Fpg comme les Nei n'excisent efficacement les bases oxydées que dans l'ADN DB tandis que les NEILs eucaryotes fonctionnent sur l'ADN SB et DB (NEIL1 et NEIL2) ou uniquement sur l'ADN simple brin (NEIL3) [194].

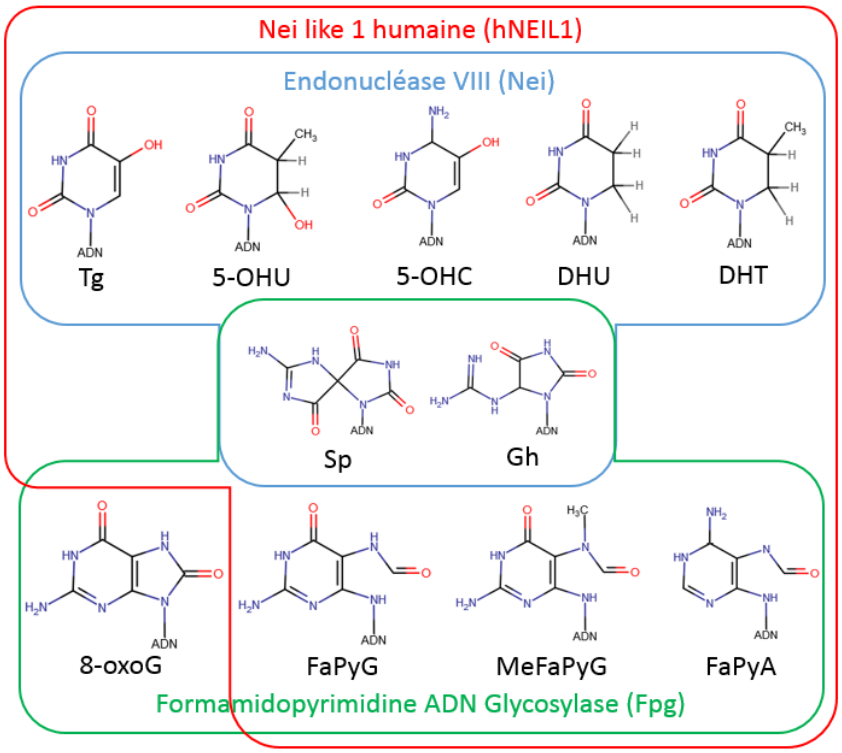


Figure 28 : Liste non exhaustive des substrats préférentiels des Fpg/Nei
Les Fpg (encadré vert) participent à la réparation de purines oxydées (8-oxoG) et de produits issus de cette oxydation (FaPyG, FaPyA, Sp, Gh), les Nei (encadré bleu) excisent uniquement les pyrimidines oxydées (Tg, 5-OHC, 5-OHU, DHU, DHT) ainsi que les purines dont le noyau pyrimidine a été démantelé par l'oxydation (Sp, Gh), NEIL1 (encadré rouge) prend en charge l'ensemble de ces substrats, excepté la 8-oxoG, qui est réparée par Ogg1 [194].

Comme nous l'avons vu plus haut, outre son spectre relativement large de purines et pyrimidines oxydées substrats, la singularité des protéines Fpg est leur capacité d'exciser le produit majeur d'oxydation des purines la 8-oxoG, une lésion extrêmement mutagène (**Figure 15** p. **50**). De ce point de vue, on aurait pu l'appeler 8-oxoG ADN glycosylase bactérienne, homologue fonctionnel de la protéine humaine OGG1 [54]. Le rôle premier de Fpg est donc d'éliminer la 8-oxoG ce qui la distingue clairement des Nei (NEILs) qui en sont incapables. De ce point de vue, Fpg ou OGG1 chez l'Homme sont des anti-mutateurs de la transversion G:C vers T:A et participent avec les protéines MutY (MUTYH chez l'Homme) et MutT (MTH1 chez l'Homme) à lutter contre l'effet mutagène de la 8-oxoG. Chez les bactéries comme chez l'Homme, l'action conjointe de Fpg(MutM)/MutY/MutT et de OGG1/MUTYH/MTH1 sont désignées sous le nom de système GO (« Guanine Oxidized ») (**Figure 29**) [54]. MutY et MUTH sont des adénine ADN glycosylases de la superfamille HhH spécifiques du mésappariement 8-oxoG.A pouvant agir après réplication mutagène de la 8-oxoG si Fpg ou hOGG1 ont été saturées. MutT et MTH1 sont des 8-oxodGTPases capablent de nettoyer le pool de désoxyribonucléotides triphosphates oxydés pouvant servir de substrats aux ADN polymérases. Le rôle fondamental des Nei chez les bactéries n'est pas très clair car sa spécificité de substrat est redondante avec les ADN glycosylases Fpg et Nth (Endonucléase III). Beaucoup de bactéries n'ont pas de Nei. La présence simultanée de Fpg, Nei et Nth chez les protéobactéries n'est pas encore expliquée.

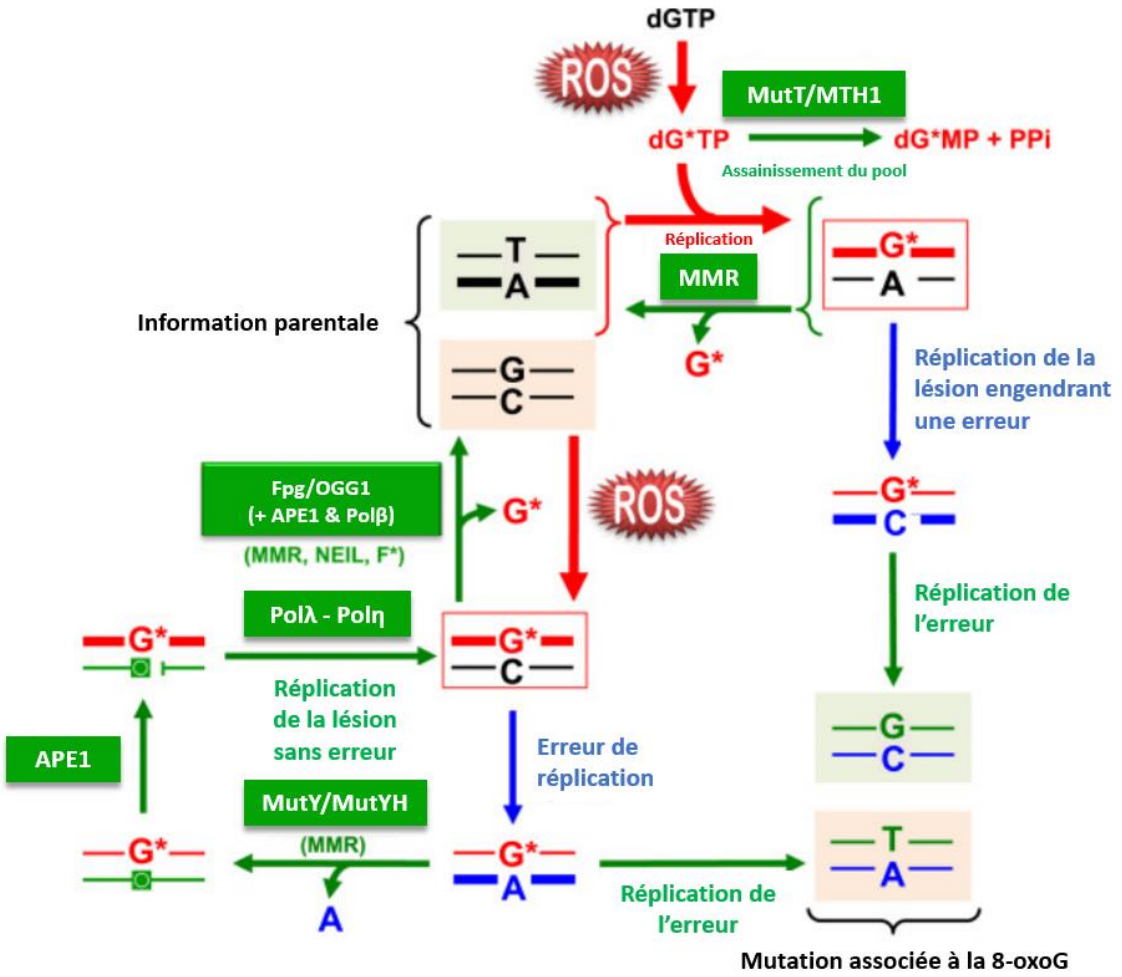


Figure 29 : Réseau de réparation du système GO des bactéries et des mamifères
Les deux voies d'apparition du 8-oxoG (G*) dans l'ADN sont indiquées par des flèches rouges. Les enzymes de réparation GO sont en lettres blanches sur fond vert (les enzymes de secours possibles sont indiquées ci-dessous en lettres vertes entre parenthèses). F* sont des facteurs additionnels tels que les ADN polymérases η et λ mammaliennes. Le brin de modèle utilisé pour la réplication est indiqué par une ligne épaisse. ◙ est pour le site AP. Les mutations associées aux 8-oxoG non réparées sont des transversions G:C à T:A et A:T à C:G [54].

I.3.3 Éléments structuraux clés des protéines Fpg/Nei

Un grand nombre de structures cristallographiques de ces protéines complexées ou non à un court fragment d'ADN double brin ont été résolues ces 25 dernières années. Pour pouvoir saisir des « captures d'images » des enzymes et les piéger avec leurs substrats la mise en place de plusieurs « stratégies » a été nécessaire.

Deux approches générales ont été utilisées pour l'analyse biochimique et structurale par cristallographie de complexes stables impliquant ADN glycosylases Fpg/Nei. En effet, pour capturer le complexe de Michaelis il faut empêcher la protéine de réaliser son activité glycosylase ou AP lyase dans

sa totalité. Dans tous les cas, la protéine est engagée dans des interactions avec des modèles de substrats ou avec les produits réactionnels intermédiaires (complexe covalent de la base de Schiff réduite) ou terminaux de la catalyse enzymatique (duplexe d'ADN contenant une lacune d'un nucléoside, produit final de réaction). L'une de ces approches a consisté à modifier par génie génétique le gène codant pour la protéine Fpg afin d'isoler des protéines de type « mutant inactif » ayant gardées la capacité de reconnaître spécifiquement les substrats. L'autre approche, de loin la plus utilisée pour l'analyse structurale de ces protéines, a consisté à fabriquer par voie chimique des analogues de substrats de ces enzymes reconnus spécifiquement par l'enzyme mais non métabolisables par celle-ci. Parmi ces analogues, mon équipe d'accueil au CBM a particulièrement utilisé des dérivés carba- de nucléosides (comme le carba-FaPyG) dans lesquels le désoxyribose a été chimiquement remplacé par un cyclopentane et des analogues de site abasique (comme le THF) (**Figure 30**). Dans ces conditions, le C1' de l'analogue ne peut être attaqué par l'amine réactive de l'enzyme. Néanmoins, ces analogues contenus dans une molécule d'ADN constituent des leurres moléculaires reconnus spécifiquement par l'enzyme et se comportent comme des inhibiteurs forts des deux activités de l'enzyme (on parle ici d'inhibiteur compétitif). Quelle que soit la stratégie utilisée, l'objectif ici est d'obtenir des complexes Enzyme/ADN stables que l'on peut cristalliser pour initier une étude par diffraction des rayons X (cristallographie).

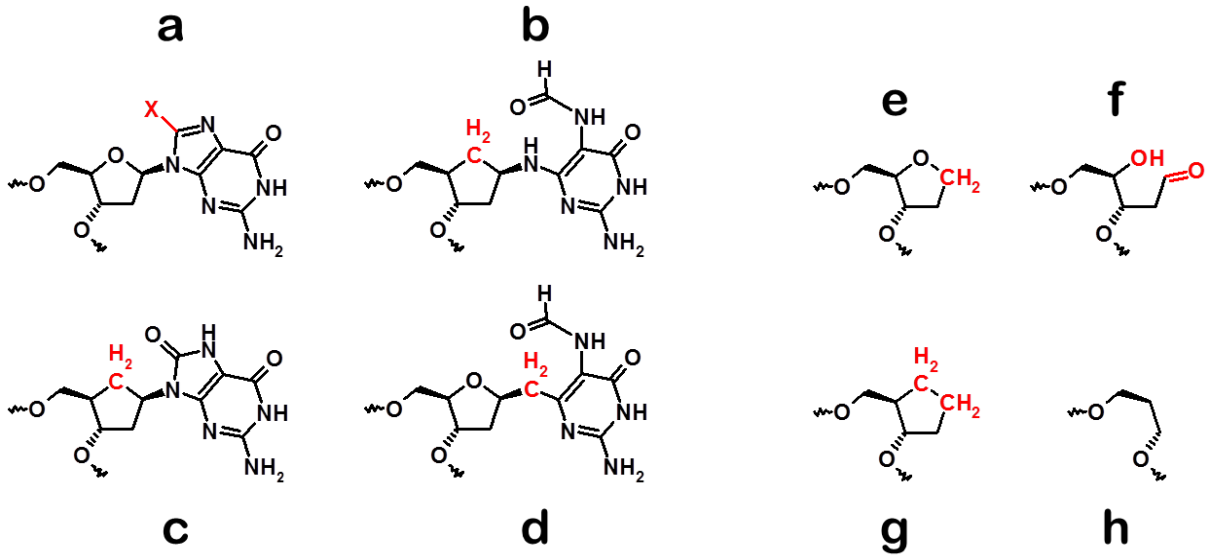


Figure 30 : Analogues de substrats des protéines Fpg/Nei
Deux types d'analogues ont été utilisés pour élucider le mode de fixation à l'ADN des protéines Fpg/Nei: (i) des duplexes d'ADN synthétiques contenant des analogues de guanines oxydées (**a**) dérivés de guanine avec X = -SH, -CH3, -NH_2 et -Br, (**b**) et (**c**) carbanucléosides cFapyG et c8-oxoG, respectivement, et (**d**) C-nucléoside et (ii) des duplexes d'ADN contenant des analogues cycliques ou non-cycliques de site abasique (**e**) tétrahydrofurane (THF), (**f**) site abasique réduit, (**g**) cyclopentanediol et (**h**) 1,3-propanediol [188, 195, 196].

Toutes ces structures ont permis de mieux appréhender le fonctionnement des Fpg/Nei. Les membres de la superfamille Fpg/Nei se caractérisent par plusieurs éléments structuraux bien conservés, notamment un domaine N-terminal composé majoritairement de feuillets β et d'un domaine C-terminal composé presque exclusivement d'hélices α, ainsi que d'autres éléments tels que les motifs ZnF/ZnLF, le motif H2TH, le site actif décrits plus en détail dans cette section (**Figure 31**).

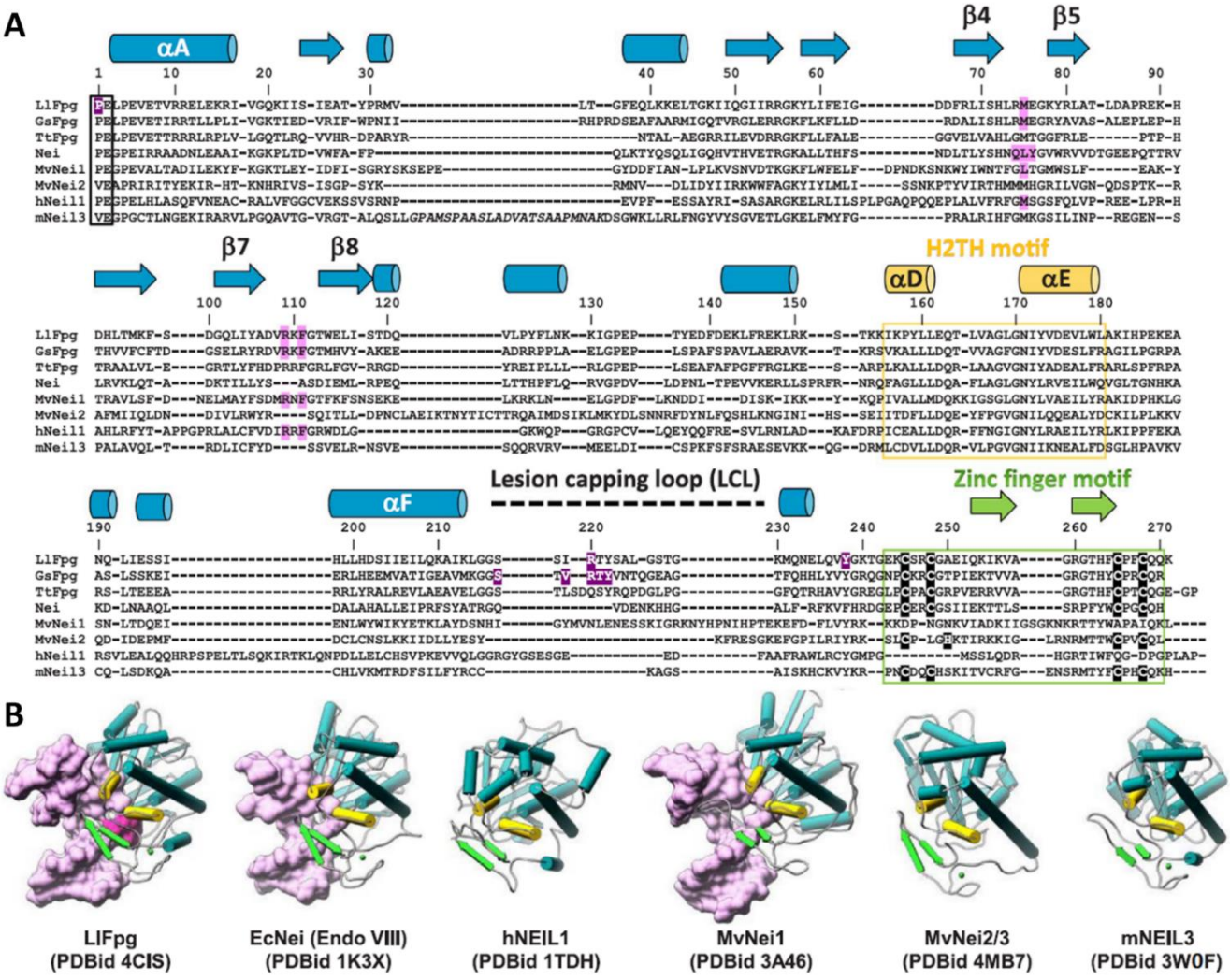


Figure 31 : Alignement des séquences protéiques et structures des protéines Fpg, Nei et NEIL
A) Structures primaires de Fpg/Nei alignées selon leurs structures 3D et structure secondaire de Fpg *L. lactis* avec les hélices α représentées par les cylindres et les brins β représentés par flèches. En encadré jaune, le motif H2TH propre à la superfamille, et en encadré vert le motif doigt à zinc. La boucle LCL est représentée en ligne pointillée. Les aa 1 et 2 catalytiques sont encadrés en noir, tandis que les aa s'intercalant dans l'ADN pour combler le gap de la lésion extraite sont surlignés en rose. Les résidus entrant en contact avec la 8-oxoG sont surlignés en magenta, et les cystéines du doigt à zinc en noir. **B**) Vue globale sur les structures des protéines Fpg/Nei. L'ADN est représenté en surface rose pâle, la protéine en cartoon bleu avec le motif H2TH présenté en cartoon jaune et le motif doigt à zinc en cartoon vert. L'ion Zn^{2+} est également représenté par une sphère verte. Les ~100 derniers aa des NEIL sont omis car seules les protéines tronquées en C-terminal peuvent être utilisées pour l'obtention de

cristaux. Les protéines MvNei2 et mNEIL3 complexée à un ADN sont présentées libres car les complexes ne sont pas encore résolus en cristallographie [54].
Abréviations : *Ll*, *Lactococcus lactis* ; *Ec*, *Escherichia coli* ; h, human ; Mv, mimivirus ; m, mouse.

I.3.3.1 Le motif doigt à zinc

Les Fpg/Nei sont des protéines comportant un « doigt de zinc » (ZnF) correspondant à un motif épingle à cheveux composé de deux feuillets β espacés par un coude (β-hairpin). L'ion métallique Zn^{2+} est coordonné par les chaines latérales de quatre Cys (ou trois Cys et une His, cf. MvNei2 **Figure 31 A**) conservées dans la séquence –C-X_2-C-$X_{16\text{-}18}$-C-X_2-C–, avec 16 aa centraux pour les protéines procaryotes et 18 pour les protéines eucaryotes. Ce motif est présent chez toutes les Fpg/Nei, cependant MvNei1 et hNEIL1 ne possèdent pas ces Cys ni d'atome de zinc. En revanche, le motif β-hairpin est tout de même bien présent, il est aussi appelé motif « doigt sans zinc » (ZnLF) (**Figure 27 C** et **Figure 31 B** motif en vert). Il a été montré au laboratoire que les protéines dont les doigts à zinc étaient déstructurés, notamment les protéines Fpg, n'étaient plus capables de se fixer à l'ADN [176, 197, 198]. Le ZnF est donc un élément important pour la fixation de la protéine à l'ADN [199, 200]. Pour *Ll*Fpg, le ZnF correspond aux positions allant de 243 à 270 et pour hNEIL1, le ZnLF correspond aux positions 267 à 285. Une Arg est extrêmenent conservée dans les motifs doigt à zinc ou doigt sans zinc, il s'agit de la position R260 pour *Ll*Fpg et R276 pour hNEIL1 (**Figure 31 A** encadré vert). Ce résidu est directement en contact avec l'ADN, et participe la torsion de l'ADN à 65° (neutralisation des phosphates côté concave de la torsion) au niveau de la base endommagée en interagissant avec les groupements phosphates de la base endomagée et du nucléotide $i+1$. (**Figure 32**). La torsion de l'ADN facilite l'extrusion de la base de la double hélice et son positionnement correct dans le site actif de l'enzyme.

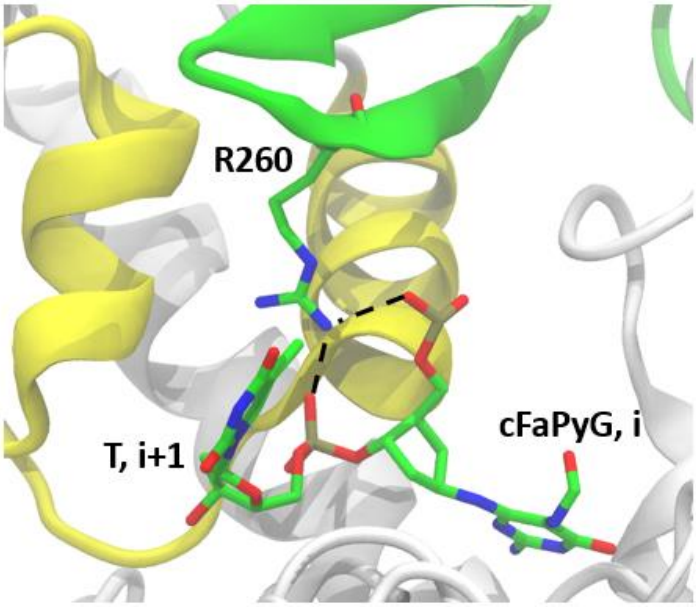


Figure 32 : Interaction entre la R260 et les groupements phostates de la base endommagée et du nucléotide suivant
Sur cette figure, le ZnF est en cartoon vert, le motif H2TH en jaune et le reste de la protéine en blanc. La R260, le cFaPyG et le nucléotide suivant sont représentés en bâtonnets [201].

I.3.3.2 Le motif H2TH

Les protéines de la superfamille Fpg/Nei comportent toutes un motif Hélice-2-Tours-Hélice (H2TH) (**Figure 31 A** motif en jaune). Les séquences de cet élément structural sont très proches parmi les individus de cette superfamille. Ce motif est connu dans la littérature pour être impliqué dans la fixation de la protéine à l'ADN [199]. Sur la protéine *Ll*Fpg, il correspond aux aa 156 à 180 et aux aa 161 à 185 sur la protéine hNEIL1.

I.3.3.3 Les acides aminés catalytiques

Pour commencer, les acides aminés catalytiques sont bien identifiés et correspondent aux aa P1 (ou V1 chez MvNei2 et MvNei3) et E2 (ou P2 et E3 si le compte des aa dans une séquence commence à partir de la méthionine initiatrice qui n'apparait jamais dans les structures cristallographiques), très conservés et localisés à l'extrémité de l'hélice α N-terminale (**Figure 31 A** encadrés en noir). Ces aa sont flanqués entre les deux domaines globulaires composants la protéine. Les mutations P1G ou ΔP1 et E2Q ou E2A sont un moyen de piéger la protéine sur son substrat car ces aa sont essentiels à l'activité glycosylase (**Tableau 10** p. **119** en fin de chapitre). La perte de l'activité glycosylase causée par la mutation de P1 ou E2 est toujours accompagnée de la perte de l'activité AP lyase.

I.3.3.4 La « Lesion Capping Loop » (LCL)

Le dernier élément structural important est la Lesion Capping Loop (LCL), initialement identifiée chez les protéines Fpg. Elle est aussi appelée boucle αF-β9/αF-β10, et se situe dans l'espace à proximité du site actif. Cette boucle est nommée ainsi car elle engage beaucoup de contacts avec les bases altérées (8-oxoG et FaPyG chez les enzymes Fpg) présentées au site actif avant leur excision par l'enzyme [183]. La boucle LCL présente des segments très flexibles (segments B et C **Figure 33 a**) mobilisés pour la reconnaissance de la base endommagée et qui peuvent être hyperdynamiques en absence de base dans le site actif (segment B **Figure 33 a**) (non résolue dans les structures cristallographiques). Les aa composant la boucle LCL (217 à 230 pour *Ll*Fpg) sont mal conservés parmi les Fpg/Nei. Cependant, d'après les structures de *Ll*Fpg complexée avec un ADN comportant une lésion carba-8-oxoG (c8-oxoG) ou carba-FaPyG (cFaPyG), il semblerait que seul le squelette de ces quatre résidus (positions *L. lactis* de 219 à 222) contacte les lésions chez Fpg. D'après l'alignement de séquences, la boucle LCL semble être différente chez Nei/hNEIL1/MvNei2/mNEIL3 pour lesquelles elle est beaucoup plus courte voire presque inexistante ou encore plus longue comme chez MvNei1 (**Figure**

31 motif en vert et **Figure 33 a**). Les structures de *Ll*Fpg ont permis d'identifier deux états de la boucle LCL, un état fermé en général en présence d'un substrat ou analogue dans le site actif comme cFaPyG (*Ll*Fpg) [184] ou une 8-oxoG (*Bst*Fpg) [183], et un état relâché observé en général en l'absence de base dans le site actif (par exemple avec un site abasique) [195] ou en présence de c8-oxoG (*Ll*Fpg) [202]. D'autres structures de Fpg *G. stearothermophilus* (*Bst*) ont montré des structures de la boucle fermée en présence de 8-oxoG dans le site actif [183] et même en l'absence de substrat dans le site actif pour Fpg issue de l'organisme *Thermus thermophilus* (*Tt*) [186] (**Figure 33**). Ceci peut s'expliquer par le fait que l'organisme dont est issue cette Fpg est un organisme thermophile et dont les protéines doivent être thermostables. Cette caractéristique facilite la production de cristaux à température ambiante et donc l'obtention de structures cristallographiques de cet élément qui semble être très flexible. Lorsque la boucle LCL est retirée par mutagénèse dirigée, la protéine Fpg n'est plus capable d'exciser la 8-oxoG, mais reconnait toujours le FaPyG [54, 185]. Des études supplémentaires impliquant la création de mutants dont la boucle LCL avait été modifiée ou délétée ont montré que cet élément est essentiel pour la reconnaissance de la 8-oxoG mais pas pour d'autres lésions [185, 203]. Les bases moléculaires impliquant la boucle LCL pour la reconnaissance de la 8-oxoG demeurent encore obscures et méritent

d'être clairement élucidées.

Figure 33 : Éléments structuraux de la boucle « Lesion Capping Loop » (LCL) des protéines Fpg/Nei
Deux états de la boucle LCL, (i) fermé (PDBid : 1XC8 et 1R2Y), (ii) relâché (PDBid : 1PM5) (iii) non résolue (PDBid : 4CIS) [54]. **a**) Alignements des structures primaires d'après les structures 3D. La boucle LCL peut être subdivisées en quatre segments : (A) une partie proximale rigide, (B) une partie hyperdynamique qui n'adopte une structure précise que lorsque Fpg reconnait la 8-oxoG ou FaPyG (souvent non résolue dans les structures cristallographiques de Fpg/Nei reconnaissant une site abasique ou des pyrimides oxydées [182, 188, 204]), (C) une partie changeant de structure entre la boucle LCL en conformation fermée et conformation relâchée et (D) la partie distale rigide de LCL. En gras (rouge ou noir) les aa relativement bien conservés. Les sphères bleues indiquent les résidus impliqués dans la reconnaissance de la 8-oxoG. **b**) Superposition de la boucle LCL en conformation fermée ou relâchée (aussi dite ouverte).

I.3.3.5 Le site actif

En plus de la boucle LCL, d'autres aa interviennent dans de nombreuses interactions avec la base en position extra-hélicale, ainsi qu'avec le groupement phosphate du nucléoside endommagé (**Figure 34**). Cependant, l'orientation de la base et les interactions que la protéine forment avec cette dernière ne sont pas les mêmes selon le type de lésion, ce qui traduit une grande flexibilité du site actif et peut expliquer la grande variété structurale des substrats reconnus par ces enzymes. On note quand même quelques redondances comme les aa stabilisant le squelette de la lésion (groupement phosphate et sucre) qui sont très conservés dans la famille Fpg/Nei et dont les positions chez *Ll*Fpg sont N171, Y238 et R260 (**Figure 34**, **A** et **B**). Les aa catalytiques P1, E2 et E5 très conservés dans la superfamille et ceux de la boucle LCL se situent à proximité de la liaison *N*-glycosidique de la base lésée, ou des aa situés sur une hélice α chez hNEIL1 sont impliqués dans de nombreuses liaisons hydrogène avec la lésion. Sur la **Figure 34**, on peut visualiser les différences de la structure secondaire au niveau de l'équivalent de la boucle LCL entre les structures de Fpg (**Figure 34 C**) et NEIL1 (**Figure 34 D**). Pour le Tg, ce sont les chaines latérales des aa chargées positivement ou aromatiques (Arg, Lys ou Phe) qui interagissent avec la base et non leur squelette, contrairement à Fpg.

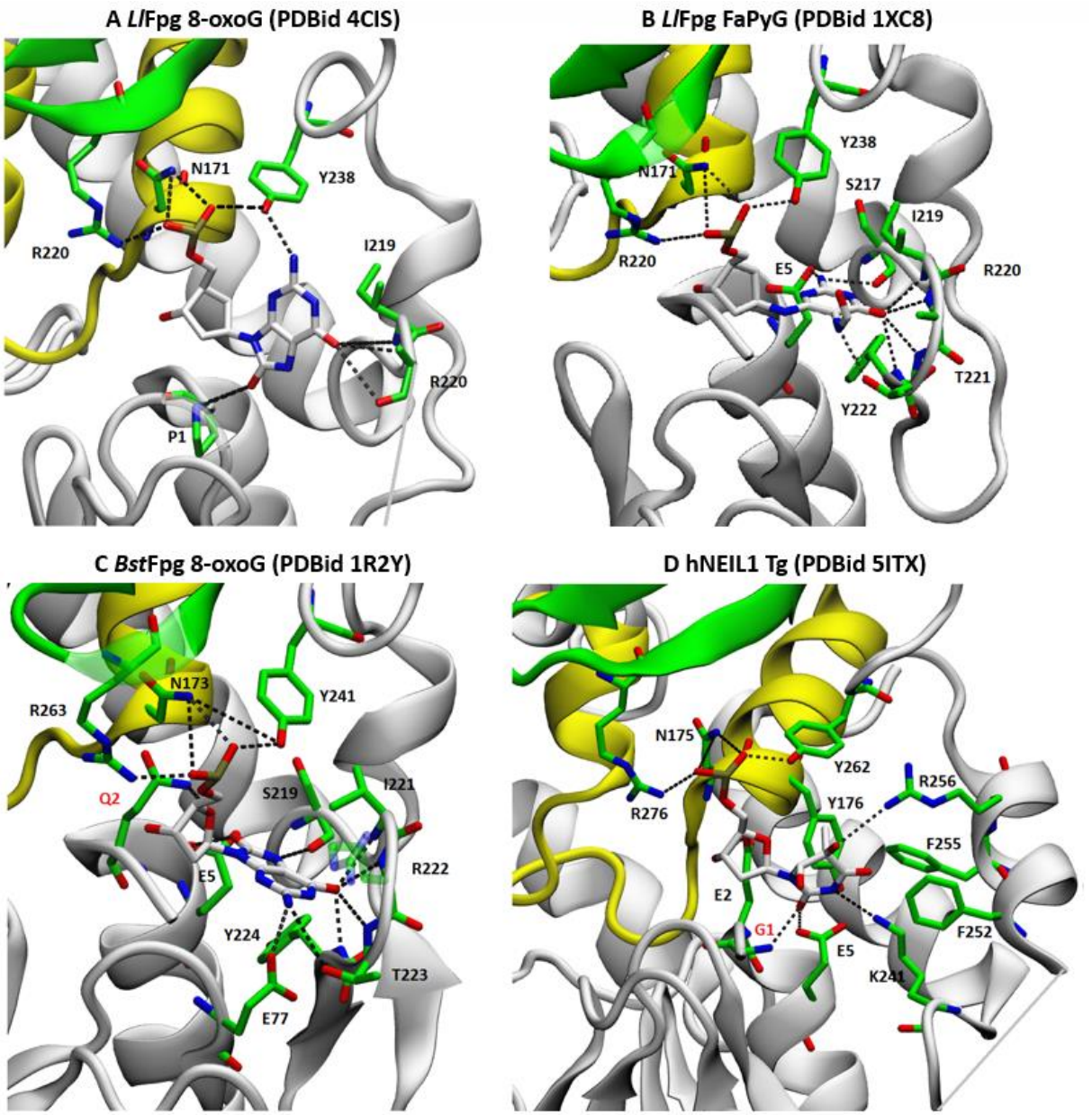


Figure 34 : Différentes conformations des substrats dans les sites actifs des protéines *Ll*Fpg, *Bst*Fpg et hNEIL1

Dans les figures de **A** à **D**, la protéine est représentée en cartoon blanc, les motifs H2TH en jaune et les ZnF et ZnLF en vert, les bases altérées et les aa à moins de 3.5 Å de cette dernière sont en bâtonnets blancs verts respectivement. Les portions de la boucle LCL non résolues sont remplacées par un trait gris. Les noms des aa sont indiqués en noirs, et les aa mutés sont inscrits en rouge. **A**) *Ll*Fpg WT complexée à ADN carba-8-oxoG en position intermédiaire entre *syn* et *anti*, la boucle LCL n'est pas résolue, on suppose qu'elle est dans l'état relâchée (PDBid : 4CIS [202]). **B**) *Ll*Fpg WT complexée à un ADN carba-FaPyG en position *syn* (PDBid : 1XC8 [201]). **C**) *Bst*Fpg muté E2Q complexée à un ADN 8-oxoG en position *anti* (PDBid : 1R2Y [183]). **D**) hNEIL1 mutée P1G complexée à un ADN Tg en position intermédiaire entre *syn* et *anti* (PDBid : 1TDH [190]).

I.3.3.6 La triade d'intercalation

Les Fpg/NEIL1 présentent certains autres aa qui, d'après les structures cristallographiques, semblent également posséder une fonction importante dans la stabilisation du complexe ADN protéine, notamment en s'intercalant dans la double hélice (positions M75, R109 et F111 sur *Ll*Fpg **Figure 35 C** et positions M80, R117 et F119 sur hNEIL1 **Figure 31**). La nature et la position de ces aa varient chez EcNei (3 aa adjacents dans la boucle entre les brins β4 et β5) ou n'existent pas chez MvNei2 et MvNei3. Pour les autres enzymes telles que Fpg et hNEIL1, ces aa sont toujours au nombre de trois (d'où triade d'intercalation), ils sont tous situés dans le domaine N-terminal de l'enzyme à la jonction entre les brins β4 et β5 puis β7 et β8 (**Figure 31** et **Figure 35**), et sont entourés d'autres aa chargés positivement (tels que la Lys et Arg) ou dont la chaine latérale est très courte (comme la Gly) rendent la protéine très flexible localement.

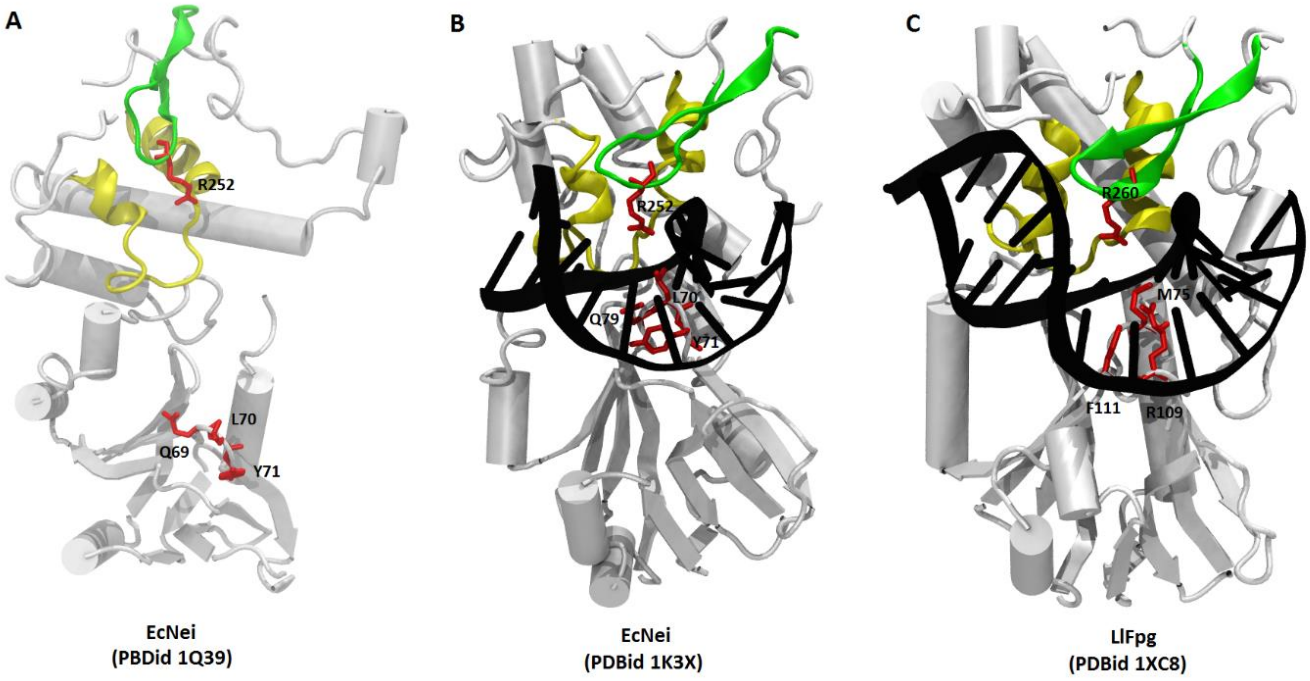


Figure 35 : Représentation de la triade d'intercalation de plusieurs Fpg/Nei libre ou complexées
En cartoon blanc la protéine, et en new carton jaune le motif H2TH, en vert le ZnF et en noir l'ADN. En rouge sont représentés les aa faisant partie de la triade d'intercalation (aa différents selon la protéine), et l'Arg dans le coude du motif doigt à zinc très conservée parmi toutes les Fpg/Nei (**Figure 31**). **A**) *Ec*Nei libre (PDBid : 1Q39 [161]) montre un relâchement et un éloignement entre les domaines N-terminal et C-terminal, la protéine est déstructurée à plusieurs endroits. La distance entre les extrémités des chaines latérales de L70 et R252 est de 28.5 Å. **B**) *Ec*Nei complexée à un ADN contenant la lésion thymine glycol bromée en C5 (PDBid : 1K3X [189]), permettant de capturer le complexe sans que la base soit coupée. Cette figure permet d'apprécier l'intercalation de Q69, L70 et Y71 dans le petit sillon de l'ADN à la place de la base extraite de la double hélice, tout en mettant en évidence l'interaction entre L70 et R252, aa faisant partie du doigt à zinc (en new cartoon vert). **C**) *Ll*Fpg complexée à un ADN contenant un carba-FaPyG (PDBid : 1XC8 [184]). Le carba-FaPyG est un analogue au FaPyG dont l'atome d'oxygène du sucre a été remplacé par un carbone, rendant la formation de la base de Schiff impossible, et empêchant donc l'excision. Sur cette figure, on constate que les aa de la

triade d'intercalations ne sont pas toujours adjacents dans la séquence (cf **B** et **C**). La thymine glycol bromée et le carba-FaPyG sont des analogues de substrats, ils ont été décrits précédemment.

Un autre aa important et très conservé chez toutes les Fpg/Nei est une Arginine (Arg) située dans le coude entre les deux brins β formant le motif ZnF ou ZnLF selon les cas (R252 chez *Ec*Fpg **Figure 35 A**-**B** et R260 chez *Ll*Fpg **Figure 35 C,** R276 chez hNEIL1) [54]. Cette Arg est importante pour stabiliser la forte torsion centrée sur le nucléoside endommagé que l'enzyme impose à l'ADN. Comme décrit précédemment, la charge positive de cette Arg interagit directement avec les charges négatives des groupements phosphates de la base endommagée et de la base en 3' de cette dernière permettant ainsi que ces phosphates se rapprochent du côté concave de la torsion [185]. Cet aa remplit la lacune laissé par la base en position extra-hélicale et forme deux liaisons hydrogènes avec la cytosine opposée au dommage, permettant de la maintenir en position intrahélicale [184]. Ces interactions permettent donc de stabiliser le complexe enzyme substrat. Elle forme également un pont à travers le gap laissé par la base extraite de la double hélice *via* une interaction avec une Leu (position *Ec*Nei L70 **Figure 35 A**-**B**, position *Ll*Fpg M75 **Figure 35 C**) d'intercalation située entre les brins β4 et β5, et participe à la torsion de l'ADN en se positionnant entre les deux groupements phosphate avant et après le nucléotide endommagé, à l'intérieur du grand sillon de l'ADN (**Figure 35**). Ce pont est brisé en l'absence d'ADN, la distance entre les extrémités des chaines latérales est de plus de 28 Å. Ceci n'est constatable qu'avec les enzymes *Ec*Nei et MvNei1 et plus récemment avec hNEIL1, car ce sont les seules qu'il est possible de cristalliser libres et complexées à l'ADN. Les protéines Fpg des organismes mésophiles sont très dynamiques lorsqu'elles sont libres (sans ADN), à l'exception de *Tt*Fpg, qui n'a pas été encore cristallisée complexée à l'ADN. De plus, une interface fortement basique de ces enzymes interagit avec les charges négatives des groupements phosphate de l'ADN.

La protéine *Ath*Fpg est un cas particulier, car il n'est aujourd'hui plus évident de la classer en tant que Fpg mais plutôt en tant que Nei/NEIL1. En effet, cette enzyme n'est pas capable de procéder à l'excision de la 8-oxoG et, elle possède un motif doigt sans zinc et le repliement correspondant à la boucle LCL est très raccourci, voire inexistant tout comme les Nei/NEIL1 [185].

I.3.4 Reconnaissance des bases oxydées par les protéines Fpg/Nei

Comme décrit dans les sections précédentes, plusieurs éléments structuraux rentrent en contact en premier lieu avec l'ADN comme le motif de reconnaissance H2TH et le motif doigt à zinc ou doigt sans zinc, puis la base endommagée en position extra-hélicale dans le site actif comme les aa catalytiques (P1, E2, E5) ainsi que ceux de la boucle LCL. Pour finir, la triade d'intercalation prend

l'espace vacant à l'intérieur de la double hélice laissé par la base extraite. D'après les études présentées ci-dessous, chacun de ces éléments participe à la reconnaissance et la stabilisation du substrat et sont essentiels pour l'activité glycosylase.

I.3.4.1 Les complexes IC, EC et LRC

Le « Disulfide Cross Linking » (DXL) est une technologie qui permet d'obtenir des modèles correspondant à différents états de la protéine appelés « Interrogation Complexes » (IC, Fpg diffusant sur l'ADN) et « Encounter Complexes » (EC, Fpg reconnaissant une 8-oxoG dans l'ADN mais avant l'extraction de cette dernière de la double hélice) difficilement observables [205]. Cette technique consiste en l'ajout d'atomes de soufre sur le squelette de la double hélice l'ADN, ce qui induit la formation de ponts disulfures avec une Cys de la protéine (C166 de *Bst*Fpg), bloquant cette dernière à l'étape précédant la reconnaissance du substrat. Les analogues de substrats et les protéines mutées permettent de proposer d'autres modèles appelés « Recognition Complexe » (RC) et « Lesion Recognition Complexes » (LRC) respectivement. Les agents réducteurs tels que le borohybride de sodium sont aussi utilisés pour piéger les complexes covalents ADN/protéine se formant pendant la catalyse et permettent ainsi d'apprécier les structures dites « Schiff base intermediates » (SBI) (**Tableau 10** p. **119** en fin de chapitre) [162, 189, 206]. Ces structures IC, EC, LRC et SBI sont à l'origine de plus amples études en modélisation moléculaire. Ces études ont permis de mieux appréhender les mécanismes de reconnaissance des lésions dans la double hélice d'ADN, le mouvement de « base extrusion » et la coupure. Tous ces phénomènes qui font partie de l'activité glycosylase des ADN glycosylases Fpg/NEIL1.

Le premier phénomène intéressant est donc la reconnaissance des bases oxydées par les ADN glycosylases Fpg/Nei. Le cas le plus étudié est notamment la distinction des 8-oxoG parmi les guanines classiques par la protéine *Bst*Fpg. D'après l'analyse de plusieurs structures cristallographiques IC, EC et LRC de *Bst*Fpg (PDBid : 3GPY, 3GO8, 3GP1, 3GPP, 3GPU, 3GPX, 3GQ3, 3GQ4, 3GQ5 [203]), les aa M77, R112 et F114 (équivalents à M75, R109 et F111 de *Ll*Fpg) sont intercalés dans la double hélice de l'ADN bien avant le mouvement permettant l'extraction de la base endommagée hors de la double hélice (**Figure 36 A** à **C**) [203]. Aussi, une différence dans la conformation du groupement phosphate de la base non extraite est mise en évidence entre les structures IC et EC. Elle semble être induite par un clash stérique non favorable entre l'O8 de la 8-oxoG et le groupement phosphate du nucléotide portant la base altérée (**Figure 36 D**). Il en résulte un éloignement de ce groupement phosphate ainsi qu'un décalage du sucre par rapport à sa position normale. Ces éléments du squelette de l'ADN déplacés sont en contact avec des aa de *Bst*Fpg qui sont différents lorsque le squelette est en position

normale, donnant un indice sur le mode de reconnaissance de la 8-oxoG par cette Fpg. Cependant comme plusieurs lots de structures IC/EC ont été produits et que ces contacts sont différents dans chacun de ces cas, il est encore difficle de les identifier avec certitude. D'après les auteurs, les différences de contact dans ces structures sont potentiellement liées à la différence dans les séquences d'ADN utilisées dans cette étude, ce qui pourrait aussi signifier que selon le type de nucléotides entourant le dommage, la protéine ne fait pas intervenir les mêmes aa dans la reconnaissance des guanines oxydées que dans la reconnaissance des autres lésions.

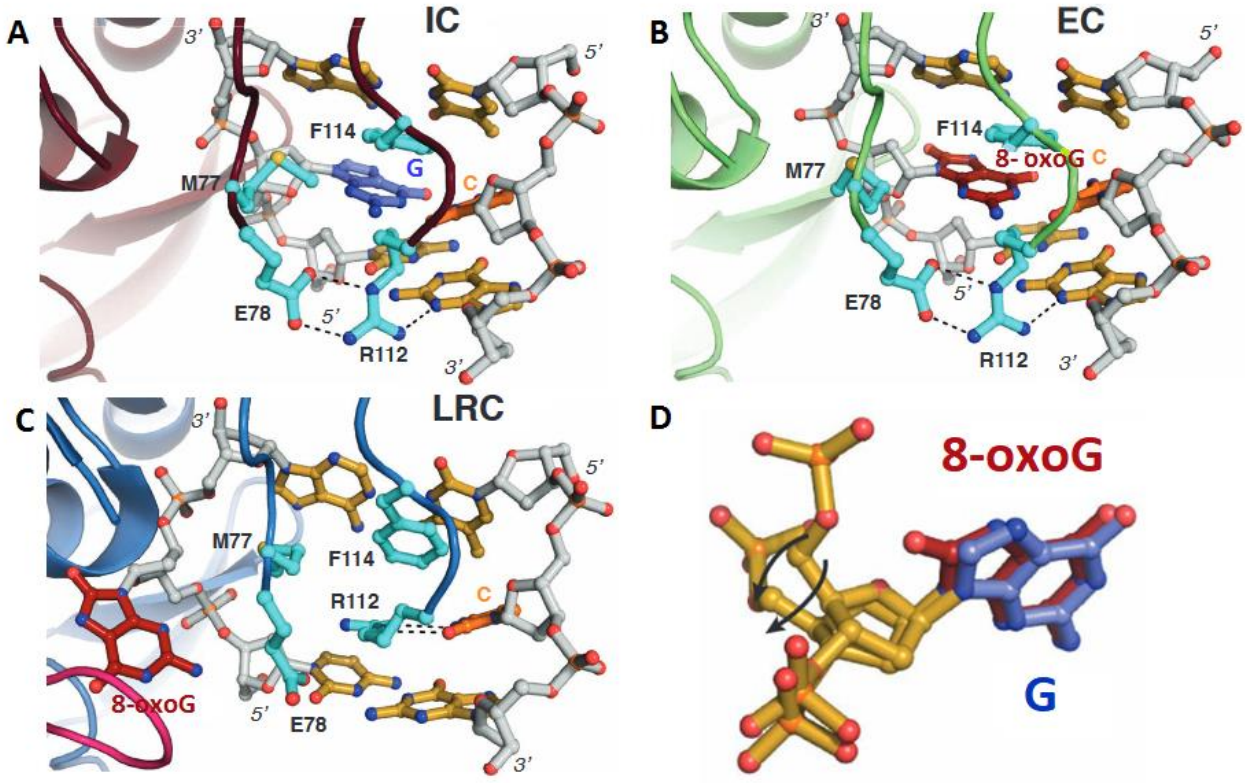


Figure 36 : Analyse de la reconnaissance d'une guanine oxydée par *Bst*Fpg
A) « Interrogation Complex » (IC) en présence d'une G (PDBid : 3GPX), **B**) « Encounter Complex » (EC) (PDBid : 3GPU), **C**) « Lesion Recognision Complex » (LRC) en présence d'une 8-oxoG par *Bst*Fpg, **D**) superposition des IC (bleu et orange) et EC (rouge et orange), mettant en évidence une différence de conformation du groupement phosphate ainsi que du sucre du nucléotide portant la base altérée [203].

Cette même étude a également pour ambition d'étudier le mouvement de « base extrusion » d'une base G classique et d'une 8-oxoG, grâce à des techniques de dynamique moléculaire ciblée et dirigée produite par le logiciel CHARMM [207]. Les auteurs mettent en évidence plusieurs étapes dans le mouvement permettant l'extraction du dommage, qui sont (i) la perturbation du mésappariement contenant l'aberration, (ii) la rotation autour de la liaison *N*-glycosidique de la base cible à partir de l'orientation *anti* jusqu'à *syn* au sein du petit sillon et (iii) retournement de la base dans une position extra-hélicale. La barrière énergétique à franchir pour lors de l'extraction d'une 8-oxoG est moindre (~4,0 kcal/mol) que pour l'extraction d'un nucléotide G classique (au-delà de 11,0 kcal/mol). De plus,

dès le passage de l'étape (i) à (ii), l'aa R112 perturbe le mésappariement et envahit la double hélice d'ADN pour interagir avec la C auparavant opposé au dommage (**Figure 36 C**). Une 8-oxoG est donc plus facile à extraire de la double hélice et à présenter au site actif qu'un nucléotide G classique et l'aa R112 est impliqué dans le passage de l'état intra- à extra-hélical de la 8-oxoG.

I.3.4.2 La discrimination entre bases normales et bases oxydées

Plus récemment, une étude de mécanique quantique/mécanique moléculaire (QMMM) évoque le fait que la discrimination entre les bases classiques et le substrat se fait avant son extraction de la double hélice ce qui appuie les résultats des études présentées précédemment. En effet les auteurs ne remarquent pas de facteur permettant à la protéine de discriminer une base normale (ici G classique) ou d'une base endommagée (ici FaPyG) dans le site actif de *Ll*Fpg. Cela signifierait que la protéine *Ll*Fpg n'effectue pas le mouvement permettant l'extraction avec toutes les bases qu'elle rencontre, seulement avec les 8-oxoG [208]. D'après plusieurs structures cristallographiques, après l'extraction de la lésion de l'ADN par Fpg, l'enzyme forme des interactions lésion spécifiques afin de la stabiliser le complexe avant la coupure [195, 201].

I.3.4.3 Le mouvement de « base extrusion »

Le mouvement de « base extrusion » d'un nucléotide G normal ou endommagé (8-oxoG) à l'intérieur du site actif de *Bst*Fpg a fait l'objet d'une autre étude basée sur des simulations de dynamique moléculaire et « Nudged Elastic Band » (NEB) combiné à de « l'Umbrella Sampling » (US) sous le logiciel AMBER [209]. Le NEB est une technique permettant de passer des barrières énergétiques en déformant le système (ici rotation de la 8-oxoG par rapport à la liaison *N*-glycosidique, changeant la valeur de l'angle dit « d'éversion ») et d'accéder à de nouveaux états du système. Ces états sont prolongés en DM classique tout en conservant les contraintes sur l'angle désiré sur un court temps pour permettre une meilleure estimation de l'énergie libre, c'est la méthode de l'US. Cette étude permet d'établir un modèle du chemin probable de la base avant son positionnement dans le site actif et d'identifier de potentiels aa intervenant spécifiquement dans la reconnaissance des 8-oxoG dans un ADN contenant majoritairement des G normales. Dans cette étude, les auteurs définissent l'angle d'éversion à partir de la liaison *N*-glycosidique de la 8-oxoG. D'après les profils énergétiques de cet angle dans les deux systèmes 8-oxoG et G classique, le mouvement permettant l'extraction du dommage se ferait selon 4 étapes durant lesquelles la base entre en contact avec différents aa (**Figure 37**). Ces étapes correspondent à plusieurs positions de la base par rapport à la double hélice, une première intra-hélicale où seuls des aa positivement chargés ou très flexibles du doigt à zinc (K257,

R263, G264) entre en contact avec la base, une deuxième et troisième phases intermédiaires où la base entame son mouvement et contacte les aa catalytiques (P1) et quelques aa stabilisateurs de la lésion (N173), et une quatrième et dernière phase durant laquelle la base entre en contact avec la boucle LCL (S219) et des boucles intercalantes (E77, R79), tout en maintenant ses interactions avec les aa catalytiques et stabilisateurs (**Figure 37**). Lors des simulations, le profil d'énergie libre de l'angle d'éversion dans le système comportant la 8-oxoG est beaucoup plus bas (< 7 kcal/mol) que le profil de la G normale (~22 kcal/mol) après le stade I (**Figure 37** et **Figure 38**), ce qui montre que *Bst*Fpg facilite sélectivement le mouvement de « base extrusion » des 8-oxoG par rapport aux G classiques, en passant par un certain nombre d'étapes intermédiaires faisant intervenir des interactions entre la base et des aa clés tels que P1, N173 et R263 qui reconnaissent le O8 (qui pour rappel fait toute la différence entre la guanine normale et la 8-oxoG). Tout comme l'étude précédente, les auteurs identifient l'aa R111 comme un participant à la stabilisation de la base opposée au dommage, ici une C, en remplissant le gap dans la double hélice laissé par la lésion en position extra-hélicale. En parallèle l'O8 de la 8-oxoG est contacté par les aa N173 et P1 dès le début des stades III et IV respectivement (**Figure 38**), évènements qui ne semblent pas se produire avec une G classique.

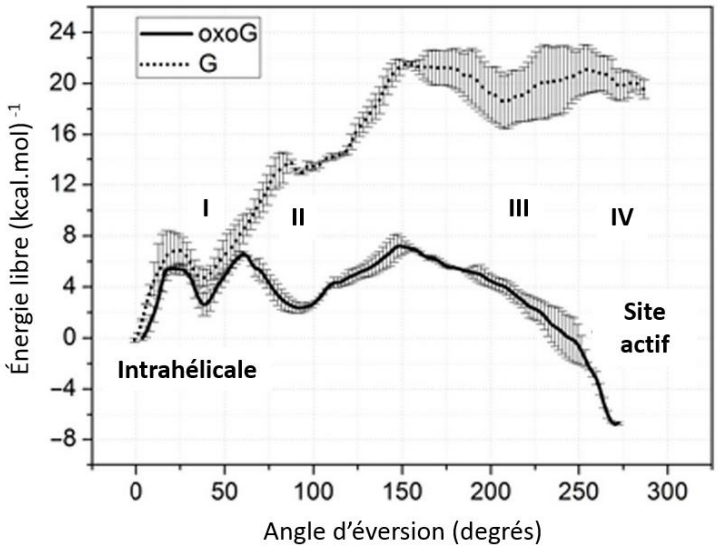


Figure 37 : Valeur d'énergie libre en fonction de l'angle d'éversion
Figure issue de l'article de Li H. *et al.*, Nucleic Acids Res. (2016) [209].

Cette étude ainsi que toutes les structures disponibles des Fpg/NEIL1 couplées à des ADN DB comportant des bases endommagées non excisées (**Tableau 10** p. **119** en fin de chapitre) permettent de comprendre comment la protéine stabilise son substrat avant la coupure. Cette vision est quelque peu incomplète car les structures avec tous les substrats possibles pour ces enzymes n'ont pas été résolues, mais permet d'avoir un premier aperçu. Les segments de la protéine impliqués dans la stabilisation du substrat dans le site actif ne semblent pas être exactement similaires sur Fpg (*Ll* : P1, E2, E5, N171, S217 I219, R220, T221, Y222, Y238, R260) et NEIL1 (humaine : P1, E2, E5, N175, Y176, K241, F252, F255, R256, Y262, R276), la différence vient surtout de la région correspondant à la boucle LCL (**Figure 31**). Cette différence dans la boucle LCL pourrait expliquer les substrats pris en charge par les deux enzymes sont distincts.

Le rôle de la boucle LCL des Fpg dans la reconnaissance des bases endommagées n'est pas encore très clair, et ce pour deux raisons. Ce motif semble très dynamique et n'est pas toujours résolue sur les structures cristallographiques. Cependant, les structures de *Bst*Fpg et *Ll*Fpg par exemple permettent de visualiser un certain nombre d'interactions entre la boucle LCL et la base endommagée 8-oxoG (PDBid : 1R2Y) et un FaPyG (PDBid : 1XC8) respectivement [183, 184, 201]. Ces structures ont d'ailleurs fait l'objet de plus amples investigations.

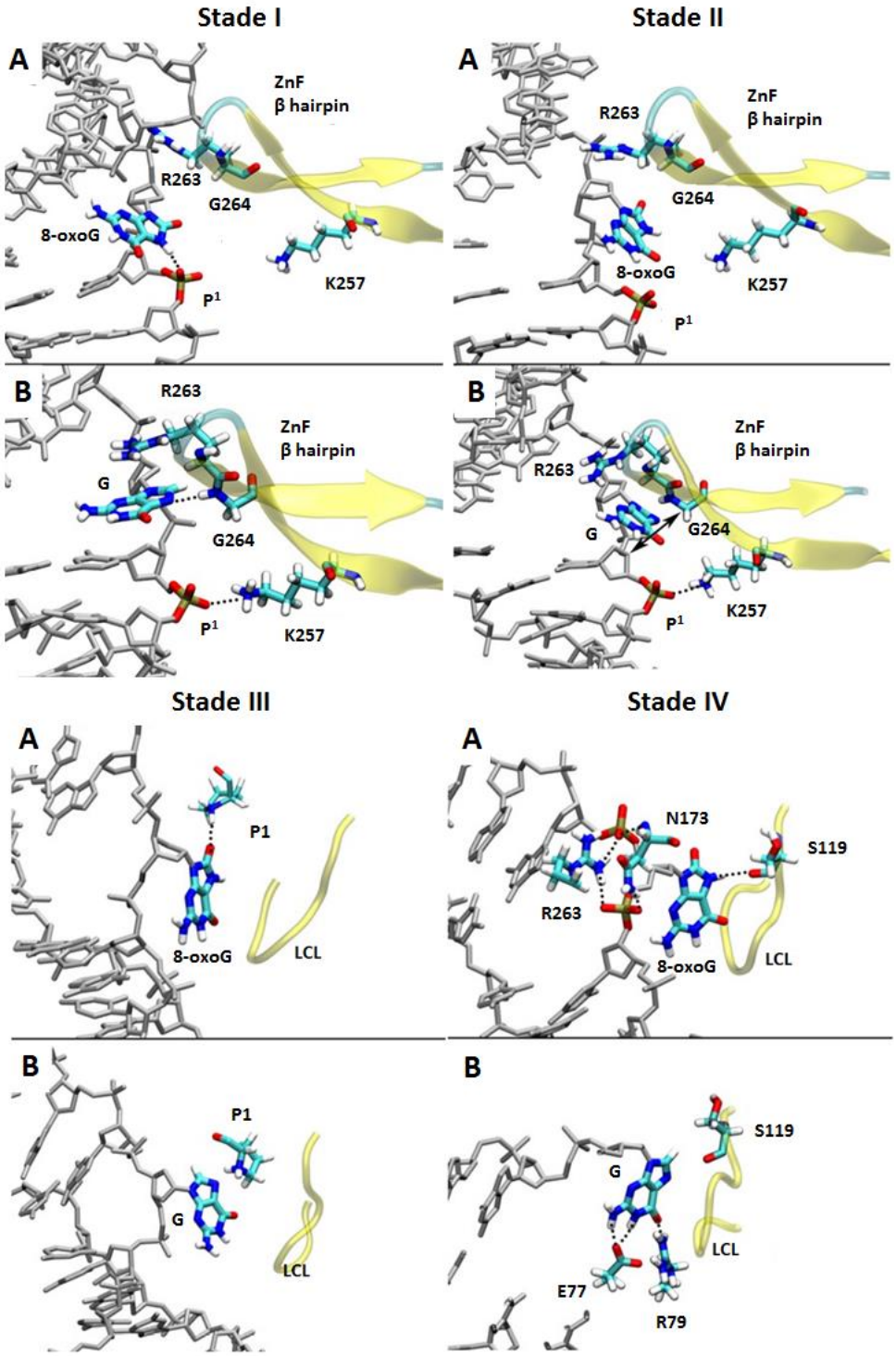


Figure 38 : Étude de la différenciation d'une 8-oxoG d'une G classique par *Bst*Fpg

Cette figure représente les quatre étapes du mouvement de « base extrusion » **A**) la 8-oxoG et **B**) d'une G classique. Chaque étape correspond à une valeur d'angle de retournement de la base : I ≈ 20°, II ≈ 75°, III ≈ 200° et IV ≈ 275° [209].

I.3.4.4 Stabilisation du substrat

Une autre étude de la boucle LCL de *Bst*- et *Ll*Fpg par simulation de dynamique moléculaire (1,5 à 3 ns) a permis de mettre en évidence différents états de cette dernière dans plusieurs contextes [141]. Plusieurs modèles différents ont été créés à partir des structures cristallographiques de *Gst*Fpg et *Ll*Fpg : en présence de 8-oxoG dans le site actif ou d'un site abasique, si la 8-oxoG est présente dans les deux orientations *anti* et *syn*, ainsi qu'en modifiant de la position dans l'espace de la chaine latérale de $Y222_{Ll}$ à l'intérieur (« in ») ou à l'extérieur (« out ») du site actif. Une étude plus globale des simulations montre que les domaines N- et C-terminaux ne semblent pas se comporter de la même façon en présence de 8-oxoG (mouvements anti-corrélés entre les deux domaines) et présence de site abasique (mouvements corrélés entre les deux domaines). L'orientation de la 8-oxoG dans le site actif ainsi que la position « in » et « out » d'$Y222_{Ll}$ semblent avoir un impact sur les interactions formées entre la protéine (notamment au niveau de la boucle LCL) et cette lésion. En se concentrant d'avantage sur la boucle LCL des différents systèmes, un pont salin entre $T224_{Bst}$ ou $T221_{Ll}$ (aa de la boucle LCL) et $E78_{Bst}$ ou $E76_{Ll}$ (aa d'une des boucles intercalantes) ainsi que la position « in » d'$Y222_{Ll}$ semblent permettre la stabilisation la boucle LCL dans un état fermé, formant deux réseaux de liaisons hydrogène différents autour de la 8-oxoG dans les deux orientations *syn* et *anti*. Cependant, les simulations ont été effectuées avec un paramètre de température de 300°K (~27°C) correspondant à la température de développement des organismes mésophiles, alors que *Bst* est un organisme thermophile. Par conséquent, cette analyse va de paire avec l'étude présentée ci-dessous.

En effet les conditions idéales pour la croissance de *Bst* nécessitent une température avoisinant les 65°C ce qui est différent des organismes mésophiles tel que *Ll* (30°C). Cet organisme vivant dans des conditions de températures différentes, son métabolisme et donc ses protéines y sont adaptées, elles sont thermostables. Une autre étude de la boucle LCL de Fpg des organismes mésophiles et thermophiles par des simulations de dynamique moléculaire a été menée dans le but de vérifier si les structures des Fpg thermophiles obtenus à température ambiante sont représentatives des structures des protéines aux températures ou se développent ces organismes extrêmophiles [210]. Plusieurs modèles de *Bst*Fpg complexée à un ADN DB comportant une 8-oxoG ou une G classique extraite et présentée au site actif ont été conçus. À partir de ces systèmes, plusieurs simulations de dynamique moléculaire ont été réalisées à des températures différentes (298 et 323°K). Les simulations à différentes températures permettent de mettre en évidence des changements globaux dans la

structure de la protéine et met en lumière le fait que la structure cristallographique de *Bst*Fpg WT obtenue à température ambiante ne suffit pas à représenter convenablement l'enzyme qui est très dynamique. En parallèle, les simulations de Fpg WT et E2Q permettent aussi de comprendre pourquoi ce mutant est inactif. En effet la protéine mutée ne forme plus le même réseau de liaisons hydrogène que dans les structures de la protéine WT, et notamment avec le O8 manquant. Dans la structure WT, l'O8 interagit avec E2 *via* le -NH du squelette et dans la structure E2Q, cet O8 interagit avec le -NH2 de la chaine latérale du Q2. En parallèle, les auteurs mettent en évidence des différences dans les interactions se créant entre la base présente dans le site actif et la boucle LCL, et notamment le contact entre le groupement carboxyle du squelette de S119 de la boucle LCL et le N7 de la 8-oxoG (tout comme les auteurs de l'étude décrite précédemment), qui n'existe plus lorsqu'elle est remplacée par une guanine classique. De plus, la boucle LCL est plus éloignée de la base azotée et semble avoir un effet répulsif en présence d'une guanine classique.

D'autres études se sont intéressées à la spécificité des protéines Fpg et hNEIL1 pour leurs substrats grâce à des simulations de dynamique moléculaire classique. La première a permis d'identifier des aa clés tel que K217 et M73 permettant tout deux la discrimination entre les substrats de *Ec*Fpg *via* la reconnaissance du cycle imidazole qu'il soit ouvert ou fermé et/ou du O8 de la base (si présent). Pour rappel, parmi les éléments étudiés, les bases excisées sont 8-oxoG, FaPyG, FaPyA, et les bases non excisées sont 8-oxoA, G classique). D'autres aa sont également identifiés comme stabilisateurs du complexe protéine substrats dans tous les contextes, comme les aa E2, E5 et/ou T214 et/ou N221. Cependant, les auteurs se sont basés sur l'unique structure d'*Ec*Fpg, dont le motif correspondant à la boucle LCL n'a pas été résolu. La boucle a donc été reconstruite aléatoirement et les interactions que les bases auraient pu se former avec cette dernière n'ont peut-être pas pu se créer dans le court temps de simulation (2 ns). L'orientation *syn* et *anti* du dommage en position extra-hélicale dans le site actif aurait également un impact sur la reconnaissance des dommages, par exemple les molécules FaPyG et FaPyA interagissent avec K217 et/ou M73 seulement lorsque leur orientation est *anti* tandis que la 8-oxoG est reconnue par ces aa dans les deux orientations *syn* et *anti*. La G classique et 8-oxoA ne forme aucune interaction avec K217 et M73, les auteurs concluent que c'est pour cette raison que ces bases ne sont pas excisées [211]. Ces données sont en bon accord avec les tests de mesure de l'activité glycosylase des mutants des aa H89 et K217de *Ec*Fpg [212]. Ces mutants ont permis d'identifier $R109_{Ec}$ comme un aa important pour les interactions et la stabilisation de la protéine sur l'ADN, et $R108_{Ec}$ comme un aa reconnaissant la base opposée au dommage. Ces Arg sont très conservées ou sont quelques fois remplacées par des Lys parmi les Fpg, aa également porteur de charge positive, une propriété compatible pour interagir avec les charges négatives du squelette d'une l'ADN. D'après les auteurs, les aa $H89_{Ec}$ et $R109_{Ec}$ interviendrait dans le mouvement de « base

extrusion » de la base en poussant cette dernière dans le site actif, ce qui a été confirmé par d'autres études présentées un peu plus tôt [212]. Une étude similaire menée sur hNEIL1 a permis d'identifier les aa P1, E2, E5, M80 et Y176/Y262 permettant la stabilisation les lésions Sp, Tg (excisées par hNEIL1) et 8-oxoG (non excisée par hNEIL1) dans le site actif de l'enzyme. Les aa E2 et E5 permettraient la stabilisation des groupements amines endo-cycliques des groupements pyrimidines tandis que le couple M80, Y176/Y262 stabiliserait les groupes exo-cycliques. On reconnait parmi ces aa la M80 qui correspond à M73 de *Bst*Fpg et M75 chez *Ll*Fpg, également faisant partis de la triade d'intercalation dans ces trois enzymes. En revanche, il ne figure pas d'aa similaire à K217 de *Bst*Fpg dans les résultats, mais les Y176/Y262 ce qui pourrait peut-être expliquer la différence dans les substrats reconnus par les deux ADN glycosylases. Une interface portant des charges positives (8 Arg) est également identifiée chez hNEIL1. Parmi les résidus basiques constituant cette interface, on retrouve R117 et R118 correspondant à R108 et R109 chez *Ec*Fpg et *Ll*Fpg. De plus, les auteurs montrent que la 8-oxoG est plus exposée au solvant que les autres lésions lorsqu'elles sont en position extra-hélicale et dans le site actif, et assimilent ce phénomène à la mauvaise capacité d'excision de la 8-oxoG par hNEIL1 [213]. Tous ces travaux ont permis de comprendre le rôle de chacun de ces éléments qui sont résumés dans le tableau ci-dessous (**Tableau 8**).

Tableau 8 : Aa clés dans la reconnaissance de l'ADN, la reconnaissance des lésions, la spécificité d'excision et la catalyse des protéines Fpg/NEIL1
Un alignement des structures cristallographiques a été réalisé avec Multiseq [214] de VMD [215], la boucle LCL ne fait pas partie du tableau car elle est partiellement non résolue dans toutes les structures et est donc remplacée par des « ? ». Le symbole « Ø » correspond à un gap dans l'alignement structural.

	Reconnaissance ADN [212, 213]	**Reconnaissance lésion et « base extrusion »** [203, 206, 216]	**Intercalation** [203, 206, 216]	**Spécificité au dommage** [211, 212]	**Catalyse**
*Bst*Fpg	Ø, Ø, H74, H92, K112, Ø, R156, Ø, Ø	M77, R111	M77, R111, F113	M77, I174, ?, Y242	P1, E2, E5
*Ll*Fpg	R31, H72, K90, H91, K110, Ø, K154, Ø, Ø	M75, R109,	M75, R109, F111	M75, I172, ?, Y238	
*Ec*Fpg	R31, H70, K88, H89, R109, Ø, Ø, Ø, Ø	M73, R108	M73, R108, F110	M73, I169, K217, Y236	
*Ath*Fpg	K32, K105, Ø, R126, Ø, Ø, Ø, Ø	M77, R125	M77, R125, F127	M77, Ø, ?, Ø	
hNEIL1	R33, R77, R94, H95, R118, R132, R158, R256, R273, R276	M80, R117	M80, R117, F119	M80, Y176, ?, Y262	

Abréviations : ● **Organismes :** *Bst*, *G. strearothermophilus ; Ll*, *L. lactis* ; *Ec*, *E. coli* ; *Ath*, *A. thaliana* ; h, human.

I.3.5 Mécanismes réactionnels des protéines Fpg/Nei

Les ADN glycosylases Fpg/Nei sont toutes bifonctionnelles, et sont donc capables de réaliser deux activités, la première est la glycosylase et la seconde l'AP lyase. L'activité glycosylase pourrait se produire selon deux modèles étudiés en mécanique quantique, et produit l'intermédiaire instable base de Schiff (**Figure 39 A**) [202]. L'intermédiaire base de Schiff (une liaison covalente entre la P1 et le C1' du sucre du nucléoside endommagé) est ensuite résolue par β,δ-élimination lors de l'activité AP lyase (**Figure 39**) [217]. Cependant, les mutants P1G ont montré que la formation de base de Schiff était toujours possible, mais l'étape β,δ-élimination est 700 à 2000 fois moins efficace que celle médiée par les enzymes WT. Cette observation souligne l'importance de la présence d'une Pro à cette position [188]. Dans la **Figure 39**, nous remarquons que deux mécanismes sont possibles pour la formation de l'intermédiaire base de Schiff. Le premier est le mécanisme nucléobase protonée et fait intervenir l'aa

E5. Le second est le mécanisme ribose protoné et fait intervenir l'aa E2 et une molécule d'eau. Dans les deux cas, l'intervention de la P1 est nécessaire. Cependant le premier semble moins favorable énergétiquement que le second (**Figure 39 B**) [202].

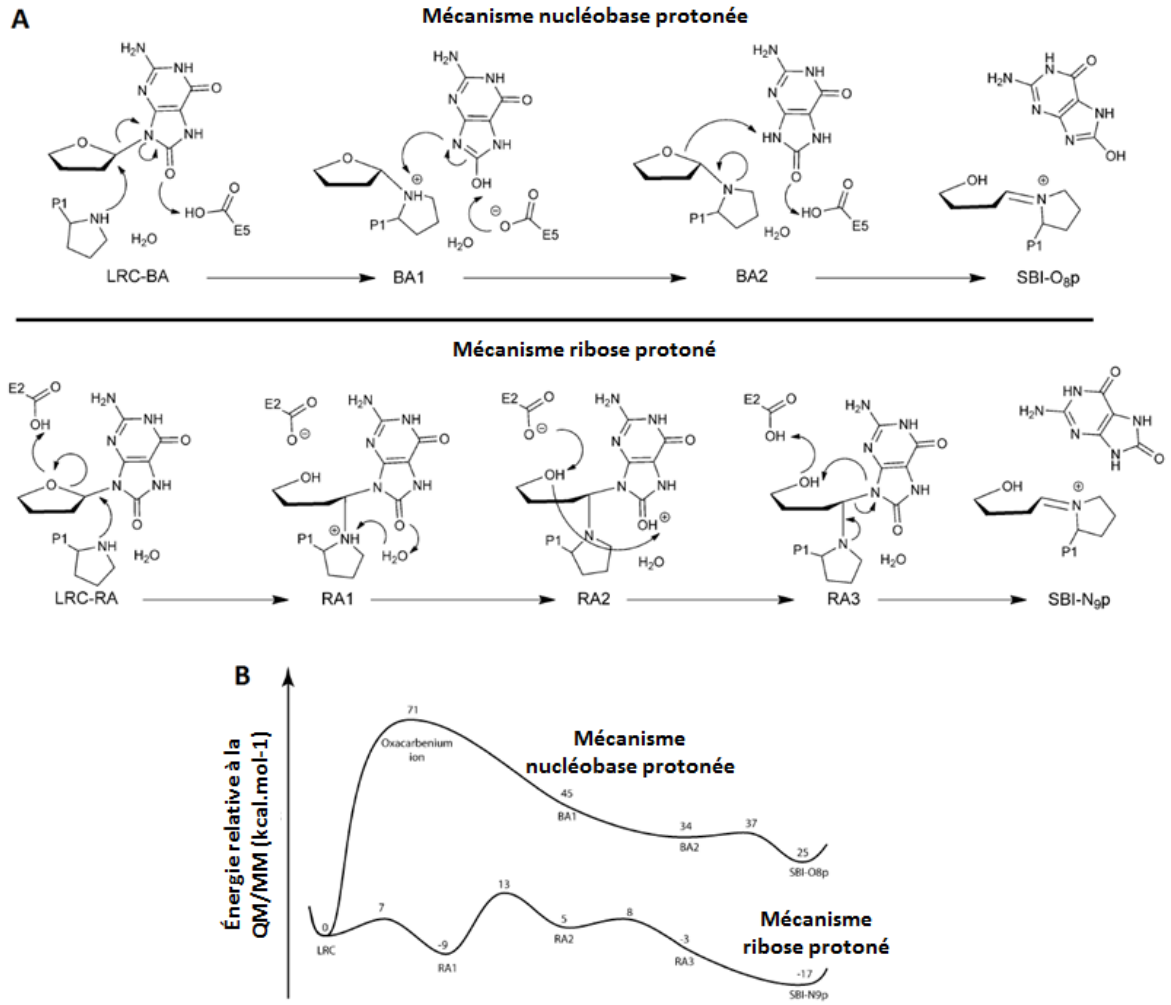


Figure 39 : Mécanismes réactionnels possibles lors de l'excision d'une 8-oxoG par *Bst*Fpg
A) Représentations schématiques des deux mécanismes et **B**) Énergies relatives à ces deux mécanismes obtenus en QM/MM, l'état de transition du pathway protonation de base semble plus élevé à franchir que l'état de transition du pathway protonation du sucre [202].
Abréviations : BA, « Base Activation » ; RA, « Ribose Activation ».

I.3.6 Inhibiteurs des enzymes Fpg/Nei

La recherche et la conception de petites molécules naturelles ou synthétiques capables de moduler (inhiber ou activer) l'activité des enzymes Fpg/Nei est un champ d'investigation très récent. La recherche d'inhibiteurs est motivée par l'observation que dans certaines situations pathologiques (cancers, maladies neurodégénératives) ces enzymes sont des cibles pharmacologiques pertinentes [147, 218, 219]. Le rationnel de la conception de ces molécules devait permettre l'inhibition compétitive des cibles. Dans ce contexte, le ligand est similaire au substrat et occupe le site actif de

l'enzyme, bloquant ainsi les activités catalytiques. C'est ainsi qu'après plusieurs criblages *in vitro* de librairies de molécules d'analogues de purines ont permis l'identification des molécules inhibitrices telles que la 2-thioxanthine (2TX) et les molécules P_n (molécules décrites dans la section **I.2.4.1** p. **77** de l'Introduction) (**Figure 40**) [153]. En collaboration avec l'équipe du Professeur Luigi Agrofoglio de l'ICOA, des dérivés synthétiques de 2TX ont été synthétisés ($2TX_n$) et certains d'entre eux se sont révélés plus efficace que 2TX (**Tableau 9**). Les IC_{50} (concentrations nécessairent pour observer une inhibition de la cible à hauteur de 50%) de ces molécules restent pour l'instant modestes (de l'ordre du µM). Ces molécules présentent une forte réactivité à cause de leur(s) atome(s) de soufre(s). Quant aux molécules P_n identifiées par Jacobs *et al.*, il semblerait qu'elles soient peu spécifiques de l'enzyme hNEIL1 et présenteraient une forte toxicité pour des cellules en culture. Par ailleurs, nous n'avons pas réussi à reproduire les mesures d'IC_{50} au laboratoire. De plus, les molécules $2TX_n$ semblent beaucoup moins efficaces sur l'homologue humain de Fpg, hNEIL1 (**Tableau 9**).

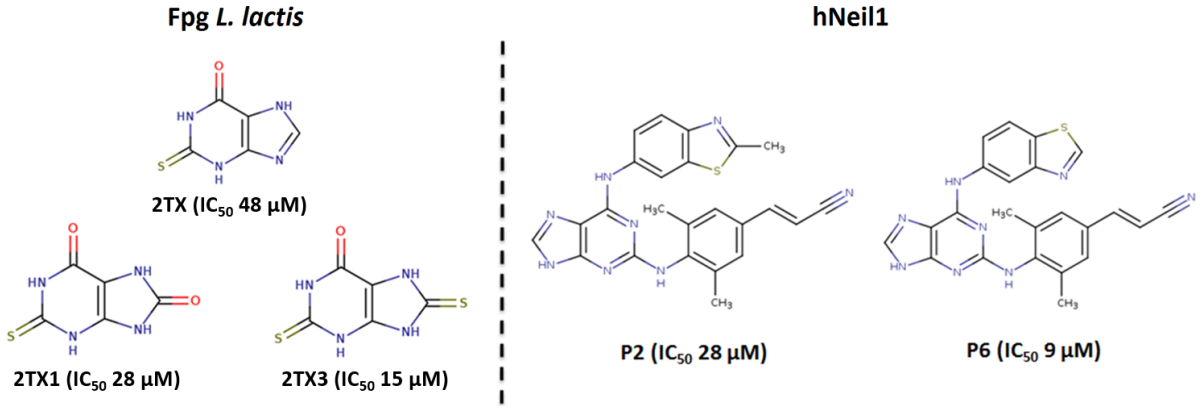


Figure 40 : Structures de quelques $2TX_n$ et P_n, analogues de substrat et présentant des effets d'inhibition sur les protéines *Ll*Fpg/hNEIL1
Les IC_{50} indiqués correspondent aux essais d'inhibition réalisés sur *Ll*Fpg à gauche, et sur hNEIL1 à droite [153].

Tableau 9 : Quelques IC_{50} (µM) traduisant des effets d'inhibitions des $2TX_n$ sur les protéines Fpg/Nei

	*Ll*Fpg	hNEIL1	MvNei1
2TX	48 ± 4	> 500	> 500
2TX1	28 ± 7	> 500	> 500
2TX2	> 500	> 500	> 500
2TX3	15 ± 1	19 ± 1	124 ± 4
2TX7	41 ± 7	21 ± 1	14 ± 1
F3CS	8 ± 1	> 250	> 500

Des études de cristallographiques et biochimiques ont montré que les molécules $2TX_n$ ont la capacité de déstructurer le ZnF de *Ll*Fpg, *Ec*Nei et hNEIL2, annihilant leur capacité à se fixer sur l'ADN. Les expériences de biochimie ont révélé que les molécules $2TX_n$ induisaient la perte du zinc des enzymes Fpg/Nei comportant un ZnF tandis que les structures obtenues en cristallographie

permettaient de visualiser la formation de pont(s) disulfure(s) entre le soufre des $2TX_n$ et les cystéines C248 et/ou C268, et entre les deux dernières cystéines C245 et C265 de *Ll*Fpg (**Figure 41**). Sachant que hNEIL1 ne possède pas de motif doigt à zinc mais un motif équivalent ZnLF sans Cys, cela peut expliquer pourquoi les $2TX_n$ sont de moins bons inhibiteurs (à quelques exceptions près) de la protéine humaine hNEIL1 [196]. Cependant, un effet d'inhibition résiduel persiste sur l'activité de hNEIL1 (qui possède 6 autres Cys libres) et de MvNei1 (qui ne possède aucune Cys). Cela qui signifie que les molécules $2TX_n$ inhibant les enzymes Fpg/Nei comportant un ZnLF (hNEIL1 et MvNei1) interagissent avec ces protéines d'une autre manière, et entravent l'activité glycosylase des enzymes Fpg/Nei.Cet effet inhibiteur semble s'exercer directement sur les enzymes car les molécules $2TX_n$ n'ont pas d'effet apparent sur les substrats respectifs de ces enzymes car nous n'observons pas de changement de la température de fusion des fragments d'ADN synthétiques utilisés comme substrats. Les modes d'interaction et d'inhibition des $2TX_n$ des enzymes hNEIL1 et MvNei1 sont donc encore inconnus, c'est pourquoi j'ai utilisé les molécules $2TX_n$ présentées succinctement dans la **Figure 40** pour une étude d'amarrage molécule sans *a priori*.

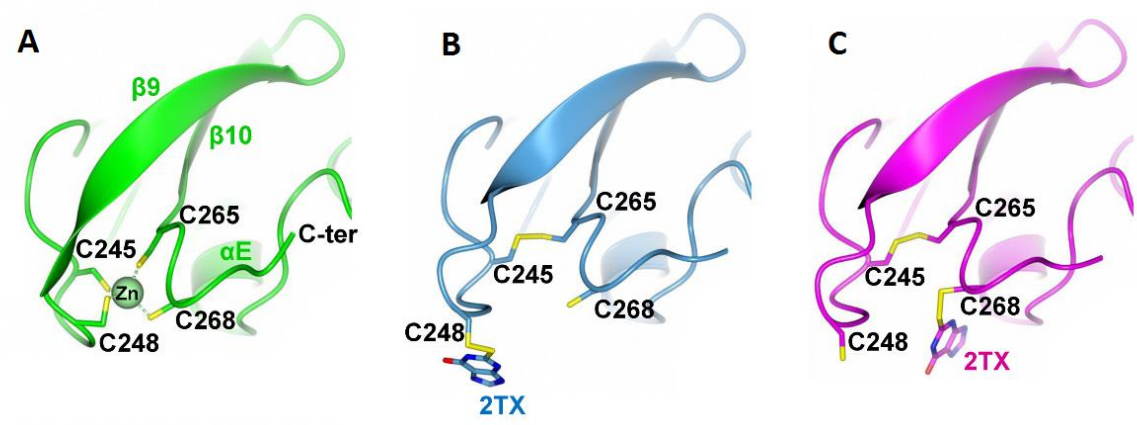


Figure 41 : Déstructuration du doigt à zinc de *Ll*Fpg par la 2TX
A) ZnF normal, l'atome de zinc est coordonné par les quatre cystéines C245, C248, C265 et C268. **B**) Formation d'un pont disulfure entre le soufre en C8 de la 2TX et la C248, **C**) ou avec la C268. Lorsque le ZnF est déstructuré, les C245 et C265 forment un pont disulfure [196]. Ces images sont des représentations du ZnF de la structure *Ll*Fpg de la structure 4PDG [196].

Un second site d'amarrage non covalent a également été observé dans les structures cristallographiques (**Figure 42**). Ce site est localisé à l'interface entre le motif H2TH et l'ADN du complexe. Les résidus entrant directement en contact avec les conformations des ligands dans ce site sont K57 (du domaine N-terminal), L161, E162, Q163 du motif (H2TH) et R260 (du ZnF). Ces données n'ont pas encore fait l'objet d'une publication.

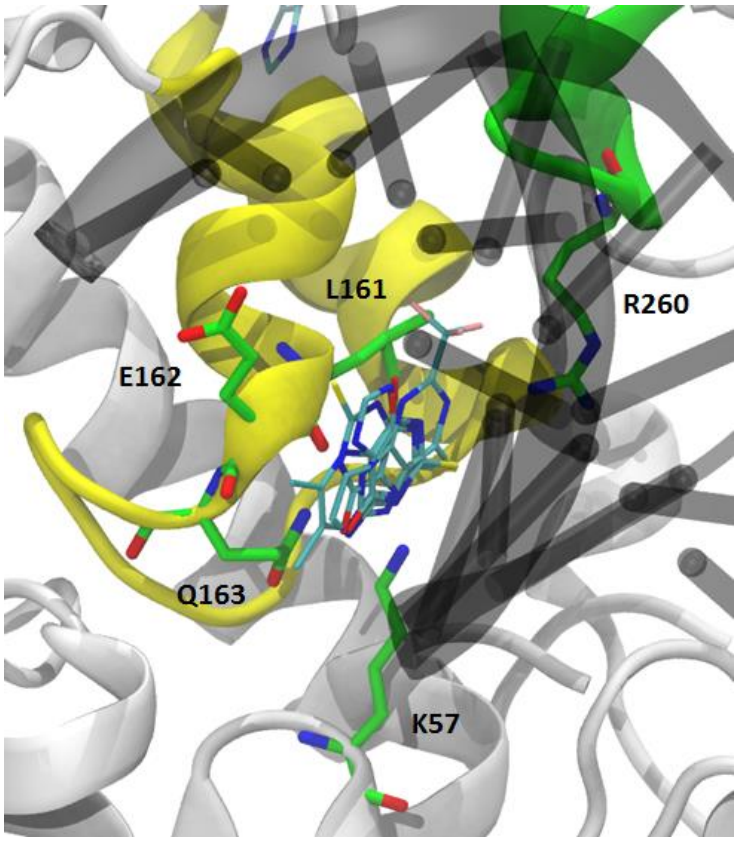


Figure 42 : Site d'interaction des 2TX, 2TX1, 2TX2 et F3CS sur *Ll*Fpg observé sur les structures cristallographiques
Le motif H2TH est représenté en cartoon jaune, le motif ZnF en vert, et le reste de la protéine est en blanc. L'ADN est représenté en cartoon noir transparent. Les aa interagissant avec les ligands sont en bâtonnets verts et les ligands sont en bâtonnets fins cyan.

Concernant *Ll*Fpg libre, hNEIL1 et MvNei1 libres et complexées à l'ADN, nous n'avons pas de données structurales expérimentales nous permettant d'appréhender les modes d'interaction entre les $2TX_n$ et ces cibles.

Tableau 10 : Mutants et molécules analogues aux substrats utilisés pour favoriser l'obtention de structures 3D des complexes ADN glycosylases Fpg/Nei avec leurs substrats (bases altérées ou site abasique)

Dans la colonne mutation, la légende (cadres et couleurs) est la même que dans la **Figure 31**, à savoir encadrés en noir les aa catalytiques, encadrés en jaune le motif H2TH et en vert le motif doigt à zinc, surlignés en rose les aa s'intercalant dans l'ADN, surlignés en mauve les aa contactant les bases en position extra-hélicale dans le site actif et en noir les cystéines coordonnant le Zn^{2+}. Cette liste ne recense pas toutes les structures des Fpg/Nei, cette recherche a été effectuée dans la Protein Data Bank (PDB) [220].

Prot.	Organisme	Mutation	Substrat	Complexe	PDB
Fpg	*T. thermophilus*	WT			1EE8 [186]
	G. stearothermo-philus	WT	G	IC	2F5O [205]
		WT	8-oxoG	SBI	1L1Z [206]
		WT	Produit fini		1L2B [206]
		E2Q	8-oxoG	LRC	3GPY, 3GQ4 [203]
		E2Q	8-oxoG	LRC	1R2Y [183]
		E2Q	DHU	LRC	1R2Z [183]
		E2Q, M76A	8-oxoG	EC	4G4O [216]
		E2Q, F113A	8-oxoG	EC	4G4R [216]
		M76A	ADN DB classique	IC	4G4N [216]
		F113A	ADN DB classique	IC	AG4Q [216]
		Δ219-234	G	IC	3GPX [203]
		Δ219-234	8-oxoG	EC	3GO8, 3GPU, 3GQ3 [203]
		V221P	8-oxoG	EC	3GP1 [203]
		T223P	G	IC	3GQ5 [203]
		T223P	8-oxoG	EC	3GPP [203]
	L. lactis	WT	cFaPyG	RC	1XC8 [184]
		WT	cbzFaPyG	RC	4PDI [196]
		WT	THF	RC	1PM5 [195]
		WT	THF*	RC	4PDG [196]

		WT	c8-oxoG	RC	4CIS [202]
		P1G	PDI	RC	1KFV [188]
		P1G	PDI	RC	1NNJ [195]
		ΔP1	cFaPyG	RC	1TDZ [201]
		W178A			Non publiée
		R246G	THF	RC	4PCZ [196]
		C248H	THF	RC	4PD2 [196]
		R252A			Non publiée
	E. coli	WT	8-oxoG	SBI	1K82 [162]
	A. thaliana	C-ter Δ88			3TWL [185]
		C-ter Δ88	THF	RC	3TWM [185]
		C-ter Δ109			3TWK [185]
Nei	*E. coli*	WT	8-oxoG	SBI	1K3X, 1K3W [189]
		E2A			1Q3C [161]
		E2Q	PED	RC	2OQ4
		R252A			1Q3B [161]
		R252A	PED	RC	2OPF
NEIL1	*Acanthamoeba polyphaga* mimivirus	E2Q	OHU	LRC	3VK7 [182]
		E2Q	TG	RC	3VK8 [182]
		G86D	THF	RC	4NRW [221]
	H. sapiens	WT			1TDH [187], 4NRV [221]
		WT	THF	RC	5ITU [190]
		P1G	TG	LRC	5ITX, 5ITY [190]
NEIL 2/3	*Acanthamoeba polyphaga* mimivirus	WT			4MB7 [199]
NEIL3	*M. Musculus*	WT			3W0F [191]

Abréviations : ● **Protéines :** Fpg, Formamidopyrimidine DNA glycosylase ; Nei : Endonucléase VIII ; NEIL1, Endonucléase VIII like 1.

● **Mutations :** WT, Wild Type ; Δ, délétion ; C-ter, C-terminale.
● **Substrats :** 8-oxoG, 8-oxoguanine ; c8-oxoG, carba-8-oxoguanine ; DHU, dihydrouracile ; cFaPyG, carba-FaPyG ; cbzFaPyG, carba-N7-benzyl-FaPyG ; THF, Tetrahydrofuranose ; OHU, 5-hydroxyuracile PDI, phosphoric acid mono-(3-hydroxy-propyl) ester ; PED, pentane-3,4-diol-5-phosphate.
● **Complexes :** SBI, Schiff Base Intermediate ; LRC, Lesion Recognition Complex ; EC, Encounter Complex ; IC, Interrogation Complex ; RC, Recognition Complex.
*Complexe co-cristallisé avec la 2-thioxantine, liée covalemment aux cystéines C248 et C268 du doigt à zinc.

II. **O**bjectifs et plan de la thèse

Les objectifs principaux de ce travail de thèse sont d'apporter de nouveaux éléments de compréhension du potentiel rôle de la boucle LCL dans l'activité glycosylase de *Ll*Fpg et la recherche et conception de nouveaux inhibiteurs sélectifs des ADN glycosylases *Ll*Fpg et hNEIL1. Ces deux questions sont complémentaires car il est nécessaire de bien comprendre la structure et la dynamique des cibles pour mieux identifier les petites molécules capables d'inhiber ces dernières. Pour apporter des solutions à ces problèmes, nous avons combiné trois domaines de compétences : (i) la modélisation moléculaire, (ii) la chimie de synthèse et (iii) la biochimie des ADN glycosylases. Mes travaux de thèses sont exclusivement réalisés grâce à des méthodes de simulations et de calculs *in silico* qui sont des techniques de modélisation moléculaire. L'aspect chimie de synthèse est apporté par l'équipe de notre collaborateur le Professeur Luigi Agrofoglio de l'Institut de Chimie Analytique et Organique (ICOA), dont le rôle est de synthétiser de nouvelles molécules qui seront ensuite testées au CBM dans notre équipe de recherche qui maitrise la biochimie des ADN glycosylases.

L'étude de la boucle LCL de *Ll*Fpg par des simulations de dynamique moléculaire peut apporter de nouveaux éléments de compréhension du rôle de cet élément dans l'activité glycosylase de l'enzyme. Des simulations de dynamique moléculaire ciblée peuvent également permettre de comprendre le rôle de la boucle LCL de *Ll*Fpg lors de la sortie du produit d'excision. Pour remplir le second objectif, nous devons comprendre quels sont les potentiels modes d'interaction entre de petites molécules et les enzymes *Ll*Fpg et hNEIL1. Pour identifier des sites de fixation des petites molécules sur les cibles, nous avons mis au point une méthode d'amarrage moléculaire flexible et aveugle. Les tests des logiciels de docking ont été effectués dans le laboratoire Matrice Extracellulaire et Dynamique Cellulaire (MEDyC) de Reims. Par la suite, nous avons réalisé un criblage virtuel haut débit de la librairie de molécules Ambinter dans le site actif de hNEIL1 en collaboration avec l'équipe du Professeur Pascal Bonnet de l'ICOA.

La **Figure 43** décrit le plan de mes travaux de thèse et les chapitres de la partie Résultats commençant à partir de la p. **173**. Les **chapitres 1** et **2** correspondent à l'étude structurale et dynamique de la boucle LCL à partir des simulations de dynamique moléculaire classique et ciblée respectivement. Les **chapitres 3** et **4** apportent de nouveaux éléments dans le projet de recherche d'inhibiteurs sélectifs de hNEIL1.

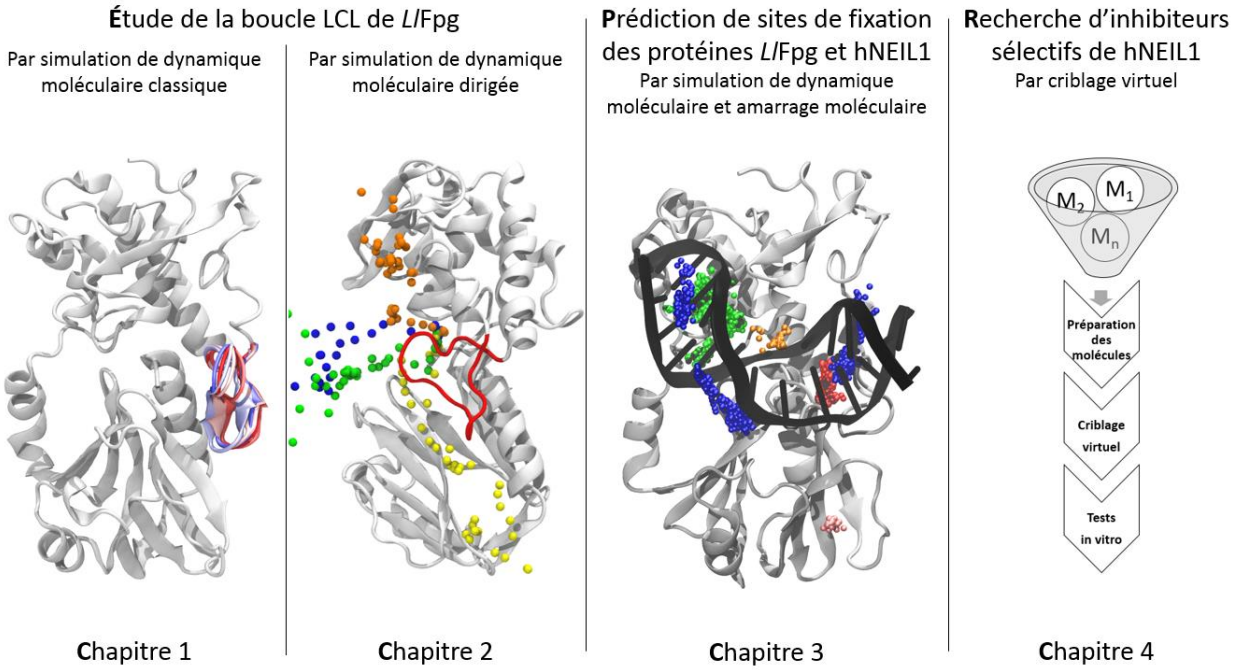


Figure 43 : Plan de la thèse
La partie Résultats sera composée de quatre chapitres décrits ci-dessous.

Dans le **chapitre 1**, nous décrirons deux états de la boucle LCL, (i) fermé et (ii) relâché. Puis, nous présenterons le comportement de la boucle LCL modélisé dans plusieurs contextes : (i) protéine libre, (ii) protéine avec un substrat et (iii) protéine avec un produit.

Dans le **chapitre 2**, nous présenterons les modèles de sortie d'une 8-oxoG libre du site actif de *Ll*Fpg dans deux contextes différents où la boucle LCL est dans l'état fermé et relâché.

Le **chapitre 3** décrira la mise au point et le test de la méthode de docking flexible et aveugle. Ce chapitre présentera les simulations de dynamique moléculaire ainsi que la méthode de classification ascendante hiérarchique utilisées pour intégrer la notion de flexibilité du système au docking. Nous décrirons ensuite le test du logiciel de docking utilisé en docking classique (zone délimitée), et en docking aveugle (sans *a priori*). Nous présenterons nos résultats que nous comparerons avec les résultats obtenus par le logiciel « Automated Molecular Inverse Docking Engine » (AMIDE) pour valider notre méthode. Par la suite, nous décrirons deux principaux sites de fixation des ligands prédits sur les protéines *Ll*Fpg et hNEIL1.

Dans le **chapitre 4**, nous présenterons les résultats préliminaires du criblage virtuel réalisé dans le site actif prédit de hNEIL1. À la suite, nous décrirons les tests *in vitro* des molécules identifiées.

III. **M**éthodologie

Chapitre 1 : La dynamique moléculaire

Chapitre 2 : L’amarrage moléculaire

III.1 **Chapitre 1 : L**a dynamique moléculaire

III.1.1 Définition

La modélisation moléculaire a pour but de prévoir la structure, la dynamique et la réactivité des molécules. Elle comporte différentes méthodes comme les méthodes quantiques (MQ), la mécanique moléculaire (MM), la dynamique moléculaire (DM). Les méthodes quantiques (« Density Functional Theory », DFT par exemple) permettent de modéliser la réactivité chimique des molécules. Les calculs se basent sur les orbitales électroniques des atomes. La complexité des calculs augmente avec le nombre d'électrons présents, ce qui permet de travailler sur des molécules à petits nombres d'atomes (ligand par exemple). La MM et la DM simplifient le modèle en réduisant les atomes à des points auxquels sont associés une masse et une charge, ne prenant pas explicitement en compte les électrons et réduisant ainsi drastiquement les coûts computationnels. Ceci permet de travailler avec des molécules plus grosses (protéines, acides nucléiques par exemple). La MM permet de calculer l'énergie d'une conformation d'une protéine, elle est en quelque sorte contenue dans la DM qui permet également de simuler le déplacement des atomes au cours du temps.

La DM est une méthode utilisée pour générer et étudier un modèle du déplacement des atomes au sein de molécules plus ou moins grosses, les mouvements globaux au sein d'une protéine et le repliement selon la durée de la simulation (**Figure 44**) [222]. Elle permet d'étudier des systèmes de plusieurs milliers d'atomes. La DM classique modélise tous les atomes par des points, elle permet de produire des simulations sur des échelles de temps allant jusqu'à la nanoseconde (10^{-9} s). La DM gros grain (« coarse-grained ») simplifie les aa et nucléotides en quelques sphères (1 à 4 sphères pour les plus gros résidus) aux propriétés bien spécifiques, ce qui permet d'explorer des échelles de temps encore plus grandes allant jusqu'à la microseconde (10^{-6} s) selon les moyens computationnels disponibles. Il existe à ce jour plusieurs dizaines de logiciels implémentant les fonctions nécessaires aux simulations de DM tels que CHARMM (M. Karplus, prix Nobel 2013) [207], GROMACS (E. Lindahl) [223] et AMBER (P. Kollmann) [224].

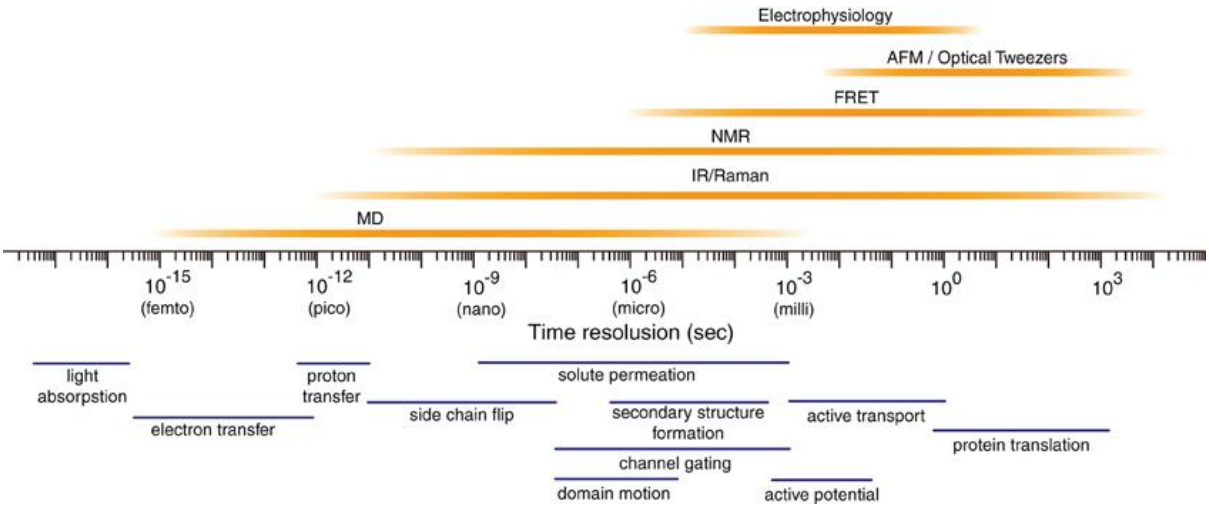


Figure 44 : Phénomènes biologiques et techniques de modélisation et expérimentales
Phénomènes observables (en haut) versus les techniques de visualisation/modélisation (en bas) en fonction de l'échelle de temps. Les moyens de calcul permettent d'effectuer une DM (Molecular Dynamics, MD), sur des échelles de temps allant de la femtoseconde à la milliseconde. Les phénomènes de déplacement de chaines latérales, de mouvement de domaine, de transport d'ions et de petites molécules *via* des canaux et la formation de structure secondaire peuvent ainsi être éventuellement modélisés [222].

III.1.2 Logiciel utilisé

Lors de mes travaux de thèse, j'ai réalisé des simulations de dynamique moléculaire grâce au logiciel AMBER [225], développé par Peter Kollman de l'université de Californie à San Fransisco dans le but d'étudier les biomolécules. La première version de l'outil est publiée en 2002 et est toujours mise à jour tous les deux ans par une collaboration très active entre plusieurs chercheurs de plusieurs universités à travers le monde, tels que David Case (université de Rutgers), Tom Cheatham (université de l'Utah), Carlos Simmerling (université de Stony Brook) entre autres. Les champs de force utilisés dans AMBER font partie d'une famille portant le même nom, pour « Assisted Model Building with Energy Refinement ». Le package AMBER met à disposition un panel d'outils permettant la création et l'optimisation de champs de force (ANTECHAMBER), la préparation des fichiers d'entrée (LEaP), la simulation et la minimisation d'énergie (SANDER, qui est le programme central), l'analyse des trajectoires issues des simulations (ptraj et cpptraj) et bien d'autres. AMBER est écrit dans les langages de programmation Fortran 95, C et C++, a été pensé pour le parallélisme (calculs traités par plus d'un « Central Processing Unit », CPU), et plus récemment adapté pour les « Graphics Processing Unit » (GPU), unités plus performantes pour le calcul numérique. Pour la production des simulations de DM, j'ai utilisé SANDER issu de la version 11 d'AMBER ainsi que le champ de force ff12SB. Pour le traitement des trajectoires, j'ai utilisé cpptraj de la version 14 d'AMBER.

III.1.3 Le champ de force

L'objet étudié est modélisé sous la forme d'un ensemble de points reliés par des ressorts, ce modèle est aussi appelé système. La DM permet de résoudre le problème suivant, qui consiste à déterminer les nouvelles positions $\vec{r_i}$ des atomes à l'instant $t + \Delta t$ (6) :

Équation 6 : Problème résolu par la Dynamique Moléculaire (DM)

$$(6)\ \vec{r_i}(t) \rightarrow \vec{r_i}(t + \Delta t)$$

Pour y arriver, le système est soumis à des forces qui dépendent du logiciel utilisé et qui tentent de reproduire au mieux ce qui se produit à l'échelle de l'atome. Dans le modèle proposé par le logiciel AMBER (**Figure 45** et **Figure 46**), les interactions entre atomes liés (séparés au plus par 3 liaisons) sont représentées par des potentiels modélisant les énergies de liaison, les énergies d'angle de valence et les énergies de torsion. Les interactions entre atomes non liés (séparés au moins de 3 liaisons) sont modélisées par un potentiel correspondant aux énergies de Van der Waals et aux énergies électrostatiques. La somme de ces énergies permet de définir une énergie potentielle comme ce qui suit (7) :

Équation 7 : Calcul de l'énergie potentielle E_p

$$(7)\ E_p = E_{liaisons} + E_{angles} + E_{dièdres} + E_{i<j}$$

Cette énergie dépend donc de plusieurs termes (8, 9, 10, 11) [226] (**Figure 45** et **Figure 46**):

Équation 8 : Calcul de l'énergie de liaison $E_{liaisons}$

$$(8)\ E_{liaisons} = \sum_{liaisons} k_b(b - b_0)^2$$

avec k_b la constante de force,

b la distance entre deux atomes,

b_0 la distance à l'équilibre

Équation 9 : Calcul de l'énergie d'angle E_{angles}

$$(9)\ E_{angles} = \sum_{angles} k_\theta(\theta - \theta_0)^2$$

avec k_θ la constante de force,

θ l'angle de valence entre trois atomes,

θ_0 l'angle à l'équilibre

Équation 10 : Calcul de l'énergie de torsion $E_{torsions}$

$$(10)\ E_{torsions} = \sum_{torsions} \frac{V_n}{2}[1 + \cos(n\varphi\ -\ \gamma)]$$

avec n multiplicité de la rotation, est égale à 1, 2, 3, 4, 6

φ l'angle de torsion entre quatre atomes,

γ la phase de l'angle φ

Équation 11 : Potentiel de Lennard-Jones

$$(11)\ E_{i<j} = \sum_{i<j}\left[\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} + \frac{q_i q_j}{\varepsilon r_{ij}}\right]$$

avec A_{ij} et B_{ij} sont les paramètres de répulsion et dispersion propres au couple d'atomes ij,

r_{ij} la distance entre les deux atomes i et j,

q_i et q_j les charges des atomes i et j,

ε la constante diélectrique du milieu

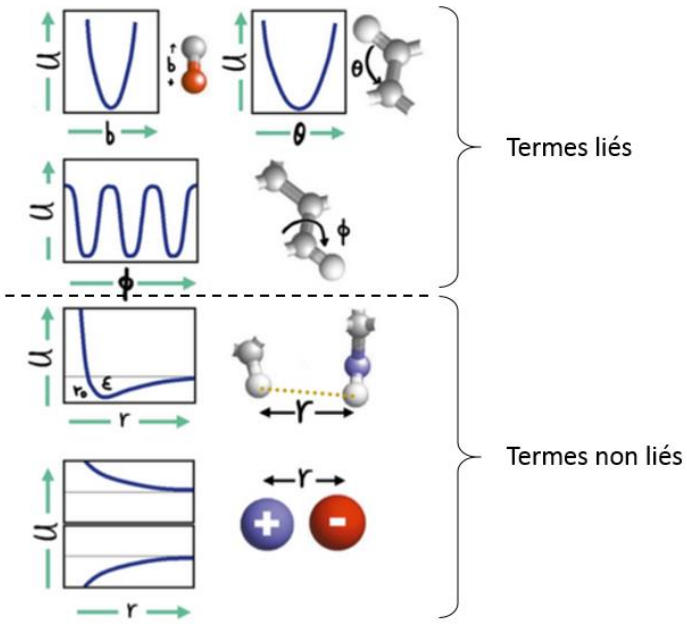


Figure 45 : Potentiels de liaison, d'angle de valence, d'angle dièdre et d'interaction à distance implémentés dans le champ de force d'Amber

Toutes ces énergies sont pondérées par les constantes de force (k_b, k_ϕ), analogues à la constante de raideur d'un ressort, les constantes d'équilibre (b_0, θ_0, γ) et les coefficients d'interaction

(A_{ij}, B_{ijj}, q_i, q_j) qui dépendent du type d'atome, ces paramètres sont rassemblés dans ce que l'on appelle un champ de force (« forcefield » en anglais). Il existe plusieurs types de champ de force, adaptés à divers systèmes (protéines, acides nucléiques). Ils sont déterminés expérimentalement (empiriques) par spectroscopie Infra Rouge (IR), grâce aux structures cristallographiques ou aux structures obtenues par Résonance Magnétique Nucléaire (RMN), et par des calculs de mécanique quantique (semi-empiriques).

Les termes représentant les liaisons et les angles de valence sont des potentiels harmoniques qui contraignent les atomes à rester à une distance ou angle d'équilibre, tel un ressort. Si les atomes s'éloignent de cet équilibre, les potentiels des termes liés forceront ces valeurs à se rapprocher des valeurs d'équilibre b_0, θ_0. Les potentiels de liaison et d'angle de valence sont de type parabolique, tandis que le potentiel de torsion est de type sinusoïdal. Ces potentiels dépendent du type d'atome mis en jeu, de l'organisation de leurs électrons sur leurs orbitales et du type de liaison (double, triple…) qui les séparent, ils permettent de maintenir la géométrie covalente de la molécule.

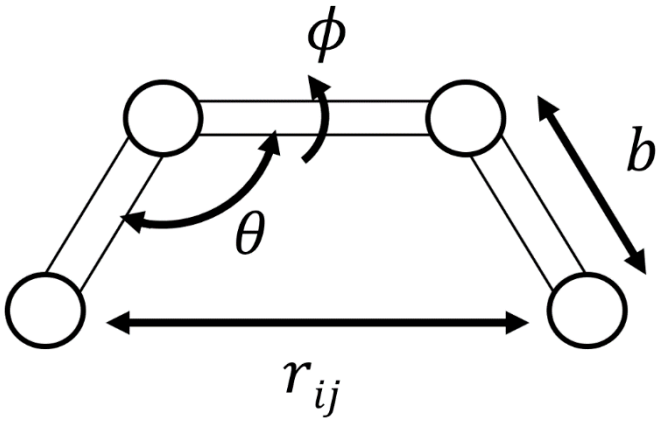


Figure 46 : Modélisation des molécules dans AMBER
Déformations autorisées par la dynamique moléculaire et termes permettant de calculer l'énergie potentielle Ep, b permettant de calculer le potentiel de liaison, θ le potentiel d'angle, ϕ le potentiel de torsion et r_{ij} la distance entre les deux atomes *i* et *j* (**Figure 45**).

Le quatrième potentiel représente les interactions de van der Waals (vdW) et électrostatiques. L'énergie de vdW est calculée grâce à une distance d'équilibre r_{0ij} et d'un puit de profondeur ε, Dans le calcul des énergies de vdW, le terme $\frac{1}{r_{ij}^{12}}$ induit un phénomène de répulsion (principe de Pauli) qui éloigne les atomes les uns des autres lorsqu'ils sont trop proches ce qui empêche les nuages électroniques de deux atomes de se chevaucher. En parallèle, le terme $-\frac{1}{r_{ij}^{6}}$ aussi connu sous le nom de dispersion de London induit un phénomène d'attraction à grande distance qui force les deux atomes à se rapprocher. Le potentiel électrostatique est généré selon la loi de Coulomb, il rapproche les atomes de charges opposées, et repousse les atomes dont les charges ont le même signe. En parallèle, l'énergie électrostatique dépend des charges partielles q des atomes i et j, de la distance

interatomique ainsi que de la constante diélectrique. Ces termes dits « liés » et « non liés » rendent compte de l'énergie nécessaire à la déformation du système.

Les forces dérivent de l'énergie potentielle E_p aussi notée U, on peut alors définir la force $\overrightarrow{F_i}$ s'exerçant sur chacun des atomes du système en utilisant la relation donnée par l'équation (12) :

Équation 12 : Calculs des forces $\overrightarrow{F_i}$

$$(12) \; -\vec{\nabla}E_p = -\frac{\partial E_p(t)}{\partial \vec{r_i}(t)} = \vec{F_i}(t)$$

Les forces $\overrightarrow{F_i}$ sont ensuite injectées dans l'équation de la deuxième loi de Newton (13) :

Équation 13 : Deuxième loi de Newton

$$(13)\; \overrightarrow{F_i} = m_i\, \overrightarrow{a_i} = m_i\, \frac{d\overrightarrow{v_i}}{dt} = m_i\, \frac{d^2\overrightarrow{r_i}}{dt^2}$$

On peut déterminer l'accélération $\overrightarrow{a_i}$ puis, par intégrations successives la vitesse $\overrightarrow{v_i}$, et enfin la nouvelle position $\overrightarrow{r_i}$ de l'atome i à l'instant t. Cette équation est résolue pour tous les atomes composant le système étudié.

III.1.4 Préparation et déroulement d'une simulation

La structure du système biologique à étudier peut provenir de la diffraction des rayons X, de la RMN, ou d'une modélisation par homologie. Les chaines latérales manquantes et les hydrogènes sont ajoutés au système si nécessaire. L'influence du solvant sur le système peut être induite implicitement, en ajoutant un nouveau terme au calcul des énergies, ou explicitement, en ajoutant directement des molécules d'eau ce qui augmente le nombre d'atomes total du système et par conséquence le temps de calcul de la simulation.

L'ensemble est intégré dans une structure périodique dont le motif élémentaire est constitué d'une « boite » contenant les molécules de soluté (protéine, acide nucléique) et n molécules de solvant nécessaires pour combler l'espace entre la surface des molécules étudiées et les côtés de la boite (dans notre cas 12 Å) (**Figure 47**). Des ions Na^+ et Cl^- sont ajoutés dans le solvant pour neutraliser le système, c'est-à-dire pour que la charge totale du système soit égale à 0. Pour la préparation de tous les systèmes étudiés lors de mes travaux de thèse, nous avons utilisé le modèle de solvant explicite

« Transferable Intermolecular Potentiel with 3 Points ») (TIP3P) [227] et le champ de force ff12SB [228] d'AMBER.

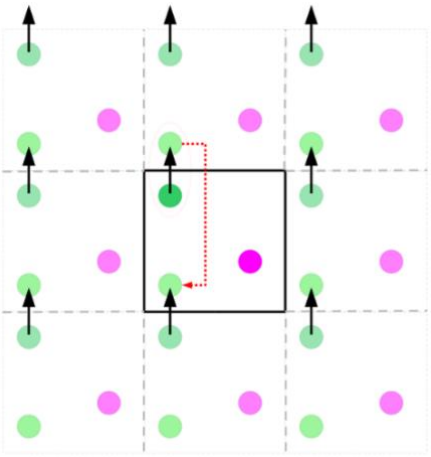


Figure 47 : Les conditions périodiques permettent de modéliser des systèmes « infinis »
Ici, les deux éléments verts interagissent indirectement ensembles (flèches noires) car ils sont très proches des extrémités de la boite centrale, malgré le fait qu'ils soient en réalité très éloignés. Ce procédé permet de ne pas contraindre la trajectoire globale de la protéine et de la limiter au centre de la boite. Cependant elle peut être emmenée à interagir indirectement avec elle-même notamment *via* les interactions à longue portée telles que les interactions électrostatiques.

L'état initial est une conformation de la molécule optimisée grâce à un algorithme de minimisation d'énergie (MM). Cet algorithme recherche l'état du système correspondant à un minimum local de l'énergie potentielle E_p. Ce minimum est dit local car il est proche de la conformation initiale du système, ce n'est pas nécessairement le minimum global. Pour la préparation de nos systèmes, nous réalisons 3 étapes de minimisation, les deux premières se font sur un maximum de 2000 itérations avec l'algorithme de « Steepest Descent », et la dernière sur 4000 itérations avec l'algorithme de « Conjugate Gradient » (décrits en **III.1.5** p. **138**). Lors de la première étape, l'objet biologique est contraint, lors de la seconde, les contraintes sont appliquées sur l'eau et lors de la troisième étape, aucune contrainte n'est appliquée au système. Une fois le système « solvaté », neutralisé et minimisé, la vitesse des atomes est nulle. Il en est de même pour les forces (14) :

Équation 14 : Forces nulles

$$(14)\ \vec{F_i} = \vec{0}$$

Dans une étape de thermalisation, de l'énergie est progressivement injectée au système jusqu'à ce que la température, initialement nulle, atteigne la valeur désirée. En général lors de simulations d'objets issus d'organismes mésophiles, la température est de 310°K (36,9°C). Cependant, pour étudier la protéine *Tt*Fpg issu d'un organisme thermophile, la température à atteindre sera de ~340°K (65°C). Lors de cette étape, l'objet biologique est sous contraintes, seule l'eau est mobile. Les vitesses initiales des atomes sont attribuées au hasard, selon une distribution de Maxwell. C'est ce qui donne la première impulsion au système, comme la première bille que l'on lâche dans le pendule balancier de Newton. Les modèles que nous avons étudiés ont été thermalisés de 0 à 310°K lors d'une unique étape de 50 ps. À la suite de la thermalisation, le système est contraint à volume et température

constants (« Substance Volume Temperature », NVT). Dans l'étape d'équilibration qui suit, le système entier est progressivement relâché à la pression constante d'1 bar (« Substance Pressure Temperature », NPT). Les systèmes que nous avons étudiés sont équilibrés en 7 étapes de 50 ps où les contraintes imposées à l'objet biologique sont dégressives (100,0% ; 50,0% ; 25,0% ; 12,5% ; 10,0% puis 5,0%) puis 2 étapes de 500 ps sans aucune contrainte. Les interactions électrostatiques sont modélisées dans les conditions périodiques par l'algorithme « Particle Mesh Ewald » (PME) [229] et les vibrations des liaisons R-H sont contraintes par l'algorithme SHAKE [230] (décrit en **III.1.7.1** p. **141**). La production est l'étape suivante, sa durée est choisie selon la nature du phénomène à étudier (**Figure 44**). Pour les analyses, nous utilisons les productions où le système est « à l'équilibre », car après l'étape d'équilibration où les contraintes sont progressivement relâchées, le système « respire » et à tendance à se transformer pour s'adapter aux nouvelles conditions de simulation. Pour savoir quand le système a atteint cet état d'équilibre, on mesure la différence structurale du système par rapport à la conformation initiale de ce dernier (calcul de RMSD dans la section **III.1.9.2** p. **143**), on considère que le système est à l'équilibre aux temps de simulation où la courbe de RMSD oscille autour d'une valeur constante et forme un plateau.

III.1.5 Algorithmes de minimisation d'énergie

Comme vu précédemment, la minimisation de l'énergie est une des premières étapes de préparation d'un système avant de réaliser la simulation de DM. Cette énergie est décrite dans l'espace cartésien par une fonction à $3\,N$ variables (3 correspondant aux 3 degrés de mouvement de chaque atome et N étant le nombre d'atomes dans le système). Elle permet de réduire les mauvais contacts entre les atomes (clashs stériques) et d'obtenir une conformation plus stable du système, correspondant à un minimum local qui ne correspond pas forcément au minimum global (**Figure 48**).

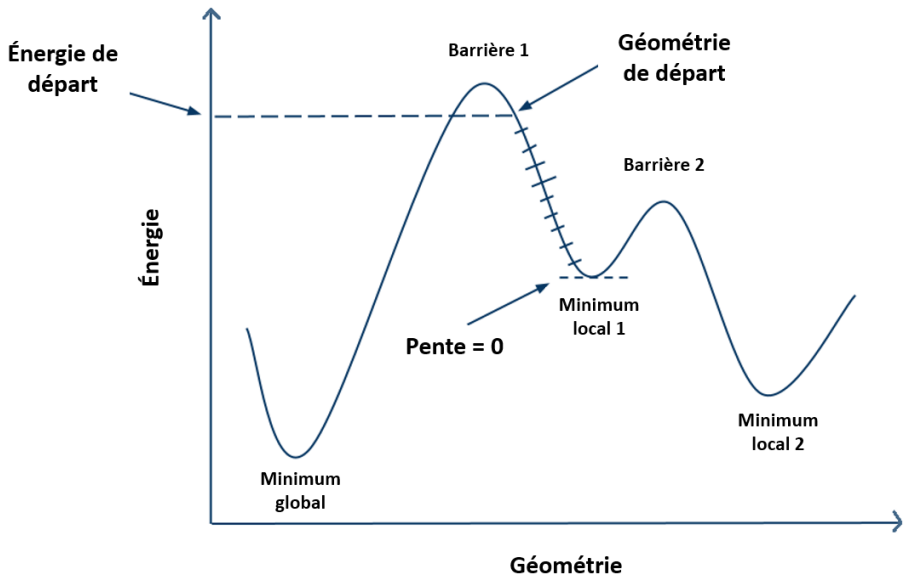


Figure 48 : Profil énergétique d'une molécule fictive
La valeur de l'énergie selon la géométrie (profil énergétique) du système met en évidence la différence entre les minima locaux et le minimum global, ainsi que le principe de la minimisation d'énergie se basant sur la recherche de la plus grande pente.

Les algorithmes les plus communément utilisés et implémentés dans AMBER sont les méthodes de la plus grande pente (« Steepest Descent » SD) [231] et du gradient conjugué (« Conjugate Gradient » CG) [232]. Lors de mes travaux de thèses, les minimisations d'énergie ont été effectuées en combinant les deux algorithmes. Les deux premières étapes sont réalisées avec SD, puis la dernière est effectuée en CG. Ces deux algorithmes consistent en la recherche itérative de la « direction » (dans le profil énergétique du système) qui mène à la conformation dont l'énergie est la plus basse (conformation la plus stable). Pour optimiser l'énergie $E(x)$ du système, il s'agit de trouver un nouveau jeu de coordonnées x^* tel que $E(x^*)$ soit inférieure à $E(x)$. Les deux méthodes SD et CG diffèrent par le choix de la direction à prendre pour atteindre la conformation à l'énergie minimale. Le SD est efficace lorsque l'état énergétique se situe sur une pente où la valeur de $E(x)$ décroit rapidement. La direction suivie sera celle indiquée par l'opposée du gradient d'énergie qui est la direction dans laquelle l'énergie diminue le plus vite localement. Cependant il n'est pas efficace pour atteindre ce minimum car il présente un comportement oscillatoire autour de ce point. C'est là qu'intervient la méthode de CG. Elle est plus couteuse en temps de calcul (facteur 2 par rapport au SD) et c'est pour cela qu'elle est utilisée pour affiner la structure après la méthode SD. L'intérêt de cet algorithme est d'éviter le comportement oscillatoire autour d'un minimum et d'accélérer la convergence. Elle est cependant inutilisable sur des structures qui présentent beaucoup de mauvais contacts, c'est également pour cela qu'elle est utilisée en deuxième, une fois que la majorité des clashs ont été éliminés par SD.

III.1.6 Algorithme d'intégration numérique

Dans notre contexte, l'algorithme de Verlet est une méthode numérique utilisée pour intégrer les équations du mouvement de Newton. Il a été utilisé pour la première fois en 1791 par le mathématicien J. B. Delambre, et a permis notamment de calculer la trajectoire de la comète de Halley en 1909 jusqu'à ce que le physicien L. Verlet l'intègre à la DM dans les années 1967 [233]. Il permet de calculer $\vec{r_i}$ à l'instant $t + \Delta t$, en supposant que les forces ne varient pas pendant l'intervalle de temps Δt. Il est donc nécessaire de prendre un Δt de l'ordre de la femtoseconde. Mais tout d'abord nous avons les forces $\vec{F_i}(t)$ s'appliquant sur l'atome i dont nous connaissons la masse m_i de l'atome, ce qui nous permet de calculer l'accélération $\vec{a_i}(t)$ de cet atome :

Équation 15 : Exploitation de la deuxième loi de Newton

$$(15)\ \vec{F_i}(t) = m_i\ \vec{a_i}(t) \Leftrightarrow \vec{a_i}(t) = \frac{\vec{F_i}(t)}{m_i}$$

L'algorithme de Verlet réduit les erreurs introduites par l'intégration numérique en calculant la position de l'atome à $t + \Delta t$ à partir des positions courantes et de la position précédente de l'atome i sans faire appel à la vitesse en utilisant deux développements de Taylor de la position $\vec{r_i}$ au deux instants $t + \Delta t$ et $t - \Delta t$ $(16, 16')$:

Équation 16 : Intégration élémentaire de Verlet

$$(16)\ \vec{r_i}(t + \Delta t) = \vec{r_i}(t) + \vec{v_i}(t)\Delta t + \frac{\vec{a_i}(t)\Delta t^2}{2} + \frac{\vec{b_i}(t)\Delta t^3}{6}$$

$$(16')\ \vec{r_i}(t - \Delta t) = \vec{r_i}(t) - \vec{v_i}(t)\Delta t + \frac{\vec{a_i}(t)\Delta t^2}{2} - \frac{\vec{b_i}(t)\Delta t^3}{6}$$

Où $\vec{r_i}$ est la position, $\vec{v_i}$ la vitesse, $\vec{a_i}$ l'accélération et $\vec{b_i}$ est le « jerk » (la secousse qui est aussi la dérivée troisième de la position par rapport au temps), les deux équations $(16, 16')$ nous donne (17) :

Équation 17 : Calcul des positions $\vec{r_i}$ à l'instant $t + \Delta t$

$$(17)\ \vec{r_i}(t + \Delta t) = 2\vec{r_i}(t) - \vec{r_i}(t - \Delta t) + \vec{a_i}(t)\Delta t^2$$

Tous les termes sont maintenant connus et permettent de déterminer les nouvelles positions $\vec{r_i}$ à l'instant $t + \Delta t$ en se basant sur la position actuelle de $\vec{r_i}$, sur la position $\vec{r_i}$ à $t - \Delta t$ et l'accélération $\vec{a_i}$ à l'instant t. La vitesse $\vec{v_i}$ ne figure plus dans cette équation. Elle n'est pas nécessaire, mais il est toujours possible de la calculer avec l'équation suivante (18) :

Équation 18 : Calcul des vitesses $\overrightarrow{v_i}$ à l'instant t

$$(18)\ \overrightarrow{v_i}(t) = \frac{\overrightarrow{r_i}(t + \Delta t) - \overrightarrow{r_i}(t - \Delta t)}{2\Delta t}$$

Le choix du pas d'intégration numérique Δt est important, il ne doit pas être trop grand pour pouvoir modéliser les mouvements les plus rapides avec précision, comme par exemple ceux de l'hydrogène. Classiquement, on utilise $\Delta t = 2\ fs$ (2.10^{-15} s).

III.1.7 Contraintes et restrictions

En anglais, les termes de « constraints » (contraintes) et « restraints » (restrictions) sont souvent confondus et sont pourtant deux choses différentes. Les contraintes imposent des critères fixes à respecter (comme une position) et ne laissent pas de liberté. Les restrictions laissent une marge de manœuvre et corrigent la valeur concernée (comme une distance) si elle sort des limites autorisées, les restrictions induisent une modification de la fonction d'énergie sans déterminer au préalable une quantité désirée.

III.1.7.1 Algorithme SHAKE

L'algorithme SHAKE [230] est utilisé uniquement au cours de la dynamique moléculaire, il permet d'appliquer une contrainte sur toutes les liaisons X-H, X étant n'importe quel type d'atome lourd. Cette contrainte permet de diminuer la fréquence d'oscillation des longueurs de ces liaisons, et permet d'accélérer le pas de production et de passer ainsi de $\Delta t = 1\ fs$ à $\Delta t = 2\ fs$. Cet algorithme permet de produire une simulation de DM deux fois plus vite.

III.1.7.2 Restrictions de distance

Dans certains contextes, il est nécessaire d'appliquer des restrictions de distance entre deux atomes, comme par exemple pour maintenir une certaine distance de 4 Å entre le N de la P1 et le C1' de la lésion dans le système *LI*Fpg complexé à un ADN contenant un THF. Cette restriction permet de mimer un système base de Schiff avec les paramètres disponibles dans le champ de force utilisé. Dans ce contexte, les restrictions qui seront appliquées à la prochaine étape sont calculées en fonction de la distance actuelle selon le potentiel harmonique (19) :

Équation 19 : Potentiel de restriction de la distance

$$(19)\ E = k(w_1\left\|\vec{r_i} - \vec{r_j}\right\| - r_0)^2$$

Où k est la constante de force, w_1 les poids, $\vec{r_i}$ et $\vec{r_j}$ sont les positions des atomes i et j et r_0 la distance à l'équilibre entre les atomes i et j définis par l'utilisateur.

III.1.8 La dynamique moléculaire ciblée

La « Targeted Molecular Dynamics » (TMD) est une méthode permettant de modéliser les structures intermédiaires entre une structure initiale et une structure finale [234]. Ce système final peut être un complexe protéine/ligand, ce qui permettrait de définir un chemin d'entrée d'un ligand vers le site d'ancrage d'une protéine. Dans le cas inverse où l'on souhaite connaitre le chemin de sortie d'un ligand hors du site actif de la protéine, la méthode utilisée est appelée TMD^{-1} (ou reverse TMD) [235]. Pour atteindre ce but, quel qu'il soit, des contraintes sont imposées au système initial jusqu'à l'obtention du système final. Cependant il n'existe pas forcément de structure correspondant à l'état final du système, comme dans le cas d'une sortie de ligand. Il est alors possible d'imposer une déformation, un changement conformationnel de tout ou partie du système initial pour forcer le déplacement du ligand, et apprécier ainsi un éventuel chemin de sortie. Le critère qui permet de forcer une déformation du système est la « Root Mean Squared Deviation » (RMSD), qui correspond à une différence entre deux structures. Si la différence entre deux structures est élevée, alors la valeur de la RMSD est élevée. Le but étant de déformer le système, il faut que nous obtenions un nouvel état du système dont la RMSD par rapport à l'état initial correspond à une certaine valeur définie au préalable. L'état initial du système est modifié progressivement par petits pas selon (20) :

Équation 20 : Potentiel de contrainte de la TMD

$$(20)\ E = \frac{1}{2}\, k_{TMD}\, N\, (RMSD_0 - RMSD)^2$$

Avec k_{TMD} la constante de force choisie par l'utilisateur, N le nombre d'atomes contenus dans les résidus (ligands, acides aminés, acides nucléiques) du système à déformer, $RMSD_0$ étant la RMSD entre le système courant et le système initial et $RMSD$ la valeur de déformation souhaitée à atteindre.

III.1.9 Analyse

Le module cpptraj d'AMBER est un outil simple et rapide pour calculer ou analyser les propriétés du système telles que des distances, des angles, l'existence ou non des liaisons hydrogène, les valeurs de RMSD, RMSF, et permet même de faire des analyses plus poussées comme des méthodes de classification par exemple. Dans cette section, je présente les analyses que nous avons utilisées pour étudier les trajectoires de DM.

III.1.9.1 Liaisons hydrogène

Les liaisons hydrogène (« Hydrogen Bonds », HB) sont des interactions entre deux atomes s'ils respectent les conditions suivantes : si ce couple d'atomes est formé d'un donneur (−NH, −OH, −SH dont la charge partielle est positive) et d'un accepteur d'hydrogène (−O, −N dont la charge partielle est négative); une distance entre ces deux atomes inférieure ou égale à 3.0 Å ; et un angle entre le donneur, l'hydrogène et l'accepteur compris entre 135° et 225°, la valeur optimale étant de 180° (**Figure 49**).

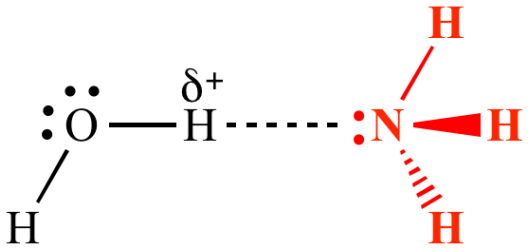


Figure 49 : Une liaison hydrogène entre une molécule d'eau et une molécule d'ammoniac
La molécule donneuse de l'hydrogène est en noire et la molécule receveuse est en rouge.

III.1.9.2 RMSD

La « Root Mean Square Deviation » (RMSD) est l'écart quadratique moyen entre deux structures, elle représente la différence structurale en Å entre deux états d'un même système. Plus cette valeur est élevée plus la différence structurale est importante. Par exemple, la RMSD entre deux conformations 1 et 2 d'un même système à n atomes se calcule ainsi (21) :

Équation 21 : Calcul de « Root Mean Square Deviation » (RMSD)

$$(21)\ \boldsymbol{RMSD}(1,2) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\vec{r}_{i1} - \vec{r}_{i2})^2}$$

III.1.9.3 RMSF

La « Root Mean Square Fluctuation » (RMSF) est la mobilité atomique moyenne pendant une simulation de DM. Pour un atome donné, plus la valeur de RMSF est élevée, plus cet atome est mobile pendant le temps de simulation analysé. La RMSF calcule donc la différence moyenne entre les T conformations produites par simulation de DM et la conformation moyenne $\bar{v}$ du système v (22) :

Équation 22 : Calcul de « Root Mean Square Fluctuation » (RMSF)

$$(22)\ RMSF(v) = \sqrt{\frac{\sum_{t=1}^{T}(v_t - \bar{v})^2}{T}}$$

III.1.9.4 Classification Ascendante Hiérarchique

La classification de données ou « clustering » en anglais est un ensemble de méthodes permettant de regrouper des objets en classes selon un ou plusieurs critères. Il existe deux types de classification : (i) la classification supervisée qui vise à classer un ou plusieurs objets dans des groupes déjà existants et qui regroupe les méthodes telles que les régressions linéaire/logistique et simple/multiple, les réseaux de neurones, les arbres de décisions, etc, (ii) la classification non supervisée qui vise à regrouper des objets dans des classes selon un critère de distance entre ces objets. La Classification Ascendante Hiérarchique (CAH) est une méthode de classification non supervisée permettant de classer des objets en un nombre choisi n de groupes en plusieurs étapes. À chaque itération, deux instances (objets et groupes) qui sont proches selon un critère (RMSD la plus faible par exemple) sont regroupées dans un même ensemble. On s'arrête lorsque l'on a atteint le nombre désiré n de groupes, aussi appelés « clusters ». A partir de chacun des clusters, il est possible d'isoler un « centroïde », un objet au centre de ce groupe. L'objet choisi pour être le centroïde d'un cluster est l'objet le plus proche de tous les autres (**Figure 50**).

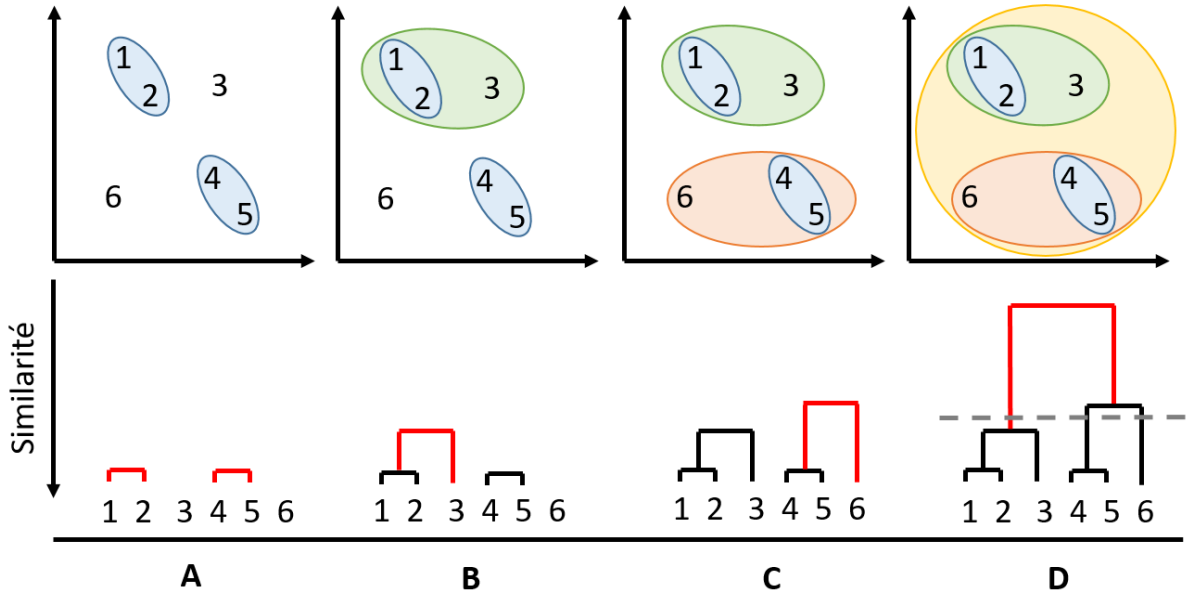


Figure 50 : Exemple de Classification Ascendante Hiérarchique (CAH)
Les 6 objets sont finalement regroupés en 3 *clusters* : (1, 2, 3) ; (4, 5) et (6). La partie haute correspond à la représentation des objets en fonction des critères choisis pour la classification, et la partie basse représente la formation d'un dendrogramme. À chaque étape **A**) **B**) **C**) et **D**) un nouveau groupe est formé. **A**) représente deux étapes jusqu'à ce que tous les objets soient rassemblés dans un même ensemble. On choisit le nombre n de *clusters* désirés de façon arbitraire ou bien pour que les éléments au sein d'un même *cluster* aient un certain degré de similarité par exemple.

III.2 **Chapitre 2 : L**'amarrage moléculaire

III.2.1 Définition

L'amarrage moléculaire (ou « Molecular Docking ») est une technique très utilisée en modélisation moléculaire, elle propose l'identification des modes d'interaction entre deux bio-molécules quelles que soient leurs tailles (protéines, acides nucléiques, sucres, lipides) (**Figure 51**). Dans le contexte d'un docking protéine/petite molécule (aussi appelée ligand), cette méthode permet de modéliser les meilleures orientations et conformations du ligand sur une surface protéique (aussi appelé récepteur ou cible), et d'évaluer sous la forme d'un score la capacité qu'ont ces deux objets à se « lier » et à « interagir » ensemble. Le docking est également utilisé en « drug design » (ou « Computer-Aided Drug Design », CADD) dans le but d'optimiser la structure du ligand et de l'adapter à la protéine.

Le problème du docking moléculaire se divise en deux parties, il s'agit tout d'abord de l'échantillonnage conformationnel du ligand et parfois du récepteur et ensuite de l'évaluation de l'affinité de liaison entre le ligand et la protéine par quantification de l'énergie libre de liaison lors d'une étape appelée le « scoring ». Le but de l'échantillonnage est d'explorer le plus de combinaisons possibles de manière à reproduire le comportement de la molécule sur la surface protéique telle quelle serait *in vivo*. Lors de cette première étape, la structure du ligand est déformée pour permettre la création des interactions les plus fortes entre les atomes du ligand et ceux de la protéine. Ces interactions sont évaluées grâce à des champs de force similaires à ceux utilisés en DM.

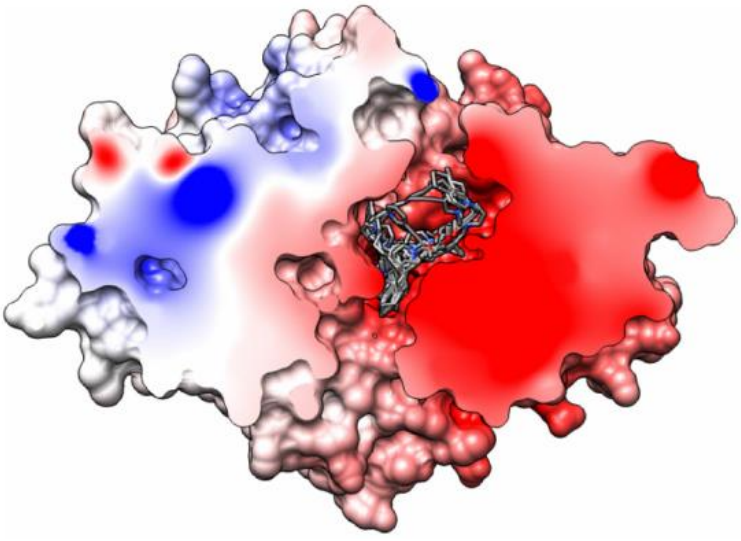


Figure 51 : Exemple d'amarrage moléculaire (« Docking »)
Docking de 44 ligands représentés en bâtonnets gris dans une cavité de la structure du transporteur 1 de la choline (ChT1) représenté en surface rouge, banche et bleue selon les charges des atomes négatives, neutres et positives respectivement. Cette structure de ChT1 est un modèle reconstruit par homologie grâce à la structure cristallographique du symporteur sodium/galactose, dont les séquences en aa sont homologues [236].

L'échantillonnage conformationnel peut être, plus particulièrement, réalisé selon deux méthodes telles que la recherche stochastique et la recherche systématique implémentées dans AutoDock 4, AutoDock Vina et Surflex-Dock.

La recherche stochastique est une recherche aléatoire, pouvant être implémentée dans plusieurs algorithmes tel que l'algorithme de Monte-Carlo (MC), de Recuit Simulé (« Simulated-Annealing », SA) ou les Algorithmes Génétiques (« Genetic Algorithm », GA) par exemple. Elle consiste en la génération au hasard de conformations putatives du ligand entier dans le volume de recherche et est parfois couplée à une étape d'optimisation par minimisation de l'énergie d'interaction protéine/ligand. La classification et la sélection par l'utilisateur des meilleures conformations sont effectuées grâce au scoring.

La recherche systématique essaye de couvrir toutes les possibilités et de choisir celles qui conviennent le mieux encore une fois grâce au scoring. Ce mode de recherche est celui des algorithmes de Construction Itérative (« Iterative Construction », IC), de Correspondance de Formes (« Shape Matching », SM) et de Forme Complémentaire (« Shape Complementary », SC) dans lesquels le ligand est construit morceau par morceau de manière itérative pour qu'il s'adapte au mieux à la surface protéique ciblée. Ces méthodes cartographient la zone de recherche et créent au préalable des figures géométriques intermédiaires correspondant aux négatifs (empreintes) de la protéine. Ce sont des motifs regroupant les points d'intérêts pouvant former des interactions protéine/ligand. Les fragments sont ensuite construits en essayant de faire correspondre autant que possible les points d'intérêts de l'empreinte et les fonctions chimiques des fragments.

La flexibilité d'un objet est sa capacité à se déformer. Les protéines et les ligands ne sont pas des objets rigides, ils sont capables de s'adapter à leur environnement. Cet aspect peut poser un problème pour déterminer le meilleur mode d'interaction protéine/ligand. En effet, la présence d'un nouvel élément sur la surface protéique peut perturber localement cette même surface et vice versa. À l'origine, le ligand et la protéine étaient traités comme des objets rigides que l'on cherchait à « emboiter » le mieux possible. Le docking semi-flexible permet la rotation de certaines liaisons du ligand assurant une déformation de la petite molécule et résolvant en partie le problème. Concernant les protéines, les structures issues de la cristallographie disponibles dans la PDB sont des clichés statiques. La DM permet d'apporter une solution à ce problème et permet de générer plusieurs conformations possibles du système. De plus, certains logiciels de docking acceptent des paramètres supplémentaires qui prennent en compte la flexibilité de seulement certaines parties de la protéine et faire un compromis sur le coût computationnel. Prendre en compte la flexibilité des molécules permet

de ne pas écarter les modes interactions protéine/ligand qui, de prime abord, ne semblent pas favorables et qui sont des faux négatifs. Dans notre contexte, un faux négatif est une conformation du ligand qui n'aurait pas un score d'affinité assez intéressant pour être sélectionnée. Ce problème intervient lorsque le récepteur n'adopte pas la meilleure conformation pour former les meilleures interactions avec le ligand, et vice versa. La flexibilité des deux objets est donc une composante importante pour la prédiction des modes d'interaction protéine/ligand.

Le score représente les interactions liées à l'électrostatique (comme les liaisons hydrogène et les interactions de vdW), les effets de solvatation et de désolvatation et l'énergie nécessaire pour déformer le ligand. Dans les logiciels que nous avons utilisés, l'affinité est évaluée grâce à la fonction de score qui calcule une approximation de l'énergie libre de liaison. Les fonctions de scores implémentées dans les logiciels que nous avons utilisés sont dérivées des champs de forces basées sur la sommation de paramètres d'énergie caractérisant les interactions atomiques. Ce sont des fonctions empiriques qui estiment l'affinité de liaison d'un complexe protéine/ligand par un ensemble de termes d'énergie pondérés par des facteurs définis par régression linéaire multiple à partir de données expérimentales de complexes protéines/ligands connus stockés dans des bases de données comme la PDBind [237]. Beaucoup de fonctions de score existent mais aucune d'entre elles n'est parfaitement précise et chacune présente ses limites [238]. Pour limiter les erreurs, il convient d'utiliser plusieurs méthodes de docking dont les fonctions de scores sont différentes et d'essayer de reproduire les résultats expérimentaux.

Le score permet de classer les ligands les uns par rapport aux autres et ainsi d'avoir une idée des molécules ayant potentiellement le plus d'effets (inhibition ou induction) sur leurs cibles biologiques. Cependant, le fait qu'une molécule ait beaucoup d'affinité pour une cible biologique ne garantit pas forcément l'inhibition ou l'induction de l'activité de cette dernière [238]. Dans le contexte de la recherche d'inhibiteurs compétitifs, le docking permet d'identifier des molécules ayant une forte affinité pour le site actif de l'enzyme.

III.2.2 Logiciels utilisés

Plus d'une centaine de logiciels référencés dans la littérature proposent aux utilisateurs plus ou moins aguerris les outils nécessaires pour réaliser des simulations de docking moléculaire [239]. Les différences entre ces logiciels sont dans les algorithmes implémentés pour la recherche et l'optimisation des orientations du ligand, la fonction de score, la performance, l'analyse des résultats et le coût de la licence car l'accès à ces logiciels n'est pas toujours offert gratuitement au public. Ils

sont produits par des laboratoires académiques ou des entreprises, se présentent sous forme de pluggins, de serveurs web ou encore sont intégrés à une suite de logiciels de modélisation moléculaire (souvent payante) proposant d'autres fonctionnalités. Pour nos calculs de docking, nous avons utilisés des logiciels dont l'utilisation est gratuite, tel que AutoDock 4 [240], AutoDock Vina [241] et AMIDE [242]. Pour la préparation des données relatives aux petites molécules, nous avons utilisé VSPrep [243]. Tous ces logiciels et les algorithmes qui y sont implémentés sont décrits à la suite de cette section.

III.2.2.1 Préparation des molécules

Préalablement au calcul d'amarrage moléculaire, les petites molécules (ligands) et les protéines (récepteurs) doivent être formatés. Les charges et liaisons rotables sont des données importantes et doivent être ajoutées aux fichiers des molécules, dont le format dépend du logiciel de docking utilisé. La suite AutoDock met à disposition le logiciel AutoDock Tools (ADT) [240] pour la visualisation, l'ajout des charges et des liaisons rotables (aux ligands et aux protéines) et la préparation d'autres fichiers concernant le volume de recherche exploré par le logiciel de docking.

Cependant, d'autres aspects sont importants lors de la préparation des ligands, tels que les tautomères et les états de protonation, les stéréoisomères et les conformères. De plus, en pharmacologie, il existe des règles préétablies pour la conception de petites molécules à effet thérapeutique. La règle des 5 de Lipinski fût créée en 1997 sur la base de structures de petites molécules médicamenteuses (drogues) connues pour pouvoir être administrées oralement aux patients. En effet, ces petites molécules sont relativement petites avec une nature lipophile modérée. En pharmacocinétique, la règle des 5 de Lipinski (« Rule of 5 », RO5) [244, 245] (**Tableau 11**) évalue en partie la capacité les petites molécules à être absorbées (A), distribuées (D), métabolisées (M) et excrétées (E) par l'organisme humain sur des critères moléculaires. Le tri des molécules avant un criblage virtuel selon le filtre de la RO5 permet d'éliminer toutes les molécules qui ne seront pas exploitées par la suite et ainsi optimiser le temps de calcul.

Outre le filtrage des ligands, la préparation consciencieuse des structures des ligands est une étape importante avant le criblage virtuel. En effet, les molécules ne sont pas des objets inertes et rigides et subissent, en fonction des propriétés chimiques de leurs environnements et de leurs structures, des transformations. L'état de protonation correspond à la position des atomes d'hydrogènes sur une molécule et dépend du pH du milieu. Une molécule peut avoir plusieurs états de protonation ainsi que plusieurs structures appelées tautomères (réarrangements au sein de la

molécule). Les molécules peuvent également adopter plusieurs structures appelées stéréoisomères (**Figure 52**). Les résultats du docking sont dépendants de la structure des ligands, l'exploration de l'ensemble des conformations des ligands est donc une étape importante de la phase de préparation.

Dans cette optique, nous avons utilisé VSPrep pour l'étape de préparation des ligands utilisés pour le criblage virtuel. VSPrep (« Virtual Screening Prep ») [243] est un workflow (WF) développé dans le logiciel KNIME [246]. L'élaboration de VSPrep a fait l'objet des travaux de thèse de Jose-Manuel Gally de l'ICOA d'Orléans. KNIME est un logiciel gratuit qui permet la construction de WF et propose des librairies utilisées dans divers domaines de recherche (tels que la chémoinformatique, la bioinformatique, les biostatistiques et le data mining entre autres). Il permet également le traitement et l'analyse de grands jeux de données. Un WF accepte divers types de données en entrée (« input ») et, à la fin d'une série d'instructions contenues dans plusieurs nœuds (« nodes »), permet d'obtenir une sortie (« output ») correspondant aux données traitées. Le WF est un outil adéquat pour la préparation des ligands avant un criblage virtuel car il permet de traiter une large quantité de données avec des outils dédiés comme par exemple les librairies RDKit [247] et ChemAxon [248].

VSPrep est un outil de préparation qui, à partir d'une librairie de ligands fournie en entrée, permet (i) la standardisation (traitement des formats), (ii) le nettoyage (suppression des ions, des molécules doublons), (iii) l'exploration conformationnelle (états de protonation et génération des tautomères, des stéréoisomères) et (iv) le filtrage des données selon plusieurs critères moléculaires et structuraux (**Tableau 11** et **Figure 52**).

Tableau 11 : Liste non exhaustive des filtres proposés dans VSPrep

Nom	Description
« Drug-like » (Règle de 5 de Lipinski [244, 245])	150 ≤ MW ≤ 500 Da xlogP ≤ 5 HD ≤ 5 HA ≤ 10
« Veber-like » [249]	« Drug-like » ET Nombre de liaisons rotables ≤ 7 PSA < 150 Å²
« Lead-like » [250]	250 ≤ MW ≤ 350 Da xlogP ≤ 3,5 Nombre de liaisons rotables ≤ 7
« Fragment-like » [251]	MW < 300 Da HD ≤ 3 HA ≤ 3 clogP ≤ 3
« Kinase-like » [252]	Molécules similaires (coefficient de Tanimoto) à un set d'inhibiteurs connus de kinases
« GPCR-like » [253]	Molécules similaires (coefficient de Tanimoto) à un set d'inhibiteurs connus de GPCR
« PAINS » [254]	Molécules similaires (sous-structures) aux « PAINS »

Abréviations : MW, « Molecular Weight » ; Da, Dalton; HD, « Hydrogen Donor »; HA, « Hydrogen Acceptor »; PSA, « Polar Surface Area » ; GPCR, « G protein–coupled receptors » ; PAINS, « Pan-Assay Interference Compounds ».

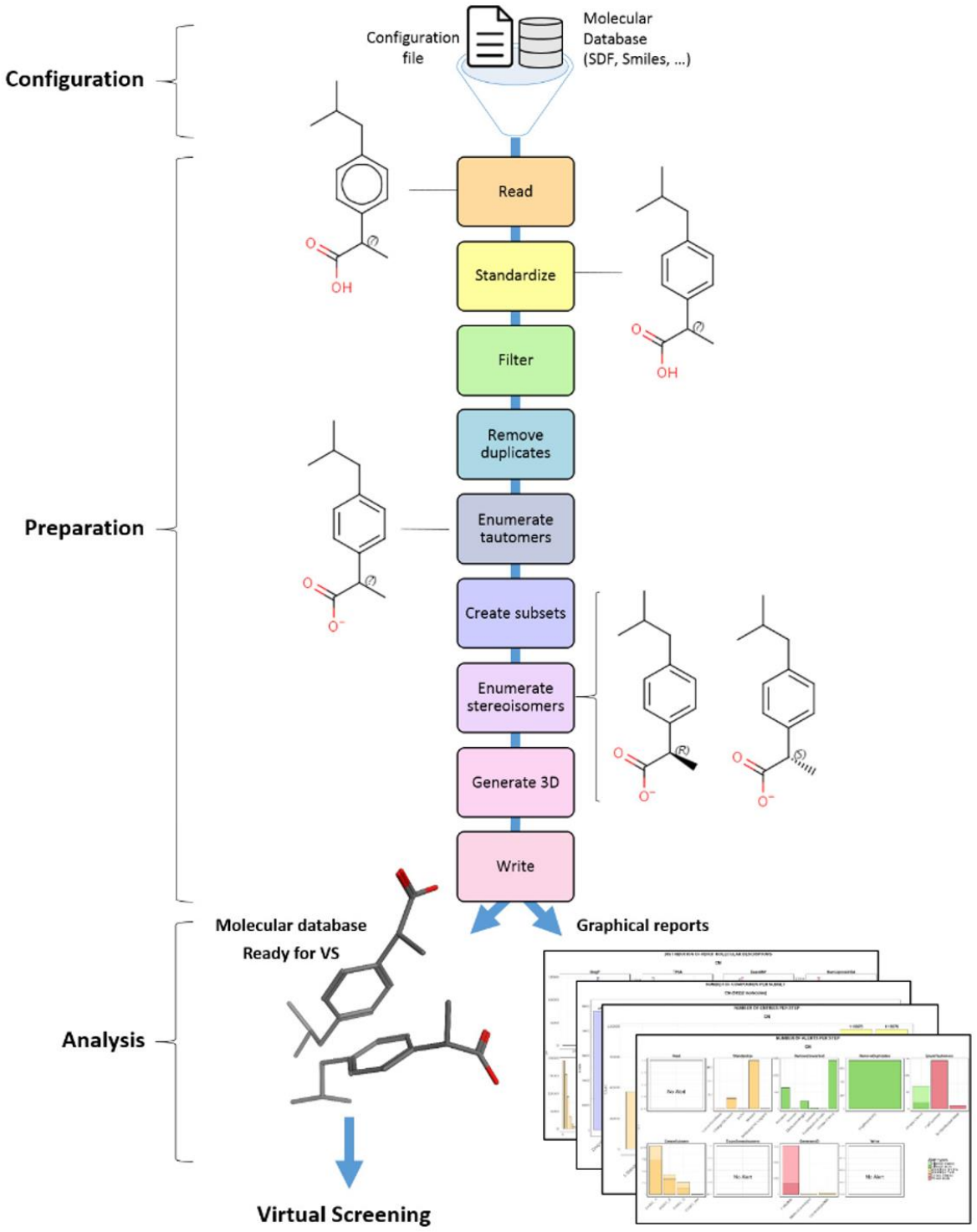


Figure 52 : Workflow intégré dans VSPrep
Il consiste en la préparation d'un set de ligands avant un criblage virtuel. Le WF génère les conformations de plus basses énergies en prenant en compte les tautomères, les stéréoisomères et les conformères. Le WF étiquette chaque ligand selon un certain nombre de filtres décrits dans le **Tableau 11** [243].

III.2.2.2 AutoDock 4 (AD4)

La première version d'AutoDock est initialement développée en 1990 par l'institut de recherche Scripps à San Diego, il fera l'objet d'un certain nombre de mises à jour jusqu'à la version la plus récente sortie en 2007 qu'est AutoDock 4 (AD4) [240]. Les outils AutoDock (AD) sont des logiciels de docking « structure-based » construits autour d'un algorithme de recherche conformationnelle du ligand et d'une fonction dévaluation de l'énergie d'interaction entre le ligand et sa cible. AD4 fourni une interface graphique permettant de visualiser les structures des molécules et les résultats tout en mettant en évidence les interactions intermoléculaires grâce à l'outil AutoDock Tools (ADT) [240]. Les logiciels ADX (X pour la version) proposent la combinaison de deux algorithmes de recherche, (i) MC et (ii) différents AG selon la version. La dernière version AD4 utilise un algorithme génétique Lamarckien (AGL) (décrit dans la section **III.2.3.5** p. **168**) qui permet la recherche conformationnelle du ligand. C'est une amélioration majeure, car elle permet aux conformations du ligand de passer les barrières énergétiques et de ne pas rester « coincées » dans des puits de potentiel et ainsi permet une recherche plus exhaustive. Cet aspect est surmonté grâce à un algorithme de recuit simulé (décrit dans la section **III.2.3.3** p. **162**) qui permet une meilleure exploration de l'espace de recherche [255]. La version AD4 propose donc un ensemble de 3 algorithmes pour la recherche conformationnelle : MC, AGL et SA.

Pour calculer l'énergie libre de liaison du complexe protéine/ligand $\Delta G_{binding}$, AD4 utilise les termes d'un champ de force selon la fonction de score (23) :

Équation 23 : Fonction de score implémentée dans AutoDock 4 (AD4)

$$(23)\ \Delta G_{binding} = \Delta G_{vdW} + \Delta G_{hbond} + \Delta G_{elec} + \Delta G_{tor} + \Delta G_{sol}$$

Les trois premiers termes ΔG_{vdW}, ΔG_{hbond} et ΔG_{elec} modélisent respectivement les énergies d'attraction/répulsion des atomes, de liaisons hydrogène et des interactions électrostatiques. Le terme ΔG_{tor} traduit l'augmentation d'énergie du système liée à la restriction des rotations libres du ligand, c'est-à-dire l'énergie à apporter pour déformer le ligand. ΔG_{sol} décrit la quantité d'énergie nécessaire pour modifier les interfaces protéine/solvant et ligand/solvant au moment de la complexation protéine/ligand. Elle correspond donc à l'énergie de désolvatation, c'est une modélisation partielle de l'effet hydrophobe. Pour estimer l'énergie libre de liaison $\Delta G_{binding}$, l'équation (23) se décompose de la façon suivante (24) :

Équation 24 : Énergie libre de liaison calculée par Autodock 4 (AD4)

$$(24)\ \Delta G_{binding} = C_{vdW} \sum_{i,j}\left(\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}}\right) + C_{hbond} \sum_{i,j} E(\theta)\left(\frac{C_{ij}}{r_{ij}^{12}} - \frac{D_{ij}}{r_{ij}^{10}} + E_{hbond}\right) + C_{elec} \sum_{i,j} \frac{Q_i Q_j}{\varepsilon(r_{ij}) r_{ij}} + C_{tor}\, N_{tor} + C_{sol} \sum_{i,j} (S_i V_j\, S_j V_i)\, e^{(r_{ij}^2 2\sigma^2)}$$

C_{vdW}, C_{hbond}, C_{elec}, C_{tor} et C_{sol} sont les 5 coefficients empiriques déterminés par l'analyse de régression linéaire multiple. Ces énergies sont calculées pour toute paire d'atomes du ligand i et de la protéine j et toutes les paires d'atomes ij du ligand séparés par au moins 3 liaisons covalentes et par la distance r_{ij}. Dans cette équation on reconnait les potentiels de Lennard-Jones $12-6$ et $12-10$ qui permettent de modéliser respectivement les interactions de vdW (dispersions/répulsions) et les liaisons hydrogène. Les coefficients A_{ij}, B_{ij}, C_{ij} et D_{ij} dépendent entre autres du couple d'atomes ij, des rayons de vdW et des types de ces atomes. Pour les liaisons hydrogène, on ajoute une constante de désolvatation E_{hbond} au potentiel qui est également pondéré en fonction de l'angle θ formé par les trois atomes suivants (i) l'accepteur, (ii) l'hydrogène et (iii) le donneur d'hydrogène. La fonction $E(\theta)$ est égale à 1 pour un angle θ égal à 180° et décroit progressivement jusqu'à devenir nulle lorsque l'angle θ est égal à 90°, traduisant une liaison hydrogène impossible. La constante E_{hbond} représente l'énergie moyenne estimée d'une liaison hydrogène entre une molécule d'eau et un atome polaire du ligand. Elle permet de pénaliser les atomes polaires du ligand qui ne forment pas de liaisons hydrogène avec le récepteur, ce qui favorise les solutions exploitant au mieux le potentiel de liaison hydrogène de la surface protéique visée. Les interactions électrostatiques sont modélisées par le potentiel de Coulomb où Q_i et Q_j sont les charges partielles des atomes d'une paire. L'influence d'une interaction électrostatique entre deux atomes ij est d'autant plus forte que la distance r_{ij} entre eux est faible, sans aller jusqu'au clash stérique. La polarité du milieu et l'effet écran sont modélisés par la fonction diélectrique $\varepsilon(r_{ij})$. La perte d'entropie du ligand lors de sa fixation sur le récepteur est prise en compte par le terme N_{tor} qui traduit la restriction des degrés de liberté conformationnelle. L'énergie nécessaire à la désolvatation du ligand lors de l'association est rapportée par le dernier terme de l'équation. Pour chaque atome du ligand, on calcule le volume V_j des atomes du récepteur qui l'entourent et on le pondère par le paramètre de solvatation atomique de l'atome du ligand S_i. Ce

terme ne concerne que les atomes de carbone du ligand en faisant bien la distinction entre les atomes de carbone aliphatiques et les atomes de carbone aromatiques. L'évaluation des atomes polaires du ligand comme O et N qui forment des liaisons hydrogène avec le solvant est intégrée au calcul de l'énergie de formation des liaisons hydrogène par la constante E_{hbond} citée plus haut.

Chacun de ces termes est pondéré par les coefficients C_{vdW}, C_{hbond}, C_{elec}, C_{tor} et C_{sol} déterminés par régression linéaire multiple. Ce modèle est construit sur la base d'un jeu de 188 complexes protéine/inhibiteur issus de la PDBind [237] pour lesquels les constantes d'inhibitions K_i sont connues. La relation liant l'énergie libre du complexe ΔG_{obs} et la constante d'inhibition K_i est donnée par l'équation (25) :

Équation 25 : Énergie libre de liaison observée

$$(25)\ \Delta G_{obs} = RT \ln K_i$$

où R est la constante des gaz parfaits ($1{,}987\ \text{cal}.K^{-1}.mol^{-1}$) et T est la température absolue.

Des informations concernant les fichiers d'entrée et de sortie d'AD4 sont en Annexe B **VI.2.1** p. **320**.

III.2.2.3 AutoDock Vina (ADV)

AutoDock Vina (ADV) [241] introduit en 2010 est la seconde génération du logiciel AutoDock (AD4) publié en 2007 par le même institut de recherche que son prédécesseur. La nature de l'algorithme de recherche et la fonction de score qui sont implémentés dans ADV sont totalement différents. Il s'agit en effet des algorithmes d'exploration conformationnelle recuit simulé et d'optimisation locale Broyden-Fletcher-Goldfarb-Shanno (BFGS), rendant le logiciel ADV plus efficace dans ses prédictions que son prédécesseur. Contrairement à AD4, ce logiciel est développé pour le calcul parallèle, ce qui le rend beaucoup plus efficace (**Figure 53**) et adapté pour le criblage virtuel haut débit [241]. Les algorithmes implémentés dans les logiciels utilisés sont décrits dans les sections suivantes.

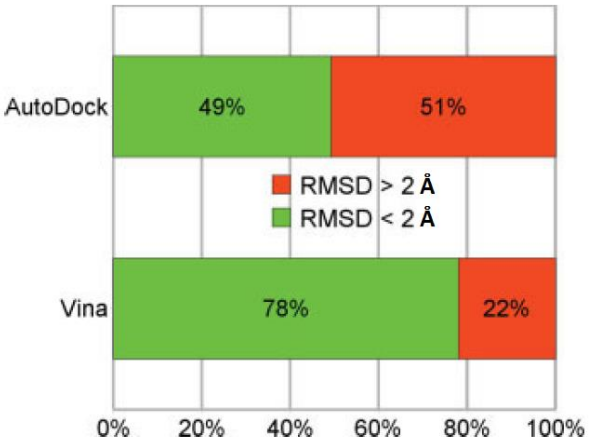


Figure 53 : Comparaison de la performance des deux logiciels de docking AD4 et ADV
AutoDock Vina semble plus efficace qu'AutoDock 4 car 78% des 190 ligands utilisés pour le test de redocking ont une RMSD < 2 Å par rapport à la structure expérimentale, contre un peu moins de 50% pour AutoDock [241].

La fonction de score empirique d'ADV a été déterminée selon la même méthode que la fonction d'AD4, c'est-à-dire en établissant un modèle de régression linéaire multiple mais en se basant sur un jeu de données beaucoup plus important : environ 1 300 complexes protéine/ligand (contre 188 pour AD4) issus de la PDBind. Le modèle et donc les pondérations des descripteurs de la fonction de score d'ADV sont différentes de celles d'AD4 (26) :

Équation 26 : Fonction de score implémentée dans AutoDock Vina (ADV)

$$(26)\ \Delta G_{binding} = \Delta G_{gauss} + \Delta G_{repulsion} + \Delta G_{hbond} + \Delta G_{hydrophobic} + \Delta G_{tors}$$

avec ΔG_{gauss} le terme d'attraction représenté par deux fonctions gaussiennes $(27, 27')$ et $\Delta G_{repulsion}$ le terme de répulsion $(27'')$:

Équation 27 : Énergie libre de liaison estimée par Autodock Vina (ADV)

$$(27)\ gauss_1(d_{ij}) = e^{-\left(d_{ij}/0{,}5\right)^2}$$

$$(27')\ gauss_2(d_{ij}) = e^{-\left((d_{ij}-3)/2\right)^2}$$

$$(27'')\ repulsion(d_{ij}) = \begin{cases} {d_{ij}}^2, si\ d_{ij} < 0 \\ 0, si\ d_{ij} \geq 0 \end{cases}$$

ΔG_{hbond} le terme des liaisons hydrogène, équivaut à 1 quand $d_{ij} < -0{,}7$ Å et 0 quand $d_{ij} > 0$ Å, et est linéairement interpolé entre les deux. Le terme $\Delta G_{hydrophobic}$ représente les interactions hydrophobes, et est égal à 1 quand $d_{ij} < 0{,}5$ Å et 0 quand $d_{ij} > 1{,}5$ Å et est également linéairement interpolé entre $0{,}5$ et $1{,}5$. Le dernier terme ΔG_{tors} est proportionnel au nombre de liaisons rotatives. Dans la fonction de score d'AD4, la distance d_{ij} entre les deux atomes considérés est définie par la distance entre les centres des atomes r_{ij} à laquelle on soustrait la somme de leurs rayons de vdW (**Figure 54**) (28) :

Équation 28 : Description de la distance d_{ij}

$$(28)\ d_{ij} = r_{ij} - (R_i + R_j)$$

Une distance d_{ij} négative signifie que les atomes i et j se recouvrent, ce qui entraine une répulsion de deux atomes l'un par rapport à l'autre selon l'équation (27''). Si d_{ij} est positive, le terme de dispersion ΔG_{gauss} est d'autant plus petit que d_{ij} est grand.

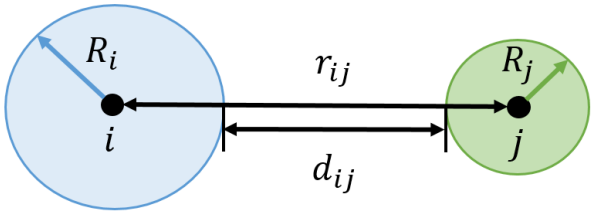


Figure 54 : Description de la distance d_{ij}
Dans les formules (27, 27' et 27''), la « distance » d_{ij} entre deux atomes, ici i et j, est la différence entre a distance r_{ij} et la somme des rayons de vdW $(R_i + R_j)$.

Les informations concernant les fichiers résultats d'ADV sont détaillés dans l'Annexe B **VI.2.2** p. **322**.

III.2.2.4 AMIDE

AMIDE [242] est l'acronyme de « Automated Molecular Inverse Docking Engine », c'est un outil développé au laboratoire Matrices Extracellulaires et Dynamique Cellulaire (MEDyC) de Reims. C'est un logiciel permettant de réaliser, comme son nom l'indique, du docking inverse.

Le docking inverse est une autre utilisation de l'amarrage moléculaire qui consiste à identifier la cible qui sera capable de former les plus fortes interactions avec le ligand utilisé pour la recherche [256]. Ainsi, on réalise un docking sans *a priori* d'un seul ligand sur les surfaces de toutes protéines dont les structures sont connues pour trouver celle qui possède la meilleure affinité pour le ligand. En pharmacologie, cette méthode permet d'identifier les molécules orphelines (dont on ne connait aucune cible) ou bien de vérifier la spécificité d'un inhibiteur pour sa cible [256].

AMIDE intègre deux moteurs de docking, (i) AD4 et (ii) ADV ainsi qu'une recherche de site de fixation par l'outil fPocket2 [257]. La stratégie d'AMIDE est de découper systématiquement la zone de recherche englobant la totalité du système en plusieurs volumes ou « boites » chevauchant(e)s (**Figure 55**). Ces volumes correspondent aux zones de recherche et le docking est réalisé dans chacun d'entre eux. Si l'utilisateur le souhaite, AMIDE peut également créer des zones de recherche autour des sites de fixations prédits par fPocket2 (**Figure 55**). AMIDE permet à l'utilisateur de choisir les paramètres

des deux moteurs de docking, que ce soit le nombre de recherches (« runs ») pour ADV ou tous les paramètres de l'algorithme génétique (dont nous parlerons plus loin) pour AD4. C'est un outil efficace car il permet la division des tâches pour des calculs sur les architectures « High Performance Computing » (HPC) de supercalculateurs. Il permet d'accomplir un grand nombre de calculs en un temps réduit, en effectuant le docking d'une molécule dans un volume de recherche par CPU (**Figure 55**). AMIDE est donc un outil adéquat pour le docking aveugle sur plusieurs structures.

Dans notre contexte, nous ne souhaitons pas réellement réaliser un docking inverse, cependant nous avons une liste de plusieurs conformations des récepteurs. Nous n'avons pas non plus d'information sur les sites de fixations des molécules $2TX_n$. C'est pourquoi nous avons utilisé le logiciel AMIDE pour une autre utilisation que le docking inverse pour laquelle il a été conçu et réalisé des tests de docking sans *a priori*.

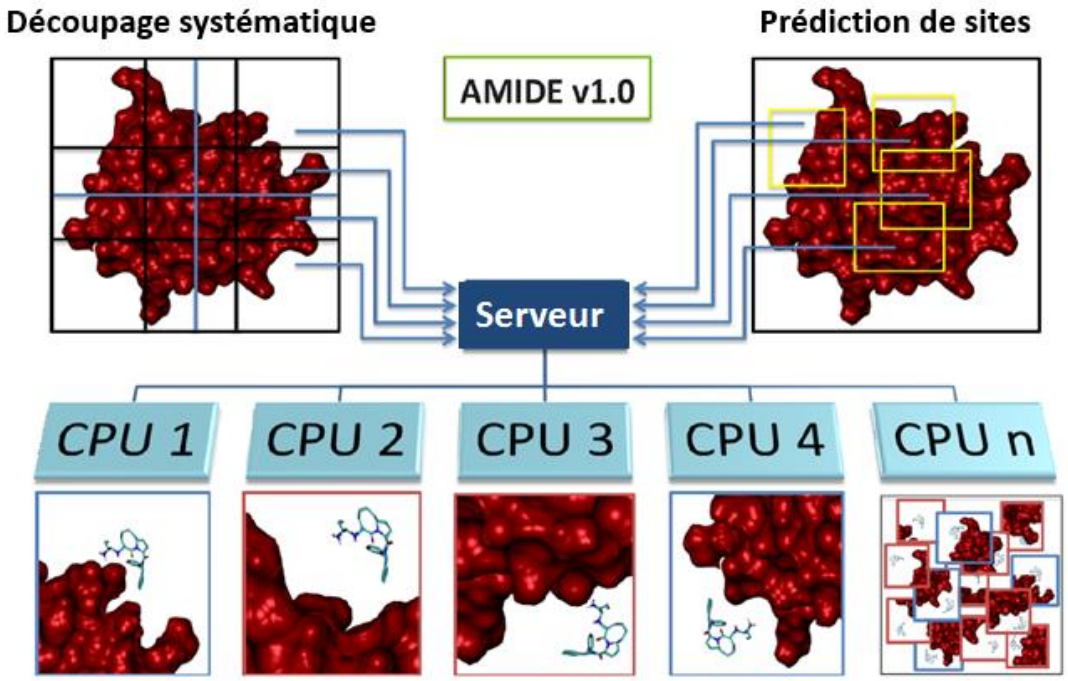


Figure 55 : Présentation du logiciel AMIDE développé au MEDyC à Reims
La première étape consiste au découpage du récepteur en volumes ou « boites » selon deux méthodes : le découpage systématique et le découpage de sites de fixation prédits par fPocket2. Ces boites correspondent aux zones de recherche qui permettront l'exploration conformationnelle du ligand avec au choix les deux moteurs de docking ADV4 et ADV. L'originalité et la puissance de l'outil résident dans sa capacité à diviser les tâches et à les distribuer sur plusieurs CPU autorisant le calcul parallèle et la diminution du temps de calcul [242].

III.2.3 Algorithmes implémentés

Dans cette partie, je décrirai les algorithmes permettant la recherche conformationnelle des ligands implémentés dans AD4 et ADV.

III.2.3.2 La méthode de Monte-Carlo Metropolis (MCM)

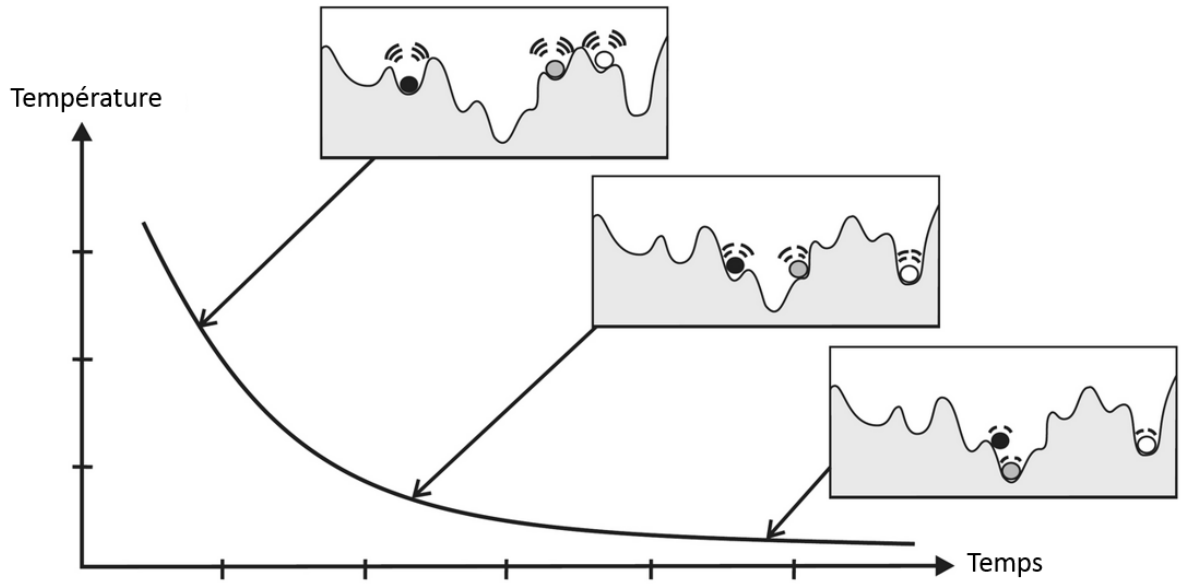
Le terme de méthode de Monte-Carlo (MC) désigne une famille d'algorithmes qui permettent d'approcher la solution d'un problème par des procédés aléatoires (techniques probabilistes). Les premiers travaux relatifs à la méthode de MC remontent à la fin de la seconde guerre mondiale dans le domaine d'étude de la matière condensée en physique. Elle fût développée par Von Neumann, Ulam et Metropolis, pionniers de l'utilisation de simulations sur ordinateur pour investiguer sur la matière, pour l'étude de la diffusion des neutrons dans un matériau fissile [258].

Cette méthode est implémentée dans les logiciels AD4 et ADV et permet la génération de la conformation initiale du ligand par des étapes aléatoires qui seront acceptées ou non selon le critère de Metropolis. On parle de simulation de Monte-Carlo Metropolis (MCM). La conformation ainsi générée est ensuite optimisée (par l'algorithme Broyden-Fletcher-Goldfarb-Shanno) pour former les plus fortes interactions avec le récepteur définies par la fonction de score. Cette étape est répétée autant de fois que nécessaire, c'est-à-dire lorsque les résultats convergent ou lorsque le nombre maximal d'essais est atteint. Cette méthode se base sur l'hypothèse qu'un nombre important de conformations initiales du ligand générées puis optimisées contient les meilleures solutions. Seule, la méthode de MCM ne garantit pas une recherche exhaustive, elle est donc couplée à une autre méthode appelée le recuit simulé.

III.2.3.3 Le recuit simulé (SA)

Le recuit simulé (« Simulated Annealing » ou SA en anglais) est un algorithme adapté de la méthode de MC et fût développé en 1979 [259]. Il est implémenté dans AD4 et ADV et permet de passer des barrières énergétiques et ainsi d'accéder à de nouvelles conformations du système qui n'auraient pas pu être obtenues sans apport d'énergie. C'est un algorithme qui permet de concevoir les conformations du ligand de manière aléatoire dans la zone de recherche, conformations qui seront ensuite modifiées pour s'adapter au mieux au récepteur lors de cycles de réchauffage (d'où le nom de la méthode) et de refroidissement lents (**Figure 56**). Cette méthode fait intervenir un algorithme de Monte-Carlo et un critère de Metropolis visant à s'approcher de la ou des meilleures solutions en

utilisant des procédés aléatoires. Cette méthode vient du constat que lorsque certains métaux refroidissent naturellement, cela ne permet pas aux atomes de se placer dans la configuration la plus solide. Cet état est atteint en maîtrisant le refroidissement et en le ralentissant par apport de chaleur



ou bien par isolation du métal.

Figure 56 : Principe du recuit simulé
Cycle de refroidissement du système lent pendant lequel la température élevée au départ permet au système de franchir des barrières énergétiques jusqu'à se stabiliser dans un des puits (ou minima locaux) du paysage énergétique, sachant que chaque énergie du système correspond à une conformation différente du système plus ou moins stable selon la valeur de cette énergie [260].

La **Figure 57**, représente les différentes étapes de l'algorithme de recuit simulé. Dans la première étape, le ligand est généré dans une position, une orientation et une conformation aléatoires. À cet instant-là, le système possède une énergie initiale E_0 (qui sera recalculée par la suite) et une température initiale élevée T_0. Cette conformation subit alors des petits changements aléatoires (réorientation du ligand, rotation des liaisons par exemple), qui entrainent successivement une variation de l'énergie du système ΔE. La suite dépend du signe de cette variation. Si elle est négative, c'est-à-dire que l'énergie du système diminue et donc qu'il est plus stable, elle est alors acceptée et appliquée à l'état courant. Si cette variation est positive, elle est acceptée avec une certaine probabilité P qui correspond au critère de sélection de Metropolis, sinon elle est rejetée. On réitère ensuite l'opération en gardant la température constante. Les itérations K sont arrêtées lorsque l'on a atteint le nombre d'itérations maximal K_{max}, ou lorsque le système est suffisament stable. La température T du système est réduite et on recommence un nouveau cycle. Ce processus est réalisé autant de fois que l'on veut générer des conformations différentes du ligand. Cela correspond au

nombre d'individus par génération (voir les algorithmes génétiques section suivante **III.2.3.4** p. **164**) dans AD4 et au nombre de runs ADV.

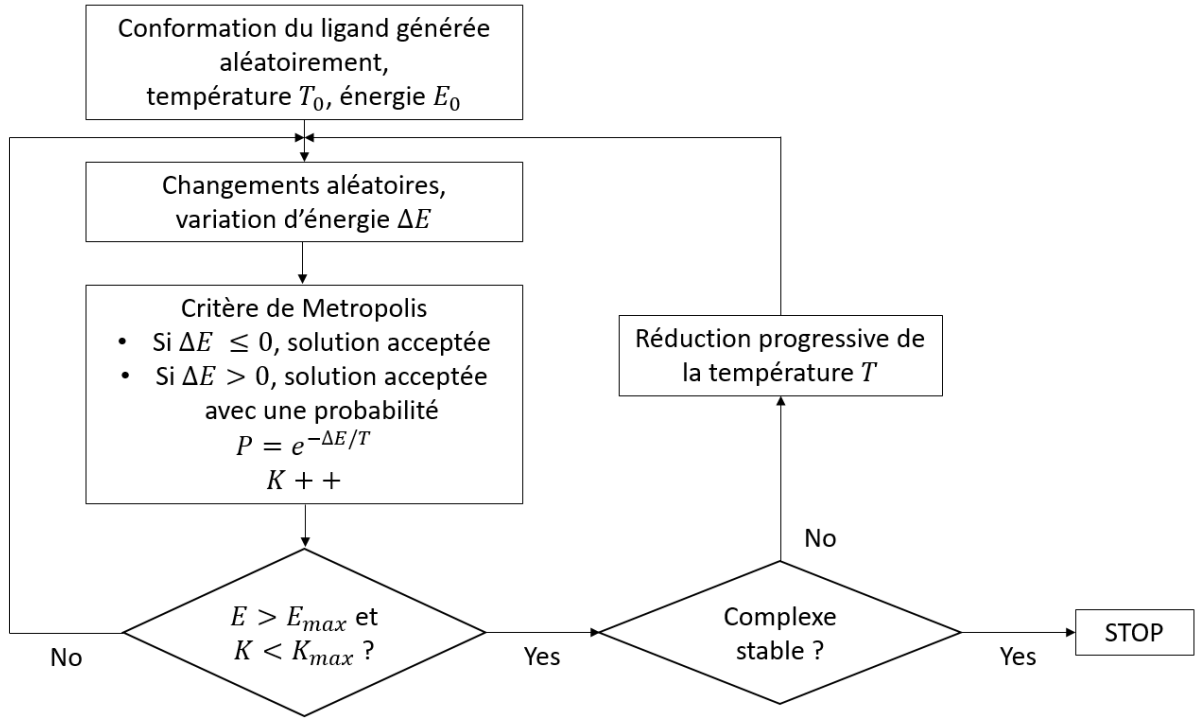


Figure 57 : Algorithme de recuit simulé
T_0 et E_0 représentent la température et l'énergie initiales du système, ΔE la variation de l'énergie du système par rapport à E_0. P est la probabilité que cette configuration du système soit acceptée et K et K_{max} sont le numéro de l'étape courante et le nombre maximal d'étapes de l'algorithme. E_{max} est l'énergie maximale du système autorisée lors de l'étape des variations conformationnelles du ligand.

III.2.3.4 Les Algorithmes Génétiques (GA)

Les Algorithmes Génétiques (GA) [255] font partie des algorithmes évolutionnistes, ce sont des méthodes stochastiques inspirées de la nature. Ils apportent une solution approchée à un problème d'optimisation lorsqu'il n'existe pas de méthode exacte. Ils se basent sur le principe de la sélection naturelle issue de la théorie de l'évolution des espèces proposée par le célèbre Charles Darwin. Cette théorie de la sélection naturelle repose sur trois principes : la mutation spontanée, l'adaptation et l'hérédité (**Figure 58**).

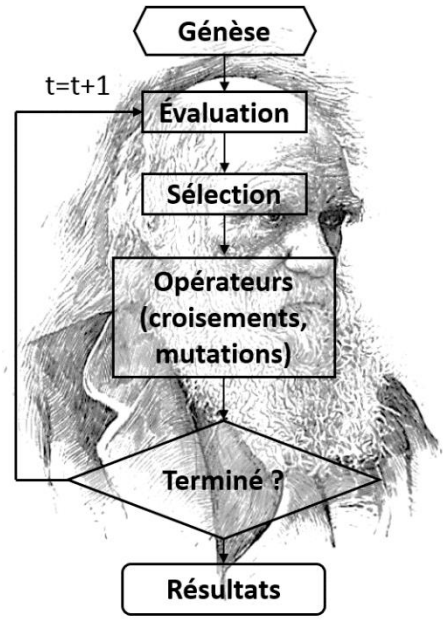


Figure 58 : Les algorithmes génétiques
Les algorithmes évolutionnistes sont inspirés du concept de sélection naturelle proposé par Charles Darwin. Le vocabulaire employé est directement calqué sur celui de la théorie de l'évolution et de la génétique. Dans cette théorie, les entités appelées individus forment une population. Chaque individu est défini par un phénotype qui est la traduction du génome, un ensemble de gènes portés sur le ou les chromosome(s). Les individus possédant la capacité à réussir une évaluation lors du processus de sélection seront soumis à des mutations qui provoquent des petites variations d'un ou plusieurs gènes. Ces individus seront rassemblés en couples et produiront la future génération qui passera par les mêmes étapes que leurs parents. Ces opérations sont répétées un nombre n de générations, jusqu'à obtention des individus les plus adaptés à leur environnement définit par les critères de sélection.

Dans le contexte de la recherche des conformations du ligand permettant les meilleures interactions protéine/ligand, le ligand est une entité « vivante » caractérisée par plusieurs descriptifs devant s'adapter à son « environnement » ou « paysage ». Cet environnement correspond à la surface protéique dans la zone de recherche. Le ligand est décrit par son génotype qui est représenté par les gènes suivants : la translation et la rotation des liaisons ainsi que les angles de torsions. Ces éléments se traduisent en phénotype qui est l'équivalent de la conformation du ligand au sein du récepteur. Dans une première étape de génération des individus, les gènes sont choisis aléatoirement. Dans la seconde étape, cette population est soumise à la sélection d'un certain nombre d'individus qui pourront persister jusqu'à l'itération suivante. Les conformations du ligand sont choisies en fonction de leur capacité à interagir de manière forte avec le récepteur, grâce à la fonction de score. Lors de la troisième étape, les ligands les plus affins subissent des mutations. Ce sont des variations de leurs gènes, elles ont un impact sur leur phénotype, et donc sur leur conformation. Dans la dernière étape, les conformations sélectionnées et mutées sont aléatoirement regroupées par paires et sont croisées pour donner la nouvelle génération de ligands. Ces individus sont ensuite les « parents » d'une nouvelle génération de conformations qui subiront à leur tour la sélection, les mutations, les croisements et ainsi de suite. Dans un tel algorithme, plusieurs paramètres sont nécessaires comme la taille maximale des populations, le seuil de sélection, le taux de mutations, le nombre d'enfants qu'aura chaque couple d'individus et le nombre de générations qui conditionnera en partie l'arrêt de l'exploration.

Dans AD4, le chromosome est composé de trois gènes correspondant : (i) aux trois coordonnées cartésiennes définissant la position du ligand sur l'axe de translation (en vert sur la **Figure**

59), (ii) aux trois coordonnées cartésiennes définissant le point « racine » sur l'axe de rotation du ligand et à la valeur de l'angle de rotation du ligand autour de cet axe (en violet sur la **Figure 59**) et (iii) aux valeurs des angles pour chaque pivots (liaisons rotables) du ligand en rotation libre (en bleu sur la **Figure 59**). Ces trois gènes permettent de générer toutes les conformations possibles du ligand dans la zone de recherche considérée.

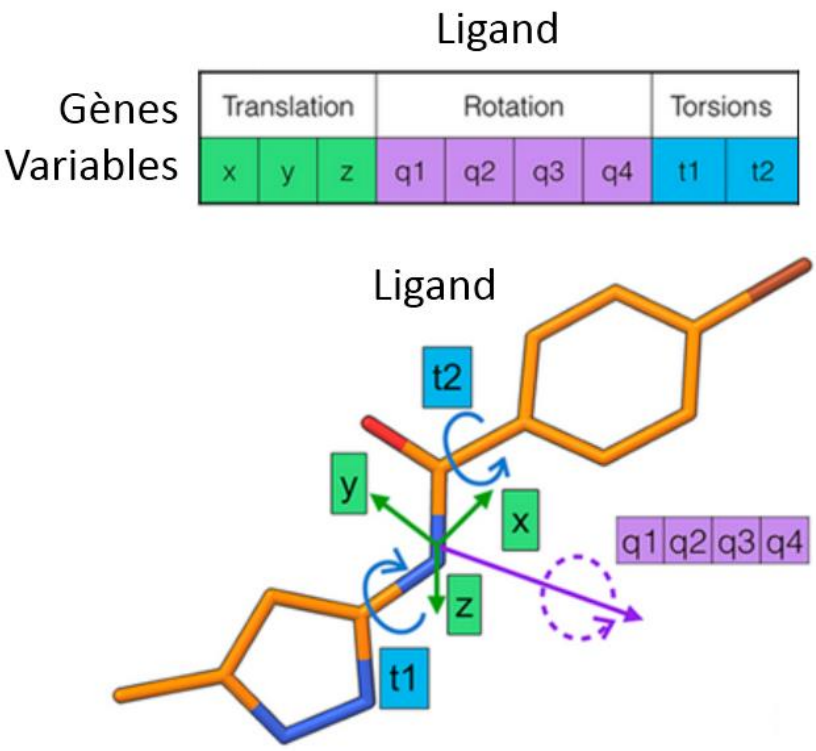


Figure 59 : Gènes composant les chromosomes du ligand
Figure issue de l'article de Ravindranath P. A. *et al.*, PLoS Comp. Biol. (2015) [261]

La phase de sélection qui suit la génération de la population détermine le nombre d'enfants qu'aura chaque individu pour former la génération qui suit, sachant que les individus ayant une capacité d'interaction avec le récepteur supérieure à la moyenne auront proportionnellement plus d'enfants. Le nombre d'enfants est donné par l'équation (29) :

Équation 29 : Condition d'arrêt de l'algorithme génétique

$$(29)\ n_i = \frac{E_b - E_i}{E_b - E_m} \text{ avec } E_b \neq E_m$$

avec n_i le nombre d'enfants qu'aura l'individu i considéré, E_i est l'énergie d'interaction de cet individu, E_b est l'énergie d'interaction de l'individu le plus mal orienté par rapport au récepteur parmi toutes les N dernières générations et E_m est l'énergie d'interaction moyenne de la génération en cours.

L'algorithme attend l'égalité $E_i = E_m$ pour considérer que la population a convergé vers la meilleure solution.

Les croisements et les mutations interviennent sur un nombre aléatoire d'individus de la population selon un taux de croisement et de mutation défini par l'utilisateur. Comme lors de la reproduction, le croisement s'opère alors en premier et les mutations ensuite. Les croisements ressemblent aux crossing-over lorsque deux chromosomes homologues s'échangent l'information contenue sur un gène. Dans notre contexte, les chromosomes parents sont découpés en 3 fragments et sont recombinés aléatoirement entre eux comme par exemple les chromosomes ABC et abc peuvent être découpés et recomposés en aBc et AbC. Le croisement donne les enfants qui remplacent les parents dans une nouvelle génération. Les mutations correspondent à une fluctuation de la valeur du gène dont la probabilité suit la distribution de Cauchy ou Loi de Lorentz (30) :

Équation 30 : Loi de Lorentz

$$(30)\ f(x; x_0, \alpha) = \frac{1}{\pi}\left[\frac{\alpha}{(x - x_0)^2 + \alpha^2}\right]$$

où les paramètres α et x_0 sont la moyenne et la variance de la distribution. Cette loi de distribution favorise les petites déviations centrées sur la moyenne mais permet des variations de plus grande amplitude avec une plus forte probabilité que n'en donne une loi Normale. La distribution de Cauchy ressemble à une loi normale, donc symétrique par rapport à x_0 et très « aplatie », où α correspond au paramètre d'étalement de la fonction, plus il est élevé et plus la fonction est aplatie.

L'Algorithme Génétique Lamarckien (AGL) [255] implémenté dans AD4 est basé sur la théorie de Jean Baptiste Lamarck selon laquelle les caractéristiques acquises par un individu durant sa vie peuvent être héréditaires. Ainsi, pendant sa vie avant de se reproduire, l'individu aura le temps de s'adapter à son environnement et pourra transmettre ses capacités d'adaptation à sa descendance. Cela signifie qu'entre l'étape de génération ou naissance d'une population et l'étape de sélection, les génomes des individus d'une fraction de la population seront soumis à un algorithme d'optimisation locale. Cette méthode essaye d'adapter l'orientation du ligand par rapport au récepteur, et est décrite dans la section suivante **III.2.3.5** p. **168**. D'après les développeurs d'AD4, l'AGL propose des résultats plus fiables en des temps de calcul plus courts [255].

III.2.3.5 L'Algorithme Génétique Lamarckien (AGL)

L'AGL est donc la combinaison d'un AG et d'un algorithme d'optimisation locale des conformations du ligand à la surface du récepteur. Cet algorithme implémente la méthode de Solis et Wets [262]. Cette méthode d'optimisation aléatoire et itérative va modifier par petits pas la valeur de tous les gènes des individus et évaluer l'énergie à chaque changement pour augmenter ou diminuer les pas, si l'énergie augmente ou diminue respectivement. La modification du phénotype se répercute sur le génotype et sera transmise à la descendance. Le nombre d'itérations lors de l'optimisation locale est un paramètre choisi par l'utilisateur, et l'algorithme s'arrête si la conformation converge vers un minimum ou si le nombre d'itération maximal est atteint. Cette méthode est une approche de recherche hydride globale-locale et permet d'obtenir des individus plus performants face à la sélection, ce qui la différencie de l'algorithme génétique classique (**Figure 60**) [255].

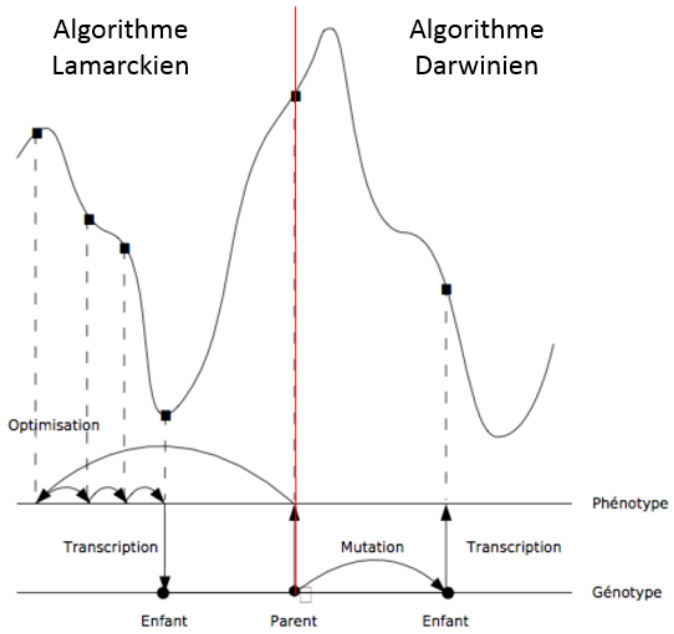


Figure 60 : Comparaison entre un algorithme génétique Lamarckien à gauche et Darwinien à droite
L'optimisation locale modifie directement le phénotype pour adapter la conformation du ligand au récepteur et réalise une transcription inverse de ces nouvelles informations du phénotype au génotype. Dans l'algorithme Darwinien, le génotype du parent est muté et donne naissance à un enfant dont le phénotype n'est pas modifié. L'algorithme génétique Lamarckien est plus efficace que l'algorithme génétique Darwinien [255].

III.2.4 Utilisations du docking

III.2.4.2 Docking sans *a priori*

Comme déjà décrit dans la section précédente **III.2.1** p. **149**, le docking est une technique communément utilisée dans un volume de recherche bien délimité par ce que l'on appelle une « boite ». L'exploration de l'espace géométrique de la surface protéique sera réalisée dans cette boite. Généralement, le docking est réalisé dans un site de fixation connu sur la cible, le plus souvent, dans le site actif. Dans notre cas, nous n'avons pas connaissance des modes d'interaction entre le ligand et

les protéines d'intérêts. Nous avons donc élargi la zone de recherche à l'ensemble de la surface du système dans une démarche que nous appelons docking aveugle (« Blind Docking ») ou docking sans *a priori* [263]. Cette méthode de docking a déjà été utilisée pour comprendre les modes d'interaction de petites molécules sur plusieurs cibles et parfois même identifier le mécanisme d'inhibition [264, 265]. AD4 et ADV ont déjà été utilisés pour ce genre de protocoles et se sont avérés être des outils efficaces pour l'identification de sites de fixations de ligands sur les protéines [263].

Lors d'un docking sans *a priori* le choix des paramètres conditionnant le nombre de runs de l'algorithme MC, le nombre de générations et d'individus par génération de l'algorithme GA est critique pour une exploration exhaustive. Si ces paramètres sont trop faibles, il est alors très probable que l'échantillon généré ne contienne pas les meilleures possibilités, elles ne pourront donc pas être identifiées. Il faut donc s'assurer que les paramètres seront suffisants pour couvrir tout l'espace conformationnel. Ces paramètres dépendent du volume à explorer, une des manières de les obtenir est de faire plusieurs essais en les faisant varier, jusqu'à ce que les résultats convergent vers la même solution. Aussi, plus le volume de recherche est important, plus le temps de calcul sera grand. Ces paramètres ont une influence sur le temps de calcul et il est donc important de bien les choisir pour obtenir la meilleure solution dans des temps computationnels raisonnables. Concernant les docking sans *a priori* réalisés dans ces travaux de thèse, nous avons toujours préféré la précision des résultats au temps de calcul. Une fois que le ou les sites de fixation ont été identifiés, il est possible de réaliser un criblage virtuel dans ces sites d'ancrage.

III.2.4.3 Criblage virtuel

Le docking est aussi utilisé pour chercher le ligand qui s'adapte au mieux à la protéine cible dans une démarche appelée criblage virtuel (« High Throughput Virtual Screening », HTVS) [266]. Dans ce contexte particulier, des bases de données (parfois commerciales) de plusieurs centaines de milliers, voire des millions de molécules sont « criblées » pour identifier la ou les molécules les plus affines du site actif de la cible. Les molécules identifiées sont appelées des « hits » et leurs effets d'inhibition ou d'induction sont *a posteriori* testés *in vitro* et *in vivo* sur les cibles par les biochimistes et/ou biologistes. Ces tests permettent de vérifier l'effet des petites molécules sur les activités enzymatiques et de mesurer par exemple les K_i (constante d'inhibition) et les IC_{50}.

III.2.5 Contrôle qualité des résultats du docking

III.2.5.2 Les courbes ROC

Aussi appelée la fonction d'efficacité du récepteur, la courbe ROC (ou « Receiver Operating Characteristic ») est un outil statistique permettant d'évaluer la précision d'un test ou d'une prédiction. Elle mesure la performance d'un classificateur binaire (0 ou 1), c'est-à-dire d'un système qui a pour objectif de catégoriser des éléments en deux groupes distincts sur la base d'une ou plusieurs caractéristiques de chacun de ces éléments. Graphiquement, la mesure ROC se représente sous la forme d'une courbe qui donne le taux de vrais positifs (fraction des positifs qui sont effectivement prédits comme des positifs, dans notre contexte les molécules actives dont les scores d'affinités sont « bons ») en fonction du taux de faux positifs (fraction de négatifs incorrectement prédits comme étant des positifs, dans notre contexte les molécules inactives dont les scores d'affinités sont « bons ») (**Figure 61**). Autour de (0,0), le classificateur (le logiciel d'amarrage moléculaire dans notre contexte) déclare toujours « négatif » et se trompe donc en créant des faux négatifs. À proximité de (1,1), le logiciel déclare toujours « positif » et se trompe donc en prédisant des faux positifs. Autour de (0,1), le logiciel prédit correctement les positifs et les négatifs et ne se trompe jamais. À proximité (1,0), le logiciel prédit les négatifs comme étant des positifs et *vice versa*, se trompant toujours. Donc, une mesure ROC dont la courbe serait sous la diagonale qualifierait des fausses prédictions, où la plupart des positifs sont des faux positifs et la plupart des négatifs sont des faux négatifs. Dans notre cas, nous nous sommes basés sur des valeurs d'IC_{50} pour définir si une molécule est « active » (ou 1, si elle a un effet) ou si elle est « inactive » (ou 0, si elle n'a pas d'effet) sur les enzymes étudiées pour évaluer les scores d'affinité issus des tests d'amarrage moléculaire.

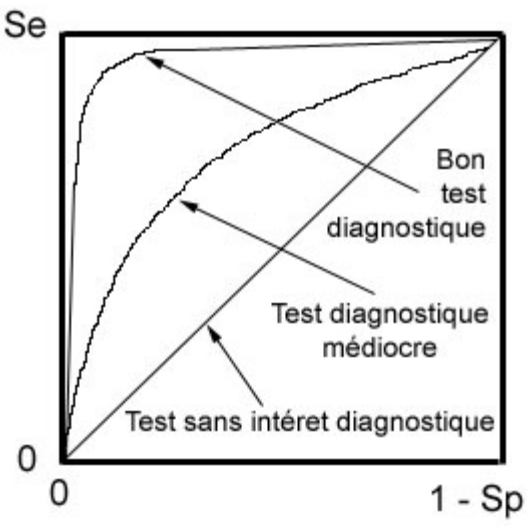


Figure 61 : Trois courbes ROC et les conclusions sur les tests ou prédictions dont elles sont issues
En abscisse nous avons « 1 - spécificité » (Sp), c'est-à-dire le taux de faux positifs, et en ordonnée nous avons la « sélectivité » (Se), c'est-à-dire le taux de vrais positifs.

IV. **R**ésultats et discussion

Chapitre 1 : Étude de la boucle LCL de *Ll*Fpg

Chapitre 2 : Étude de la sortie de la 8-oxoG hors du site actif de *Ll*Fpg

Chapitre 3 : Prédiction de sites de fixation sur les protéines *Ll*Fpg/hNEIL1

Chapitre 4 : Résultats préliminaires du criblage virtuel de hNEIL1

IV.1 **Chapitre 1 : É**tude de la boucle LCL de *Ll*Fpg

Comme décrit dans les sections **I.3.3.4** p. **100** et **I.3.4.4** p. **111** de l'Introduction, la boucle LCL semble être un élément essentiel dans la reconnaissance et la stabilisation du substrat [206]. Beaucoup de données issues de la cristallographie permettent de décrire la structure de cet élément mais de manière rigide. Cependant, la boucle LCL est très flexible [183] et des études de la dynamique de cet objet sont nécessaires pour mieux comprendre son rôle dans le fonctionnement de Fpg. Il existe quelques études *in silico* de la dynamique de la boucle LCL de Fpg dans la littérature [141, 210]. Ce motif structural a été identifié exclusivement dans les structures des protéines Fpg, et ne possède pas de structure similaire dans les structures de hNEIL1. D'après les structures cristallographiques disponibles à ce jour, cette boucle est très dynamique et peut adopter deux états [54]. Dans ce chapitre, je vais exposer les éléments nécessaires pour définir l'état de la boucle LCL, décrire la construction des différents systèmes de *Ll*Fpg utilisés pour comprendre les fonctions de la boucle LCL et présenter les résultats des simulations de dynamique moléculaire classique de ces systèmes.

IV.1.1 Les deux états de la boucle LCL

Dans les structures cristallographiques, la boucle LCL de Fpg est observée dans deux états, un état relâché (PDBid 1PM5) et un état fermé (PDBid 1XC8, 1R2Y, 1EE8) (**Figure 62**). Il est difficile d'associer l'état de la boucle LCL avec la présence ou non d'ADN ou de substrat ou d'analogue de substrat dans le site actif (**Tableau 12**).

Tableau 12 : Structures cristallographiques de Fpg libre ou complexée à un ADN

Protéine	PDBid	ADN	Orientation du substrat	État de la boucle
*Ll*Fpg	1PM5* (WT)	ADN THF	-	Relâché
	4CIS (WT)	ADN c8-oxoG	*Anti*/*Syn*	Partiellement non résolue
	1XC8* (WT)	ADN cFaPyG	*Syn*	Fermé
*Bst*Fpg	1R2Y (E2Q)	ADN 8-oxoG	*Anti*	Fermé
*Tt*Fpg	1EE8 (WT)	-	-	Fermé

* Structures cristallographiques présentées dans la **Figure 62** représentant les deux états de la boucle LCL.
WT : « Wild Type » ou type sauvage.

L'état fermé de la boucle LCL est observé dans des structures de Fpg complexée à un ADN cFaPyG provenant de l'organismes mésophile *L. lactis* (PDBid 1XC8) (**Figure 31 B**), où encore dans les structures de *Tt*Fpg libre ou du mutant *Bst*Fpg E2Q complexé à un ADN 8-oxoG (PDBid 1EE8, 1R2Y respectivement). L'état relâché de la boucle LCL a été observé dans la structure cristallographique de la protéine *Ll*Fpg WT complexée à un ADN contenant un analogue de site abasique appelé THF (PDBid 1PM5) (**Figure 62 A**). Dans d'autres structures, elle n'est pas entièrement résolue, comme notamment les chaines latérales des aa 220 à 222 de la structure de *Ll*Fpg WT complexée à un ADN contenant une c8-oxoG (PDBid 4CIS), elle est ainsi considérée comme flexible. Dans la structure 4CIS, les parties résolues de la boucle LCL se superposent avec l'état relâché de la boucle LCL de la structure 1PM5 (**Figure 33**), on en déduit qu'elle est donc dans un état relâché dans la structure 4CIS, et que cet état relâché est un état flexible. D'après les structures cristallographiques, la boucle LCL ne forme aucune interaction avec l'analogue de substrat lorsqu'elle est dans un état relâché.

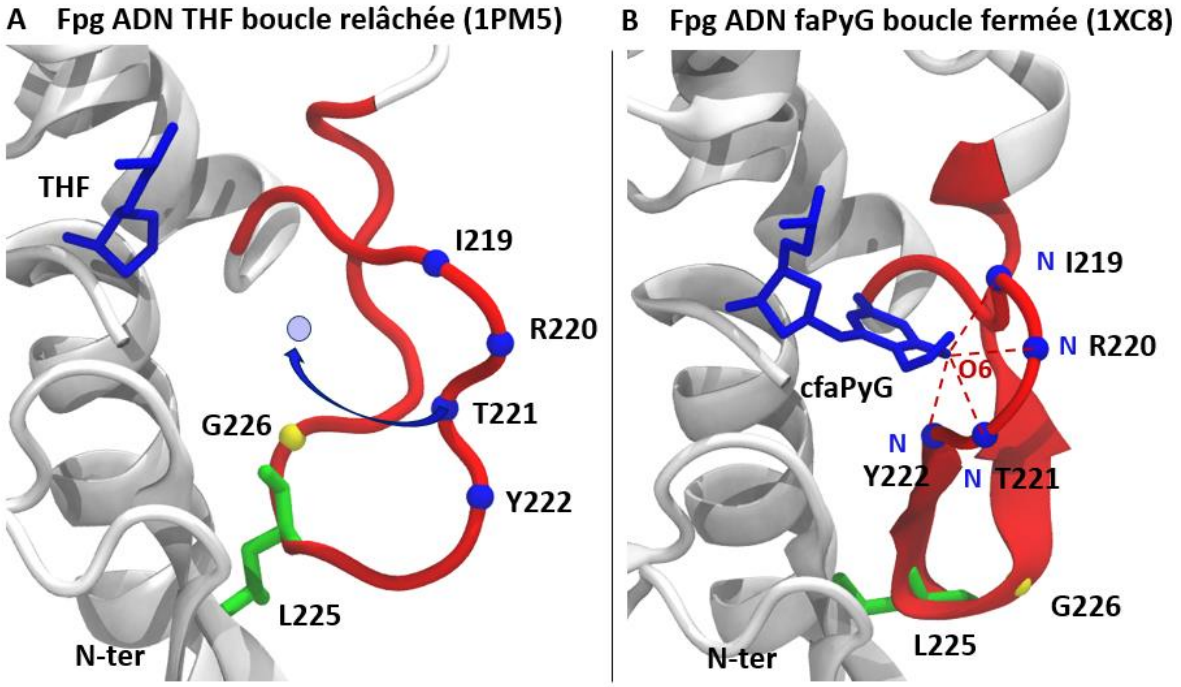


Figure 62 : Site actif de *Ll*Fpg et différence structurale de la boucle LCL dans deux structures cristallographiques 1PM5 et 1XC8

A) Boucle LCL dans un état relâché (PDBid 1PM5 [195]), et **B)** boucle LCL dans un état fermé (PDBid 1XC8 [184]). La protéine est représentée en cartoon et l'ADN est masqué pour plus de clarté. Le THF et le cFaPyG sont représentés en bâtonnets bleus. La boucle LCL est colorée en rouge, le reste visible de la protéine est en blanc. Les éléments utilisés pour décrire l'état de la boucle LCL (fermé ou relâché) sont représentés en sphères (N des squelettes de I219, R220, T221, Y222 et Cα de G226) ou en bâtonnets (A225). Les points-tillés rouges représentent des liaisons hydrogène entre le O6 du cFaPyG et des N des aa I219, R220, T221 et Y222. Les flèches correspondent au mouvement que doivent effectuer la G226 (en jaune) et la T221 (en bleu) pour passer de la structure présentée en **A** à la structure présentée en **B**.

Dans un état fermé, la boucle LCL forme une couronne de liaisons hydrogène entre l'O6 du substrat (cFaPyG) impliquant les amines secondaires du squelette des aa I219, R220, T221 et Y222 (**Figure 62 B**). On note aussi la présence d'un pont β dans la structure secondaire de la boucle LCL, impliquant les aa allant d'Y222 à T228 (**Figure 62 B**). Dans les structures de la **Figure 62**, la L225 semble agir comme une charnière, elle maintient la boucle à proximité du site actif. En effet, quel que soit l'état de la boucle, la chaine latérale de L225 est flanquée entre l'hélice α 1 et le feuillet β 5 du domaine N-terminal. Autour de cette charnière, on observe une rotation de la boucle LCL qui permet de passer d'un état fermé à un état relâché. Lorsque la boucle LCL est fermée, les aa I219 à Y222 sont orientés vers l'intérieur et G226 à l'extérieur du site actif. On observe l'inverse lorsque la boucle LCL est relâchée. Nous analyserons le comportement de ces éléments remarquables présentés dans la **Figure 62** ci-dessus dans les simulations de dynamique moléculaire ce qui nous permettra d'ajouter des éléments à la compréhension de la dynamique de la boucle LCL.

IV.1.2 Constructions des systèmes étudiés

Sept systèmes (ou modèles) de *Ll*Fpg ont été créés pour réaliser des simulations de dynamique moléculaire classique et ainsi mieux comprendre les mécanismes fonctionnels impliquant cette boucle, ils sont résumés dans le **Tableau 13** ci-après :

Tableau 13 : Description des 7 systèmes créés pour l'étude de la boucle LCL de *Ll*Fpg

Nom du système	PDBid d'origine	État initial de la boucle LCL	Ligand dans la structure d'origine	Modifications apportées	Temps de simulation (ns)	Contraintes
Fpg ADN THF boucle relâchée (1PM5)	1PM5	Relâché	ADN THF	-	200	-
Fpg ADN FaPyG boucle fermée (1XC8)	1XC8	Fermé	ADN FaPyG	-	200	-
Fpg ADN FaPyG (1XC8) boucle relâchée (1PM5)	1PM5, 1XC8	Relâché	ADN FaPyG	Insertion de la boucle LCL (217-232) relâchée de 1PM5 dans 1XC8	200	-
Fpg libre boucle relâchée (1PM5 sans ADN)	1PM5	Relâché	ADN THF	Retrait ADN	40	-
Fpg libre boucle fermée (1XC8 sans ADN)	1XC8	Fermé	ADN FaPyG	Retrait ADN	40	-

Fpg ADN THF boucle relâchée (1PM5) 8-oxoG libre (1R2Y)	1PM5, 1R2Y	Relâché	ADN THF	Insertion de la 8-oxoG libre de 1R2Y dans le site actif de 1PM5	40	Distance de 4 Å maximale entre le N de la P1, et le C1' du THF, contrainte pseudo Schiff
Fpg ADN THF (1PM5) boucle fermée (1XC8) 8-oxoG libre (1R2Y)	1PM5, 1R2Y, 1XC8	Fermé	ADN THF	Insertion de la boucle LCL (217-232) fermée de 1XC8 dans 1PM5 et insertion de la 8-oxoG libre de 1R2Y dans le site actif de 1PM5	~35*	

*La production du système Fpg ADN THF 8-oxoG libre boucle fermée s'arrête à ~35,0 ns après une dégénérescence du complexe ADN/protéine à ~25,0 ns, les détails sont dans la partie analyse.

Le FaPyG et la 8-oxoG libre ont été modélisés à partir des cFaPyG et 8-oxoG des structures 1XC8 et 1R2Y (**Figure 63**). Pour obtenir la structure du FaPyG, nous avons remplacé le C4' de la fonction carba par un O4' et ainsi obtenir un sucre classique désoxyribose. La 8-oxoG libre est obtenue en copiant uniquement les coordonnées des atomes du nucléotide de la 8-oxoG de la structure 1R2Y. Certains de ces systèmes comportent des molécules non répertoriées dans les champs de force d'AMBER, et notamment les molécules FaPyG, THF, 8-oxoG libre, l'aa P1 neutre et l'atome de zinc. Les paramètres du FaPyG en position extra-hélicale [267] et de la P1 neutre en position N-terminale [211] sont extraits de la littérature. Les paramètres des charges et de géométrie de la 8-oxoG libre et du Zn^{2+} ont été générés par le logiciel de calculs quantiques Antechamber [268] et Gaussian [269] dans le cadre de notre collaboration avec l'ICOA. Ces paramètres ne seront pas présentés dans ce manuscrit car ils n'ont pas encore fait l'objet d'une publication. La lésion FaPyG est considérée comme un substrat, car il s'agit d'une base endommagée encore reliée à l'ADN, et la 8-oxoG libre est considérée comme un produit d'excision car il s'agit d'une base endommagée coupée.

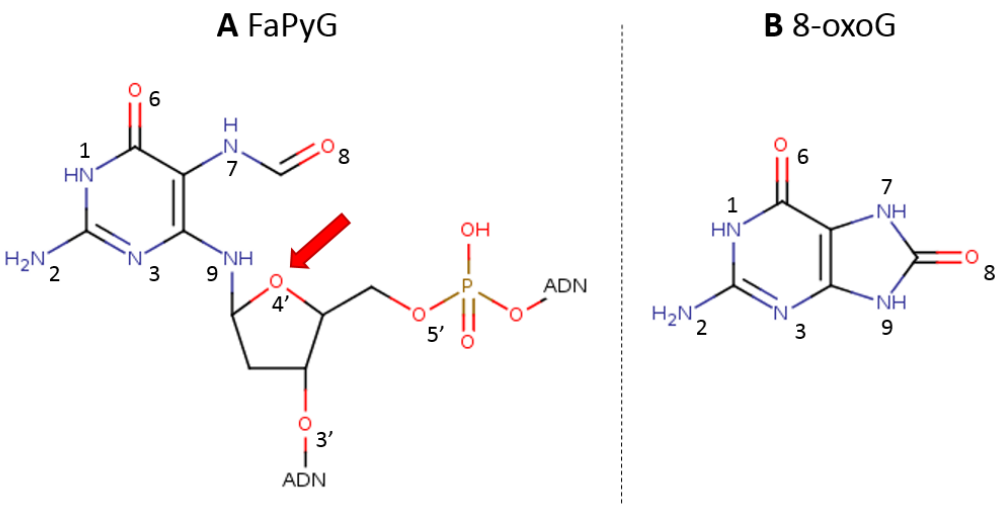
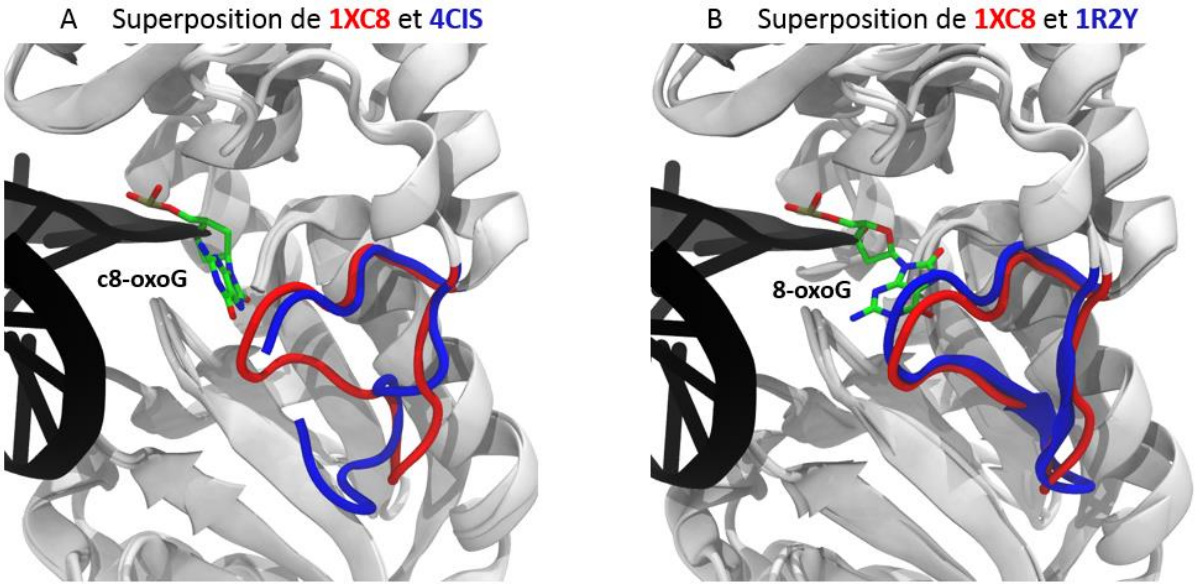


Figure 63 : Structures du FaPyG et de la 8-oxoG
Les structures sont issues de 1XC8 [196] et 1R2Y [183], respectivement. Ces molécules ont été utilisées pour construire les systèmes Fpg ADN FaPyG boucle fermée (1XC8), Fpg ADN FaPyG (1XC8) boucle relâchée (1PM5), Fpg ADN THF (1PM5) boucle fermée (1XC8) 8-oxoG libre (1R2Y) et Fpg ADN THF boucle relâchée (1PM5) 8-oxoG libre (1R2Y). Pour obtenir un FaPyG, nous avons modifié le C4' en O4' de la structure du cFaPyG de 1XC8 (flèche rouge). Quant à la 8-oxoG libre, nous avons simplement copié les coordonnées de la base azotée 8-oxoG sans prendre le sucre et le groupement phosphate qui y était relié.

Il convient de noter que le système que l'on désire étudier est *Ll*Fpg car c'est la protéine la mieux maîtrisée par les biochimistes de l'équipe, ainsi, les modèles de la protéine reconstruits sont basés exclusivement sur des structures de cette enzyme. Les structures cristallographiques initiales utilisées pour construire les modèles correspondent aux PDBids 1PM5, 1XC8 et 1R2Y. Il s'agit de structures des protéines *Ll*Fpg (1PM5, 1XC8) et *Bst*Fpg (1R2Y). Les structures 1PM5 et 1XC8 ont servi de bases à la conception des sept systèmes car elles contiennent les coordonnées 3D de la protéine *Ll*Fpg complexée à un ADN THF (analogue de site abasique) ou à un ADN FaPyG respectivement. Ces deux structures possèdent chacune un état différent de la boucle LCL, relâchée (1PM5) et fermée (1XC8) (**Tableau 13 et Figure 62**). La structure 1R2Y est une *Bst*Fpg qui est un organisme thermophile, elle contient une structure de la protéine entièrement résolue et complexée à un ADN 8-oxoG en position extra-hélicale dans le site actif de l'enzyme avec une orientation *anti*. Cette position de la 8-oxoG est compatible avec un état fermé de la boucle LCL de *Ll*Fpg, c'est-à-dire que la boucle LCL de 1XC8 recouvre parfaitement la 8-oxoG de 1R2Y, sans engendrer de clash stérique (**Figure 64 B**). Nous aurions pu utiliser les coordonnées de la c8-oxoG de la structure 4CIS, cependant, comme mentionné précédemment, une partie de la boucle LCL dans cette structure n'est pas résolue, et la base en position extra-hélicale dans le site actif possède une orientation intermédiaire entre *syn* et *anti* qui, superposée avec la boucle fermée de 1XC8, est légèrement décalée avec la boucle LCL qui ne couvre

pas le dommage (**Figure 64 A**). Nous avons donc écarté la structure 4CIS pour la construction des systèmes.

Dans les modèles comportant une 8-oxoG libre, nous avons appliqué une contrainte de distance de 4 Å entre le C1' du THF et le N de la P1 afin de mimer une base de Schiff. Cela est justifié par le fait que c'est le complexe qui se forme après l'excision de la base endommagée par l'enzyme et le fait que le site actif contient un produit d'excision et non plus un substrat dans ces modèles [217]. Comme décrit dans l'introduction, la base de Schiff est une liaison covalente entre le C1' de la base excisée et le N de la P1 de l'enzyme. Une liaison covalente entre ces deux atomes engendre une distance de ~1,5 Å, nous avons cependant choisi une contrainte de distance de 4 Å car c'est la valeur de la distance entre ces deux atomes dans la structure de 1PM5 et que nous souhaitons ne pas induire



de clash stérique dans nos modèles.

Figure 64 : Boucle LCL dans les structures de *Ll*Fpg et de *Bst*Fpg
A) Superposition des structures de *Ll*Fpg 1XC8 [201] et 4CIS [202]. La protéine est représentée en cartoon blanc, la boucle LCL de 1XC8 est représentée en rouge et la boucle partiellement résolue en cristallographie de 4CIS est en bleu. Le c8-oxoG issue de la structure de 4CIS est représenté en bâtonnets. L'ADN est représenté en cartoon noir. Cette représentation montre que l'orientation intermédiaire *syn*/*anti* de la c8-oxoG issue de 4CIS est légèrement décalée par rapport à la boucle LCL fermée de 1XC8. **B**) Superposition des structures de *Ll*Fpg 1XC8 [201] et de *Bst*Fpg 1R2Y [183]. La protéine est représentée en cartoon blanc, la boucle LCL de 1XC8 est représentée en rouge et de 1R2Y en bleu. Le 8-oxoG issue de la structure de 1R2Y est représenté en bâtonnets. L'ADN est représenté en cartoon noir. Dans cette représentation on remarque que la boucle LCL de 1XC8 recouvre totalement la 8-oxoG issue de la structure 1R2Y.

Les nouveaux modèles 3D ont été créés par l'assemblage des coordonnées 3D des atomes désirés issues de structures cristallographiques obtenues de la PDB. Les éléments non résolus (chaines

latérales et aa entiers) dans les structures cristallographiques (notamment les aa 221-224 de 1PM5 et la chaine latérale de R220 de 1XC8) ont été reconstruits avec le module xLeap d'AMBER et grâce au champ de force d'AMBER ff12SB. Les systèmes sont solvatés, neutralisés, minimisés, thermalisés et équilibrés selon la procédure décrite dans la section **III.1.4** p. **136** de la partie méthode. Les systèmes sont ensuite produits en DM sur des durées allant de 40 à 200 ns avec un pas d'intégration de 2 fs et en utilisant l'algorithme SHAKE pour contraindre la vibration des liaisons R-H. Les modèles sont dans des conditions de pression et de température constante (NPT) ainsi que dans des conditions périodiques dans lesquelles les interactions électrostatiques sont modélisées par l'algorithme *Particle Mesh Ewald* (PME).

IV.1.3 Analyse des simulations de dynamique moléculaire

IV.1.3.1 Interruption du système Fpg ADN THF (1PM5) boucle LCL fermée (1XC8) 8-oxoG libre (1R2Y)

Comme présenté dans le tableau précédent, le système correspondant à Fpg ADN THF (1PM5) boucle LCL fermée (1XC8) 8-oxoG libre (1R2Y) a subi des déformations importantes lors de l'étape de production de la simulation de DM. Sur la **Figure 65 A**, on observe en premier lieu, un éloignement des deux aa formant le pont stabilisateur du complexe ADN/protéine M75 et R260 décrit dans l'introduction, puis un éloignement de la P1 du THF. Sur le graphique, on remarque que c'est d'abord la distance entre les deux aa du pont M75-R260 qui augmente de ~6,5 Å à ~10 Å à partir de ~24 ns (**Figure 65 A**, rouge), puis la distance entre les éléments clés P1 et le THF augmente à son tour de ~4 Å jusqu'à ~30 Å à partir de ~26 ns (**A**, noir). Nous avons préféré arrêter la production de la simulation. En même temps que l'augmentation de la distance entre M75 et R260 survenant à ~24 ns, on observe une augmentation de la RMSD des atomes lourds des aa M75 et R109 (**Figure 65 B**), traduisant une augmentation de la mobilité de ces aa intercalants de l'ADN. Il en est de même pour R260 après ~28 ns, tandis que F111 ne semble pas impactée. Cela indique que la perturbation débute dans un premier temps auprès des aa M75 et R109, puis s'étend dans un second temps à R260 qui interagit fortement avec M75. Pour résumer, on observe d'abord une instabilité de M75 et R109 qui sont les aa intercalants de l'ADN, puis une rupture du pont M75-R260 ce qui met en évidence un mécanisme de destruction du complexe ADN/protéine initial. Cette observation est cohérente avec les hypothèses formulées à partir des structures cristallographiques des Fpg, qui était que M75, R109 et R260 sont des aa importants pour la stabilisation du complexe ADN/protéine. Nous avons contrôlé la distance entre les bases opposées aux extrémités de l'acide nucléique et ce phénomène n'est pas lié à des effets de bords, c'est-à-dire que l'ADN reste bien bicaténaire pendant la durée de la simulation. Ces éléments

appuient l'hypothèse selon laquelle M75 et R109 sont essentiels au maintien du complexe ADN/protéine. Cet évènement est surprenant, et n'apparait que dans la simulation d'un des sept modèles, et peut être dû au fait que le système est atypique. En effet, il est la combinaison de 3 structures cristallographiques différentes (protéine et ADN THF de 1PM5, boucle LCL de 1XC8 de *Ll*Fpg et 8-oxoG libre 1R2Y de *Bst*Fpg). Une seconde hypothèse est que la combinaison 8-oxoG libre dans le site actif de *Ll*Fpg et boucle LCL fermée n'est pas stable et ne peut donc pas exister *in vivo.* Cela est en accord avec la structure cristallographique de *Ll*Fpg complexée à un ADN contenant une c8-oxoG (PDBid 4CIS) dans laquelle la boucle LCL est non résolue plutôt que dans un état fermé. Rappelons que les coordonnées de la 8-oxoG sont issues de la structure de Fpg de l'organisme thermophile *Bst*Fpg (1R2Y). Ces résultats pourraient donc suggérer que l'orientation *anti* de la 8-oxoG dans le site actif est propre aux Fpg thermophiles telles que *Bst*Fpg et *Tt*Fpg.

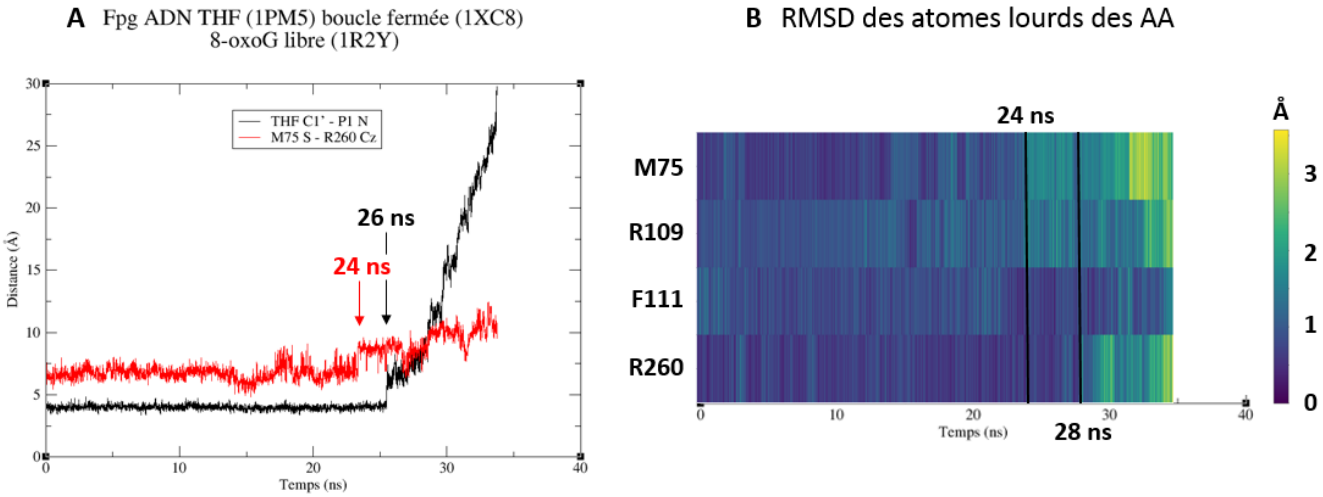


Figure 65 : Détails du système Fpg ADN THF 8-oxoG boucle fermée
A) Distances entre le C1' du THF et le N catalytique de la P1 (noir), ainsi qu'entre le soufre de la M75 et le Cz de R260 formant le pont stabilisateur du complexe ADN/protéine (rouge) et **B)** RMSD des atomes lourds des aa stabilisateurs du complexe ADN/protéine, notamment les aa du pont M75 et R260 et les aa intercalants tels que R109 et F111. On aurait donc 1) une perturbation auprès des aa M75 et R109 (aa intercalants de l'ADN), puis 2) la transmission de cette perturbation à R260 (aa du doigt à zinc tordant l'ADN) et 3) déformation du complexe ADN/protéine.

IV.1.3.2 Flexibilité globale de la protéine

La *Root Mean Square Fluctuation* (RMSF) est la fluctuation moyenne d'un atome autour d'une position moyenne au cours du temps de simulation de DM considéré. Plus un atome a tendance à s'éloigner d'une position moyenne au cours de la simulation de DM, plus la valeur de RMSF associée est élevée. Elle se calcule en général sur les Cα de la protéine, car leur mobilité est influencée par celle de la chaine latérale qu'ils portent, sur le temps de simulation où le système est considéré à l'équilibre,

c'est-à-dire à partir de 10 ns pour tous nos systèmes (Annexe C **VI.3** p. **324**). Dans cette section, nous avons calculé les RMSF des Cα des protéines dans les différents modèles (**Figure 66**).

Les aa intercalants de l'ADN, M75, R109, F111 et le doigt à zinc semblent plus flexibles dans les systèmes des protéines libres (**Figure 66 C** bleu et orange) que dans les systèmes où la protéine est complexée à l'ADN. Ceci est cohérent car ces résidus ne sont plus en interaction avec l'ADN et sont donc par conséquent beaucoup plus mobiles. Nous remarquons également que l'hélice α entre les deux domaines N- et C-terminaux est très flexible dans les systèmes Fpg ADN THF boucle relâchée et Fpg ADN THF boucle relâchée (1PM5) 8-oxoG libre (1R2Y) (**Figure 66 A** noir et **D** cyan). De plus, la boucle LCL dans les systèmes « boucle fermée » semble plus stable que la boucle LCL des modèles « boucle relâchée », ce qui est en bon accord avec le fait que les boucles LCL relâchées ne soient pas entièrement résolues dans certaines structures cristallographiques. Le détail de la RMSF de la boucle LCL est présenté sur la **Figure 67** de la section suivante **IV.1.3.3** p. **185**.

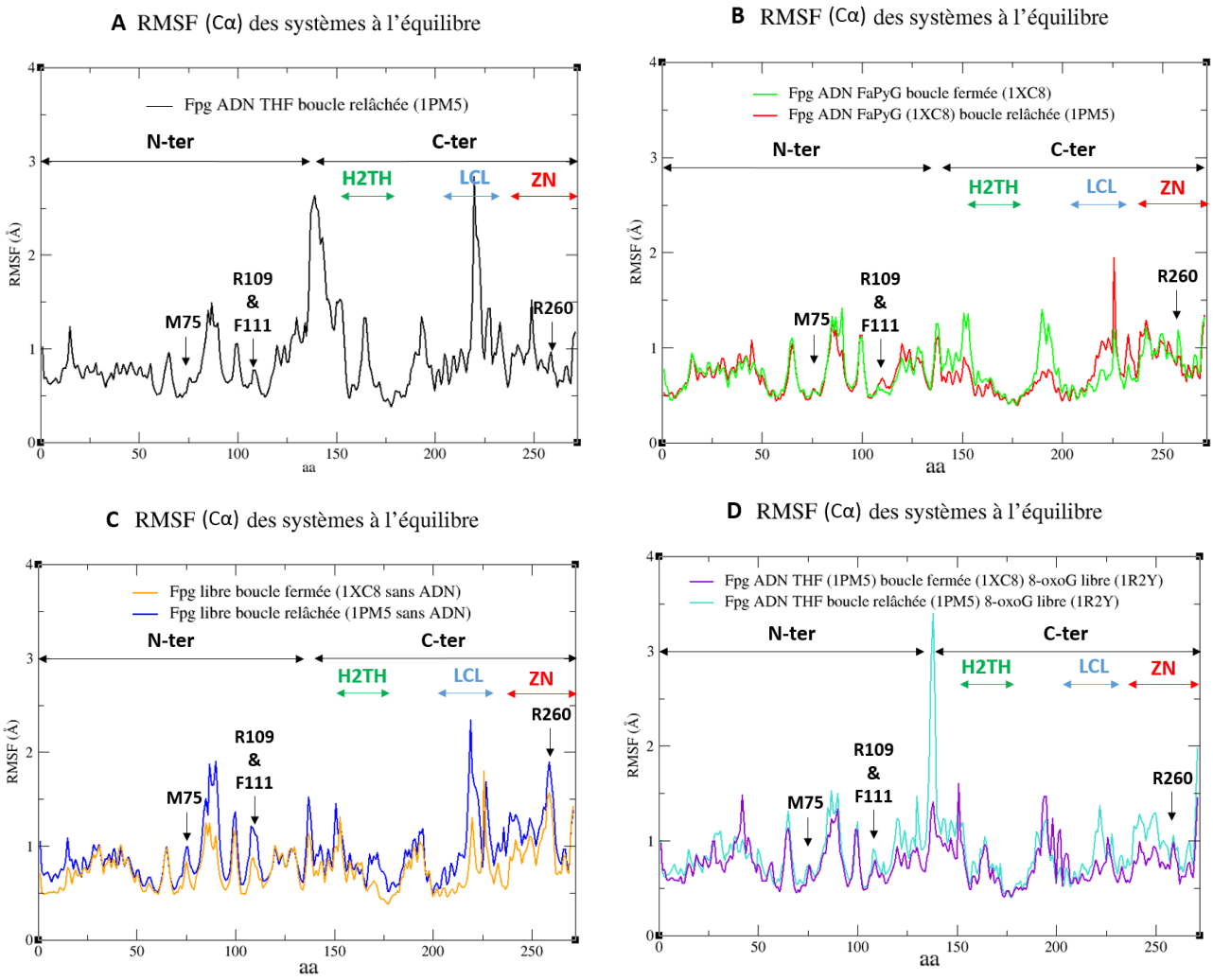


Figure 66 : RMSF (sur les Cα) des 7 systèmes produits en simulation de DM
A) Fpg ADN THF boucle relâchée (1PM5), **B)** modèles de Fpg ADN FaPyG boucle fermée (1XC8) et boucle relâchée (1XC8, 1PM5), **C)** modèles de Fpg libres boucle relâchée (1PM5) et boucle fermée (1XC8) et

D) modèles de Fpg ADN THF boucle relâchée (1PM5) et boucle fermée (1PM5, 1XC8) contenant une 8-oxoG libre (1R2Y) dans leurs sites actifs. Les annotations correspondent aux différents motifs identifiés chez Fpg et décrits dans l'introduction, à savoir les domaines N-ter et C-ter, le motif H2TH (156-180 en vert), la boucle LCL (216-230 en bleu) et le ZnF (243-270 en rouge), ainsi que les aa de la triade d'intercalation M75, R109 et F111 et le pont traversant l'ADN composé des aa M75 et R260.

IV.1.3.3 Flexibilité de la boucle LCL

La **Figure 67** représente les RMSF de la boucle LCL de *Ll*Fpg des différents systèmes étudiés. Il semblerait que L225 soit un aa stable quel que soit le système ce qui est cohérent avec les structures cristallographiques décrites précédemment, où cet aa était présenté comme une charnière autour de laquelle le reste de la boucle LCL se déforme. Pour faciliter l'analyse, nous diviserons la boucle LCL en deux parties, la partie précédant L225 que l'on appellera « Partie I, 216-224 » et la partie suivant L225 que l'on qualifiera du nom de « Partie II, 226-230 ».

Dans les systèmes issus de la structure 1PM5 dont le site actif et vide et dont la boucle LCL est relâchée (**Figure 67 A** noir et **E** bleu), quelques aa de la partie I de la boucle tel que I219, R220, T221 et Y222 semblent se déplacer au cours du temps de la simulation, ce qui traduit une certaine flexibilité de cette partie de la boucle LCL (RMSF > 2 Å). La partie I des autres systèmes dont le site actif est occupé et/ou dont la boucle LCL est fermée (**Figure 67 B** vert, **C** rouge, **D** orange, **F** violet et **G** cyan) semble moins flexible (RMSF ≈ 1 Å). De manière générale, lorsque la boucle LCL est relâchée, la partie I de la boucle est plus flexible que dans le système homologue dont la boucle LCL est fermée (**Figure 67 B** vert et **C** rouge, puis **D** orange et **E** bleu, puis **F** violet et **G** cyan). Cela est en accord avec le fait que lorsqu'elle est dans un état fermé, la boucle LCL est plus stable que lorsqu'elle est dans un état relâché. Quel que soit le système, la position charnière de la L225 semble toujours très stable (**Figure 67** de **A** à **G**). La G226 de la partie II de la boucle LCL dans certains des systèmes issus de la structure 1XC8 (**Figure 67 C** vert et **D** orange) est modérément flexible selon les cas (RMSD ≈ 2 Å) et, en de moindres mesures, la S227 et T228 de la partie II de la boucle LCL des systèmes issus de 1PM5 (**Figure 67 A** noir et **E** bleu) (RMSF ≈ 1,5 Å). Il semblerait que la partie II de la boucle LCL soit plus stable dans les systèmes comportant la 8-oxoG libre que dans les autres (**Figure 67 F** magenta et **G** cyan). Les aa de la partie II de la boucle pourraient donc peut être permettre à la boucle LCL de reconnaitre le produit d'excision (base coupée) d'un substrat (base en position extra-hélicale).

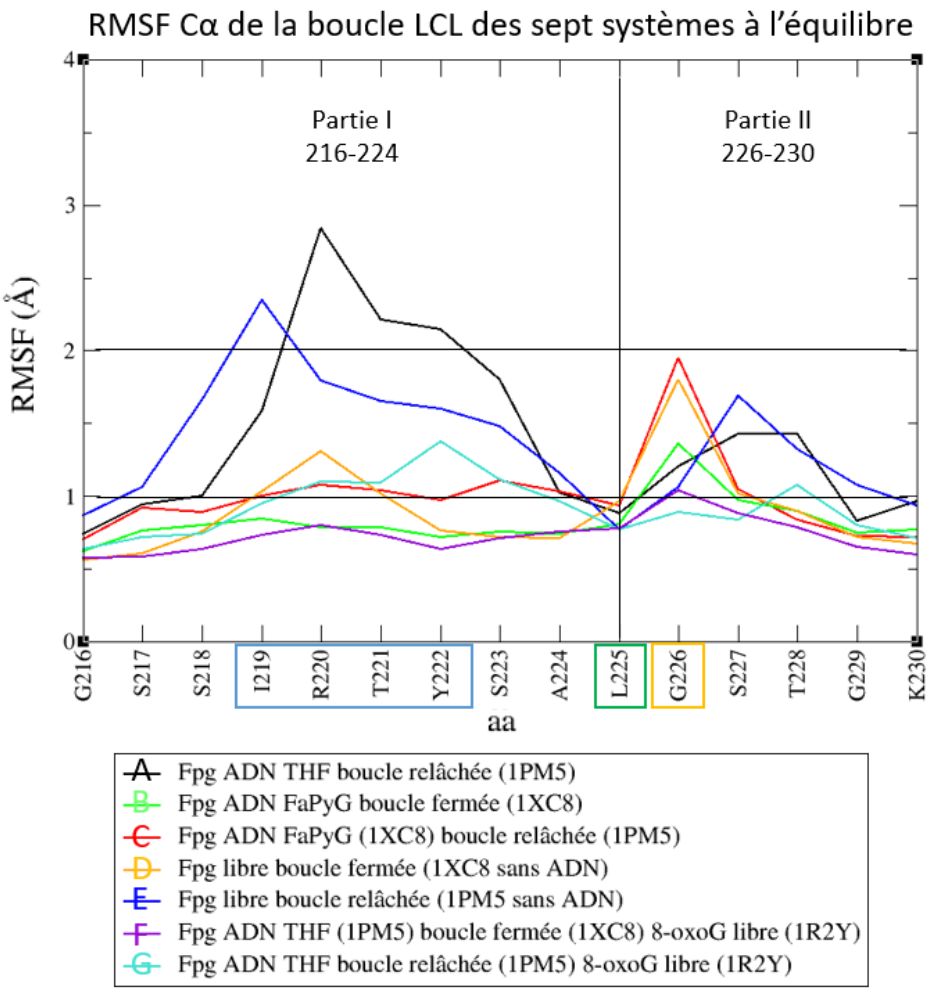


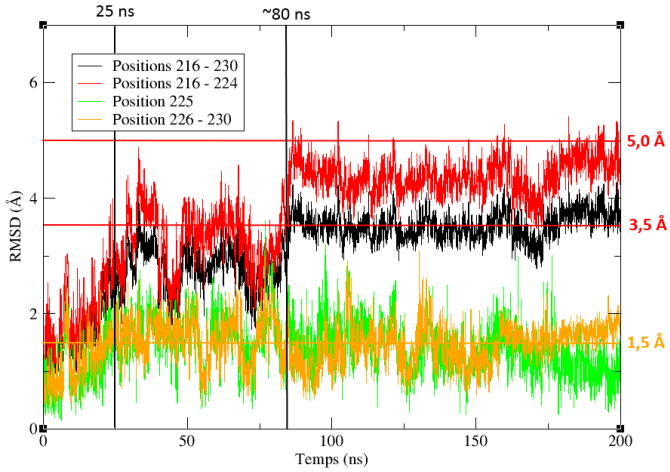
Figure 67 : RMSF de la boucle LCL
La RMSF est calculée sur les Cα de la boucle LCL, en ne considérant que les temps de simulation où la protéine est à l'équilibre, à savoir après 10 ns de simulation de DM classique. Quel que soit le système, l'aa L225 (encadrée en vert) reste une charnière stable. Cet aa permet de visualiser deux parties de la boucle LCL se comportant de manières différentes selon les systèmes, la partie 216-224 et la partie 226-230. Les aa I219, R220, T221 et Y222 (encadrés en bleu) sont, dans les structures cristallographiques directement en interaction avec la base endommagée d'après les structures cristallographiques. La G226 (encadrée en jaune) est un indicateur de l'état de la boucle LCL, elle est orientée à l'intérieur du site actif dans une boucle relâchée et à l'extérieur du site actif dans une boucle fermée. Nous avons attribué des lettres à chacun des sept systèmes, nous garderons ce code dans la suite de ce chapitre.

Contrairement aux calculs de RMSF, les calculs de RMSD permettent d'appréhender la déformation structurale d'un objet au cours du temps de simulation de DM en comparant sa structure à l'instant t par rapport à sa conformation initiale ou à l'équilibre. Ci-dessous, nous présentons les RMSD pour les parties I et II ainsi que la charnière de la boucle LCL. La **Figure 68** représente la différence structurale de la boucle LCL au cours du temps par rapport à sa structure initiale. Dans les systèmes où la boucle LCL est dans un état initial relâché (**Figure 68 A, C, E** et **G**), la RMSD de la boucle LCL est plus important (RMSD allant de 1,7 à 5 Å) que dans les systèmes correspondants où la boucle LCL est dans

un état initial fermé (RMSD compris entre 1 et 1,2 Å) (**Figure 68 D** et **F**). Cela est cohérent avec le fait que lorsque la boucle LCL est dans un état fermé, elle est plus stable que lorsqu'elle est dans un état relâché où elle semble d'avantage flexible et dynamique. Lorsqu'il n'y a pas d'ADN dans le système et que la boucle LCL est relâchée (**Figure 68 E**), on remarque que la RMSD globale de la boucle LCL varie entre 1,2 et 3,9 Å. D'autre part, l'ADN est présent mais que le site actif est vide, la RMSD globale de la boucle s'élève jusqu'à 3,5 Å (**Figure 68 A**). Lorsque le site actif de la protéine est occupé par un substrat (FaPyG) ou un produit (8-oxoG libre) (**Figure 68 C** et **G**), la RMSD globale de la boucle ne dépasse pas 2,4 Å, montrant que la présence d'une molécule dans le site actif semble stabiliser la boucle LCL. Nous remarquons également que dans les systèmes où l'état initial de la boucle est relâché (**Figure 68 A, C,** E et **G**), la partie I (216-224) de la boucle LCL induit une hausse du RMSD global de cette dernière. La partie II (226-230) de la boucle LCL est quant à elle beaucoup moins flexible que la partie I et ne semble pas impactée par la présence ou non d'ADN dans le système ou de la présence ou non d'une molécule (substrat ou produit) dans le site actif de l'enzyme. Il est possible que nous n'ayons pas un aperçu de toute la flexibilité que peut avoir la boucle LCL dans les systèmes **E** et **G** car le temps de simulation est égal à 40 ns, car on remarque que dans le système **A**, les changements conformationnels importants de la boucle LCL semblent survenir à partir de 40 puis 80 ns de simulation de DM.

Dans la **Figure 68 A**, il semblerait que les RMSD de la boucle LCL entière et de la partie I (216-224) de la boucle LCL atteignent plusieurs paliers où elle oscille autour d'une valeur constante, notamment entre 25 et 80 ns (RMSD ~3,5 Å) puis entre ~80 et 200 ns (RMSD ~5 Å). Cela suggère que les aa 216 à 224 de la boucle LCL subissent des déformations importantes à 25 puis à 80 ns de la simulation de DM, nous les décrirons plus loin dans ce chapitre.

A Fpg ADN THF boucle relâchée (1PM5)



(Suite page suivante)

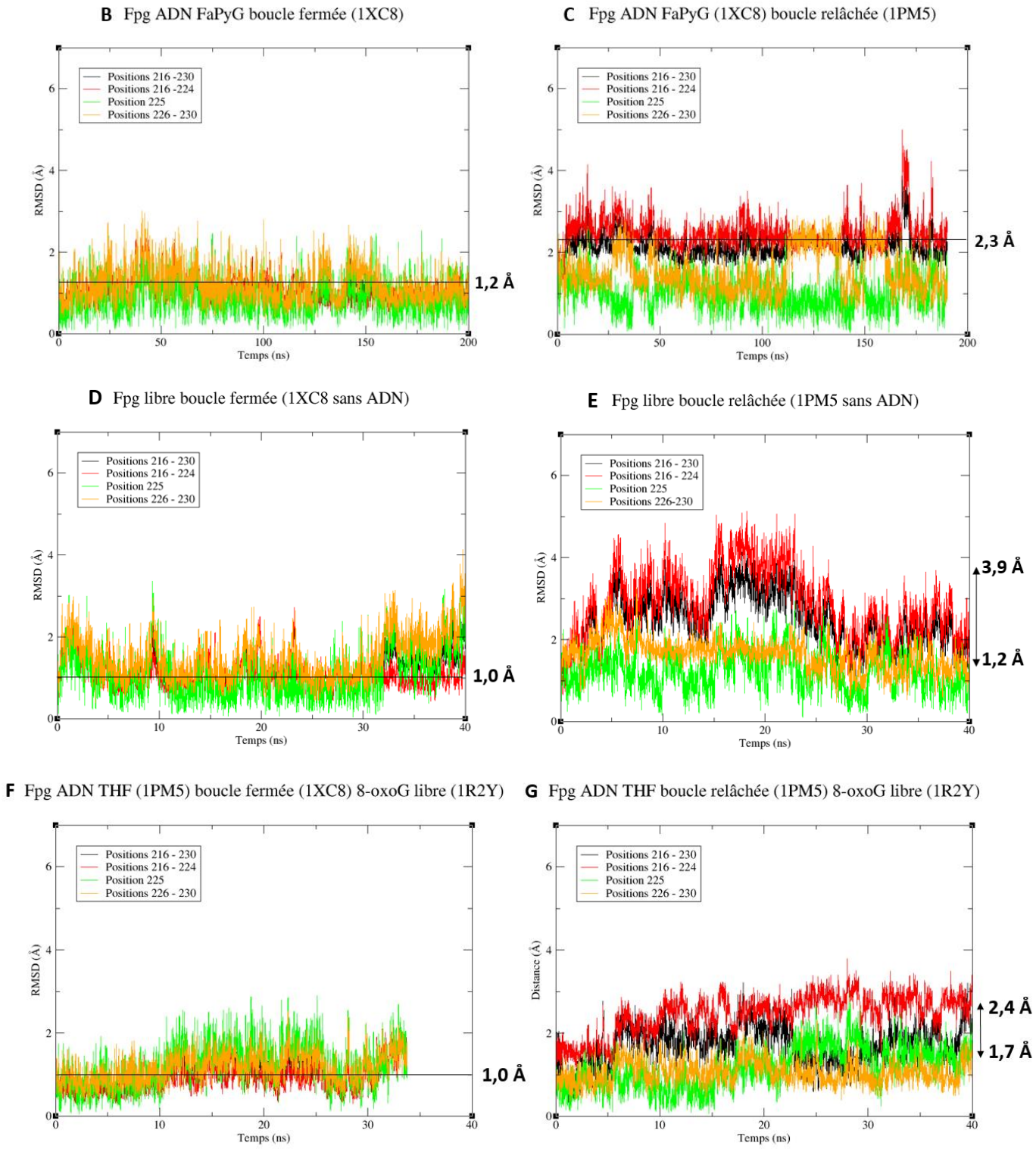


Figure 68 : RMSD de la boucle LCL en fonction du temps de tous les modèles étudiés
La RMSD est calculée sur les Cα des résidus uniquement. En noir, la boucle LCL en entier, de la position 216 à 230. En rouge, la RMSD de la partie I de la boucle LCL, de la position 216 à 224. En vert, la RMSD de la position 225, qui est également la position charnière. En orange, la RMDS de la partie II de la boucle LCL, de la position 226 à 230.

Les courbes de RMSF et RMSD sont utiles pour identifier les zones de la protéine et les aa flexibles au cours de la simulation de dynamique moléculaire, mais elles ne nous permettent pas de

qualifier cette flexibilité. Il peut s'agir en effet d'une oscillation plus ou moins grande de la boucle autour d'une position moyenne ou bien d'une déformation globale de cette boucle. D'autres analyses peuvent ajouter des informations à ce schéma partiel, comme des calculs de distances ou l'identification de liaisons hydrogène.

IV.1.3.4 Interactions boucle LCL/substrat

D'après les structures cristallographiques, la boucle LCL dans l'état fermée forme 4 liaisons hydrogène avec les substrats 8-oxoG et FaPyG. La **Figure 69** nous permet d'appréhender les liaisons hydrogène qui se forment entre la protéine et la molécule contenue dans le site actif, ainsi que les fréquences d'apparition de ces liaisons hydrogène sur la durée totale de simulation.

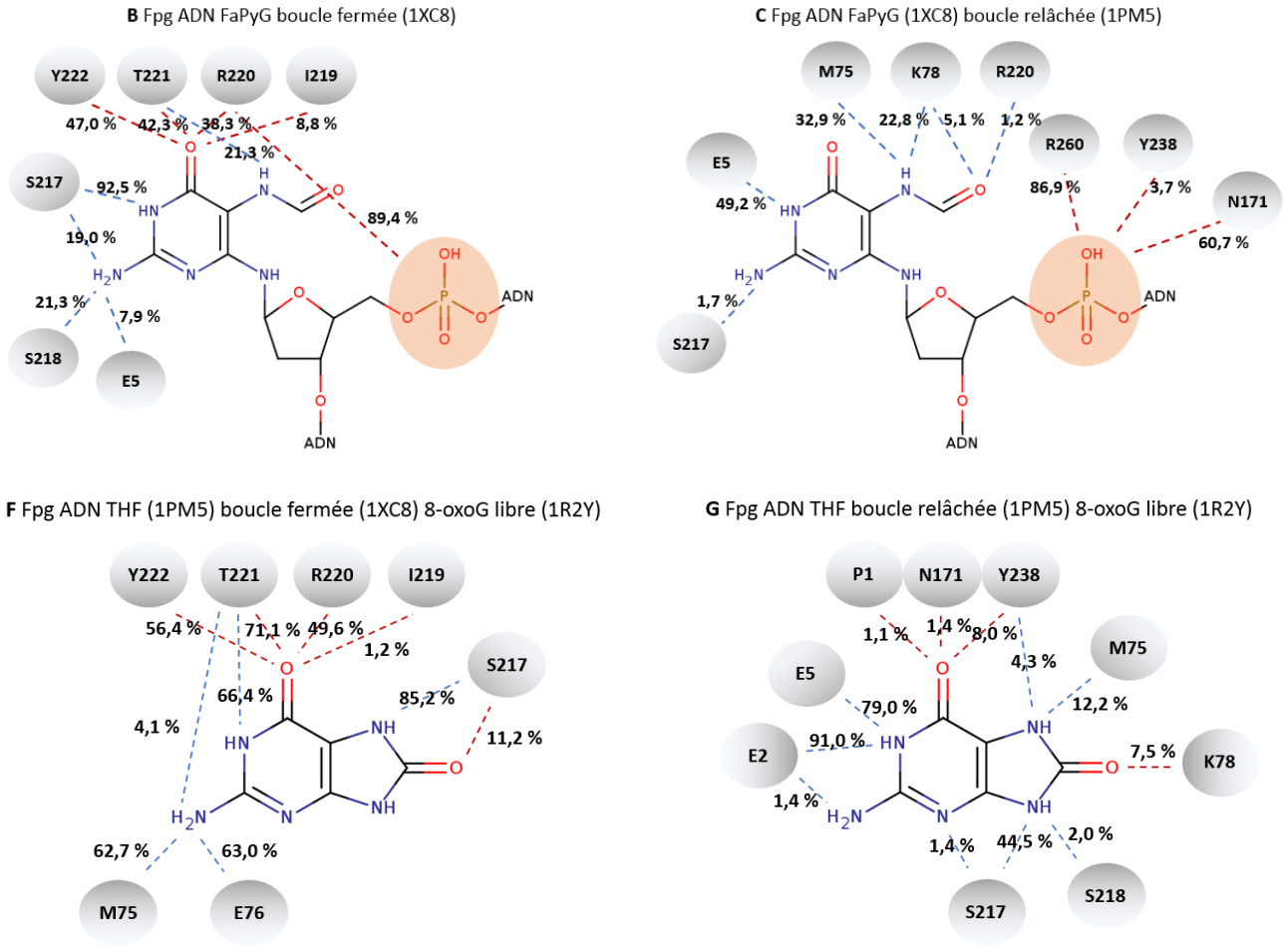


Figure 69 : Liaisons hydrogène créées entre les aa du site actif de la protéine et le substrat (FaPyG) ou produit d'excision (8-oxoG libre) ainsi que leur fréquence (% du temps de simulation)
Les systèmes A, D et E ne sont pas présentés ici car le site actif de *Ll*Fpg dans ces systèmes est vide. **B** et **F**, sont les systèmes dont la boucle LCL est fermée, **C** et **G**, les systèmes dont la boucle LCL est relâchée. **B** et **C**, sont les systèmes comportant le FaPyG et **F** et **G**, sont les systèmes comportant la 8-oxoG coupée. Les liaisons hydrogène sont représentées par des points-tillés colorés en fonction du type de l'atome de la molécule mobilisé pour cette dernière, en rouge s'il s'agit d'un oxygène et en bleu s'il s'agit d'un azote.

D'après la **Figure 69** selon l'état de la boucle LCL et la nature de la molécule présente dans le site actif, les liaisons hydrogène qui se forment entre cette dernière et la protéine sont différentes. On remarque que lorsque la boucle LCL est fermée, les liaisons hydrogène persistent entre l'O6 de la molécule et les aa I219, R220, T221 et Y222 de la boucle LCL (**Figure 69 B** et **F**). Ce n'est plus le cas lorsque la boucle LCL est relâchée, car ces aa sont trop éloignés de la petite molécule (**Figure 69 C** et **G**). On remarque que ces interactions formées entre la boucle LCL et la molécule occupant le site actif n'ont pas les mêmes fréquences d'apparitions dans les deux cas. La T221 semble participer davantage à la stabilisation d'une 8-oxoG libre que d'un FaPyG (**Figure 69 B** et **F**). D'autre part, on remarque que quelques aa tels que P1, K78, N171, Y238 et R260 forment des interactions avec les molécules, en revanche les fréquences des liaisons hydrogène formées entre l'O6 et ces aa sont très faibles (**Figure 69 C** et **G**). L'aa S217 et parfois S218 semblent être importants également pour la stabilisation de la molécule dans le site actif, indépendant de la nature de cette dernière (**Figure 69**). On remarque aussi que lorsque la boucle LCL est fermée, c'est R220 (de la boucle LCL) qui interagit avec le groupement phosphate de la lésion (**Figure 69 B**). Lorsque la boucle LCL est relâchée, ce sont N171 et Y238 ainsi que R260 du ZnF qui forment des liaisons hydrogène avec ce dernier (**Figure 69 C**). Le groupement phosphate de la lésion soit dans tous les cas en interaction avec une Arg (**Figure 69 B** et **C**). Nous remarquons que lorsque la boucle LCL est fermée sur le produit d'excision 8-oxoG, E76 et T221 interagissent avec le N2 de la molécule (**Figure 69 F**), ce qui n'est pas le cas avec le substrat FaPyG (**Figure 69 B**). Nous parlerons plus tard des interactions pouvant avoir lieu entre E76 et T221 et de leurs impacts sur la structure de la boucle LCL. Nous ne remarquons pas d'interaction exclusive à l'un des quatre cas présentés dans la **Figure 69**. En se penchant un peu plus sur le cas de la 8-oxoG libre, on remarque que l'état relâchée de la boucle LCL semble avoir un impact sur l'orientation du produit d'excision dans le site actif. En effet, ce ne sont pas les mêmes atomes qui sont en interaction avec les aa M75 et S217 dans les deux cas (**Figure 69 F** et **G**). La 8-oxoG libre dans le système de Fpg avec la boucle relâchée a changé d'orientation durant la simulation de DM ce qui peut nous amener à penser que la boucle LCL, lorsqu'elle est fermée, possède bien un effet stabilisateur sur la petite molécule libre. À la suite de ce manuscrit, nous concentrerons nos analyses sur les aa I219 à Y222 de la boucle LCL.

Dans le système Fpg libre boucle fermée (1XC8 sans ADN), la boucle LCL et notamment les aa I219, R220 et T221 sont faiblement flexibles au cours de la simulation de DM (**Figure 67 E**) alors que la protéine n'est plus complexée à son substrat, comme si la boucle LCL gardait une emprunte du substrat après qu'il ait été retiré (**Figure 68 D** et **Figure 70**). Le substrat étant absent, la boucle LCL ne devrait plus être stabilisée par les interactions qu'elle formait avec ce dernier (**Figure 62 B**). Cependant, dans ce même système, la boucle LCL n'est pas aussi flexible que dans les systèmes boucle relâchée. Cela

signifie que d'autres interactions, probablement entre la boucle LCL et le reste de la protéine et/ou la boucle LCL et l'ADN pourraient permettre de maintenir un état pseudo stable au sein de la boucle LCL. Il est aussi possible que nous n'ayons pas observé de grand changement conformationnel dans la boucle LCL du système Fpg libre boucle fermée (1XC8 sans ADN) car nous n'avons potentiellement pas exploré des échelles de temps suffisamment grandes où peut être que le passage d'une boucle LCL de l'état fermé à l'état relâché demande un apport énergétique et que nous pourrons donc pas observer un tel changement grâce à des simulations de MD classique mais plutôt grâce à des TMD.

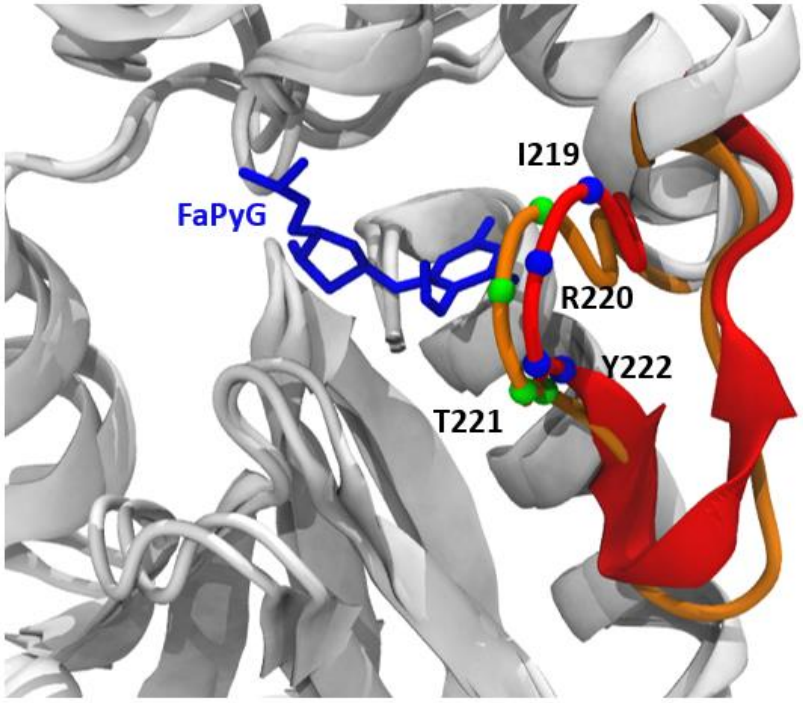


Figure 70 : Comparaison entre les boucles LCL des systèmes Fpg ADN FaPyG boucle fermée (1XC8) et Fpg libre boucle fermée (1XC8 sans ADN)
La boucle LCL fermée (1XC8) est représentée en cartoon rouge et sphères bleues (t=200 ns) et la boucle LCL fermée (1XC8 sans ADN) en cartoon orange et sphères vertes (t=40 ns). Les sphères correspondent aux Cα des aa I219 à Y222. Le FaPyG est représenté en bâtonnets bleus.

IV.1.3.5 Interactions boucle LCL/domaine N-terminal

La boucle LCL interagit *via* plusieurs aa avec le reste de la protéine, comme déjà décrit dans la section **IV.1.1** p. **175**, notamment l'aa L225 flanqué entre l'hélice α 1 et le feuillet β 5 du domaine N-terminal. Cette interaction est conservée dans toutes les structures cristallographiques de *Ll*Fpg et dans tous les systèmes que nous avons étudiés par simulation de DM. La création d'autres interactions entre les aa de la boucle LCL (I219-Y222) et d'autres aa du domaine N-ter (P1 et E76) de la protéine est également observée dans certains systèmes. Pour évaluer l'existence et l'importance de ces

interactions, nous avons mesuré les distances entre les atomes Cd de E76, Cz de R220 et Cb de T221 ainsi que les distances entre les Cα de la P1 catalytique et les Cα des aa I219-Y222 (**Figure 71**).

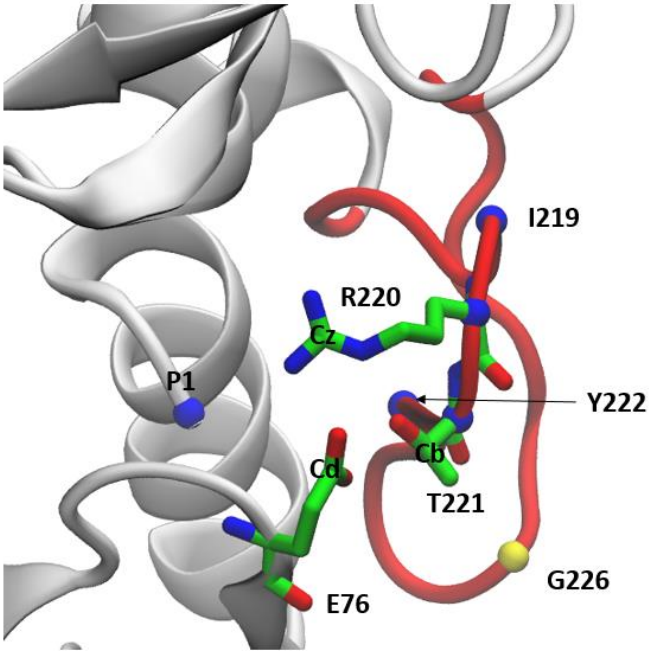


Figure 71 : Site actif de Fpg ADN THF (1PM5) boucle LCL fermée (1XC8) 8-oxoG libre (1R2Y)
La boucle LCL est représentée en cartoon rouge. Les aa E76, R220 et T221 sont représentés en bâtonnets, et les Cα des aa P1, I219, R220, T221, Y222 et G226 sont représentés en sphères bleues et jaune. Les carbones des chaines latérales Cd E76, Cz R220 et Cb T221 sont inscrits sur cette figure, ils sont les atomes utilisés pour mesurer les distances entres les chaines latérales de ces aa présentées dans la **Figure 72**.

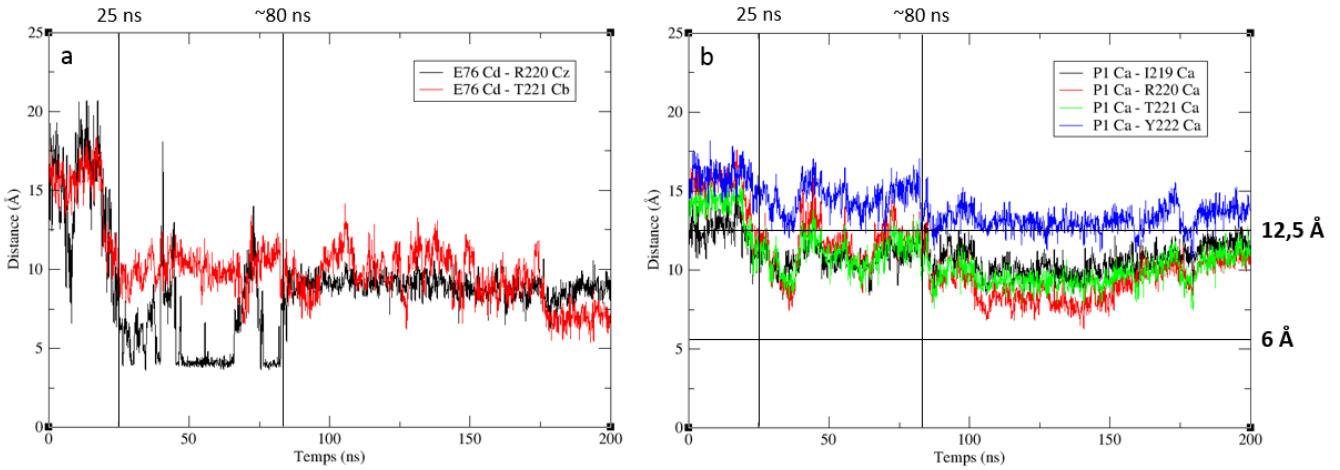
D'après la **Figure 72**, dans les systèmes « boucle fermée » des interactions entre les aa E76 et R220 puis E76 et T221 se forment alternativement ou simultanément (**Figure 72 B.a**, **D.a** et **F.a**). Lorsque les distances entre E76-R220 et E76-T221 sont inférieures à 5 Å, cela traduit la formation de liaisons hydrogène avec pour accepteurs les oxygènes du groupement carboxyle d'E76 et pour donneurs les amines secondaires de R220 ou de l'alcool de T221 sur les chaines latérales. Dans cette conformation, la boucle LCL est proche du site actif et les distances entre les Cα de la P1, et les Cα de I219, R220, T221 et Y222 sont comprises entre 6 et 12,5 Å (**Figure 72 B.b, D.b** et **F.b).** Dautre part, l'aa le plus proche de la P1 catalytique est l'Y222, puis la T221 et ensuite les aa I219/R220 (**Figure 72 B.b, D.b** et **F.b).**

Dans certains systèmes où la boucle LCL est relâchée, on observe la création de l'interaction E76-R220 mais jamais de l'interaction E76-T221. Ces systèmes correspondent à Fpg ADN THF boucle relâchée (1PM5) et à Fpg ADN FaPyG (1XC8) boucle relâchée (1PM5) (**Figure 72 A.a** et **C.a**). La formation de cette interaction E76-R220 s'accompagne de la réduction des distances entre la boucle LCL et la P1 catalytique (Figure 72 A.b et C.b). Dans le cas de Fpg ADN THF boucle relâchée (1PM5), les aa I219, R220 et T221 se situe à ~8 Å de la P1 catalytique. Concernant Fpg ADN FaPyG (1XC8) boucle relâchée (1PM5), seuls les aa I219 et R220 sont à moins de 12,5 Å de la P1 catalytique (**Figure 72 C.b**). Ce rapprochement suggère potentiellement un mouvement de fermeture. Cependant, nous ne retrouvons pas d'interaction similaire dans le système Fpg libre boucle relâchée (1PM5 sans ADN) ni le

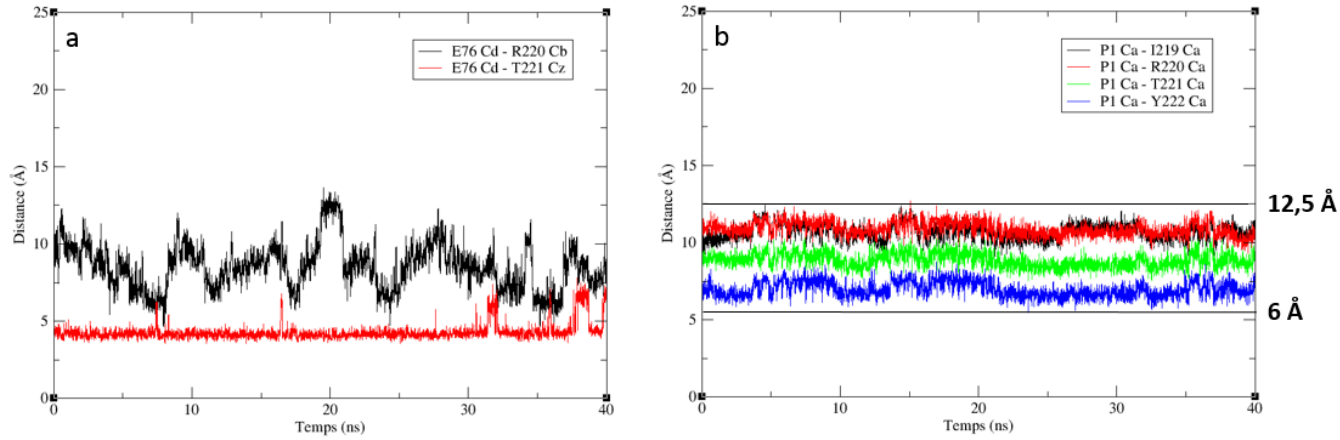
système Fpg ADN THF boucle relâchée (1PM5) 8-oxoG libre (1R2Y) (**Figure 72 E.a** et **G.a**). Cela suggère que l'interaction E76-R220 ne se forme que lorsque la protéine Fpg est complexée à un acide nucléique et que son site actif est vide ou contient un substrat. Cette interaction ne se forme pas lorsque la protéine est libre ou lorsque le site actif contient un produit d'excision.

Dans le système Fpg ADN THF boucle relâchée (1PM5), l'interaction E76-R220 se forme uniquement de 25 à 80 ns, et se défait (**Figure 72 A.a**). Nous verrons plus loin que la R220 contacte d'autres éléments du système après s'être rapprochée du domaine N-terminal. Après ~80 ns, les Cα des aa I219, R20 et T221 se rapproche à moins de 12,5 Å du Cα de la P1 catalytique, mais pas la Y222 (**Figure 72 A.b**). Concernant le système Fpg ADN FaPyG (1XC8) boucle relâchée (1PM5), la R220 interagit avec E76 à partir de ~15 ns, cette interaction persiste jusqu'à ~70 ns (**Figure 72 C.a**).

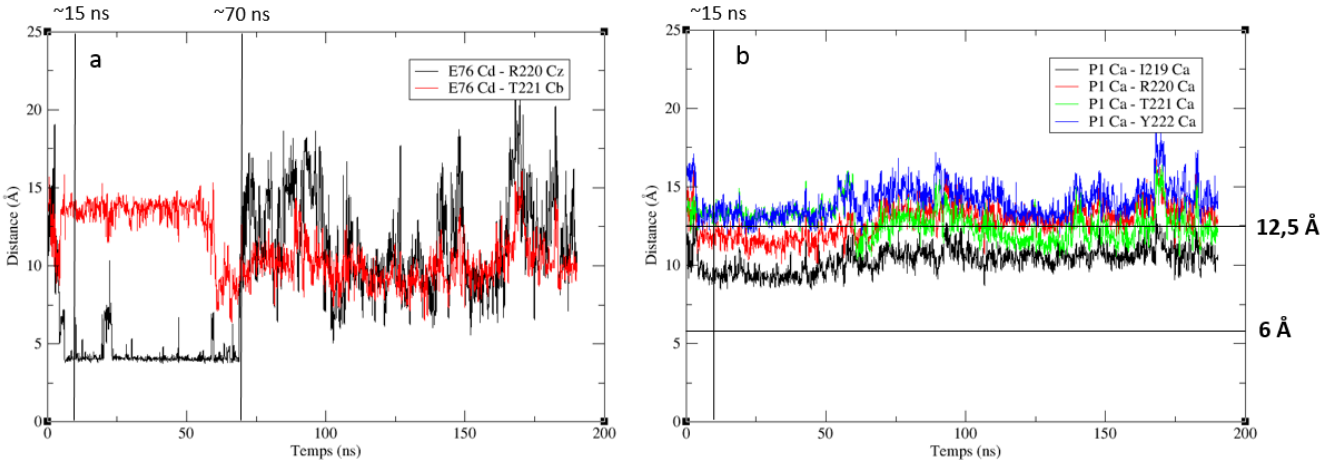
A Fpg ADN THF boucle relâchée (1PM5)



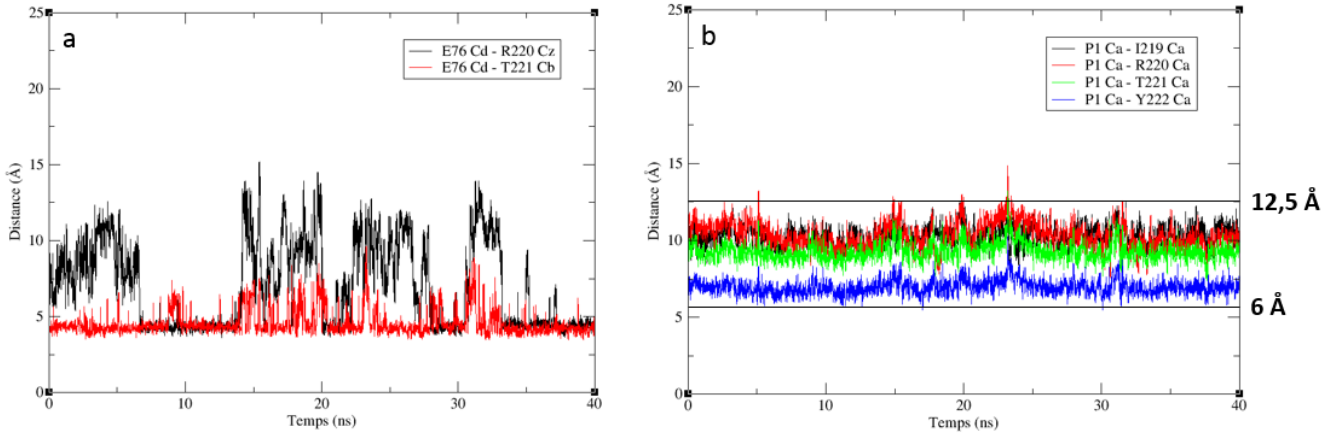
B Fpg ADN FaPyG boucle fermée (1XC8)



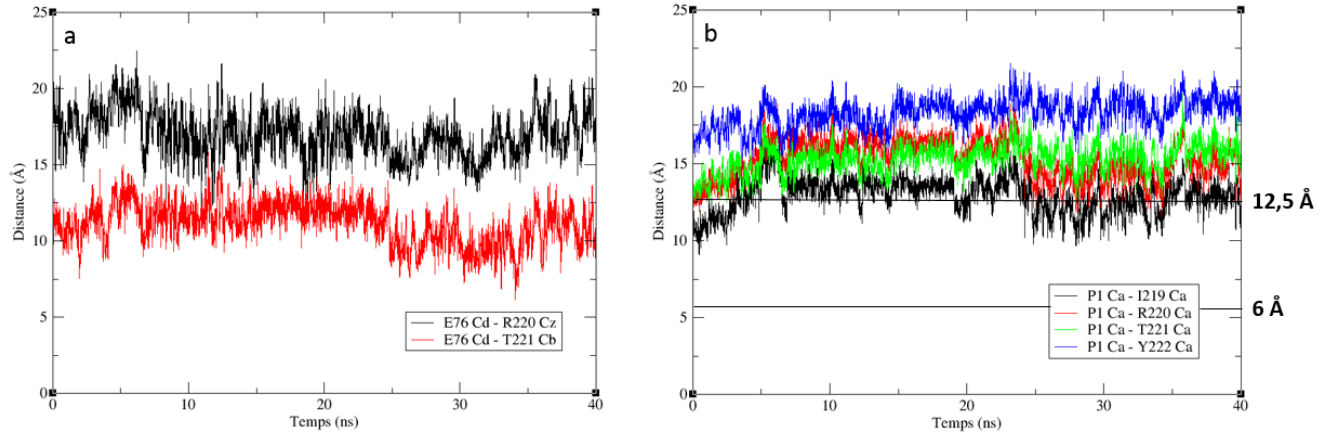
C Fpg ADN FaPyG (1XC8) boucle relâchée (1PM5)



D Fpg libre boucle fermée (1XC8 sans ADN)



E Fpg libre boucle relâchée (1PM5 sans ADN)



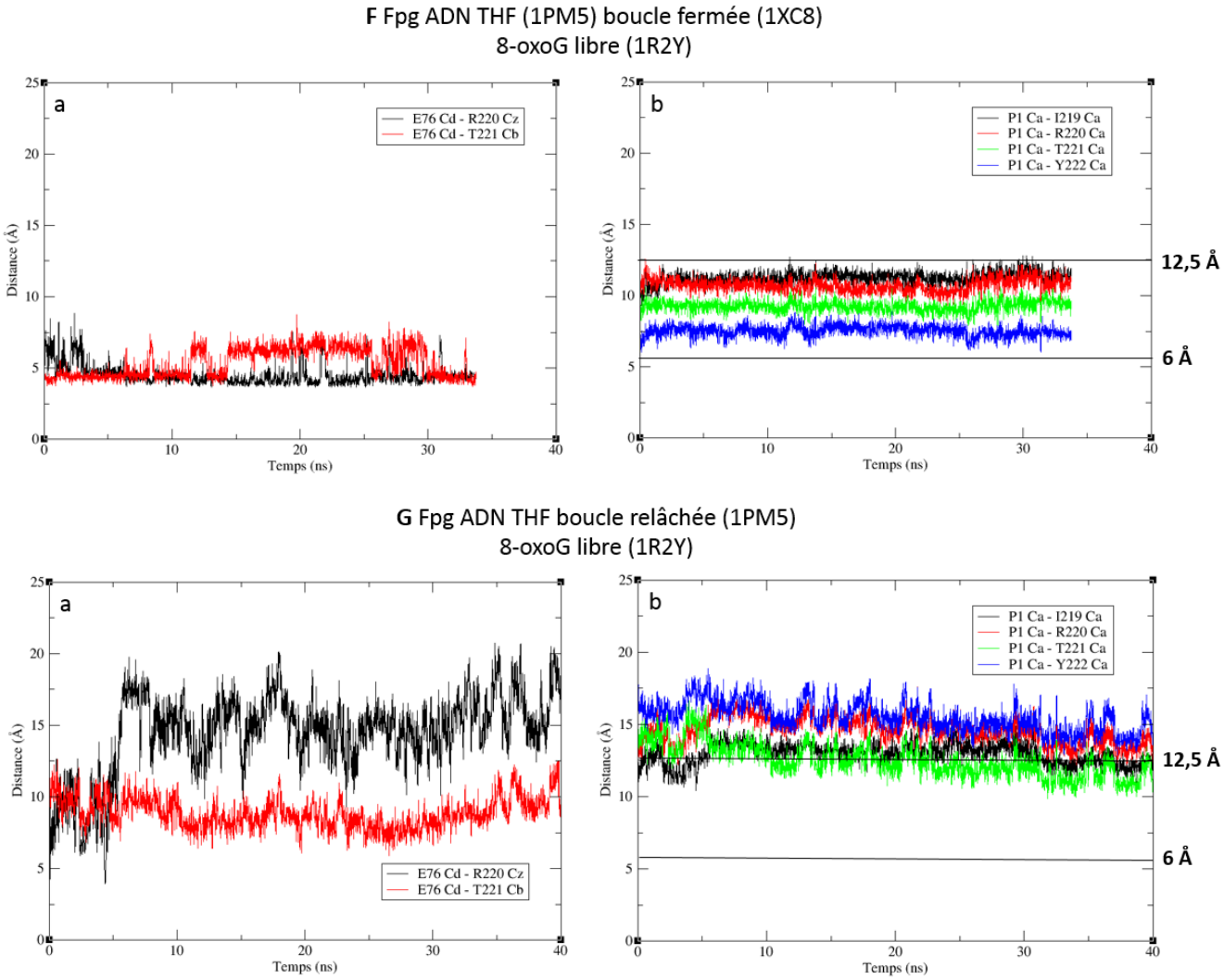


Figure 72 : Distances entre les éléments I219, R220, T221 et Y222 de la boucle LCL et les aa P1 catalytique et E76 du domaine N-terminal
De **A** à **G**, les sept systèmes étudiés dans ce chapitre. **a**) Distances entre le Cd de E76 (noir) et le Cz de R220 (rouge), et le Cb de T221. **b**) Distances entre les Cα des aa I219 (noir), R220 (rouge), T221 (vert) et Y222 (bleu) et le Cα de la P1 catalytique.

Par conséquent, il semblerait que les aa E76, R220 et T221 aient donc un rôle important dans la stabilisation de la boucle LCL en position fermée (**Figure 72 B.a**, **D.a** et **F.a**) et dans certains cas, dans le rapprochement (**Figure 72 A.a, A.b, C.a** et **C.b**) de la boucle LCL relâchée vers le site actif. En parallèle, la boucle LCL semble également interagir avec l'ADN.

IV.1.3.6 Interactions boucle LCL/ADN et boucle LCL/P1

La boucle LCL n'interagit pas uniquement avec la molécule produit/substrat et le domaine N-terminal de la protéine. En effet, la R220 (porteuse de charges positives) peut interagir avec certaines charges négatives du squelette de l'ADN. La **Figure 73** représente l'évolution au cours des simulations

des distances entre la R220 et les phosphores du nucléotide G27 et du THF ainsi qu'avec la P1 catalytique (représentés dans la **Figure 74**).

D'après la **Figure 73**, les interactions THF-R220 et G27-R220 sont créées dans les systèmes Fpg ADN THF boucle relâchée (1PM5) et Fpg ADN FaPyG boucle fermée (1XC8) (**Figure 73 A** et **B**).

Dans le système Fpg ADN THF boucle relâchée (1PM5), les interactions THF-R220 et G27-R220 se forment par alternance entre 25 et 35 ns (**Figure 73 A** et **Figure 74 A** et **B**), puis se rompent au profit d'une nouvelle interaction entre P1 et R220 (**Figure 73 A** et **Figure 74 C**). L'interaction P1-R220 n'est possible que lorsque la boucle LCL est proche du site actif et lorsque celui-ci est vide ce qui n'est pas le cas des autres modèles. Le mouvement qu'effectue R220 peut être relié à la valeur de RMSF élevée de cet aa sur la **Figure 67 A** et la hausse de RMSD de la partie I (216-224) de la boucle LCL de la **Figure 74 A** entre 25 et 80 ns. Sur la **Figure 74 A**, la boucle LCL subit des transformations de 10 à 80 ns de la simulation. Dans la section suivante, nous décrirons les modifications de structures subies par la boucle LCL dans le système Fpg ADN THF boucle relâchée (1PM5).

Dans le système Fpg ADN FaPyG boucle fermée (1XC8) les interactions FaPyG-R220 et G27-R220 sont observées de façon transitoire et alternée (**Figure 73 B** et **Figure 75 A** et **B**). D'autre part, dans le système Fpg ADN FaPyG (1XC8) boucle relâchée (1PM5), nous remarquons la formation sur de courtes durées de l'interaction G27-R220 à partir de 80 ns (**Figure 73 C**). Cependant, elles n'existent pas dans le système de Fpg ADN THF (1PM5) boucle fermée (1XC8) 8-oxoG libre (1R2Y) car les distances oscillent autour de 10 Å (**Figure 73 F**). Ceci peut s'expliquer par le fait que, dans ce modèle, R220 semble interagir exclusivement qu'avec E76 (**Figure 72 F.a**).

Dans les systèmes contenant la 8-oxoG libre, c'est-à-dire Fpg ADN THF (1PM5) boucle fermée (1XC8) 8-oxoG libre (1R2Y) et Fpg ADN THF boucle relâchée (1PM5) 8-oxoG libre (1R2Y), les distances entre la R220 et l'ADN oscillent autour de 9 Å et 18 Å respectivement (**Figure 73 F** et **G**). La R220 reste donc à distance de l'ADN lorsque le site actif contient un produit d'excision. L'absence de ces interactions peut être liée à la différence structurale entre le FaPyG et la 8-oxoG contenue dans le site actif de l'enzyme ou bien au fait que l'un est un substrat (FaPyG) et l'autre un produit d'excision (8-oxoG libre).

Le système Fpg ADN THF boucle relâchée (1PM5) 8-oxoG libre (1R2Y) présente des distances G27-R220 et THF-R220 autour de 15 et 20 Å respectivement (**Figure 73 G**). Il n'y a donc pas de rapprochement de R220 vers l'ADN. Nous avons également noté que les interactions E76-R220 et E76-

T221 n'existaient pas dans ce système (**Figure 72 G.a**) et que la boucle LCL ne se rapprochait pas non plus du site actif (**Figure 72 G.b**). Il semblerait donc que la boucle LCL reste dans un état relâché et que par conséquent, la présence d'un produit d'excision (base libre) et non d'un substrat dans le site actif de *Ll*Fpg ne permet pas un mouvement de fermeture de la boucle LCL.

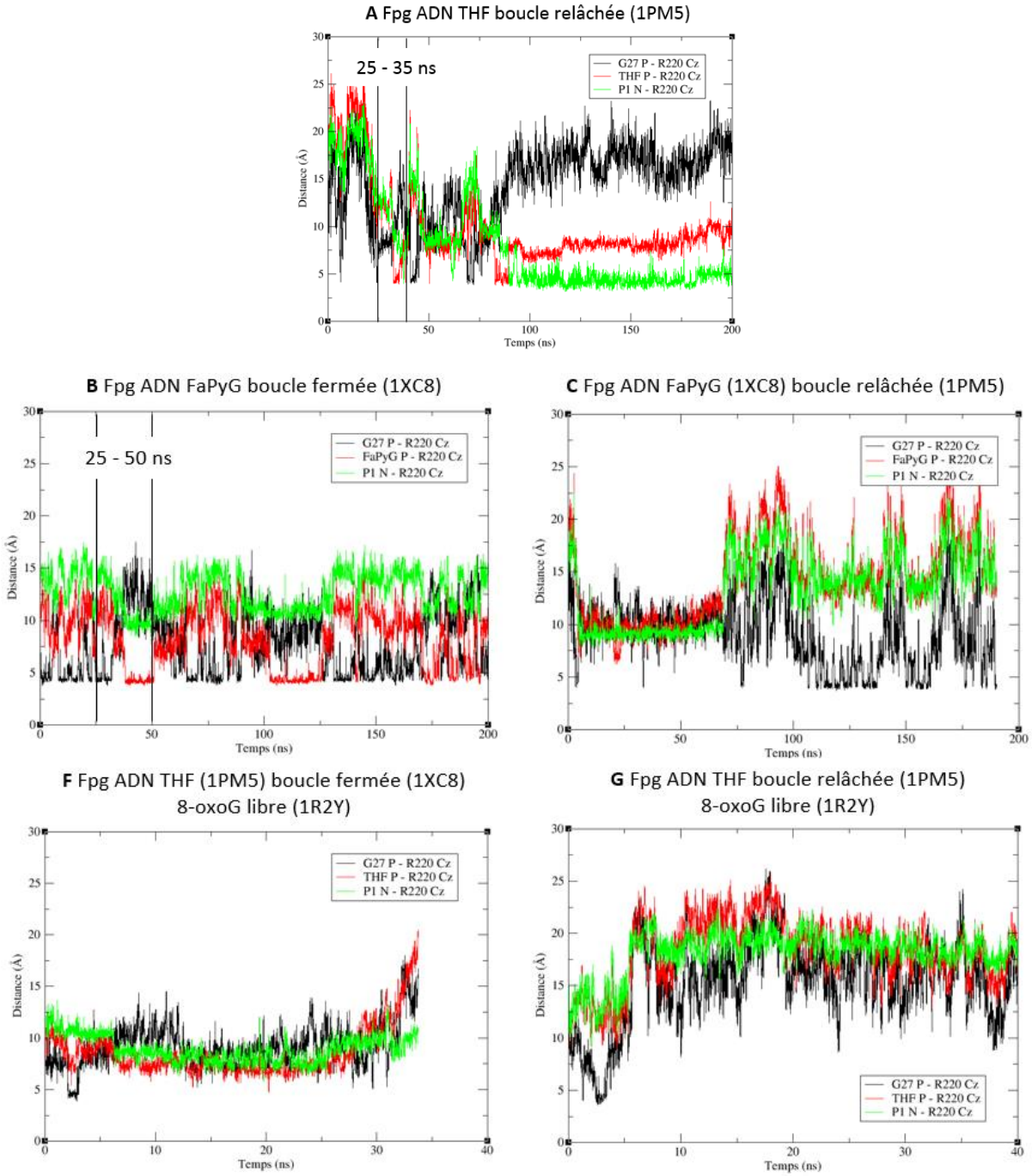


Figure 73 : Interaction entre la R220 de la boucle LCL et l'ADN

Ici, nous représentons les distances entre le Cz de R220 et les phosphores (P) des nucléotides G27 (noir) et THF (rouge) dans tous les systèmes contenant de l'ADN. Pour le système Fpg ADN THF boucle relâchée, une troisième distance est calculée, P1 N-R220 Cz (vert), car c'est le seul système où la protéine est complexée avec un ADN et dont le site actif est vide, et permet donc une interaction entre les aa P1 et R220.

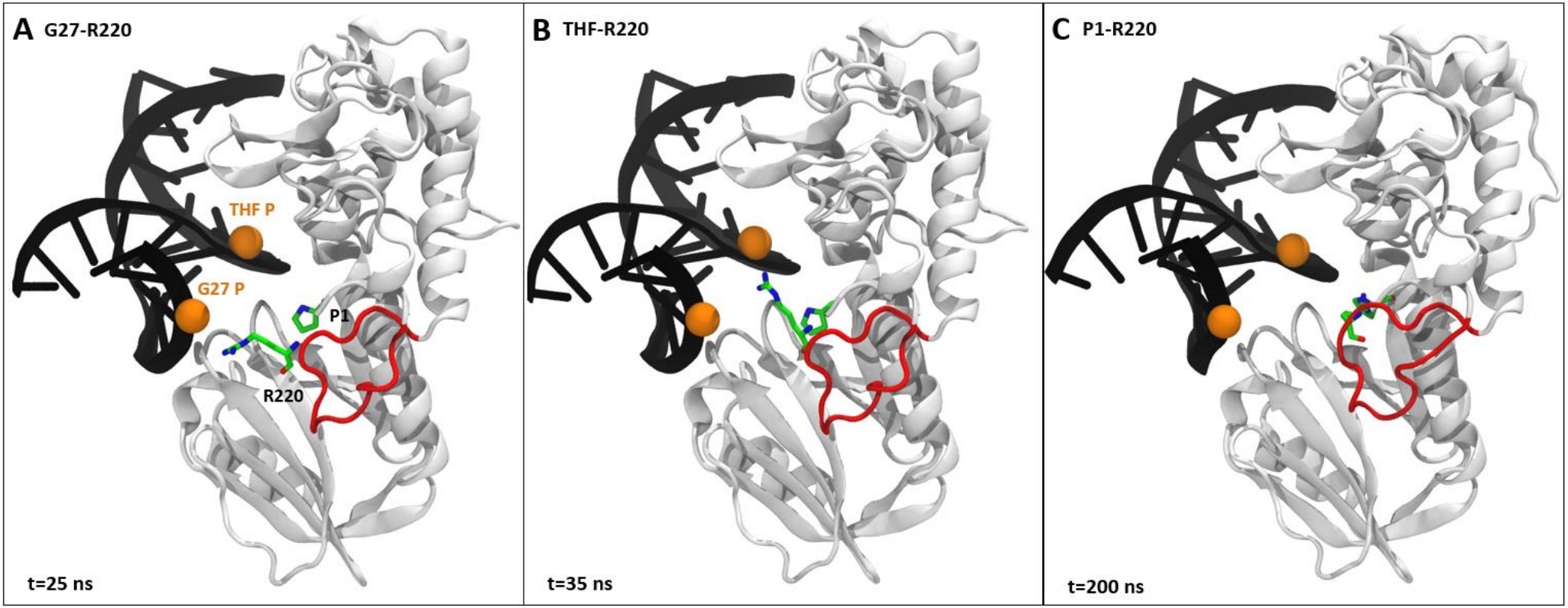


Figure 74 : Interaction de la boucle LCL avec l'ADN via la R220 dans le système Fpg ADN THF boucle relâchée (1PM5)
A) À t=25 ns, **B**) t=35 ns et **C**) t=200 ns. La protéine est représentée en carton blanc, l'acide nucléique en cartoon noir. Les aa P1 et R220 sont représentés en bâtonnets, et les phosphates du nucléotide G27 et du THF en sphères orange.

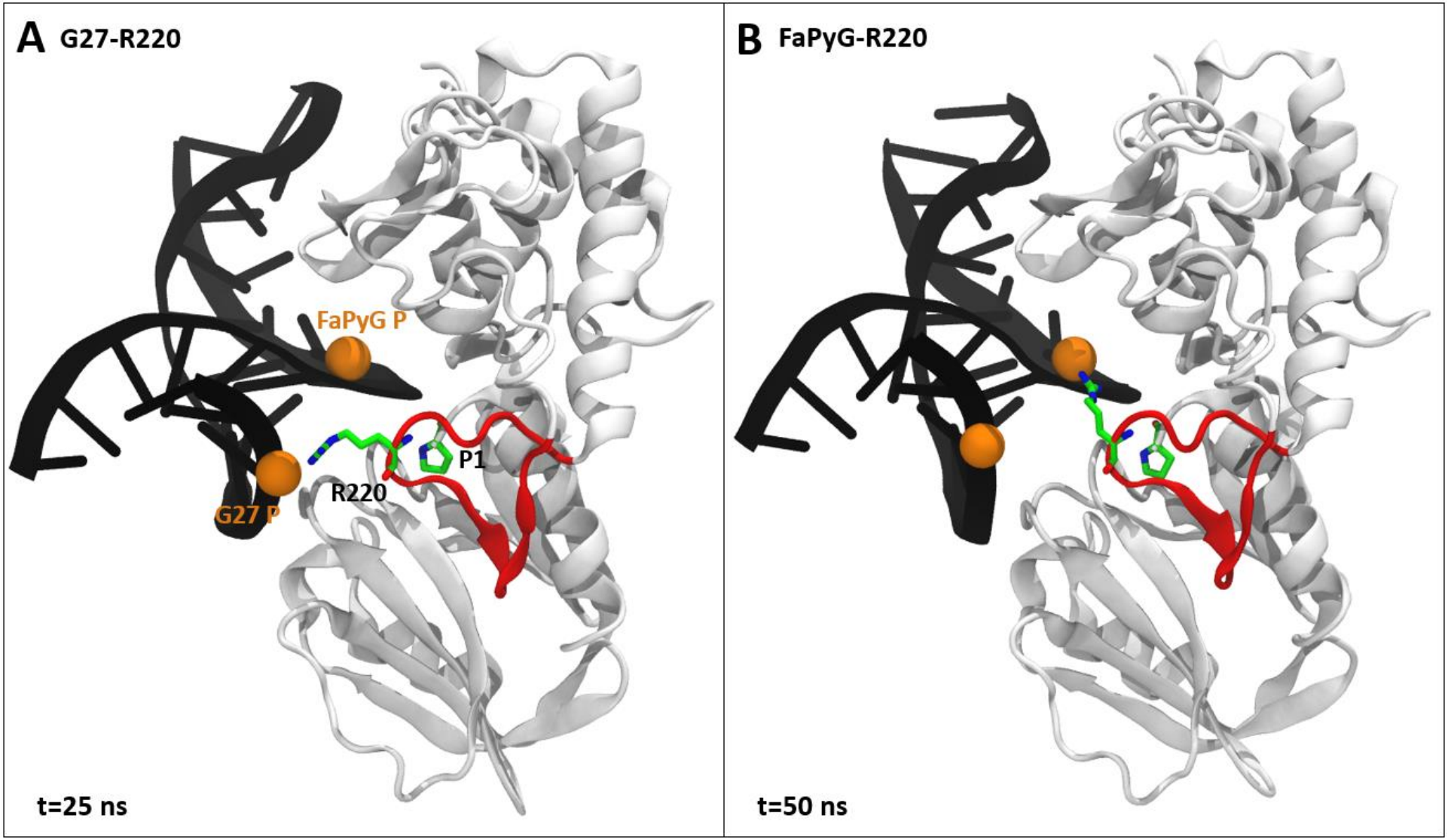


Figure 75 : Interaction de la boucle LCL avec l'ADN via la R220 dans le système Fpg ADN FaPyG boucle fermée (1XC8)
A) À t=25 ns et **B**) t=50 ns. La protéine est représentée en carton blanc, l'acide nucléique en cartoon noir. Les aa P1 et R220 sont représentés en bâtonnets, et les phosphates du nucléotide G27 et du FaPyG en sphères orange.

IV.1.3.7 Un mouvement de pseudo-fermeture

Nous avons vu que la boucle LCL du système Fpg ADN THF boucle relâchée (1PM5) subissait des changements importants de sa conformation (**Figure 67 A** et **Figure 68 A**), et que ces changements étaient accompagnés par la formation des interactions E76-R220 (**Figure 72 A.a**), THF-R220, G27-R220 puis P1-R220 (**Figure 73 A**). Dans cette section, nous allons comparer les structures des boucles LCL des systèmes Fpg ADN THF boucle relâchée (1PM5) et Fpg ADN FaPyG boucle fermée (1XC8) au cours des simulations de DM. La Figure 76 ci-dessous représente la différence structurale des deux boucles LCL sur 200 ns de simulation de DM.

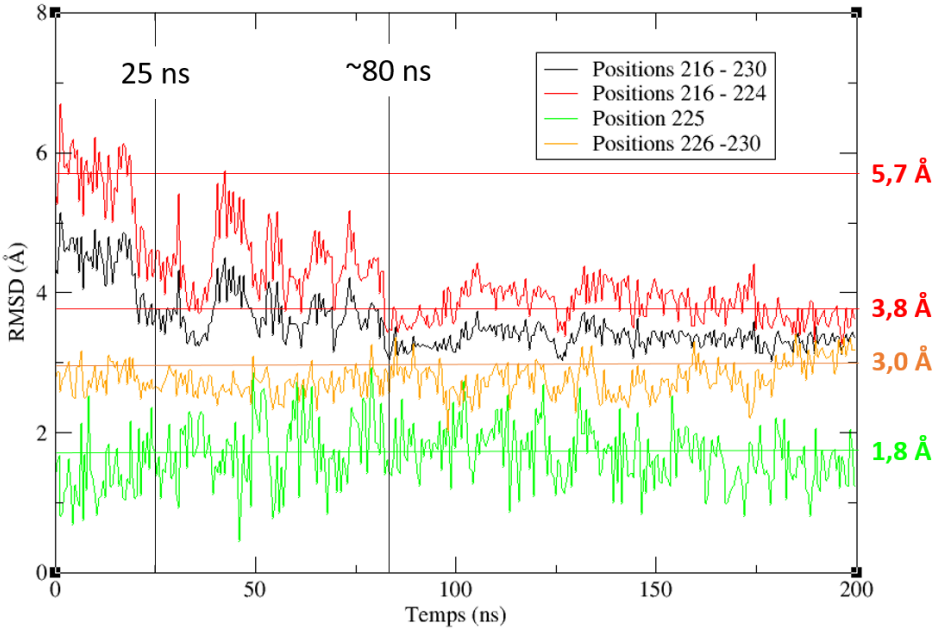


Figure 76 : Différence structurale de la boucle LCL du modèle Fpg ADN THF boucle relâchée (1PM5) au cours du temps sur la référence du modèle Fpg ADN FaPyG boucle fermée (1XC8) à l'équilibre
Nous considérons que le système est à l'équilibre à partir de 10 ns de simulation de MD classique. La RMSD est calculée sur les Cα des aa correspondants. En noir la RMSD totale de la boucle LCL, en rouge la RMSD de la portion 216-224, en vert la RMSD de l'aa L225 et en orange la RMSD de la partie 226-230 de la boucle LCL.

La RMSD de la partie I (216-224) de la boucle LCL diminue au cours de la simulation évoluant de ~ 5,7 Å à 3,8 Å (**Figure 76** courbe en rouge). De plus, à partir de 80 ns de simulation, elle se rapproche sensiblement de la structure de la boucle LCL fermée, d'après cette figure, la partie I de la boucle LCL (216-224) soit plus éloignée que la partie II (226-230) de la structure de référence. L'aa charnière L225 et la partie II de la boucle LCL semblent peu évoluer au cours du temps par rapport à la structure de référence (RMSD ~1,8 et 3,0 Å respectivement). Cependant cette figure ne nous permet pas de définir quels sont les aa qui présentent les plus grands changements structuraux afin de se rapprocher de la conformation de la boucle LCL du système Fpg ADN FaPyG boucle fermée (1XC8). La **Figure 77**

représente la différence structurale individuelle des Cα des boucles LCL des systèmes Fpg ADN THF boucle relâchée (1PM5) et Fpg ADN FaPyG boucle fermée (1XC8). D'après cette figure, on remarque que ce sont les positions I219 à Y222 qui semblent subir le plus de changements conformationnels (la RMSD évolue de 11 à 4 Å dans la **Figure 77**) pour se rapprocher d'une structure similaire à la boucle LCL fermée du système de référence. La L225 semble peu évoluer avec une RMSD ≈ 1,8 Å ce qui suggère à nouveau le rôle charnière de cet aa (**Figure 76** courbe en vert). D'autre part, les G226 et S227 sont des aa de la partie II de la boucle LCL et ne semblent pas se rapprocher significativement de la structure de référence au cours du temps de simulation. Le mouvement de pseudo fermeture correspond donc à une déformation des positions 219-223 de la partie I de la boucle LCL.

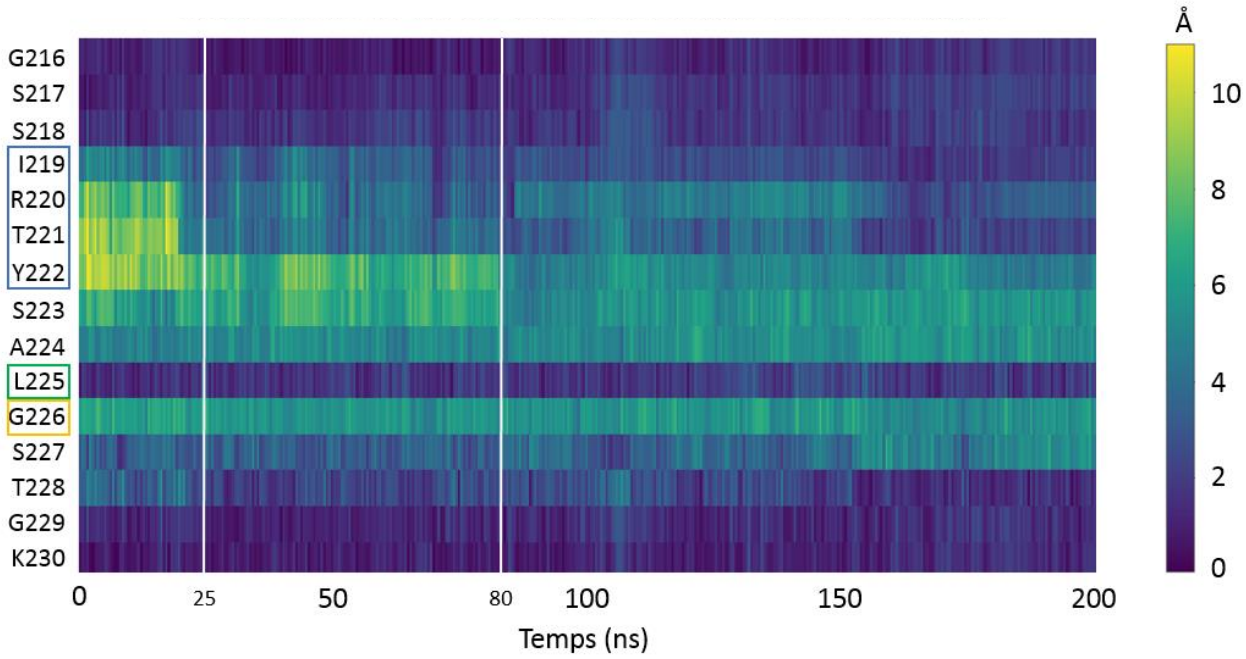


Figure 77 : Différence structurale individuelle des Cα des aa des boucles LCL des systèmes Fpg ADN THF boucle relâchée (1PM5) et Fpg ADN FaPyG boucle fermée (1XC8)
Les éléments importants de la boucle LCL et décrits dans la **Figure 62** p. **30** sont encadrés par différentes couleurs. Les aa formant la couronne d'interactions avec l'O6 du substrat en bleu, l'aa charnière en vert et l'aa dont l'orientation (à l'intérieur ou à l'extérieur du site actif) est un indicateur de l'état de la boucle en jaune.

Sur les sept systèmes produits en simulation de DM, un seul présente un mouvement relativement important de la boucle LCL. En effet, la partie I de la boucle LCL dans le système Fpg ADN THF boucle relâchée (1PM5) semble être flexible (**Figure 67 A**), ce qui traduit des changements conformationnels. Dans la **Figure 78**, nous comparons qualitativement les deux structures de la boucle LCL dans les systèmes Fpg ADN THF boucle relâchée (1PM5) et Fpg ADN FaPyG boucle fermée (1XC8). Comme décrit dans la **Figure 62 B**, la G226 est localisée à l'extérieur du site actif dans la structure cristallographique de *Ll*Fpg présentant une boucle LCL fermée. Elle est située à l'intérieur du site actif lorsque la boucle LCL est relâchée (**Figure 62 A**). Lors de la simulation du système de Fpg ADN THF

boucle relâchée (1PM5), l'orientation de la G226 de l'intérieur jusqu'à l'extérieur du site actif (Figure 78 C et D), sans se déplacer. Ce changement d'orientation est lié au déplacement des aa I219-S223 de la partie I de la boucle LCL. D'autre part, la chaine latérale d'Y222 est orientée à l'extérieur du site actif (**Figure 78 D**) contrairement au système boucle LCL fermée où cette chaine latérale est orientée vers l'intérieur du site actif (**Figure 78 B**). Cela peut expliquer pourquoi l'Y222 est à une distance supérieure à 12,5 Å de la P1 catalytique (**Figure 72 A.b**). De plus, la couronne d'interaction faisant intervenir les N des aa I219 à Y222 n'est pas reformée car le site actif de l'enzyme est vide. La boucle LCL ne rassemble donc pas tous les critères de l'état fermé. Cependant, la structure finale de la boucle LCL de Fpg ADN THF boucle relâchée (1PM5) (Figure 78 D) semble se rapprocher de la structure de la boucle LCL de Fpg ADN FaPyG boucle fermée (1XC8) (**Figure 78 A**).

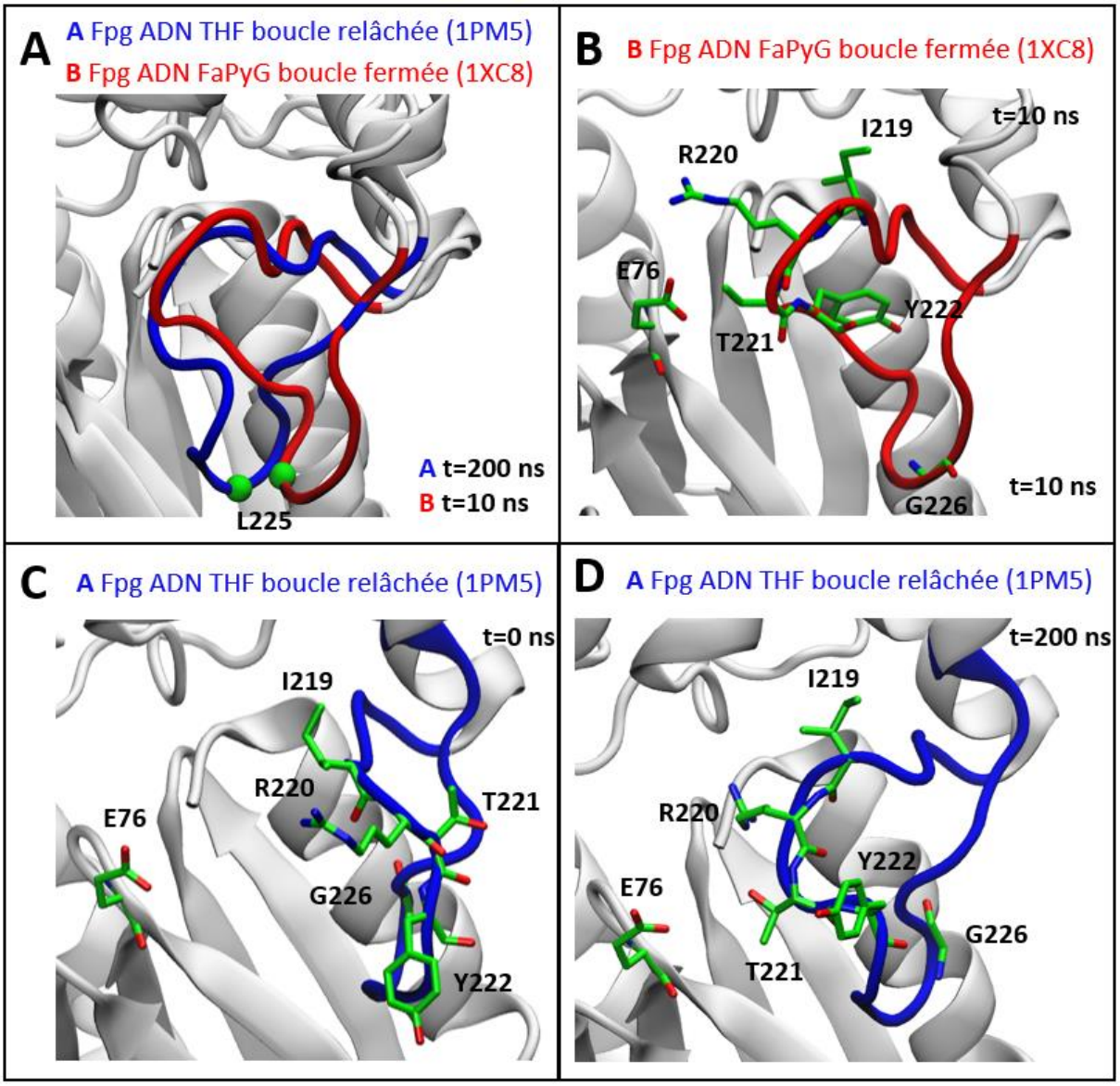


Figure 78 : Comparaison entre les systèmes Fpg ADN FaPyG boucle fermée et Fpg ADN THF boucle relâchée à différents temps de simulation

A) Superposition des deux systèmes, avec Fpg ADN FaPyG boucle fermée à l'équilibre (10 ns) en rouge et Fpg ADN THF boucle relâchée après 200 ns de simulation de DM en bleu. La sphère verte représente

le Cα de L225. **B**) Fpg ADN FaPyG boucle fermée à l'équilibre (10 ns) de simulation, la différence structurale de la boucle LCL entre les deux systèmes est calculée en fonction du temps dans les **Figure 77** et **Figure 78**. **C**) Fpg ADN THF boucle relâchée à 0 ns et **D**) à 200 ns de simulation. Les deux images du bas montrent la différence de structure de la boucle LCL entre 0 ns et après 200 ns de simulation de DM dans le système Fpg ADN THF boucle relâchée. La boucle semble amorcer un mouvement de fermeture, permettant le rapprochement des éléments de la boucle du site actif, notamment de la P1. Cependant, la nouvelle conformation de la boucle LCL (en bas à gauche) ne correspond pas à un état fermé de la boucle (les deux images en haut), car des éléments tels que la chaine latérale d'Y222 orientée à l'extérieur du site actif ou encore l'aa G226 situé à l'intérieur du site actif sont des caractéristiques d'une boucle dans l'état relâchée. Après 200 ns de simulation de DM, le système Fpg ADN THF boucle initialement relâchée présente donc une boucle LCL pseudo-fermée ou pseudo-relâchée.

Dans la simulation du système correspondant à Fpg ADN THF boucle relâchée (1PM5), on remarque d'abord la création par intermittence et alternance la création des interactions E76-R220 (entre ~25 et ~80 ns) (**Figure 72 A.a** p. **195**), G27-R220 (entre ~25 et ~30 ns) et THF-R220 (entre ~30 ns et ~35 ns), qui seront à leur tour rompues au profit d'une nouvelle interaction P1-R220 (à partir de ~90 ns) (**Figure 73 A** p. **197** et **Figure 74 A**-**C** p. **198**). La R220 semble initiatrice de ce mouvement en interagissant successivement avec des aa du domaine N-ter puis avec l'ADN lui-même, ainsi qu'avec la P1 catalytique pour terminer. Le déplacement de la R220 semble entrainer la partie I de la boucle LCL forçant cette dernière à adopter une structure proche de l'état fermé, dans une conformation « pseudo-fermée ». Il semblerait que la présence d'ADN soit importante pour un tel mouvement car il n'est pas observé dans le système Fpg libre (1PM5 sans ADN). De plus, nous n'avons pas observé de mouvement similaire dans les systèmes Fpg ADN FaPyG (1XC8) boucle relâchée (1PM5) ou Fpg ADN THF boucle relâchée (1PM5) 8-oxoG libre (1XC8), suggérant que la présence d'ADN avec un site abasique et un site actif vide permettrait ce mouvement, c'est-à-dire uniquement lorsque la base est coupée, et qu'elle ait quitté le site actif de l'enzyme.

IV.1.3.8 Conclusion

D'après les interactions que la boucle LCL peut former avec la molécule (FaPyG ou 8-oxoG, substrat ou produit) contenue dans le site actif décrites dans la section **IV.1.3.4** p. **189**, les aa I219, R220, T221 et Y222 interviennent dans la stabilisation de la molécule lorsque la boucle LCL est fermée. De plus, la T221 est plus impliquée dans la stabilisation d'une 8-oxoG libre que d'un FaPyG. Ceci est potentiellement lié à la nature de la molécule contenue dans le site actif et provient soit de la différence structurale entre un FaPyG et une 8-oxoG, soit du fait que l'un est encore liée à l'ADN et l'autre non, auquel cas la boucle LCL permettrait également de différencier le substrat du produit d'excision. Pour valider ou infirmer l'une ou l'autre de ces hypothèses, il faudrait réaliser les mêmes

analyses mais sur deux systèmes supplémentaires, un contenant un ADN THF et FaPyG libre, et l'autre un ADN 8-oxoG.

Dans les sections **IV.1.3.5** p. **191** et **IV.1.3.6** p. **195**, nous avons décrit les interactions formées entre la boucle LCL et le domaine N-terminal de la protéine ainsi que l'ADN. Dans le système Fpg ADN THF boucle relâchée (1PM5) 8-oxoG libre (1R2Y), nous n'observons pas les interactions E76-R220, G27-R220, ni THF-R220. Nous n'observons pas non plus de rapprochement de la boucle de la P1 catalytique, la boucle LCL reste donc dans un état relâché tout au long des 40 ns de simulation de DM. On peut alors émettre deux hypothèses, la première est que la boucle LCL n'interagit pas naturellement avec une 8-oxoG libre dans le site actif de *Ll*Fpg (ce qui est en bon accord avec le fait qu'il n'existe pas de structure cristallographique de *Ll*Fpg complexée à un ADN 8-oxoG avec un état « fermé » de la boucle LCL), soit que la boucle LCL n'interagit pas avec une molécule du type produit d'excision comme la 8-oxoG libre dans le site actif. Dans le cas de la seconde hypothèse, la boucle LCL serait donc un stabilisateur de substrat et non de produit d'excision, ce qui pourrait faciliter la sortie de la base après la coupure. La boucle LCL pourrait donc être capable de distinguer les différentes molécules à chaque étape de la catalyse. Pour valider ou infirmer l'une ou l'autre de ces hypothèses, il faudrait réaliser les mêmes analyses mais sur deux systèmes supplémentaires, l'un contenant un ADN THF et FaPyG libre, et l'autre un ADN 8-oxoG.

Dans la section **IV.1.3.7** p. **200**, nous avons fait un lien entre les interactions E76-R220 et E76-T221, puis G27-R220, THF-R220 puis P1-R220 avec un changement conformationnel de la boucle LCL dans le système de Fpg ADN THF boucle relâchée (1PM5). La boucle LCL initie un mouvement de fermeture mais n'accède pas totalement à un état fermé, car des éléments ne remplissent pas les conditions nécessaires (décrites dans la **Figure 62** p. **176**). L'orientation de la G226 évolue bien de l'intérieur jusqu'à l'extérieur du site actif. Nous avons noté que la G226 ne se déplaçait pas, c'est la partie I de la boucle LCL qui se déplace, ce qui permet tout de même à la G226 de changer d'orientation par rapport au site actif. Cependant, le site actif de l'enzyme est vide dans le modèle de Fpg ADN THF boucle relâchée (1PM5), par conséquent la couronne d'interactions impliquants les N des squelettes des aa I219, R220, T221 et Y222 ne peut être reformée. Entre autres, la chaine latérale d'Y222 reste orientée vers l'extérieur du site actif au lieu d'être enfouie. La chaine latérale d'une Tyr est encombrante et peut être que son enfouissement requiert d'autres conditions supplémentaires. Les aa E76 et R220 ainsi que l'ADN semblent importants pour amorcer un mouvement de fermeture de la boucle LCL et le maintien de cet état fermé, état permettant la stabilisation du substrat dans le site actif avant l'activité glycosylase. Des tests d'activités sur des enzymes comportant des mutations ciblées sur les aa E76 et R220 pourraient permettre de vérifier cette hypothèse.

Les interactions boucle LCL/protéine et boucle LCL/ADN pourraient favoriser le rapprochement de la boucle vers le site actif, et donc permettre la stabilisation du contenu dans le site actif de l'enzyme par la boucle LCL. On observe notamment un mouvement de rapprochement de la boucle LCL dans le système Fpg ADN THF boucle relâchée (1PM5), mais pas dans les autres systèmes où cette dernière est relâchée. Cela suggère que ce mouvement de pseudo-fermeture peut se produire spontanément après la coupure et le relargage de la base coupée, lorsque le site actif est vide.

Pour conclure, la boucle LCL est potentiellement un élément de reconnaissance permettant à la protéine de différencier un substrat d'un produit d'excision. Cet élément structural est très flexible, suffisamment pour adopter deux états conformationnels, (i) fermé et (ii) relâché. Le passage d'un état vers l'autre s'effectue grâce à des interactions boucle LCL/protéine, puis boucle LCL/ADN. Nous avons pu observer un mouvement de pseudo-fermeture de la boucle LCL et émettre l'hypothèse que ce dernier se produit lorsque la base a été coupée et une fois que le site actif est vacant. Nous n'avons cependant pas observé un mouvement d'ouverture de la boucle LCL, peut-être parce que celui-ci se produit sur des échelles de temps plus grandes que nous n'avons pas explorées ici ou qu'il ne peut survenir de manière spontanée sans apport énergétique. Dans le prochain chapitre, je décrirai les simulations de TMD^{-1} que nous avons réalisées afin d'étudier la sortie de la base libre après la coupure et d'appréhender l'éventuel rôle de la boucle LCL dans le mécanisme de sortie du produit d'excision.

IV.2 **Chapitre 2 : É**tude de la sortie de la 8-oxoG hors du site actif de *Ll*Fpg

La 8-oxoG est une altération de l'ADN courante, elle apparait suite à l'oxydation d'une guanine. C'est également un substrat d'excision de la protéine Fpg. Ogg1 et Fpg sont toutes les deux des ADN glycosylases bifonctionnelles, elles réalisent deux activités catalytiques (i) Glycosylase et (ii) AP lyase. Ogg1 est une enzyme qui possède un « turnover » (relatif au temps entre deux activités catalytiques) élevé (temps de demi-vie du complexe > 2h) [270, 271], ce qui suggère qu'elle a une tendance à rester inactive tant que le produit d'excision est présent dans le site actif, et bloque ainsi l'activité AP lyase. Ogg1 est inhibée par son propre produit de réaction, ce qui n'est pas le cas pour Fpg [196]. Cela signifie que la protéine Fpg possède potentiellement un mécanisme actif ou passif de libération du produit d'excision après la coupure pour permettre l'activité suivante de l'enzyme, à savoir l'activité AP lyase. Comme décrit dans les sections **I.3.3.4** p. **100** et **I.3.4.4** p. **111** de l'Introduction, la boucle LCL et un élément structural très conservé chez les Fpg et interagit fortement avec le substrat de l'enzyme. Elle pourrait être un élément impliqué dans le phénomène de relargage du produit d'excision. Afin de mieux comprendre ce que devient une base altérée après la coupure, nous avons réalisé plusieurs modèles de deux systèmes de Fpg complexée à un ADN contenant un analogue de site abasique (THF), ainsi que la molécule 8-oxoG libre dans le site actif de l'enzyme. À partir de ces simulations nous avons décrit les modes de sortie du produit d'excision, les aa interagissant avec la base libre, ainsi que le rôle de la boucle LCL dans la libération du produit de réaction.

IV.2.1 Protocole

IV.2.1.1 Construction des systèmes

Nous cherchons à comprendre le rôle éventuel de la boucle LCL ainsi que d'autres éléments potentiels dans la sortie du produit d'excision du site actif. Cependant, nous ne possédons pas de données structurales concernant l'état de la boucle LCL en présence du produit d'excision. En effet, il est possible de « capturer » uniquement le complexe enzyme mutée/substrat (mutants décrits dans le **Tableau 10** p. **119** à la fin de l'Introduction) ou enzyme/analogue de substrat (analogues décrits dans **Figure 30** p. **97** de l'Introduction). Par conséquent, les structures cristallographiques représentent la protéine uniquement avant la coupure lors de la formation de la base de Schiff ou avant l'activité AP lyase. Le moment où le substrat devient un produit d'excision lors de l'activité Glycosylase étant très court, nous ne connaissons pas encore de moyens de le « capturer » et d'obtenir des structures

cristallographiques correspondant à cet état. De plus, comme nous l'avons vu lors du dernier chapitre, la boucle LCL peut posséder deux états, (i) fermé ou (ii) relâché. Nous avons donc créé deux modèles différents de Fpg contenant un ADN THF (analogue de site abasique) et une 8-oxoG libre dans le site actif de l'enzyme.

Les deux systèmes comportent donc la protéine *Ll*Fpg complexée à un ADN THF issue de la structure cristallographique 1PM5 et la 8-oxoG libre dans le site actif dont la position initiale est issue de la structure cristallographique 1R2Y. La boucle LCL étant relâchée dans la structure 1PM5, nous avons copié la boucle fermée de la structure 1XC8 pour réaliser le modèle de Fpg ADN THF (1PM5) boucle fermée (1XC8) 8-oxoG libre (1R2Y). Comme décrit dans le chapitre précédent, le système Fpg ADN THF 8-oxoG libre boucle relâchée est un mélange des coordonnées de 1PM5 et 1R2Y, tandis que le système Fpg ADN THF 8-oxoG libre boucle fermée est un mélange des structures 1PM5, 1XC8 et 1R2Y (**Tableau 14**).

Tableau 14 : Description des 2 systèmes créés pour l'étude de la sortie de la 8-oxoG du site actif de *Ll*Fpg

Nom du système	PDBid(s) d'origine(s)	État initial de la boucle LCL	Ligand dans la structure d'origine	Modifications apportées	Temps de simulation (ns)	Contraintes
Fpg ADN THF (1PM5) boucle fermée (1XC8) 8-oxoG libre (1R2Y)	1PM5, 1R2Y, 1XC8	Fermé	ADN THF	Insertion de la boucle LCL (217-232) fermée de 1XC8 dans 1PM5 et insertion de la 8-oxoG libre de 1R2Y dans le site actif de 1PM5	~35*	Distance de 4 Å maximale entre le N de la P1, et le C1' du THF
Fpg ADN THF boucle relâchée (1PM5) 8-oxoG libre (1R2Y)	1PM5, 1R2Y	Relâché	ADN THF	Insertion de la 8-oxoG libre de 1R2Y dans le site actif de 1PM5	40	

*La production du système Fpg ADN THF 8-oxoG libre boucle fermée s'arrête à ~35 ns après une déformation du complexe ADN protéine déjà décrit dans la section **IV.1.3.1** p. **182** du Chapitre 1 de la partie Résultats.

IV.2.1.2 Simulations de dynamique moléculaire classique

Le protocole de préparation des simulations de DM est identique à celui utilisé dans le Chapitre 1. Les complexes ADN/protéines + produit d'excision sont solvatés et neutralisés par l'ajout d'un solvent explicite et d'ions Cl^- ou Na^+. Une contrainte de distance de 4 Å est imposée entre le C1' du THF et le N de la P1 dans les deux systèmes. Cette contrainte de distance est modélisée par le potentiel harmonique décrit dans la section **III.1.7.2** p. **141** de la partie Méthodologie. Les étapes de minimisation, de thermalisation et d'équilibration sont les mêmes que celles décrites dans la section **III.1.4** p. **136**. Après production de simulation de DM classique des deux systèmes, nous avons réalisé plusieurs simulations de TDM^{-1}. Les structures de départ pour les TMD^{-1} sont choisies toutes les 5 ns à partir des simulations de DM classique, à partir du temps où le système est à l'équilibre. Le temps où le système est à l'équilibre est identifié grâce aux courbes de RMSD (Cα) des systèmes, c'est-à-dire à partir de 10 ns (**Figure 79**). Sur cette figure, on constate aussi que le système Fpg ADN THF boucle relâchée (1PM5) 8-oxoG libre (1R2Y) (**Figure 79 B**) semble globalement plus flexible que le système Fpg ADN THF (1PM5) boucle fermée (1XC8) 8-oxoG libre (1R2Y) (**Figure 79 A**). La stabilité du système entier dépend potentiellement de l'état de la boucle LCL.

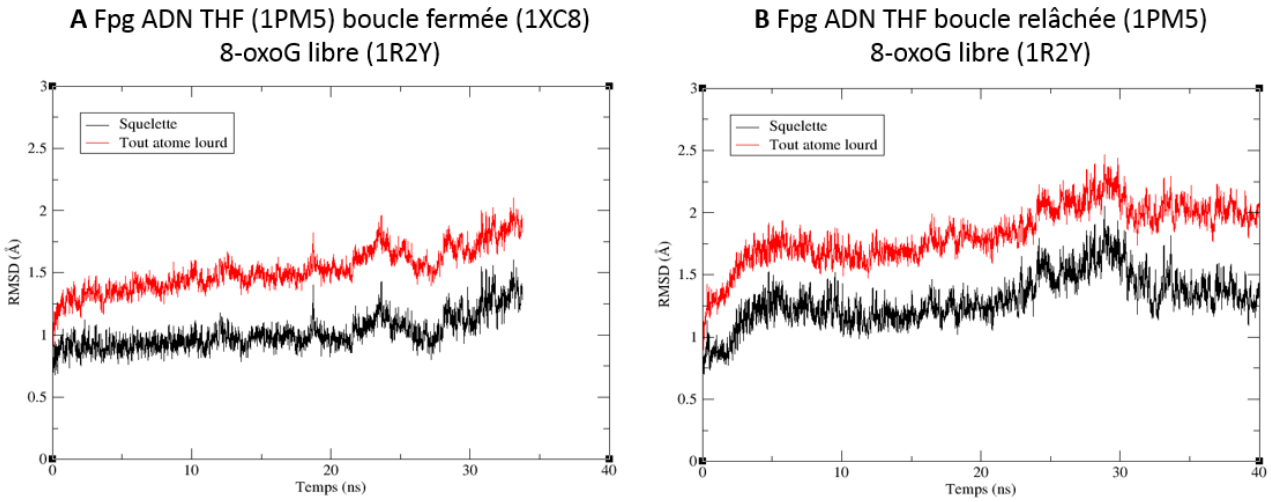


Figure 79 : Courbes de RMSD des deux systèmes produits en dynamique moléculaire classique
A) Système Fpg ADN THF (1PM5) boucle fermée (1XC8) 8-oxoG libre (1R2Y) et **B**) Fpg ADN THF boucle relâchée (1PM5) 8-oxoG libre (1R2Y). En noir la RMSD sur le squelette (C, Cα, N et O) de la protéine et en rouge sur tous les atomes lourds (tous les atomes sauf les hydrogènes).

Pour chacune des simulations de DM, nous contrôlons également la distance entre le THF et la P1 devant respecter la contrainte de 4 Å (**Figure 80 A**) ainsi que les énergies appliquées sur le système. Nous remarquons que la contrainte de distance n'est plus respectée à environ 25 ns, et simultanément, les énergies de contraintes augmentent drastiquement. Cette observation est liée au phénomène de rupture du complexe ADN/protéine, comme déjà évoqué dans la section **IV.1.3.1** p. **182** du Chapitre 1. Par conséquent, nous n'avons exploité la trajectoire de la simulation de ce système que de 10 à 25 ns, ce qui nous fait quatre points de départ pour les TMD^{-1}. Pour avoir la même quantité de données à analyser, nous avons décidé d'exploiter la simulation du système de Fpg ADN THF boucle relâchée (1PM5) 8-oxoG libre (1R2Y) sur le même temps de simulation, de 10 à 25 ns. Nous allons donc par la suite produire et analyser huit simulations de TMD^{-1}, quatre pour chacun des deux systèmes.

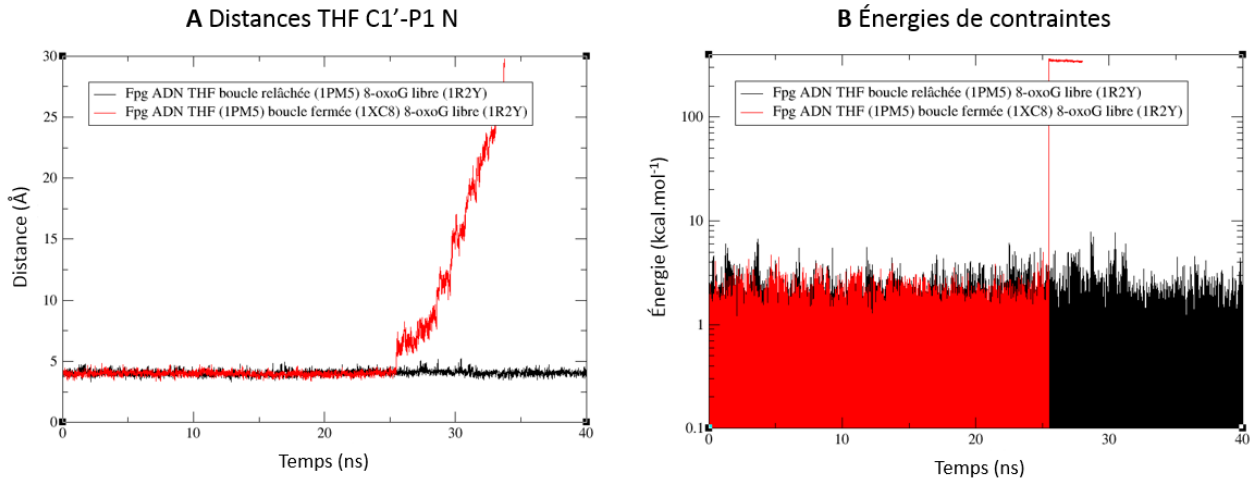


Figure 80 : Contrainte de distance de 4 Å entre la P1 et le THF
A) Distance entre le C1' du THF et le N de la P1 et **B**) Energies de contraintes appliquées aux systèmes pour que la contrainte de distance soit respectée. En noir, les courbes concernant le système Fpg ADN THF boucle relâchée (1PM5) 8-oxoG libre (1R2Y) et en rouge, les courbes concernant le système Fpg ADN THF (1PM5) boucle fermée (1XC8) 8-oxoG libre (1R2Y).

IV.2.1.3 Simulations de dynamique moléculaire ciblée inverse (TMD^{-1})

Les huit structures de départ générées par les simulations de DM classique ont été utilisées pour réaliser autant de TMD^{-1}. Nous avons utilisé une contrainte basée sur la RMSD d'un masque d'atomes du système. La RMSD des atomes du masque augmentera progressivement au cours du temps de simulation de la TMD^{-1}. La RMSD sera calculée sur les atomes du masque suivant : E2, E5 (30 atomes lourds) qui sont des aa catalytiques et la 8-oxoG libre (16 atomes lourds), pour un total de 46 atomes lourds. Les atomes du masque subissent des forces générées par un potentiel harmonique (décrit dans la section **III.1.8** p. **142** de la partie Méthodologie) leur imposant un changement conformationnel. Les TMD^{-1} ont été réalisées sur 2 ns. La RMSD entre la structure courante et la structure initiale du masque augmente de 0,375 Å toutes les 50 ps (soit un taux de 0,0075 $Å.ps^{-1}$)

jusqu'à la valeur seuil de 15 Å, avec la constante de force $k_{TMD} = 10\ kcal.mol^{-1}.\text{Å}^{-2}$. La contrainte de distance de 4 Å entre le C1' du THF et le N de la P1 imposée lors des simulations de DM classique est également appliquée dans les simulations de TMD^{-1}.

Nous avons vérifié les énergies de contraintes (la contrainte de distance entre le THF et la P1 et la contrainte de RMSD) des systèmes en fonction du temps, pour évaluer les TMD^{-1}, sachant que les énergies de contrainte appliquées lors d'une TMD^{-1} ne doivent pas être trop élevées et ne pas altérer la structure secondaire globale de la protéine par exemple (**Figure 81**). Cette figure nous permet aussi d'évaluer l'énergie nécessaire pour déformer les systèmes, nous discuterons cet aspect dans la partie suivante concernant les chemins de sorties de la 8-oxoG libre. D'après la **Figure 81**, il semblerait que pour toutes les simulations de TMD^{-1} quel que soit l'état de la boucle LCL, les plus grands apports en énergie sont nécessaires en toute première partie de simulation. L'énergie ajoutée aux systèmes varie entre 4 et 10 kcal.mol^{-1} de 0 à 50 ps. Au-delà, on constate que l'énergie injectée aux systèmes est modérée voire faible, > 3 kcal.mol^{-1}. La sortie du ligand n'est donc pas un mécanisme passif, car de l'énergie est nécessaire. Par conséquent, il semblerait qu'il y ait au moins une barrière énergétique à franchir. Après 50 ps de simulation, on constate plusieurs pics d'énergie injectée dans les différents systèmes. Ils ne se produisent pas en même temps et ne dépassent pas les 3 kcal.mol^{-1}. Ils peuvent être induits par un phénomène de stagnation de la base libre à la surface de la protéine avec laquelle la 8-oxoG libre forme de nouvelles interactions qu'il faut briser pour lui permettre de se déplacer à nouveau. Plus les interactions sont fortes, plus l'ajout d'énergie au système est requis pour les rompre. On peut donc en déduire que c'est au début des simulations que les interactions les plus fortes sont détruites, et donc que c'est dans le site actif que la base libre forme les plus fortes interactions avec l'enzyme.

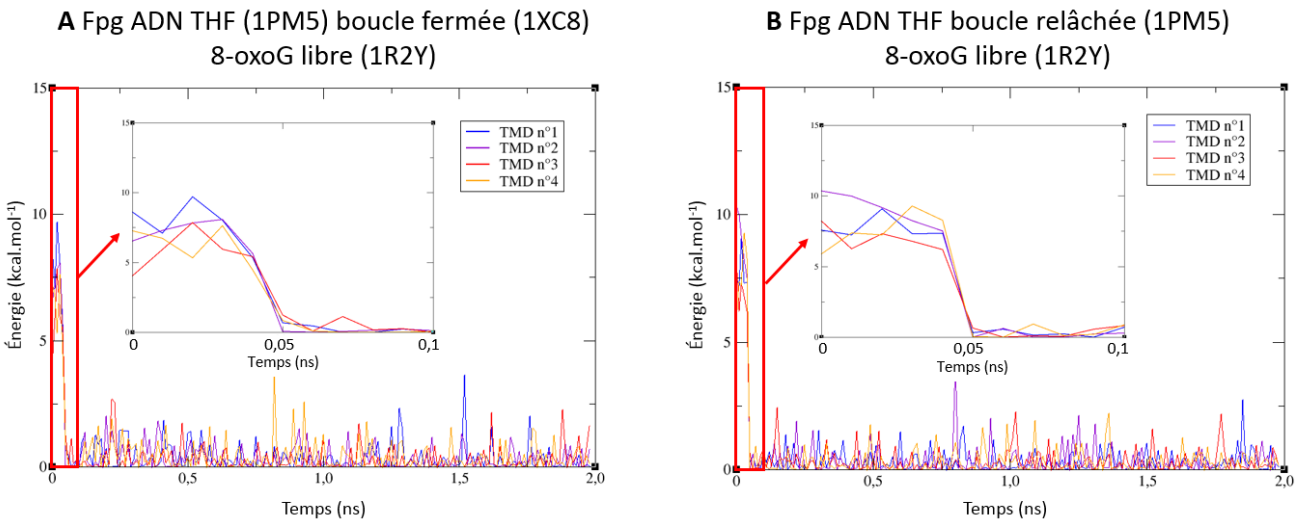


Figure 81 : Energies de contraintes en fonction du temps des TMD^{-1}
Ces énergies de contraintes sont relatives aux huit simulations de TMD^{-1} issues des quatre conformations à 10 (bleu), 15 (violet), 20 (rouge) et 25 ns (orange) des simulations de DM classique des systèmes **A**) Fpg ADN THF boucle fermée (1XC8) 8-oxoG (1R2Y) et **B**) Fpg ADN THF boucle relâchée (1PM5) 8-oxoG (1R2Y). Ces contraintes sont appliquées sur le masque composé des atomes des aa E2, E5 et 8-oxoG. Ces graphes permettent de s'assurer que les énergies de contrainte ne dépassent pas 15 kcal.mol^{-1} et de déterminer l'existence de potentielles barrières énergétiques que doivent franchir les systèmes lors de la sortie du ligand du site actif de l'enzyme.

Par la suite, nous analyserons les déformations des systèmes causées par ces apports en énergie. Nous évaluerons la flexibilité de la protéine grâce à des calculs de RMSF, identifierons les éléments structuraux mobiles et les interactions formées entre la protéine et la base libre au moment de sa sortie.

IV.2.2 Chemins de sortie de la 8-oxoG libre

En sortant du site actif, la base coupée entre en interaction avec différents éléments de la protéine, et entraine de légères déformations de certains éléments de sa structure. Dans les deux sections qui vont suivre, nous représentons les chemins de sortie du produit et nous listons les aa intervenant dans la formation de liaisons hydrogène pour chacun de ces chemins.

IV.2.2.1 Système boucle LCL fermée

Les structures initiales utilisées pour les simulations de TMD^{-1} sont très similaires **(Figure 82**). En effet, comme la boucle LCL est fermée sur la 8-oxoG libre, la base est entre autres stabilisée par la couronne d'interactions entre le O6 de la base libre et les N des aa I219, R220, T221 et Y222 (**Figure 69 F** et **Figure 83**). Ces interactions devront être rompues pour que la base libre puisse sortir du site actif.

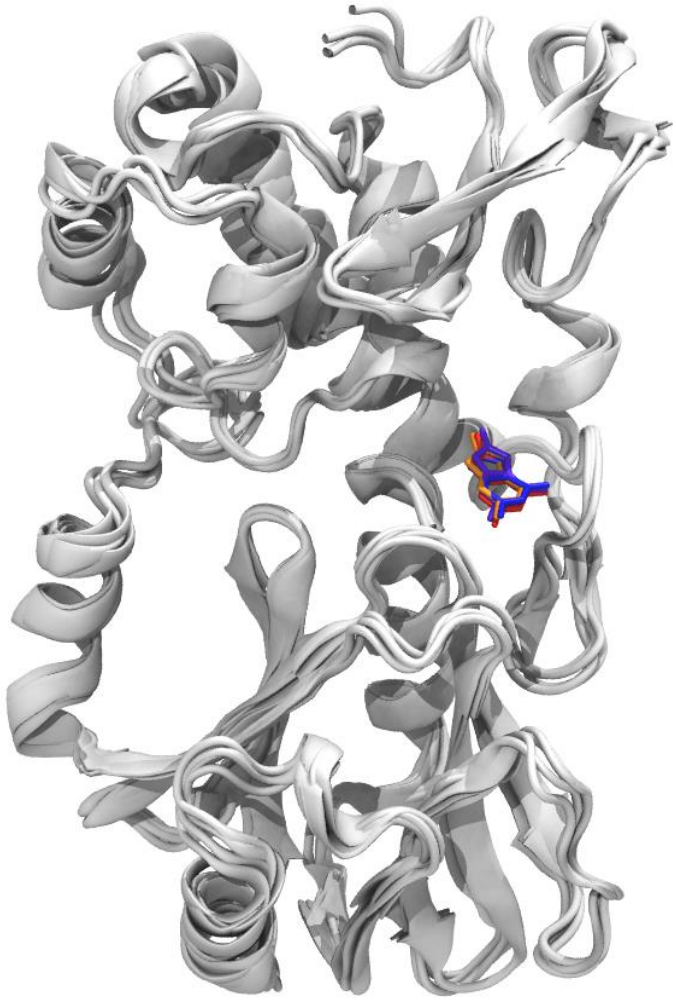


Figure 82 : Conformations initiales de la 8-oxoG libre dans les quatre simulations de TMD^{-1} du système boucle LCL fermée

Dans le système de Fpg ADN THF (1PM5) boucle LCL fermée (1XC8) 8-oxoG libre (1R2Y), la base libre forme beaucoup d'interactions avec la boucle LCL. Cette dernière a donc peu de liberté de mouvement dans la simulation de DM classique, donnant quatre points de départ similaires pour les quatre simulations TMD^{-1}. Sur la figure, les quatre conformations de la protéine sont affichées en cartoon blanc. Les quatre conformations de la 8-oxoG sont représentées en bâtonnets colorés, en bleu la TMD^{-1} n°1, en violet la TMD^{-1} n°2, en rouge la TMD^{-1} n°3 et en orange la TMD^{-1} n°4. L'ADN n'est pas représenté.

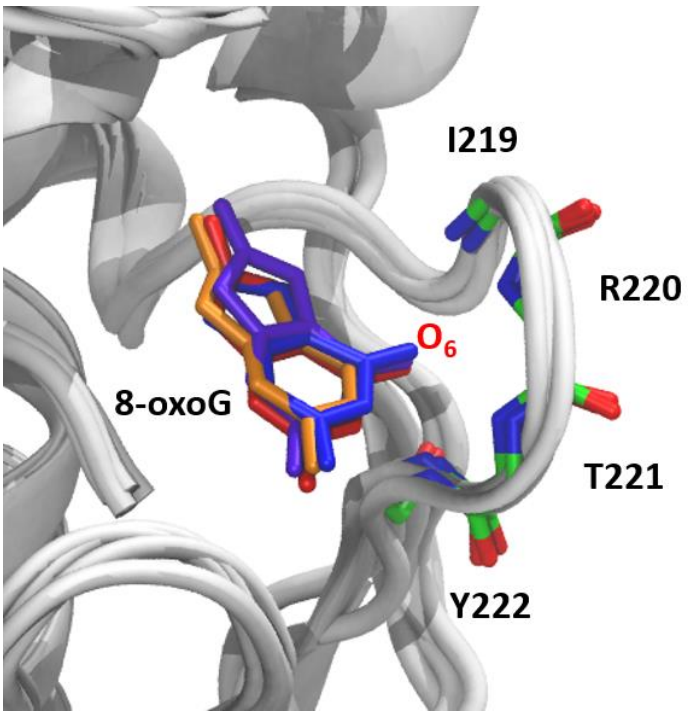
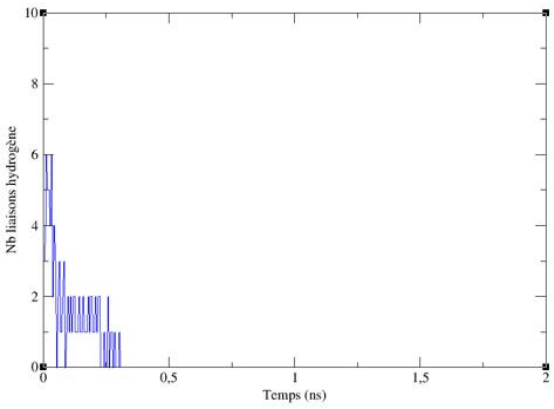
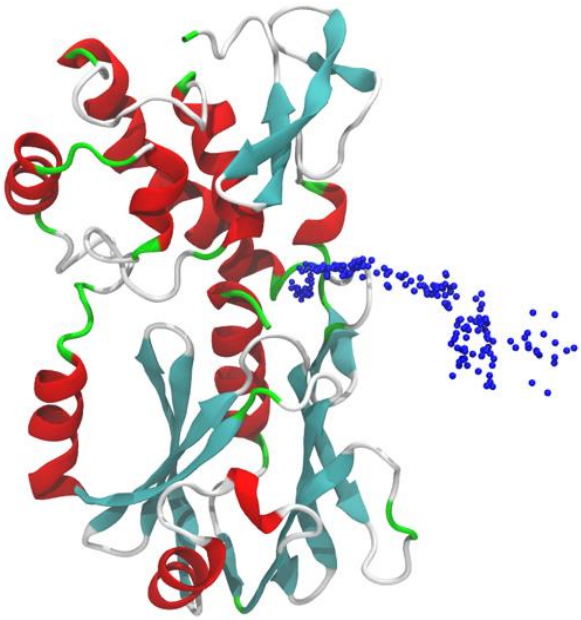


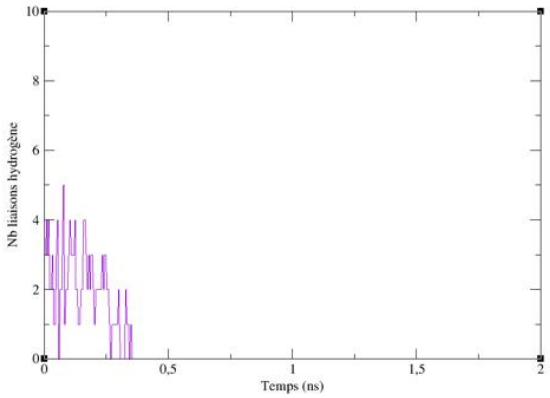
Figure 83 : Conservation de la couronne d'interactions de la boucle LCL dans les structures initiales du système boucle fermée

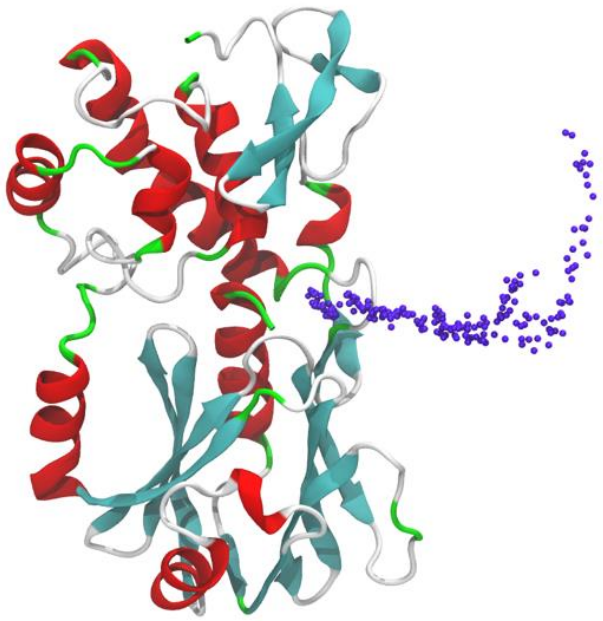
Ces interactions entre la base coupée et la boucle LCL fermée devront être rompues pour permettre la sortie de la 8-oxoG en dehors du site actif. Les quatre conformations de la 8-oxoG sont représentées en bâtonnets colorés, en bleu la TMD^{-1} n°1, en violet la TMD^{-1} n°2, en rouge la TMD^{-1} n°3 et en orange la TMD^{-1} n°4.

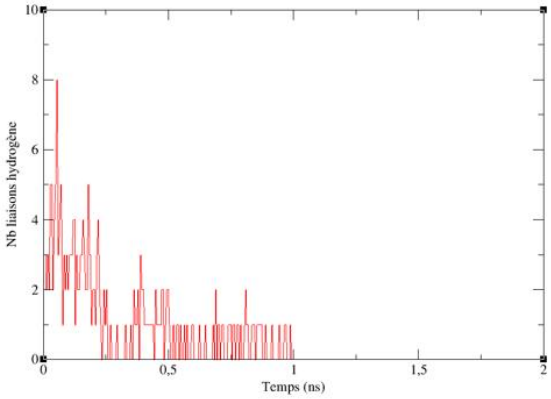
D'après la **Figure 84**, lorsque la boucle LCL est fermée, au moins trois différents scénarii sont possibles. Nous n'avons pas de moyen de discriminer ces scénarii, ils ont tous la même probabilité d'exister. Les chemins de sortie empruntés par le produit d'excision au cours des TMD^{-1} n°1 et 2 sont similaires. La 8-oxoG libre suit la même trajectoire, et on constate aussi que les liaisons hydrogène se rompent très vite et que la base coupée n'interagit plus avec la protéine très tôt dans les simulations, à partir de 0,4 ns (**Figure 84** bleu et violet). Dans ces deux scénarii, la base libre interagit avec le petit sillon de l'ADN avant de diffuser dans le solvant. Dans le scénario n°3 (**Figure 84** rouge), la 8-oxoG libre quitte le site actif et longe le domaine N-terminal de la protéine. Dans cette simulation, la base libre forme des liaisons hydrogène jusqu'à ~1 ns de simulation. Dans le scénario n°4 (**Figure 84** orange), le produit d'excision longe le domaine C-teminal et forme des liaisons hydrogène sur la quasi-totalité du temps de simulation. Les aa intervenant dans les liaisons hydrogène formées au cours des quatre simulations de TMD^{-1} sont reportés dans le **Tableau 15**.

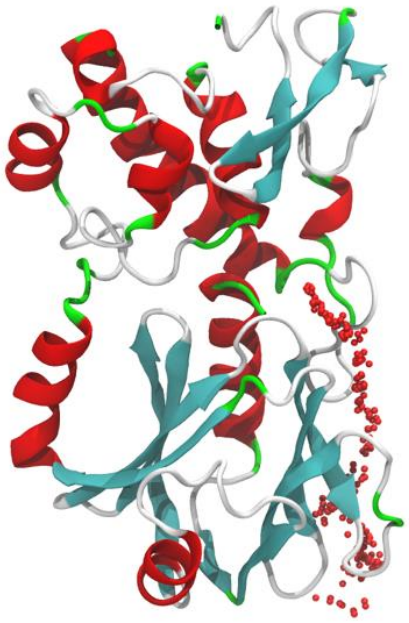





Nb liaisons hydrogène
Temps (ns)
0
2
4
6
8
10
0
0,5
1
1,5
2




Nb liaisons hydrogène
Temps (ns)
0
2
4
6
8
10
0
0,5
1
1,5
2



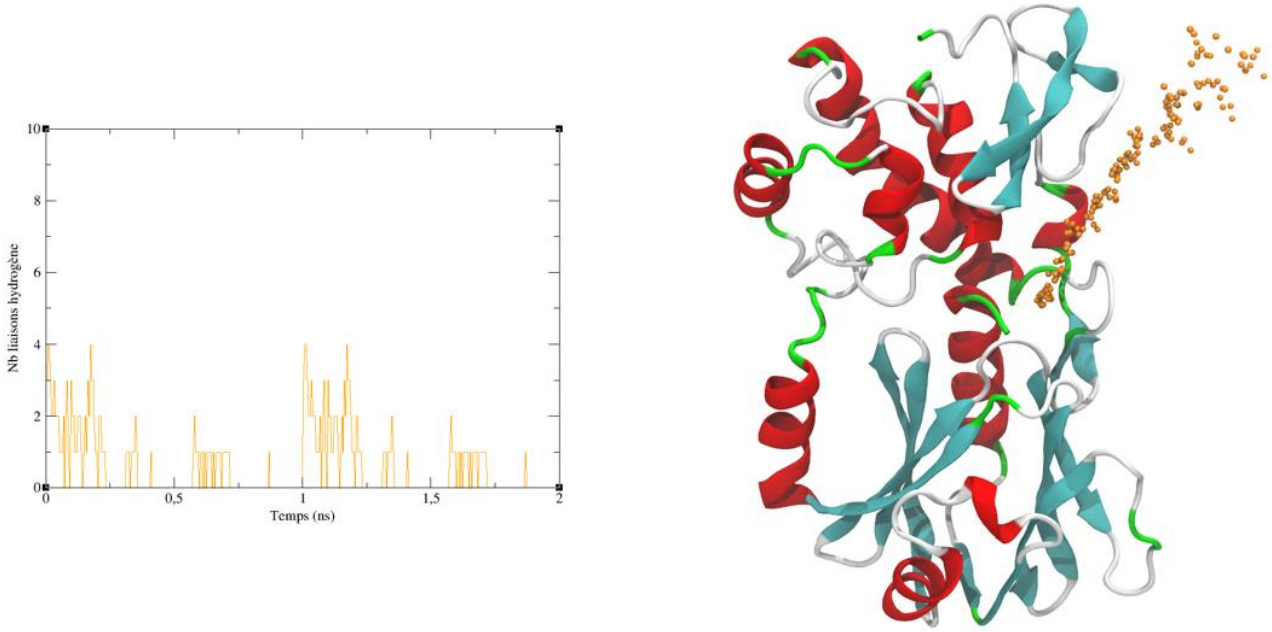


Figure 84 : Description des chemins de sortie empruntés par la base libre dans le système Fpg ADN THF (1PM5) boucle fermée (1XC8) 8-oxoG libre (1R2Y)
À gauche, le nombre total des liaisons hydrogène formées entre la base libre et la protéine au cours du temps. Sur les images à droite, la protéine est en cartoon colorée en fonction de sa structure secondaire, l'ADN n'est pas représenté. Les sphères représentent l'évolution du Centre Géométrique (CG) de la 8-oxoG au cours du temps, à intervalle de 10 ps.

Comme déjà décrit dans le Chapitre 1 de la partie Résultat (p. 175), d'après la simulation de DM classique de Fpg ADN THF (1PM5) boucle fermée (1XC8) 8-oxoG libre (1R2Y), lorsque la boucle LCL de Fpg est fermée, celle-ci forme beaucoup d'interactions avec la base coupée (**Figure 69** F et **Figure 83**). En se concentrant maintenant sur les simulations de TMD^{-1}, nous remarquons que la couronne d'interactions entre le O6 de la 8-oxoG libre et les N des squelettes des aa I219, R220, T221 et Y222 est toujours présente au début des simulations (**Figure 83** et **Tableau 15**). Dans les quatre scénarii, la base libre interagit avec au moins un des aa catalytiques P1, E2 et E5, ainsi qu'avec les aa S217, T221 et Y222 de la boucle LCL (**Tableau 15**). Dans trois scénarii sur quatre, l'aa intercalant de l'ADN M75 entre aussi en interaction avec la 8-oxoG libre lors de sa sortie. Il en est de même avec l'aa E76. L'aa Y238 forme des liaisons hydrogène avec la base libre dans deux scénarii sur quatre. D'autres aa interagissent également avec la 8-oxoG libre dans au moins un des quatre scénarii tel que les aa R16, Q20, K78, Y79, L81, A82, A99, D100, S223, L225, Y240, E243 et K244. Dans le scénario issu de la TMD^{-1} n°3, la base libre semble interagir avec 18 aa différents, cela est dû au fait qu'elle longe le domaine N-terminal de la protéine après sa sortie du site actif dont la M75 qui est un aa impliqué dans le remplissage du vide laissé par la base endommagée en position extra-hélicale dans le site actif de l'enzyme.

Tableau 15 : Aa formant des liaisons hydrogène avec la 8-oxoG libre lors de sa sortie du site actif de *Ll*Fpg ADH THF (1PM5) boucle fermée (1XC8)
Les aa retrouvés dans les quatre scénarii sont soulignés. Les aa sont colorés en fonction de leur rôle dans la protéine, en rouge les aa catalytiques, en vert les aa intercalant de l'ADN, en bleu les aa de la boucle LCL et en noir les autres.

	Aa
TMD^{-1} n°1	**E5, M75, E76, S217, R220, T221, Y222, Y238**
TMD^{-1} n°2	**P1, M75, E76, S217, T221, Y222**
TMD^{-1} n°3	**P1, E2, R16, Q20, M75, E76, K78, Y79, L81, A82, A99, D100, S217, R220, T221, Y222, S223, L225**
TMD^{-1} n°4	**P1, E5, S217, R220, T221, Y222, Y238, K240, E243, K244**

Dans tous ces scénarii, la boucle LCL subit des déformations, car elle interagit fortement avec la 8-oxoG libre, ces déformations sont décrites dans la section **IV.2.2.3** (p. **224**).

IV.2.2.2 Système boucle LCL relâchée

Lors de la simulation de DM classique du système de Fpg ADN THF boucle relâchée (1PM5) 8-oxoG libre (1R2Y), le produit d'excision ne forme que peu d'interaction avec la boucle LCL (**Figure 69 G**). Par conséquent, la base coupée est mobile au cours du temps dans la simulation de DM classique. Dans les structures issues de la DM classique utilisées pour les simulations de TMD^{-1}, la position et l'orientation de la 8-oxoG n'est donc pas la même (**Figure 85**).

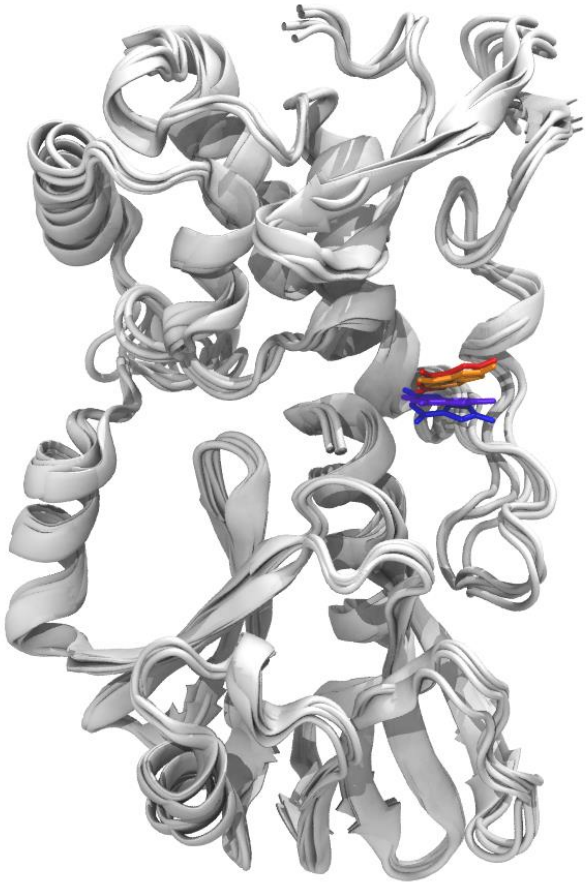
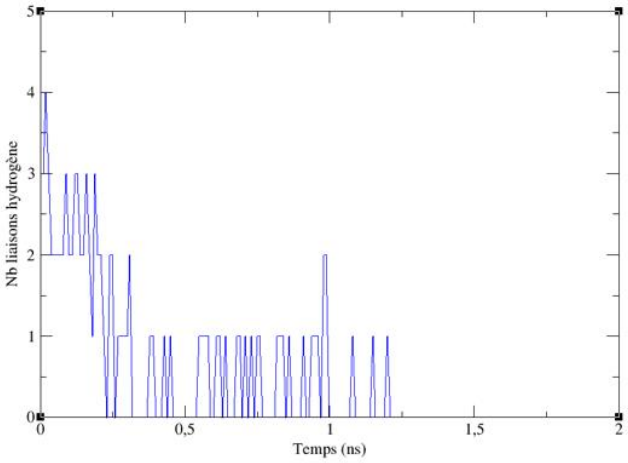
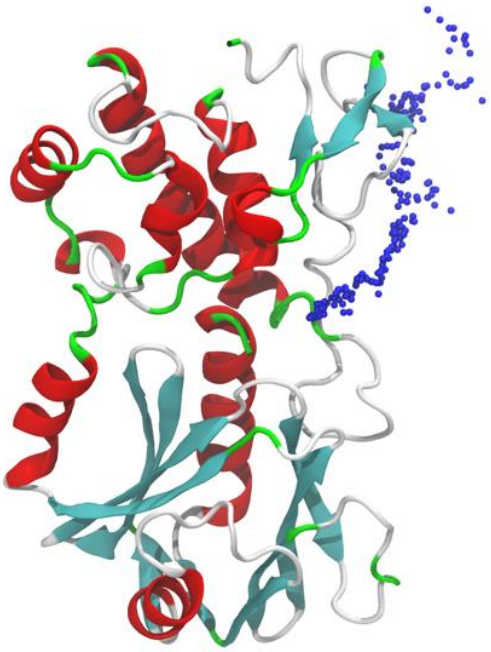


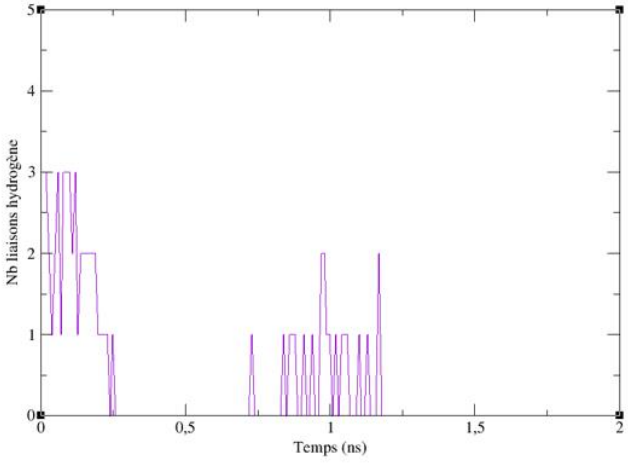
Figure 85 : Conformations initiales de la 8-oxoG libre dans les quatre simulations de TMD^{-1} du système boucle LCL relâchée
Dans le système de Fpg ADN THF boule relâchée (1PM5) 8-oxoG libre (1R2Y), la base libre n'est pas en interaction avec la boucle LCL. Cette dernière a donc plus de liberté de mouvement dans la simulation de DM classique, donnant quatre points de départ différents pour les quatre simulations TMD^{-1}. Sur la figure, les quatre conformations de la protéine sont affichées en cartoon blanc. Les quatre conformations de la 8-oxoG sont représentées en bâtonnets colorés, en bleu la TMD^{-1} n°1, en violet la TMD^{-1} n°2, en rouge la TMD^{-1} n°3 et en orange la TMD^{-1} n°4. Dans la suite de ce manuscrit, nous garderons ce code couleur. L'ADN n'est pas représenté.

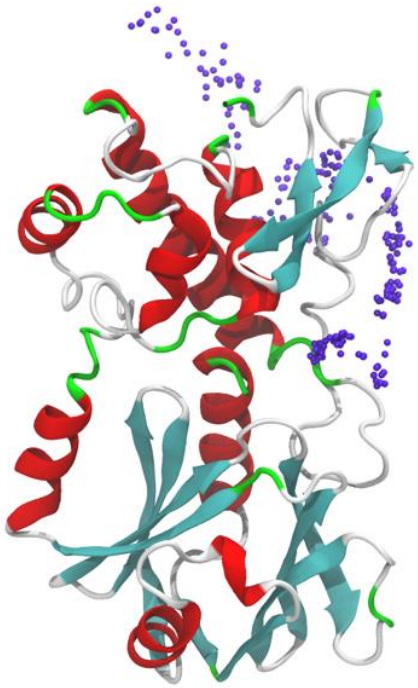
La **Figure 86** représente les chemins de sortie de la 8-oxoG dans les quatre scénarii produits par la TMD^{-1} ainsi que le nombre de liaisons hydrogène formées entre le produit d'excision et la protéine au cours du temps. Le nombre de liaisons hydrogène diminue au cours du temps dans les quatre scénarii, montrant que la base coupée s'éloigne bien de la protéine et que les quatre simulations modélisent bien des phénomènes de sortie de la base du site actif. Dans certains scénarii, comme les TMD^{-1} n° 2, 3 et 4 (**Figure 86** violet, rouge et orange à gauche), les liaisons hydrogène se rompent pendant une durée d'environ 0,5 ns et se reforment. Cela peut être expliqué par le déplacement de la base libre à la surface de la protéine (**Figure 86** violet, rouge et orange à droite). Dans ces trois cas, on a d'abord un éloignement suffisant entre le produit d'excision et la protéine pour ne plus permettre la création des liaisons hydrogène puis un rapprochement permettant à nouveau la formation de ces interactions entre les deux entités. Dans toutes les simulations, les liaisons hydrogène sont définitivement rompues après 1,5 ns. Lorsque nous regardons l'évolution de la position du CG de

la base libre lors de sa sortie du site actif, nous remarquons que les quatre TMD^{-1} modélisent quatre chemins de sortie différents. Dans les scénarii des TMD^{-1} n°1 et 2 (**Figure 86** bleu et violet), la base libre longe le domaine C-terminal de la protéine. Dans le scénario de la TMD^{-1} n°3 (**Figure 86** rouge**)**, le produit d'excision diffuse le long du petit sillon de l'ADN et continue sa trajectoire dans le solvant. Dans le scénario de la TMD^{-1} °4 (**Figure 86** orange), la 8-oxoG libre longe le domaine N-terminal de la protéine, mais n'interagit pas avec la M75 de la triade d'intercalation de l'ADN (**Tableau** 16).


Nb liaisons hydrogène
Temps (ns)
0
0,5
1
1,5
2
1
2
3
4
5




Nb liaisons hydrogène
Temps (ns)
0
0,5
1
1,5
2
1
2
3
4
5



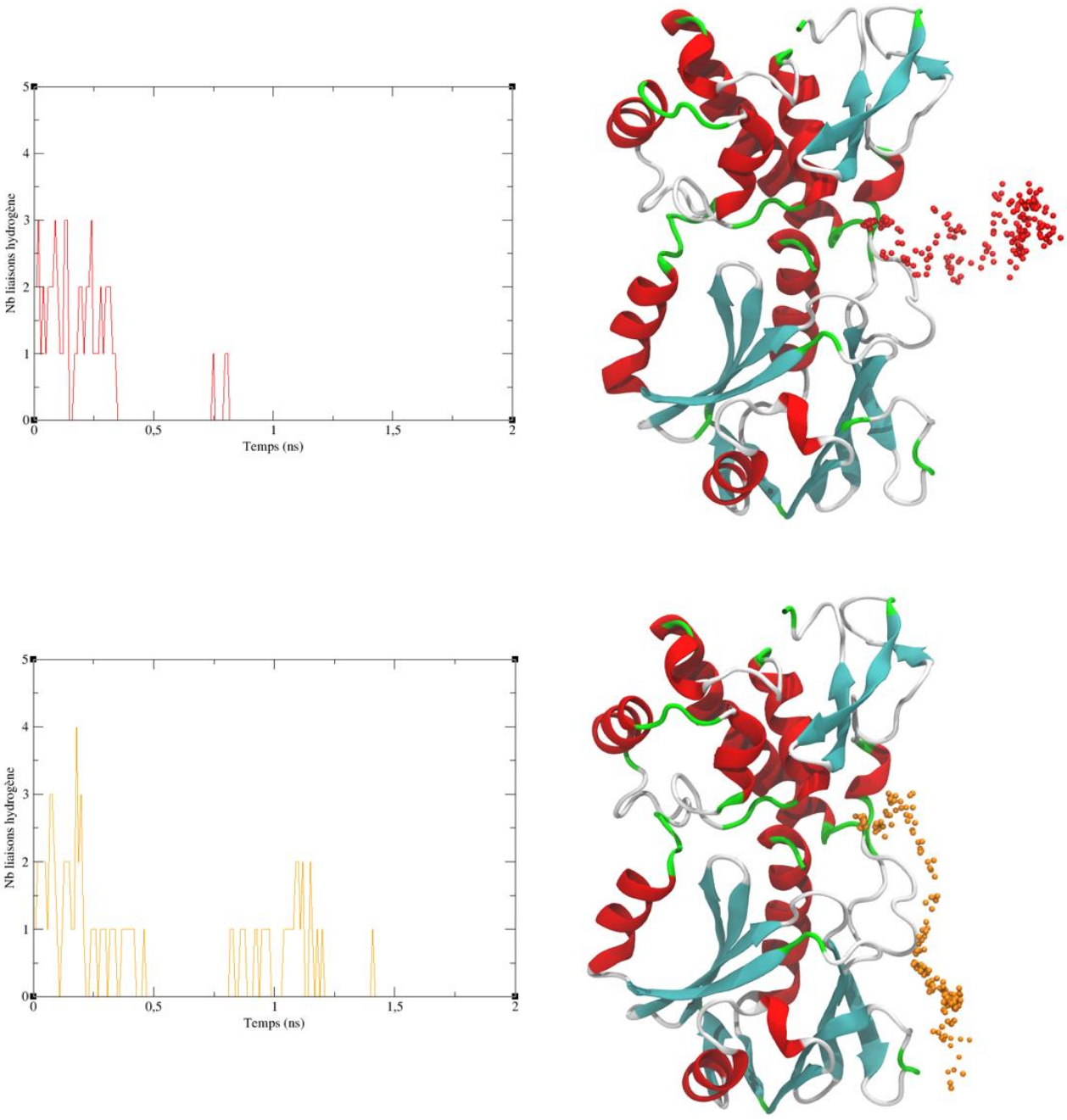


Figure 86 : Description des chemins de sortie empruntés par la base libre dans le système Fpg ADN THF boucle relâchée (1PM5) 8-oxoG libre (1R2Y)
À gauche, le nombre total des liaisons hydrogène formées entre la base libre et la protéine au cours du temps. Sur les images à droite, la protéine est en cartoon colorée en fonction de sa structure secondaire, l'ADN n'est pas représenté. Les sphères représentent la position de la 8-oxoG au cours du temps, à intervalle de 10 ps.

Comme déjà décrit dans le Chapitre 1 de la partie Résultat, d'après la simulation de DM classique de Fpg ADN THF boucle relâchée (1PM5) 8-oxoG libre (1R2Y), lorsque la boucle LCL de Fpg est relâchée, la 8-oxoG située dans le site actif de l'enzyme n'interagit que très peu avec la boucle LCL (**Figure 69 G**, **Figure 85** et **Tableau 16**). En regardant maintenant les simulations de TMD^{-1}, seuls

quelques aa de la boucle LCL tels que S217, S218 et R220 interagissent avec la base coupée lors des quatre scénarii de sortie. Les aa catalytiques tels qu'E2 et E5 sont fréquemment en interaction avec la base libre lors de sa sortie. Les aa E2 et S217 interviennent dans les quatre scénarii, suivis par E5 et S218 et Y238 dans trois scénarii, puis R220 et N233 et K240 dans deux scénarii. Les autres aa tels que K78, E89, K90, H91, Q232, S236, et S246 dans un scénario seulement. Il est intéressant de comparer ces données avec celles obtenues à partir des TMD^{-1} des systèmes où la boucle LCL est fermée (voir section **IV.2.2.3** p. **224**).

Tableau 16 : Aa formant des liaisons hydrogène avec la 8-oxoG libre lors de sa sortie du site actif de *Ll*Fpg ADH THF boucle relâchée (1PM5)
Les aa retrouvés dans les quatre scénarii sont soulignés. Les aa sont colorés en fonction de leur rôle dans la protéine, en rouge les aa catalytiques, en bleu les aa de la boucle LCL et en noir les autres.

	Aa
TMD^{-1} n°1	**E2, E5, S217, R220, N233, Y238, K240**
TMD^{-1} n°2	**E2, E5, S217, S218, Q232, N233, S236, K240, S246**
TMD^{-1} n°3	**E2, E5, S217, S218, R220, Y238**
TMD^{-1} n°4	**E2, K78, E89, K90, H91, S217, S218, Y238**

IV.2.2.3 Implication de la boucle LCL dans le mécanisme de sortie de la 8-oxoG libre

D'après les **Figure 84** et **Figure 86**, quel que soit l'état de la boucle LCL (fermé ou relâché), plusieurs scénarii sont possibles pour permettre la sortie de la base libre. Tous ces scénarii sont probables, nous ne pouvons pas les discriminer. Dans cette section, nous comparons les quatre chemins de sortie en fonction de l'état de la boucle LCL.

Nous avons vu dans les deux sections précédentes que la base libre était beaucoup plus contrainte lorsque la boucle LCL est fermée que lorsqu'elle est relâchée (**Figure 82**, **Figure 83** et **Figure 85).** En effet, la couronne d'interactions formée entre l'O6 et le squelette des aa I219, R220, T221 et Y222 existe en début de simulation et, pour permettre la sortie de la base libre, les interactions devront être rompues. C'est un élément supplémentaire qui suggère que la boucle LCL dans un état fermé stabilise le contenu du site actif. En calculant la RMSF (Cα) de la boucle LCL, nous remarquons que la partie I (216-224) de la boucle LCL du système Fpg ADN THF (1PM5) boucle fermée (1XC8) 8-oxoG libre (1R2Y) peut être plus ou moins flexible, avec un RMSF maximale de 1,7 Å (**Figure 87 A**). Nous remarquons notamment que les aa R220 et T221 sont plus flexibles dans les TMD^{-1}s n°2 et 4 avec des

valeurs de RMSF ≈ 1,5 Å (**Figure 87 A**). Dans ces systèmes, la partie II de la boucle LCL (226-230) semble avoir des valeurs de RMSF très similaires, avec une valeur de RMSF maximale ≈ 1 Å.

Concernant le système Fpg ADN THF boucle relâchée (1PM5) 8-oxoG libre (1R2Y), la courbe de RMSF de la partie I de la boucle LCL semble avoir des profils similaires dans toutes les simulations de TMD^{-1}, avec des valeurs de RMSF maximale comprise entre ~0,8 et ~1,3 Å (**Figure 87 B**). Les courbes de RMSF de la partie II de la boucle LCL semblent avoir des profils différents avec des valeurs de RMSF variant de ~0,7 à ~1 Å (**Figure 87 B**). Il apparait donc que certains aa de la partie I de la boucle LCL soient plus flexibles dans le système boucle fermée que boucle relâchée. Nous allons décrire ce que traduit cette flexibilité par la suite.

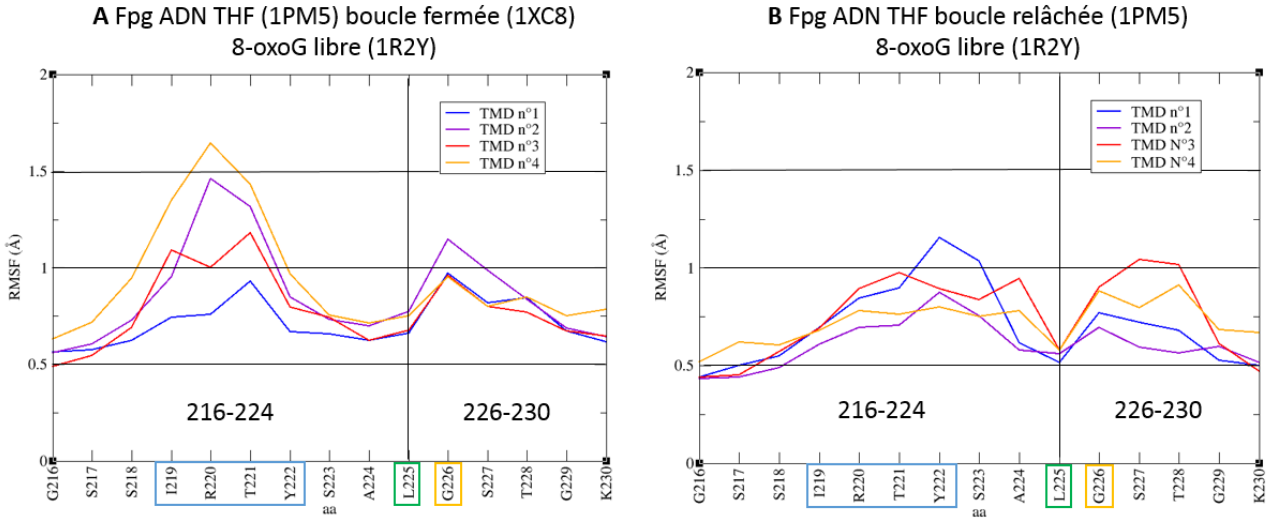


Figure 87 : RMSF des Cα de la boucle LCL des TMD^{-1}s
A) Système Fpg ADN THF boucle relâchée (1PM5) boucle fermée (1XC8) 8-oxoG libre (1R2Y) et **B**) Système Fpg ADN THF boucle relâchée (1PM5) 8-oxoG libre (1R2Y). Les aa clés de la boucle LCL sont encadrés : en bleu les aa de la couronne d'interactions avec l'O6 du substrat, en vert l'aa charnière et en jaune la G226 témoignant de l'état de la boucle LCL, fermé ou relâché.

Sur la **Figure 88**, nous représentons le site actif de l'enzyme à différents temps de simulations lors des quatre simulations de TMD^{-1} de Fpg ADN THF (1PM5) boucle fermée (1XC8) 8-oxoG libre (1R2Y). Dans les images du site actif des quatre TMD^{-1} à 4 temps différents, nous remarquons que les interactions E76-R220 et E76-T221 sont présentes à t = 0 ns, avant l'application de contraintes pour induire une évolution du système. Dans les images **B** et **D** de la **Figure 88**, nous voyons que ces interactions sont rompues à la fin des TMD^{-1} n° 2 et 4 du système Fpg ADN THF (1PM5) boucle fermée (1XC8) 8-oxoG libre 1R2Y), ce qui pourrait expliquer la valeur de RMSF élevée pour les aa R220 et T221 dans ces systèmes (**Figure 87** violet et jaune). Concernant la TMD^{-1} n°3, seule l'interaction E76-T221

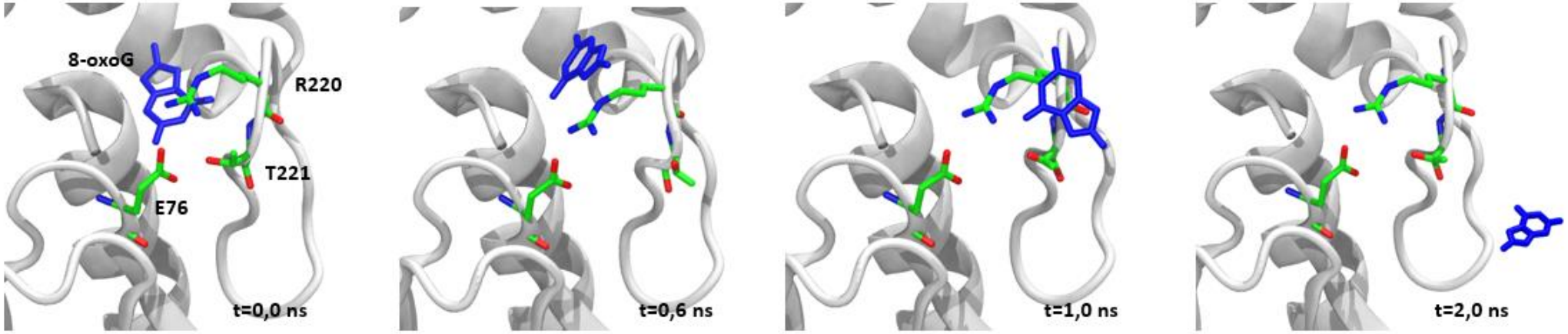
est brisée (**Figure 88 C**), ce qui est également cohérent avec une RMSF moyenne de T221 (**Figure 87 A**). Pour finir, lors de la TMD^{-1} n°1 les deux interactions E76-R220 et E76-T221 sont conservées (**Figure 88 A**), d'où le fait que la RMSF de la partie I de la boucle soit la plus faible des quatre simulations (**Figure 87 A**). Ici, on remarque qu'il y a quatre scénarii bien distincts. Lors de la TMD^{-1} n°1, la base libre emprunte un chemin sans altérer les interactions E76-R220 et E76-T221 sans les altérer (**Figure 88 A**). Dans la TMD^{-1} n°2, la 8-oxoG libre force le passage et brise ces deux interactions (**Figure 88 B**). Lors de la TMD^{-1} n°3, le produit d'excision n'altère pas l'interaction E76-R220 mais rompt l'interaction E76-T221 lors de son passage (**Figure 88 C**). Pour terminer, dans la TMD^{-1} n°4, la base libre semble suivre le même chemin que dans la TMD^{-1} n°1 (**Figure 88 A**), ce qui engendre quand même la destruction des deux interactions E76-R220 et E76-T221 (**Figure 88 D**). Nous avons vu dans le Chapitre précédent que les interactions E76-R220 et E76-T221 sont importantes pour la stabilisation de la boucle dans l'état fermé. D'après ces quatre scénarii, il semblerait que l'ouverture de la boucle LCL ne soit pas nécessaire à la sortie de la base libre. Cependant, à la sortie de cette dernière, deux fois sur quatre, les deux interactions sont rompues, une fois sur quatre l'interaction E76-T221 est brisée, et une fois sur quatre les deux interactions sont conservées. D'après la **Figure 81 A**, les TMD^{-1} peuvent être classées dans l'ordre croissant en fonction de la valeur maximale des apports en énergie apportées aux systèmes dans les 50 premières ps : TMD^{-1} n° 4 (7,6 $kcal.mol^{-1}$) < n°3 (7,9 $kcal.mol^{-1}$) < n°2 (8,1 $kcal.mol^{-1}$) n°4 < 1 (9,8 $kcal.mol^{-1}$). Il semblerait donc que lors de la sortie du produit d'excision, les scénarii les moins couteux en énergie sont ceux durant lesquels l'interaction E76-T221 ou les interactions E76-R220 et E76-T221 sont rompues. Si la base libre sort du site actif sans détruire ces interactions, d'avantage d'énergie doit être injectée au système.

La **Figure 89** représente le site actif de Fpg ADN THF boucle relâchée (1PM5) 8-oxoG libre (1R2Y) lors de 4 TMD^{-1}. Sur les images du site actif de l'enzyme Fpg à t = 0 ns de chaque simulation de TMD^{-1}, on remarque que les interactions E76-R220 et E76-T221 n'existent pas. La conformation des aa E76 et R220/T221 de la boucle LCL ne semblent pas évoluer au cours des quatre TMD^{-1} (**Figure 89 A-D**). On peut en déduire que lorsque la boucle LCL est relâchée, la sortie de la base libre n'influence pas les aa E76, R220 et T221. D'après la **Figure 81 B**, les TMD^{-1} peuvent être classées dans l'ordre croissant en fonction des valeurs maximales des énergies nécessaires à l'évolution des systèmes dans les 50 premières ps : TMD^{-1} n° 3 (8,3 $kcal.mol^{-1}$) < n°1 (9,1 $kcal.mol^{-1}$) < n°4 (9,3 $kcal.mol^{-1}$) n°2 < 1 (10,4 $kcal.mol^{-1}$).

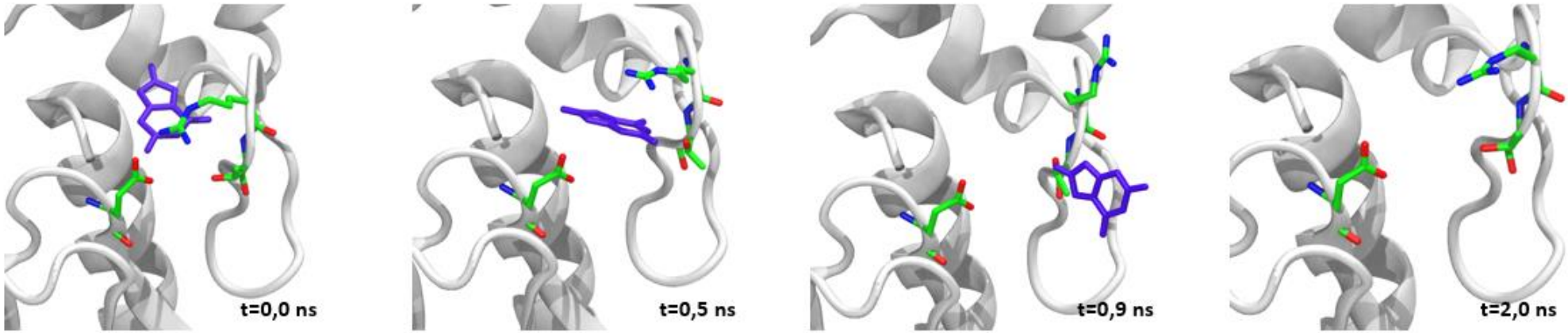
Les énergies maximales injectées aux systèmes lors des huit TMD^{-1} semblent légèrement plus importantes dans le cas de la boucle LCL relâchée. Cependant le nombre de TMD^{-1} réalisé dans chaque cas, boucle LCL fermée et relâchée n'est pas suffisant pour conclure que la différence d'énergie

apportée au système soit statistiquement significative et liée à l'état de la boucle LCL. Pour conclure à ce sujet, il faudrait réaliser d'autres simulations de TMD^{-1} et obtenir un échantillon d'au moins 30 simulations pour réaliser une étude statistique solide. Les simulations de TMD^{-1} produites à cet effet pourraient également permettre d'identifier la trajectoire de sortie de la base libre la plus représentée dans l'échantillon.

A Fpg ADN THF (1PM5) boucle fermée (1XC8) 8-oxoG (1R2Y) TMD^{-1} n°1



B Fpg ADN THF (1PM5) boucle fermée (1XC8) 8-oxoG (1R2Y) TMD^{-1} n°2



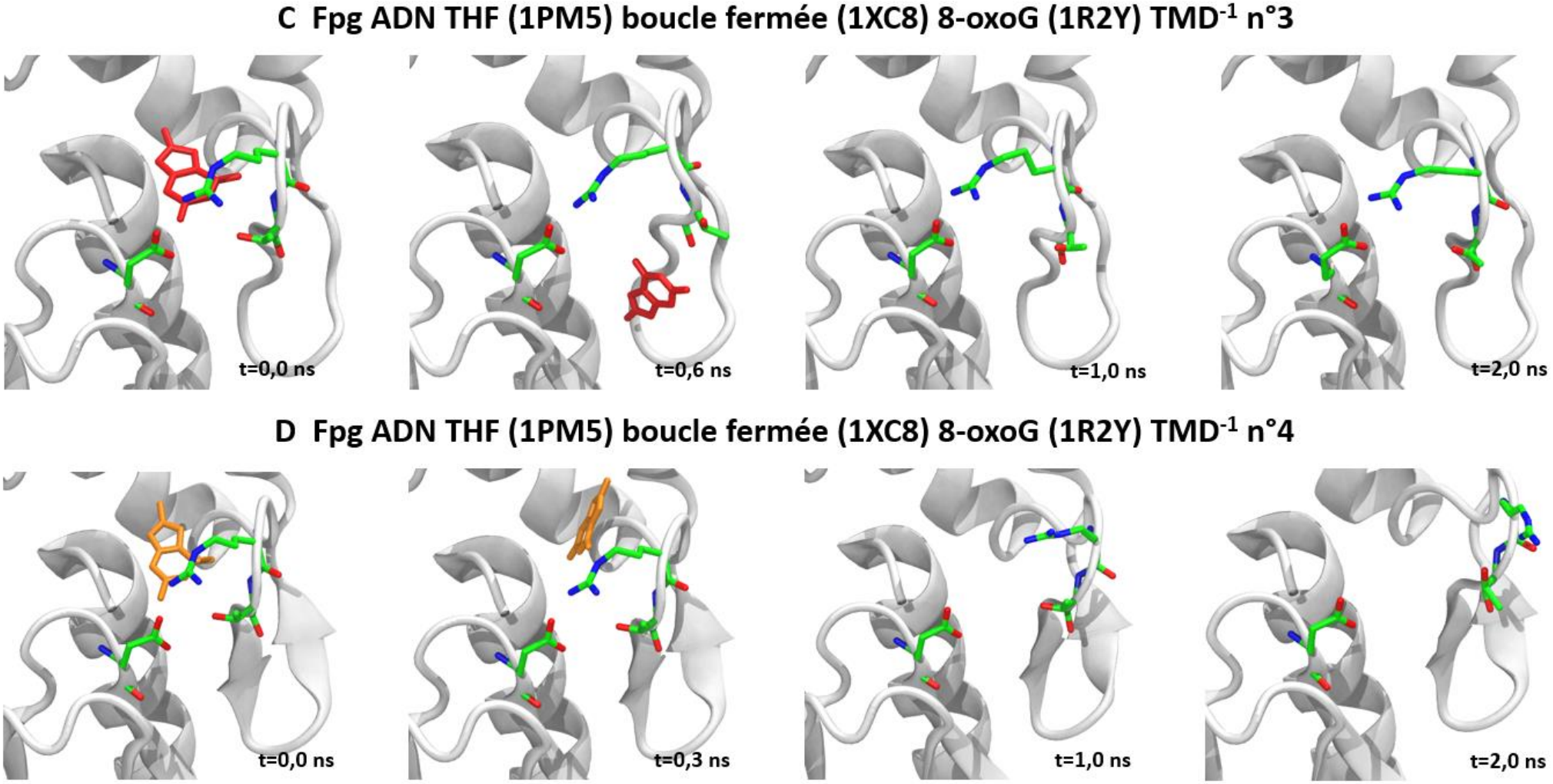
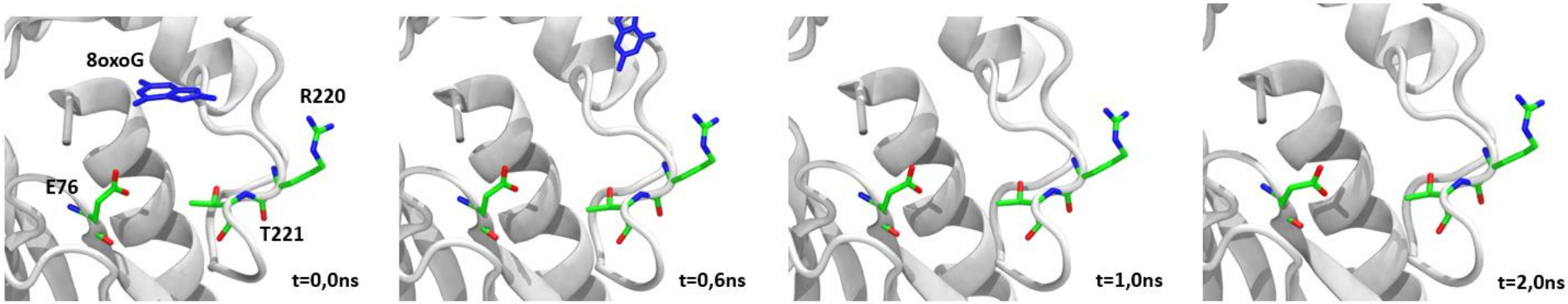


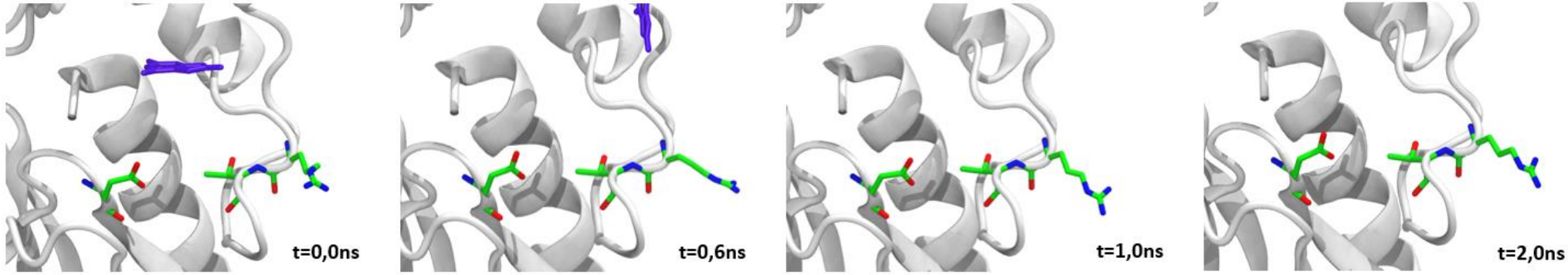
Figure 88 : Évolution de la position de la 8-oxoG dans les quatre TMD^{-1}s du système Fpg ADN THF (1PM5) boucle fermée (1XC8) 8-oxoG libre (1R2Y)

A) TMD^{-1} réalisées à partir de la conformation extraite à 10 ns, **B**) à 15 ns, **C**) à 20 ns, **D**) à 25 ns de la simulation de DM classique du système Fpg ADN THF (1PM5) boucle fermée (1XC8) 8-oxoG libre (1R2Y).

A Fpg ADN THF boucle relâchée (1PM5) 8-oxoG (1R2Y) TMD^{-1} n°1



B Fpg ADN THF boucle relâchée (1PM5) 8-oxoG (1R2Y) TMD^{-1} n°2



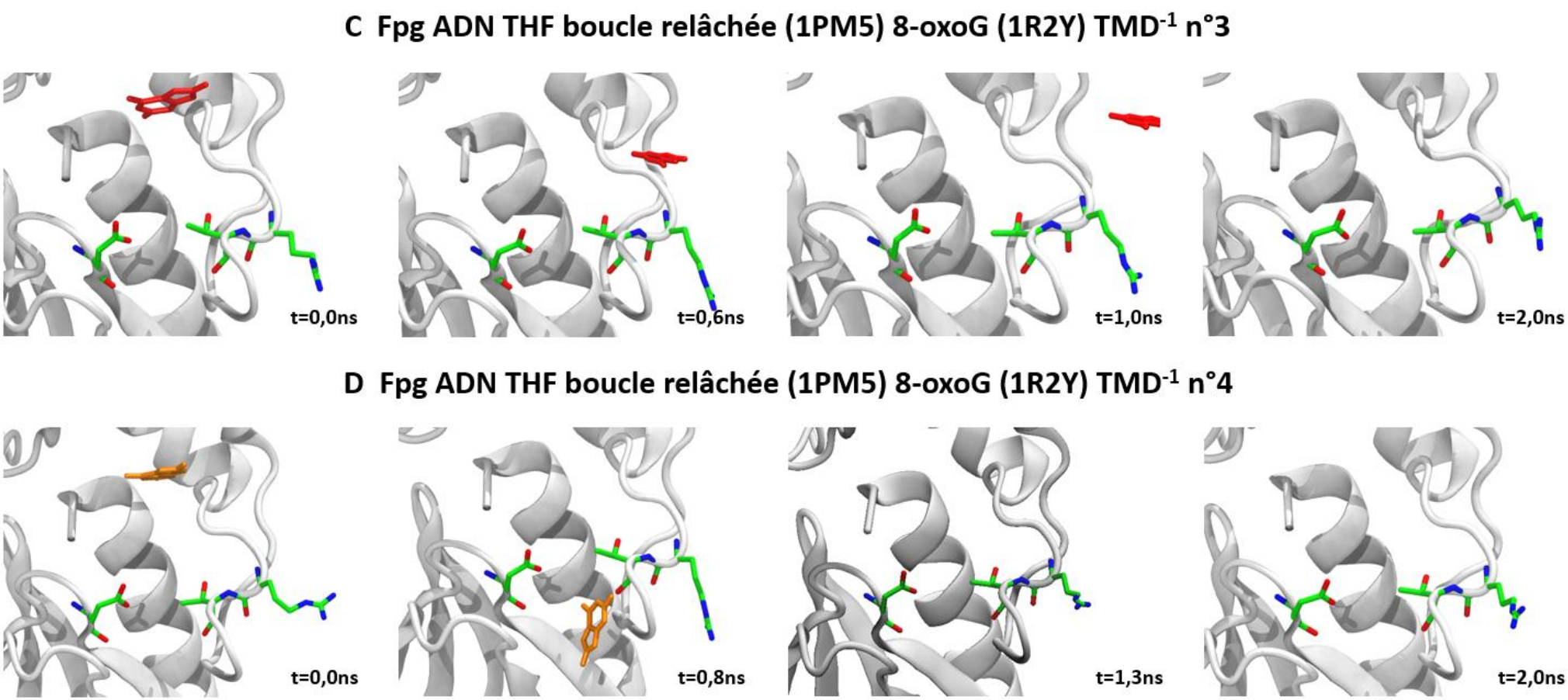


Figure 89 : Évolution de la position de la 8-oxoG dans les quatre TMD^{-1}s du système Fpg ADN THF boucle relâchée (1PM5) 8-oxoG libre (1R2Y)
A) TMD^{-1} réalisées à partir de la conformation extraite à 10 ns, **B**) à 15 ns, **C**) à 20 ns, **D**) à 25 ns de la simulation de DM classique du système Fpg ADN THF boucle relâchée (1PM5) 8-oxoG libre (1R2Y).

IV.2.3 Conclusion et perspectives

Dans les sections **IV.2.2.1** p. **214** et **IV.2.2.2** p. **220**, nous avons décrit différents chemins de sortie de la 8-oxoG libre selon les deux états possibles de la boucle LCL (fermée ou relâchée) à partir de 8 simulations de TMD^{-1}. La base libre suit trois trajectoires différentes, la 8-oxoG libre (i) longe le domaine C-terminal de la protéine ou (ii) diffuse dans le petit sillon de l'ADN puis dans le solvent ou encore (iii) longe le domaine N-terminal de la protéine. Ces trois scénarii sont observés dans les simulations de TMD^{-1} des deux systèmes. La trajectoire suivie par la base libre semble donc indépendante de l'état de la boucle LCL. Quel que soit l'état de la boucle LCL, les valeurs des énergies de contraintes apportées aux systèmes sont trop similaires ce qui ne nous permet pas de discriminer les scénarii observés dans les huit simulations de TMD^{-1}. Elles ne nous permettent pas non plus d'identifier l'état le plus probable de la boucle LCL, fermé ou relâché, en présence d'une 8-oxoG libre.

Ensuite, dans la section **IV.2.2.3** p. 224, nous nous sommes intéressés à l'impact de la sortie de la base libre sur la boucle LCL. Nous avons noté la présence des interactions E76-R220 et E76-T221 exclusivement dans le contexte de la boucle LCL fermée. Ces interactions E76-R220 et E76-T221 sont rompues 3 fois sur 4, cette rupture s'accompagne un mouvement d'ouverture de la boucle LCL pour permettre à la base libre de quitter le site actif. Les aa E76, R220 et T221 sont donc importants pour maintenir un état fermé de la boucle LCL, et cette dernière s'adapte à la trajectoire de la base libre en rompant les interactions qu'elle crée avec le domaine N-terminal de l'enzyme. Dans le cas où la boucle LCL est relâchée, les interactions E76-R220 et E76-T221 n'existent pas. Dans ce contexte, les aa E76, R220 et T221 ne sont pas impliqués dans la sortie de la base hors du site actif. De ce fait, la boucle LCL ne semble pas avoir un rôle dans la sortie de la base libre dans le contexte où elle est relâchée.

Un nombre plus important de simulations de TMD^{-1} ainsi que l'analyse des données obtenues par une méthode de classification pourrait nous permettre entre autres d'identifier un ou plusieurs chemins de sortie préférentiels de la base libre. En outre, la création de protéines *Ll*Fpg mutantes sur les aa E76, R220 et T221 ainsi que des tests d'activité de ces mutants pourraient nous permettre de déterminer si la boucle LCL joue un rôle dans la sortie du produit d'excision.

IV.3 **Chapitre 3 : P**rédiction de sites de fixation sur les protéines *Ll*Fpg/hNEIL1

Nous avons vu dans la section **I.2.4.1** p. **77** de l'Introduction que la protéine hNEIL1 est une cible thérapeutique intéressante dans les stratégies anti-cancers. Dans ce cadre, plusieurs molécules appelées $2TX_n$ ont été conçues à l'ICOA et testées dans l'équipe au CBM dans le but d'identifier des inhibiteurs de la protéine *Ll*Fpg. Le projet de recherche d'inhibiteurs a été initié sur la protéine bactérienne car les essais biochimiques (tests d'activité et cristallographie) sont mieux maitrisés sur ce modèle biologique. Les molécules semblent avoir des effets d'inhibition sur *Ll*Fpg avec des IC_{50} de l'ordre du µM. Les molécules $2TX_n$ n'ont que des effets faibles ou inexistants sur l'enzyme humain. Dans la continuité du projet de recherche d'inhibiteurs sélectifs de l'enzyme hNEIL1, nous avons utilisé la modélisation moléculaire pour chercher à comprendre pourquoi les molécules présentent des effets d'inhibition sur l'enzyme bactérien et non sur l'enzyme humain. Les modes d'interactions des $2TX_n$ avec les deux protéines testées ne sont pas bien compris. Pour approfondir le sujet, nous avons mis au point une méthode de prédiction de sites de fixation prenant en compte la structure du ligand grâce au docking. Le docking est classiquement utilisé pour explorer une surface délimitée (poche ou cavité enfouie) de la protéine. Dans notre contexte, nous ne connaissons qu'un seul site de fixation des $2TX_n$ sur les cibles, le motif H2TH. Cependant, il n'est pas encore établi que les effets d'inhibition des $2TX_n$ sur les cibles soient liés à ce site d'ancrage. Dans le but d'identifier d'autres sites de fixation, nous avons donc élargis la zone de recherche à l'ensemble de la protéine. Dans ce contexte, le docking est dit « sans *a priori* », comme décrit dans la section **III.2.4.2** p. **168** de la partie Méthodologie. Comme énoncé, dans la section **III.2.1** p. **149** de la partie Méthodologie, les protéines sont des objets flexibles et dynamiques. Afin de tenir compte de cet aspect, nous avons réalisé des simulations de DM classique à partir desquelles ont été extraites un petit nombre de structures représentatives de la flexibilité des systèmes étudiés. Ces structures ont été utilisées pour le docking sans *a priori*. Ainsi, nous avons exploré la totalité de la surface de plusieurs conformations possibles des systèmes à la recherche de sites de fixation des $2TX_n$ grâce à une méthode originale de docking flexible et sans *a priori*.

IV.3.1 Simulations de dynamique moléculaire classique

Quatre systèmes ont été créés pour ce projet. Deux d'entre eux correspondent à *Ll*Fpg complexée à un ADN THF et à la protéine libre, de même pour hNEIL1 (**Tableau 17**). Au moment de l'initiation de ce projet, il n'existait pas de structure de *Ll*Fpg libre ni de structure de hNEIL1 complexée à un acide nucléique. Ce n'est qu'un an plus tard qu'une structure de hNEIL1 mutante complexée à un

ADN Tg a été résolue et publiée [190]. Lors de l'initiation du projet, la création de ces deux systèmes a donc été nécessaire. Pour *Ll*Fpg libre, nous avons utilisé la structure de *Ll*Fpg complexée à un ADN THF (PDBid : 1PM5) [195], et nous avons retiré l'ADN. Pour la conception du modèle hNEIL1 libre (PDBid : 1TDH) [187], une seule structure cristallographique de l'enzyme était disponible dans la PDB au début du projet. Cette structure est incomplète et ne représente d'une partie de la protéine. Il s'agit en effet d'une enzyme tronquée 1-289 au lieu de 1-389 (100 aa manquant en C-ter). Cependant, d'après la littérature, ceci n'affecte pas les activités glycosylases ou AP lyase de l'enzyme [187]. Les 100 derniers aa ayant une nature désordonnée, l'obtention de la structure cristallographique de hNEIL1 entière est difficile [187]. La fonction de la queue C-ter n'est toujours pas connue à ce jour. Nous n'avons donc pas cherché à reconstruire les 100 aa manquants. Par ailleurs, une autre partie de la protéine n'a pas été résolue dans la structure 1TDH, il s'agit des aa 203 à 207. Nous avons reconstruit cette portion de la protéine avec l'outil Prime de la suite Schrödinger [272]. Pour la conception du modèle hNEIL1 ADN THF, nous avons fusionné les structures de hNEIL1 libre et de MvNei1 ADN THF (PDBid : 4NRW) [221] après superposition de ces dernières grâce à un alignement structural selon l'outil MultiSeq [214] de VMD [215].

Les modèles ont été solvatés, neutralisés, thermalisés, minimisés, équilibrés et produits avec AMBER en suivant le même protocole déjà évoqué dans les Chapitres 1 et 2 de la partie Résultats et décrit dans la section **III.1.4** p. **136** de la partie Méthode.

Tableau 17 : Description des 4 systèmes créés pour la prédiction des sites de fixation des $2TX_n$ sur les protéines *Ll*Fpg et hNEIL1

Nom du système	PDBid(s) d'origine(s)	Ligand dans la structure d'origine	Modifications apportées	Temps de simulation analysé (ns)
Fpg ADN THF (1PM5)	1PM5	ADN THF	-	40
Fpg libre (1PM5 sans ADN)	1PM5	-	Retrait ADN	40
hNEIL1 (1TDH) ADN THF (4NRW)	1TDH, 4NRW	ADN THF	Ajout ADN	40
hNEIL1 libre (1TDH)	1TDH	-	-	40

IV.3.2 Vérification des simulations

Pour chaque simulation de DM, la stabilité des systèmes protéines libres et des complexes ADN/protéine est évaluée par calcul de la RMSD (**Figure 90**). Ces graphiques permettent de déterminer le temps de simulation à partir duquel le système a atteint l'équilibre. Les analyses qui vont suivre portent sur les simulations une fois l'équilibre du système atteint, c'est-à-dire à partir de 10 ns pour les quatre systèmes (**Figure 90 A-D**).

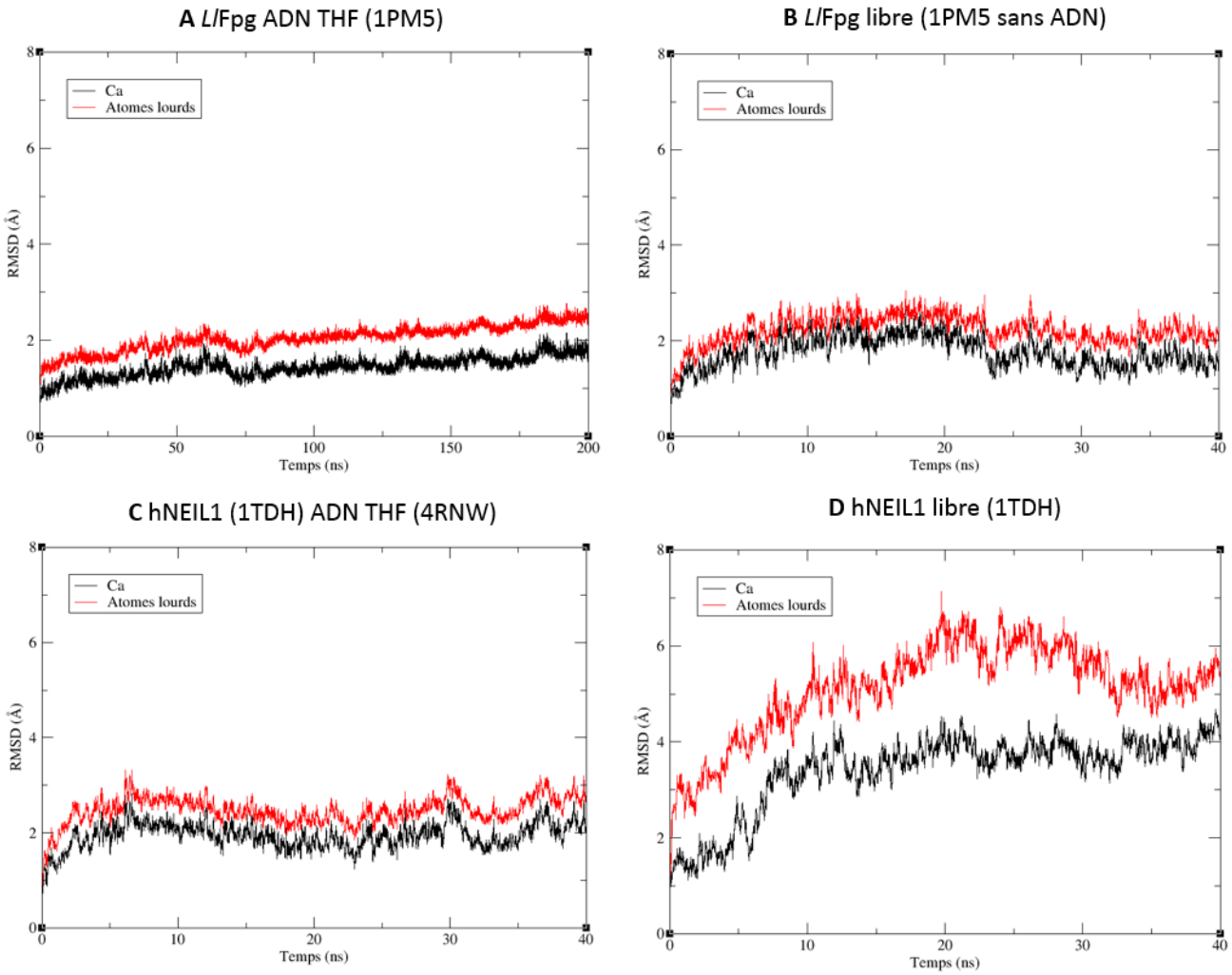


Figure 90 : Courbes de RMSD des quatre systèmes de *Ll*Fpg et hNEIL1
En noir RMSD de Cα et en rouge RMSD de tous les atomes lourds. Ces graphiques permettent de définir le temps de simulation à partir duquel les systèmes sont à l'équilibre, soit à 10 ns.

Le système hNEIL1 libre (1TDH) montre une RMSD beaucoup plus élevée (~5 Å) que les 3 autres systèmes (~2 Å), traduisant un certain changement conformationnel de la protéine à partir de 10 ns (**Figure 90 D**). En faisant le détail de la RMSD de ce système, on peut s'apercevoir qu'une région de la protéine est particulièrement flexible, il s'agit des positions 201 – 223 (**Figure 91 A**). Cette portion de la protéine correspond aux aa 203-207 qui ont été reconstruits car ils n'étaient pas résolus dans la

structure cristallographique. Le fait que ces aa soient très flexibles est donc en bon accord avec les données expérimentales. Elle correspond à un enchainement hélice α, boucles, hélice α (**Figure 91 B**). Nous ne remarquons pas une telle flexibilité dans le système hNeil1 ADN THF (4NRW), (**Figure 90 C**) ce qui peut signifier que l'ADN THF est un élément qui permet de stabiliser cette partie de la protéine, avec laquelle l'acide nucléique n'est pourtant pas en contact direct.

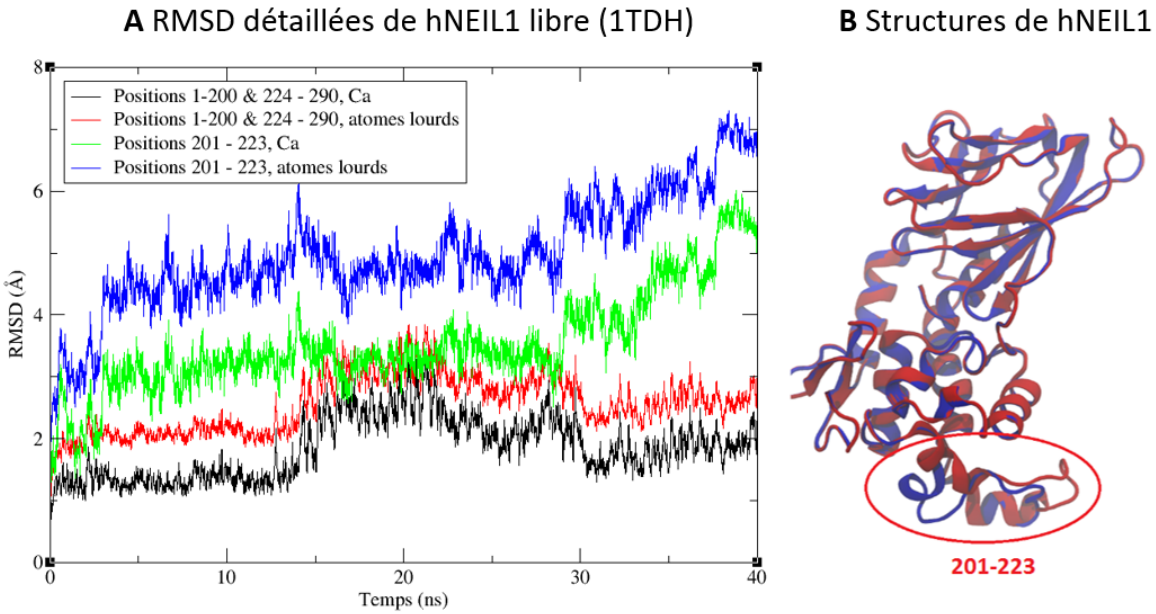


Figure 91 : Détails de la RMSD du système hNEIL1 libre (1TDH)
A) La RMSD de la protéine sans les positions 201-223 est calculée sur les Cα de la protéine (noir) et pour les atomes lourds (rouge). La RMSD des positions 201-223 est également calculée sur les Cα de la protéine (vert) et pour les atomes lourds (bleu). **B**) Représentation cartoon de la structure PDB 1TDH (bleu) et de la structure reconstruite avec Prime (rouge). Les positions 201 – 223 comportant les aa 203 à 207 reconstruits sont entourées en rouge.

IV.3.3 Echantillonnage structural

Les simulations de DM produisent un grand nombre de conformations des systèmes respectifs soient 4000 structures par système. Ce nombre est beaucoup trop élevé pour les calculs d'amarrage moléculaire. L'échantillonnage structural est une solution pour sélectionner les structures, qui, dans ces « pools », sont les plus diverses et les plus représentatives de la flexibilité du système. Pour cette étape, nous considérons l'ensemble du système, c'est-à-dire la protéine mais aussi l'ADN lorsque celui-ci est présent. Pour identifier ces structures, nous avons utilisé une méthode de classification non supervisée appelée classification ascendante hiérarchique (CAH) décrite dans la section **III.1.9.4** p. **144** de la partie Méthodologie. Ce procédé de classification a pour prérequis la connaissance du nombre de groupes dans lesquels nous souhaitons classer nos structures. Nous choisissons de classer les structures selon un critère de RMSD calculé sur tous les atomes lourds. Pour estimer ce nombre de

groupes, nous avons produit des cartes de RMSD 2D (**Figure 92** et **Figure 93**). Elles représentent les valeurs de RMSD de toutes les conformations (des systèmes à l'équilibre) les unes par rapport à toutes les autres. La qualité d'une CAH peut être évaluée par le calcul des distances moyennes entre les structures au sein d'un même cluster (**Tableau 18**, **Tableau 19**, **Tableau 20** et **Tableau 21**). Si cette distance est grande, cela signifie que les structures au sein de ce cluster sont très dissimilaires, et ce n'est pas souhaitable car nous voulons extraire une structure représentative de toutes les autres. Cette structure est appelée « centroïde » et doit donc être proche des autres structures du cluster auquel elle appartient. Pour réduire la distance moyenne inter-structures au sein des clusters, il est nécessaire d'augmenter progressivement le nombre de clusters et de segmenter d'avantage l'échantillon jusqu'à atteindre une faible diversité structurale au sein des clusters.

Sur la **Figure 92 A**, nous distinguons 4 clusters et remarquons que la RMSD inter-structures monte jusqu'à 9 Å, ce qui est élevé. Cela est lié au fait que nous avons considéré tous les atomes lourds du système, et comme déjà mentionné dans la section précédente, la portion 201 à 223 de la protéine est très flexible (**Figure 91**)**,** et a donc augmenté les RMSD globales. Pour réduire la distance moyenne inter-structure, nous avons décidé de segmenter le cluster composé des conformations du système aux temps allant de 10 à 24 ns en 3 clusters, ce qui nous fait au total 6 clusters. D'après les cartes RMSD 2D **Figure 92** et **Figure 93**, nous remarquons que les clusters peuvent être discontinus par rapport au temps, comme par exemple le cluster n°4 en vert semble entouré par le cluster n°2 en noir sur la **Figure 92 A**. D'après le **Tableau 21**, on observe que les distances moyennes inter-structures au sein des clusters sont assez élevées. La portion 201-223 ne fait pas partie du site actif de l'enzyme, mais comme nous souhaitions tenir compte de la flexibilité globale du récepteur, nous devions garder ces résidus dans le calcul des cartes RMSD 2D et la CAH. Cependant la grande dynamique de cette partie de l'enzyme biaise la carte RMSD 2D et par conséquent le choix du nombre de clusters. Les centroïdes issus de cette CAH ne sont donc pas les structures les plus représentatives de la flexibilité globale du système compte tenu de la grande influence des aa 201-223 sur les valeurs de RMSD. Nous aurions peut-être dû effectuer cette analyse pour les aa 201 à 223 et le reste de l'enzyme de façon individuelle.

Concernant les trois autres systèmes, nous avons choisis, à partir des cartes de RMSD 2D des **Figure 92 B**, **Figure 93 A** et **B**, de segmenter les pools de structures en 8 clusters chacun. Les **Tableau 18**, **Tableau 19**, **Tableau 20** et **Tableau 21** permettent de juger la qualité des CAH. Les distances moyennes inter-structures au sein des 24 clusters n'excèdent pas 2 Å. On constate que certains de ces clusters représentent une faible partie de la totalité des conformations produites en DM (< 5%). Même

si la portion de la DM représentée par ces clusters est faible, la diversité conformationnelle qu'ils représentent n'est pas négligeable.

Tableau 18 : Classification des structures produites par simulation de DM de *Ll*Fpg ADN THF (1PM5) en 8 groupes, et identification de 8 structures centroïdes

N° du cluster	Nb de structures	Fraction (%)	Distance moyenne inter-structures (Å)	Ecart type distance inter-structures (Å)	Centroïde (structure X/4000)	Distance moyenne au centroïde (Å)
1	616	20,5	1,6	0,2	3720	2,2
2	531	17,7	1,5	0,2	1873	2,0
3	524	17,5	1,5	0,2	1184	2,2
4	440	14,7	1,4	0,2	2627	2,0
5	414	13,8	1,5	0,2	2240	2,0
6	200	6,7	1,5	0,3	3145	2,0
7	149	5,0	1,4	0,2	2956	2,0
8	127	4,2	1,3	0,2	3505	2,1

Tableau 19 : Classification des structures produites par simulation de DM de *Ll*Fpg libre (1PM5 sans ADN) en 8 groupes, et identification de 8 structures centroïdes

N° du cluster	Nb de structures	Fraction (%)	Distance moyenne inter-structures (Å)	Ecart type distance inter-structures (Å)	Centroïde (structure X/4000)	Distance moyenne au centroïde (Å)
1	727	24,2	1,5	0,2	1978	1,9
2	599	20,0	1,5	0,2	3553	1,9
3	520	17,3	1,5	0,2	3159	1,9
4	510	17,0	1,5	0,2	1197	2,0
5	245	8,2	1,4	0,2	2699	1,9
6	239	8,0	1,4	0,2	2442	1,8
7	138	4,6	1,4	0,3	2305	1,9
8	23	0,8	1,1	0,1	2619	2,0

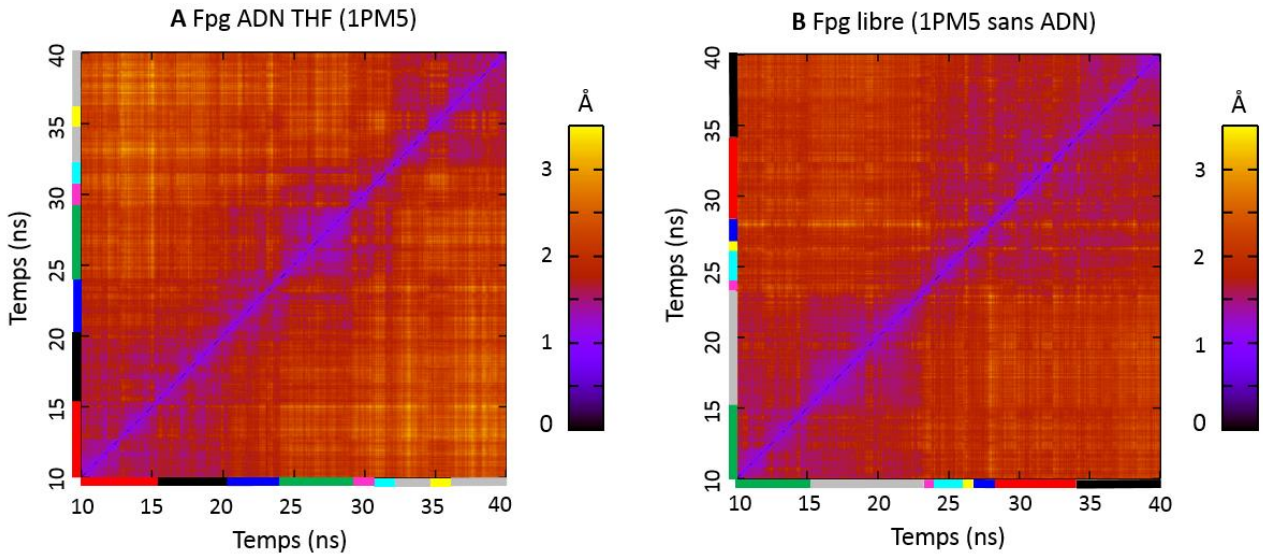


Figure 92 : Cartes RMSD 2D des simulations de *Ll*Fpg
A) Fpg ADN THF (1PM5) et **B**) Fpg libre (1PM5 sans ADN). Comme nous n'avons analysé que les portions de trajectoire où le système protéine ou ADN/protéine est à l'équilibre, les cartes ne représentent la RMSD du système que pour les temps allant de 10 à 40 ns. Les barres colorées au niveau de l'abscisse et de l'ordonnée correspondent aux clusters générés par la CAH, le code couleur est le même que pour les **Tableau 18** et **Tableau 19** correspondants.

Tableau 20 : Classification des structures produites par simulation de DM de hNEIL1 (1TDH) avec ADN (4NRW) en 8 groupes, et identification de 8 structures centroïdes

N° du cluster	Nb de structures	Fraction (%)	Distance moyenne inter-structures (Å)	Ecart type distance inter-structures (Å)	Centroïde (structure X/4000)	Distance moyenne au centroïde (Å)
1	1062	35,4	1,9	0,3	1660	2,7
2	751	25,0	2,0	0,3	2607	2,4
3	382	12,7	1,9	0,3	3628	2,7
4	317	10,6	1,8	0,3	3093	2,6
5	153	5,1	1,6	0,3	2149	2,3
6	147	4,9	1,6	0,3	2290	2,6
7	102	3,4	1,4	0,2	3938	2,7
8	87	2,9	1,7	0,3	2899	2,5

Tableau 21 : Classification des structures produites par simulation de DM de hNEIL1 libre (1TDH) en 6 groupes, et identification de 6 structures centroïdes

N° du cluster	Nb de structures	Fraction (%)	Distance moyenne inter-structures (Å)	Ecart type distance inter-structures (Å)	Centroïde (structure X/4000)	Distance moyenne au centroïde (Å)
1	808	26,9	4,6	0,9	3385	8,7
2	735	24,5	4,8	0,9	1439	7,3
3	541	18,0	4,2	0,8	2757	8,3
4	380	12,7	4,6	0,9	1932	7,9
5	290	9,7	4,4	0,9	1105	7,1
6	247	8,2	4,5	1,1	2593	7,7

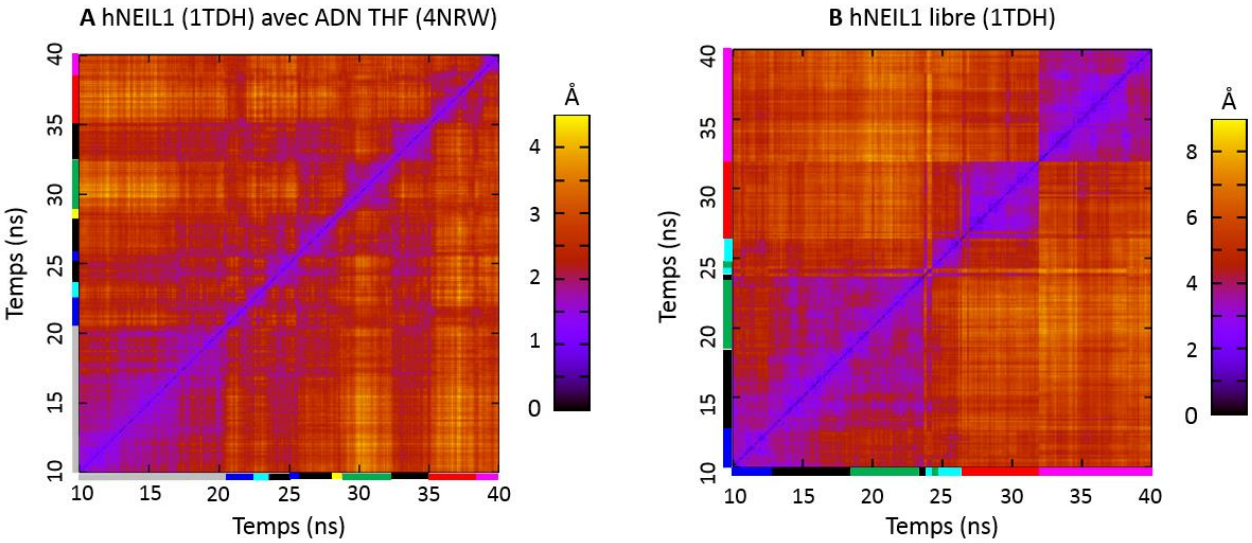


Figure 93 : Cartes RMSD 2D des simulations de hNEIL1
A) hNEIL1 avec ADN THF (4NRW) et **B**) hNEIL1 libre (1TDH). Comme nous n'avons analysé que les portions de trajectoire où le système protéine ou ADN/protéine est à l'équilibre, les cartes ne représentent la RMSD du système que pour les temps allant de 10 à 40 ns. Les barres colorées au niveau de l'abscisse et de l'ordonnée correspondent aux clusters générés par la CAH, le code couleur est le même que pour les **Tableau 20** et **Tableau 21** correspondants.

À partir de la CAH des structures produites par les simulations de DM des quatre systèmes, nous avons extrait 8 structures représentatives de la flexibilité des systèmes *Ll*Fpg ADN THF (1PM5), *Ll*Fpg libre (1PM5 sans ADN) et hNEIL1 (1TDH) ADN THF (4RNW) ainsi que 6 structures à partir du système hNEIL1 libre (1TDH) (**Figure 92** et **Figure 93**). Au total, nous avons donc utilisé 30 structures pour la recherche de site de fixation sur les cibles *Ll*Fpg et hNEIL1 grâce à une méthode de docking.

IV.3.4 Tests des méthodes de docking

Dans cette partie je décrirai les différents protocoles utilisés afin de tester et d'éprouver les logiciels de docking. Ensuite, je détaillerai les résultats de la prédiction de sites de fixation des $2TX_n$ sur les différents systèmes obtenus par docking sans *a priori*.

IV.3.4.1 Préparation des molécules

La préparation des ligands et des récepteurs est nécessaire avant le docking. Si plusieurs chaines protéiques sont écrites dans le fichier contenant la structure du récepteur, une seule d'entre elles est conservée. Dans le cas où les structures des récepteurs sont issues des simulations de DM, les molécules d'eau sont retirées. En revanche concernant les « redockings » (tentative de reproduction des résultats expérimentaux) lors des tests des logiciels, nous avons réalisé chaque essai deux fois, avec et sans eau structurale. Cela se justifie par le fait que les molécules d'eau contenues dans les structures de complexes protéine/ligand sont parfois impliquées dans la stabilisation de ces derniers [273]. Toutes les molécules tierces (ions, lipides…) ajoutées pour faciliter la cristallisation sont supprimées. Les structures des ligands sont dessinées avec la suite Marvin de la suite ChemAxon [248]. Nous avons utilisé VSPrep (décrit dans la section **III.2.2.1** p. **152** de la partie Méthodologie) pour générer les tautomères des ligands et nous avons gardé la molécule dont le pourcentage de distribution est le plus élevé à pH = 7,4 pour chacune des molécules dockées. Les étapes qui suivent sont les mêmes pour les protéines et les ligands avec le logiciel et les scripts AutoDock Tools [240]. Seuls les hydrogènes polaires (portés par n'importe quel atome sauf le carbone) sont conservés. Les charges de Kollman sont ajoutées à la protéine, et les charges de Gasteiger sont calculées et assignées aux atomes des acides nucléiques et des ligands. La préparation des ligands et des protéines est la même pour tous les tests de docking décrit à la suite de ce paragraphe, quel que soit le logiciel de docking utilisé.

IV.3.4.2 Test d'ADV en docking « classique »

Avant d'effectuer un docking sans *a priori*, nous avons testé le logiciel ADV lors d'un docking classique, dans un petit volume de recherche. En effet, des structures de *Ll*Fpg complexée à un ADN THF co-cristallisées avec les molécules 2TX, 2TX2, 2TX3 et F3CS, obtenues dans l'équipe de recherche au CBM (**Figure 94 A**), ont permis l'identification d'un site de fixation. Ce site d'interaction est localisé sur le motif H2TH de *Ll*Fpg, à l'interface entre la protéine et l'ADN (**Figure 94 B**). Nous avons donc tenté de reproduire ces résultats expérimentaux avec le logiciel de docking ADV. Pour cela, nous avons

préparé les structures cristallographiques de la protéine complexée à un ADN THF desquelles nous avons retiré les ligands que nous avons ensuite « redockés ». Pour évaluer l'importance de l'eau structurale dans la stabilisation des complexes protéine/2TX$_n$, nous avons effectué deux tests, le premier sans et le second avec les molécules d'eau structurales issues de la structure cristallographique. Nous avons défini le paramètre « `num_modes` » à 5 et laissé le paramètre `exhaustiveness` à 8 (valeur par défaut). Pour rappel, les paramètres d'ADV pour ce docking sont résumés dans l'Annexe B **VI.2** p. **320**.

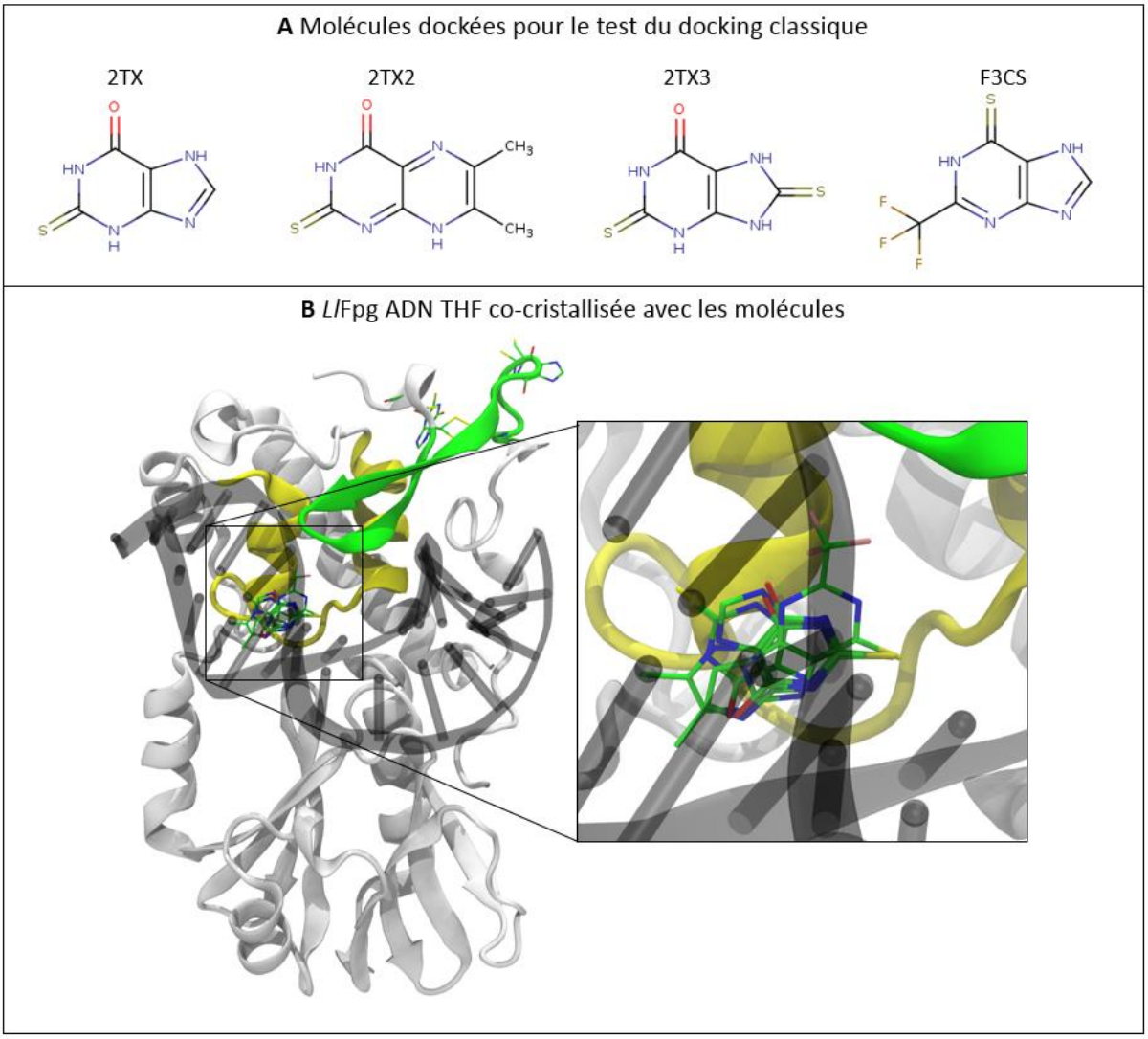


Figure 94 : Données expérimentales exploitées pour le test de docking classique
A) Structures des molécules 2TX, 2TX2 2TX3 et F3CS ainsi que **B**) structures de *Ll*Fpg co-cristallisée avec les molécules présentées en **A**. La protéine est représentée en cartoon blanc, les motifs H2TH et le doigt à zinc en jaune et vert respectivement. L'ADN est représenté en cartoon translucide noir. Les structures des molécules 2TX, 2TX2, 2TX3 et F3CS sont représentées en bâtonnets fins.

IV.3.4.2.1 Comparaison des résultats aux données expérimentales

Pour mieux évaluer la différence entre les résultats des deux tests de docking classique dans le récepteur avec ou sans eau structurale et les données expérimentales, nous avons calculé la RMSD de chacune des conformations du ligand générées par le docking avec la conformation du ligand observée dans les structures cristallographiques (**Tableau 22**). Les poses du docking les plus proches des données expérimentales sont présentées dans la **Figure 95**.

Tableau 22 : Comparaison des résultats issus du docking aux données expérimentales
Les résultats les plus proches de données expérimentales sont surlignés en jaune et sont aussi présentés en partie dans la **Figure 95**.

	Sans eau structurale		Avec eau structurale	
	Score (kcal.mol^{-1})	**RMSD structure expérimentale (Å)**	**Score (kcal.mol^{-1})**	**RMSD structure expérimentale (Å)**
2TX	-5,6	3,2	-6,2	0,5
	-5,4	0,6	-5,8	2,7
	-5,2	4,0	-4,7	3,4
	-5,2	4,5	-3,3	4,1
	-4,9	3,2	-3,1	4,0
2TX2	-6,4	4,5	-5,8	3,2
	-6,2	4,9	-5,5	1,4
	-6,0	3,4	-4,9	4,6
	-5,9	3,8	-3,8	5,3
	-5,9	5,4	-0,5	3,2
2TX3	-5,4	3,6	-6,1	0,4
	-5,3	3,5	-5,8	4,1
	-5,3	3,6	0,4	2,5
	-5,3	3,4	1,3	4,7
	-5,1	4,1	1,8	2,6
F3CS	-5,6	3,0	-5,9	3,9
	-5,5	0,7	-5,8	1,0
	-5,5	1,3	-5,6	1,3
	-5,5	3,8	-5,6	4,0
	-5,4	4,0	-5,5	4,5

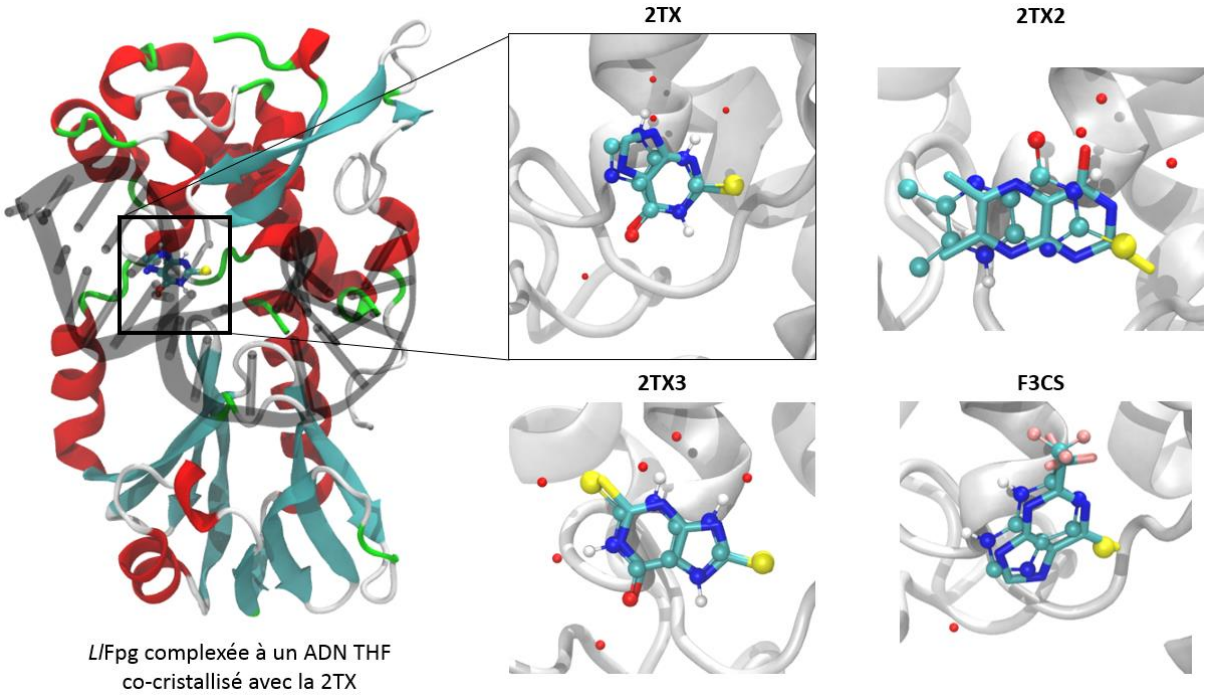


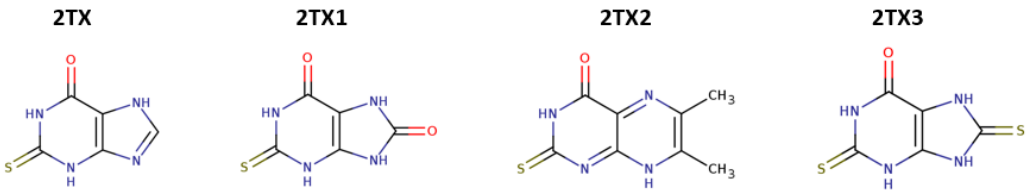
Figure 95 : Poses des ligands, issues du docking, les plus proches des données expérimentales
Les données issues du docking sont représentées par les sphères et les bâtonnets fins et les données expérimentales sont représentées par les bâtonnets épais. Sont présentés ici les résultats du docking effectué dans les structures contenant l'eau structurale (représentée par des sphères rouges).

Dans le **Tableau 22** nous observons que la présence d'eau structurale dans la structure du récepteur permet, pour certaines molécules comme la 2TX2 et la 2TX3, de retrouver plus facilement des conformations du ligand similaires aux données expérimentales. On remarque aussi qu'en présence d'eau structurale, les scores d'affinité sont légèrement plus élevés (-0,3 à -0,7 kcal.mol^{-1}), sauf dans le cas de la 2TX2 où l'on observe le phénomène inverse. Ces variations de score ainsi que l'analyse des poses obtenues permettent de penser que, d'après ces modèles, la présence de l'eau permet l'obtention de complexes protéine/ligand plus stables (**Figure 95 2TX**, **2TX2** et **2TX3**). D'autre part, les valeurs de RMSD sont plus faibles lorsque nous prenons en compte l'eau structurale. Cela valide le logiciel de docking car nous avons pu reproduire les résultats expérimentaux avec une bonne fidélité pour toutes les molécules. On remarque également que moins il y a d'eau structurale autour du ligand dans les structures expérimentales, moins la présence de cette eau a une influence sur les scores du docking (**Tableau 22** et **Figure 95 F3CS**). Grâce à ces résultats préliminaires nous avons validé le logiciel ADV qui a pu reproduire avec une certaine précision les données expérimentales dans une approche de docking classique, tout en notant que la présence de l'eau structurale semble importante dans la stabilisation du complexe protéine/ligand. Néanmoins pour les tests de docking qui suivront, nous ne pourrons pas prendre en compte l'eau structurale car elle risque de former un obstacle à la génération des conformations des ligands. Dans le cas des complexes protéine/ligand obtenus par cristallographie, l'eau structurale fait partie du système et s'est adaptée à la présence du ligand, ce

que nous ne pourrons pas modéliser par la suite. Nous enlèverons donc les molécules d'eau des structures des protéines pour les tests de docking sans *a priori*.

IV.3.4.3 Tests de plusieurs méthodes de docking sans *a priori*

Afin d'éprouver le docking sans *a priori*, nous avons testé deux logiciels ADV et AMIDE. AMIDE permet de réaliser un docking inverse, c'est-à-dire un docking d'un ligand sur plusieurs récepteurs, et propose deux moteurs de docking différents, ADV et AD4. Nous avons pu utiliser AMIDE dans le cadre d'une collaboration avec le laboratoire MEDyC de Reims. L'intérêt d'AMIDE réside dans l'exploration de plusieurs cibles en une seule fois, dans le découpage de ces cibles en plusieurs zones de recherche (permettant une recherche plus précise) et dans la répartition les tâches sur plusieurs CPUs (réduisant le temps de calcul nécessaire). Nous avons fait le choix de tester les deux moteurs de docking d'AMIDE, ce qui nous fait au total trois tests de docking sans *a priori* : un test avec ADV qui nous appellerons « test n°1, ADV », un deuxième test avec AMIDE et le moteur de recherche ADV « test n°2, AMIDE ADV » et un troisième test avec AMIDE et le moteur de recherche AD4 « test 3, AMIDE AD4 ». Lors du premier test nous n'explorons qu'une seule grande zone de recherche. Pour les tests n°2 et n°3 réalisés avec AMIDE, le système est découpé en 12 zones de recherches chevauchantes. Ces trois tests nous permettent de savoir s'il est nécessaire de découper la zone de recherche pour être exhaustif et de comparer les résultats des deux logiciels de docking ADV et AD4. Pour ces trois tests du docking sans *a priori*, seules les molécules 2TX, 2TX1, 2TX2 et 2TX3 ont été dockées dans les 30 structures représentatives de la flexibilité des quatre systèmes présentés précédemment (**Figure 96**). Nous avons choisi ces quatre molécules car elles présentent toutes les quatre différents IC_{50} sur les cibles *Ll*Fpg et hNEIL1. Nous souhaitons discriminer ces molécules grâce au docking et identifier un site de fixation lié aux effets d'inhibition observés *in vitro*. Les paramètres utilisés pour les trois tests sont décrits en Annexe D **VI.4** p. **328**.



	IC_{50} (µM)			
*Ll*Fpg	48	28	>500	15
hNEIL1	>500	>500	>500	19

Figure 96 : Ligands utilisés pour les tests du docking sans *a priori*

IV.3.4.3.1 Traitement des données

Les données sont déjà traitées en sortie de docking, que ce soit avec ADV ou AMIDE et les deux moteurs de docking proposés. À cela nous ajoutons un traitement par Classification Ascendante Hiérarchique (CAH, expliquée dans la section **III.1.9.4** p. **144** de la partie Méthodologie) pour identifier les sites de fixation des molécules sur les cibles. Les n conformations du ligand sont classifiées sur un critère de RMSD et seuls les représentants des *clusters* sont retournés à l'utilisateur. Les résultats sont donc un ensemble de conformations les plus diverses possibles générées par le docking tout autour du système ciblé. Dans un souci de clarté lors de la visualisation des résultats, chacune des conformations modélisées par le docking est réduite à son Centre Géométrique (CG). Ces CG sont classifiés par une méthode de CAH en n *clusters* selon un critère de distance euclidienne dans l'espace 3D (**Figure 97**). Chacun des *clusters* de CG peut être visualisé à la surface du système (**Figure 97**). Pour le calcul des CG et la CAH nous avons dû développer un script bash qui fait appel au langage R (Annexe E **VI.5** p. **331**). Après la CAH, il est possible d'extraire les scores des conformations du ligand, ainsi que d'identifier les aa en interaction (à 3,5 Å) avec ces dernières pour chaque site.

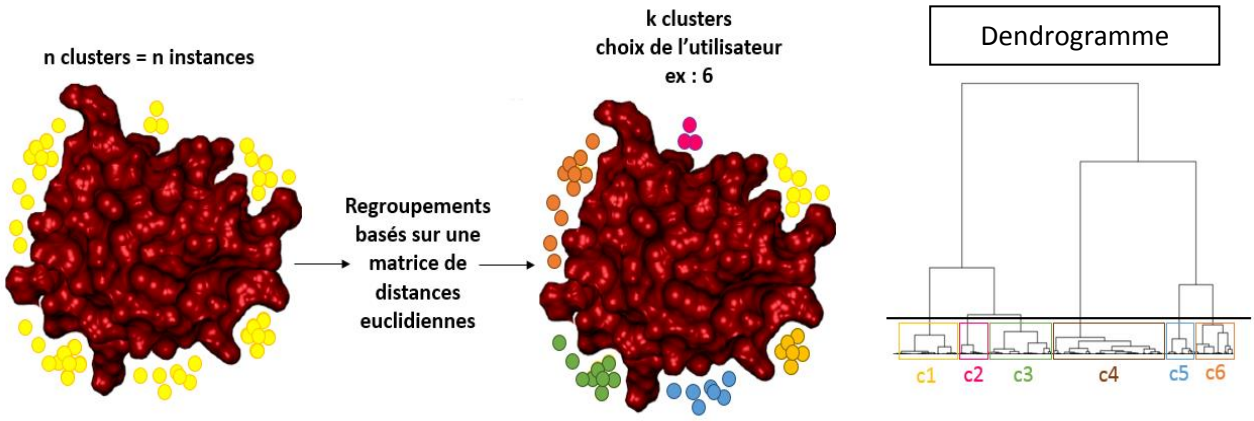


Figure 97 : Méthode de Classification Ascendante Hiérarchique (CAH) utilisées pour classifier les données issues des trois tests du docking sans *a priori*
À gauche, les sphères jaunes correspondent aux centres géométriques (CG) des conformations du ligand, elles ne sont alors pas encore classifiées. Au centre, les CG classifiés en 6 clusters. À droite, le dendrogramme correspondant à cette classification. La CAH est décrite dans la section **III.1.9.4** p. **144** de la partie Méthodologie.

IV.3.4.3.2 Comparaison des résultats des trois tests du docking sans *a priori*

La **Figure 98** présente les distributions des scores obtenus dans les trois tests de la méthode de docking sans *a* priori. Pour rappel, les scores d'affinité les plus négatifs indiquent une plus forte interaction entre le ligand et la protéine. Les poses issues du test n°1 (ADV) sont moins nombreuses

que les poses issues des tests n°2 (AMIDE, ADV) et n°3 (AMIDE, AD4) individuellement (**Figure 98 A**-**C**). En effet ADV ne permet l'obtention que de 20 conformations du ligand par boite au maximum. En sachant que nous avons réalisé la recherche dans 1 seule boite dans le test n°1, et dans 12 boites pour les tests n°2 et n°3 réalisés avec AMIDE, nous obtenons ~12 fois plus de poses du ligand dans ces deux tests. En comparant les images **A** et **B** de la **Figure 98**, on remarque que les scores minima sont les mêmes mais que les scores maxima ont une valeur plus élevée. Cela signifie que le test n°1 est aussi bon que le test n°2 pour trouver les meilleures conformations du ligand, mais que le test n°2 fourni d'autres poses du ligand aux scores plus élevés que ne retourne pas le test n°1. Cela peut être encore expliqué par le fait qu'ADV ne retourne que les 20 meilleures conformations. Cela signifie que le test n°2 est adapté pour générer une plus grande gamme de conformations des ligands, du plus stable au moins stable. En comparant B et C de la **Figure 98**, on remarque que les gammes de scores sont différentes. ADV semble proposer des conformations aux scores d'affinité compris entre -7,3 et 3,0 kcal.mol^{-1} et AD4 entre -6,3 et -2,0 kcal.mol^{-1}. Cette différence est reliée aux deux fonctions de scores d'ADV et d'AD4 qui n'évaluent pas les interactions avec les mêmes poids et avec les mêmes potentiels.

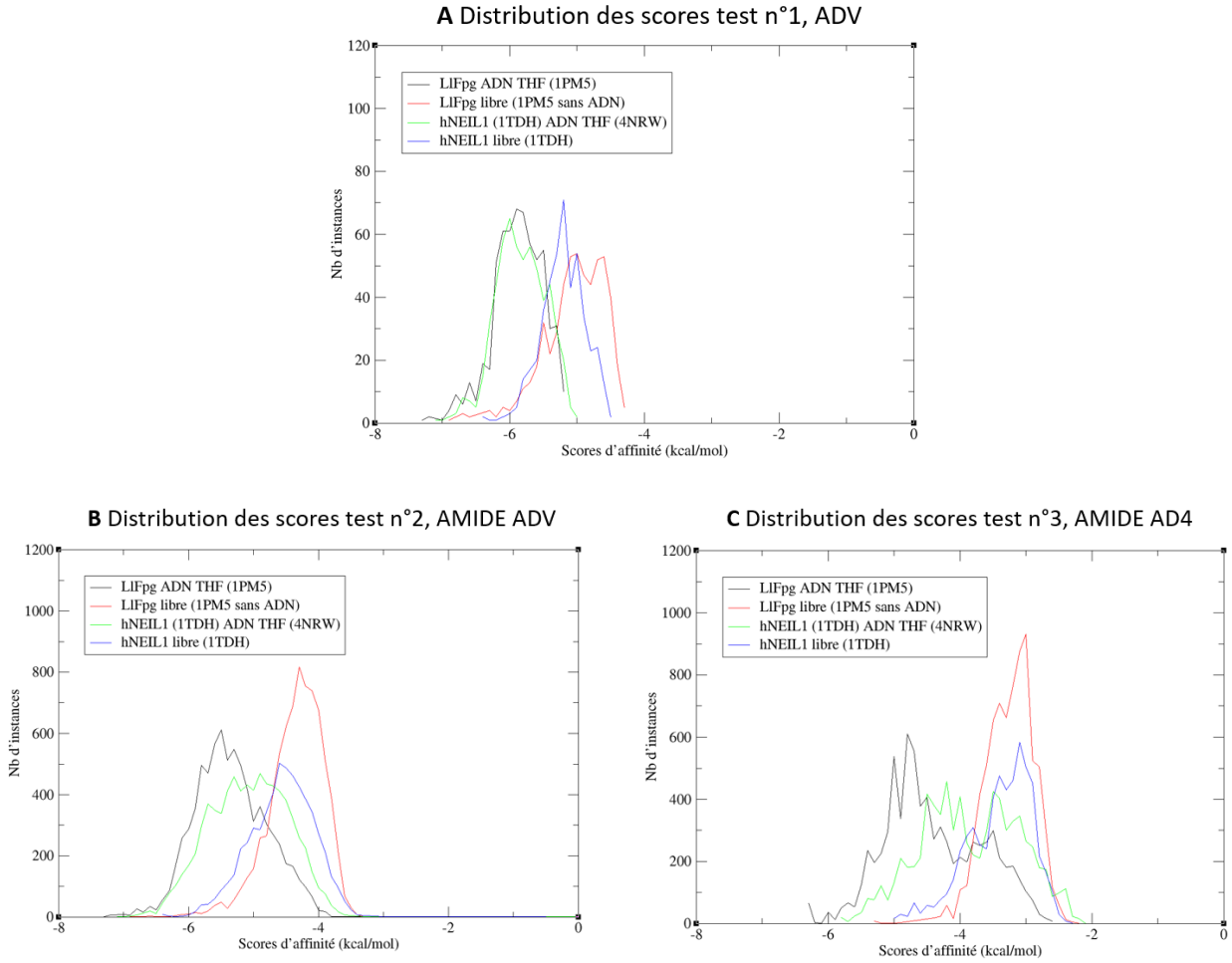


Figure 98 : Distributions des scores d'affinité obtenus lors du docking des 2TX, 2TX1, 2TX2 et 2TX3 dans les 4 modèles selon les différentes méthodes de docking utilisées
A) Test n°1, ADV, **B**) test n°2, AMIDE ADV et **C**) test n°3, AMIDE AD4.

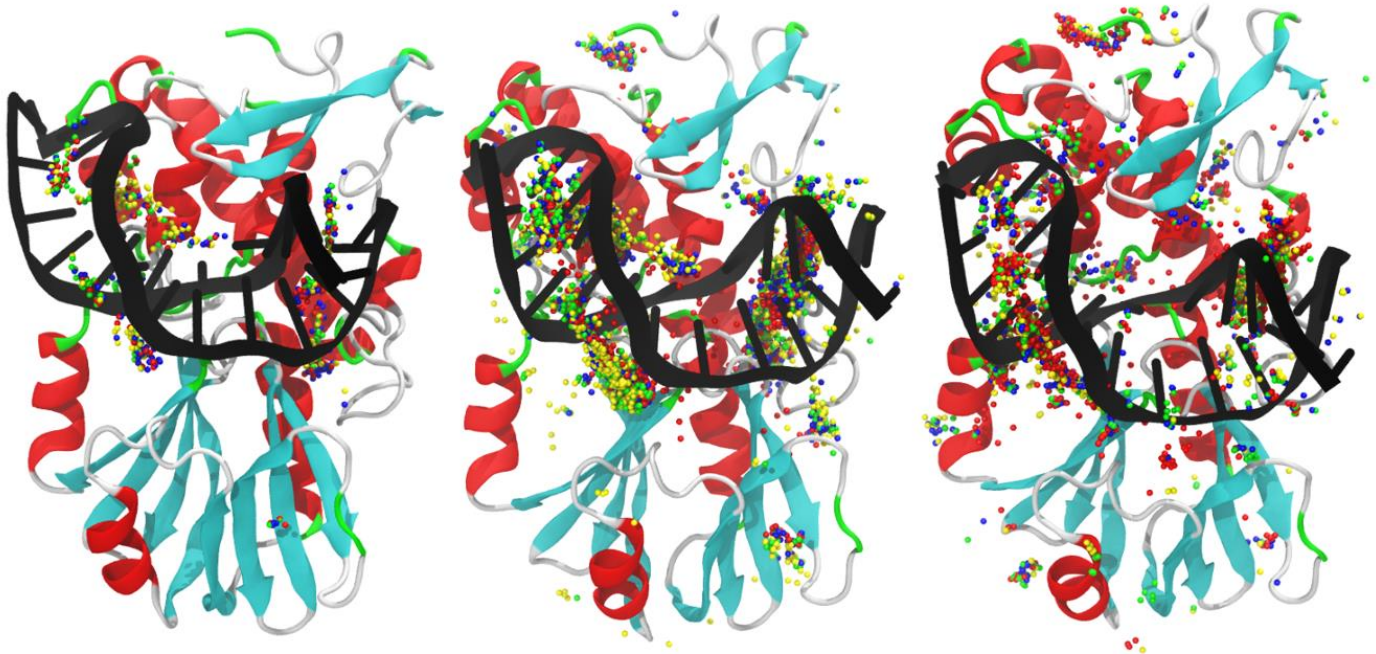
La comparaison des résultats de docking du test n°1 (ADV, 1 seule zone de recherche) avec les résultats des tests n°2 (AMIDE ADV, 12 zones de recherche) nous permet de conclure que le test n°1 est suffisamment exhaustif pour retrouver les meilleures conformations du ligand. De plus, nous remarquons que les différences dans les fonctions de scores d'ADV et d'AD4 ont une incidence sur les résultats car elles ne détectent pas les interactions protéine/ligand selon les mêmes critères et ne leur donne pas les mêmes poids dans le score.

D'autres part, en observant A, B et C de la **Figure 98**, les modèles contenant de l'ADN THF semblent contenir les conformations des ligands avec les scores d'affinité plus négatifs, ce qui suggère l'implication de l'ADN dans la création de sites interfaces plus affins que lorsque la protéine est libre. Ces sites d'interaction peuvent impliquer l'ADN (sites sur l'ADN ou bien à l'interface ADN/protéine), ou bien uniquement impliquer la protéine qui peut changer de conformation en présence d'ADN et ainsi proposer des sites d'interaction plus affins pour les ligands que lorsqu'elle est libre.

La **Figure 99** correspond à la représentation des centres géométriques (CG) des conformations des 2TX, 2TX1, 2TX2 et 2TX3 autour des protéines *Ll*Fpg et hNEIL1 libre et complexée à un ADN THF. Les CG permettent de visualiser plus clairement les résultats du docking. En observant cette figure, on remarque que les CG forment des groupes à la surface des récepteurs. Quatre couleurs sont assignées aux molécules 2TX, 2TX1, 2TX2 et 2TX3. Nous remarquons que les répartitions des quatre molécules sur la surface des quatre systèmes de *Ll*Fpg et hNEIL1 sont homogènes, signifiant que les sites de fixations prédits peuvent être occupés par les quatre molécules. Cela ne nous permet pas de définir un site de fixation comme étant lié aux effets d'inhibition des quatre molécules sur les cibles. On remarque également qu'en présence d'ADN, les molécules ont tendance à former des sites de fixation autour de ce dernier, faisant parfois intervenir la protéine dans ces sites dit à « l'interface » (composés de la protéine et de l'ADN). Sur cette figure, nous pouvons observer qu'une grande partie des surfaces des protéines interagit avec les poses des ligands. Néanmoins, nous pouvons affiner l'analyse sur les meilleurs points d'ancrage en sélectionnant les CG possédant les scores d'affinités les plus négatifs, traduisant une plus forte affinité protéine/ligand.

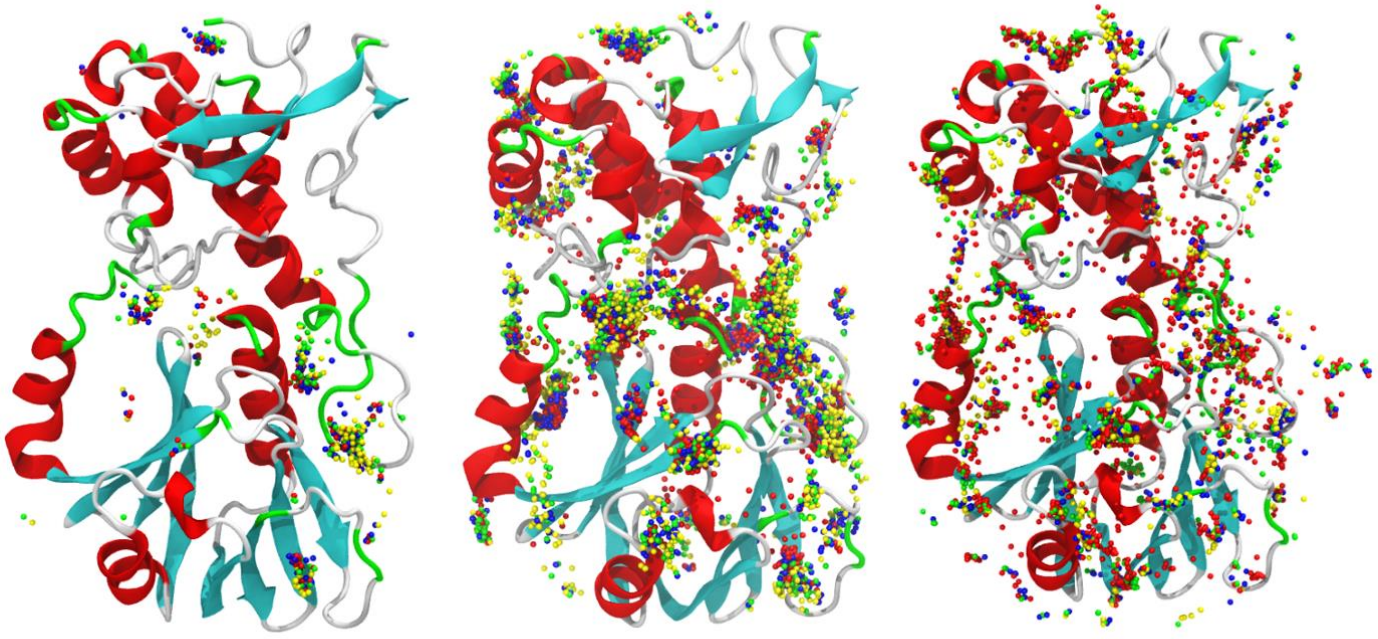
D'après la **Figure 99 A** et **C**, des conformations des ligands 2TX, 2TX1, 2TX2 et 2TX3 sont retrouvées sur la quasi-totalité de la surface du petit sillon ainsi qu'au niveau du grand sillon de l'ADN quelle que soit la méthode de docking. Sur les images **A** et **B**, on retrouve également des poses des ligands à proximité du ZnF avec les méthodes de docking réalisées avec ADV exclusivement. En dehors de l'ADN, plusieurs sites d'ancrage sont prédits sur les deux protéines tels que le site actif et le motif H2TH. Ces sites d'ancrages sont prédits sur les deux protéines *Ll*Fpg et hNEIL1, avec et sans ADN, quelle que soit la méthode de docking utilisée ADV ou AD4 (sauf le motif H2TH sur hNEIL1 (1TDH) ADN THF (4RNW) avec AD4). D'autres sites de fixation sont également prédits et correspondent à l'extrémité C-terminale par exemple, mais les scores sont les plus élevés et nous avons préféré ne pas les inclure dans l'analyse qui va suivre. Les meilleurs scores d'affinité par ligand et par site en fonction du système et de la méthode de docking utilisé sont résumés dans le **Tableau 23**.

A *Ll*Fpg ADN THF (1PM5)



Test n°1, ADV | Test n°2, AMIDE ADV | Test n°3, AMIDE AD4

B *Ll*Fpg libre (1PM5 sans ADN)



Test n°1, ADV | Test n°2, AMIDE ADV | Test n°3, AMIDE AD4

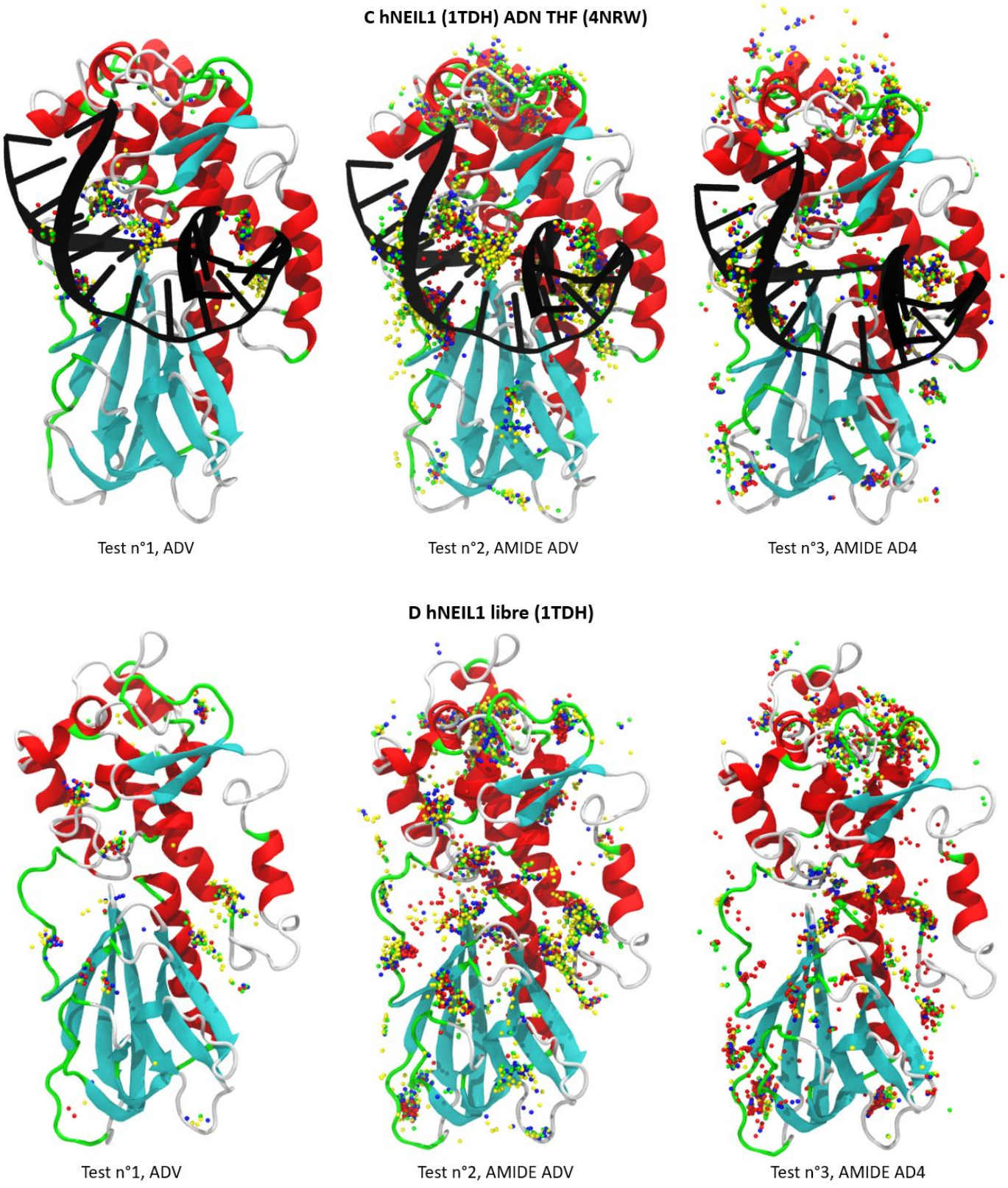


Figure 99 : Résultats des trois méthodes de docking, les sphères correspondent aux centres géométriques (CG) des quatre molécules dockées

Les sphères bleues représentent les CG de la 2TX, les rouges de la 2TX1, les jaunes de la 2TX2 et les vertes de la 2TX3. La protéine est représentée en cartoon et colorée en fonction de la structure secondaire, et en cartoon noir est représenté l'ADN.

Tableau 23 : Meilleurs scores des sites d'ancrage prédits dans les modèles étudiés de *Ll*Fpg et d'hNEIL1 libres ou complexés

Les scores (kcal.mol^{-1}) sont colorés du meilleur en vert au moins bon en rouge. Lorsqu'aucune conformation du ligand n'a été prédite, la case est laissée vide. Le site secondaire représente le motif H2TH. Le site petit sillon 1 est situé du côté du site actif et le site petit sillon 2 est localisé au niveau du motif H2TH, à proximité du site secondaire.

		***Ll*Fpg ADN THF (1PM5)**			***Ll*Fpg libre (1PM5 sans ADN)**			**hNEIL1 (1TDH) ADN THF (4NRW)**			**hNEIL1 libre (1TDH)**		
		Test n°1	**Test n°2**	**Test n°3**	**Test n°1**	**Test n°2**	**Test n°3**	**Test n°1**	**Test n°2**	**Test n°3**	**Test n°1**	**Test n°2**	**Test n°3**
Site actif	**2TX**	-6,3	-6,3	-4	-6,2	-6,2	-3,8	-6,1	-6,1	-3,6	-5,8	-5,8	-4,1
	2TX1	-7,3	-7,3	-4,2	-6,7	-6,7	-3,9	-6,6	-6,6	-3,9	-6,3	-6,4	-4
	2TX2	-6,9	-6,9	-4,4	-6,7	-5,9	-5,3	-6,7	-6,6	-4,4	-6,4	-6,4	-4,4
	2TX3	-6,4	-6,4	-4,1	-6,1	-6,1	-4	-6,5	-6,5	-3,8	-5,9	-5,9	-4
Site secondaire	**2TX**	-6,2	-6,2	-4,8	-5,2	-5,2	-3,7	-6	-6		-5,8	-5,8	-3,7
	2TX1	-6,5	-6,5	-4,6	-5,6	-5,6	-3,8	-6,4	-6,4		-6,4	-6,4	-3,5
	2TX2	-6,8	-6,8	-5,2	-5,9	-5,9	-3,9	-6,9	-6,9		-5,9	-5,9	
	2TX3	-6,4	-6,4	-4,7	-5,2	-5,2	-3,7	-6,2	-6,2		-5,6	-5,6	-4
ADN petit sillon 1	**2TX**	-6	-6,1	-5,4				-6,2	-6,2	-4,9			
	2TX1	-6,9	-6,9	-5,5				-7	-7	-5,5			
	2TX2	-6,6	-6,5	-6				-6,7	-6,7	-5,3			
	2TX3	-6,2	-6,2	-5,3				-6,8	-6,8	-5,5			
ADN petit sillon 2	**2TX**	-5,9	-6	-5,2				-6,3	-6,3	-5,6			
	2TX1	-7,2	-7,2	-5,6				-6,7	-6,8	-5,2			
	2TX2	-7	-7	-6,3				-7,1	-7,1	-5,8			
	2TX3	-6,3	-6,2	-5,5				-6,3	-6,3	-5,4			
ADN grand sillon	**2TX**	-6	-6					-6,2	-6,2				
	2TX1	-6,1	-6,2					-6,5	-6,5				
	2TX2	-6,6	-6,6					-6,8	-6,8				
	2TX3	-5,7	-5,7					-6,1	-6,1				
ZnF ou ZnLF	**2TX**							-5,5	-5,5	-4,4	-5,6	-5,6	-4,6
	2TX1							-5,8	-5,8	-4,2	-6,2	-6,2	-4,9
	2TX2								-5,6	-4,9	-5,6	-5,6	-5
	2TX3								-5,4	-4,4	-5,6	-5,7	-5

Dans le **Tableau 23**, nous remarquons que les sites d'interaction prédits par les 3 tests sont les mêmes, à l'exception du site secondaire dans le modèle de hNEIL1 (1TDH) ADN THF (4RNW) et du grand sillon de l'ADN qui ne sont pas prédits par le logiciel AD4. On observe une différence dans les scores d'affinités prédits par ADV (tests n°1 et n°2) et AD4 (test n°3), cela est lié aux différences dans les fonctions de score implémentées dans les deux logiciels (décrites dans la section **III.2.2** p. **151** partie Méthode), de ce fait, les scores d'affinités des tests n°2 et 3 sont difficiles à comparer. Les scores d'affinités prédits par les tests n°1 (1 zone de recherche) et n°2 (12 zones de recherches) pour un ligand donné et dans un modèle donné sont presque identiques (à 0,1 kcal.mol^{-1} près). Cela signifie que la méthode ADV est tout aussi exhaustive que la méthode AMIDE pour trouver les meilleurs modes d'interaction protéine/ligand. Les résultats donnés par le test n°1 sont donc satisfaisants ce qui est en bon accord avec ce que nous avions observés sur les images **A** et **B** de la **Figure 98**. On remarque également que le logiciel AD4 fourni des scores d'affinités des sites de fixation sur l'ADN plus faibles que sur les protéines pour tous les ligands, ce qui est cohérent avec les images **B** et **C** de la **Figure 98**. On peut remarquer que les scores d'affinités du site secondaire augmentent (de 0,9 à 1,2 kcal.mol^{-1} selon les molécules) dans les modèles des protéines libres, suggérant une forte implication de l'ADN dans la formation de ce site que l'on peut désormais considérer comme une interface ADN/protéine. De plus, d'après les scores d'affinité générés par ADV, les molécules 2TX1 et 2TX2 semblent être les molécules les plus affines pour les sites de fixation prédits. Nous remarquons que les scores de 2TX1 sont meilleurs dans le site actif de *Ll*Fpg que dans le site actif de hNEIL1. Par ailleurs, d'après les scores d'affinités générés par AD4, c'est la molécule 2TX2 qui est la plus affine pour tous les sites de fixation prédits, quel que soit le modèle, *Ll*Fpg ou hNEIL1, libre ou complexées. Or, nous savons que 2TX1 et 2TX2 possèdent des IC_{50} de 28 µM et > 500 µM sur *Ll*Fpg et de > 500 µM et > 500 µM sur hNEIL1 (**Tableau 9** de l'Introduction p. **116**). Cela signifie les données issues d'ADV sont d'avantage corrélées aux données expérimentales que les données résultant de AD4. Cependant, les trois méthodes ont prédit les scores d'affinité équivalents (± 0,4 kcal.mol^{-1}) pour 2TX et 2TX3 dans tous les sites de fixation, alors que les IC_{50} de ces molécules sont de 28 µM et 15 µM sur *Ll*Fpg et hNEIL1 (**Tableau 9** de l'Introduction p. **116**). Les trois tests de docking ont donc échoué dans l'identification de 2TX3 comme un meilleur inhibiteur de ces cibles que 2TX.

Les **Figure 100** et **Figure 101** représentent les différences structurales entre les conformations des ligands résultant des tests n°1 et n°2 et des tests n°2 et n°3 respectivement. Ces figures nous permettent d'appréhender la différence structurale dans les conformations du ligand générées par les 3 tests du docking sans *a priori*. Ces figures ont été obtenues en calculant les RMSD de toutes les poses générées dans les 4 systèmes par le test n°X par rapport à toutes les poses générées par le test n°Y

pour un ligand donné et en gardant la valeur de RMSD la plus faible. Ce calcul est réalisé pour chacun des 4 ligands individuellement. Après observations des conformations issues du docking, nous considérons qu'une RMSD ≤ 1 Å entre deux conformations du ligand indique qu'elles sont très proches. Entre 1 et 6 Å, les conformations sont dans le même site de fixation mais n'ont pas la même orientation, et au-delà de 6 Å, les deux conformations du ligand sont dans deux sites de fixation différents. Dans la **Figure 100**, nous comparons les tests n°1 et n°2. Les conformations très similaires représentent presque 10% des résultats pour les deux tests. Les conformations des ligands dans le même site de fixation mais pas dans la même orientation représentent 43% des résultats. Les conformations étant dans deux sites différents représentent 47% des résultats. Si des 90% poses sont dans le même site de fixation mais pas dans la même orientation ou bien dans deux sites différents, c'est parce que le test n°2 permet de modéliser une plus grande gamme de modes d'interaction que le test n°1 qui ne fournit que les 20 meilleures pour la totalité de la surface du système, et que 12 fois plus de poses sont produites par le test n°2. La comparaison des résultats des tests n°2 et n°3 ne sera pas biaisée par ce fait car les deux tests ont produit des nombres similaires de conformations des ligands. La **Figure 101** montre que seulement 5% des conformations sont identiques, 42% sont dans le même site de fixation et 53% dans deux sites différents. Cela indique que très peu de conformations sont très similaires, 95% des résultats sont dans des orientations différentes dans un même site ou dans deux sites différents. Cela est également lié à la différence dans les fonctions de scores implémentées dans les deux logiciels ADV et AD4 et dans les algorithmes de recherches qui sont aussi très différents.

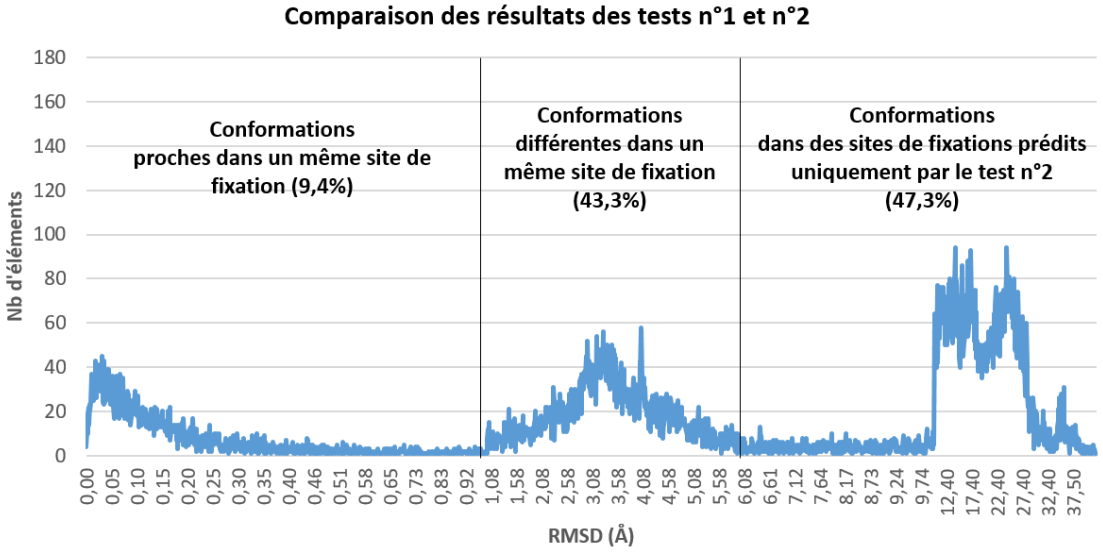


Figure 100 : Comparaison conformationnelle des poses générées par les tests n°1 et n°2 du docking sans *a priori*

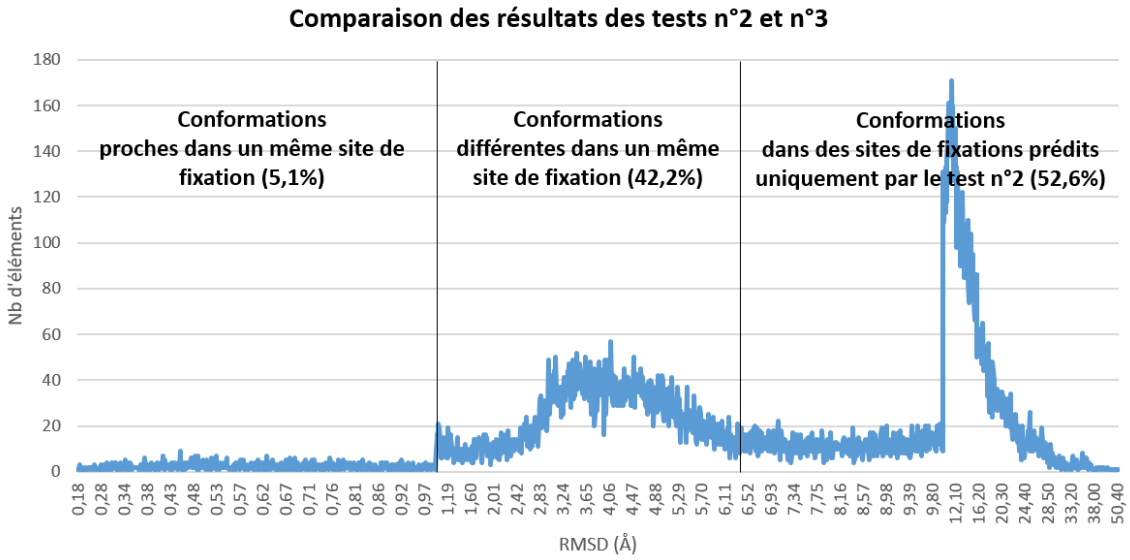


Figure 101 : Comparaison conformationnelle des poses générées par les tests n°2 et n°3 du docking sans *a priori*

Par la suite, nous n'avons étudié que les sites de fixation présentant un intérêt pour la conception d'inhibiteurs compétitifs, c'est-à-dire le site actif et le site secondaire. Un inhibiteur compétitif est une molécule capable de se lier à sa cible dans le site où se réalise la catalyse. Si le site actif est déjà occupé par une molécule, le substrat ne peut s'y placer et la catalyse ne peut pas avoir lieu. Le site secondaire est un site de fixation des ligands lorsque les protéines *Ll*Fpg et hNEIL1 sont libres. Il est localisé au niveau du motif H2TH, motif de fixation à l'ADN et est donc également un site intéressant pour la conception d'inhibiteur compétitif. Par conséquent, nous n'avons pas poursuivi l'analyse des sites de fixation au niveau des motifs ZnF/ZnLF, sur le petit et le grand sillon de l'ADN. Nous avons analysé les résultats des tests n°2 et n°3 car nous avons vu précédemment que les résultats retournés par AMIDE fournissaient une plus grande gamme de modes conformations. Pour vérifier que les sites d'amarrage prédits sont bien les mêmes quel que soit le moteur de docking utilisé (ADV ou AD4), nous avons comparé les aa formant les surfaces des sites de fixation. Pour cela nous avons identifié tous les résidus localisés à une distance inférieure ou égale à 3,5 Å des conformations des ligands générées lors des 3 tests du docking sans *a priori*. Ces aa sont représentés dans les **Figure 102** et **Figure 103** et sont listés dans le **Tableau 24**.

D'après le **Tableau 24**, les compositions en aa des sites actif et secondaire sont similaires, indépendamment du moteur de docking utilisé, ADV (test n°2) et AD4 (test n°3). Nous ne présenterons pas les résultats relatifs au test n°1 car ils sont très similaires à ceux obtenus dans le test n°2. Il semblerait que les aa qui ne sent pas retrouvés par les deux logiciels sont ceux situés en bordure des sites prédits, car les formes des surfaces des sites sont légèrement différentes sur les **Figure 102** et

Figure 103. Les aa composant le site actif sur *Ll*Fpg complexée et libre sont les aa catalytiques P1, E2 et E5 ainsi que les aa intercalants de l'ADN M75 et parfois F111, ainsi que certains aa de la boucle LCL G216-K230 (**Tableau 24 A**-**B**). D'autres aa forment la surface du site actif de *Ll*Fpg et sont localisés sur d'autres parties de la protéine dont le rôle n'est pas encore bien connu. Dans les modèles de hNEIL1, les aa catalytiques P1 et E5 ne font partie du site actif uniquement lorsque l'enzyme est libre (**Tableau 24 C**-**D**). D'autres aa interviennent dans la composition du site actif, nous verront plus loin dans le manuscrit qu'ils sont relatifs à un motif équivalent à la boucle LCL chez *Ll*Fpg. Le site secondaire chez *Ll*Fpg semble plus petit lorsque l'enzyme est complexée que lorsqu'elle est libre : 9-10 aa contre 14-26 respectivement (**Tableau 24 A**-**B**). Il est composé de L161, E162 et Q163 qui sont les aa avec lesquels interagissaient les ligands dans les structures de *Ll*Fpg co-cristallisée avec 2TX, 2TX2, 2TX3 et F3CS. Dans les modèles de hNEIL1, ces aa correspondent à la L165, D166 et Q167 (**Tableau 24 C**-**D**) qui font également partis du site secondaire prédit dans les modèles de l'enzyme complexée et libre (**Figure 125** p. **338** de l'Annexe H **VI.8** p. **338**). La prédiction du site secondaire est donc validée par les données expérimentales. Concernant le site actif, nous ne possédons pas de structure de *Ll*Fpg co-cristallisée avec les $2TX_n$ présentant ces molécules dans le site catalytique. Ce fait suggère que le site actif est très flexible, notamment par la présence de la boucle LCL.

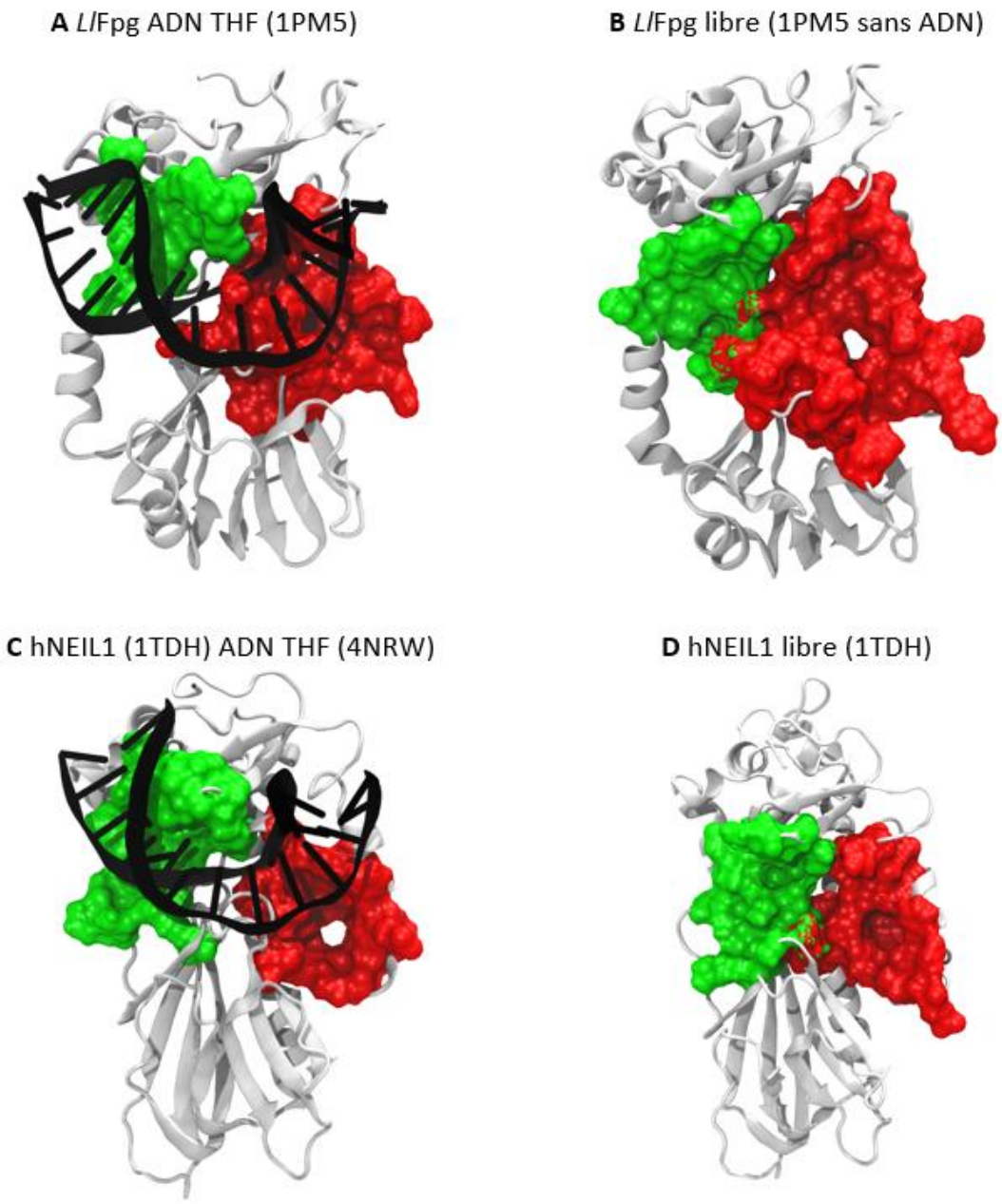


Figure 102 : Sites de fixation prédits par le test n°2 du docking sans *a priori*
Site actif (surface rouge) et site secondaire (surface verte) prédits par ADV par le deuxième test du docking sans *a priori* dans le modèle **A**) *Ll*Fpg ADN THF (1PM5), **B**) *Ll*Fpg libre **C**) hNEIL1 (1TDH) ADN THF (4NRW) et **D**) hNEIL1 libre (1TDH).

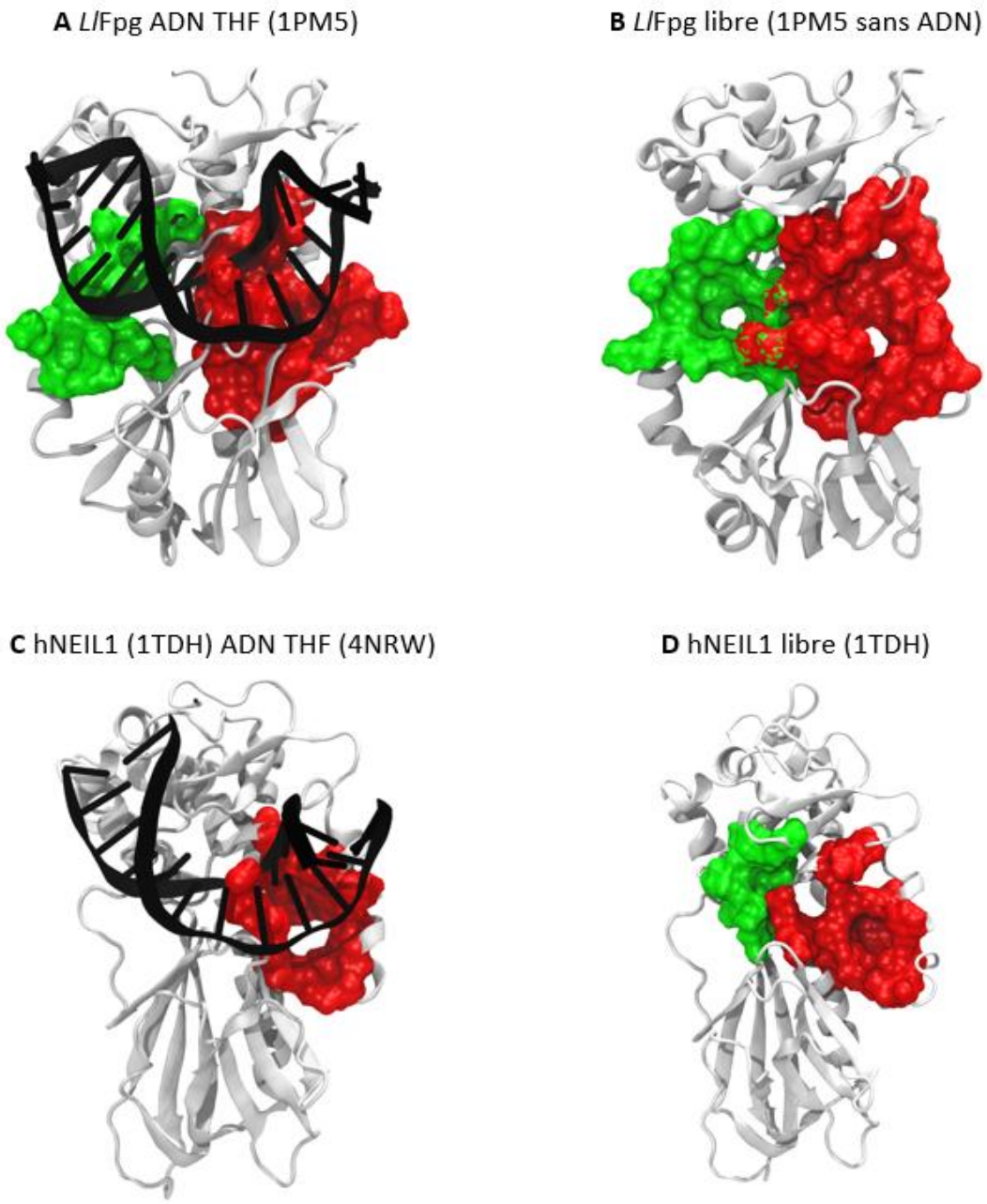


Figure 103 : Sites de fixation prédits par le test n°3 du docking sans *a priori*
Site actif (surface rouge) et site secondaire (surface verte) prédits par AD4 par le dernier test du docking sans *a priori* dans les modèles **A**) *Ll*Fpg ADN THF (1PM5), **B**) *Ll*Fpg libre **C**) hNEIL1 (1TDH) ADN THF (4NRW) et **D**) hNEIL1 libre (1TDH).

Tableau 24 : Aa composant le site actif et le site secondaire prédits par les tests n°2 et n°3 du docking sans *a priori*

Les aa retrouvés dans les deux méthodes sont mis **en gras**. Dans la suite de ce manuscrit, nous avons assigné la couleur rouge au site actif et la couleur verte au site secondaire.

		Site actif	Site secondaire
A FD	**Test n°2 (ADV)**	**P1**, **E2**, P4, **E5**, V6, T8, **L73**, R74, **M75**, **E76**, **G77**, **K78**, **Y79**, V108, **I172**, **Y173**, G215, G216, **S217**, **S218**, **I219**, **R220**, **T221**, **Y222**, S223, **A224**, **L225**, **G226**, **S227**, Q232, **Y238**	**K128**, K154, P158, Y159, **L161**, **E162**, **Q163**, A258, G259, R260
	Test n°3 (AD4)	**P1**, **E2**, **E5**, V9, **L73**, **M75**, **E76**, **G77**, **K78**, **Y79**, **I172**, **Y173**, **S217**, **S218**, **I219**, **R220**, **T221**, **A224**, **L225**, **G226**, **S227**, L235, **Y238**	Y58, Y125, **K128**, K129, K130, I131, **L161**, **E162**, **Q163**
B FS	**Test n°2 (ADV)**	**P1**, **E2**, P4, **E5**, **V6**, **V9**, **K57**, **L73**, **R74**, **M75**, **E76**, **G77**, **K78**, **Y79**, H91, A106, D107, V108, F111, G112, **I172**, **Y173**, G216, **S217**, **S218**, **I219**, R220, **T221**, Y222, **A224**, **L225**, **G226**, **S227**, G229, M231, **Q232**, **L235**, **Y238**	P1, E2, L3, R55, **G56**, **K57**, **Y58**, **H72**, **R74**, **K129**, **K130**, **I131**, P158, Y159, L160, **L161**, **E162**, **Q163**, **T164**, V166, A167, **G168**, L169, G170, I171, R260
	Test n°3 (AD4)	**P1**, **E2**, **E5**, **V6**, **V9**, **K57**, **L73**, **R74**, **M75**, **E76**, **G77**, **K78**, **Y79**, N171, **I172**, **Y173**, **S217**, **S218**, **I219**, **T221**, **A224**, **L225**, **G226**, **S227**, **Q232**, **L235**, **Y238**	**G56**, **K57**, **Y58**, **H72**, **R74**, K128, K129, K130, I131, **L161**, **E162**, **Q163**, **T164**, **G168**
C ND	**Test n°2 (ADV)**	**Y176**, **V235**, **V236**, Q237, **L238**, **G239**, **G240**, R241 **G242**, **Y243**, **G244**, **S245**, **G248**, E249, **D251**, **F252**, A253, **F255**	Q129, P130, G131, R132, R158, E152, A153, L165, D166, Q167, R168, G174, D272, R273, H274, R276
	Test n°3 (AD4)	G79, M80, **Y176**, **V235**, **V236**, **L238**, **G239**, **G240**, **G242**, **Y243**, **G244**, **S245**, **G248**, **D251**, **F252**, **F255**	-
D NS	**Test n°2 (ADV)**	**P1**, P4, **E5**, L6, **L8**, **Y176**, **V235**, **L238**, **G239**, **G242**, **Y243**, **G244**, **S245**, E246, G248, **D251**	P1, **E2**, **G3**, R51, **G52**, **K53**, E54, R77, G79, R132, **L165**, D166, **Q167**, **F170**, **N171**, **G172**, **I173**, G174, N175, H274, **R276**
	Test n°3 (AD4)	**P1**, E2, **E5**, **L8**, **Y176**, **V235**, **L238**, **G239**, **G242**, **Y243**, **G244**, **S245**, **D251**, F252, F255, Y262	**E2**, **G3**, **G52**, **Y53**, **L165**, **Q167**, **F170**, **N171**, **G172**, **I173**, **R276**

Abréviations : ● **Modèles** : FD, *Ll*Fpg ADN THF (1PM5) ; FS, *Ll*Fpg libre (1PM5 sans ADN) ; ND, hNEIL1 (ATDH) ADN THF (4NRW) ; NS, hNEIL1 libre (1TDH).

Pour comparer ces sites d'interaction entre les deux protéines *Ll*Fpg et hNEIL1, nous avons réalisé un alignement des aa communs prédits (en **gras** dans le **Tableau 24**) lors des tests n°2 (ADV) et

n°3 (AD4) du docking sans *a priori* sur les deux enzymes complexées (**Figure 104**) ou libre (**Figure 105**). L'alignement est basé sur les structures cristallographiques de 1PM5 et 1TDH correspondant respectivement à ces deux protéines (voir alignement complet à l'Annexe F **VI.6** p. **335**). Ces deux figures montrent qu'il semble exister des différences dans les séquences en aa rentrant dans la composition des sites actif et secondaire selon la présence ou l'absence d'ADN dans les modèles. En effet, le site d'interaction correspondant au site actif de hNEIL1 avec ADN THF ne fait pas intervenir les aa catalytiques P1, E2 et E5 contrairement au système hNEIL1 libre suggérant une différence dans la conformation de ce site entre les deux modèles. Nous ne retrouvons pas cette différence chez *Ll*Fpg. Le site de fixation correspondant au site actif fait intervenir la M75 de la triade d'intercalation chez *Ll*Fpg uniquement. De plus, il semblerait que la boucle LCL (216-230) uniquement présente chez *Ll*Fpg intervienne aussi dans ce site de fixation. L'alignement 1PM5/1TDH ne présente pas de structure homologue à la boucle LCL chez hNEIL1, cependant la portion de la protéine hNEIL1 alignée sur la boucle LCL de *Ll*Fpg est également en interaction avec les ligands. Ces résultats suggèrent qu'un élément structural différent de la boucle LCL de *Ll*Fpg intervient dans la fixation de molécules dans le site actif de hNEIL1. En regardant les structures de hNEIL1, on remarque que cette portion de la protéine correspond à une partie des hélices G et H ainsi que du Tour de 5 aa localisé aux positions G241 - S246 entre ces deux hélices (**Figure 124** p. **337** en Annexe G **VI.7** p. **337**). La boucle LCL de *Ll*Fpg est également composée d'un tour de 11 aa entre les positions S218 - G229. Cette différence dans ces éléments structuraux entrant en interaction avec les objets présents dans le site actif peut expliquer pourquoi *Ll*Fpg et hNEIL1 ne peuvent pas exciser les mêmes substrats, suggérant l'implication de la boucle LCL chez *Ll*Fpg et d'une structure Hélice Tour Hélice chez hNEIL1 dans la reconnaissance et/ou la stabilisation du substrat.

Le site considéré comme le site secondaire fait intervenir les aa L161, E162 et Q163 chez *Ll*Fpg en présence ou en l'absence d'ADN, et, comme vu précédemment, L165, D166 et Q167 chez hNEIL1 (**Figure 125** en Annexe H **VI.8** p. **338**). Ces aa sont strictement alignés sur les **Figure 104** et **Figure 105**, ils correspondent aux résidus en interaction avec les $2TX_n$ co-cristallisées avec *Ll*Fpg ADN THF, et font partie des deux tours formants le motif H2TH propre à la superfamille Fpg/Nei qui est, pour rappel, un motif de fixation de la protéine à l'ADN. Les données expérimentales représentant le site de fixation H2TH observé dans les structures cristallographiques de *Ll*Fpg avec les $2TX_n$ en Annexe H **VI.8** p. **338** permettent de valider la prédiction du site secondaire, car les aa composant le site d'interaction dans ces structures sont L161, E162 et Q163.

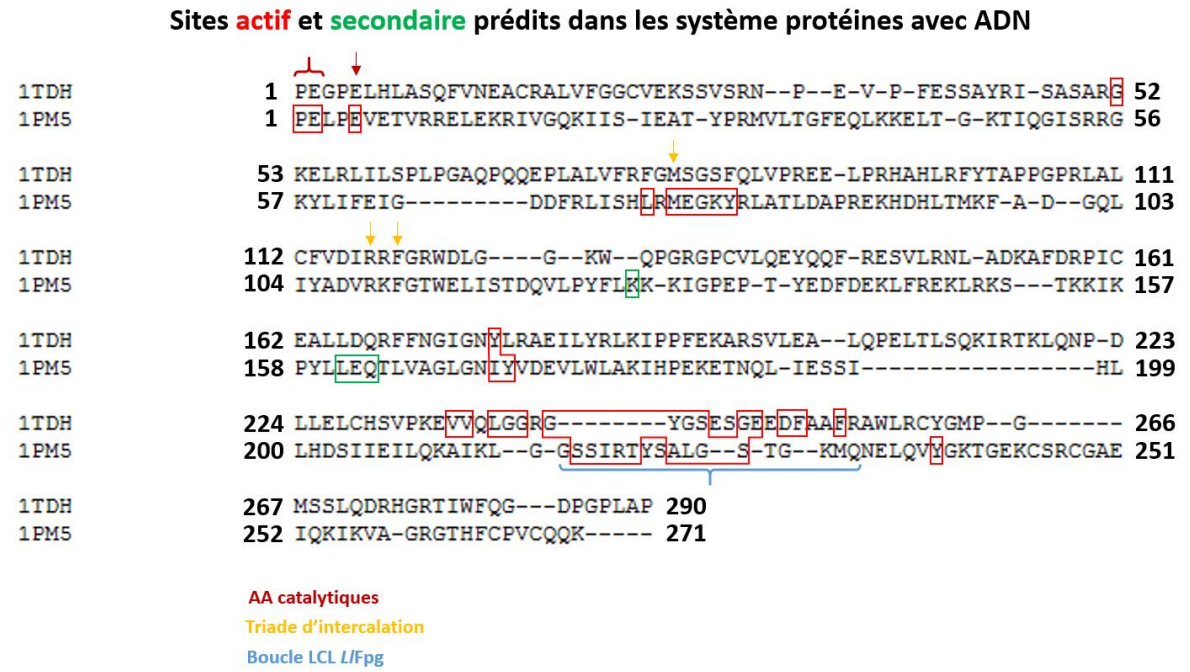


Figure 104 : Site actif et site secondaire prédits sur les protéines complexées à un ADN THF représentés sur un alignement structural de *Ll*Fpg (1PM5) et de hNEIL1 (1TDH)

Les aa encadrés correspondent aux résidus rentrant dans la composition du site actif (en rouge) et du site secondaire (en vert) sur les modèles des protéines complexées à un ADN THF (systèmes **A FD** et **C ND**) du **Tableau 24**. Ces aa sont prédits par les deux logiciels ADV et AD4 comme étant des points d'interaction entre les 4 ligands 2TX, 2TX1, 2TX2 et 2TX3 avec les protéines.

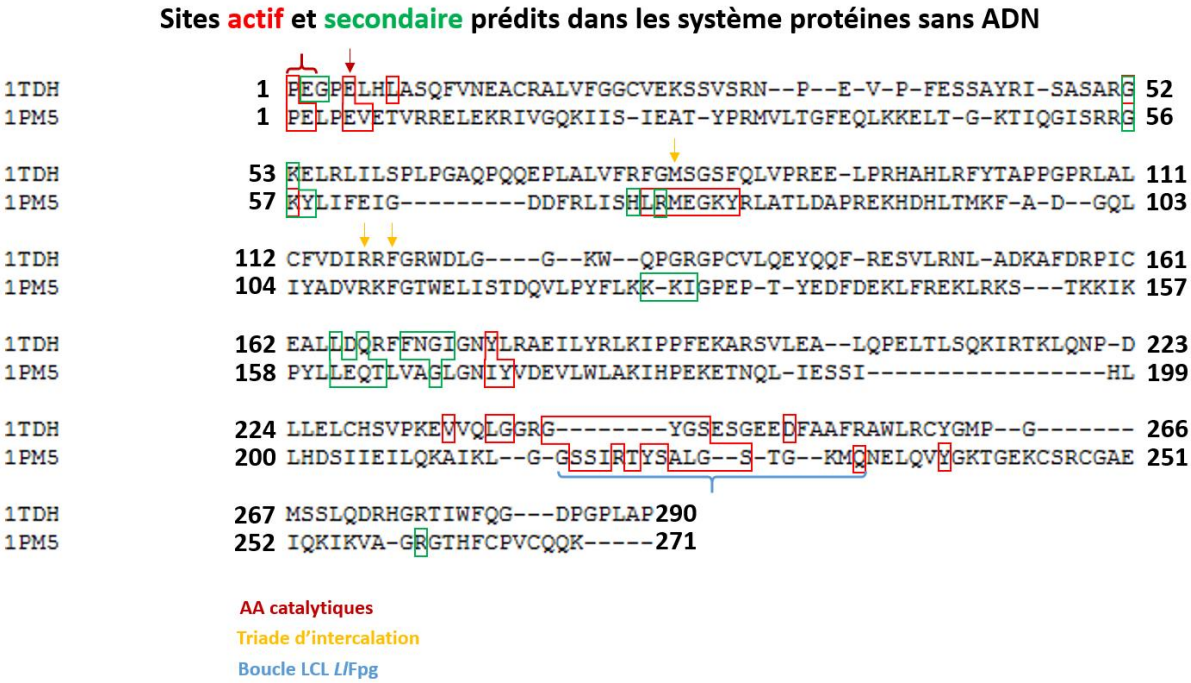


Figure 105 : Site actif et site secondaire prédits sur les protéines libres représentés sur un alignement structural de *Ll*Fpg (1PM5) et de hNEIL1 (1TDH)

Les aa encadrés correspondent aux résidus rentrant dans la composition du site actif (en rouge) et du site secondaire (en vert) sur les modèles des protéines libres (systèmes **B FS** et **D NS**) du **Tableau 24**. Ces aa sont prédits par les deux logiciels ADV et AD4 comme étant des points d'interaction entre les 4 ligands 2TX, 2TX1, 2TX2 et 2TX3 avec les protéines.

IV.3.4.4 Conclusion sur les tests du docking sans *a priori*

Dans cette section j'ai présenté les résultats des trois tests de docking sans *a priori*. Pour cela, nous avons choisi de prédire les sites d'amarrage des molécules 2TX, 2TX1, 2TX2 et 2TX3 sur les protéines *Ll*Fpg et hNEIL1 complexées à un ADN THF puis libres. Nous avons d'abord réalisé un premier test avec le logiciel ADV dans une seule boite de recherche, test réalisé au CBM. Puis, au MEDyC de Reims, nous avons pu utiliser un second protocole de docking grâce au logiciel AMIDE qui permet de réaliser un docking dans 12 boites de recherche chevauchantes avec les deux moteurs de recherche conformationnel ADV et AD4, correspondant aux tests n°2 et n°3. À partir des résultats des trois tests, nous avons pu alors isoler 6 sites d'interaction des 4 ligands sur les 4 systèmes, à savoir le site actif, un autre site appelé le site secondaire localisé à côté du motif H2TH, un troisième à proximité du ZnLF (hNEIL1 uniquement), ainsi que trois derniers sites d'amarrage localisés dans le petit et le grand sillon de l'ADN.

Nous avons pu identifier deux sites de fixation intéressants pour la conception d'inhibiteurs compétitifs, à savoir le site actif et le site secondaire sur les 4 systèmes étudiés. Ces deux sites de fixation ont des conformations différentes selon la présence ou non d'ADN dans le système, notamment le site actif de hNEIL1. En examinant les aa composant le site actif, nous avons également pu identifier une structure équivalente à la boucle LCL de *Ll*Fpg sur hNEIL1. D'autre part, cette structure est composée de deux hélices espacées par un tour. La structure secondaire de cet élément structural est donc très différente de celle de la boucle LCL. Nous avons également comparé le site secondaire au site de fixation des $2TX_n$ dans les structures de *Ll*Fpg co-cristallisées avec la 2TX, 2TX2, 2TX3 et F3CS. Les aa L161, E162 et Q163 forment le site prédit et sont les résidus qui interagissent avec les ligands dans les structures cristallographiques. Nous en avons déduit que la prédiction du site secondaire était validée par les données expérimentales. Dans la partie suivante, nous allons étudier en détail le site actif et le site secondaire.

En comparant les scores d'affinités ainsi que les poses des ligands produits par les trois tests de docking sans *a priori*, nous avons vu que le test n°1 est aussi précis que le test n°2 pour trouver les meilleurs modes d'interaction protéines/ligands. AMIDE permet donc de générer un plus grand panel de mode d'interactions autant et moins fort que le logiciel ADV seul. Les résultats des tests n°2 et n°3 ne sont pas similaires, cela est lié au fait que les fonctions de scores et les algorithmes de recherche des logiciels ADV et AD4 sont très différents. Cependant, les sites prédits par les deux logiciels sont très similaires, même si les conformations au sein de ces sites sont différentes. Comme le protocole du test n°1 est efficace pour prédire les meilleurs modes d'interaction protéine/ligand, nous utiliserons

par la suite cette même méthode pour prédire des modes d'interaction les plus forts entre les $2TX_n$ et les cibles *Ll*Fpg et hNEIL1 complexées et libres.

IV.3.5 Docking sans *a priori* des molécules $2TX_n$ sur *Ll*Fpg et hNEIL1

Pour notre prédiction de site de fixation des $2TX_n$ sur les cibles *Ll*Fpg et hNEIL1, nous avons réalisé un autre docking aveugle d'autres molécules dont les structures sont décrites dans la **Figure 106**. Le docking a été réalisé dans les 30 structures extraites par CAH à partir des simulations de DM. Nous avons analysé la composition en résidus du site actif et du site secondaire prédits sur *Ll*Fpg et hNEIL1.

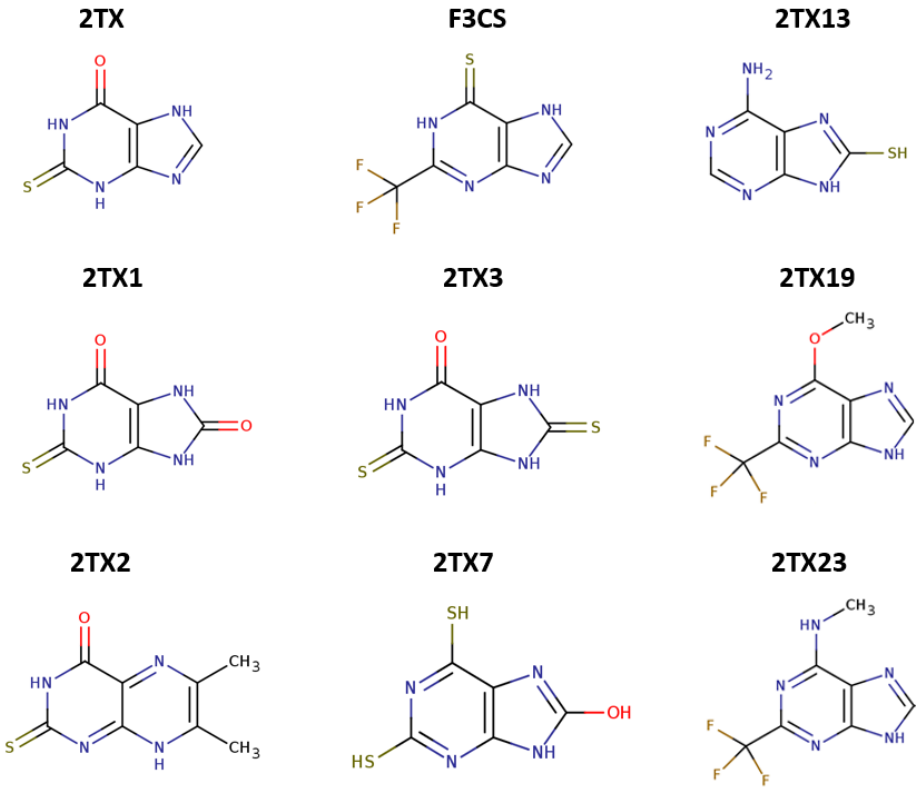


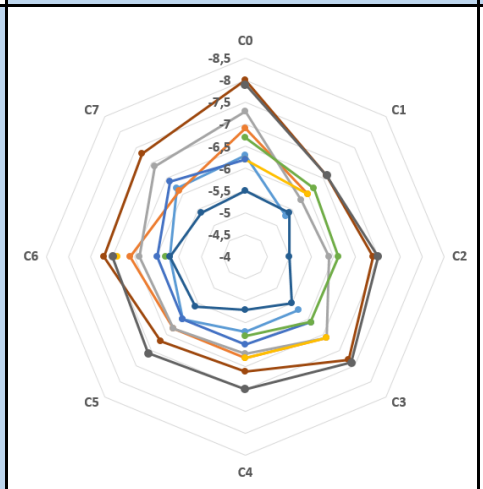
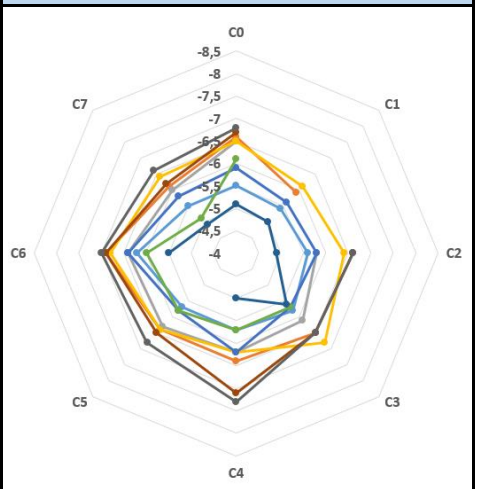
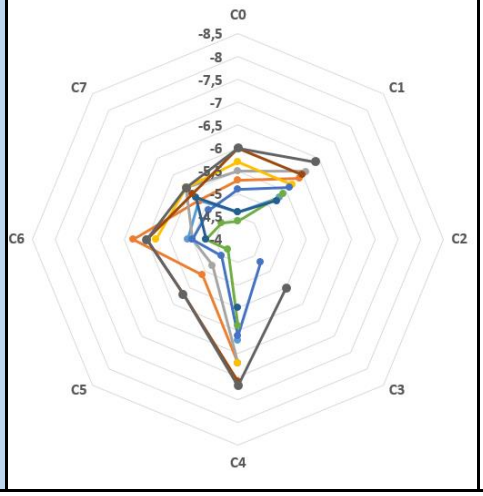
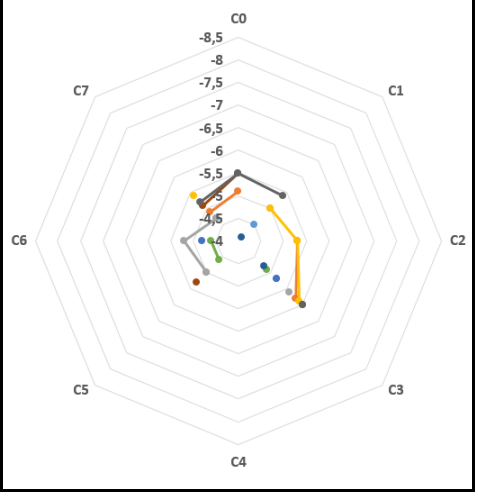
Figure 106 : Structures des molécules dockées pour la prédiction de site de fixation sur les cibles *Ll*Fpg et hNEIL1

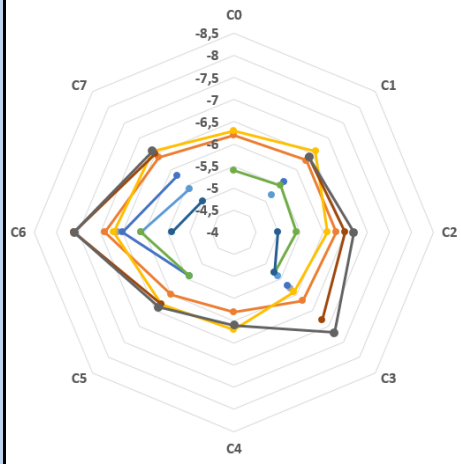
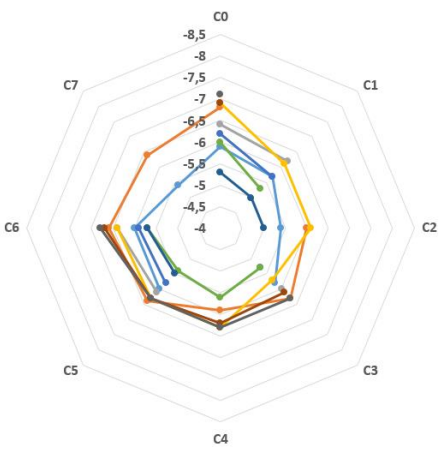
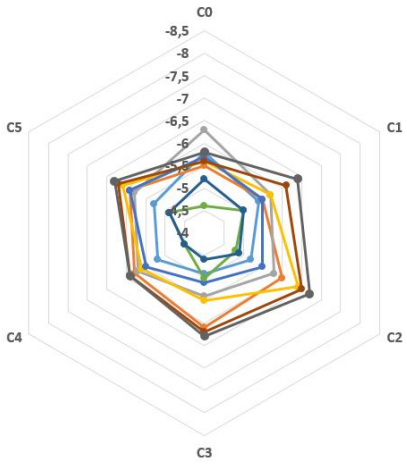
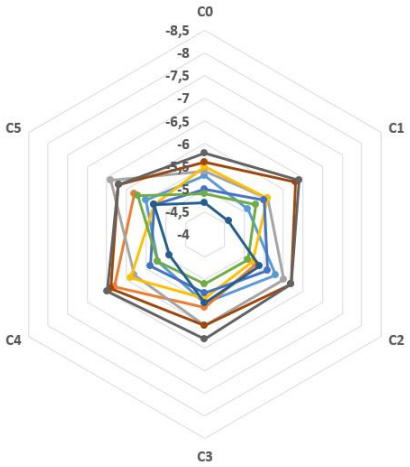
IV.3.5.2 Analyse des sites de fixation prédits sur *Ll*Fpg et hNEIL1

Dans le **Tableau 25**, nous remarquons déjà l'importance d'avoir réalisé le docking dans plusieurs structures extraites par CAH des simulations de DM. Si on regarde par exemple le système *Ll*Fpg libre (1PM5 sans ADN), on remarque qu'aucune molécule n'a été dockée dans le site actif dans le centroïde 2, ni dans le site secondaire (sauf 2TX2) dans ce même centroïde. Cela signifie d'abord que si nous n'avions exploité qu'une seule structure du système pour le docking, nous n'aurions peut-être

pas prédit le site actif comme site de fixation des molécules dans ce système. L'absence de prédiction dans un site de fixation résulte soit de la prédiction d'un site plus affin pour la molécule, soit de la flexibilité de ce dernier qui peut devenir trop étroit pour permettre l'interaction avec un ligand. Dans le cas du centroïde 2 de *Ll*Fpg libre (1PM5 sans ADN), un ou plusieurs autres sites de fixation que le site actif et le site secondaire ont été prédits, nous nous y intéresserons par la suite. Aussi, nous observons des différences dans les scores d'affinité pour une molécule selon le centroïde, ce qui suggère une certaine flexibilité des sites de fixation prédits.

Tableau 25 : Meilleurs scores recensés dans les deux sites de fixation principaux
Meilleurs scores (kcal.mol^{-1}) des poses générées dans le site actif et le site secondaire, par centroïde (CX) et par molécule dans les quatre systèmes étudiés.

	Site actif	**Site secondaire**
***Ll*Fpg ADN THF (1PM5)**		
***Ll*Fpg libre (1PM5 sans ADN)**		

hNEIL1 (1TDH) **ADN THF (4RNW)**		
hNEIL1 libre **(1TDH)**		

IV.3.5.3 Comparaison des résultats du docking sans a priori aux données expérimentales

Dans cette section, nous mettrons en parallèle les scores d'affinité calculés par le docking et les IC_{50} mesurés lors des tests d'activité *in vitro*.

La **Figure 107** représente les énergies d'affinité traduisant les plus fortes interactions et les IC_{50} obtenus dans l'équipe sur *Ll*Fpg et hNEIL1. D'après cette figure, si nous devions classer les molécules d'après leurs IC_{50} (une valeur faible traduit un fort effet d'inhibition sur la cible) sur *Ll*Fpg de l'effet le plus fort à l'effet le plus faible, nous aurions : F3CS < 2TX3 < 2TX13 < 2TX1 < 2TX7 < 2TX et un effet quasiment nul de 2TX2, 2TX19 et 2TX23. Les molécules ayant un plus fort effet sur hNEIL1 sont : 2TX3 < 2TX7 < 2TX13 < F3CS et un effet quasiment nul des molécules 2TX, 2TX1, 2TX2, 2TX19, 2TX23. Si on s'intéresse maintenant aux meilleurs scores d'affinité, nous remarquons qu'il n'y a pas de corrélation entre ces derniers dans le site actif et le site secondaire et les IC_{50}. Cela signifie que les effets d'inhibitions des molécules sur les cibles ne peuvent pas être expliqués par docking. Les raisons de cette observation seront discutées plus loin.

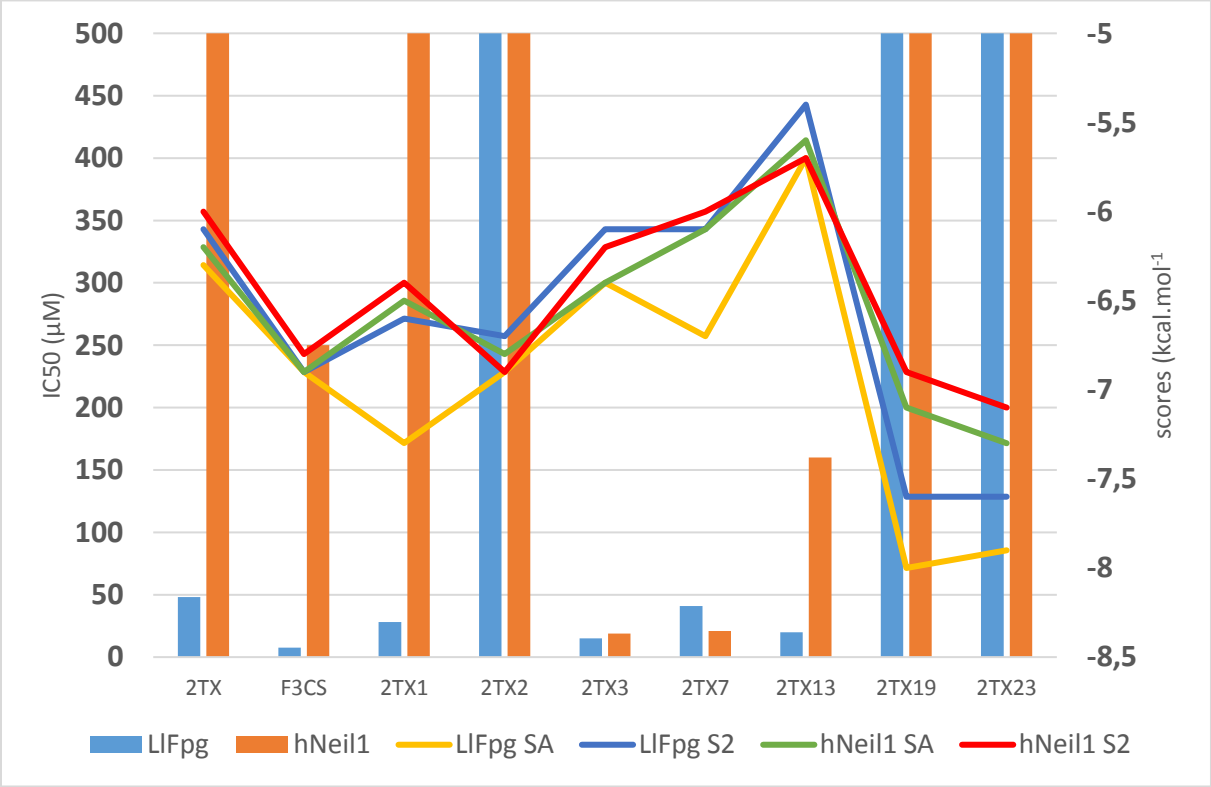


Figure 107 : Comparaison des scores de docking et des IC_{50} des molécules $2TX_n$ sur les cibles *Ll*Fpg et hNEIL1

IC_{50} des molécules obtenues dans l'équipe (barres pleines) et scores d'affinités (meilleurs scores) obtenues avec le docking aveugle (traits) dans le site actif et le site secondaire de *Ll*Fpg et de hNEIL1.
Abréviations : SA, Site Actif ; S2, Site secondaire.

Pour évaluer la qualité des prédictions nous avons tracé des courbes ROC (*Receiver Operating Characteristic*) qui permettent de visualiser le taux de vrais positifs en fonction du taux de faux positifs (**Figure 108**). Pour la production de ces courbes ROC, nous avons dû assigner arbitrairement une valeur binaire « active » (a un effet sur la cible, IC_{50} < 100 µM) ou « inactive » (n'a pas d'effet sur la cible, IC_{50} ≥ 100 µM) à chacune des molécules testées, et cela selon les IC_{50} mesurés sur les cibles étudiées (**Tableau 26**).

Tableau 26 : IC_{50} des molécules sur les cibles étudiées, ainsi que l'assignation des valeurs « active » et « inactive » en fonction de la valeur de ces IC_{50}

	***Ll*Fpg**		**hNEIL1**	
	IC_{50} (µM)	**Valeur**	**IC_{50} (µM)**	**Valeur**
2TX	48 ± 4	**Active**	> 500	**Inactive**
F3CS	8 ± 1	**Active**	> 250	**Inactive**
2TX1	28 ± 7	**Active**	> 500	**Inactive**
2TX2	> 500	**Inactive**	> 500	**Inactive**
2TX3	15 ± 1	**Active**	19 ± 1	**Active**
2TX7	41 ± 7	**Active**	21 ± 1	**Active**
2TX13	20 ± 2	**Active**	160 ± 16	**Inactive**
2TX19	> 500	**Inactive**	> 500	**Inactive**
2TX23	> 500	**Inactive**	> 500	**Inactive**

La **Figure 108** représente les courbes ROC permettant d'évaluer la qualité des résultats du docking. La courbe ROC est un outil statistique décrit dans la section **III.2.5.2** p. **170** de la partie Méthodologie Lorsque la courbe est en dessous de la diagonale, les prédictions sont mauvaises car nous prédisons plus de faux positifs que de vrais positifs et vice versa. Ici, les prédictions sont très mauvaises, nous discutons par conséquent des causes qui pourraient altérer les prédictions.

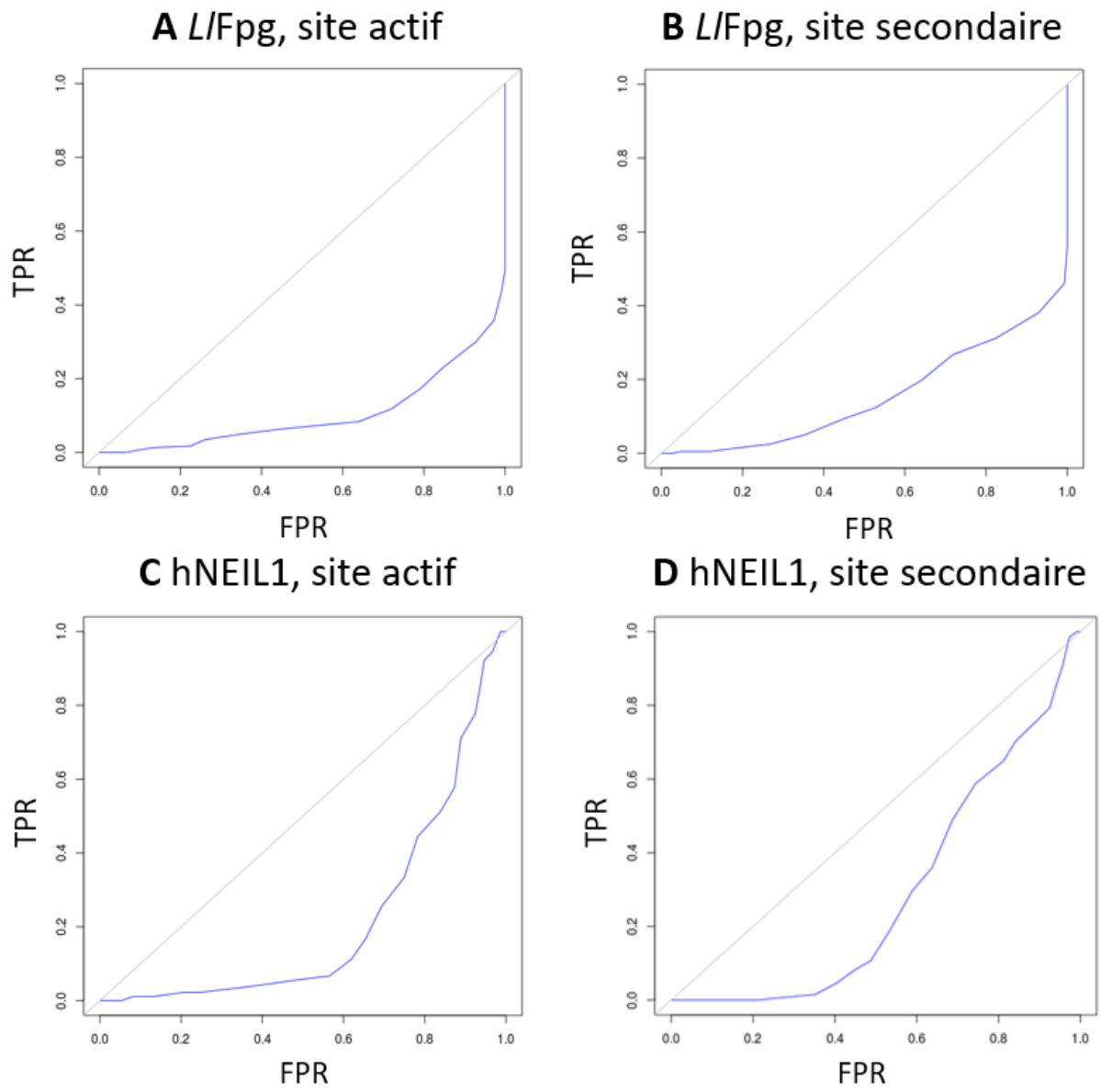


Figure 108 : Courbes ROC permettant d'évaluer la qualité de la prédiction du docking
Abréviations : TPR, « True Positive Rate » ; FPR, «False Positive Rate ».

La première hypothèse est que le choix du seuil séparant les molécules dites actives et inactives lors de la création des courbes ROC n'est pas le bon, faussant ainsi l'évaluation des prédictions.

La deuxième hypothèse est que l'inhibition des cibles par ces molécules passe par un autre mécanisme que celui que l'on suppose comme par exemple la formation de liaison covalente, ce qui est possible car les $2TX_n$ comportent toutes un groupement thione (=S) ou thiol (-SH) hautement réactif. Ce groupement fonctionnel aurait la capacité de former des liaisons covalentes avec les Cys de la protéine comme décrit dans la section **I.3.6** p. **115**. Pour rappel, il a déjà été observé dans des structures cristallographiques la formation de ponts disulfures entre les $2TX_n$ et les C248 et C268 du

ZnF de *Ll*Fpg, induisant l'incapacité de l'enzyme à se fixer à l'ADN (**Figure 41** p. **117**). Ce phénomène pourrait expliquer les effets d'inhibitions des $2TX_n$ sur *Ll*Fpg. En revanche, hNEIL1 ne possède pas de Cys dans son motif ZnLF et ne peut pas être affectée par le même mécanisme d'inhibition. Elle possède néanmoins 6 Cys dans sa structure. On note tout de même que la 2TX3 possède un effet d'inhibition sur cette cible dont le mécanisme pourrait passer par l'interaction de cette petite molécule avec le site actif de l'enzyme. Les mécanismes d'inhibition étant différents pour les deux cibles et pouvant être induits par des liaisons covalentes protéine/ligand, les courbes ROC seraient donc biaisées. Des études de docking covalent pourraient nous permettre de mieux comprendre quels sont les mécanismes impliqués dans l'inhibition de *Ll*Fpg et hNEIL1 par les $2TX_n$. Par ailleurs, des structures cristallographiques de hNEIL1 avec les molécules $2TX_n$ nous permettraient de comprendre les effets d'inhibitions de la 2TX3 sur cette enzyme.

La troisième hypothèse est que les $2TX_n$, du fait de la présence d'un atome de soufre très réactif dans leurs structures, pourraient se recombiner en polymères (des trimères pour la 2TX3 et 2TX7 car elle comporte deux atomes de soufre et des dimères pour toutes les autres car elles ne comportent qu'un seul atome soufre). Dans ce contexte, les molécules ayant un effet d'inhibition sur les cibles seraient potentiellement les dimères où trimères et non pas les monomères. Nous reviendrons sur cette hypothèse dans la section suivante.

La quatrième hypothèse se base sur le postulat que c'est bien l'interaction des molécules $2TX_n$ avec le site actif qui permet un effet d'inhibition des cibles. Dans ce contexte, le site actif n'est pas dans les meilleures conformations dans les structures extraites des simulations de DM. Cela ne permet pas de générer les conformations des ligands de plus basse énergie dans ce site d'amarrage. Un échantillonnage structural plus fin des systèmes produits par simulation de DM nous permettrait peut-être d'améliorer la qualité de nos résultats.

La cinquième hypothèse est que la fonction de score d'ADV, optimisée pour le criblage haut débit, est trop approximative pour évaluer l'affinité entre une petite molécule et la protéine cible. Une possibilité pour affiner les scores d'affinité est de réaliser un « rescoring » avec une méthode tierce (MM-PBSA par exemple).

IV.3.5.4 Aa composant les sites de fixation de $2TX_n$ sur *Ll*Fpg et hNEIL1

Dans le but de mieux comprendre les modes d'interaction des ligands dans les sites principaux prédits sur *Ll*Fpg et hNEIL1, nous avons analysé la composition en aa des sites de fixation des ligands

prédits par le docking (**Figure 110** et **Figure 112)**. Pour tous les résultats de docking obtenus dans les centroïdes des deux systèmes *Ll*Fpg et hNEIL1 et pour toutes les molécules, nous avons identifié les aa les plus fréquemment en contact (distance atome lourd/atome lourd ≤ 3,5 Å) avec les conformations des ligands prédites par le docking.

Dans la **Figure 110** nous remarquons que les aa catalytique P1 (22%) et E5 (25%) sont fréquemment en contact avec les poses générées dans le site actif par le docking dans le modèle de *Ll*Fpg avec ADN THF et dans une moindre mesure, les aa S217 (6%), I219 (9%) et T221 (10%). Lorsque *Ll*Fpg est libre, les aa catalytiques P1 (17%) et E5 (13%) interagissent moins fréquemment avec les poses des ligands, tandis que l'aa K78 (18%) à proximité de la M75 intercalante, et les aa S217 (10%), S218 (19%) de la boucle LCL forment plus souvent des contacts avec les poses des ligands localisées dans le site actif. Concernant hNEIL1 complexée à l'ADN, l'aa D251 (24%) est fréquemment contacté par les poses du ligand, ainsi que plus faiblement, les aa E5 (5%), L8 (12%), V235 (9%), L238 (7%), G242 (5%), Y243 (10%), F252 (6%) et F255 (5%). Concernant l'enzyme humain libre, les aa E5 (13%), Y243 (20%), D251 (19%), F252 (13%) et F255 (21%) semble interagir fréquemment avec les poses situées dans le site actif. Les aa V235-F255 font partie de la structure équivalente à la boucle LCL de *Ll*Fpg. D'après ces résultats, nous pouvons dire que les sites actifs de *Ll*Fpg et de hNEIL1 sont flexibles, car ils ne sont pas composés par les mêmes aa si les enzymes sont complexés ou libres. Les aa catalytiques P1, E2 et E5 composent en partie le site actif de *Ll*Fpg et seulement E5 le site actif de hNEIL1. La boucle LCL et la structure équivalente de hNEIL1 font partie du site actif. Les aa clés (fréquence d'interaction > 5%) du site actif de *Ll*Fpg sont P1, E5, K78, S217, S218, I219 et T221 (**Figure 109 A**), concernant hNEIL1, il s'agit de E5, L8, V235, L238, G242, Y243, D251, F252 et F255 (**Figure 109 B**).

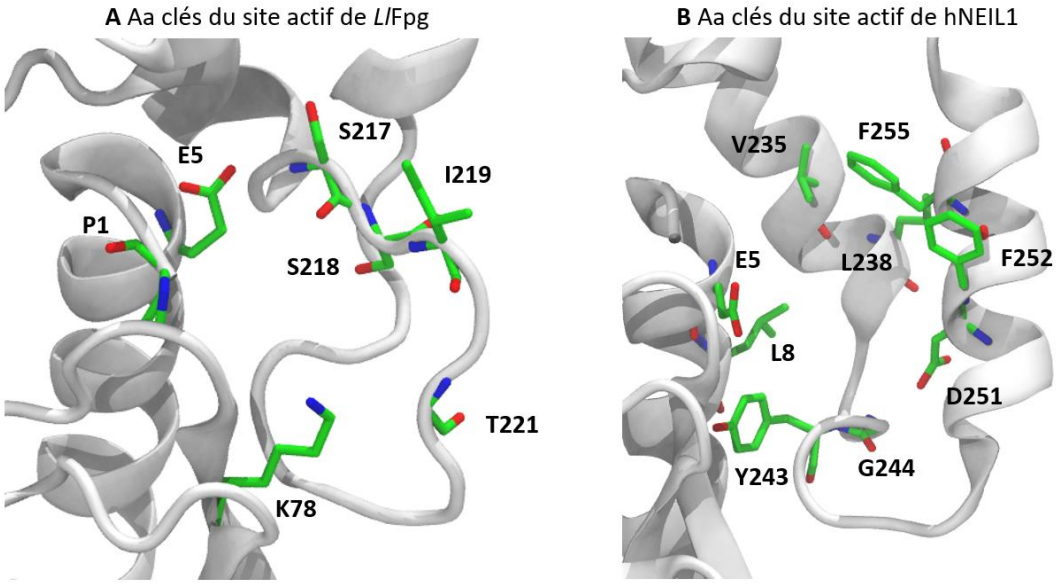
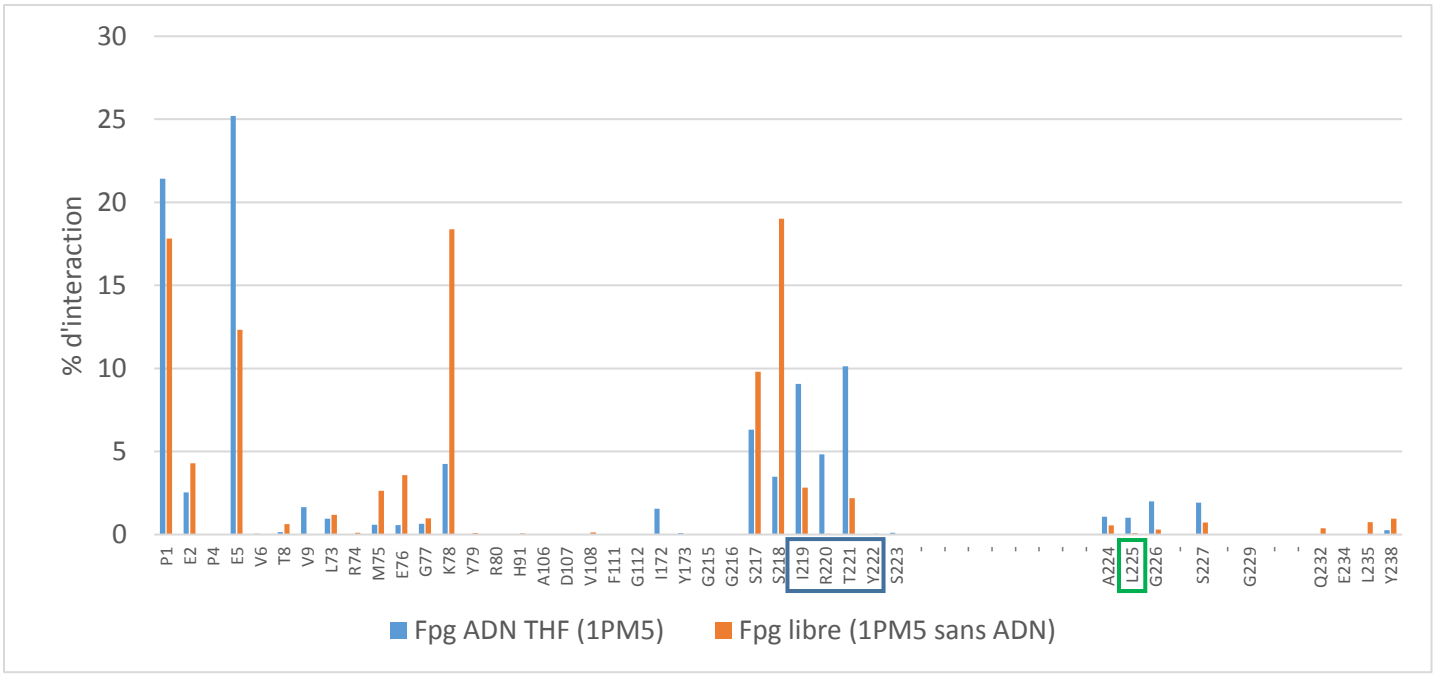


Figure 109 : Aa clés du site actif de *Ll*Fpg et de hNEIL1
La protéine en blanc et les aa clés sont représentés en bâtonnets.



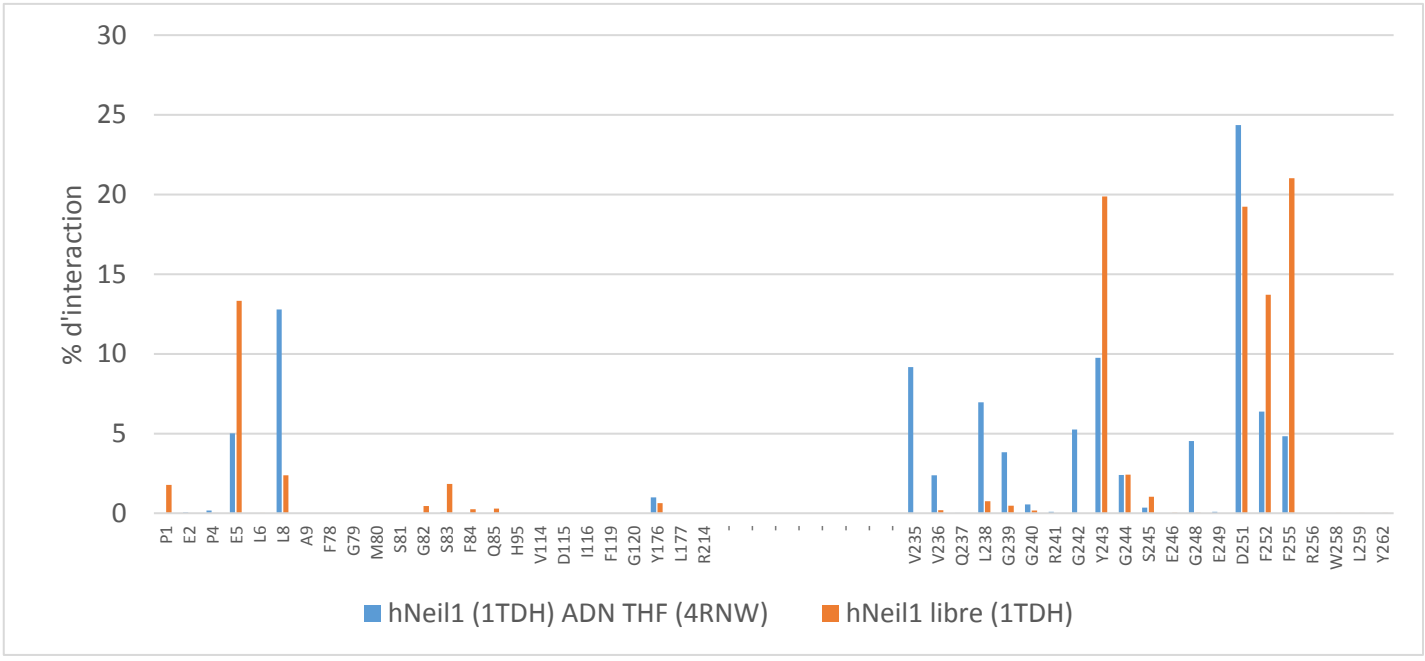


Figure 110 : Aa composant le site actif des modèles étudiés
Les systèmes avec ADN THF sont en bleu et ceux sans ADN sont en orange. Les aa en abscisses correspondent à un alignement de *Ll*Fpg et de hNEIL1 basé sur les structures cristallographiques de 1PM5 et 1TDH respectivement (voir alignement complet dans l'Annexe F **VI.6** p. **335**). Pour les modèles de *Ll*Fpg, nous avons encadré les aa importants de la boucle LCL identifiés dans le Chapitre 1 de la partie Résultats p. **175**, à savoir en bleu I219, R220, T221 et Y222, en vert l'aa charnière L225.

Le site secondaire de *Ll*Fpg est le plus souvent composé des aa K154 (6%), P158 (14%), L161 (23%), E162 (40%) et R260 (6%) lorsque l'enzyme est complexée et de K57 (14%), Y58 (8%), K129 (10%), E162 (8%) et Q163 (31%) lorsqu'elle est libre. Concernant hNEIL1, les aa composant le plus

fréquemment le site secondaire sont L165 (26%), D166 (22%), Q167 (18%), H274 (14%) de l'enzyme complexée et E2 (32%), K53 (5%), L165 (24%), Q167 (26%) et I173 (6%) de l'enzyme libre. Les aa clés du site secondaire de *Ll*Fpg sont donc K57, Y58, K129, I131, K154, P158, L161, E162, Q163 et R260 (**Figure 111 A**) et de hNEIL1, E2, K53, L165, D166, Q167, I173 et H274 (**Figure 111 B**). Le site secondaire est composé du motif H2TH et d'une petite partie des ZnF/ZnLF.

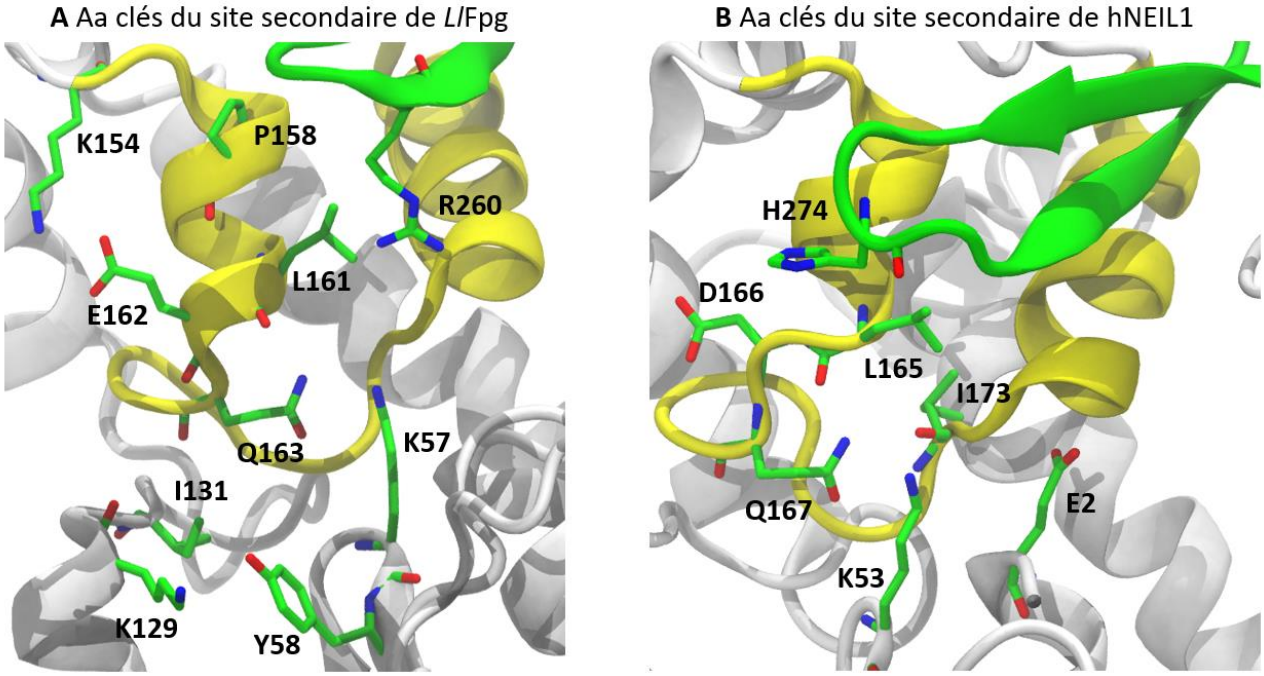
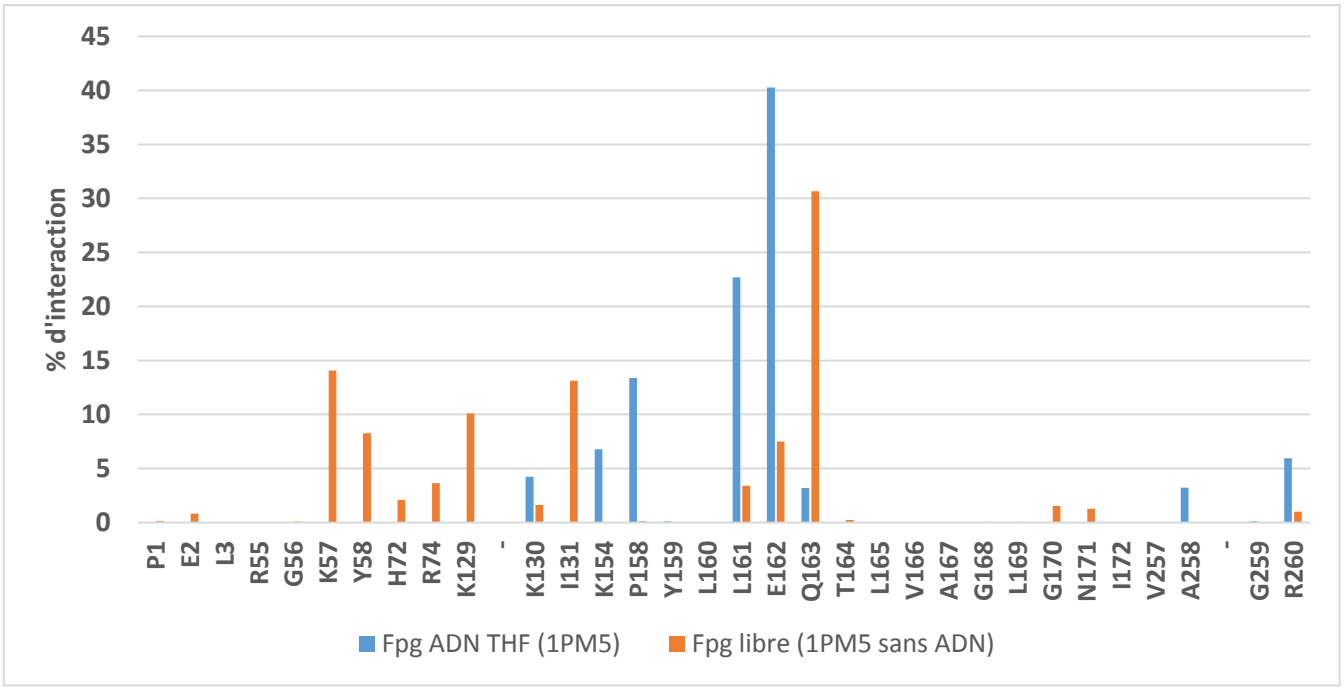


Figure 111 : Aa clés du site secondaire de *Ll*Fpg et hNEIL1
Le motif H2TH et le ZnF/ZnLF sont représentés en cartoon jaune et vert respectivement, le reste de la protéine en blanc. Les aa clés sont représentés en bâtonnets



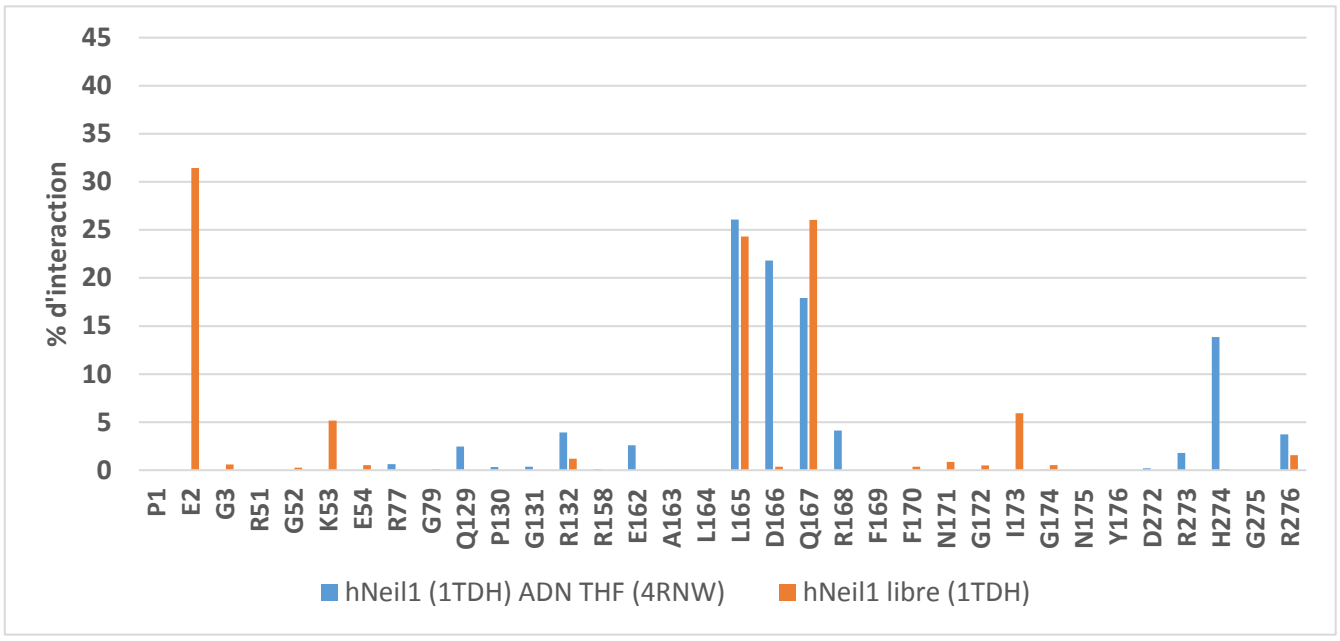


Figure 112 : Aa composant le site secondaire des modèles étudiés
Les systèmes avec ADN THF sont en bleu et sans ADN en orange. Les aa en abscisses correspondent à un alignement de *Ll*Fpg et de hNEIL1 basé sur les structures cristallographiques de 1PM5 et 1TDH respectivement (voir alignement complet dans l'Annexe F **VI.6** p. **335**).

IV.3.6 Docking des oligomères de 2TX et 2TX3

Comme suggéré dans la section précédente **IV.3.5.3** p. **271**, les molécules $2TX_n$ pourraient se combiner en polymères suite à la formation de ponts disulfures *via* leurs groupements thiones (**Figure 113**). Devant cette problématique, l'équipe de recherche a préféré investiguer sur la nature du contenu exact des stocks des molécules $2TX_n$ utilisés lors des tests d'activité. D'après les analyses de spectroscopie de masse (résultats non présentés), les échantillons de $2TX_n$ seraient bien en partie composés de polymères. Les tests biochimiques effectués sur *Ll*Fpg et hNEIL1 seraient donc biaisés car nous ne pouvons pas savoir si les effets d'inhibition observés sont causés par les monomères ou les polymères. De plus, nous avons vu dans la section précédente qu'il n'est pas possible de corréler les résultats de docking sans *a priori* et les IC_{50}. Sachant tout cela, nous avons collaboré avec l'entreprise Orléanaise GreenPharma pour réaliser une étude plus approfondie des polymères des molécules 2TX et $2TX_3$ et prédire leurs potentiels sites d'interaction sur *Ll*Fpg et hNEIL1 par une méthode de docking sans *a priori*. GreenPharma a réalisé le docking sans *a priori* avec le logiciel Surflex-Dock [274] de la suite Sybyl, à partir des 30 structures des cibles obtenues grâce aux simulations de DM des quatre systèmes de *Ll*Fpg ADN THF (1PM5), *Ll*Fpg libre (1PM5 sans ADN), hNEIL1 (1TDH) ADN THF (4NRV) et hNEIL1 libre (1TDH). Surflex-Dock est un logiciel de docking dont la fonction de score empirique est brevetée (brevet US6470305).

Plusieurs conformations des molécules ont été générées par 3 méthodes d'exploration conformationnelle différentes, à savoir (i) « Random Search » qui fait évoluer les angles de torsion de manière stochastique et qui effectue une « relaxation » par minimisation d'énergie, (ii) « GA search » qui est une recherche effectuée par un algorithme génétique et (iii) « Confort » qui est une recherche exhaustive, générant toutes les conformations possibles du ligand et sélectionnant seulement celles de plus basse énergie. Ces méthodes ont été utilisées pour faire varier les angles de torsions C-S-S-C des polymères et ainsi obtenir au total 51 conformations (pour les 7 ligands) les plus stables selon un critère énergétique. Toutes les conformations obtenues par les trois méthodes complémentaires citées précédemment ont été regroupées et classifiées afin d'écarter toutes les conformations redondantes grâce à une méthode de CAH. Dans le logiciel Surflex-Dock, le docking est guidé par un « protomol » (liste de toutes les interactions possibles dans une zone de recherche) préalablement calculé par le logiciel. Plusieurs conformations du ligand prédites par le logiciel sont ensuite générées tout en essayant d'optimiser le score d'affinité grâce au protomol.

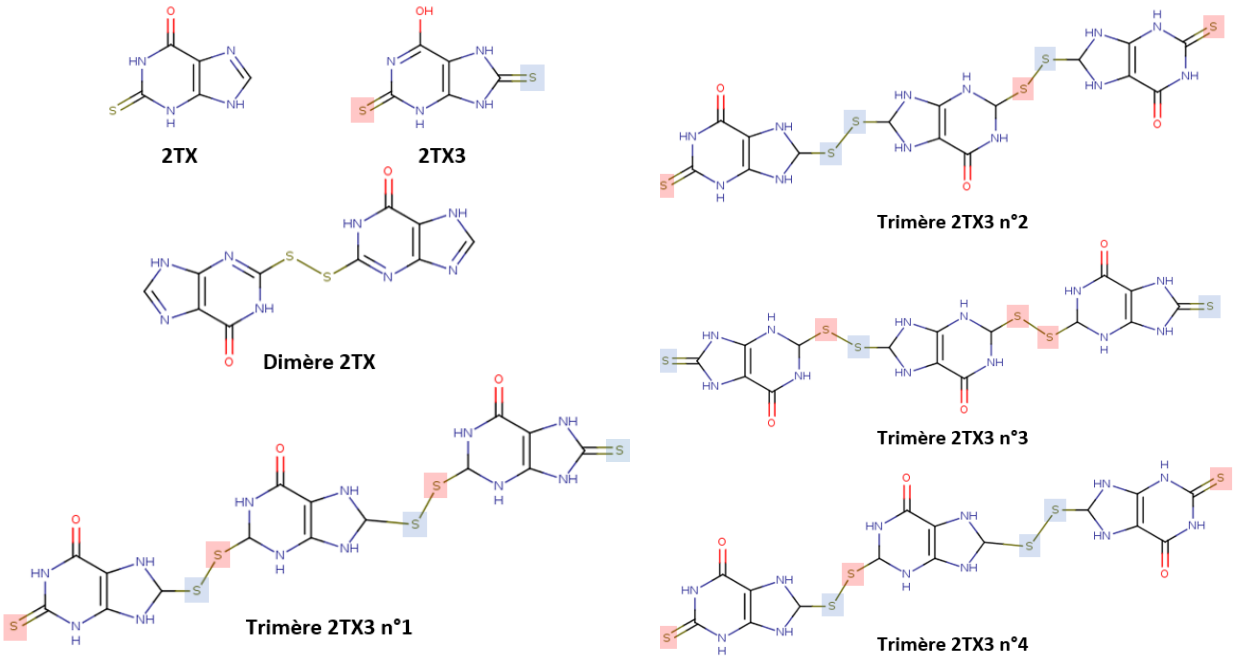


Figure 113 : Structures des monomères et des polymères de 2TX et 2TX3 utilisées pour le docking sans a priori sur les 30 structures de *Ll*Fpg et hNEIL1
Un seul dimère de la 2TX est possible car la molécule ne possède qu'un seul thione, quant à la 2TX3, elle peut former quatre trimères car elle présente deux groupements thiones.

Les résultats du docking réalisé par GreenPharma ont permis l'identification du site actif et du site secondaire des protéines *Ll*Fpg et hNEIL1 comme des sites de fixation des monomères et des polymères de 2TX et 2TX3. Les résultats relatifs aux monomères sont en bon accord avec ce que nous avions obtenu précédemment. Les aa en contact (< 3,5 Å) et formant des liaisons hydrogène avec les poses des ligands sont listés dans le **Tableau 27** tous ces résidus ont déjà été identifiés par le docking sans *a priori* présenté précédemment (**Figure 110** et **Figure 112**).

De plus, parmi ces résidus, GreenPharma a identifié P1 et L161 de *Ll*Fpg et E5, Y243 et D251 de hNEIL1 comme étant des aa d'arrimage privilégiés par les ligands. Des expériences de mutagénèse dirigée pourraient nous permettre de valider ou d'invalider la participation de ces résidus dans les modes d'interaction protéine/ligand prédits par le docking. Ils notent également l'interaction des ligands avec l'ADN lorsqu'il est présent dans le système.

La **Figure 114** représente une partie des résultats du docking des monomères et polymères dans un des centroïdes de hNEIL1 libre. Sur cette figure, nous remarquons que les polymères peuvent occuper à la fois le site actif et le site secondaire. Les molécules sont assez grandes pour occuper les deux sites en même temps, ce qui est un nouveau mode d'interaction que nous n'avions pas observé jusqu'à présent. Cependant, d'après les résultats de GreenPharma (non représentés ici), ce mode d'interaction n'est observé que lorsque l'ADN n'est pas présent dans le système car l'ADN encombre

la partie intermédiaire localisée entre le site actif et le site secondaire. Les effets d'inhibitions des 2TX et 2TX3 sur les cibles pourraient donc être lié à la fixation de polymères de ces molécules dans le site actif et le site secondaire simultanément.

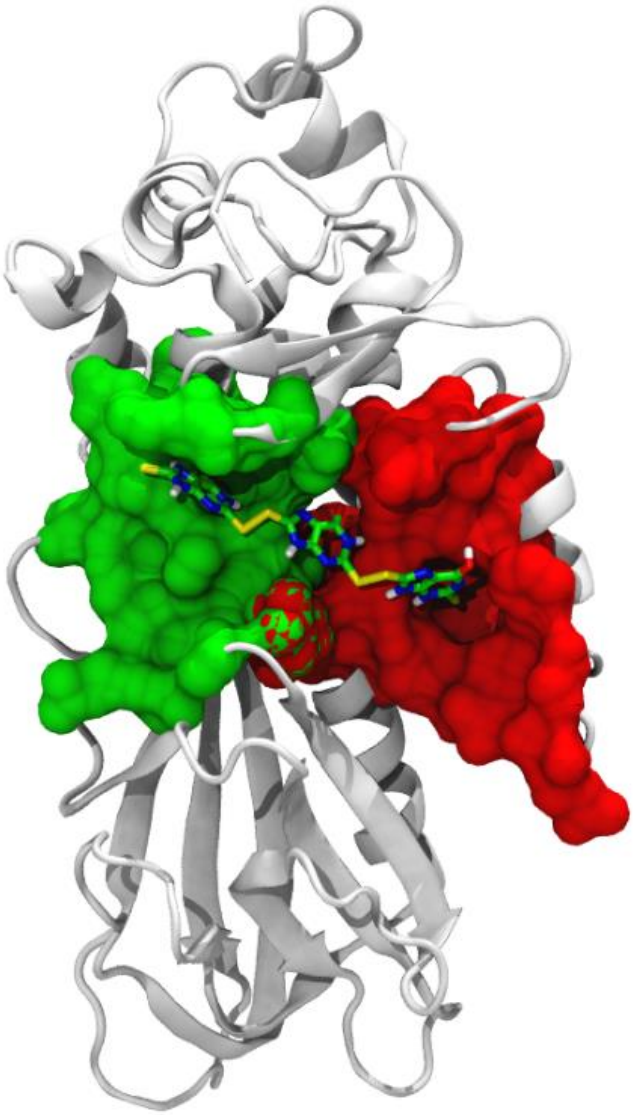


Figure 114 : Conformation du trimère n°1 2TX3 sur un des centroïdes de hNEIL1
Le site actif de la protéine est représenté en surface rouge et le site secondaire en surface vert. Le reste de la protéine est représenté en cartoon blanc. La conformation du trimère n°1 de 2TX3 issu du docking est représentée en bâtonnets.

Tableau 27 : Résidus en contact avec les conformations des monomères et polymères de 2TX et $2TX_3$ prédites par SurflexDock à la surface de *Ll*Fpg et hNEIL1
Les aa intervenant dans des liaisons hydrogène sont notés en **gras**. Les aa notés en rouge et à gauche de « | » font partie du site actif et les résidus notés en vert et à droite, composent le site secondaire.

Molécule	*Ll*Fpg	hNEIL1
Monomère 2TX	P1, M75 | L161, Q163, **R260**	E2, **E5**, L8, Y243, **G244**, D251 | L165, R276
Monomère $2TX_3$	P1, E5, V9, L73, K78, **L225** | L161, E162, Q163, R260	P1, **E5**, **G82**, S83, **Y243**, G244, **S245**, G248, **D251**, F252 | L165
Dimère 2TX	P1, E5, M75, E76, I219 | K130, L161, E162, **R260**	**P1**, **E2**, **E5**, **R132**, **Y176**, V235, **Y243**, **D251**, F252, F255 | **L165**, **N167**, H274, **R276**
Trimère n°1 $2TX_3$	P1, **M75**, E76, K78 | K130, **L161**, **E162**, Q163, **R260**	P1, **E2**, **E5**, L8, **M80**, **G82**, S83, **Y176**, V235, V236, L238, G239, Y243, **G244**, **D251**, F252, F255 | L165, N167, H274, **R276**
Trimère n°2 $2TX_3$	P1, E76, K78 | K130, L161, **E162**, Q163, **R260**	P1, **E2**, **E5**, L8, **M80**, **G82**, S83, **Y176**, V235, V236, L238, **Y243**, **G244**, G248, **D251**, F252, F255 | R132, L165, N167, R276
Trimère n°3 $2TX_3$	P1, E5, E76, K78, I219 | K130, L161, **E162**, **R260**	P1, **E2**, **E5**, L8, **M80**, **G82**, S83, **Y176**, V235, V236, L238, **Y243**, G244, D251, F252, F255 | L165, H274, **R276**
Trimère n°4 $2TX_3$	P1, E5, E76 | K130, L161, **E162**, Q163, **R260**	P1, **E2**, **E5**, L8, **M80**, G82, S83, **Y176**, V235, L238, **Y243**, 244, **D251**, F252, F255 | L165, **R276**

Par ailleurs GreenPharma possède une base de données de complexes protéine/ligand connus appelée ciblothèque. GreenPharma a effectué le redocking de chacun des ligands de la ciblothèque dans leurs cibles respectives et a évalué les scores d'affinités avec Surflex-Dock. Les scores d'affinités sont divisés par le nombre d'atomes N du ligand afin d'évaluer la « contribution moyenne » des atomes. Ces scores d'affinités pondérés par N obtenus grâce à la ciblothèque nous permettent d'estimer par comparaison si les monomères et polymères de $2TX_n$ possèdent ou non une bonne affinité pour les cibles. GreenPharma nous a fourni une analyse des scores de docking des monomères et des polymères des $2TX_n$. Les scores pondérés par le nombre d'atomes des molécules $2TX_n$ sont ensuite comparés aux scores obtenus à partir de la ciblothèque pour *Ll*Fpg (**Tableau 28**) et hNEIL1 (**Tableau 29**).

Nous remarquons que les scores moyens pondérés par le nombre d'atomes du ligand (N) sont plus élevés dans les systèmes des enzymes complexées à l'ADN que dans le cas des systèmes des enzymes libres, ce qui est en bon accord avec ce que nous avions constaté un peu plus tôt. De plus, il semblerait que, d'une manière générale, quel que soit le ligand et la cible, les scores pondérés par N soient bien en dessous des valeurs obtenues par le redocking des complexes protéines/ligand connus (environ divisé par deux). Cela signifie que les monomères et polymères de 2TX et 2TX3 ont une affinité plus faible pour *Ll*Fpg et hNEIL1 que les ligands pour leurs cibles respectives issus de la ciblothèque. Nous ne pouvons donc pas conclure sur l'origine des effets d'inhibition des molécules sur les cibles *Ll*Fpg et hNEIL1, ces effets peuvent être induits par les monomères et/ou les polymères de 2TX et 2TX3.

Tableau 28 : Rapport score d'affinité/nombre d'atomes de la molécule des docking des monomères et des polymères de 2TX et $2TX_3$ sur la cible *Ll*Fpg avec et sans ADN THF

Les valeurs à gauche correspondent aux scores du docking des monomères et polymères de 2TX et 2TX3. Les valeurs à droite « | » correspondent aux valeurs obtenues par le re-docking des ligands dans leurs cibles respectives à partir des complexes protéine/ligand issus de la ciblothèque.

		***Ll*Fpg ADN THF (1PM5)**		***Ll*Fpg libre (1PM5 sans ADN)**	
		Score	**Score/N**	**Score**	**Score/N**
Monomère 2TX N=11	**Moyenne**	3,75 | 5,82	0,34 | 0,53	3,18 | 5,82	0,29 | 0,53
	Médiane	3,68 | 5,42	0,33 | 0,49	3,17 | 5,42	0,29 | 0,49
	Écart-type	0,99	0,09	0,57	0,05
Monomère $2TX_3$ N=12	**Moyenne**	3,00 | 6,27	0,25 | 0,52	3,36 | 6,27	0,28 | 0,52
	Médiane	3,05 | 5,25	0,25 | 0,44	3,09 | 5,25	0,26 | 0,44
	Écart-type	0,47	0,04	0,56	0,05
Dimère 2TX N=22	**Moyenne**	5,17 | 9,00	0,24 | 0,41	5,04 | 9,00	0,23 | 0,41
	Médiane	5,30 | 8,67	0,24 | 0,39	5,15 | 8,67	0,23 | 0,39
	Écart-type	0,58	0,03	0,33	0,01
Trimère 1 $2TX_3$ N=36	**Moyenne**	6,24 | 12,08	0,17 |0,34	6,98 |12,08	0,19 | 0,34
	Médiane	6,12 | 11,33	0,17 |0,31	6,98 | 11,33	0,19 | 0,31
	Écart-type	0,94	0,03	0,88	0,02
Trimère 2 $2TX_3$ N=36	**Moyenne**	6,47 | 12,08	0,18 | 0,34	7,41 | 12,08	0,21 | 0,34
	Médiane	6,58 | 8,67	0,18 | 0,31	7,42 | 11,33	0,21| 0,31
	Écart-type	0,92	0,03	0,88	0,03
Trimère 3 $2TX_3$ N=36	**Moyenne**	6,28 | 12,08	0,17 | 0,34	7,01 | 12,08	0,20 | 0,34
	Médiane	6,45 | 11,33	0,18 | 0,31	7,10 | 11,33	0,20 | 0,31
	Écart-type	0,97	0,03	0,77	0,02
Trimère 4 $2TX_3$ N=36	**Moyenne**	6,03 |12,08	0,17 | 0,34	6,77 | 12,08	0,19 | 0,34
	Médiane	6,34 | 11,33	0,18 | 0,31	6,77 | 11,33	0,19 | 0,31
	Écart-type	1,00	0,03	0,68	0,02

Tableau 29 : Rapport score d'affinité/nombre d'atomes de la molécule des docking des monomères et des polymères de 2TX et $2TX_3$ sur la cible hNEIL1 avec et sans ADN THF

Les valeurs à gauche correspondent aux scores du docking des monomères et polymères de 2TX et 2TX3. Les valeurs à droite « | » correspondent aux valeurs obtenues par le re-docking des ligands dans leurs cibles respectives à partir des complexes protéine/ligand issus de la ciblothèque.

		hNEIL1 (1TDH) ADN THF (4RNW)		hNEIL1 libre (1DTH)	
		Score	Score/N	Score	Score/N
Monomère 2TX N=11	Moyenne	3,75 | 5,82	0,34 | 0,53	3,18 | 5,82	0,29 | 0,53
	Médiane	3,68 | 5,42	0,33 | 0,49	3,17 | 5,42	0,29 | 0,49
	Écart-type	0,99	0,09	0,57	0,05
Monomère $2TX_3$ N=12	Moyenne	3,00 | 6,27	0,25 | 0,52	3,36 | 6,27	0,28 | 0,52
	Médiane	3,05 | 5,25	0,25 | 0,44	3,09 | 5,25	0,26 | 0,44
	Écart-type	0,47	0,04	0,56	0,05
Dimère 2TX N=22	Moyenne	5,17 | 9,00	0,24 | 0,41	5,04 | 9,00	0,23 | 0,41
	Médiane	5,30 | 8,67	0,24 | 0,39	5,15 | 8,67	0,23 | 0,39
	Écart-type	0,58	0,03	0,33	0,01
Trimère 1 $2TX_3$ N=36	Moyenne	6,24 | 12,08	0,17 | 0,34	6,98 | 12,08	0,19 | 0,34
	Médiane	6,12 | 11,33	0,17 | 0,31	6,98 | 11,33	0,19 | 0,31
	Écart-type	0,94	0,03	0,65	0,02
Trimère 2 $2TX_3$ N=36	Moyenne	6,47 | 12,08	0,18 | 0,34	7,41 | 12,08	0,21 | 0,34
	Médiane	6,58 | 8,67	0,18 | 0,31	7,42 | 11,33	0,21 | 0,31
	Écart-type	0,92	0,03	0,88	0,03
Trimère 3 $2TX_3$ N=36	Moyenne	6,28 | 12,08	0,17 | 0,34	7,01 | 12,08	0,20 | 0,34
	Médiane	6,45 | 11,33	0,18 | 0,31	7,10 | 11,33	0,20 | 0,31
	Écart-type	0,97	0,03	0,77	0,02
Trimère 4 $2TX_3$ N=36	Moyenne	6,03 |12,08	0,17 | 0,34	6,77 | 12,08	0,19 | 0,34
	Médiane	6,34 | 11,33	0,18 | 0,31	6,77 | 11,33	0,19 | 0,31
	Écart-type	1,00	0,03	0,68	0,02

IV.3.7 Conclusion et perspectives

Nous avons tout d'abord validé le logiciel ADV en approchant les données expérimentales lors d'un docking classique dans la section **IV.3.4.2** p. **245**. Nous avons noté que l'eau structurale pouvait être importante dans la stabilisation des complexes protéines/ligand. Nous avons ensuite testé trois méthodes de docking sans *a priori* avec les logiciels (i) ADV, (ii) AMIDE ADV et (iii) AMIDE AD4. Nous en avons conclu qu'AMIDE est un logiciel permettant d'obtenir une plus grande gamme de conformations des ligands, et que, pour nos systèmes, ADV seul suffisait à prédire les modes d'interaction les plus forts. Les scores et les modes d'interactions prédits par AMIDE ADV et AMIDE AD4 étaient différents, ceci est lié aux différences dans les fonctions de score des deux logiciels. Les sites d'amarrage prédits étaient similaires dans tous les tests du docking sans *a priori*. Pour la suite, nous avons utilisé ADV pour le docking car nous souhaitions obtenir les meilleurs modes d'interaction protéine/ligand.

Dans la section **IV.3.5.2** p. **268** nous avons souligné l'intérêt de prendre en compte la flexibilité de la protéine dans les méthodes de docking. Nous avons également prédit deux sites d'interaction principaux des $2TX_n$ sur *Ll*Fpg et hNEIL1, à savoir le site actif et le site secondaire. D'après les scores d'affinité, l'ADN entre dans la composition du site secondaire ce qui fait de ce dernier un site à l'interface. De plus, les modes d'interaction prédits dans le site actif des enzymes complexées à l'ADN sont plus forts que dans les protéines libres, cela signifie que le site actif adopte différentes conformations en fonction de la présence ou de l'absence d'ADN dans le système. Dans la section suivante **IV.3.5.3** p. **271**, nous avons comparé les scores d'affinité obtenus par le docking et les IC_{50} mesurés expérimentalement. Nous en avons conclu qu'il n'était pas possible de faire une corrélation entre les données obtenues *in silico* et *in vitro*, et émis plusieurs hypothèses. L'une de ces hypothèses serait que les molécules $2TX_n$ comportant un groupement thione sont capables de former des polymères, et que les effets d'inhibition mesurés expérimentalement sont en fait induits par ces polymères et non les monomères.

Nous avons ensuite identifié les aa les plus impliqués dans la composition du site actif et du site secondaire de *Ll*Fpg et hNEIL1 complexées et libres. Des expériences de mutagénèse dirigées sur ces aa ainsi que des structures des cibles co-cristallisées par les ligands pourraient nous permettre de valider ou non la participation de ces derniers dans les modes d'interaction protéine/ligand prédits par le docking. Les deux sites de fixation des $2TX_n$ sont localisés à proximité l'un de l'autre, ce qui pourrait être un point de départ dans la conception par fragments d'inhibiteurs plus affins pour les deux cibles. Les $2TX_n$ sont des petites molécules et s'apparentent à des fragments et peuvent servir de bases dans

cette démarche. Il existe des librairies de fragments telle que la ChemBridge [275] qui pourrait être également utilisée à cet effet.

Pour répondre à la question des polymères de $2TX_n$, nous avons collaboré avec GreenPharma qui a pu reproduire des résultats similaires à ce que nous avions obtenus avec un autre logiciel de docking, Surflex-Dock. Les monomères et les polymères de 2TX et 2TX3 semblent interagir avec les sites actifs et les sites secondaires de *Ll*Fpg et hNEIL1. Nous avons également noté que les polymères pouvaient interagir avec le site actif et le site secondaire des cibles simultanément. En revanche les affinités de ces molécules pour les cibles estimées par le logiciel de docking sont très faibles, ce qui est en bon accord avec le fait que les molécules 2TX et 2TX3 ont des IC_{50} seulement de l'ordre du µM sur les cibles. De plus, il semblerait que les polymères présentent une affinité équivalente pour les cibles à celle des monomères. Les effets d'inhibition de *Ll*Fpg et hNEIL1 peuvent donc être induits par les monomères et/ou les polymères de 2TX et 2TX3. Les molécules $2TX_n$ sont extrêmement réactives et pourraient également réagir avec les Cys de hNEIL1, tout comme avec les Cys du ZnF de *Ll*Fpg dont nous avions parlé dans la section **I.3.6** p. **115** de l'Introduction. Cependant les logiciels de docking que j'ai utilisés ne sont pas des outils adéquats pour prédire les interactions covalentes entre protéines et petites molécules, pour valider cette hypothèse il faudra utiliser d'autres méthodes de docking covalent.

À la suite de cette prédiction de sites de fixation des monomères et polymères de $2TX_n$, nous avons identifié plusieurs aa fréquemment en contact avec les ligands. Des expériences de mutagénèses dirigées pourraient nous permettre de valider ou non les modes d'interaction des ligands dans les sites actifs et les sites secondaires prédits par le docking sur *Ll*Fpg et hNEIL1.

IV.4 **Chapitre 4 : R**ésultats préliminaires du criblage virtuel de hNEIL1

En parallèle des projets d'étude fonctionnelle et structurale des enzymes *Ll*Fpg et hNEIL1, nous avons réalisé un criblage virtuel d'une partie de la base de données Ambinter de l'entreprise GreenPharma [276] sur la protéine hNEIL1. Cette base de données contient 20 millions de petites molécules dont 150 000 sont des composés d'origine naturelle (extraits de plantes). Le but de ce projet est d'identifier des « hits » sur la base du score d'affinité généré par le docking et ensuite de tester ces molécules afin d'évaluer leurs effets sur l'enzyme hNEIL1 grâce à des tests *in vitro* réalisés au CBM. Le criblage *in silico* est une option peu coûteuse contrairement au criblage *in vitro*. Pour ce projet, nous avons décidé d'explorer uniquement le site actif de l'enzyme lors du docking, dans le but d'identifier des molécules spécifiques pouvant être de potentiels inhibiteurs compétitifs aux substrats de l'enzyme. En effet, si le site actif de l'enzyme est déjà occupé par une autre molécule, ici l'inhibiteur, on suppose que le mouvement de « base extrusion » de la base endommagée décrit dans la section **I.3.4.3** p. **108** de l'Introduction ne sera donc plus possible et par conséquent l'excision non plus. Ce projet est une collaboration entre le CBM, GreenPharma et l'ICOA d'Orléans.

IV.4.1 Protocole

Pour réaliser le criblage en des temps de calcul réduits, nous avons choisi de sélectionner une partie des 20 millions de composés proposés par Ambinter avec VSPrep (outil décrit dans la section **III.2.2.1** p. **152** de la partie Méthodologie). Nous avons donc appliqué le filtre « simple ». Ces molécules ne sont pas composées de cycle aromatique saturé pouvant adopter les conformations « chaise » ou « bateau » ainsi que des conformations intermédiaires. Ces cycles sont problématiques car la plupart des logiciels de docking (c'est le cas de ADV) ne sont pas capables de générer ces conformations et d'évaluer les affinités des molécules « chaise » ou « bateau », diminuant la précision des résultats et par conséquent biaisant les prédictions. En n'incluant pas les molécules cycliques insaturées dans le pool de ligands à docker, nous contournons ce problème. En outre, nous avons appliqué deux filtres supplémentaires afin de sélectionner les molécules « Veber-like » et éliminer les « PAINS » (décrits dans le **Tableau 11** p. **154** de la partie Méthodologie). Les molécules Veber-like répondent aux critères de la RO5 de Lipinski et ont également un nombre maximal de liaisons rotables et une faible surface polaire (une molécule dont la surface polaire est trop élevée, au-delà de 140 A^2 ne pourra pas franchir la membrane cellulaire). Les molécules PAINS sont des molécules connues pour être toxiques car elles sont très peu spécifiques et interagir avec un certain nombre de cibles biologiques, et ressortent très souvent comme des faux positifs lors des criblages virtuels. Au total, c'est 2 688 869 molécules

composées de cycles rigides ou sans cycles qui ont été sélectionnées pour le criblage virtuel du site actif de hNEIL1.

IV.4.1.1 Préparation des ligands et du récepteur

La préparation des molécules a donc été réalisée avec l'outil VSPrep [243], puis le formatage des fichiers a été réalisé avec l'outil OpenBabel [277].

Concernant le récepteur, nous avons choisi la structure de hNEIL1 libre (PDBid : 1TDH). Le protocole de préparation du récepteur est le même que dans le Chapitre 3. Cependant, contrairement aux dockings réalisés dans le chapitre précédent, une seule conformation du récepteur a été utilisée car nous ne pouvions pas réaliser le criblage virtuel dans plusieurs structures compte tenu du coût computationnel.

IV.4.1.2 Définition du site actif

La zone de recherche englobe plusieurs aa sélectionnés en se basant sur les structures de *Ll*Fpg complexée à un analogue de substrat. Ces aa ont été déterminés après un alignement structural de hNEIL1 (1TDH) et de *Ll*Fpg complexée à un ADN contenant un cFaPyG (1XC8). Les résidus en contact direct (distance entre atomes lourds ≤ à 3,5 Å) avec la base endommagée sont : P1, E2, E5, M74, E76, I172, S217, I219, R220, T221, Y222 et Y238 [201]. Les deux protéines comportent des différences au niveau du site actif. En effet la boucle LCL est un motif qui n'existe que chez *Ll*Fpg, et nous n'avons pas pu sélectionner d'aa d'un motif équivalent chez hNEIL1. Nous avons donc identifié des aa des hélices G et H pour compléter le site actif de hNEIL1 partiellement identifié grâce à l'alignement structural *Ll*Fpg/hNEIL1. Nous avons défini une zone de recherche centrée autour du site actif, dans une boite suffisamment grande pour englober les aa suivants : P1, E2, E5, L8, M80, S81, Y176, R241, G242, Y243, G244, S245, E246, S247, G248, E249, E250, D251, F252, A253, F255 et Y262 (**Figure 115**). L'alignement des deux structures (voir Annexe F **VI.6** p. **335**) a été réalisé avec l'outil MultiSeq [214] du logiciel VMD [215].

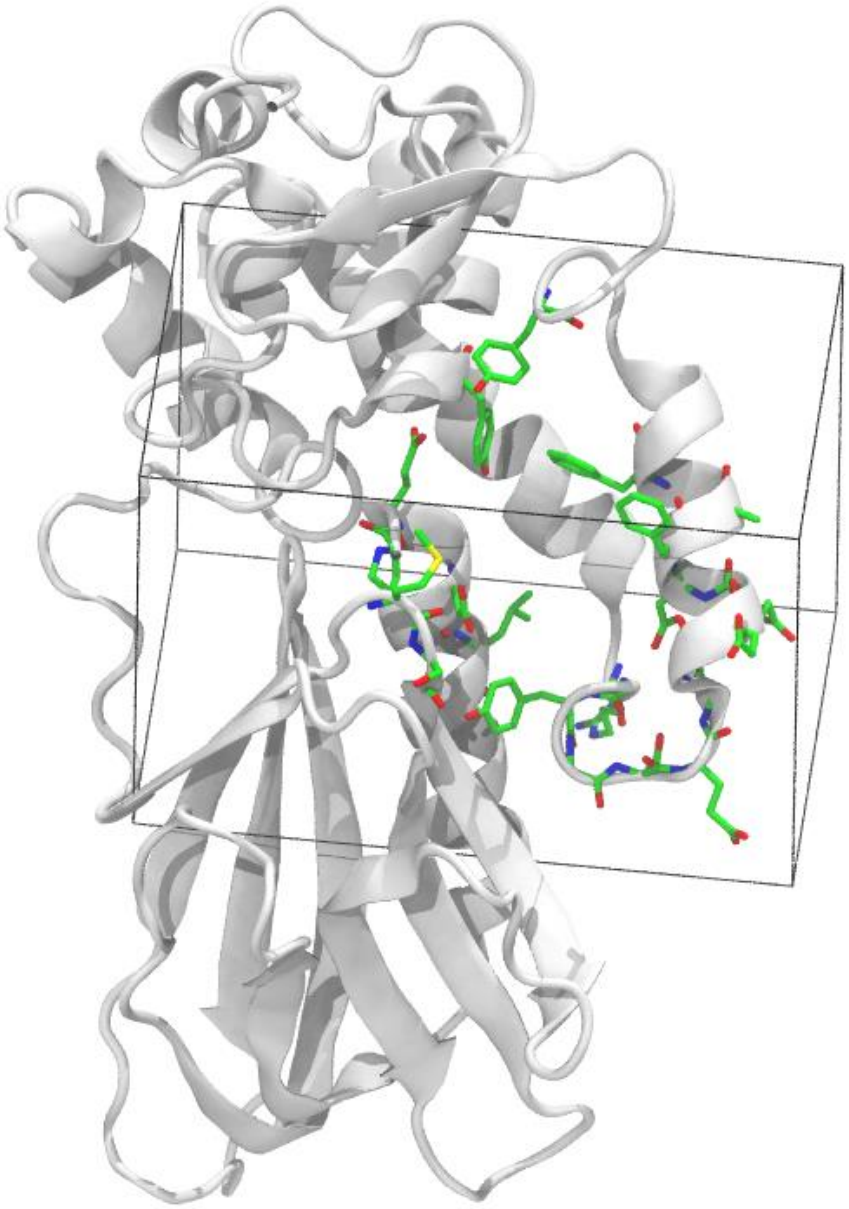


Figure 115 : Site actif de hNEIL1 criblé
La structure de hNEIL1 est représentée en cartoon blanc et les aa importants sont en bâtonnets. Le cube noir correspond à la zone de recherche du criblage virtuel.

IV.4.1.3 Criblage virtuel

Nous avons choisi ADV comme moteur de docking car ce logiciel est parfaitement adapté pour le criblage virtuel du fait de ses fonctionnalités. En effet, ADV est un outil qui, d'après ses développeurs, est beaucoup plus rapide et plus efficace qu'AD4 [241] et semble donc être un outil adéquat pour permettre un écrémage des 2 688 869 molécules sélectionnées à partir d'Ambinter. De plus, les tâches sont effectuées en parallèle sur plusieurs processeurs ce qui permet d'accélérer la production des données et peuvent être lancées dans une boucle automatisée en ligne de commande.

Nous avons réalisé le docking des molécules dans un volume restreint défini dans la section précédente et nous avons choisi une valeur faible du paramètre « `exhautiveness` » égale à 8 runs afin de réduire le temps de calcul. Le criblage virtuel a été réalisé sur le cluster Artemis du Centre de Calcul Scientifique en région Centre (CCSC).

IV.4.1.4 Identification des « hits »

Les molécules criblées sont classées par score d'affinité, et seules les molécules présentant les « meilleurs scores » d'affinité allant de -13,1 à -10,0 kcal.mol^{-1} sont extraites, ce qui représente un total de 1 005 molécules. La distribution des scores est représentée **Figure 116**. La solubilité de ces substances est parfois un obstacle au test de leurs effets sur une cible enzymatique. La capacité des molécules à se solubiliser dans l'eau pure est évaluée par une valeur appelée « logP ». C'est une mesure de solubilité différentielle de la molécule en question dans deux solvants qui sont l'octanol (o) et l'eau (w) (31) :

Équation 31 : Coefficient de partage

$$(31)\ logP = \log([o].[w])$$

Le logP peut être obtenu expérimentalement ou calculé avec plus ou moins de précision *in silico*. Le logP représente aussi la capacité d'absorption d'une molécule par l'organisme et du mode d'administration (oral, percutané, sublingual) qu'il faut privilégier. Dans la base de données Ambinter, les logP sont calculés grâce aux outils Indigo et Bingo [278, 279]. Nous avons extrait les valeurs de logP pour les 1 005 molécules sélectionnées et nous avons utilisé d'autres outils de prédictions du logP fournis par les logiciels MOE [280] et DataWarrior [281]. Nous avons reclassé les molécules en fonction du logP moyen prédit par les 3 méthodes et sélectionné 5 molécules aux logP moyens les plus faibles

et qui sont donc censés être les molécules les plus solubles (**Tableau 30** de la section suivante **IV.4.2** p. **295**). Cette sélection a pour but de faciliter les tests *in vitro*.

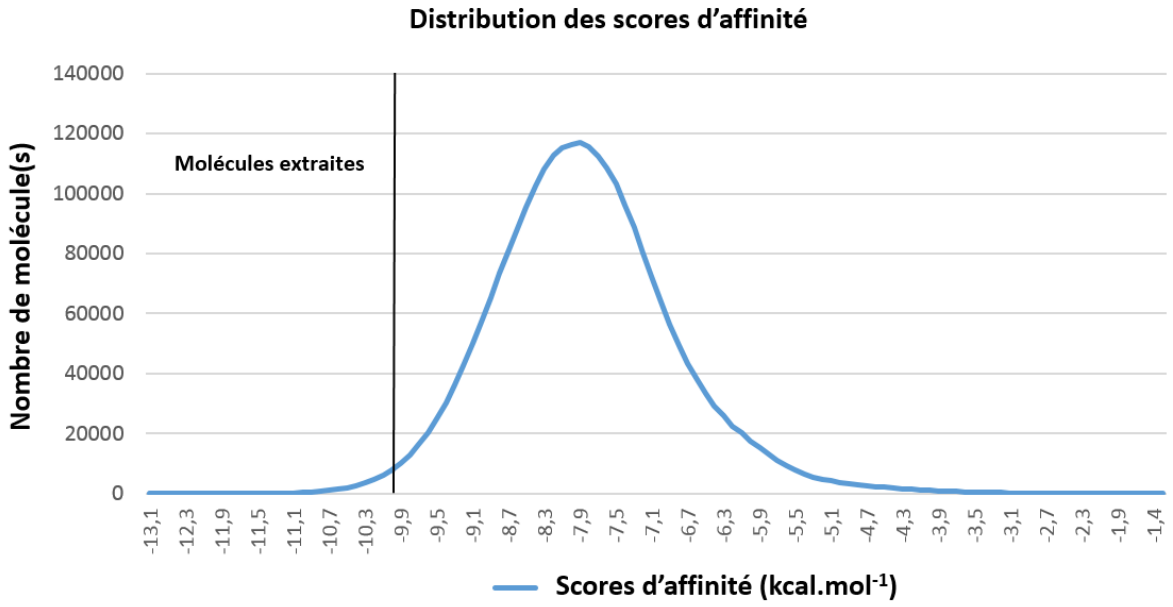


Figure 116 : Distribution des scores d'affinité issus du criblage des molécules sélectionnées à partir d'Ambinter dans le site actif de hNEIL1
Le score minimum est de -13,1 kcal.mol^{-1} et le score maximum est de -1,2 kcal.mol^{-1}. Le score moyen de cette distribution est de -7,8 kcal.mol^{-1} et l'écart type est égal à 1,0 kcal.mol^{-1}.

IV.4.2 Hits identifiés

Les 5 molécules les plus solubles et aux meilleurs scores d'affinités pour le site actif de hNEIL1 sont présentées dans le **Tableau 30**. On peut remarquer que beaucoup de ces molécules comportent des motifs communs tels que (i) des imides (rouge), (ii) des benzènes (violet), (iii) des amines secondaires (jaune), (iv) des hydrazides (vert) et (v) des benzhydryles (bleu). Ces groupements fonctionnels sont présents dans au moins deux molécules différentes et permettent d'identifier les éléments structuraux formant des interactions récurrentes avec la protéine. Les sous-structures les plus représentées à travers ces 5 molécules sont, dans l'ordre décroissant d'apparition, les imides/benzènes/amines secondaires (4), les hydrazides (3) et les benzhydryles (2). Les molécules Amb489941 et Amb119329 comportent toutes les deux une partie de leurs structures composées d'imides, amines secondaires, d'hydrazides et benzhydryles, et présentent donc une certaine similarité.

IV.4.2.1 Tests *in vitro*

Les effets des 5 molécules présentées dans le **Tableau 30** ont été testés *in vitro* sur la cible hNEIL1 selon deux protocoles différents. Deux sondes ADN contenant du DHU (dihydrouracile), altération de l'ADN que hNEIL1 est capable d'exciser, ont été conçues pour les tests biochimiques d'activité de hNEIL1. La première est un ADN comportant un fluorochrome sur le DHU et un « désactivateur » (ou « quencher ») sur l'extrémité du brin porteur de la lésion. Lorsque la base est excisée par l'enzyme, le fluorochrome porté par cette dernière est éloigné du quencher et émet une fluorescence dont on mesure l'intensité. À partir de cette fluorescence, il est possible de déterminer l'activité de l'enzyme lorsqu'elle est en relation avec une petite molécule. La seconde sonde est radiomarquée, c'est-à-dire que les extrémités des deux brins d'ADN sont marquées au P_{32}, et émettent une radiation β^-. Lorsque l'enzyme réalise les activités glycosylase et AP lyase, elle va créer un brin tronqué en position n^{-1} de la position de la base endommagée. La différence de taille des deux brins d'ADN peut être révélée grâce aux rayonnements β^- émis par le P_{32} après une migration sur gel acrylamide en conditions dénaturantes. Si un effet d'inhibition est détecté à haute concentration en petite molécule sur l'activité de hNEIL1, de nouveaux tests biochimiques d'activité à concentration variable sont effectués pour déterminer les IC_{50}.

D'après le **Tableau 30**, une seule molécule présente un effet d'inhibition de l'activité de hNEIL1, Amb6417670, avec un IC_{50} de 158 µM et un écart type de 34 µM. Cette molécule est également la seule à posséder deux groupes fonctionnels imide (rouge). Nous pourrions extraire du criblage virtuel d'autres molécules possédant ce groupe fonctionnel pour les tester sur hNEIL1. Cette valeur d'IC_{50} est obtenue par mesure de l'activité glycosylase résiduelle de hNEIL1 en présence de différentes concentrations de la petite molécule (**Figure 117**). D'après les tests d'activités, Amb6417670 n'a pas d'effet sur les activités des enzymes *Ll*Fpg et MvNei1, faisant de cette molécule un faible inhibiteur sélectif de hNEIL1.

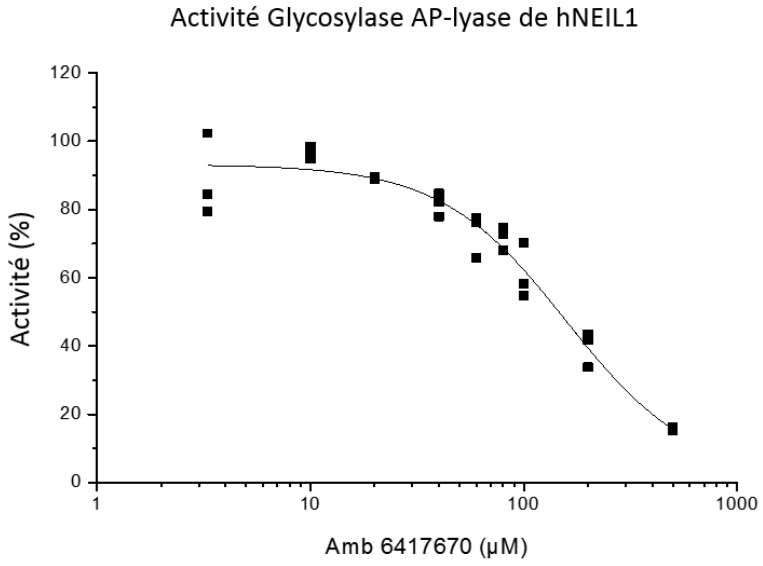
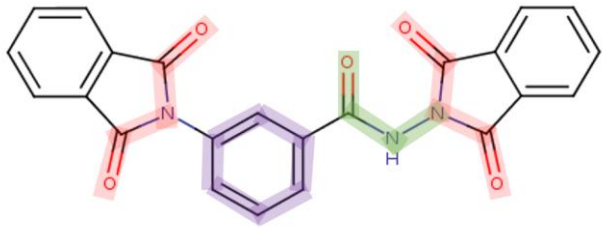
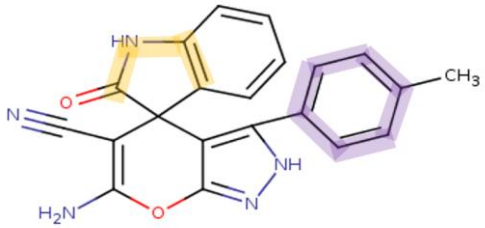
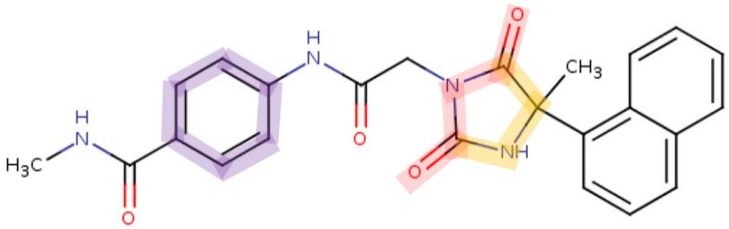
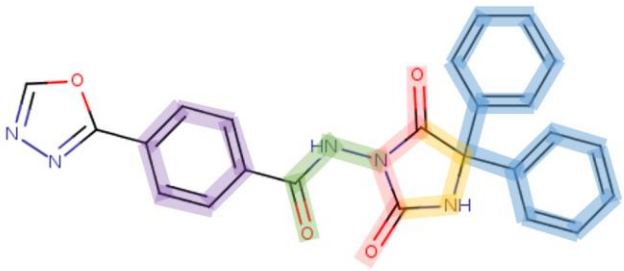
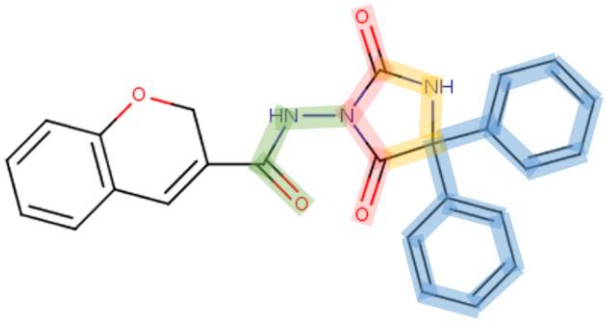


Figure 117 : Mesure de l'IC_{50} de la molécule Amb6417670 sur hNEIL1
L'activité glycosylase AP lyase de hNEIL1 est réduite de 50% à une concentration en Amb6417670 de 158 µM

Tableau 30 : Molécules testées *in vitro* sur l'activité de hNEIL1 ainsi que les IC_{50} associés à ces essais
NM signifie que l'effet d'inhibition de la molécule sur l'activité de hNEIL1 est trop faible et que l'IC_{50} ne peut pas être mesuré. Les logP indiqués dans le tableau correspondent aux valeurs extraites d'Ambinter.

Structure 2D et AmbinterId	Score d'affinité (kcal.mol^{-1})	LogP (o/w)	IC_{50} (µM)
Amb6417670 	-11,0	2,8	158 ± 34
Amb2655562 	-11,1	3,5	NM
Amb525825 	-11,5	3,3	NM

Amb489941 	-11,1	3,5	NM
Amb119329 	-11,5	3,6	NM

IV.4.2.2 Analyse des molécules testées *in vitro*

Les **Figure 118** et **Figure 119** représentent les meilleures poses obtenues dans le site actif de hNEIL1 et les interactions formées entre les deux objets détectées par l'outil Maestro de la suite Schrödinger [272].

Amb641760 est la seule molécule parmi les 5 à posséder un groupe imide. Nous pourrions extraire d'autres molécules possédant ce groupe fonctionnel à partir du criblage virtuel et tester leurs effets sur l'activité de hNEIL1 *in vitro*.

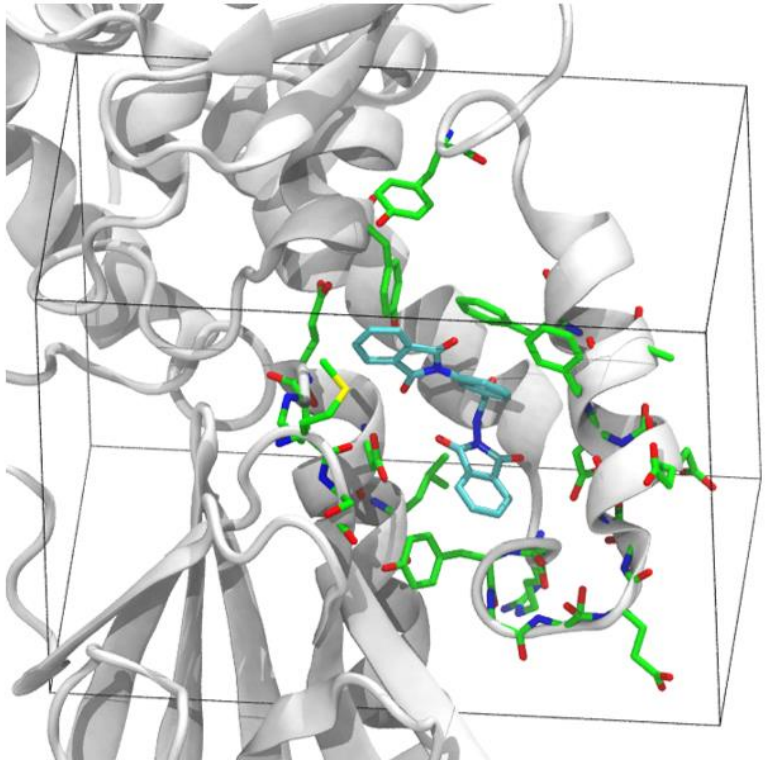
Nous savons que la seule molécule ayant un effet sur l'activité de hNEIL1 est Amb6417670. D'après les **Figure 118 A** et **Figure 119 A**, Amb641760 est également la seule molécule à être assez proche de la P1 catalytique, et à former également un stacking avec Y176. Ces aspects sont potentiellement liés à l'effet d'inhibition de cette molécule sur hNEIL1 et peuvent permettre la sélection de nouvelles molécules à tester à partir des résultats du criblage virtuel.

Sur la **Figure 119**, nous remarquons également qu'aucune liaison hydrogène n'est détectée par Maestro, en revanche quelques aa forment des interactions de type stacking avec les 5 molécules testées, à savoir notamment Y176 (1 fois sur 5), Y243 (2/5) et F255 (5/5). F255 semble former un stacking dans les 5 cas et notamment avec la molécule inhibitrice de hNEIL1 Amb6417670, ce qui fait de cette interaction un autre élément intéressant pour la sélection de plus de molécules à tester sur hNEIL1 à partir des résultats du criblage virtuel.

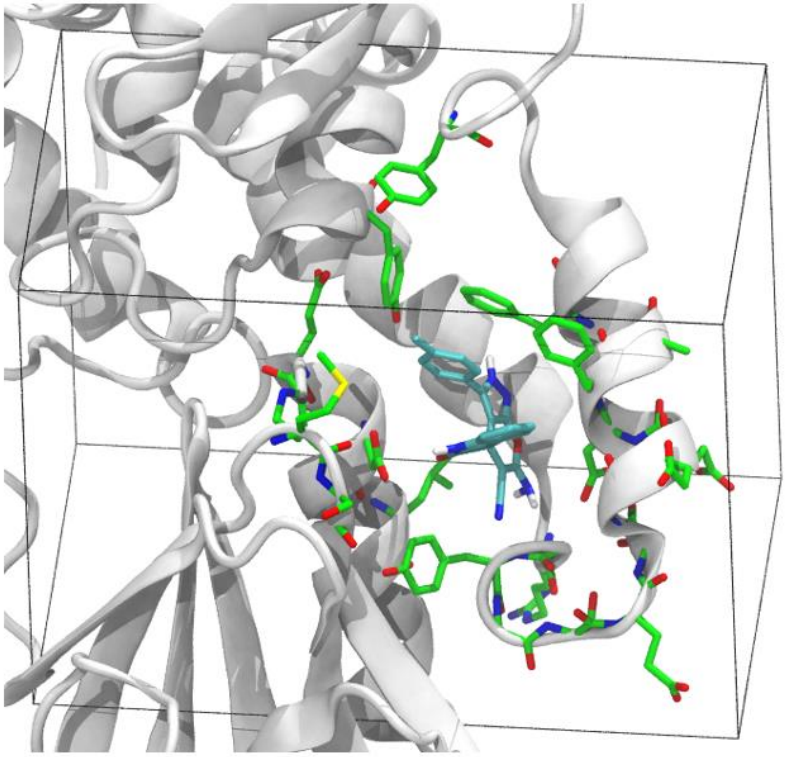
D'autres aa peuvent potentiellement former des interactions avec les petites molécules dans le site de hNEIL1. Il y a notamment F252 qui pourrait former un stacking. D'autre part, les aa chargés tels que E2, E5, S83, Q85, S245, D251 pourraient également être exploités pour la conception de nouveaux inhibiteurs de hNEIL1.

La recherche d'autres interactions fréquentes comme les liaisons hydrogène et les stackings entre les ligands et hNEIL1 dans les résultats du criblage virtuel pourraient nous permettre d'identifier les aa clés jouant un rôle dans l'inhibition compétitive de hNEIL1.

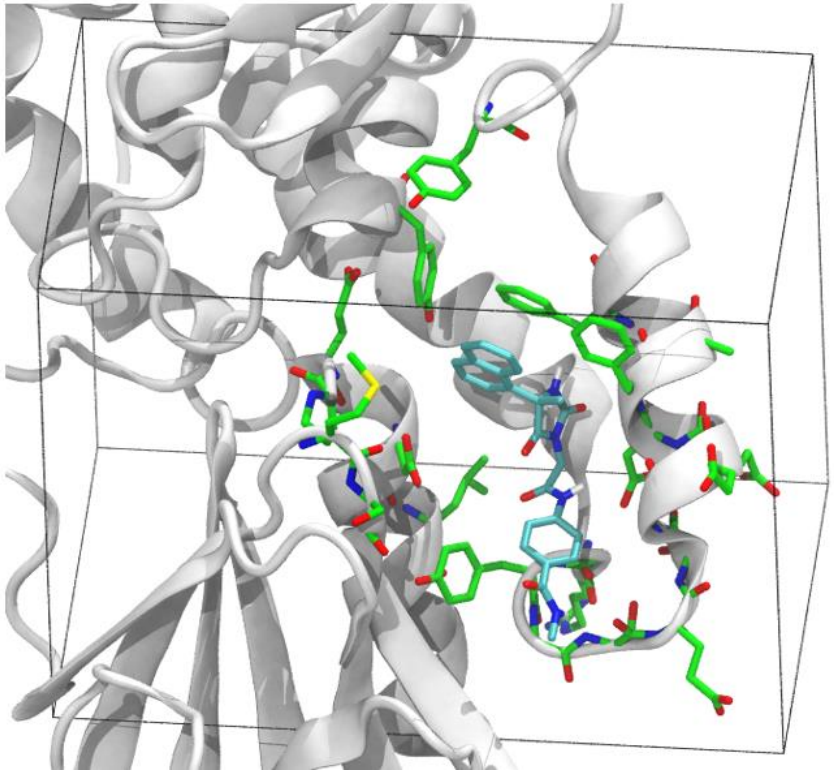
A Amb6417670



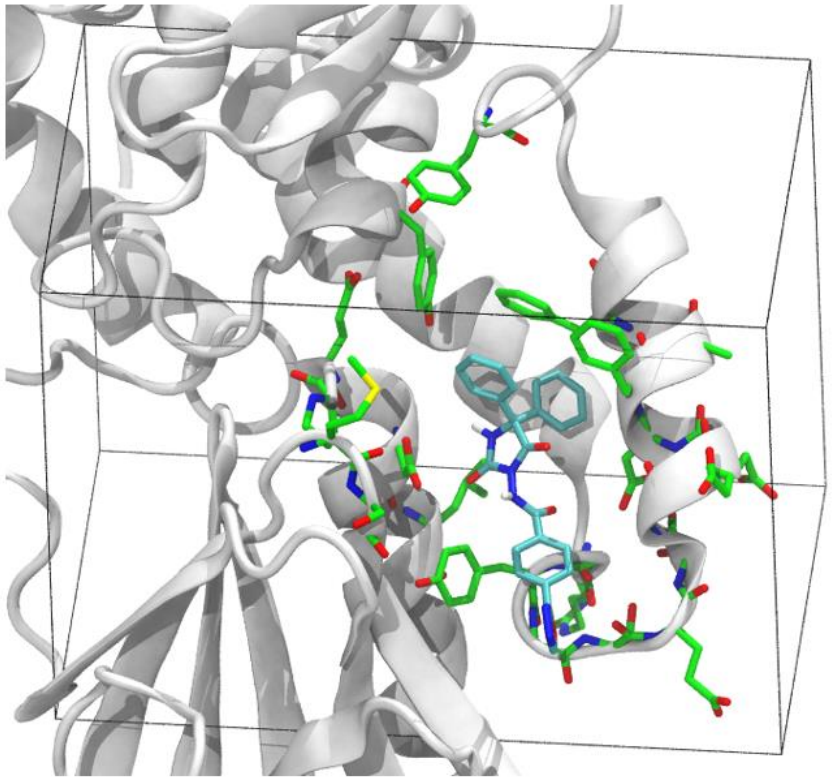
B Amb2655562



C Amb525825



D Amb489941



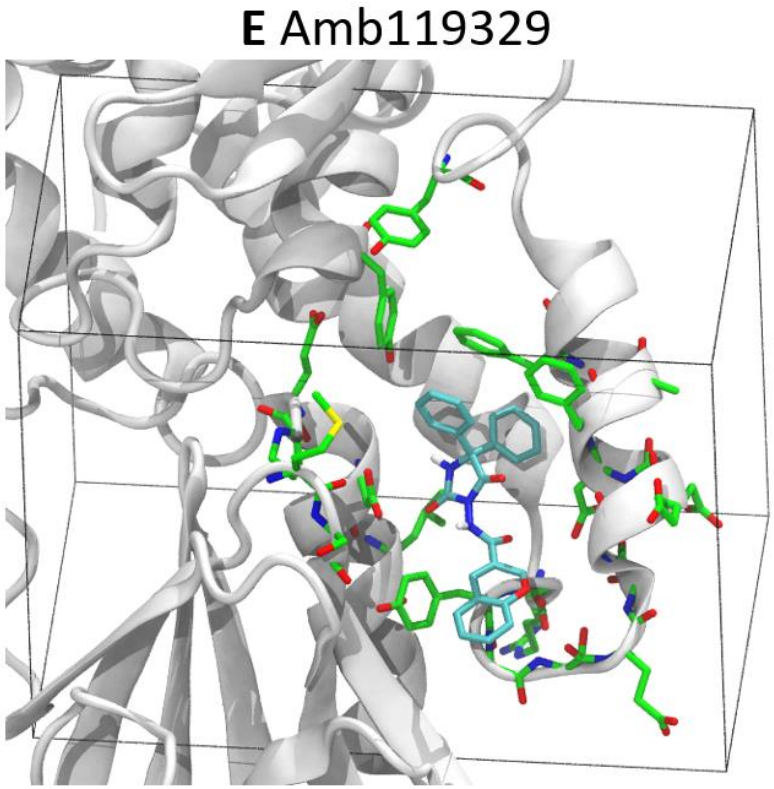
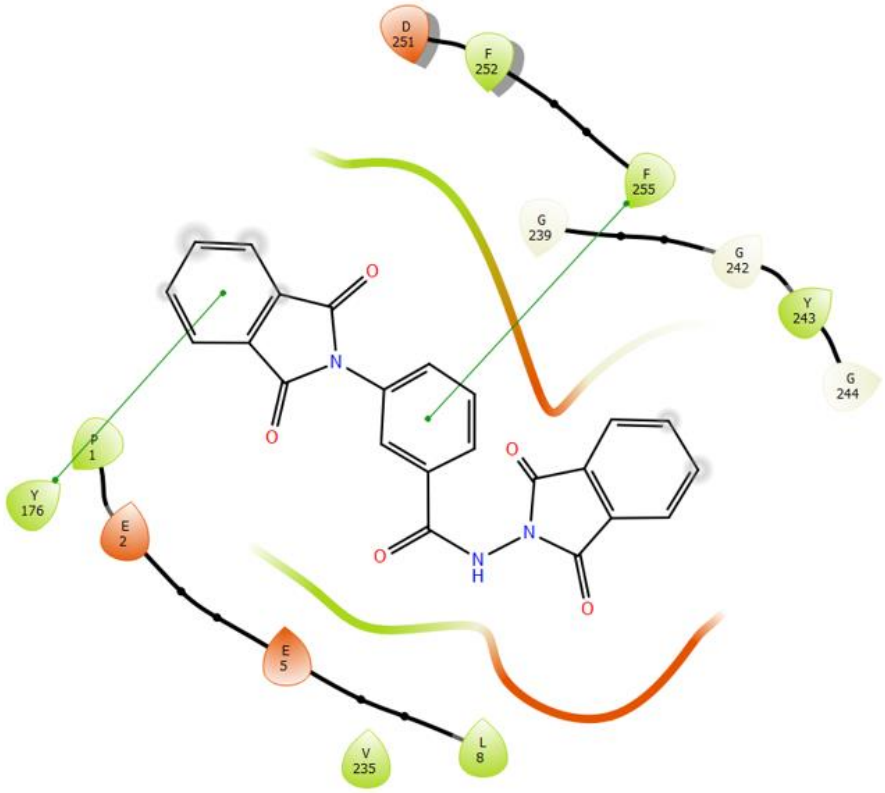
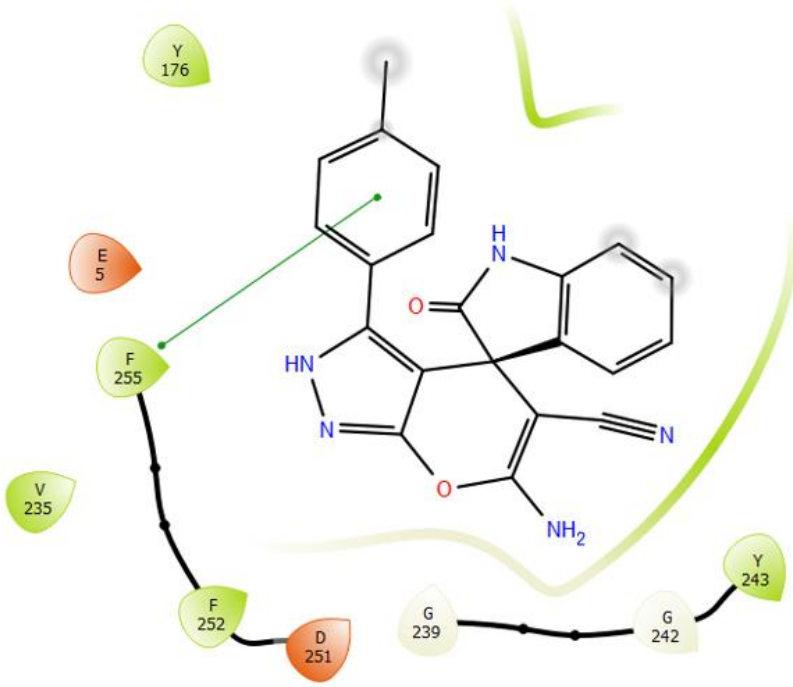


Figure 118 : Meilleures poses de docking des molécules testées avec le site actif de hNEIL1
La protéine est représentée en cartoon blanc, les aa formant le site actif sont en bâtonnets verts, et le cube noir correspond à la zone de recherche. La molécule issue du criblage est en bâtonnets cyan.

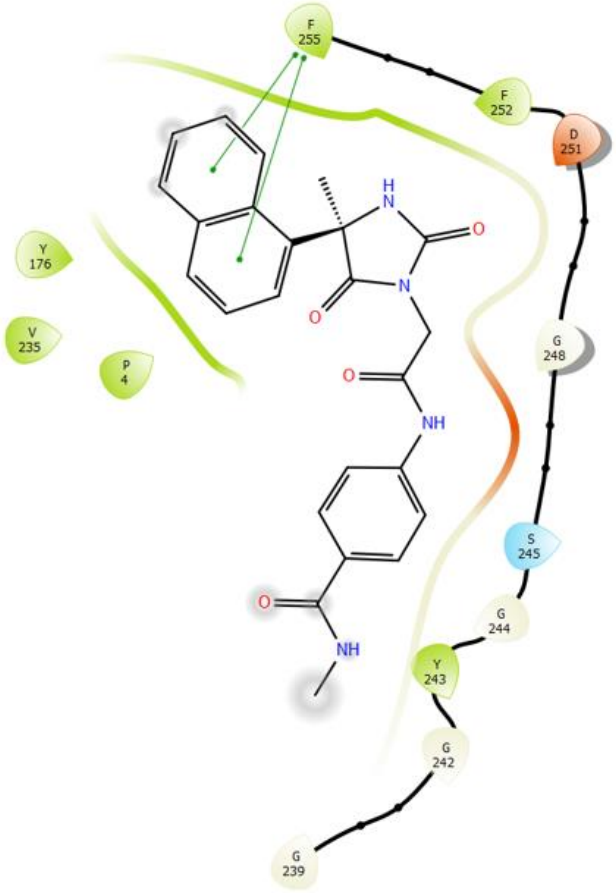
A Amb6417670



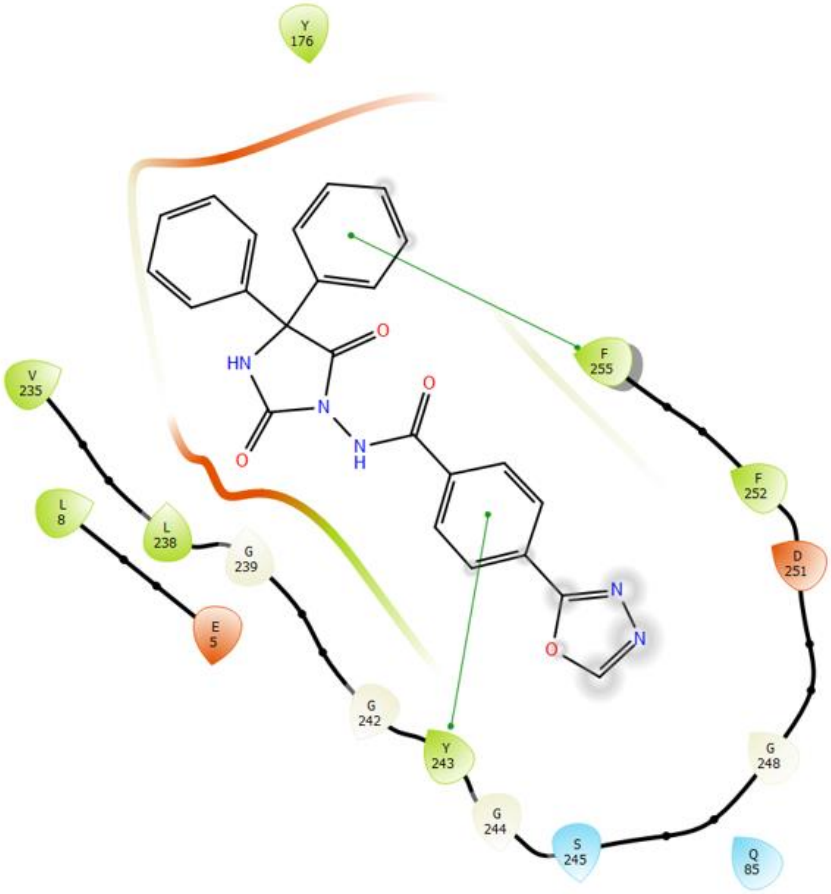
B Amb2655562



C Amb525825



D Amb119329



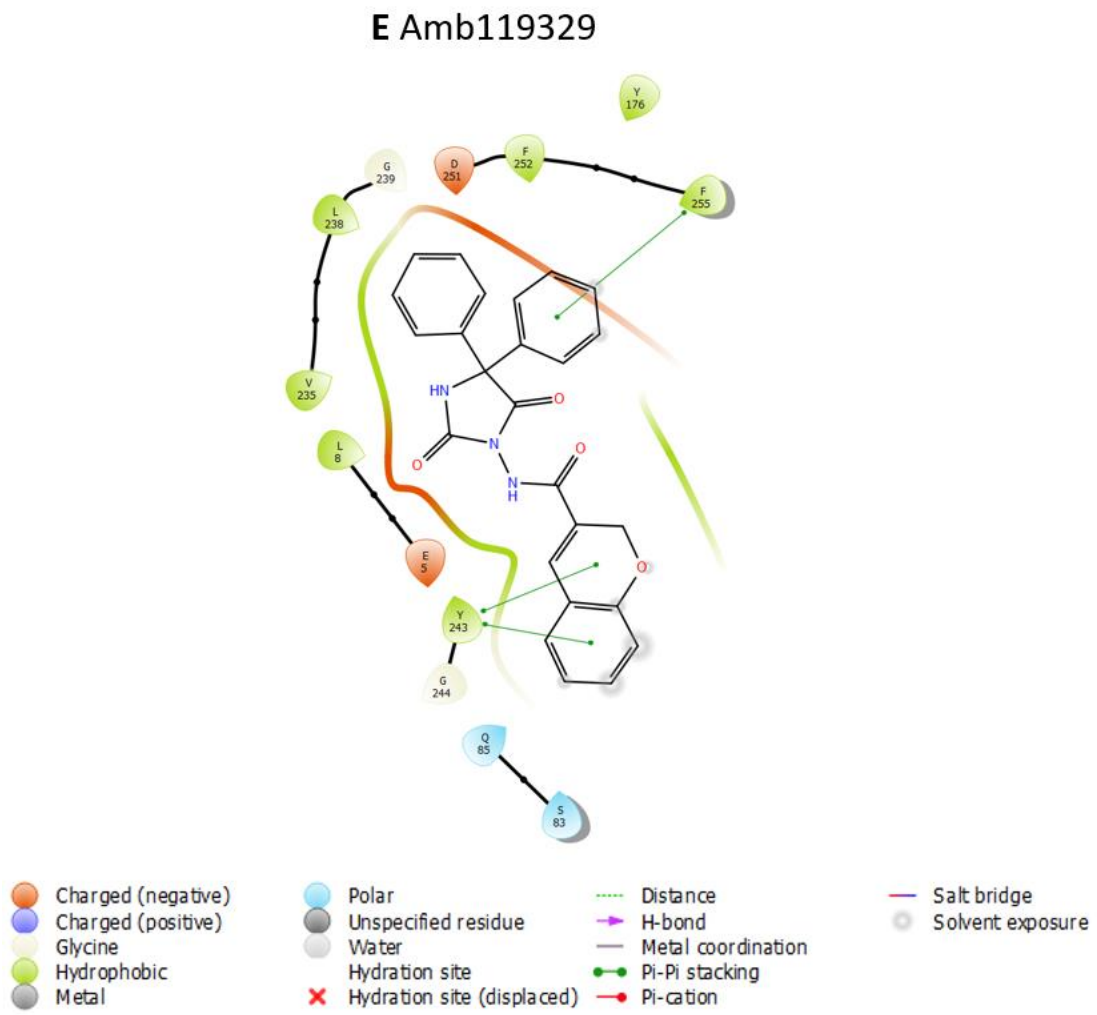


Figure 119 : Interactions formées entre les molécules testées et le site actif de hNEIL1

Ces représentations ont été générées par l'outil Maestro de la suite Schrödinger [272]. Les aa à moins de 3 Å de distance des petites molécules sont représentés sur ces schémas.

IV.4.3 Conclusion et perspectives

Le criblage virtuel de plus de 2 600 000 de molécules dans le site actif hNEIL1 a été réalisé en collaboration avec l'ICOA et sur les ressources computationnelles du Centre de Calcul Scientifique de la région Centre (cluster Artemis). L'analyse préliminaire du criblage virtuel nous a permis d'identifier une molécule avec un faible pouvoir d'inhibition sur hNEIL1, Amb6417670. Nous avons mis en évidence le fait que cette molécule formait quelques interactions de type stacking avec les cycles aromatiques d'Y176 et F255 et aucune liaison hydrogène. De plus, il semblerait que la molécule doit être capable d'interagir avec la P1 catalytique de la cible pour avoir un effet d'inhibition sur cette dernière. L'optimisation par pharmacomodulation de cette molécule pourrait nous permettre d'obtenir d'autres molécules possédant d'avantage d'affinité pour la cible et avec un pouvoir inhibant plus puissant.

L'équipe de recherche au CBM prévoit de réaliser très prochainement d'autres tests d'activités *in vitro* à partir de nouvelles molécules sélectionnées par le criblage virtuel et le tri des logP, car jusqu'à présent, seulement 5 molécules sur les 1005 sélectionnées ont été testées *in vitro*.

Les données issues du criblage virtuel peuvent également être exploitées de manière plus approfondie. À partir de toutes les conformations générées par le criblage, il est également possible d'identifier des interactions fréquentes entre certains groupements fonctionnels ou fragments et certains aa de hNEIL1 comme ce qui a été fait à partir du **Tableau 30** mais dans l'espace 3D du site actif de l'enzyme (**Figure 120**). Ces redondances peuvent permettre la construction d'un « échafaudage » (ou « scaffold ») qui peut servir de base pour la conception d'une nouvelle série de molécules optimisées pour s'adapter spécifiquement au site actif de hNEIL1, toujours dans l'optique de la conception d'inhibiteurs compétitifs de cet enzyme.

Aussi, de nouveaux tests de mutagénèse dirigée peuvent être envisagés pour mieux comprendre et valider ou non les modes d'interaction entre les potentiels futurs inhibiteurs et hNEIL1.

À cela s'ajoute le criblage du reste de la base de données Ambinter car nous n'avions criblé qu'une petite partie de la libraire (< 13%).

A Superposition des poses du criblage virtuel **B** Exemple de création d'un scaffold

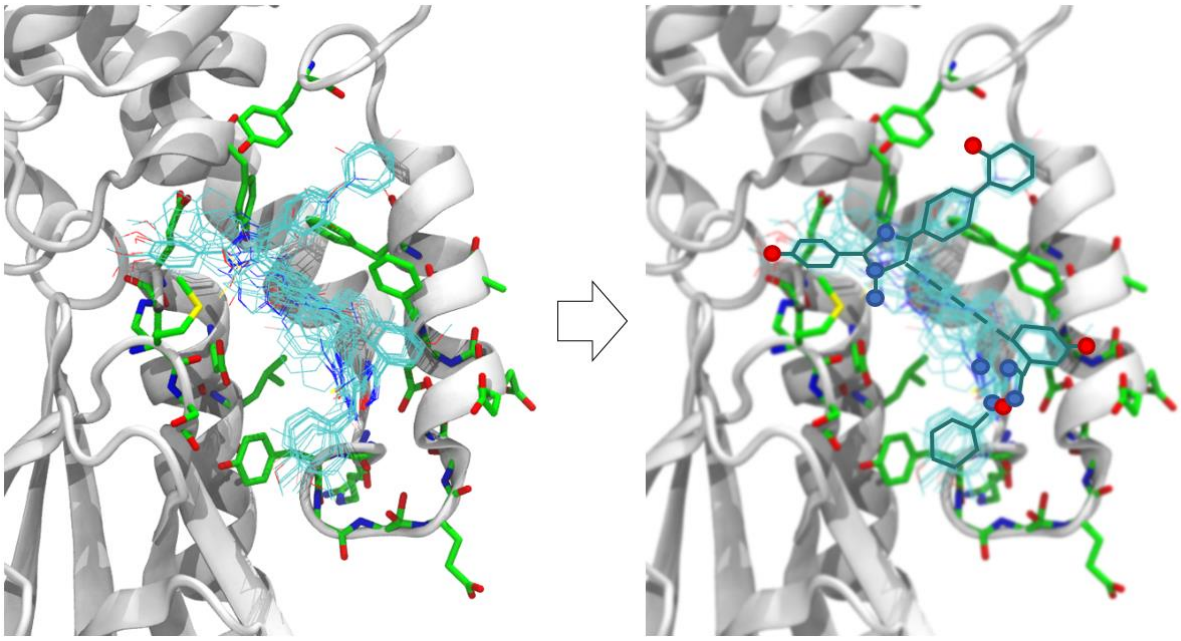


Figure 120 : Exemple de création d'un scaffold à partir des résultats du criblage virtuel du site actif de hNEIL1

La protéine est représentée en cartoon blanc et les aa formant le site actif sont en bâtonnets verts. Les molécules issues du criblage est en bâtonnets cyan fins, et le scaffold est en bâtonnets cyan épais.

V. **C**onclusion et perspectives générales

Le support de l'information génétique subit constamment des altérations issues de facteurs endogènes (métabolisme cellulaire) et exogènes (UV, radiations, produits chimiques). Pour maintenir l'intégrité de l'ADN, les cellules possèdent plusieurs systèmes de réparation de l'ADN dont le système de réparation par excision de base (BER) initié par les ADN glycosylases. La radio- et chimiothérapie sont classiquement utilisées pour endommager l'ADN des cellules cancéreuses et ainsi induire leur apoptose. Dans ce contexte, les ADN glycosylases réparent l'ADN des cellules cancéreuses et induisent une résistance face à ces traitements. C'est pourquoi les ADN glycosylases sont des cibles thérapeutiques intéressantes dans des stratégies anti-cancers combinatoires. Pendant ces trois années de thèse, je me suis tout particulièrement intéressée aux ADN glycosylases de la superfamille structurale Fpg/Nei. Nous avons utilisé plusieurs approches de modélisation moléculaire afin de comprendre la dynamique du site actif des protéines *Ll*Fpg et hNEIL1 et ainsi concevoir des molécules inhibitrices de ces enzymes de façon rationnelle. Ce projet de thèse est financé par le ministère de l'Enseignement Supérieur et de la Recherche et s'inscrit dans le projet « Modulation de l'activité des ADN glycosylases et Applications Connexes » (MAAC2) financé par la Région Centre-Val de Loire.

Les objectifs de ces travaux de thèses étaient d'identifier le potentiel rôle de la boucle LCL dans l'activité glycosylase de *Ll*Fpg et rechercher et concevoir de nouveaux inhibiteurs sélectifs de l'ADN glycosylase hNEIL1. Pour répondre à ces questions, nous avons utilisé des simulations de dynamique moléculaire classique et ciblée, une méthode de docking flexible et aveugle, et réalisé le criblage virtuel d'une partie de la librairie de molécules Ambinter dans le site actif de hNEIL1.

Nous avons donc étudié et reconstruit quand nécessaire les structures des protéines *Ll*Fpg libres et complexées à l'ADN. Les simulations de dynamique moléculaire de sept systèmes nous ont permis dans un premier temps d'apporter de nouveaux éléments de compréhension concernant le rôle de la boucle LCL de *Ll*Fpg. La conclusion de cette étude est que la boucle LCL est un objet flexible, pouvant adopter deux états et que le passage de l'état relâché à l'état fermé est potentiellement dépendant de la présence d'ADN dans le système ainsi que du contenu du site actif. La présence ou non d'interactions entre certains aa clés de la boucle LCL tels que R220 et T221 et deux aa du domaine N-terminal de la protéine, P1 et E76, ou encore avec l'analogue de site abasique THF et la cytosine opposée au dommage (E76-R220, E76-T221, THF-R220, G27-R220 et P1-R220) traduisent plusieurs comportements différents de la boucle LCL. Parmi ces comportements nous avons mis en évidence un

mouvement de pseudo-fermeture de la boucle LCL dans le système de la protéine *Ll*Fpg complexée à un ADN THF et dont le site actif était vacant. Nous avons également remarqué que la boucle LCL adoptait différents comportements en présence d'un substrat ou d'un produit d'excision dans le site actif de l'enzyme, ce qui pourrait indiquer que la boucle LCL serait capable de faire la différence entre les deux types de molécules. La boucle LCL serait potentiellement un stabilisateur de substrat et non de produit d'excision, ce qui pourrait faciliter la sortie de la base après la coupure. D'autre part, nous n'avons pas observé de mouvement d'ouverture de la boucle LCL ou de passage spontané de l'état fermé à relâché de cette dernière dans les simulations de dynamique moléculaire classique. Nous en avons conclu que sans apport d'énergie au système, nous ne pouvions pas faire une telle observation.

Dans le but d'identifier les mécanismes d'ouverture de la boucle LCL, et en partant du postulat que la boucle LCL de *Ll*Fpg est impliquée dans le mécanisme de sortie de la base libre après sa coupure, nous avons réalisé huit simulations de dynamique moléculaire dirigée des systèmes de *Ll*Fpg complexée à un ADN THF et comportant une 8-oxoG libre dans le site actif. La boucle LCL était dans un état fermé dans les quatre premières simulations et dans un état relâché dans les quatre autres simulations. Nous avons identifié trois chemins de sortie de la base libre du site actif, (i) un premier où la 8-oxoG libre diffuse le long du petit sillon d'ADN avant d'évoluer dans le solvant, (ii) un deuxième où la base libre longe le domaine C-terminal et (iii) un dernier où le produit d'excision longe le domaine N-terminal. Dans les systèmes boucle LCL relâchée, nous n'avons pas vu de mouvement de cette dernière. Concernant les simulations des systèmes comportant une boucle LCL fermée, nous avons remarqué la rupture d'interactions identifiées précédemment telles que E76-R220 et E76-T221 dans certains cas, accompagnées de légères déformations de la boucle LCL sans que cette dernière n'atteigne un état relâché.

Les informations obtenues dans les deux premières études nous ont permis de conclure que la boucle LCL de *Ll*Fpg est un élément flexible (ce qui est en bon accord avec les structures cristallographiques), et par conséquent le site actif de *Ll*Fpg possède plusieurs conformations. Pour mieux identifier quels sont les sites d'interaction protéine/petites molécules sur *Ll*Fpg et hNEIL1, nous avons mis en place une méthode de docking flexible aveugle en combinant des simulations de dynamique moléculaire classiques et une méthode d'amarrage moléculaire sans *a priori*. Cette méthode a été éprouvée en comparant les résultats obtenus avec des données expérimentales (structures cristallographiques) et avec les résultats issus du logiciel AMIDE. Nous avons remarqué que la prise en compte de la flexibilité des protéines est importante pour une exploration exhaustive de l'espace conformationnel des protéines. Grâce à notre méthode de docking flexible et aveugle, nous avons prédit deux sites de fixation principaux sur *Ll*Fpg et hNEIL1 correspondant aux sites actifs et aux

motifs H2TH (aussi appelés sites secondaires) des deux enzymes. Nous n'avons en revanche pas pu faire de lien entre les scores de docking et les IC_{50}. Cette difficulté provient probablement du fait que nous n'avons pas pu mettre en évidence de poses des ligands $2TX_n$ au voisinage du ZnF de *Ll*Fpg lors du docking.

L'observation par spectrométrie de masse de l'existence de formes oxydées des thionucléobases (un dimère pour 2TX et quatre trimères pour 2TX3) dans les préparations utilisées pour les tests d'inhibition des enzymes (données non présentées) nous a conduit à faire l'hypothèse que les formes actives de ces molécules pourraient être aussi leurs formes oxydées expliquant par exemple l'attaque du ZnF de *Ll*Fpg selon un processus de « thiol/disulfide exchange » (**Figure 121**).

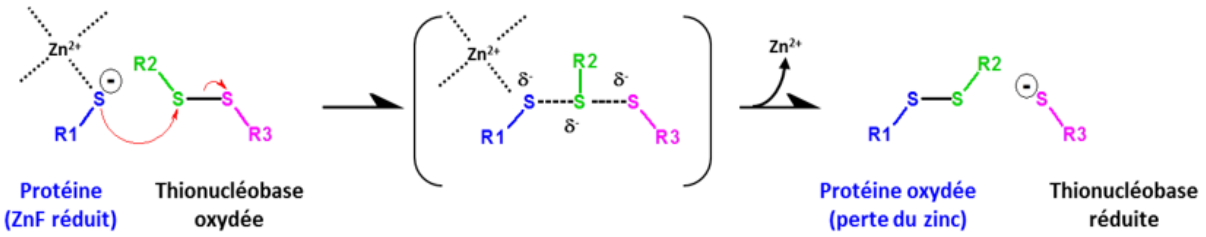


Figure 121 : Modèle proposé pour l'attaque du ZnF de *Ll*Fpg par les molécules $2TX_n$ oxydées

L'inhibition des protéines comportant un motif ZnLF comme hNEIL1 et MvNei1 suit un mécanisme obligatoirement en partie différent de celui impliqué dans l'inhibition de *Ll*Fpg. Cela pourrait suggérer que ces formes oxydées se fixent plus facilement à hNEIL1 et MvNei1 que les formes réduites monomériques des $2TX_n$. Cette hypothèse a aussi été renforcée par l'observation au laboratoire que l'inhibition de hNEIL1 et de MvNei1 était levée par le TCEP, un puissant réducteur des ponts disulfures (données non présentées). Pour évaluer cette hypothèse, le docking du dimère de 2TX et des trimères possibles de 2TX3 a été réalisé. Néanmoins, nous n'avons pas observé de conformations des ligands, monomères comme trimères des 2TX et 2TX3, à proximité du ZnF, ce qui ne permet pas de valider cette hypothèse. Grâce au docking flexible et aveugle, nous avons également mis en évidence la possibilité que les polymères de 2TX et 2TX3 puissent se fixer sur les sites actifs et les sites secondaires de *Ll*Fpg et hNEIL1 simultanément. D'autre part, les molécules testées ont peu d'affinité pour *Ll*Fpg et hNEIL1, ce qui signifie que l'effet d'inhibition peut être induit par des molécules monomériques et/ou polymériques.

D'autre part, le site actif et le site secondaire sont très proches dans l'espace. En perspective, la prédiction de sites de fixation de petites molécules sur les cibles *Ll*Fpg et hNEIL1 pourrait nous permettre de concevoir des inhibiteurs de nouvelle génération par une méthode de « Fragment Based

Drug Design » (FBDD). Cette méthode consiste à docker des petites molécules de type fragment (moins de 15 atomes) dans chacun des deux sites prédits, et d'identifier ceux qui possèdent le plus d'affinité pour chacun des deux sites. Ces fragments peuvent ensuite être reliés par un « linker », lui aussi optimisé en fonction de la surface protéique.

Le criblage virtuel d'une partie de la librairie de molécules d'Ambinter (~2 700 000 composés) dans le site actif de hNEIL1 nous a permis d'identifier les molécules les plus affines pour le site actif de l'enzyme. Parmi ces composés, nous avons sélectionné 5 molécules dont les valeurs de logP étaient les plus faibles afin de faciliter les tests d'activité sur l'enzyme hNEIL1. Nous avons identifié la petite molécule Amb6417670 comme étant un inhibiteur modéré (IC_{50} = 158 µM) de hNEIL1. À partir de cette molécule et de sa configuration dans le site actif de l'enzyme, nous avons identifié les interactions potentiellement importantes pour impacter l'activité catalytique de la cible, à savoir la proximité avec la P1 catalytique et les stackings avec les résidus Y176 et F255. Nous avons également remarqué que le groupe fonctionnel Imide est potentiellement impliqué dans l'inhibition de hNEIL1. La sélection d'autres molécules à partir de ces informations et des résultats du criblage virtuel pour de nouveaux tests d'activités devrait permettre d'identifier d'autres inhibiteurs potentiellement plus efficaces qu'Amb6417670. Aussi, à partir des données générées par le criblage virtuel, la construction d'un « scaffold » selon les interactions fréquemment formées entre le récepteur et les ligands peut également être envisagée. Ce scaffold peut servir de base et être pharmacomodulé pour la création d'un inhibiteur spécifique au site actif de la cible.

Nous avons étudié deux membres de la superfamille structure Fpg/Nei, et nous avons vu dans l'Introduction qu'il existait d'autres superfamilles structurales des ADN glycosylases. La recherche d'inhibiteurs peut également s'appliquer aux enzymes telles que MDB4 et OGG1 de la superfamille structurale HhH et faire l'objet d'autres recherches. À ce titre, un projet sera prochainement financé sur trois ans par la Région Centre-Val de Loire sur l'étude de hNEIL1 et hOGG1 ainsi qu'une bourse de thèse par La Ligue Contre le Cancer sur l'étude de hOGG1.

VI. **A**nnexes

Annexe A : Détail des « Latent Structural Clusters »

Annexe B : Paramètres, fichiers d'entrée et de sortie d'AD4 et d'ADV

Annexe C : RMSD des simulations de DM

Annexe D : Paramètres et fichiers de sortie des trois tests de docking sans a priori

Annexe E : Scripts de CAH des résultats des docking sans a priori

Annexe F : Alignement structural de *Ll*Fpg et de hNEIL1

Annexe G : Sites principaux prédits par le docking aveugle

Annexe H : Site secondaire dans les structures de *Ll*Fpg co-cristallisée avec 2TX, 2TX2, 2TX3 et F3CS

VI.1 **Annexe A : D**étail des « Latent Structural Clusters »

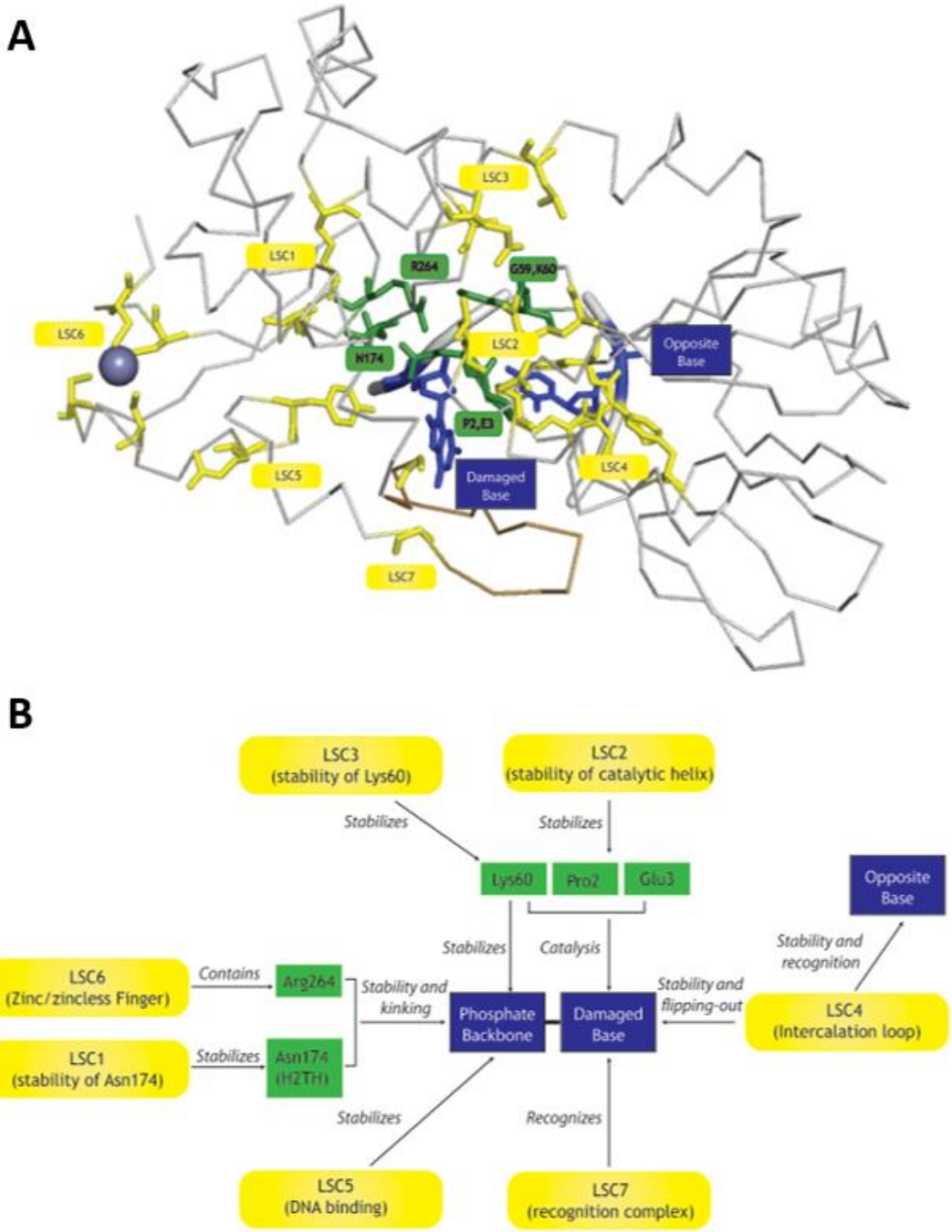


Figure 122 : Détails des LSC et aa impliqués dans la stabilisation de l'ADN
A) Structure de la protéine ainsi que les aa dont les chaines latérales sont associées au LSC 1 à 6. Les aa sont représentés en bâtonnets, le squelette de la protéine en fil de fer. **B**) Diagramme correspondant aux interactions (stabilisation, reconnaissance, catalyse) entre ces aa et l'ADN. Les aa en vert correspondent aux résidus les plus conservés dans la famille Fpg/Nei, à savoir R264 (contenu dans le LSC6), N174 (stabilisé par le LSC1), et K60 (stabilisé par LSC2/LSC3). Ces résidus stabilisent le groupement phosphate de la base endommagée. P2 et E3 font partie des aa catalytiques. L'hélice α composée de P2 et E3 sont également stabilisés par LSC2. L'enzyme extrait le dommage de la double hélice d'ADN et la boucle d'intercalation (LSC4) rempli le vide et interagit avec la base opposée. D'autres résidus importants ne sont pas inclus ici, tel que H74 et E6. L'aa Y242 impliqué dans la liaison de la protéine à l'ADN (faisant parti du LSC5) n'est pas déjà décrit dans la littérature [179].

VI.2 **Annexe B : P**aramètres, fichiers d'entrée et de sortie d'AD4 et d'ADV

VI.2.1 AD4

AD4 requiert au préalable la génération des grilles qui correspondent aux points de charge contenus dans la zone de recherche, celles-ci sont générées par le module `autogrid4` et nécessitent en entrée le fichier `.gpf` (« Grid parameter file »). Ce fichier contient les paramètres de grille comme le nombre de points dans les dimension x, y et z, l'espace en Å entre ces points, le type d'atome du ligand et du récepteur, ainsi que la liste des fichiers grilles `.map` a générer par la commande :

```
autogrid4 –p X.gpf –l X.glg
```

Une fois les grilles générées, le docking à proprement parler requiert la création d'un fichier `.dpf` contenant les noms des fichiers des grilles et les paramètres relatifs à l'AGL. Voici un exemple ci-dessous :

```
autodock_parameter_version 4.2       # used by autodock to validate
parameter set
outlev 1                             # diagnostic output level
intelec                              # calculate internal
electrostatics
seed pid time                        # seeds for random generator
ligand_types A HD N OA SA C P        # atoms types in ligand
fld 1PM5.maps.fld                    # grid_data_file
map 1PM5.A.map                       # atom-specific affinity map
map 1PM5.HD.map                      # atom-specific affinity map
map 1PM5.N.map                       # atom-specific affinity map
map 1PM5.OA.map                      # atom-specific affinity map
map 1PM5.SA.map                      # atom-specific affinity map
map 1PM5.C.map                       # atom-specific affinity map
map 1PM5.P.map                       # atom-specific affinity map
elecmap 1PM5.e.map                   # electrostatics map
desolvmap 1PM5.d.map                 # desolvation map
move VR_4_charges_avogadro.pdbqt     # small molecule
about 14.4083 -3.8894 -0.0498        # small molecule center
tran0 random                         # initial coordinates/A or
random
axisangle0 random                    # initial orientation
dihe0 random                         # initial dihedrals (relative)
or random
tstep 2.0                            # translation step/A
qstep 50.0                           # quaternion step/deg
dstep 50.0                           # torsion step/deg
```

```
torsdof 0                               # torsional degrees of freedom
rmstol 2.0                              # cluster_tolerance/A
extnrg 1000.0                           # external grid energy
e0max 0.0 10000                         # max initial energy; max
number of retries
ga_pop_size 150                         # number of individuals in
population
ga_num_evals 2500000                    # maximum number of energy
evaluations
ga_num_generations 27000                # maximum number of generations
ga_elitism 1                            # number of top individuals to
survive to next generation
ga_mutation_rate 0.02                   # rate of gene mutation
ga_crossover_rate 0.8                   # rate of crossover
ga_window_size 10                       #
ga_cauchy_alpha 0.0                     # Alpha parameter of Cauchy
distribution
ga_cauchy_beta 1.0                      # Beta parameter Cauchy
distribution
set_ga                                  # set the above parameters for
GA or LGA
sw_max_its 300                          # iterations of Solis & Wets
local search
sw_max_succ 4                           # consecutive successes before
changing rho
sw_max_fail 4                           # consecutive failures before
changing rho
sw_rho 1.0                              # size of local search space to
sample
sw_lb_rho 0.01                          # lower bound on rho
ls_search_freq 0.06                     # probability of performing
local search on individual
set_psw1                                # set the above pseudo-Solis &
Wets parameters
unbound_model extended                  # state of unbound ligand
ga_run 10                               # do this many hybrid GA-LS
runs
analysis                                # perform a ranked cluster
analysis
```

Le docking est effectué par la commande :

```
autodock4 -p X.dpf -l X.dlg
```

Les fichiers de sortie de AD4 sont au format `.dlg`, ils contiennent les structures des poses du ligand générées par le docking, les informations relatives à l'estimation de l'énergie d'affinité par le scoring et d'autres informations concernant l'AGL.

VI.2.2 ADV

Les paramètres « `receptor` » et « `ligand` » correspondent aux noms des fichiers au format `.pdbqt` (propre aux logiciels AutoDock) de la protéine et de la petite molécule à docker, respectivement. Le logiciel ADV effectue une recherche dans une zone délimitée par une « boite ». Cette boite est définie par un centre aux coordonnées *x*, *y* et *z*, ainsi qu'une taille dans ces 3 même dimensions. Le centre de cette boite dépend des paramètres « `center_x` », « `center_y` » et « `center_z` », et sa taille en fonction des paramètres « `size_x` », « `size_y` » et « `size_z` ». Le nombre de pose(s) retournée(s) à l'utilisateur est défini par le paramètre « `num_modes` » et l'écart de score autorisé entre la pose possédant la meilleure énergie d'affinité et la moins bonne est défini par le paramètre « `energy_range` ». Le nombre de run(s) effectué par un docking est conditionné par le paramètre « `exhaustiveness` » dont la valeur par défaut est de 8. L'utilisateur peut s'il le souhaite définir une « graine » dans le paramètre « `seed` », si ce paramètre est laissé vide, il est tiré au hasard. Ce paramètre conditionne l'aspect stochastique de l'algorithme de MC. Les variables « `out` » et « `log` » conditionnent le nom des fichiers de sorties des poses du docking et des scores respectivement. Le paramètre « `cpu` » défini les ressources que l'utilisateur souhaite allouer au calcul. L'accès à d'autres paramètres disponibles est possible en utilisant la commande `vina` suivi de l'argument `--advanced_help`. Les paramètres d'entrée d'ADV sont sauvés dans un fichier `config.txt`.

Autodock vina est lancé par la commande suivante :

```
vina --config config.txt
```

Les résultats retournés par ADV se présentent sous la forme d'un fichier `.pdbqt` contenant la liste de toutes les conformations du ligand générées (aussi appelées « poses ») dans la zone de recherche, ainsi qu'un fichier `.log` contenant les scores de chacune de ces conformations. ADV fourni également deux valeurs de RMSD par pose, représentant la différence structurale de ces conformations par rapport à la meilleure prédiction. La première RMSD est nommée « RMSD upper bond », elle correspond à un calcul de RMSD classique effectué entre l'atome i de la molécule A et l'atome i' correspondant de la molécule B. La « RMSD lower bound » (RMSD *lb*) à tendance à être plus faible que la « RMSD upper bound » (RMSD *ub*) car elle prend en compte la symétrie de la molécule et ne calcule que les distances les plus faibles entre les paires d'atome (**Figure 123**).

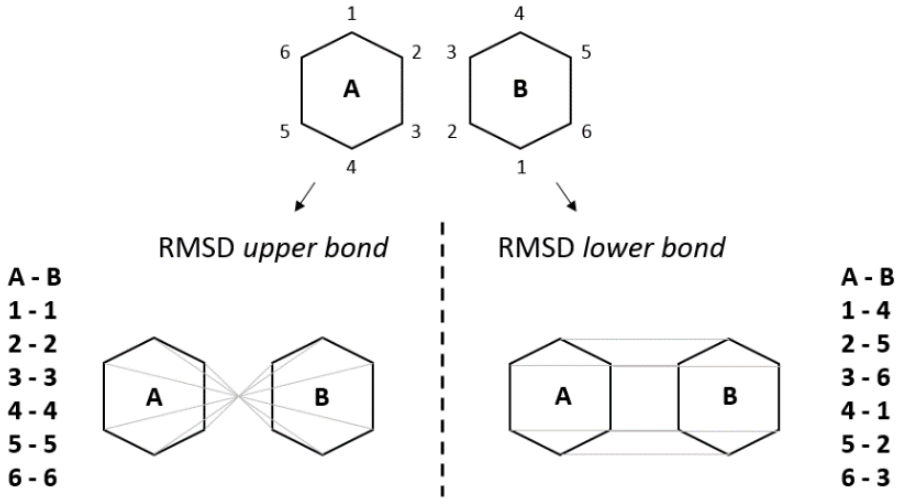
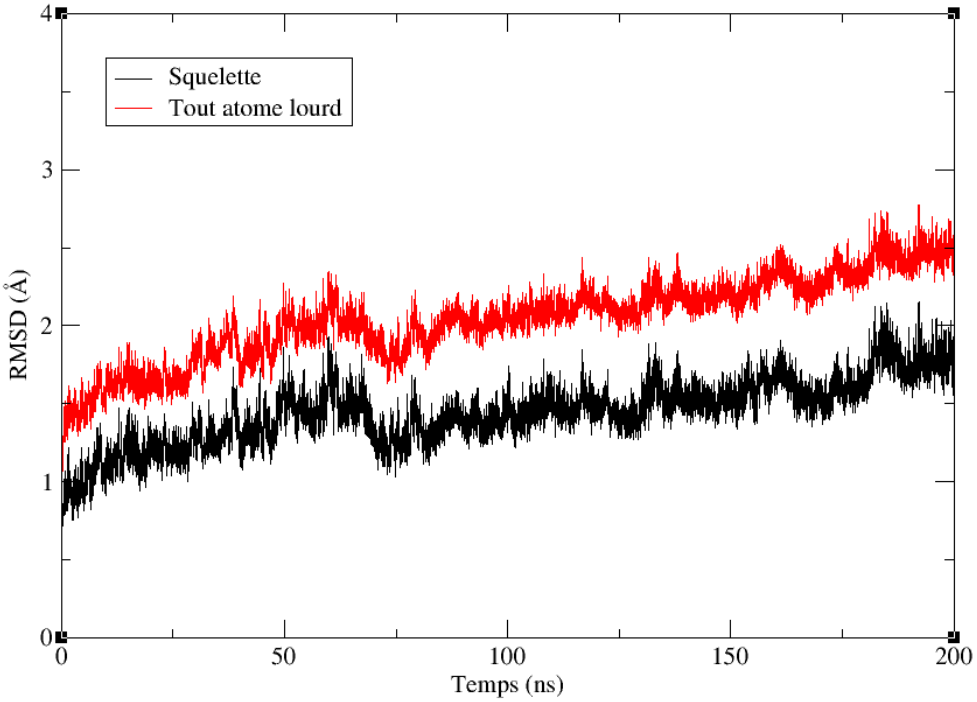


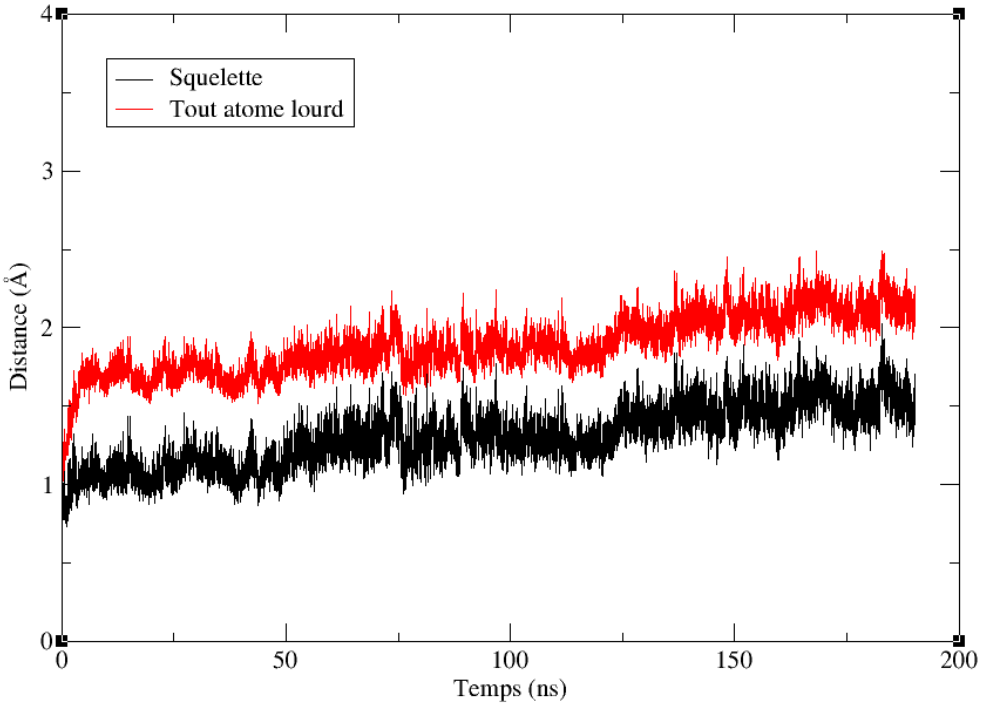
Figure 123 : Principe du calcul de la RMSD *up* et de la RMSD *lb* dans ADV
La RMSD *up* est calculée en respectant le numéro de l'atome, tandis que ce sont les distances les plus faibles entre deux atomes qui sont gardées pour le calcul de RMSD *lb*.

VI.3 **Annexe C : R**MSD des simulations de DM

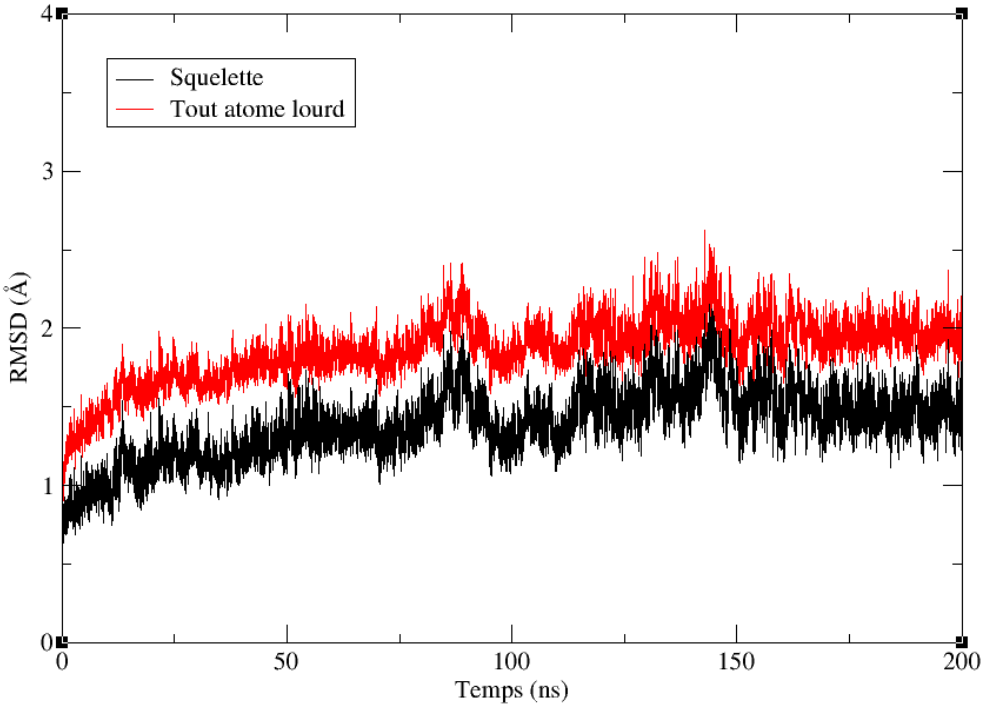
Fpg ADN THF boucle relâchée (1PM5)



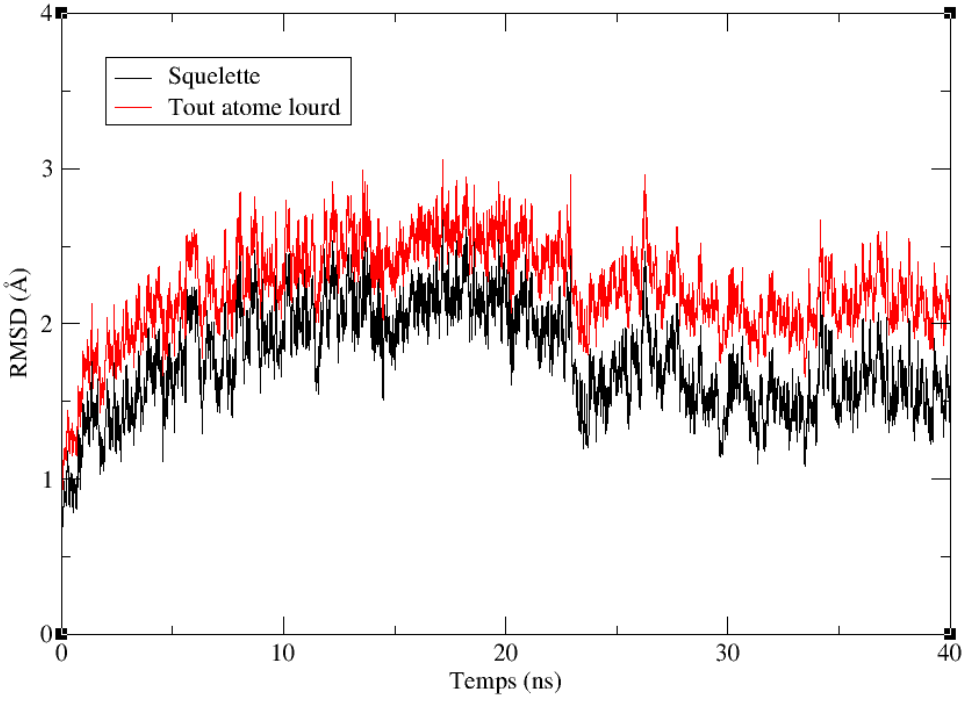
Fpg ADN FaPyG (1XC8) boucle LCL relâchée (1PM5)



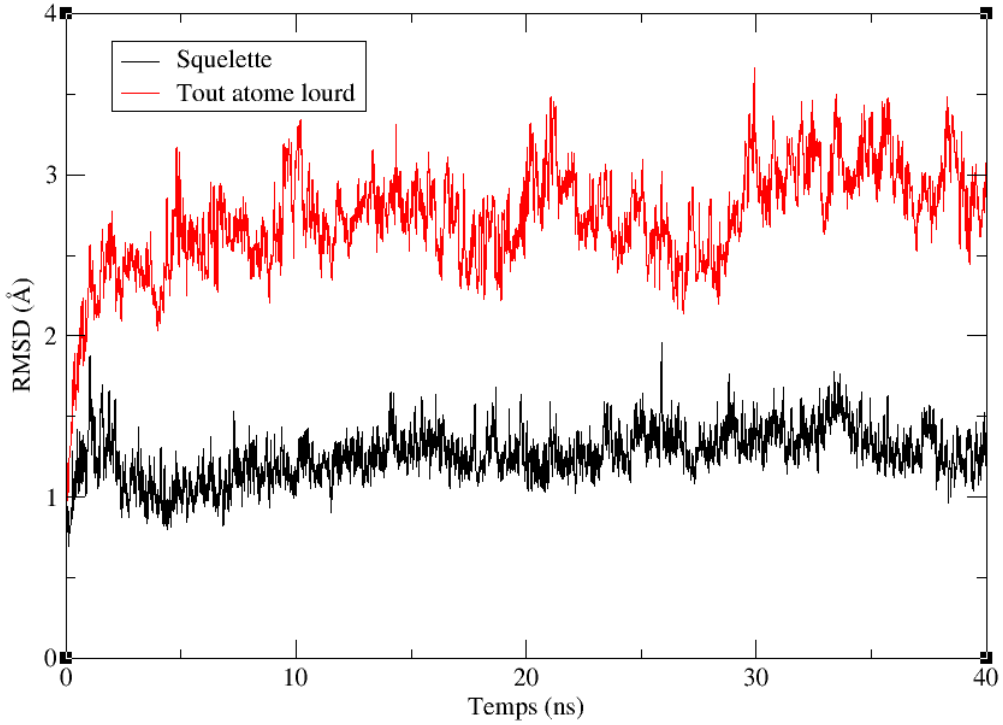
Fpg ADN faPyG boucle fermée (1XC8)



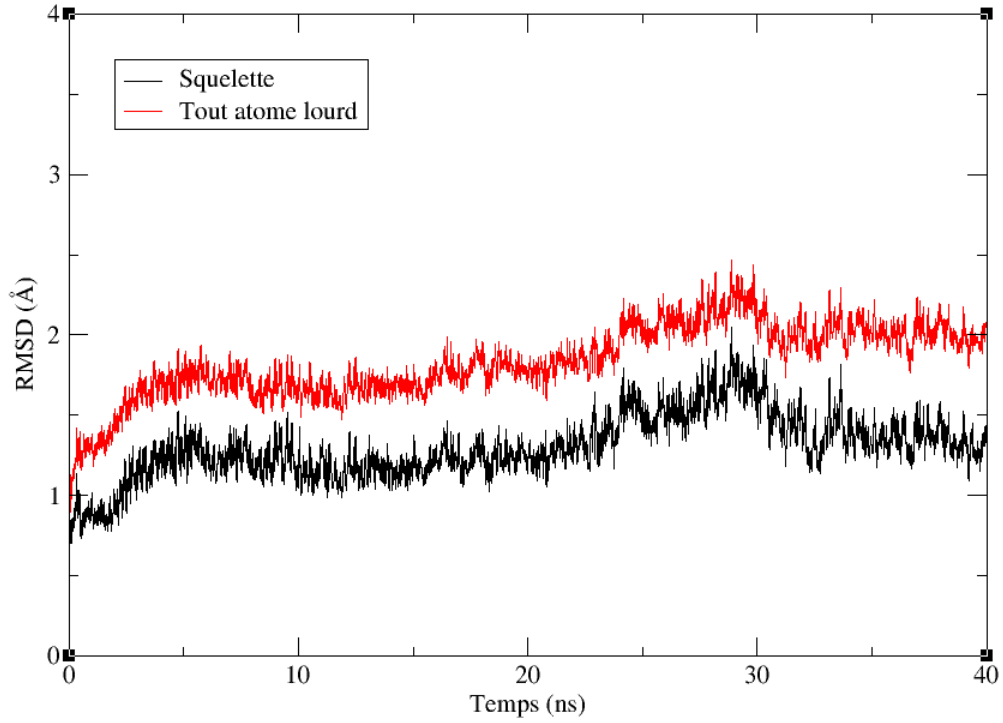
Fpg libre boucle relâchée (1PM5 sans ADN)



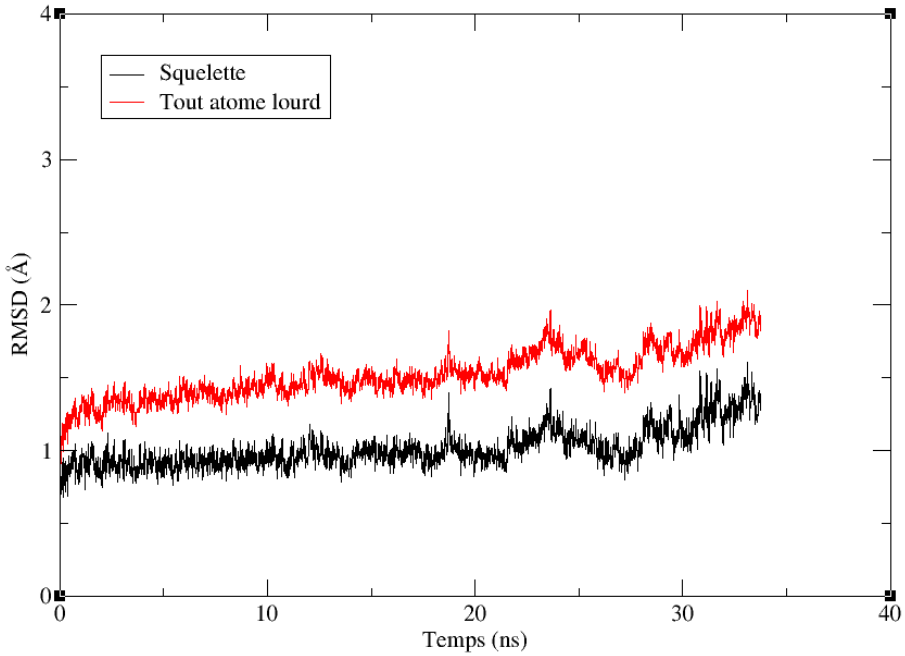
Fpg libre boucle fermée (1XC8 sans ADN)



Fpg ADN THF boucle relâchée (1PM5) 8-oxoG libre (1R2Y)



Fpg ADN THF (1PM5) boucle fermée (1XC8) 8-oxoG (1R2Y)



VI.4 **Annexe D : P**aramètres et fichiers de sortie des trois tests de docking sans *a priori*

VI.4.1 Paramètres

VI.4.1.2 Test n°1, AutoDock Vina

Le test n°1 du docking sans *a priori* consiste à explorer en une fois la totalité du système, ce qui fait une grande zone de recherche. Le volume de recherche est donc contenu dans une seule boite. Les paramètres pour des boites sont évidemment différents pour chacune des 30 structures produites par les simulations de DM et sont définis manuellement grâce à l'outil AutoDock Tools. Dans ce premier test, comme la boite englobe tout le système (protéine seule ou complexe ADN/protéine selon les cas), son volume dépasse 27 000 $Å^3$ ce qui fait l'objet d'un message d'avertissement par ADV lors du docking. Le paramètre « `exhaustiveness` » est comme son nom l'indique le paramètre d'exhaustivité, et correspond au nombre de « runs » qui seront effectués, par défaut, ADV effectue 8 runs par docking. Un run correspond à la génération d'une conformation du ligand et à son optimisation pour maximiser le nombre et la force des interactions qu'elle établit avec la protéine. Plus le nombre de *runs* est élevé, plus le temps de calcul l'est aussi. Plus la recherche est exhaustive et plus il y a de chance de générer les meilleures solutions, c'est-à-dire les meilleures interactions système/ligand possibles. Nous avons choisi d'effectuer 10000 runs, afin d'être sûr de bien explorer la totalité de la surface du système. Le paramètre « `energy_range` » autorise un écart de 20 $kcal.mol^{-1}$ entre les scores de docking le plus bas et le plus élevé. Le nombre de conformations du ligand générées est égal au paramètre « `num_mode` » et est égal à 20, la valeur maximale pour ce paramètre. ADV propose à l'utilisateur de choisir la valeur de la graine qui permet de générer les conformations du ligand lors des runs, nous avons choisi le paramètre « `random seed` », c'est-à-dire que les conformations du ligand seront générées aléatoirement.

VI.4.1.3 Tests n°2 et n°3, AMIDE

Les tests n°2 et n°3 ont été mise en place grâce au logiciel AMIDE. Cet outil permet de diviser la zone à explorer afin de permettre une recherche exhaustive, et de répartir les calculs sur plusieurs CPUs diminuant le temps nécessaire pour l'obtention des résultats. Nous avons pu utiliser AMIDE sur les ressources proposées par l'Université de Reims (supercalculateur ROMEO). Les paramètres dépendent du moteur de docking utilisé, à savoir AD4 ou ADV, et sont décrits dans l'Annexe B. Pour

ces deux tests, nous avons choisi de découper chacun des 30 systèmes en 12 boites chevauchantes. Les paramètres « center_x », « center_y », « center_z », « size_x », « size_y » et « size_z » de ces 12 boites sont générés automatiquement par AMIDE et sont adaptés à chacun des récepteurs fournis en entrée. Dans chacune de ces boites, le docking des tests n°2 et n°3 sont réalisés avec des graines générées aléatoirement, comme pour le test n°1.

VI.4.1.3.1 Moteur de docking : AutoDock 4

AMIDE prend un certain nombre de paramètres qui dépendent d'AD4 tel que le nombre de *runs* « Nb_runs » fixé sur 20, qui produira donc 20 conformations du ligand dans à la surface du système dans chacune des boites.

Les paramètres de l'Algorithme Génétique Lamarckien (AGL) utilisé pour le docking réalisé par AMIDE et AD4 en moteur de docking sont les paramètres par défaut et sont décrits dans la thèse de Romain Vasseur, docteur de l'université de Reims Champagne-Ardenne [282] :

ga_pop_size (taille de la population) : 150

ga_num_evals (nombre maximum d'évaluations de l'énergie) : 2 500 000

ga_num_generations (nombre maximum de générations) : 27 000

ga_elistim (nombre des meilleurs individus qui survivent d'une génération à l'autre) : 1

ga_mutation_rate (taux de mutations) : 0,02

ga_crossover_rate (taux de croissements) : 0,8

ga_cauchy_alpha (paramètre α de la distribution de Cauchy pour la mutation des gènes) : 0,0

ga_cauchy_beta (paramètre β de la distribution de Cauchy pour la mutation des gènes) : 1,0

ga_run (nombre de *runs*) : 10

Les paramètres de l'optimiseur local par défaut dans AMIDE sont les suivants :

sw_max_it (nombre d'itérations d'optimisation locale par génération) : 300

sw_max_succ (nombre de succès consécutifs de l'optimisation avant de modifier la taille du pas de l'opérateur de recherche locale (*p*)) : 4

sw_max_fail (nombre d'échecs consécutifs de l'optimisation avant de modifier la taille du pas de l'opérateur de recherche local (*p*)) : 4

sw_rho (opérateur d'optimisation) : 1

sw_lb_rho (taille de pas limite de *p*) : 0,01

ls_search_freq (probabilité qu'un individu soit sujet à l'optimisation locale) : 0,06

`set_psw1` (utilisation des paramètres d'optimisation pseudo-Solis et Wets)

`set_sw1` (utilisation des paramètres d'optimisation Solis et Wets)

VI.4.1.3.2 Moteur de docking : AutoDock Vina

Pour ce second docking, nous avons utilisés les paramètres ADV suivant :

`exhaustiveness` (nombre de conformations du ligand générés) : 40

`num_modes` (nombre de conformations du ligand en fin de docking) : 20

`energy_range` (différence maximale entre le score d'affinité le plus bas et le plus élevé) : 50

VI.4.2 Fichiers de sorties d'AutoDock 4 et d'AutoDock Vina

Les résultats retournés par ADV se présentent sous la forme d'un fichier `.pdbqt` contenant la liste de toutes les conformations du ligand générées dans la zone de recherche, ainsi qu'un fichier `.log` contenant les scores de chacune de ces conformations. ADV fourni également deux valeurs de RMSD par pose représentant la différence structurale de ces conformations par rapport à la meilleure prédiction (voir description dans la section **VI.2** p. **320** de l'Annexe B).

VI.5 **Annexe E : S**cripts de CAH des résultats des docking sans *a priori*

VI.5.1 Script docking_result_classification.sh

```
# Author : Charlotte Rieux, University of Orléans, France
#!/usr/bin/env bash

#$1 = au nombre de pose qu'on veut analyser par docking, min =
1, max = 20

#Liste des systèmes et des molécules
sys=$(echo "FD FS ND NS" | sed 's/ /\n/g')
mol=$(echo "2TX 2TX2 2TX3 2TX7 2TX13 2TX19 2TX23 F3CS" | sed
's/ /\n/g')

#Traitements
for i in $sys
do
     echo -n "Traitement système $i... "
     ./centres_masse.sh $i
     cat centres_masse_$i.pdb | awk '{print($7, $8, $9)}' >
centres_masse_$i.txt
    if [[ $i =~ "FD"  ]]; then
         Rscript hac.r centres_masse_$i.txt 10
    elif [[ $i =~ "FS"  ]]; then
        Rscript hac.r centres_masse_$i.txt 15
    elif [[ $i =~ "ND"  ]]; then
        Rscript hac.r centres_masse_$i.txt 9
    elif [[ $i =~ "NS"  ]]; then
        Rscript hac.r centres_masse_$i.txt 12
    fi
     sed -i '1d' cdm_classifiés_$i.txt
     sed -i s/\"//g cdm_classifiés_$i.txt
     echo "Done."
done

#Formatage
for i in $sys
do
     echo -n "Formatage système $i... "
     cat centres_masse_$i.pdb > centres_masse_classifiés_$i.pdb
     for j in $(cat cdm_classifiés_$i.txt | sed 's/ /,/g')
     do
          #echo $j
          com=$(echo $j | cut -d "," -f 1)
          clust=$(echo $j | cut -d "," -f 2)
          #echo "$com $clust"
          sed -i
'"$com"'s/^\(.*\)\(0.00\)\(.*\)$/\1'"$clust"'.00\3/g'
centres_masse_classifiés_$i.pdb
```

```
     done
     sed -i 's/^\(.\{60\}\)\(0\)\(.*\)$/\1\3/g'
centres_masse_classifiés_$i.pdb
     rm cdm_classifiés_$i.txt
     echo "Done."
done
rm Rplots.pdf

#Construction du tableau
centroid=$(cat centres_masse_classifiés_FD.pdb | awk
'{print($6)}' | sort -V | uniq)
echo -n "" > tableau_recap.txt
for i in $centroid
do
     for j in $sys
     do
          drug_list=$(cat centres_masse_classifiés_$j.pdb | awk
'{print($4)}' | sort | uniq | grep -v "xoG" | grep -v "pyG")
        first_drug=$(echo "$drug_list" | head -n 1)
          for k in $drug_list
          do
            if [[ $k =~ $first_drug ]]; then
                  text1_to_print=$(cat
centres_masse_classifiés_$j.pdb | grep $k | grep "A   $i" |
awk '{print($10, $11)}' | awk '{print("'"C$i"'", "'"$j"'", $1,
$2)}' | sort -t " " -rVk 4 | head -n $1)
                if [ $(echo "$text1_to_print" | wc -l) -ne $1
]; then
                    diff=$(($1 - $(echo "$text1_to_print" | wc
-l)))
                    l=1
                    while [ $l -le $diff ]; do
                        text1_to_print=$(echo -e
"$text1_to_print\nC$i $j 0.00 0.00")
                        l=$(($l + 1))
                    done
                fi
            else
                  text2_to_print=$(cat
centres_masse_classifiés_$j.pdb | grep $k | grep "A   $i" |
awk '{print($10, $11)}' | sort -t " " -rVk 2 | head -n $1)
                if [ $(echo "$text2_to_print" | wc -l) -ne $1
]; then
                    diff=$(($1 - $(echo "$text2_to_print" | wc
-l)))
                    l=1
                    while [ $l -le $diff ]; do
                        text2_to_print=$(echo -e
"$text2_to_print\n0.00 0.00")
                        l=$(($l + 1))
                    done
```

```
                fi
                text1_to_print=$(paste <(echo
"$text1_to_print") <(echo "$text2_to_print") --delimiters ' ')
            fi
          done
        if [[ $text1_to_print != '     ' ]]; then
            echo $drug_list >> tableau_recap.txt
            echo "$text1_to_print" >> tableau_recap.txt
        fi
     done
done
```

VI.5.2 Script centres_masses.sh appelé dans docking_result_classification.sh

```
# Author : Charlotte Rieux, University of Orléans, France
#!/usr/bin/env bash

echo -n "" > centres_masse_$1.pdb

#Calcul de tous les centres de masse de toutes les poses du docking
for file in $1/*/*.pdb
do

     atoms=$(cat $file)
     atom_nb=$(echo "$atoms" | wc -l)
     #echo "$atom_nb"

     #echo -e "MODEL $j\n$atoms"
     x_mean=$(echo "$atoms" | awk '{print(sum += $7)}' | tail -n 1 |
awk '{print($1/'"$atom_nb"')}' | sed 's/\./,/g')
     x_mean=$(printf "%.3f" $x_mean | sed 's/,/./g')
     #echo "$x_mean"
     y_mean=$(echo "$atoms" | awk '{print(sum += $8)}' | tail -n 1 |
awk '{print($1/'"$atom_nb"')}' | sed 's/\./,/g')
     y_mean=$(printf "%.3f" $y_mean | sed 's/,/./g')
     #echo "$y_mean"
     z_mean=$(echo "$atoms" | awk '{print(sum += $9)}' | tail -n 1 |
awk '{print($1/'"$atom_nb"')}' | sed 's/\./,/g')
     z_mean=$(printf "%.3f" $z_mean | sed 's/,/./g')
     #echo "$z_mean"

     score=$(echo $file | cut -d "_" -f 5 | cut -d '.' -f 1,2)
     mol=$(echo $file | cut -d "/" -f 2 | sed
's/\(.*\)\(...\)/\2/g')
     centroide=$(basename $file | cut -d "_" -f 1)
     conf=$(basename $file | cut -d "_" -f 3)

     #echo "$atoms"
     echo "$x_mean $y_mean $z_mean $score $mol $centroide $conf" |
column -t  >> centres_masse_$1.pdb

done
```

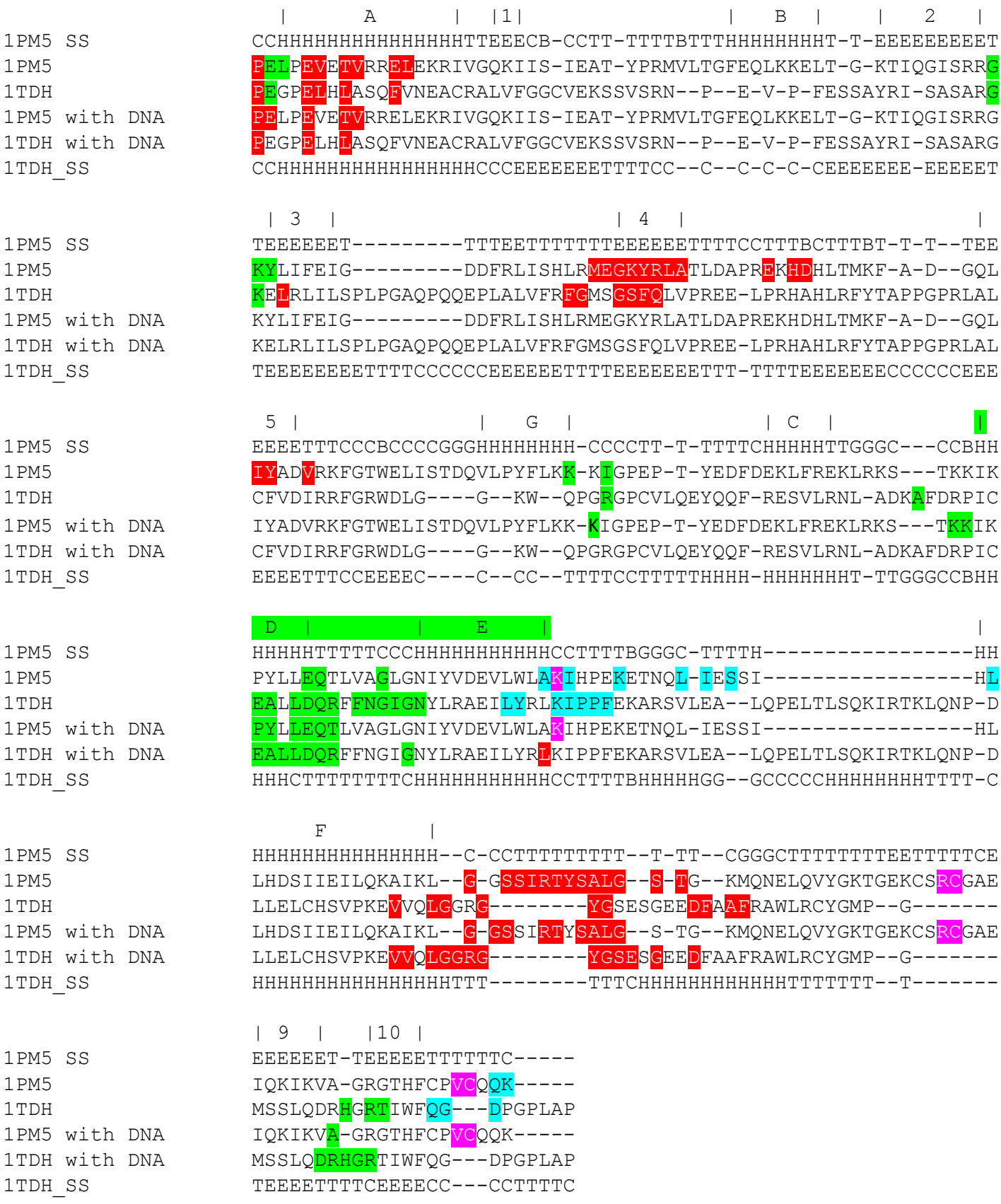
```
#Format PDB
x=$(cat centres_masse_$1.pdb | awk '{print($1)}' | cut -c 1-6 )
#echo "$x"
y=$(cat centres_masse_$1.pdb | awk '{print($2)}' | cut -c 1-6)
#echo "$y"
z=$(cat centres_masse_$1.pdb | awk '{print($3)}' | cut -c 1-6)
#echo "$z"
score=$(cat centres_masse_$1.pdb | awk '{print($4)}')
mol=$(cat centres_masse_$1.pdb | awk '{print($5)}')
centroide=$(cat centres_masse_$1.pdb | awk '{print($6)}' | cut -c 2)
conf=$(cat centres_masse_$1.pdb | awk '{print($7)}')

coordonnees=$(paste <(echo "$x") <(echo "$y")  <(echo "$z")  <(echo
"$score")  <(echo "$mol")  <(echo "$centroide") <(echo "$conf") --
delimiters ' ')
#echo "$coordonnees"

echo "$coordonnees" | awk '{print("'"ATOM      1  C   "'"$5"'" A
"'"$6"'"      "'"$1"'"  "'"$2"'"  "'"$3"'"  0.00 "'"$4, $7)}'  >
centres_masse_$1.pdb
```

VI.5.3 Script hac.r appelé dans docking_result_classification.sh

```
#!/usr/bin/Rscript


oldw <- getOption("warn")
options(warn = -1)

args <- commandArgs(TRUE)
file_name <- args[1]
cluster <- as.numeric(args[2])

data <- read.table(file_name, sep = " ")
dist <- as.dist(dist(data))
tree <- hclust(dist, method="ward")
plot(tree, xlab="")
dist$hcluster <- as.factor((cutree(tree, k=cluster) - 2) %% cluster
+ 1)
file_name <- sub("centres_masse_", "", file_name)
file_name <- sub(".txt", "", file_name)
file_name <- paste("cdm_classifiés_", file_name, sep = "")
file_name <- paste(file_name, ".txt", sep = "")
write.table(dist$hcluster, file = file_name)

options(warn = oldw)
```

VI.6 **Annexe F : A**lignement structural de *L*/Fpg et de hNEIL1

Alignement structural (MultiSeq [214], VMD [215]) de *L*/Fpg (1PM5) and hNEIL1 (1TDH):



Légende :
Site actif
Site secondaire (motif H2TH)
Doigt à zinc ou motif sans zinc

Les éléments annotés par des lettres sont des hélices, et ceux notés par des chiffres sont des feuillets beta.

Légende :
B = pont β (single pair b-sheet conformation, 2 aa de long au minimum)
E = brin β (parallèle ou anti-parallèle, 2 aa de long au minimum)
G = hélice 3_{10} (hélice à 3 tours, 3 aa de long au minimum)
H = hélice α (4 aa de long au minimum)
I = hélice pi (hélice à 5 tours, 5 aa de long au minimum)
S = coude (pas de liaison hydrogène)
T = tour (3, 4 ou 5 tours)
C = Pelote (aucune des conformations ci-dessus)

VI.7 **Annexe G : S**tructure équivalente à la boucle LCL chez hNEIL1

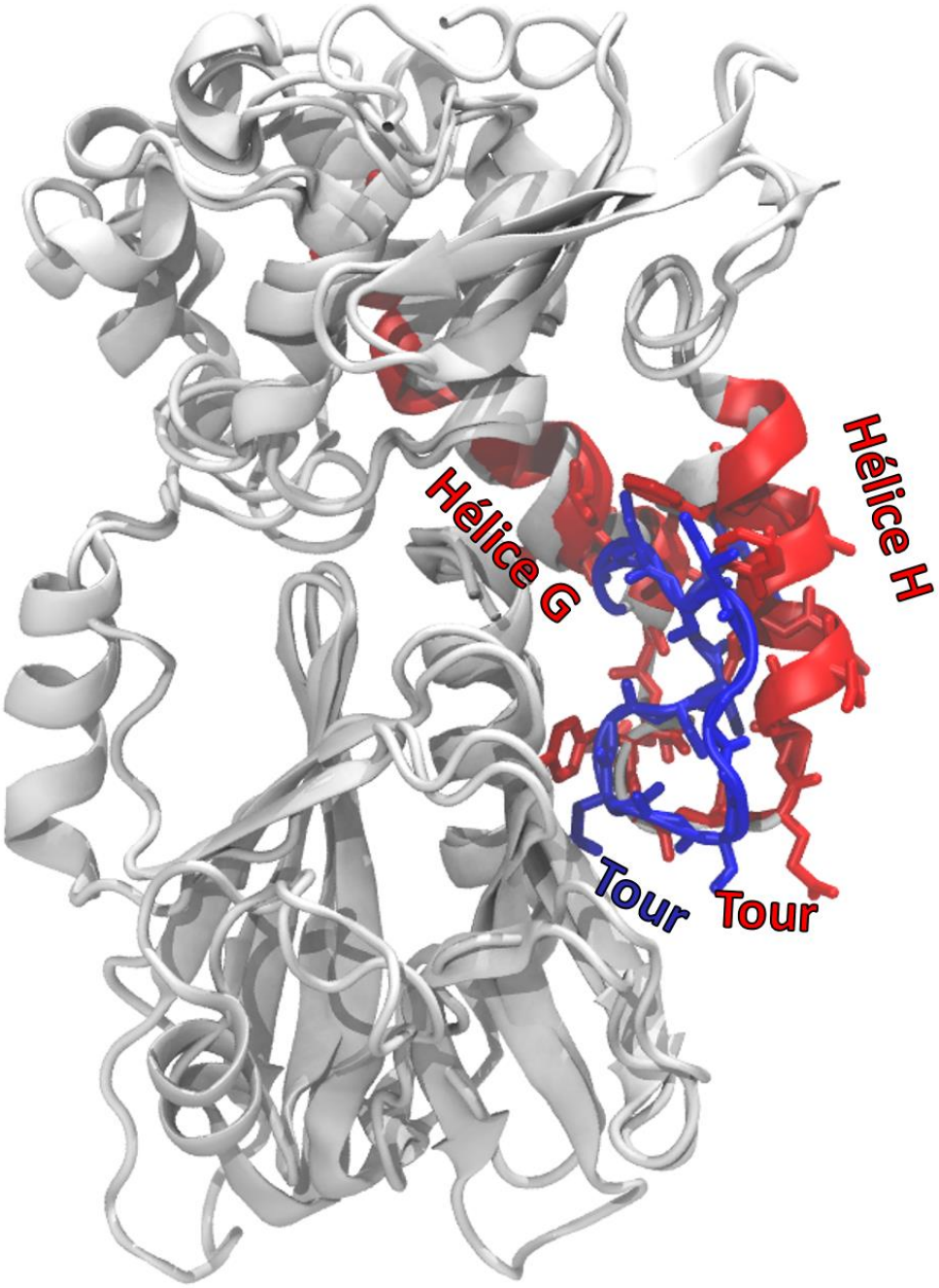


Figure 124 : Mise en évidence de la différence structurale entre la boucle LCL de *Ll*Fpg et la structure équivalente chez hNEIL1

La boucle LCL de *Ll*Fpg est représentée en bleu et l'objet structural équivalent de hNEIL1 (d'après l'alignement de l'annexe F) est représenté en rouge. Cette structure équivalente à la boucle LCL dans l'enzyme humain est elle aussi composée d'un Tour, cependant plus court (5 aa contre 11 aa chez *Ll*Fpg), intégré entre les deux Hélices α G et H.

VI.8 **Annexe H : S**ite secondaire dans les structures de *L*/Fpg co-cristallisée avec 2TX, 2TX2, 2TX3 et F3CS

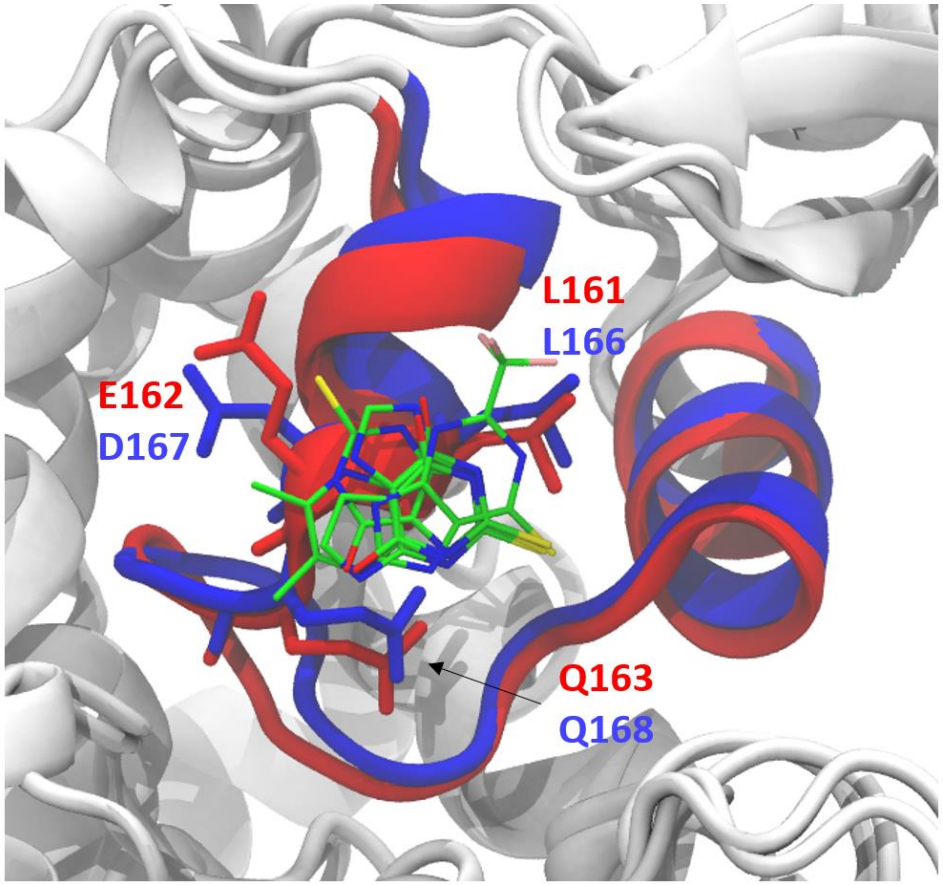


Figure 125 : En rouge le motif H2TH de *L*/Fpg et en bleu le motif H2TH de hNEIL1
Les $2TX_n$ co-cristallisées avec la structure de *L*/Fpg sont représentées en vert et interagissent avec les aa L161, E162 et Q163 de *L*/Fpg qui correspondent aux aa L166, D167 et Q168 de hNEIL1 après alignement structural des deux protéines. Ces aa correspondent aux positions principales en interaction avec les ligands dockés dans ces protéines libres ou complexées à l'ADN, validant en partie les résultats du docking aveugle présenté dans la section **IV.3.4.3.2** de la partie Résultats p. **251**.

VII. **B**ibliographie

1. Darwin, C., *L'Origine des espèces*. 1859.
2. Mendel, G., *Versuche über Pflanzenhybriden.* Verhandlungen des naturforschenden Vereines in Brünn, 1866. **IV für das Jahr 1865**: p. 3-47.
3. Watson, J.D. and F.H. Crick, *The structure of DNA.* Cold Spring Harb Symp Quant Biol, 1953. **18**: p. 123-31.
4. Wilkins, M.H., A.R. Stokes, and H.R. Wilson, *Molecular structure of deoxypentose nucleic acids.* Nature, 1953. **171**(4356): p. 738-40.
5. Hoogsteen, K., *The crystal and molecular structure of a hydrogen-bonded complex between 1-methylthymine and 9-methyladenine.* Acta Crystallographica, 1963. **16**(9): p. 907-916.
6. Duca, M., et al., *The triple helix: 50 years later, the outcome.* Nucleic Acids Res, 2008. **36**(16): p. 5123-38.
7. Dickerson, R.E., et al., *The anatomy of A-, B-, and Z-DNA.* Science, 1982. **216**(4545): p. 475-85.
8. Lindahl, T., *Instability and decay of the primary structure of DNA.* Nature, 1993. **362**(6422): p. 709-15.
9. Wilhelm, M., et al., *Mass-spectrometry-based draft of the human proteome.* Nature, 2014. **509**(7502): p. 582-7.
10. Chambers, I., et al., *The structure of the mouse glutathione peroxidase gene: the selenocysteine in the active site is encoded by the 'termination' codon, TGA.* EMBO J, 1986. **5**(6): p. 1221-7.
11. Mathews, C.K.v.H., K. E.; Appling, D. R.; Anthony-Cahill, S. J. , *Biochemistry (4th Edition)*. 4th ed. 2011.
12. Kabsch, W. and C. Sander, *Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features.* Biopolymers, 1983. **22**(12): p. 2577-637.
13. Lodovici, M. and E. Bigagli, *Oxidative stress and air pollution exposure.* J Toxicol, 2011. **2011**: p. 487074.
14. Sinha, R.P. and D.P. Hader, *UV-induced DNA damage and repair: a review.* Photochem Photobiol Sci, 2002. **1**(4): p. 225-36.
15. Mehta, A. and J.E. Haber, *Sources of DNA double-strand breaks and models of recombinational DNA repair.* Cold Spring Harb Perspect Biol, 2014. **6**(9): p. a016428.
16. Neeley, W.L. and J.M. Essigmann, *Mechanisms of formation, genotoxicity, and mutation of guanine oxidation products.* Chem Res Toxicol, 2006. **19**(4): p. 491-505.
17. Kalam, M.A., et al., *Genetic effects of oxidative DNA damages: comparative mutagenesis of the imidazole ring-opened formamidopyrimidines (Fapy lesions) and 8-oxo-purines in simian kidney cells.* Nucleic Acids Res, 2006. **34**(8): p. 2305-15.
18. Ward, J.F., *The complexity of DNA damage: relevance to biological consequences.* Int J Radiat Biol, 1994. **66**(5): p. 427-32.
19. Malkova, A. and J.E. Haber, *Mutations arising during repair of chromosome breaks.* Annu Rev Genet, 2012. **46**: p. 455-73.

20. Barker, S., M. Weinfeld, and D. Murray, *DNA-protein crosslinks: their induction, repair, and biological consequences.* Mutat Res, 2005. **589**(2): p. 111-35.

21. Gonzalez, V.M., et al., *Is cisplatin-induced cell death always produced by apoptosis?* Mol Pharmacol, 2001. **59**(4): p. 657-63.

22. Casares, C., et al., *Reactive oxygen species in apoptosis induced by cisplatin: review of physiopathological mechanisms in animal models.* Eur Arch Otorhinolaryngol, 2012. **269**(12): p. 2455-9.

23. Stordal, B. and M. Davey, *Understanding cisplatin resistance using cellular models.* IUBMB Life, 2007. **59**(11): p. 696-9.

24. Kelland, L.R., *New platinum antitumor complexes.* Crit Rev Oncol Hematol, 1993. **15**(3): p. 191-219.

25. Alberto, M.E., V. Butera, and N. Russo, *Which one among the Pt-containing anticancer drugs more easily forms monoadducts with G and A DNA bases? A comparative study among oxaliplatin, nedaplatin, and carboplatin.* Inorg Chem, 2011. **50**(15): p. 6965-71.

26. Goodsell, D.S., *The molecular perspective: ultraviolet light and pyrimidine dimers.* Oncologist, 2001. **6**(3): p. 298-9.

27. Setlow, R.B., *Cyclobutane-type pyrimidine dimers in polynucleotides.* Science, 1966. **153**(3734): p. 379-86.

28. Kiefer, J., *Effects of Ultraviolet Radiation on DNA*, in *Chromosomal Alterations: Methods, Results and Importance in Human Health*, G. Obe and Vijayalaxmi, Editors. 2007, Springer Berlin Heidelberg: Berlin, Heidelberg. p. 39-53.

29. Lerman, L.S., *Structural considerations in the interaction of DNA and acridines.* J Mol Biol, 1961. **3**: p. 18-30.

30. Lerman, L.S., *The structure of the DNA-acridine complex.* Proc Natl Acad Sci U S A, 1963. **49**: p. 94-102.

31. Efferth, T., et al., *Molecular target-guided tumor therapy with natural products derived from traditional Chinese medicine.* Curr Med Chem, 2007. **14**(19): p. 2024-32.

32. Waring, M.J., *Complex formation between ethidium bromide and nucleic acids.* J Mol Biol, 1965. **13**(1): p. 269-82.

33. Strekowski, L. and B. Wilson, *Noncovalent interactions with DNA: an overview.* Mutat Res, 2007. **623**(1-2): p. 3-13.

34. Elhamamsy, A.R., *DNA methylation dynamics in plants and mammals: overview of regulation and dysregulation.* Cell Biochem Funct, 2016. **34**(5): p. 289-98.

35. O'Brown, Z.K. and E.L. Greer, *N6-Methyladenine: A Conserved and Dynamic DNA Mark.* Adv Exp Med Biol, 2016. **945**: p. 213-246.

36. Bellacosa, A. and A.C. Drohat, *Role of base excision repair in maintaining the genetic and epigenetic integrity of CpG sites.* DNA Repair (Amst), 2015. **32**: p. 33-42.

37. Mattes, W.B., *DNA sequence selectivity of guanine-N7 alkylation by nitrogen mustards.* Nucleic Acids Res, 1986. **14**(7): p. 2971-2987.

38. Dong, Q., et al., *A structural basis for a phosphoramide mustard-induced DNA interstrand cross-link at 5'-d(GAC).* Proc Natl Acad Sci U S A, 1995. **92**(26): p. 12170-4.

39. Guainazzi, A. and O.D. Scharer, *Using synthetic DNA interstrand crosslinks to elucidate repair pathways and identify new therapeutic targets for cancer chemotherapy.* Cell Mol Life Sci, 2010. **67**(21): p. 3683-97.

40. Gates, K.S., *An overview of chemical processes that damage cellular DNA: spontaneous hydrolysis, alkylation, and reactions with radicals.* Chem Res Toxicol, 2009. **22**(11): p. 1747-60.

41. Maxam, A.M. and W. Gilbert, *A new method for sequencing DNA.* Proc Natl Acad Sci U S A, 1977. **74**(2): p. 560-4.

42. Hayyan, M., M.A. Hashim, and I.M. AlNashef, *Superoxide Ion: Generation and Chemical Implications.* Chem Rev, 2016. **116**(5): p. 3029-85.

43. Fenton, H.J.H., *Oxidation of tartaric acid in presence of iron.* J. Chem. Soc, Trans., 1894. **65**(65): p. 899-911.

44. Koppenol, W.H., *The Haber-Weiss cycle--70 years later.* Redox Rep, 2001. **6**(4): p. 229-34.

45. Jackson, J.H., et al., *Damage to the bases in DNA induced by stimulated human neutrophils.* J Clin Invest, 1989. **84**(5): p. 1644-9.

46. Monastyrska, I. and D.J. Klionsky, *Autophagy in organelle homeostasis: peroxisome turnover.* Mol Aspects Med, 2006. **27**(5-6): p. 483-94.

47. Bartosz, G., *Reactive oxygen species: destroyers or messengers?* Biochem Pharmacol, 2009. **77**(8): p. 1303-15.

48. Valko, M., et al., *Free radicals and antioxidants in normal physiological functions and human disease.* Int J Biochem Cell Biol, 2007. **39**(1): p. 44-84.

49. Till, A., et al., *Pexophagy: the selective degradation of peroxisomes.* Int J Cell Biol, 2012. **2012**: p. 512721.

50. Dalhus, B., et al., *DNA base repair--recognition and initiation of catalysis.* FEMS Microbiol Rev, 2009. **33**(6): p. 1044-78.

51. Jena, N.R. and P.C. Mishra, *Is FapyG mutagenic?: Evidence from the DFT study.* Chemphyschem, 2013. **14**(14): p. 3263-70.

52. O'Connor, T.R., S. Boiteux, and J. Laval, *Ring-opened 7-methylguanine residues in DNA are a block to in vitro DNA synthesis.* Nucleic Acids Res, 1988. **16**(13): p. 5879-94.

53. Burrows, C.J., et al., *Structure and potential mutagenicity of new hydantoin products from guanosine and 8-oxo-7,8-dihydroguanine oxidation by transition metals.* Environ Health Perspect, 2002. **110 Suppl 5**: p. 713-7.

54. Boiteux, S., F. Coste, and B. Castaing, *Repair of 8-oxo-7,8-dihydroguanine in prokaryotic and eukaryotic cells: Properties and biological roles of the Fpg and OGG1 DNA N-glycosylases.* Free Radic Biol Med, 2016.

55. Brooks, S.C., et al., *Recent advances in the structural mechanisms of DNA glycosylases.* Biochim Biophys Acta, 2013. **1834**(1): p. 247-71.

56. Yi, C. and C. He, *DNA repair by reversal of DNA damage.* Cold Spring Harb Perspect Biol, 2013. **5**(1): p. a012575.

57. Maul, M.J., et al., *Crystal structure and mechanism of a DNA (6-4) photolyase.* Angew Chem Int Ed Engl, 2008. **47**(52): p. 10076-80.

58. Lindahl, T., et al., *Regulation and expression of the adaptive response to alkylating agents.* Annu Rev Biochem, 1988. **57**: p. 133-57.

59. Fedeles, B.I., et al., *The AlkB Family of Fe(II)/alpha-Ketoglutarate-dependent Dioxygenases: Repairing Nucleic Acid Alkylation Damage and Beyond.* J Biol Chem, 2015. **290**(34): p. 20734-42.

60. Watson JD, B.S., Gann A, Levine M, Losick R, *Molecular Biology of the Gene*. CSHL Press 5th ed. ed. 2004.

61. Alberts, B.J., A.; Lewis, J.; Raff, M.; Roberts, K.; Walter, P., *Molecular Biology of the Cell*. 5th ed. ed. 2008.

62. Ciccia, A. and S.J. Elledge, *The DNA damage response: making it safe to play with knives.* Mol Cell, 2010. **40**(2): p. 179-204.

63. Marcon, E. and P.B. Moens, *The evolution of meiosis: recruitment and modification of somatic DNA-repair proteins.* Bioessays, 2005. **27**(8): p. 795-808.

64. Mimitou, E.P. and L.S. Symington, *Nucleases and helicases take center stage in homologous recombination.* Trends Biochem Sci, 2009. **34**(5): p. 264-72.

65. Renkawitz, J., C.A. Lademann, and S. Jentsch, *Mechanisms and principles of homology search during recombination.* Nat Rev Mol Cell Biol, 2014. **15**(6): p. 369-83.

66. Heyer, W.D., K.T. Ehmsen, and J. Liu, *Regulation of homologous recombination in eukaryotes.* Annu Rev Genet, 2010. **44**: p. 113-39.

67. Sung, P. and H. Klein, *Mechanism of homologous recombination: mediators and helicases take on regulatory functions.* Nat Rev Mol Cell Biol, 2006. **7**(10): p. 739-50.

68. Andersen, S.L. and J. Sekelsky, *Meiotic versus mitotic recombination: two different routes for double-strand break repair: the different functions of meiotic versus mitotic DSB repair are reflected in different pathway usage and different outcomes.* Bioessays, 2010. **32**(12): p. 1058-66.

69. McVey, M. and S.E. Lee, *MMEJ repair of double-strand breaks (director's cut): deleted sequences and alternative endings.* Trends Genet, 2008. **24**(11): p. 529-38.

70. Iyer, R.R., et al., *DNA mismatch repair: functions and mechanisms.* Chem Rev, 2006. **106**(2): p. 302-23.

71. Larrea, A.A., S.A. Lujan, and T.A. Kunkel, *SnapShot: DNA mismatch repair.* Cell, 2010. **141**(4): p. 730 e1.

72. Michaels, M.L., et al., *Evidence that MutY and MutM combine to prevent mutations by an oxidatively damaged form of guanine in DNA.* Proc Natl Acad Sci U S A, 1992. **89**(15): p. 7022-5.

73. Qiu, R., et al., *MutL traps MutS at a DNA mismatch.* Proc Natl Acad Sci U S A, 2015. **112**(35): p. 10914-9.

74. Kanehisa, M. and S. Goto, *KEGG: kyoto encyclopedia of genes and genomes.* Nucleic Acids Res, 2000. **28**(1): p. 27-30.

75. Putnam, C.D., *Evolution of the methyl directed mismatch repair system in Escherichia coli.* DNA Repair (Amst), 2016. **38**: p. 32-41.

76. Le May, N., J.M. Egly, and F. Coin, *True lies: the double life of the nucleotide excision repair factors in transcription and DNA repair.* J Nucleic Acids, 2010. **2010**.

77. de Melo, J.T., et al., *XPC deficiency is related to APE1 and OGG1 expression and function.* Mutat Res, 2016. **784-785**: p. 25-33.

78. Costa, R.M., et al., *The eukaryotic nucleotide excision repair pathway.* Biochimie, 2003. **85**(11): p. 1083-99.

79. Daley, J.M., C. Zakaria, and D. Ramotar, *The endonuclease IV family of apurinic/apyrimidinic endonucleases.* Mutat Res, 2010. **705**(3): p. 217-27.

80. Dalhus, B., I. Alseth, and M. Bjoras, *Structural basis for incision at deaminated adenines in DNA and RNA by endonuclease V.* Prog Biophys Mol Biol, 2015. **117**(2-3): p. 134-42.

81. Prorok, P., et al., *Uracil in duplex DNA is a substrate for the nucleotide incision repair pathway in human cells.* Proc Natl Acad Sci U S A, 2013. **110**(39): p. E3695-703.

82. Timofeyeva, N.A., et al., *Kinetic mechanism of human apurinic/apyrimidinic endonuclease action in nucleotide incision repair.* Biochemistry (Mosc), 2011. **76**(2): p. 273-81.

83. Couvé-Privat, S., et al., *Nucleotide Incision Repair: An Alternative and Ubiquitous Pathway to Handle Oxidative DNA Damage*, in *Oxidative Damage to Nucleic Acids*, M.D. Evans and M.S. Cooke, Editors. 2007, Springer New York: New York, NY. p. 54-66.

84. Krokan, H.E. and M. Bjoras, *Base excision repair.* Cold Spring Harb Perspect Biol, 2013. **5**(4): p. a012583.

85. Fortini, P. and E. Dogliotti, *Base damage and single-strand break repair: mechanisms and functional significance of short- and long-patch repair subpathways.* DNA Repair (Amst), 2007. **6**(4): p. 398-409.

86. Caldecott, K.W., *XRCC1 and DNA strand break repair.* DNA Repair (Amst), 2003. **2**(9): p. 955-69.

87. Caldecott, K.W., *Single-strand break repair and genetic disease.* Nat Rev Genet, 2008. **9**(8): p. 619-31.

88. Le Page, F., et al., *Poly(ADP-ribose) polymerase-1 (PARP-1) is required in murine cell lines for base excision repair of oxidative DNA damage in the absence of DNA polymerase beta.* J Biol Chem, 2003. **278**(20): p. 18471-7.

89. Jagtap, P. and C. Szabo, *Poly(ADP-ribose) polymerase and the therapeutic effects of its inhibitors.* Nat Rev Drug Discov, 2005. **4**(5): p. 421-40.

90. Khodyreva, S.N., et al., *Apurinic/apyrimidinic (AP) site recognition by the 5'-dRP/AP lyase in poly(ADP-ribose) polymerase-1 (PARP-1).* Proc Natl Acad Sci U S A, 2010. **107**(51): p. 22090-5.

91. Horton, J.K., et al., *XRCC1 and DNA polymerase beta in cellular protection against cytotoxic DNA single-strand breaks.* Cell Res, 2008. **18**(1): p. 48-63.

92. Petermann, E., C. Keil, and S.L. Oei, *Roles of DNA ligase III and XRCC1 in regulating the switch between short patch and long patch BER.* DNA Repair (Amst), 2006. **5**(5): p. 544-55.

93. Burkovics, P., et al., *Human Ape2 protein has a 3'-5' exonuclease activity that acts preferentially on mismatched base pairs.* Nucleic Acids Res, 2006. **34**(9): p. 2508-15.

94. Unk, I., et al., *3'-phosphodiesterase and 3'-->5' exonuclease activities of yeast Apn2 protein and requirement of these activities for repair of oxidative DNA damage.* Mol Cell Biol, 2001. **21**(5): p. 1656-61.

95. Mol, C.D., et al., *DNA-bound structures and mutants reveal abasic DNA binding by APE1 and DNA repair coordination [corrected].* Nature, 2000. **403**(6768): p. 451-6.

96. Feng, J.A., C.J. Crasto, and Y. Matsumoto, *Deoxyribose phosphate excision by the N-terminal domain of the polymerase beta: the mechanism revisited.* Biochemistry, 1998. **37**(27): p. 9605-11.

97. Stucki, M., et al., *Mammalian base excision repair by DNA polymerases delta and epsilon.* Oncogene, 1998. **17**(7): p. 835-43.

98. Balakrishnan, L. and R.A. Bambara, *Flap endonuclease 1.* Annu Rev Biochem, 2013. **82**: p. 119-38.

99. Arakawa, H., et al., *Functional redundancy between DNA ligases I and III in DNA replication in vertebrate cells.* Nucleic Acids Res, 2012. **40**(6): p. 2599-610.

100. Copeland, W.C., *The mitochondrial DNA polymerase in health and disease.* Subcell Biochem, 2010. **50**: p. 211-22.

101. Wallace, S.S., *Base excision repair: a critical player in many games.* DNA Repair (Amst), 2014. **19**: p. 14-26.

102. Smith, Z.D. and A. Meissner, *DNA methylation: roles in mammalian development.* Nat Rev Genet, 2013. **14**(3): p. 204-20.

103. Wu, H. and Y. Zhang, *Reversing DNA methylation: mechanisms, genomics, and biological functions.* Cell, 2014. **156**(1-2): p. 45-68.

104. Deaton, A.M. and A. Bird, *CpG islands and the regulation of transcription.* Genes Dev, 2011. **25**(10): p. 1010-22.

105. Drohat, A.C. and C.T. Coey, *Role of Base Excision "Repair" Enzymes in Erasing Epigenetic Marks from DNA.* Chem Rev, 2016. **116**(20): p. 12711-12729.

106. Maiti, A. and A.C. Drohat, *Thymine DNA glycosylase can rapidly excise 5-formylcytosine and 5-carboxylcytosine: potential implications for active demethylation of CpG sites.* J Biol Chem, 2011. **286**(41): p. 35334-8.

107. He, Y.F., et al., *Tet-mediated formation of 5-carboxylcytosine and its excision by TDG in mammalian DNA.* Science, 2011. **333**(6047): p. 1303-7.

108. Ito, S., et al., *Tet proteins can convert 5-methylcytosine to 5-formylcytosine and 5-carboxylcytosine.* Science, 2011. **333**(6047): p. 1300-3.

109. Muller, U., et al., *TET-mediated oxidation of methylcytosine causes TDG or NEIL glycosylase dependent gene reactivation.* Nucleic Acids Res, 2014. **42**(13): p. 8592-604.

110. Hwang, J.K., F.W. Alt, and L.S. Yeap, *Related Mechanisms of Antibody Somatic Hypermutation and Class Switch Recombination.* Microbiol Spectr, 2015. **3**(1): p. MDNA3-0037-2014.

111. Maul, R.W. and P.J. Gearhart, *Refining the Neuberger model: Uracil processing by activated B cells.* Eur J Immunol, 2014. **44**(7): p. 1913-6.

112. Zharkov, D.O., *Base excision DNA repair.* Cell Mol Life Sci, 2008. **65**(10): p. 1544-65.

113. Yang, B., et al., *Virion-associated uracil DNA glycosylase-2 and apurinic/apyrimidinic endonuclease are involved in the degradation of APOBEC3G-edited nascent HIV-1 DNA.* J Biol Chem, 2007. **282**(16): p. 11667-75.

114. Mostoslavsky, R., et al., *Genomic instability and aging-like phenotype in the absence of mammalian SIRT6.* Cell, 2006. **124**(2): p. 315-29.

115. Jia, P., C. Her, and W. Chai, *DNA excision repair at telomeres.* DNA Repair (Amst), 2015. **36**: p. 137-45.

116. Massaad, M.J., et al., *Deficiency of base excision repair enzyme NEIL3 drives increased predisposition to autoimmunity.* J Clin Invest, 2016. **126**(11): p. 4219-4236.

117. Wilson, D.M., 3rd and V.A. Bohr, *The mechanics of base excision repair, and its relationship to aging and disease.* DNA Repair (Amst), 2007. **6**(4): p. 544-59.

118. Senejani, A.G., et al., *Mutation of POLB causes lupus in mice.* Cell Rep, 2014. **6**(1): p. 1-8.

119. Howard, J.H., et al., *Epigenetic downregulation of the DNA repair gene MED1/MBD4 in colorectal and ovarian cancer.* Cancer Biol Ther, 2009. **8**(1): p. 94-100.

120. Suzuki, T., H. Harashima, and H. Kamiya, *Effects of base excision repair proteins on mutagenesis by 8-oxo-7,8-dihydroguanine (8-hydroxyguanine) paired with cytosine and adenine.* DNA Repair (Amst), 2010. **9**(5): p. 542-50.

121. Shinmura, K., et al., *Inactivating mutations of the human base excision repair gene NEIL1 in gastric cancer.* Carcinogenesis, 2004. **25**(12): p. 2311-7.

122. Do, H., et al., *A critical re-assessment of DNA repair gene promoter methylation in non-small cell lung carcinoma.* Sci Rep, 2014. **4**: p. 4186.

123. Mao, G., et al., *Identification and characterization of OGG1 mutations in patients with Alzheimer's disease.* Nucleic Acids Res, 2007. **35**(8): p. 2759-66.

124. Liu, Y., et al., *Coordination between polymerase beta and FEN1 can modulate CAG repeat expansion.* J Biol Chem, 2009. **284**(41): p. 28352-66.

125. Kemmerich, K., et al., *Germline ablation of SMUG1 DNA glycosylase causes loss of 5-hydroxymethyluracil- and UNG-backup uracil-excision activities and increases cancer predisposition of Ung-/-Msh2-/- mice.* Nucleic Acids Res, 2012. **40**(13): p. 6016-25.

126. Rada, C., et al., *Immunoglobulin isotype switching is inhibited and somatic hypermutation perturbed in UNG-deficient mice.* Curr Biol, 2002. **12**(20): p. 1748-55.

127. Nilsen, H., et al., *Gene-targeted mice lacking the Ung uracil-DNA glycosylase develop B-cell lymphomas.* Oncogene, 2003. **22**(35): p. 5381-6.

128. Imai, K., et al., *Human uracil-DNA glycosylase deficiency associated with profoundly impaired immunoglobulin class-switch recombination.* Nat Immunol, 2003. **4**(10): p. 1023-8.

129. Cortazar, D., et al., *Embryonic lethal phenotype reveals a function of TDG in maintaining epigenetic stability.* Nature, 2011. **470**(7334): p. 419-23.

130. Millar, C.B., et al., *Enhanced CpG mutability and tumorigenesis in MBD4-deficient mice.* Science, 2002. **297**(5580): p. 403-5.

131. Sidorenko, V.S. and D.O. Zharkov, *[The role of glycosylases of the base excision DNA repair in pathogenesis of hereditary and infectious human diseases].* Mol Biol (Mosk), 2008. **42**(5): p. 891-903.

132. Elder, R.H., et al., *Alkylpurine-DNA-N-glycosylase knockout mice show increased susceptibility to induction of mutations by methyl methanesulfonate.* Mol Cell Biol, 1998. **18**(10): p. 5828-37.

133. Klungland, A., et al., *Accumulation of premutagenic DNA lesions in mice defective in removal of oxidative base damage.* Proc Natl Acad Sci U S A, 1999. **96**(23): p. 13300-5.

134. Halsne, R., et al., *Lack of the DNA glycosylases MYH and OGG1 in the cancer prone double mutant mouse does not increase mitochondrial DNA mutagenesis.* DNA Repair (Amst), 2012. **11**(3): p. 278-85.

135. Kovtun, I.V., et al., *OGG1 initiates age-dependent CAG trinucleotide expansion in somatic cells.* Nature, 2007. **447**(7143): p. 447-52.

136. Kohno, T., et al., *Genetic polymorphisms and alternative splicing of the hOGG1 gene, that is involved in the repair of 8-hydroxyguanine in damaged DNA.* Oncogene, 1998. **16**(25): p. 3219-25.

137. Xie, Y., et al., *Deficiencies in mouse Myh and Ogg1 result in tumor predisposition and G to T mutations in codon 12 of the K-ras oncogene in lung tumors.* Cancer Res, 2004. **64**(9): p. 3096-102.

138. Al-Tassan, N., et al., *Inherited variants of MYH associated with somatic G:C-->T:A mutations in colorectal tumors.* Nat Genet, 2002. **30**(2): p. 227-32.

139. Jones, S., et al., *Increased frequency of the k-ras G12C mutation in MYH polyposis colorectal adenomas.* Br J Cancer, 2004. **90**(8): p. 1591-3.

140. Ocampo, M.T., et al., *Targeted deletion of mNth1 reveals a novel DNA repair enzyme activity.* Mol Cell Biol, 2002. **22**(17): p. 6111-21.

141. Amara, P., et al., *Insights into the DNA repair process by the formamidopyrimidine-DNA glycosylase investigated by molecular dynamics.* Protein Sci, 2004. **13**(8): p. 2009-21.

142. Vartanian, V., et al., *The metabolic syndrome resulting from a knockout of the NEIL1 DNA glycosylase.* Proc Natl Acad Sci U S A, 2006. **103**(6): p. 1864-9.

143. Mollersen, L., et al., *Neil1 is a genetic modifier of somatic and germline CAG trinucleotide repeat instability in R6/1 mice.* Hum Mol Genet, 2012. **21**(22): p. 4939-47.

144. Chakraborty, A., et al., *Neil2-null Mice Accumulate Oxidized DNA Bases in the Transcriptionally Active Sequences of the Genome and Are Susceptible to Innate Inflammation.* J Biol Chem, 2015. **290**(41): p. 24636-48.

145. Regnell, C.E., et al., *Hippocampal adult neurogenesis is maintained by Neil3-dependent repair of oxidative DNA lesions in neural progenitor cells.* Cell Rep, 2012. **2**(3): p. 503-10.

146. Sejersted, Y., et al., *Endonuclease VIII-like 3 (Neil3) DNA glycosylase promotes neurogenesis induced by hypoxia-ischemia.* Proc Natl Acad Sci U S A, 2011. **108**(46): p. 18802-7.

147. Frosina, G., *Overexpression of enzymes that repair endogenous damage to DNA.* Eur J Biochem, 2000. **267**(8): p. 2135-49.

148. Helleday, T., et al., *DNA repair pathways as targets for cancer therapy.* Nat Rev Cancer, 2008. **8**(3): p. 193-204.

149. Taricani, L., et al., *Phenotypic enhancement of thymidylate synthetase pathway inhibitors following ablation of Neil1 DNA glycosylase/lyase.* Cell Cycle, 2010. **9**(24): p. 4876-83.

150. Nijman, S.M., *Synthetic lethality: general principles, utility and detection using genetic screens in human cells.* FEBS Lett, 2011. **585**(1): p. 1-6.

151. Bridges, K.A., et al., *Niraparib (MK-4827), a novel poly(ADP-Ribose) polymerase inhibitor, radiosensitizes human lung and breast cancer cells.* Oncotarget, 2014. **5**(13): p. 5076-86.

152. Srinivasan, A. and B. Gold, *Small-molecule inhibitors of DNA damage-repair pathways: an approach to overcome tumor resistance to alkylating anticancer drugs.* Future Med Chem, 2012. **4**(9): p. 1093-111.

153. Jacobs, A.C., et al., *Inhibition of DNA glycosylases via small molecule purine analogs.* PLoS One, 2013. **8**(12): p. e81667.

154. Speina, E., et al., *Inhibition of DNA repair glycosylases by base analogs and tryptophan pyrolysate, Trp-P-1.* Acta Biochim Pol, 2005. **52**(1): p. 167-78.

155. Donley, N., et al., *Small Molecule Inhibitors of 8-Oxoguanine DNA Glycosylase-1 (OGG1).* ACS Chem Biol, 2015. **10**(10): p. 2334-43.

156. Thayer, M.M., et al., *Novel DNA binding motifs in the DNA repair enzyme endonuclease III crystal structure.* EMBO J, 1995. **14**(16): p. 4108-20.

157. Bjoras, M., et al., *Reciprocal "flipping" underlies substrate recognition and catalytic activation by the human 8-oxo-guanine DNA glycosylase.* J Mol Biol, 2002. **317**(2): p. 171-7.

158. Lingaraju, G.M., et al., *A DNA glycosylase from Pyrobaculum aerophilum with an 8-oxoguanine binding mode and a noncanonical helix-hairpin-helix structure.* Structure, 2005. **13**(1): p. 87-98.

159. Fromme, J.C. and G.L. Verdine, *Structure of a trapped endonuclease III-DNA covalent intermediate.* EMBO J, 2003. **22**(13): p. 3461-71.

160. Bruner, S.D., D.P. Norman, and G.L. Verdine, *Structural basis for recognition and repair of the endogenous mutagen 8-oxoguanine in DNA.* Nature, 2000. **403**(6772): p. 859-66.

161. Golan, G., et al., *Structure of the uncomplexed DNA repair enzyme endonuclease VIII indicates significant interdomain flexibility.* Nucleic Acids Res, 2005. **33**(15): p. 5006-16.

162. Gilboa, R., et al., *Structure of formamidopyrimidine-DNA glycosylase covalently complexed to DNA.* J Biol Chem, 2002. **277**(22): p. 19811-6.

163. Barrett, T.E., et al., *Crystal structure of a G:T/U mismatch-specific DNA glycosylase: mismatch recognition by complementary-strand interactions.* Cell, 1998. **92**(1): p. 117-29.

164. Parikh, S.S., et al., *Base excision repair initiation revealed by crystal structures and binding kinetics of human uracil-DNA glycosylase with DNA.* EMBO J, 1998. **17**(17): p. 5214-26.

165. Lau, A.Y., et al., *Crystal structure of a human alkylbase-DNA repair enzyme complexed to DNA: mechanisms for nucleotide flipping and base excision.* Cell, 1998. **95**(2): p. 249-58.

166. Dalhus, B., et al., *Structural insight into repair of alkylated DNA by a new superfamily of DNA glycosylases comprising HEAT-like repeats.* Nucleic Acids Res, 2007. **35**(7): p. 2451-9.

167. Nyaga, S.G. and R.S. Lloyd, *Two glycosylase/abasic lyases from Neisseria mucosa that initiate DNA repair at sites of UV-induced photoproducts.* J Biol Chem, 2000. **275**(31): p. 23569-76.

168. Golan, G., et al., *Structure of T4 pyrimidine dimer glycosylase in a reduced imine covalent complex with abasic site-containing DNA.* J Mol Biol, 2006. **362**(2): p. 241-58.

169. Schrock, R.D., 3rd and R.S. Lloyd, *Reductive methylation of the amino terminus of endonuclease V eradicates catalytic activities. Evidence for an essential role of the amino terminus in the chemical mechanisms of catalysis.* J Biol Chem, 1991. **266**(26): p. 17631-9.

170. Kahl, G., *HEAT repeat domain (Huntingtin, elongation factor 3 (EF3), protein phosphatase 2A (PP2A), yeast kinase TOR1 domain)*, in *The Dictionary of Genomics, Transcriptomics and Proteomics*. 2015, Wiley-VCH Verlag GmbH & Co. KGaA.

171. Rubinson, E.H., et al., *A new protein architecture for processing alkylation damaged DNA: the crystal structure of DNA glycosylase AlkD.* J Mol Biol, 2008. **381**(1): p. 13-23.

172. Wyatt, M.D., et al., *3-methyladenine DNA glycosylases: structure, function, and biological importance.* Bioessays, 1999. **21**(8): p. 668-76.

173. Pearl, L.H., *Structure and function in the uracil-DNA glycosylase superfamily.* Mutat Res, 2000. **460**(3-4): p. 165-81.

174. Denver, D.R., S.L. Swenson, and M. Lynch, *An evolutionary analysis of the helix-hairpin-helix superfamily of DNA repair glycosylases.* Mol Biol Evol, 2003. **20**(10): p. 1603-11.

175. Chetsanga, C.J. and T. Lindahl, *Release of 7-methylguanine residues whose imidazole rings have been opened from damaged DNA by a DNA glycosylase from Escherichia coli.* Nucleic Acids Res, 1979. **6**(11): p. 3673-84.

176. Castaing, B., et al., *Cleavage and binding of a DNA fragment containing a single 8-oxoguanine by wild type and mutant FPG proteins.* Nucleic Acids Res, 1993. **21**(12): p. 2899-905.

177. Tchou, J., et al., *Substrate specificity of Fpg protein. Recognition and cleavage of oxidatively damaged DNA.* J Biol Chem, 1994. **269**(21): p. 15318-24.

178. Jiang, D., et al., *Escherichia coli endonuclease VIII: cloning, sequencing, and overexpression of the nei structural gene and characterization of nei and nei nth mutants.* J Bacteriol, 1997. **179**(11): p. 3773-82.

179. Barrantes-Reynolds, R., S.S. Wallace, and J.P. Bond, *Using shifts in amino acid frequency and substitution rate to identify latent structural characters in base-excision repair enzymes.* PLoS One, 2011. **6**(10): p. e25246.

180. Zharkov, D.O., G. Shoham, and A.P. Grollman, *Structural characterization of the Fpg family of DNA glycosylases.* DNA Repair (Amst), 2003. **2**(8): p. 839-62.

181. Raoult, D., et al., *The 1.2-megabase genome sequence of Mimivirus.* Science, 2004. **306**(5700): p. 1344-50.

182. Imamura, K., et al., *Structural characterization of viral ortholog of human DNA glycosylase NEIL1 bound to thymine glycol or 5-hydroxyuracil-containing DNA.* J Biol Chem, 2012. **287**(6): p. 4288-98.

183. Fromme, J.C. and G.L. Verdine, *DNA lesion recognition by the bacterial repair enzyme MutM.* J Biol Chem, 2003. **278**(51): p. 51543-8.

184. Coste, F., et al., *Bacterial base excision repair enzyme Fpg recognizes bulky N7-substituted-FapydG lesion via unproductive binding mode.* Chem Biol, 2008. **15**(7): p. 706-17.

185. Duclos, S., et al., *Structural and biochemical studies of a plant formamidopyrimidine-DNA glycosylase reveal why eukaryotic Fpg glycosylases do not excise 8-oxoguanine.* DNA Repair (Amst), 2012. **11**(9): p. 714-25.

186. Sugahara, M., et al., *Crystal structure of a repair enzyme of oxidatively damaged DNA, MutM (Fpg), from an extreme thermophile, Thermus thermophilus HB8.* EMBO J, 2000. **19**(15): p. 3857-69.

187. Doublie, S., et al., *The crystal structure of human endonuclease VIII-like 1 (NEIL1) reveals a zincless finger motif required for glycosylase activity.* Proc Natl Acad Sci U S A, 2004. **101**(28): p. 10284-9.

188. Serre, L., et al., *Crystal structure of the Lactococcus lactis formamidopyrimidine-DNA glycosylase bound to an abasic site analogue-containing DNA.* EMBO J, 2002. **21**(12): p. 2854-65.

189. Zharkov, D.O., et al., *Structural analysis of an Escherichia coli endonuclease VIII covalent reaction intermediate.* EMBO J, 2002. **21**(4): p. 789-800.

190. Zhu, C., et al., *Tautomerization-dependent recognition and excision of oxidation damage in base-excision DNA repair.* Proc Natl Acad Sci U S A, 2016. **113**(28): p. 7792-7.

191. Liu, M., et al., *Structural characterization of a mouse ortholog of human NEIL3 with a marked preference for single-stranded DNA.* Structure, 2013. **21**(2): p. 247-56.

192. Zhou, X., et al., *OGG1 is essential in oxidative stress induced DNA demethylation.* Cell Signal, 2016. **28**(9): p. 1163-71.

193. Pani, B. and E. Nudler, *Mechanistic insights into transcription coupled DNA repair.* DNA Repair (Amst), 2017. **56**: p. 42-50.

194. Prakash, A., S. Doublie, and S.S. Wallace, *The Fpg/Nei family of DNA glycosylases: substrates, structures, and search for damage.* Prog Mol Biol Transl Sci, 2012. **110**: p. 71-91.

195. Pereira de Jesus, K., et al., *Structural insights into abasic site for Fpg specific binding and catalysis: comparative high-resolution crystallographic studies of Fpg bound to various models of abasic site analogues-containing DNA.* Nucleic Acids Res, 2005. **33**(18): p. 5936-44.

196. Biela, A., et al., *Zinc finger oxidation of Fpg/Nei DNA glycosylases by 2-thioxanthine: biochemical and X-ray structural characterization.* Nucleic Acids Res, 2014. **42**(16): p. 10748-61.

197. O'Connor, T.R., et al., *Fpg protein of Escherichia coli is a zinc finger protein whose cysteine residues have a structural and/or functional role.* J Biol Chem, 1993. **268**(12): p. 9063-70.

198. Tchou, J., et al., *Function of the zinc finger in Escherichia coli Fpg protein.* J Biol Chem, 1993. **268**(35): p. 26738-44.

199. Prakash, A., et al., *Structural investigation of a viral ortholog of human NEIL2/3 DNA glycosylases.* DNA Repair (Amst), 2013. **12**(12): p. 1062-71.

200. Klug, A. and J.W. Schwabe, *Protein motifs 5. Zinc fingers.* FASEB J, 1995. **9**(8): p. 597-604.

201. Coste, F., et al., *Structural basis for the recognition of the FapydG lesion (2,6-diamino-4-hydroxy-5-formamidopyrimidine) by formamidopyrimidine-DNA glycosylase.* J Biol Chem, 2004. **279**(42): p. 44074-83.

202. Sadeghian, K., et al., *Ribose-protonated DNA base excision repair: a combined theoretical and experimental study.* Angew Chem Int Ed Engl, 2014. **53**(38): p. 10044-8.

203. Qi, Y., et al., *Encounter and extrusion of an intrahelical lesion by a DNA repair enzyme.* Nature, 2009. **462**(7274): p. 762-U79.

204. Le Bihan, Y.V., et al., *5-Hydroxy-5-methylhydantoin DNA lesion, a molecular trap for DNA glycosylases.* Nucleic Acids Res, 2011. **39**(14): p. 6277-90.

205. Banerjee, A., W.L. Santos, and G.L. Verdine, *Structure of a DNA glycosylase searching for lesions.* Science, 2006. **311**(5764): p. 1153-7.

206. Fromme, J.C. and G.L. Verdine, *Structural insights into lesion recognition and repair by the bacterial 8-oxoguanine DNA glycosylase MutM.* Nat Struct Biol, 2002. **9**(7): p. 544-52.

207. Brooks, B.R., et al., *CHARMM: the biomolecular simulation program.* J Comput Chem, 2009. **30**(10): p. 1545-614.

208. Blank, I.D., K. Sadeghian, and C. Ochsenfeld, *A base-independent repair mechanism for DNA glycosylase--no discrimination within the active site.* Sci Rep, 2015. **5**: p. 10369.

209. Li, H., et al., *A dynamic checkpoint in oxidative lesion discrimination by formamidopyrimidine-DNA glycosylase.* Nucleic Acids Res, 2016. **44**(2): p. 683-94.

210. Amara, P. and L. Serre, *Functional flexibility of Bacillus stearothermophilus formamidopyrimidine DNA-glycosylase.* DNA Repair (Amst), 2006. **5**(8): p. 947-58.

211. Perlow-Poehnelt, R.A., et al., *Substrate discrimination by formamidopyrimidine-DNA glycosylase: distinguishing interactions within the active site.* Biochemistry, 2004. **43**(51): p. 16092-105.

212. Zaika, E.I., et al., *Substrate discrimination by formamidopyrimidine-DNA glycosylase: a mutational analysis.* J Biol Chem, 2004. **279**(6): p. 4849-61.

213. Jia, L., et al., *Lesion specificity in the base excision repair enzyme hNeil1: modeling and dynamics studies.* Biochemistry, 2007. **46**(18): p. 5305-14.

214. Roberts, E., et al., *MultiSeq: unifying sequence and structure data for evolutionary analysis.* BMC Bioinformatics, 2006. **7**: p. 382.

215. Humphrey, W., A. Dalke, and K. Schulten, *VMD: visual molecular dynamics.* J Mol Graph, 1996. **14**(1): p. 33-8, 27-8.

216. Sung, R.J., et al., *Structural and biochemical analysis of DNA helix invasion by the bacterial 8-oxoguanine DNA glycosylase MutM.* J Biol Chem, 2013. **288**(14): p. 10012-23.

217. Tchou, J. and A.P. Grollman, *The catalytic mechanism of Fpg protein. Evidence for a Schiff base intermediate and amino terminus localization of the catalytic site.* J Biol Chem, 1995. **270**(19): p. 11671-7.

218. Dianov, G.L., *Base excision repair targets for cancer therapy.* Am J Cancer Res, 2011. **1**(7): p. 845-51.

219. Goula, A.V. and K. Merienne, *Abnormal base excision repair at trinucleotide repeats associated with diseases: a tissue-selective mechanism.* Genes (Basel), 2013. **4**(3): p. 375-87.

220. Rose, P.W., et al., *The RCSB Protein Data Bank: views of structural biology for basic and applied research and education.* Nucleic Acids Res, 2015. **43**(Database issue): p. D345-56.

221. Prakash, A., et al., *Genome and cancer single nucleotide polymorphisms of the human NEIL1 DNA glycosylase: activity, structure, and the effect of editing.* DNA Repair (Amst), 2014. **14**: p. 17-26.

222. Ode, H., et al., *Molecular dynamics simulation in virus research.* Front Microbiol, 2012. **3**: p. 258.

223. Van Der Spoel, D., et al., *GROMACS: fast, flexible, and free.* J Comput Chem, 2005. **26**(16): p. 1701-18.

224. Cornell, W.D., et al., *A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules.* Journal of the American Chemical Society, 1995. **117**(19): p. 5179-5197.

225. Wang, J., et al., *Development and testing of a general amber force field.* J Comput Chem, 2004. **25**(9): p. 1157-74.

226. Fadda, E. and R.J. Woods, *Molecular simulations of carbohydrates and protein-carbohydrate interactions: motivation, issues and prospects.* Drug Discov Today, 2010. **15**(15-16): p. 596-609.

227. Jorgensen, W.L., Chandrasekhar, J., Madura, J. D., Impey, R. W., *Comparison of simple potential functions for simulating liquid water.* J. Chem. Phys., 1983. **79**: p. 926-935.

228. Case, D.A., Darden, T.A., Cheatham III, T.E., Simmerling, C.L. , Wang, J., Duke, R.E., Luo, R., Walker, R.C., Zhang, W., Merz, K.M., Roberts, Jr., B., Hayik, S., Roitberg, A., Seabra, G., Swails, J., Goetz, A.W., Kolossvary, I., Wong, K.F., aesani, F., Vanicek, J., Wolf, R.M., Liu, J., Wu, X., Brozell, S.R., Steinbrecher, T., Gohlke, H., Cai, Q., Ye, X., Wang, J., Hsieh, M.-J., Cui, G., Roe, D.R., Mathews, D.H., Seetin, M.G., Salomon-Ferrer, R., Sagui, C., Babin, V., Luchko, T., Gusarov, S., Kovalenko, A., Kollman P.A. , *The FF12SB force field.* AmberTools 13 Reference Manual, 2013: p. 27-29.

229. Batcho, P.F., D.A. Case, and T. Schlick, *Optimized particle-mesh Ewald/multiple-time step integration for molecular dynamics simulations.* The Journal of Chemical Physics, 2001. **115**(9): p. 4003-4018.

230. Ryckaert, J.-P., G. Ciccotti, and H.J.C. Berendsen, *Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes.* Journal of Computational Physics, 1977. **23**(3): p. 327-341.

231. Deift, P. and X. Zhou, *A Steepest Descent Method for Oscillatory Riemann--Hilbert Problems. Asymptotics for the MKdV Equation.* Annals of Mathematics, 1993. **137**(2): p. 295-368.

232. Hestenes, M., R.; Stiefel, E., *Methods of conjugate gradients for solving linear systems.* J. Res. Natl. Bur. Stand., 1952. **49**(6): p. 409.

233. Verlet, L., *Computer "'Experiments"' on Classical Fluids. I. Thermodynamical Properties of Lennard-Jones Molecules*. Vol. 159. 1967. 98.

234. Schlitter, J., M. Engels, and P. Kruger, *Targeted molecular dynamics: a new approach for searching pathways of conformational transitions.* J Mol Graph, 1994. **12**(2): p. 84-9.

235. Wolf, R.M., *Extracting ligands from receptors by reversed targeted molecular dynamics.* J Comput Aided Mol Des, 2015. **29**(11): p. 1025-34.

236. Shityakov, S. and C. Forster, *In silico predictive model to determine vector-mediated transport properties for the blood-brain barrier choline transporter.* Adv Appl Bioinform Chem, 2014. **7**: p. 23-36.

237. Wang, R., et al., *The PDBbind database: methodologies and updates.* J Med Chem, 2005. **48**(12): p. 4111-9.

238. Jain, A.N., *Scoring functions for protein-ligand docking.* Curr Protein Pept Sci, 2006. **7**(5): p. 407-20.

239. Pagadala, N.S., K. Syed, and J. Tuszynski, *Software for molecular docking: a review.* Biophys Rev, 2017. **9**(2): p. 91-102.

240. Morris, G.M., et al., *AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility.* J Comput Chem, 2009. **30**(16): p. 2785-91.

241. Trott, O. and A.J. Olson, *AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading.* J Comput Chem, 2010. **31**(2): p. 455-61.

242. Vasseur, R.*e.a.*, *AMIDE – Automatic Molecular Inverse Docking Engine for Large-Scale Protein Targets Identification.* International Journal on Advances in Life Sciences, 2014. **6**((3 & 4)): p. 325-337.

243. Gally, J.M., et al., *VSPrep: A General KNIME Workflow for the Preparation of Molecules for Virtual Screening.* Mol Inform, 2017.

244. Lipinski, C.A., et al., *Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings.* Adv Drug Deliv Rev, 2001. **46**(1-3): p. 3-26.

245. Lipinski, C.A., *Lead- and drug-like compounds: the rule-of-five revolution.* Drug Discov Today Technol, 2004. **1**(4): p. 337-41.

246. Berthold, M.R., Cebron, N., Dill, F., Gabriel, T. R., Kötter, T., Meinl, T., Ohl, P., Thiel, K., Wiswedel, B. , *KNIME - the Konstanz information miner: version 2.0 and beyond.* ACM SIGKDD Explorations Newsletter, 2009. **11**(1): p. 26-31.

247. *RDKit : Open-source chemoinformatics*. 2016; Available from: http://www.rdkit.org.

248. *ChemAxon - Software for Chemistry and Biology*. 2016; Available from: https://www.chemaxon.com/.

249. Veber, D.F., et al., *Molecular properties that influence the oral bioavailability of drug candidates.* J Med Chem, 2002. **45**(12): p. 2615-23.

250. Teague, S.J., et al., *The Design of Leadlike Combinatorial Libraries.* Angew Chem Int Ed Engl, 1999. **38**(24): p. 3743-3748.

251. Congreve, M., et al., *A 'rule of three' for fragment-based lead discovery?* Drug Discov Today, 2003. **8**(19): p. 876-7.

252. Metz, J.T., et al., *Navigating the kinome.* Nat Chem Biol, 2011. **7**(4): p. 200-2.

253. Gatica, E.A. and C.N. Cavasotto, *Ligand and decoy sets for docking to G protein-coupled receptors.* J Chem Inf Model, 2012. **52**(1): p. 1-6.

254. Baell, J.B. and G.A. Holloway, *New substructure filters for removal of pan assay interference compounds (PAINS) from screening libraries and for their exclusion in bioassays.* J Med Chem, 2010. **53**(7): p. 2719-40.

255. Garret M. Morris, D.S.G., Robert S. Halliday, Ruth Huey, William E. Hart, Richard. K. Belew, Arthur J. Olson, *Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function.* Journal of Computational Chemistry, 1998. **19**(14): p. 1639-1662.

256. Lee, A., K. Lee, and D. Kim, *Using reverse docking for target identification and its applications for drug discovery.* Expert Opin Drug Discov, 2016. **11**(7): p. 707-15.

257. Schmidtke, P., et al., *fpocket: online tools for protein ensemble pocket detection and tracking.* Nucleic Acids Res, 2010. **38**(Web Server issue): p. W582-9.

258. Metropolis, N. and S. Ulam, *The Monte Carlo method.* J Am Stat Assoc, 1949. **44**(247): p. 335-41.

259. Kirkpatrick, S., C.D. Gelatt, Jr., and M.P. Vecchi, *Optimization by simulated annealing.* Science, 1983. **220**(4598): p. 671-80.

260. Ledesma, S., *Computer and Information Science*. 2012.

261. Ravindranath, P.A., et al., *AutoDockFR: Advances in Protein-Ligand Docking with Explicitly Specified Binding Site Flexibility.* PLoS Comput Biol, 2015. **11**(12): p. e1004586.

262. Solis, F., J. & Wets, R., J-B., *Minimization by random search techniques.* Mathematics of Operation Research, 1981. **6**(1): p. 19-30.

263. Hetenyi, C. and D. van der Spoel, *Toward prediction of functional protein pockets using blind docking and pocket search algorithms.* Protein Sci, 2011. **20**(5): p. 880-93.

264. Eleftheriou, P., et al., *Prediction of enzyme inhibition and mode of inhibitory action based on calculation of distances between hydrogen bond donor/acceptor groups of the molecule and docking analysis: An application on the discovery of novel effective PTP1B inhibitors.* SAR QSAR Environ Res, 2015. **26**(7-9): p. 557-76.

265. Pedretti, A., L. Villa, and G. Vistoli, *Modeling of binding modes and inhibition mechanism of some natural ligands of farnesyl transferase using molecular docking.* J Med Chem, 2002. **45**(7): p. 1460-5.

266. Rester, U., *From virtuality to reality - Virtual screening in lead discovery and lead optimization: a medicinal chemistry perspective.* Curr Opin Drug Discov Devel, 2008. **11**(4): p. 559-68.

267. Song, K., et al., *Molecular mechanics parameters for the FapydG DNA lesion.* J Comput Chem, 2008. **29**(1): p. 17-23.

268. Wang, J., Wang, W., Kollman, P., Case, D., *Antechamber, An Accessory Software PackageFor Molecular Mechanical Calculation.* J. Comput. Chem., 2005. **25**: p. 1157-1174.

269. Glendening, E.D., Reed, A. E. , Carpenter, J. E. Weinhold F., *Gaussian 09*.

270. Vidal, A.E., et al., *Mechanism of stimulation of the DNA glycosylase activity of hOGG1 by the major human AP endonuclease: bypass of the AP lyase activity step.* Nucleic Acids Res, 2001. **29**(6): p. 1285-92.

271. Morland, I., et al., *Product inhibition and magnesium modulate the dual reaction mode of hOgg1.* DNA Repair (Amst), 2005. **4**(3): p. 381-7.

272. *Prime, Schrödinger, LLC, New York, NY.* 2017.

273. Poornima, C.S. and P.M. Dean, *Hydration in drug design. 1. Multiple hydrogen-bonding features of water molecules in mediating protein-ligand interactions.* J Comput Aided Mol Des, 1995. **9**(6): p. 500-12.

274. Spitzer, R. and A.N. Jain, *Surflex-Dock: Docking benchmarks and real-world application.* J Comput Aided Mol Des, 2012. **26**(6): p. 687-99.

275. *ChemBridge Corporation.* 2017.

276. *Ambinter, Greenpharma S.A.S.* 2017.

277. O'Boyle, N.M., et al., *Open Babel: An open chemical toolbox.* J Cheminform, 2011. **3**: p. 33.

278. *Indigo, LifeSciences unit of EPAM Systems, Inc.* 2014.

279. *Bingo, LifeSciences unit of EPAM Systems, Inc.* 2014.

280. *Molecular Operating Environment (MOE), 2015.08; Chemical Computing Group Inc. 1010 Sherbooke St. West, Suite #910, Montreal, QC, Canada, H3A 2R7.* 2017.

281. Sander, T., et al., *DataWarrior: an open-source program for chemistry aware data visualization and analysis.* J Chem Inf Model, 2015. **55**(2): p. 460-73.

282. Vasseur, R., *Développements HPC pour une nouvelle méthode de docking inverse : applications aux protéine matricielles*, in *Ecoles Doctorales URCA*. 2015, Université de Reims Champagne-Ardenne: Reims. p. 266.

Charlotte RIEUX

Étude des ADN glycosylases de la superfamille structurale Fpg/Nei par modélisation moléculaire, de nouvelles cibles thérapeutiques potentielles dans les stratégies anti-cancer

Résumé : L'ADN, support de l'information génétique, est constamment altéré par des agents physiques ou chimiques d'origines endogènes (métabolisme) et exogènes (UV, radiations ionisantes, produits chimiques) dont les effets sont génotoxiques. Ces modifications structurales délétères de l'ADN sont éliminées par de nombreux mécanismes de réparation. Parmi eux, le système de réparation par excision de bases (BER) est initié par les ADN glycosylases qui reconnaissent et éliminent les bases endommagées. Dans certaines stratégies anti-cancéreuses, l'utilisation de la chimiothérapie et la radiothérapie ont pour but la destruction des cellules cancéreuses en altérant leur ADN. Dans ce contexte, les ADN glycosylases réparent l'ADN des cellules traitées et induisent une résistance non désirée au traitement, faisant de ces enzymes des cibles thérapeutiques intéressantes. Le but de ces travaux est d'approfondir la compréhension des mécanismes de réparation des ADN glycosylases de la superfamille structurale Fpg/Nei grâce à la modélisation moléculaire et de pouvoir identifier et concevoir des inhibiteurs de ces enzymes. Les simulations de dynamique moléculaire (DM) nous ont permis d'étudier la « Lesion Capping Loop » (LCL) et de l'associer à la stabilisation de la base endommagée positionnée dans le site actif. Nous avons également étudié les chemins de sortie possibles de la base après coupure par l'enzyme et l'implication de la boucle LCL dans ce phénomène grâce à des simulations de DM ciblée (TMD^{-1}). De plus, les simulations de DM couplées à un protocole d'amarrage moléculaire « aveugle » nous ont permis d'identifier 2 sites de fixations possibles majoritaires pour des petites molécules potentiellement inhibitrices. Un de ces sites correspondant au site actif de hNEIL1 a fait l'objet d'un criblage virtuel d'une partie de la base de molécules Ambinter. Ceci nous a permis d'identifier des molécules potentiellement inhibitrices dont les effets seront prochainement testés *in vitro* dans l'équipe sur la protéine humaine hNeil1.

Mots-clés : Réparation de l'ADN, ADN glycosylase, Simulation de dynamique moléculaire, Amarrage moléculaire, Criblage virtuel

Study of DNA glycosylases from Fpg/Nei structural superfamilly by molecular modeling, new potential therapeutic target for anti-cancer strategies

Summary: The DNA, genetic information support, is frequently damaged by physical or chemical agents from endogenous (cell metabolism) and exogenous (UV, ionizing radiations, chemicals) factors whose effects are genotoxic. These deleterious DNA structural alterations are removed by many DNA repair mechanisms. Among them, the base excision repair (BER) is initiated by DNA glycosylases which recognize and remove damaged bases. In some anti-cancer strategies, the use of chemo- and radiotherapy is aimed to cancerous cells destruction by altering their DNA. In that specific context, DNA glycosylases repair the DNA of treated cells and induce unwanted resistance to treatments, making these enzymes interesting therapeutic targets. The purpose of this work is to deepen the repair mechanism knowledge of Fpg/Nei structural superfamily of DNA glycosylases using molecular modeling and designing inhibitors of these enzymes. Molecular dynamic simulations allowed us to study the « Lesion Capping Loop » (LCL) and to associate its role to substrate stabilization in the enzyme active site. We also studied some possible excision's product release pathways and LCL implication in this phenomena by targeted molecular dynamic simulations (TMD^{-1}). Furthermore, molecular dynamic simulations coupled to a blind molecular docking protocol allowed us to identify 2 possible main binding sites of potential inhibitiors. One of these binding sites corresponding to the hNEIL1 active site has been the object of a virtual screening of the Greenpharma database. This allowed us to identify potential inhibitors whom effects will be soon tested *in vitro* on the humain protein hNEIL1.

Keywords: DNA Repair, DNA Glycosylase, Molecular dynamic simulation, Molecular Docking, Virtual screening

Centre de Biophysique Moléculaire
rue Charles Sadron
45071 Orléans