UNIVERSITÀ DEGLI STUDI DI MILANO

DIPARTIMENTO DI CHIMICA

PhD COURSE IN CHEMISTRY - XXXIV CYCLE

STRUCTURE BASED DESIGN OF RSH INHIBITORS

CRESCENZO COPPA

Tutor: Prof. Sara Sattin
Co-tutor: Dr. Monica Civera

Coordinator: Prof. Emanuela Licandro

Academic Year 2020/2021

Contents

Acronyms

3D: three dimensional

ACT: Aspartate kinase, Chorismate mutase TyrA

AH: Alpha Helical

AMP-CPP: adenosine 5’-(α,β-methylene)triphosphate

ANT: anthranilic acid

AMP: Adenosine monophosphate

ATP: Adenosine triphosphate

Av: average

B. subtilis: *Bacillus Subtilis*

C. glutamicum: *Corynebacterium glutamicum*

CFU: colony-forming unit

CTD: C-terminal domain

dAMP: deoxy-adenosine monophosphate

DMSO: dimethyl sulfoxide

DNA: deoxyribonucleic acid

dt: time step

FBDD: fragment-based drug design

GDP: guanosine diphosphate

Glide: Grid-based Ligand Docking with Energetics

GltX: glutamyl-tRNA-synthetase

GTP: guanosine triphosphate

E. coli: *Escherichia coli*

E. faecalis: *Enterococcus faecalis*

EF: enrichment factor

EMA: European medicines agency

ERC: European research council

FDA: Food and Drug Administration

Ff: force field

Hpf: hibernation-promoting factor

HTS: high-throughput screening

HTVS: high-throughput virtual screening

Gscore: Glide score

H-bond: hydrogen bond

Hip: high persistence

HSS: High Solubility Subset

HYD: hydrolase

K_D: dissociation constant

LCPO: Linear Combinations of Pairwise Overlaps

LE: ligand efficiency

LJ: Lennard-Jones

LPS: lipopolysaccharide

M. smegmalis: *Mycobacterium smegmalis*

MM: molecular mechanics

MW: mass weight

NMR: nuclear magnetic resonance

NPT: isobaric-isothermal ensemble (constant number of atoms, pressure and temperature)

NTD: N-terminal domain

NVT: canonical ensemble (constant number of atoms, volume and temperature)

NuDiX: nucleoside diphosphate linked moiety X

M. tuberculosis: Mycobacterium tuberculosis

MD: molecular dynamics

PAINS: Pan-assay interference compounds

PBC: periodic boundary conditions

PBS: Phosphate-buffered saline

PC: principal component

PCA: principal component analysis

PDB: protein data bank

PME: particle mesh Ewald

pGpp: guanosine-5’monophosphate-3’-diphosphate

ppApp: adenosine-5’diphosphate-3’-diphosphate

ppGpp: guanosine-5’diphosphate-3’-diphosphate

ppG2’:3’p: guanosine-5'-diphosphate 2':3'-cyclic monophosphate

PPi: pyrophosphate

pppGpp: guanosine-5’triphosphate-3’-diphosphate

(p)ppGpp: guanosine penta-(pppGpp) or tetra-phosphate (ppGpp)

PVC: *Planctomycetes, Verrucomicrobia* and *Chlamydiae*

QM: quantum mechanics

QM/MM: quantum mechanics/ molecular mechanics

Rel_{Mtb} : Long RSH from *Mycobacterium tuberculosis*

Rel_{Seq}: Long RSH from *Streptococcus dysgalactiae subsp. equisimilis*

Rel_{Tt}: Long RSH from *Thermus termophilus*

RIS: Ribosome Inter subunit

Ro3: rule of three

RoG: radius of gyration

RSH: RelA/SpoT homolog

Rmf: ribosomal modulation factor

RMSD: root mean quare deviation

RMSF: root mean square fluctuation

RRM: RNA Recognition Motif

rRNA: ribosomial ribonucleic acid

RNA: ribonucleic acid

RNAP: ribonucleic acid polymerase

S. aureus: *Staphylococcus aureus*

SAR: Structure–activity relationship

SASA: solvent acesible surface area

SD: standard deviation

SID: simulation interaction diagram

SP: standard precision

SPR: surface plasmon resonance

SR: Stringent response

STD-NMR: saturation-transfer difference-nuclear magnetic resonance

SYNTH: synthetase

TGS: Threonyl-tRNA synthetase GTPase Spot

TNCG: truncated Newton conjugated gradient

TPSA: topological polar surface area

PSA: polar surface area

TS: termal shift

TSA: thermal shift assay

vdW: van der Waals

VS: virtual screening

XP: extra precision

Thesis Overview

Persistence is one of the biological mechanisms by which bacteria can avoid to be killed by antibiotic treatment. This phenotypic variant is characterized by a slowdown of cell metabolism that promotes bacteria dormant state. The molecular mechanisms leading to persisters formation have not been elucidated, yet. One of the mechanisms that was thought to be involved in the persister formation is the stringent response, but this connection was retracted. However, the first step of the stringent response is the accumulation of (p)ppGpp (guanosine tetra or pentaphosphate), alarmone synthesised by a family of enzymes called RelA/SpoT Homologue (RSH), that has pleiotropic effect on the cell including the formation of persister cells.0

This PhD thesis is part of a multidisciplinary research project (ERC-StG ERACHRON, grant n. 758108) whose aim is hampering persister formation by blocking the stringent response at its early stage, inhibiting the RSH proteins synthetase activity.

Specifically, the aim of this thesis was to identify, by *in silico* approaches, specific chemotypes able to interact with the synthetase active site of Rel_{Seq}, a RSH protein from *Streptococcus equisimilis* (Rel_{Seq}). Starting from the X-ray structure of Rel_{Seq}, virtual screening campaigns and molecular dynamics (MD) simulations were carried out. The identified chemotypes were then used to generate potential Rel_{Seq} ligands able to inhibit (p)ppGpp synthesis. *In silico* predictions and the activity of selected compounds were experimentally determined by thermal shift assays. Moreover, the role of GDP and Mn^{2+} in modulating the 3D conformation and the dynamic behaviour of Rel_{Seq} was also studied by means of molecular dynamics simulations.

The thesis is organized as follows:

- In chapter 1 a background of the topic (persistence) and an overview of the targets investigated in this thesis (RSH superfamily) are provided. The chapter is focussed on: i) the behaviour of persister cells compared to ‘wild type’ phenotype; ii) the accumulation of the alarmone guanosine tetra and pentaphosphate ((p)ppGpp), one of the possible reasons involved in the formation of persistent cells; iii) the 3D conformation and the role of Rel_{Seq} and some other RSH enzymes; iv) a description of some known structures of RSH family inhibitors.
- In chapter 2 the computational methods applied, in particular fragment based drug design (FBDD), molecular docking simulations, MD simulations, are briefly described.
- In chapter 3 the MD simulations of the X-ray complex structure are discussed. Systems with a different occupancy of the active sites were studied.

- In chapter 4 the results of the FBDD workflow were discussed. Seven fragment libraries were used and the most representative chemotypes were tested by thermal shift assays and MD simulations were carried out to assess binding stability. Staring from one fragment, a small library of ligands was designed and docked into the synthetase site, and tested in both thermal shift assays and MD simulations.
- In chapter 5 the selected fragments were also studied in the hydrolase site of of Rel_{Seq} by molecular docking calculation followed by MD simulations.
- In chapter 6 a general discussion of the results and the conclusions of this thesis are provided.

CHAPTER 1: INTRODUCTION

In this chapter I am going to provide a background about the topic and the targets investigated in this thesis: persistence and RelA/SpoT homolog (RSH) superfamily. I am focusing on i) the behaviour of persister cells compared to 'wild type' phenotype; ii) the accumulation of the alarmone guanosine tetra and pentaphosphate ((p)ppGpp), one of the possible reasons involved in the formation of persistent cells; iii) the 3D conformation and the role of my target (Rel$_{Seq}$) and some other RSH enzymes; iv) the structures of some inhibitors of the RSH family.

1.1 Background

In the last decades, antimicrobial resistance has become one of the most widespread threats to global health, after cancer and degenerative diseases. Bacterial resistance is caused by one or more genetic alterations that confer to bacterial cells the ability to survive antibiotic treatments. Unfortunately, the increasing number of bacteria resistant to antibiotics does not correspond to an increase in the number of new approved antibiotic compounds, that has been very low in the last thirty years. In fact, in the period between July 2017 and 2020 only eleven new antibiotics were approved by FDA or EMA or both, and only two of them are of a new class.[1] More important none of them hits a new target.[1] According to these data, a so low number of new antimicrobial drugs is not sufficient to obstruct the fast resistance mutation rate.

In this scenario, the discovery of new bacterial targets is of the outmost importance, with antibiotic persistence being an interesting phenomenon to be investigated. The formation of persisters, a phenotypic variant of bacterial cells (not connected to a genetic modification) that reverts their active state to a dormant one, is induced as survival response to several stress conditions, such as antibiotic treatments, nutrient starvation and even darkness.[2] Persisters play a role in recurring and chronic infections, such as in the case of cystic fibrosis,[3] candidiasis,[4] and tuberculosis.[5]

In the next sections, a more detailed description of the current knowledge about persistence is reported and discussed.

1.2 Persistence

Persistence is a phenotypic variation of the bacterial cell, that causes a metabolic slowdown and a temporary 'resilience' to the treatment, inactivating the replication, until the stress condition is over and the cell can potentially revert to the awake state. Persistence must not be confused with resistance, that is instead a genetic modification that deactivates the antibiotic action, leading to normal bacterial growth even in its presence (Figure 1).

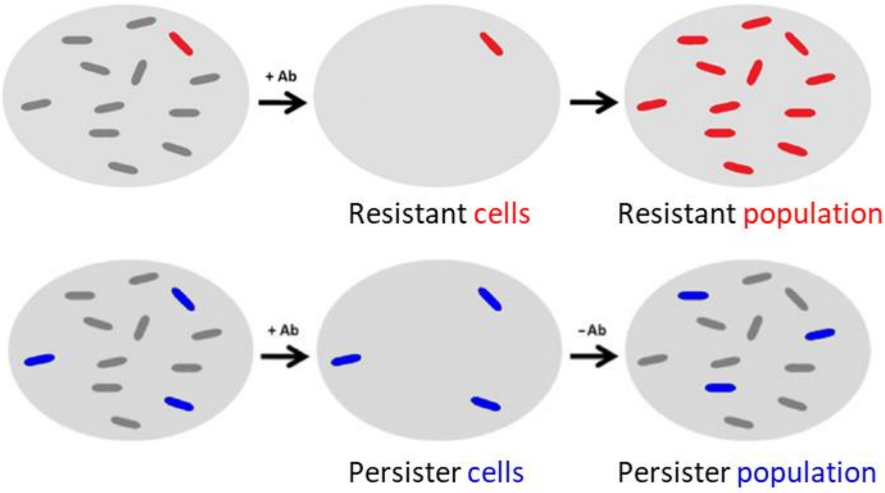


*Figure 1: **Resistance vs persistence.** In the top panel, a colony containing a resistant cell (red) is treated with antibiotic (Ab). The entire population is killed except for the resistant cell that replicates to form a new drug resistant population. In the bottom panel, a colony with persistent cells (blue). Once an antibiotic (Ab) is used only persisters survive, generating afterwards a population identical to the initial one, i.e. antibiotic sensitive. (figure modified from [4])*

Considering an isogenic bacterial population floating in their environment, namely the planktonic state, a little fraction of bacteria (10^{-6} – 10^{-5} Colony-forming unit, CFU) adopts the persistence state. This bacterial population can be killed by an appropriate antibiotic dosage that leads to a biphasic killing curve, typical of bacterial populations with persistent cells (Figure 2).[6] Following this trend, the most of bacterial cells die within the first couple of hours, while persisters survive until they awake. When they resurge, if there are favourable conditions (i.e. absence of the stress condition that induced persistence state) persisters can form a new population identical to the previous one, so antibiotic sensitive, or, if the antibiotic is still present, they start dying.[7]

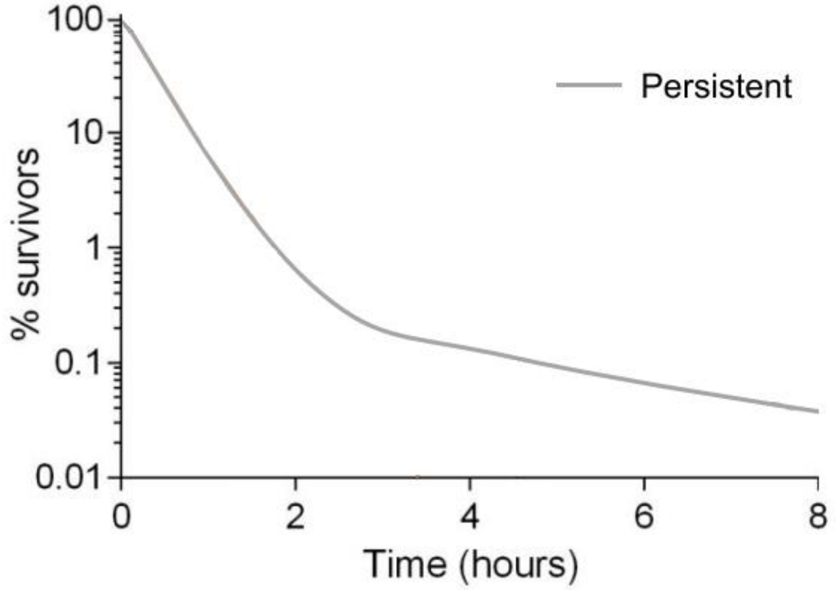


*Figure 2: **Killing-rate profile of a bacteria population with persister cells**. The graph illustrates the percentage of bacterial cells surviving antibiotic treatment (y-axis) vs time (hours; x-axis). The 'sensitive population' (the most of the colony) is killed within the first 2 hours. If the antibiotic is present in low concentration the persister population awakes over time and dies due to the antibiotic presence. A killing plateau is visible as persister cells remain viable (modified from [6])*

The first time this phenomenon was observed was in 1944 by Joseph W. Bigger who reported that a little fraction of a *Staphylococcus aureus* population well tolerated antibiotic treatments even in absence of genetic modification that could cause resistance to the antibiotic itself.[8] He called this singularity 'persistence' due to the ability of these cells to 'persist' over antibiotic treatment. Unfortunately, the technological limitations of the time stopped Bigger from further investigating the problem and the persistence was so long forgotten due to the increasing interest in the most worrying resistance problem. Only recently, several research groups all over the world started studying the phenotypic variation. In 1983 Moyed H. and colleagues[9] identified a gene that if mutated increased the formation of persister cells in *E. coli*. They isolated *Hip* (high persistence) mutant cells that showed an increased persisters ratio (10^{-2} CFU Hip mutant vs $10^{-6}/10^{-5}$ CFU in wild type) with an unchanged susceptibility to antibiotics after resurgence. It took twenty years to characterize the *hipA7* as the allele conferring this higher persistence frequency.[10]

Once single-cell techniques became available, new aspects of the persister cells were evaluated. The use of microfluid devices helped the Balaban's group[11] to investigate the persistence occurrence in *hipA7* single *E. coli* cells. The study focused on the growth of individual bacteria under normal conditions and antibiotic treatments, analysing the history of the survivors. Balaban and colleagues were able to distinguish persister cells, even before ampicillin treatment, due to the reduced growth rate. This work elucidated the link between persisters and the inherent heterogeneity of growth rate in the bacterial population. Hip mutation was observed in bacterial cells of Cystic fibrosis (CF) patients. In the study performed in 2006 by Smith and collaborators[12] on 35 longitudinal clinical isolates of single CF patients (from 8 to 96 months) they demonstrated that persisters formation of *P. aeruginosa* increased by 100-fold due to the *hip* gene mutation.

Despite the different ways they tolerate antibiotic treatments, persister and resistant cells are somehow connected: if persister cells regrow in an environment where the antibiotic is still present in a concentration that cannot kill them fast, genetic modification can occur provoking the formation of resistant cells. This event can happen due to the increasing production of DNA repair proteins connected to the SOS and stress responses that awake persister cells.[13] In particular, the expression of error-prone polymerase, induced by SOS signalling, increases the chance of mutation during persister infections, producing mutated and resistant bacteria. Moreover, it was demonstrated that persister cells prolong the antibiotic treatments time promoting resistance mutation.[14]

Bacteria have developed this ability to ensure survival under selective pressure. Therefore, it can be seen as a defensive mechanism implemented by several microorganisms against environmental changes. Unfortunately, the mechanism of persisters formation is still unclear. Moreover, it is also unclear precisely how the dormant persister phenotype can revert to the “awake” state. Elucidating both mechanisms could help the development of new therapeutic treatments.

1.3 Persisters Formation

The dormant state of a persister cell appears to be a hibernation state where no biochemical activity takes place. However, as reported in a work performed on *Mycobacterium smegmatis,*[15] the cell might replicate, albeit very slowly, even during this dormant state.

There are different working hypothesis for the molecular mechanisms at the base of persisters formation. The downstream of the stringent response, i.e. the redundant toxin-antitoxin systems, was first postulated as responsible of the overall metabolic slowdown, only to be retracted few years later due to phage contamination in several of the reported studies.[16] The other main school of thought linked the insurgence of the persister phenotype with a sudden drop of ATP levels within the cell.[17,18] Our research group decided to focus on the upstream of the stringent response as described hereafter.

1.3.1 Stringent Response

The stringent response (SR) is a signalling cascade that activates within the bacterial cell in response to stressful conditions such as nutrient starvation, darkness,[19] temperature shift, osmotic shock,[20] pH downshift[21] and oxygen variation[22] or sub-lethal doses of antibiotic treatment.[23]

The cascade is triggered by the accumulation of guanosine tetra- and pentaphosphate collectively called (p)ppGpp (a.k.a. alarmone or 'magic spot'[24]) and ends with the activation of the toxin-antitoxin pathways.

As already said the linkage between this cascade and the formation of persister cells was retracted. [16] However the accumulation of (p)ppGpp can induce tolerance and persistence in bacterial cells.[2]

1.3.2 (p)ppGpp

The enzymes involved in the synthesis, and hydrolysis, of (p)ppGpp belong to the RelA/SpoT homolog (RSH) superfamily (see below) and the process is summarized in Figure 3.

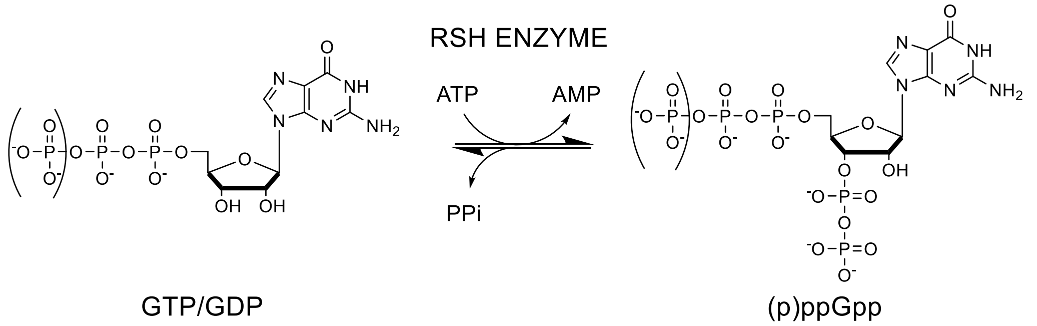


Figure 3: Schematic representation of (p)ppGpp synthesis and hydrolysis mediated by RSH proteins

(p)ppGpp affects a wide variety of cellular processes with pleiotropic effects on the cell (Figure 4). The alarmone (p)ppGpp exerts a very important role in controlling the energetic metabolism of bacterial cells[25,26] and in virulence and immune evasion.[25] It behaves in different ways in different organisms. In *E. coli* (p)ppGpp activates RpoS and Rpoe that respond to stress for misfolded proteins in the periplasm, expressing or silencing about 500 genes.[27] Another of its roles in *E. coli* is the inhibition of DNA primase[28] and it could inhibit rRNA synthesis, affecting translation in general, by regulating the transcription of Rmf (ribosomal modulation factor).[29] (p)ppGpp can also induce persistence via Hpf (hibernation-promoting factor) and Rmf that inactivate ribosomes converting 90s ribosomes into 100s ribosomes in *E. coli.*[30] Rmf and Hpf overexpression not only increases persister cell formation but also avoids cell resuscitation.[2] Furthermore, the connection between these alarmones and the induction of persistence *via* ribosomes dimerization was highlighted by Wood and colleagues that demonstrated the ability of (p)ppGpp to inhibit the ribosome-associated GTPase Era, one of the protein involved in the late assembly process of ribosome 30S subunit, causing problems in ribosomal assemblation and cell growth arrest in S. *aureus.*[31,32] Several other works performed on *B. subtilis* showed the role of ribosome dimerization in persister formation.[33,34] In gram-positive bacteria, (p)ppGpp role is to bind RNA polymerase, provoking a signal to the site of the catalytic Mg^{2+}, which change the gene expression profile.[2] All together these changes induce metabolic slowdown and reduce cellular growth.[2]

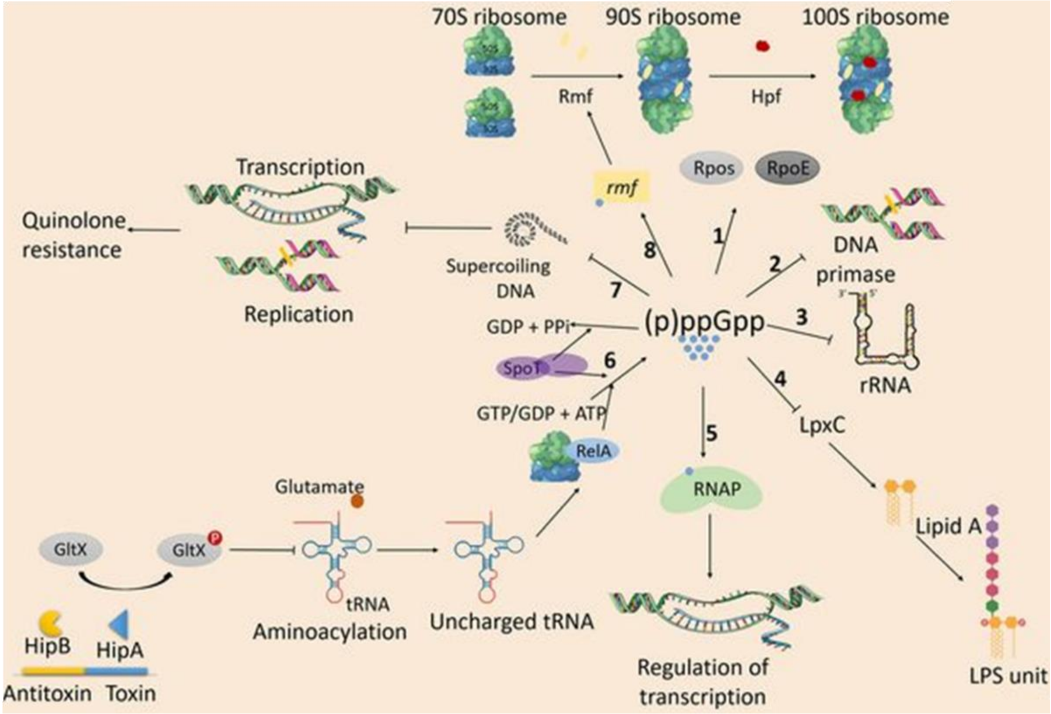


Figure 4: ***P*****athways regulated by (p)ppGpp in *E. coli.*** **1**, accumulation of (p)ppGpp induces transcription of RpoS and RpoE; **2**, DNA primase is inhibited by (p)ppGpp and thus chromosome replication; **3**, (p)ppGpp affects general translation inhibiting transcription of rRNA; **4**, (p)ppGpp affects LpxC regulation, decreasing lipopolysaccharide (LPS) formation; **5**, (p)ppGpp affects transcription regulation of many genes by binding to RNA polymerase (RNAP), regulating the transcription of many genes; **6**, RSH enzymes are triggered by empty tRNA to synthesise (p)ppGpp. Empty tRNAs are generated by phosphorylation and inactivation of glutamyl-tRNA-synthetase (GltX) induced by HipA toxin; **7**, resistance to quinolones is induced by (p)ppGpp by inhibiting supercoiling of DNA in *E. coli*; **8**, (p)ppGpp also affects ribosome dimerization by inducing the transcription of the ribosomal modulation factor (Rmf) and hibernation-promoting factor (Hpf). (p)ppGpp is also involved in human pathogenesis, virulence and immune evasion (modified from [2])

1.3.2 pGpp

Even though its presence in amino acid starved bacteria was already verified in the second half of 1970s,[35,36] the characterization and the role of guanosine-5'monophosphate-3'-diphosphate (pGpp) as a possible third alarmone involved in stringent response have been studied only recently. Its synthesis is due to the presence of guanosine-monophosphate (GMP) instead of guanosine-diphosphate (GDP) or guanosine-triphosphate (GTP) in the synthetic site of long RSH of some bacteria such as *M. smegmalis,*[37] *S. aureus,*[38] *C. glutamicum*[39] and *E. coli.*[40] The formation of pGpp can also be caused by the enzymatic degradation of pppGpp and ppGpp by a hydrolase enzyme called 'NahA' from the NuDiX (nucleoside diphosphate linked moiety X) hydrolase family (Figure 5).[41] More enzymes of the NuDiX family from *E. coli*[42,43] and *T. thermophilis*[44] can form pGpp and even degrade it into pGp. pGpp, as well as ppGpp and pppGpp, inhibits enzymes involved in the GTP synthetase pathway and the transcription of the 16s rRNA gene rrnB in *E. faecalis* and *E. coli* respectively.[45]

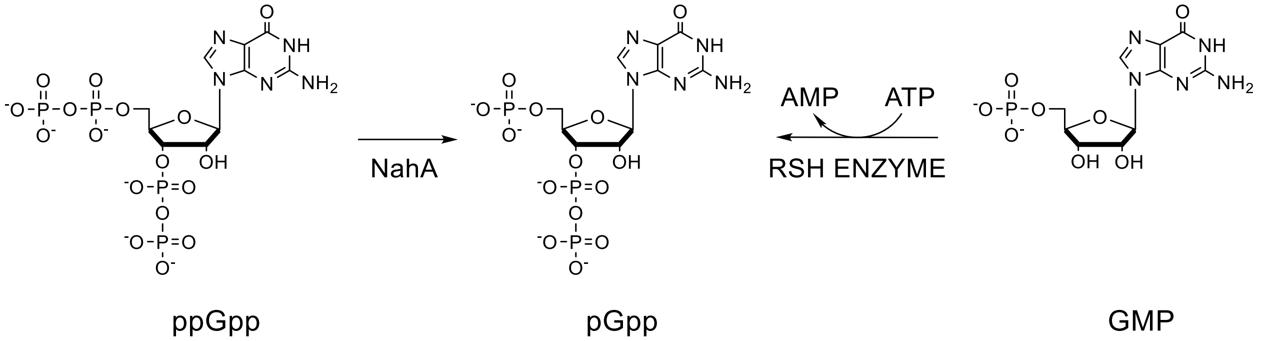


Figure 5: pGpp formation. From the left hydrolysis of ppGpp by NahA enzyme from the right synthesis of pGpp by RSH enzyme using GMP and ATP

The most interesting aspect in targeting accumulation of (p)ppGpp is the absence of RSH enzymes able to synthesise these alarmones in the mammalian cells[2] so if a selective inhibitor is found the risk of side and off-target effects in human patients, or more in general in mammalians, should be sensibly lower than in other therapeutic strategies.

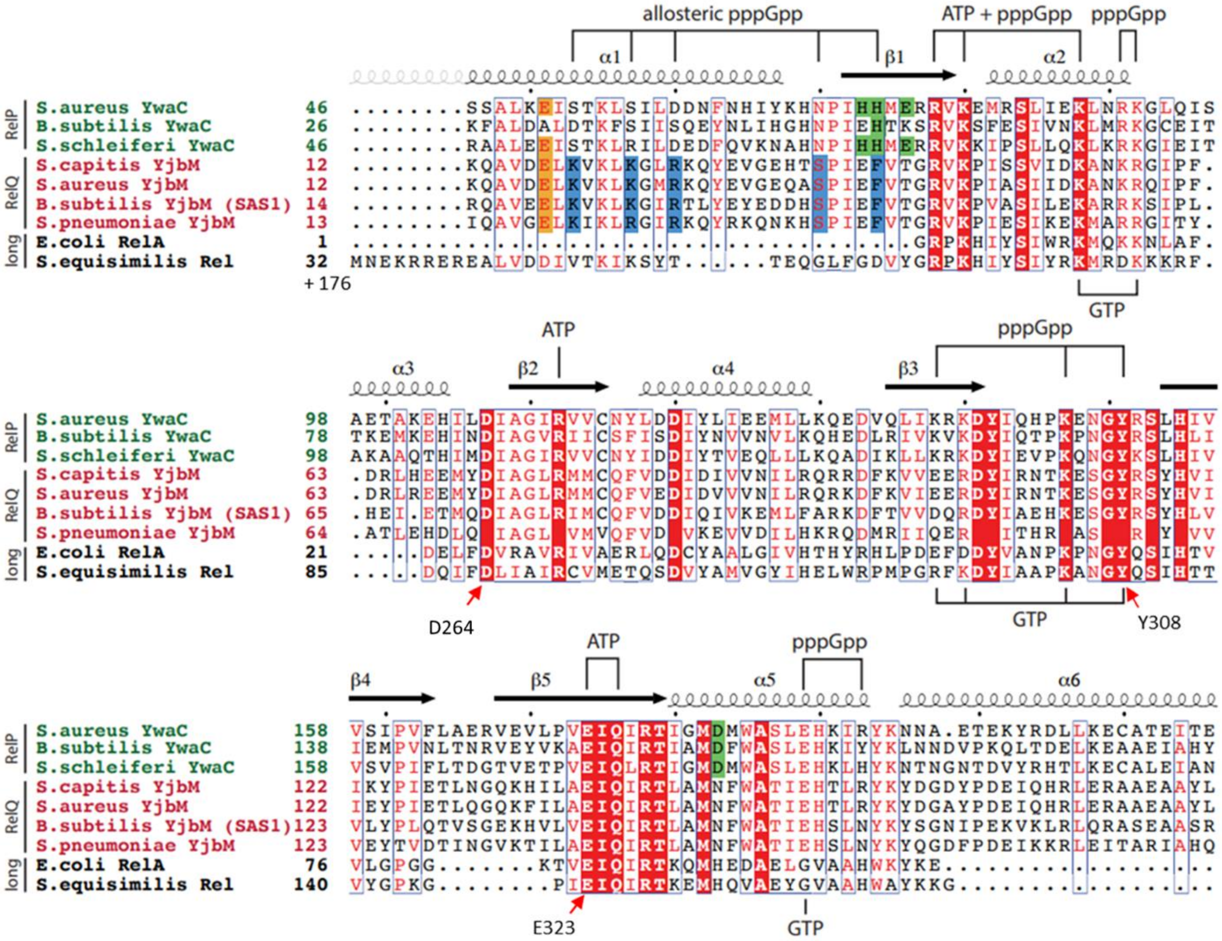
1.4 RSH superfamily

RSHs constitute a superfamily of enzymes able to synthesise and/or hydrolyse (p)ppGpp, where RelA and SpoT are the two Rel enzymes first discovered in *E. coli.*[46]

(p)ppGpp is synthesised by the transfer of a pyrophosphate group (PPi) from ATP to the 3'-OH group of a molecule of GDP or GTP, yielding ppGpp or pppGpp, respectively, in a Mg^{2+}-dependent fashion. The hydrolysis instead converts (p)ppGpp into GDP (from ppGpp) or GTP (from pppGpp) and a molecule of PPi (Figure 3).

These enzymes are widespread in the bacterial kingdom, with at least one form of the protein present in each specie with some exceptions, i.e. bacteria belonging to the PVC (*Planctomycetes, Verrucomicrobia* and *Chlamydiae*) superphylum, and bacteria that proliferate in stable microenvironments.[23] According to the phylogenetic classification, the RSH superfamily is organized into three subclasses:[46] long RSHs, Small Alarmone Hydrolases (SAHs) and Small Alarmone Synthetases (SASs).

The three subclasses share residues that are conserved within species (Figure 6).



allosteric pppGpp
ATP + pppGpp
pppGpp
α1
β1
α2
RelP
RelQ
long
S.aureus YwaC 46 ........SSALKEISTKLSILDDNFNHIYKHNPIHHMERRVKEMRSLIEKLNRKGLQIS
B.subtilis YwaC 26 ........KFALDALDTKFSIISQEYNLIHGHNPIEHTKSRVKSFESIVNKLMRKGCEIT
S.schleiferi YwaC 46 ........RAALEEISTKLRILDEDFQVKNAHNPIHHMERRVKKIPSLLQKLKRKGIEIT
S.capitis YjbM 12 ........KQAVDELKVKLKGLRKQYEVGEHTSPIEFVTGRVKPISSVIDKANKRGIPF.
S.aureus YjbM 12 ........KQAVDELKVKLKGMRKQYEVGEQASPIEFVTGRVKPIASIIDKANKRQIPF.
B.subtilis YjbM (SAS1) 14 ........RQAVEELKVKLKGIRTLYEYEDDHSPIEFVTGRVKPVASILEKARRKSIPL.
S.pneumoniae YjbM 13 ........IQAVGELKIKLRGIRKQYRKQNKHSPIEFVTGRVKPIESIKEKMARRGITY.
E.coli RelA 1 ....GRPKHIYSIWRKMQKKNLAF.
S.equisimilis Rel 32 MNEKRREREALVDDIVTKIKSYT......TEQGLFGDVYGRPKHIYSIYRKMRDKKRF.
+176
GTP
ATP
pppGpp
α3
β2
α4
β3
S.aureus YwaC 98 AETAKEHILDIAGIRVVCNYLDDIYLIEEMLLKQEDVQLIKRKDYIQHPKENGYRSLHIV
B.subtilis YwaC 78 TKEMKEHINDIAGVRIICSFISDIYNVVNVLKQHEDLRIVKVKDYIQTPKPNGYRSLHLI
S.schleiferi YwaC 98 AKAAQTHIMDIAGIRVVCNYIDDIYTVEQLLLKQADIKLLKRKDYIEVPKQNGYKSLHIV
S.capitis YjbM 63 .DRLHEEMYDIAGLRMMCQFVDDIDIVVNILRQRRDFKVVEERDYIRNTKESGYRSYHVI
S.aureus YjbM 63 .DRLREEMYDIAGLRMMCQFVEDIDVVVNILRQRKDFKVIEERDYIRNTKESGYRSYHVI
B.subtilis YjbM (SAS1) 65 .HEI.ETMQDIAGLRIMCQFVDDIQIVKEMLFARKDFTVVDQRDYIAEHKESGYRSYHLV
S.pneumoniae YjbM 64 .ATLEHDLQ IAGL VMVQFVD VKEVVDILHKRQDMRIIQER ITHR AS R Y VV
E.coli RelA 21 .....DELFDVRAVRIVAERLQDCYAALGIVHTHYRHLPDEFDDYVANPKPNGYQSIHTV
S.equisimilis Rel 85 .....DQIFDLIAIRCVMETQSDVYAMVGYIHELWRPMPGRFKDYIAAPKANGYQSIHTT
D264
GTP
Y308
ATP
pppGpp
β4
β5
α5
α6
S.aureus YwaC 158 VSIPVFLAERVEVLPVEIQIRTIGMDMWASLEHKIRYKNNA.ETEKYRDLLKECATEITE
B.subtilis YwaC 138 IEMPVNLTNRVEYVKAEIQIRTIAMDFWASLEHKIYYKLNNDVPKQLTDELKEAAEIAHY
S.schleiferi YwaC 158 VSVPIFLTDGTVETPVEIQLRTIGMDMWASLEHKLHYKNTNGNTDVYRHTLKECALEIAN
S.capitis YjbM 122 IKYPIETLNGQKHILAEIQIRTLAMNFWATIEHTLRYKYDGDYPDEIQHRLERAAEAAYL
S.aureus YjbM 122 IEYPIETLQGQKFILAEIQIRTLAMNFWATIEHTLRYKYDGAYPDEIQHRLERAAEAAYL
B.subtilis YjbM (SAS1) 123 VLYPLQTVSGEKHVLVEIQIRTLAMNFWATIEHSLNYKYSGNIPEKVKLRLQRASEAASR
S.pneumoniae YjbM 123 VEYTVDTINGVKTILAEIQIRTLAMNFWATIEHSLNYKYQGDFPDEIKKRLEITARIAHQ
E.coli RelA 76 VLGPGG.......KTVEIQIRTKQMHEDAELGVAAHWKYKE...................
S.equisimilis Rel 140 VYGPKG........PIEIQIRTKEMHQVAEYGVAAHWAYKKG..................
E323
GTP

Figure 6: Sequence alignment of RSH synthetase domains of different bacteria. Residue highlighted in red are conserved within species while letter in red represent those amino acids whose mutation do not change polarity or charge. Secondary structure of RelP is shown in curvy lines (elixes) and arrow (β-sheet). In the figure residues involved in ligand binding are marked with black lines (figure modified from [47]).

1.4.1 Long RSHs

From a structural perspective, long RSH enzymes are composed of two macrodomains: the N-terminal domain (NTD) or enzymatic domain, consisting of both the Hydrolase (HYD) and Synthetase (SYNTH) domains, and the C-terminal domain (CTD), also called regulatory domain, comprising several other subdomains, namely: Threonyl-tRNA synthetase GTPase Spot (TGS), Alpha Helical (AH), Ribosome Inter subunit (RIS) and Aspartate kinase, Chorismate mutase TyrA (ACT) domains (Figure 7).[48] Among these, ACT is the part of the protein able to bind to the ribosome and for this reason it is also called RNA Recognition Motif (RRM).[49]

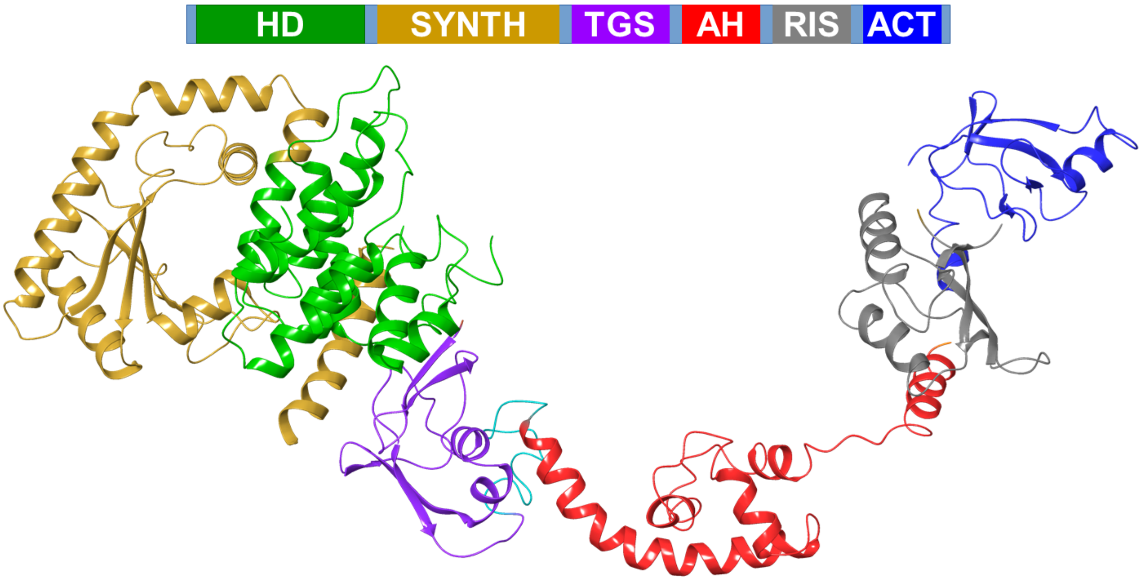


*Figure 7: **3D representation of the long RSH RelA from E. coli.** In the top panel the schematic representation of domains that compose RelA is shown; the bottom panel shows the 3D structure of RelA (PDB entry: 5KPX) rendered as ribbons coloured according to the same color code reported in the top panel.*

Despite the presence of both HYD and SYNTH domains, some long RSHs can present one of the two domains catalytically inactive. For example, in RelA from *E. coli* (Figure 7) and in the long RSHs from gammaproteobacteria and betaproteobacteria the HYD domain is inactive. This feature leads to a pseudo-hydrolase domain, that is structurally and evolutionarily conserved across these bacteria species. However, the loss of the hydrolase functionality is counterbalanced by the presence of a second long RSH that maintains (p)ppGpp level controlled.[46,50,51] SYNTH and HYD domains work in harmony, by alternating their activation, to regulate (p)ppGpp concentration depending on environmental conditions. The switch between the two domains, i.e. the activation of the HYD and the inactivation of the SYNTH or *vice versa*, is controlled by substrate interaction[52] and/or by the binding between the regulatory domain and cellular or nuclear components, such as ribosomes.[53–55] Tamman and collegues[52] demonstrated that in *Thermus Thermophilus* when GDP and ATP enter in the synthetase domain, forming a specific interaction network within the binding pocket, the enzyme takes on an “open” conformation able to inhibit hydrolysis. On the other hand, when (p)ppGpp binds to the HYD site, the enzyme rearranges in a “closed” conformation activating hydrolysis and occluding the SYNTH pocket. Similar opened and closed conformations can be seen when long RSHs bind to or do not bind to the ribosome, respectively,[56,57] and also in this case the closed conformation promotes (p)ppGpp hydrolysis while the binding to the ribosome activates its synthesis.

Substrate preference or specificity (i.e. for GDP or GTP) of these enzymes seems to lie in two conserved motifs: EXDD and RXKD (where X is any amino acid).[40] It seems that the more acidic

motif (EXDD) is more suitable for GDP, therefore, RSH proteins with EXDD motif, such as RelA from *E. coli*, prefer this substrate instead of GTP. GTP results to be more suitable for the second motif, that has been found for example in the long RSH from *M. tuberculosis* (Rel$_{\mathrm{Mtb}}$).[40]
As described above, long RSHs can respond to different stimuli: SpoT responds to fatty acid starvation,[58,59] while RelA is activated by stalled ribosomes.[60,61]

1.4.1.1 Rel$_{Seq}$: the target

We chose Rel$_{Seq}$, the long RSH from *Streptococcus dysgalactiae subsp. equisimilis*, as our reference structure, since it has been the first long RSH for which an X-ray structure was made available (PDB entry 1VJ7).[62] The entire enzyme sequence consists of 739 amino acids (UniProtKB - Q54089), but the crystals were obtained from a truncated version of it (residues 1-385) where the C-terminal regulatory domain was absent.
From a structural point of view, the N-terminal HYD domain (residues 1-159, green in Figure 8a) is connected to the SYNTH domain (residues 176-385, yellow in Figure 8a) by a central 3 helix bundle (residues 135-195, red in Figure 8a). Two different conformations with posited opposite catalytic activities were detected, namely Chain A and B. Chain A was assumed to have a Synthetase ON / Hydrolase OFF conformation, while the opposite was hypothesized for Chain B (Synthetase OFF / Hydrolase ON). Indeed, both SYNTH sites bear the GDP substrate molecule and, while both HYD domains feature a Mn^{2+} ion, only the HYD domain of Chain B shows the peculiar guanosine-5'-diphosphate 2':3'-cyclic monophosphate (ppG2':3'p), probably a hydrolysis reaction byproduct or intermediate.
The catalytic sites are more than 30 Å apart and require two different metal cofactors: while the already mentioned Mn^{2+} ion in the HYD domain seems stably bound, the Mg^{2+} ion required for (p)ppGpp synthesis was not detected in the crystal structure. In addition, it has been postulated that binding of the ligand(s) in one of the catalytic sites induces conformational changes in the protein, probably reducing or inactivating the other domain functionality.[62] Indeed, it was reported that in the presence of ppGpp, ATP, GDP and both ions, Rel$_{Seq}$ is able to both synthesise and hydrolyse the alarmone, in a so called 'futile cycle' (Figure 8b), where ATP is consumed while ppGpp and GDP concentrations remain constant.[63]

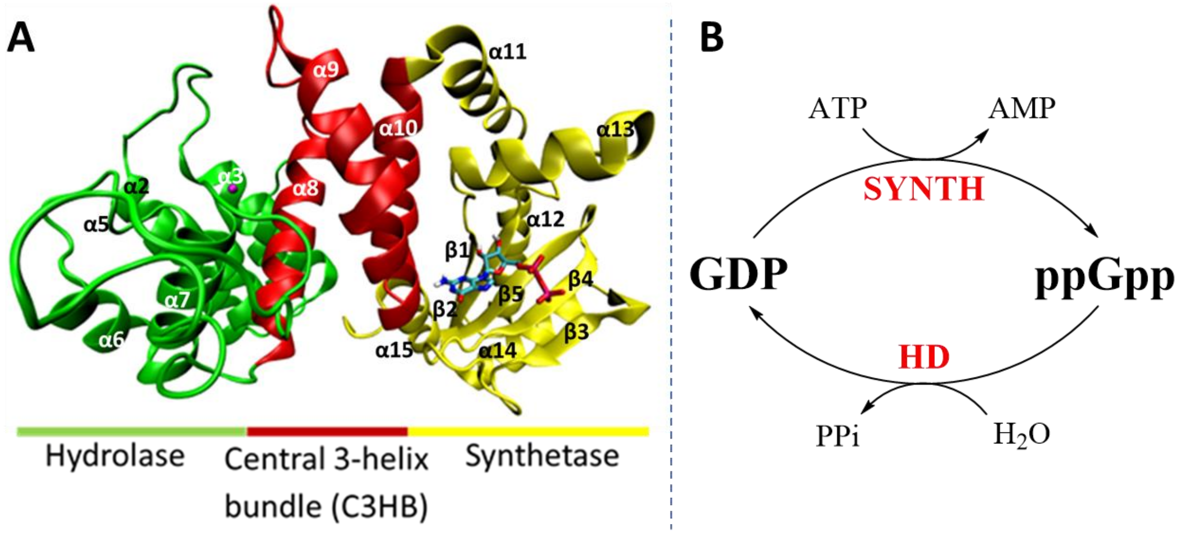


Figure 8: A) 3D representation of Rel_{Seq} NTD portion. HYD domain is shown in green, 3-helix bundle is shown in red, SYNTH domain is shown in yellow, Mn^{2+} in the HYD site is shown as a purple ball and GDP in the SYNTH site is shown as sticks. B) schematic representation of the futile cycle

Mutagenesis studies performed by Hogg and collaborators[62] on the truncated version of the enzyme crystallized already presented, showed that single point mutations of residues belonging to the HYD or SYNTH domains can only influence the domain activity they belong to. Figure 9 shows mutations that affect hydrolytic and synthetic activity of Rel_{Seq}. In detail, single point mutations shown above and below the primary structure sequence inhibit (p)ppGpp hydrolysis or synthesis, respectively.

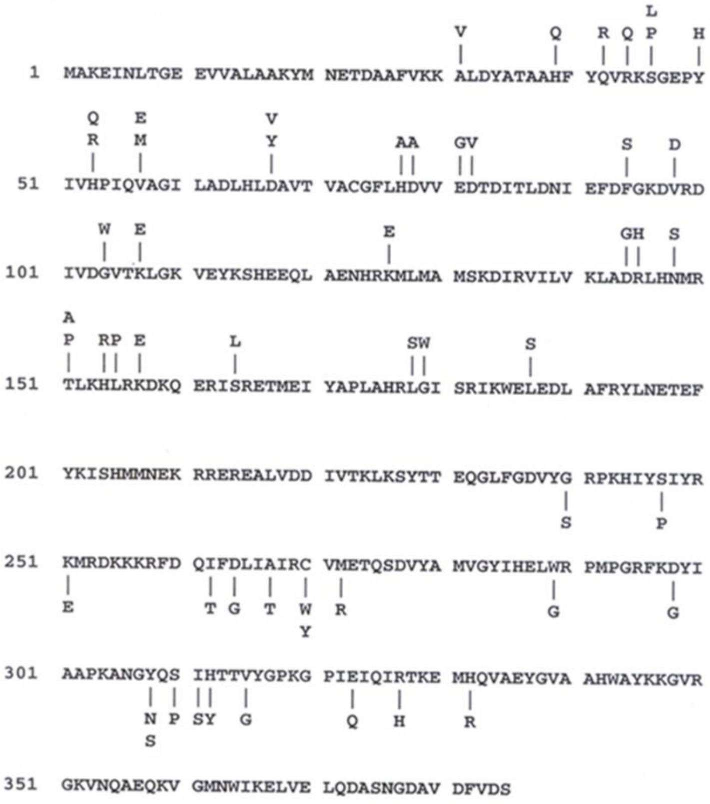

V Q R Q L P H
1 MAKEINLTGE EVVALAAKYM NETDAAFVKK ALDYATAAHF YQVRKSGEPY
Q R E M V Y AA GV S D
51 IVHPIQVAGI LADLHLDAVT VACGFLHDVV EDTDITLDNI EFDFGKDVRD
W E E GH S
101 IVDGVTKLGK VEYKSHEEQL AENHRKMLMA MSKDIRVILV KLADRLHNMR
A P RP E L SW S
151 TLKHLRKDKQ ERISRETMEI YAPLAHRLGI SRIKWELEDL AFRYLNETEF
201 YKISHMMNEK RREREALVDD IVTKLKSYTT EQGLFGDVYG RPKHIYSIYR
S P
251 KMRDKKKRFD QIFDLIAIRC VMETQSDVYA MVGYIHELWR PMPGRFKDYI
E T G T W Y R G G
301 AAPKANGYQS IHTTVYGPKG PIEIQIRTKE MHQVAEYGVA AHWAYKKGVR
N S P SY G Q H R
351 GKVNQAEQKV GMNWIKELVE LQDASNGDAV DFVDS

Figure 9: Single point mutation that affect enzymatic activity of Rel$_{Seq}$. In detail mutations marked above the sequence reduce or inhibit hydrolysis activity while mutations shown below the sequence reduce synthesis activity (figure modified from [62])

Comparing the mutagenesis studies shown in Figure 9 and the X-ray GDP binding mode (Figure 10), described in detail in the next paragraph, the only residues involved in this binding that cause synthetic activity inhibition if mutated are Y308 and H312. Therefore, these residues are crucial for substrate binding and synthetic activity.

1.4.1.1.1 Rel$_{Seq}$ synthetase site

The SYNTH domain of Rel$_{Seq}$, such as the SYNTH domain of other RSH enzyme,[52,47,64,65] is composed of five antiparallel β-sheets (β1- β5) surrounded by α helices (α12-α15 Figure 8).

In the X-ray structure, GDP binds to the protein G-loop (Y299-S310) forming a π-π stacking interaction between the guanine ring and the side chain of Y308, further stabilized by hydrogen bonds (H-bonds) between guanine N7 and the K304 side chain (Nζ), guanine N2 and the backbone (C=O) of A335 and guanine O6 and the side chain of N306 (Nδ2). GDP β-phosphate group forms salt bridge with K297 side chain (Nζ) and H-bonds with Y299 side chain (OH) and H312 side chains (Nδ) (Figure 10).

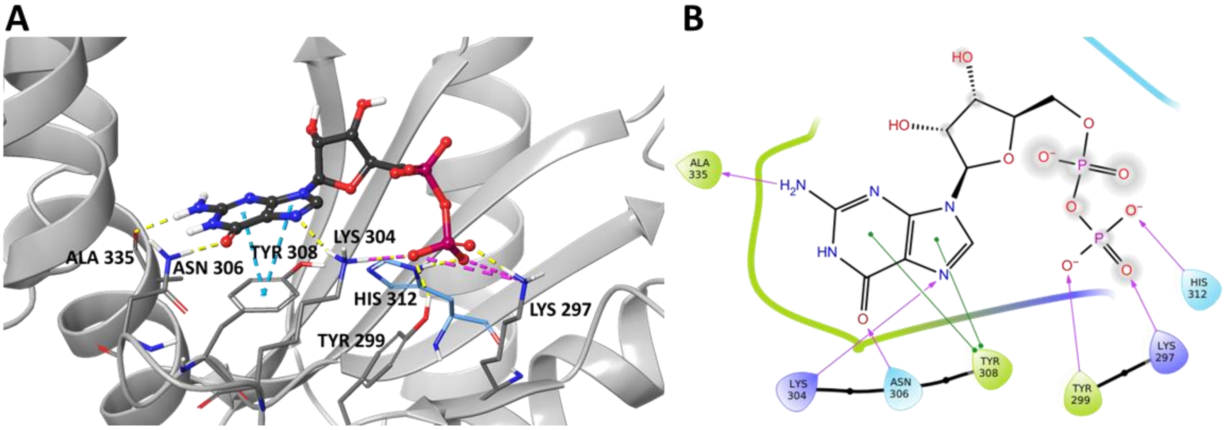


Figure 10: A) Interaction of GDP (ball and stick in dark grey carbons) in the X-ray structure of Rel$_{\text{Seq}}$ *(1VJ7.pdb, chain A): residues involved in the interaction are represented in sticks and the protein is represented in ribbons. B) 2D Ligand interaction diagram of GDP into Rel*$_{\text{Seq}}$ *X-ray complex. Green lines represent π-π stacking, purple arrows represent H-bonds, blue to red lines represent salt bridges*

Analysing Rel$_{Seq}$ SYNTH catalytic site, it is possible to notice how it lacks the space needed to accommodate the pyrophosphate donor ATP, and how the supposedly catalytic residues D264 and E323 are not correctly oriented to promote the reaction. We therefore deduced that the synthetase ON conformation is indeed not fully switched ON.

Only in recent years the X-ray structure of a SAS from *S. aureus*, RelP, was made available in a true pre-catalytic state (PDB entry 6EWZ),[47] RelP catalytic site shows its substrate GTP, a non-hydrolysable ATP analogue, AMP-CPP (adenosine 5'-(α,β-methylene)triphosphate), and the required Magnesium (Figure 11).[47] When a superposition of the SYNTH domains of Rel$_{Seq}$ and RelP is performed, a conformational change of helix 2 of RelP, which is rotated compared to the corresponding alpha 13 of Rel$_{Seq}$, can be seen (Figure 11b). The rotation frees enough space to accommodate the ATP molecule. Moreover, the conformational change and the presence of the Mg^{2+} allow for the correct orientation of RelP catalytic residues D107 (coordinating the Mg^{2+} ion) and E174 differently from their analogues in Rel$_{Seq}$ (D264 and E323) which are in an out position, non-catalytically correct to bind to the Mg^{2+} (Figure 11c). Therefore, the correct behaviour of these conserved residues in Rel$_{Seq}$ (Figure 6), essential for the synthetic activity (Figure 9), cannot be described by using the X-ray structure. A work performed by our group,[66] in which a chimera was generated by the homology modelling of Rel$_{Seq}$ combining its HYD domain 3D conformation and RelP SYNTH domain 3D conformation, can be instead a good starting point to better study the behaviour of these two amino acids, such as the behaviour of a real synthetase ON Rel$_{Seq}$.

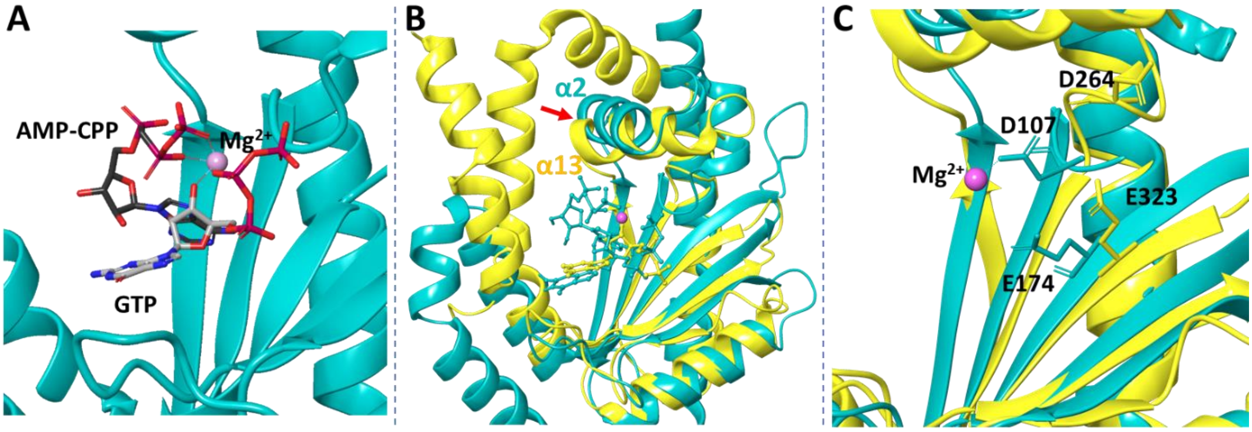


Figure 11: A) 3D representation of the SYNTH site of RelP (cyan ribbons) binding GTP (white carbon) and a modified not hydrolysable ATP (AMP-CPP in dimgrey carbons) in the presence of Mg^{2+} (pink ball) (PDB entry: 6EWZ) B) Rel_{Seq} (yellow ribbon) RelP (cyan ribbon) superposition. Helix 2 (red arrow) is rotated in RelP to better accommodate ligands into the pocket. GTP and AMP-CPP are represented in cyan ball and stick, GDP is represented in yellow ball and stick, magnesium is represented in magenta ball C) comparison of catalytical residues involved in Mg^{2+} coordination into the SYNTH site of Rel_{Seq} (yellow) and RelP (cyan). Mg^{2+}, from RelP crystal structure, is represented in magenta ball.

1.4.1.1.2 Rel_{Seq} hydrolase site

The HYD domain of Rel_{Seq} is composed of six α helices and its structure is similar to the human phosphodiesterase 4B2B as claimed by Hogg and collegues.[62] The putative hydrolase ON conformation of the crystalized Rel_{Seq} (Chain B PDB entry 1VJ7), shows a HYD domain 3D conformation very similar to the one of the long RSH from *Thermus termophilus* (Rel_{Tt}) obtained in 2020 (PDB entry 6S2T),[52] despite the absence of the natural (p)ppGpp ligand. Comparing instead the HYD domains of these enzymes to the pseudo-hydrolase domain of RelA, the main difference is presented by helix 6, helix 7 and the loop connecting these two helices. In fact, helices 6 and 7 are partially disordered in Rel_{Seq} and Rel_{Tt} to help the ligand to enter into the pocket while they are more rigid in RelA with the loop connecting the two helices that completely block the HYD site.[52]

The binding mode of ppG2':3'p in the HYD site of Rel_{Seq} chain B is shown in Figure 12. The guanine ring is sandwiched between the side chains of K45 and L155 on one side, and the side chains of N148 and R44 on the other side forming a π-cationic interaction with the guanidinium group. The ring is further stabilized by H-bonds formed with K45 backbone (N7), N146 side chain (NH2) and T151 side chain (N1). The 5' phosphate groups form salt bridges with side chains of K45 and K159 while 2':3'-cyclic phosphate coordinates Mn^{2+} via a water bridge. R44, S46, N148 and T151 that stabilize the ligand into the pocket, are conserved residues (Figure 6) that if mutated induce inhibition of the hydrolytic activity of the enzyme (Figure 9).[62]

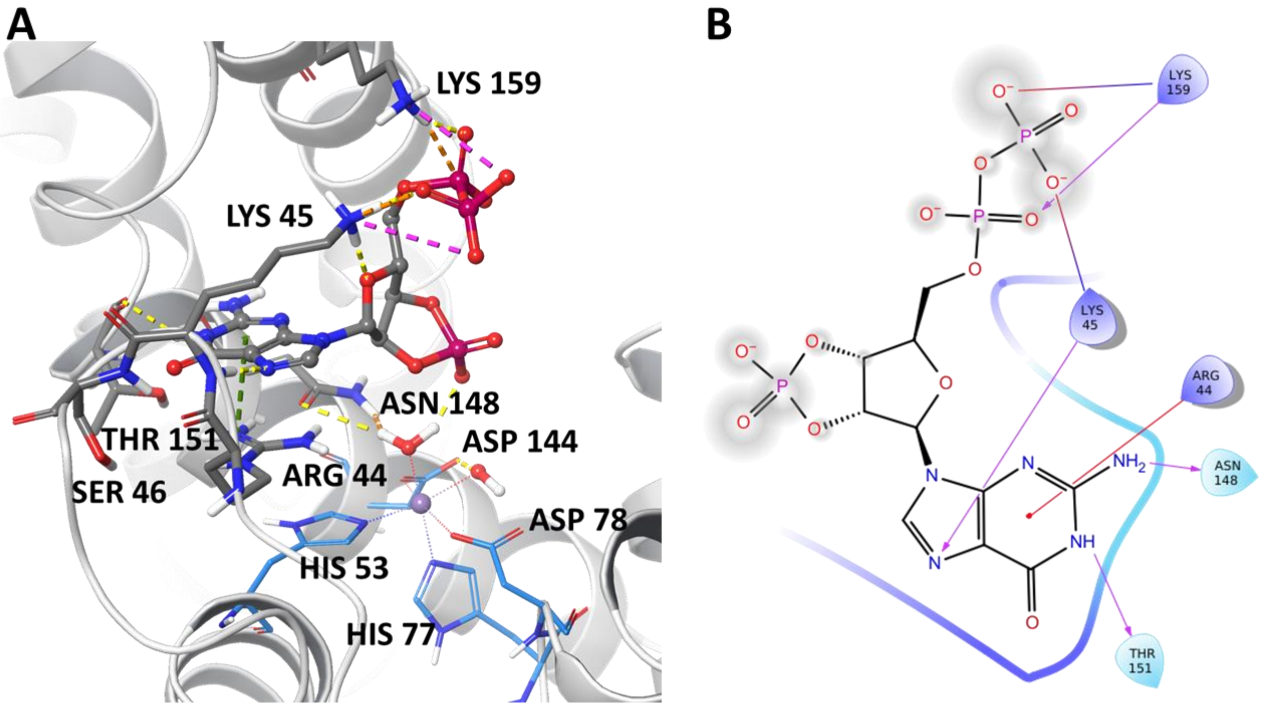


Figure 12: A) ppG2':3'p interactions in HYD site of Rel$_{Seq}$ crystallographic structure (chain B). ppG2':3'p is represented in ball and stick, the amino acids involved in the ppG2':3'p interaction are shown in grey sticks, the amino acids involved in Mn^{2+} coordination are shown in cyan sticks and the Mn^{2+} is shown in purple sphere B) 2D Ligand interaction diagram of ppG2':3'p::Rel$_{Seq}$ complex. The red line represents the π-cationic interaction, purple arrows represent H-bonds, blue and red lines represent salt bridges

Compared to the X-ray structure of Rel$_{Tt}$ where ppGpp is cocrystalized within the enzyme (Figure 13), it is possible to see that residues involved in the guanine ring and 5' pyrophosphate group binding mode are conserved within the two enzymes (only the numbering is different due to an indel mutation) while the 3' pyrophosphate group of ppGpp directly coordinates manganese without water bridges. Moreover, the close contact between the Mn^{2+} and an oxygen atom of the ppGpp 3' α-phosphate group in this crystal structure, confirms the role of the metal ion in the activation of phosphorous centre for a nucleophilic attack. The most important difference between the active pockets of Rel$_{Seq}$ and Rel$_{Tt}$, apart from the ligands, consists in the presence of a second metal (Mg^{2+}) into the HYD pocket of Rel$_{Tt}$ whose role was not clarified. Furthermore, no water molecules coordinate the ion in presence of ppGpp in Rel$_{Tt}$.

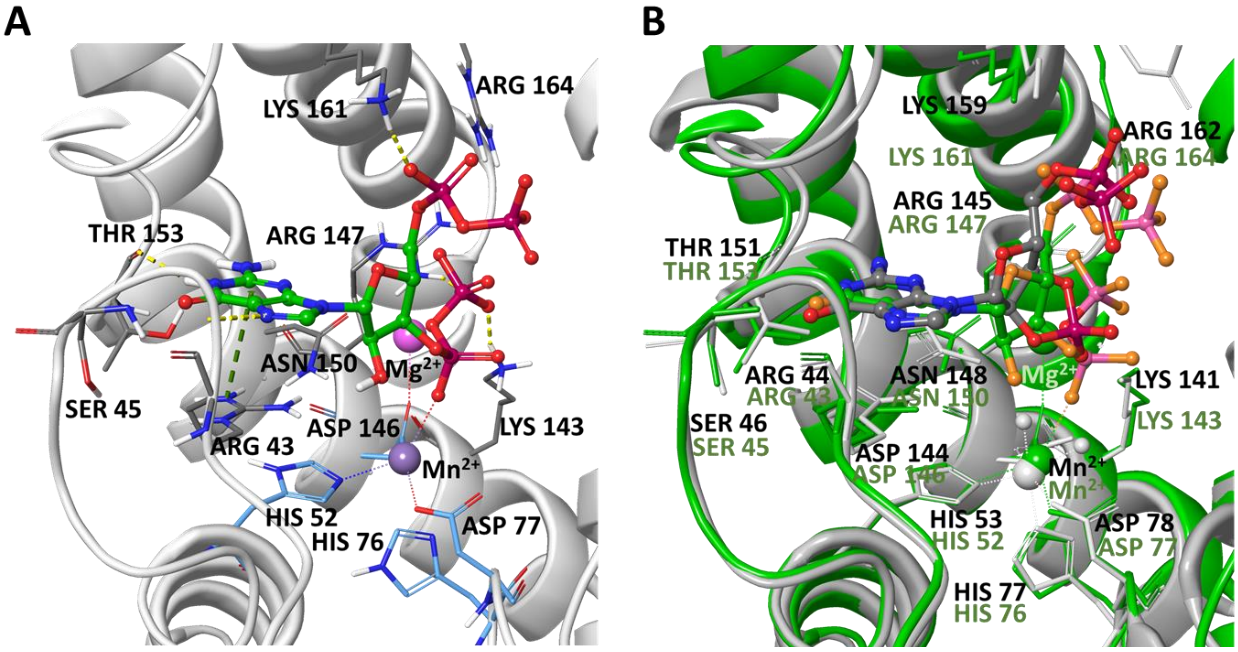


Figure 13: A) ppGpp interaction in HYD site of the long RSH from Thermus thermophilus. ppGpp is shown in green carbons ball and sticks, Mn^{2+} is shown in violet sphere, Mg^{2+} is rendered as magenta sphere. B) Superposition of the two HYD domains of Rel$_{Seq}$ (chain B in white) and RSH from Thermus thermophilus (green) crystallographic structures (RMSD 1.11 Å). ppG2':3'p in Rel$_{Seq}$ is shown in grey carbon balls and sticks, ppGpp in Rel$_{Tt}$ is shown in green carbons ball and sticks, Rel$_{Seq}$ residues are shown in white sticks and black letter, Rel$_{Tt}$ residues are shown in green sticks and letter

The binding mode shown in Rel$_{Tt}$ structure (Figure 13a) suggests how the natural ligand should bind to the pocket but does not give any information about how the hydrolase reaction should work. Thus, a plausible RSH hydrolysis-based mechanism of action can be hypothesized starting from the one reported by Zimmerman and collegues.[67] Specifically, the proposed mechanism involves the 5′-deoxyribonucleotidase YfbR from *E. coli* whose structure (PDB entry 2PAQ and 2PAU)[67] is shown in Figure 14a. The hydrolysis mechanism of Rel$_{Seq}$ was hypothesized according to the structural superposition of the chain B of Rel$_{Seq}$ and YfbR (Figure 14b) and considering the residues involved in YfbR hydrolysis reaction (Figure 14d). Residue D82 in Rel$_{Seq}$ replaces residue D72 in YfbR while E81 should deprotonate the water molecule coordinated to the Mn^{2+} and involved in the nucleophilic attack. This mechanism can be considered plausible not only because of the 3D superposition of residues involved in the reaction (with both YfbR and Rel$_{Tt}$), but also because of the mutagenesis studies,[62] i.e., mutations of E81 and D82 in Gly and Val, respectively, induce inhibition of the hydrolysis activity.

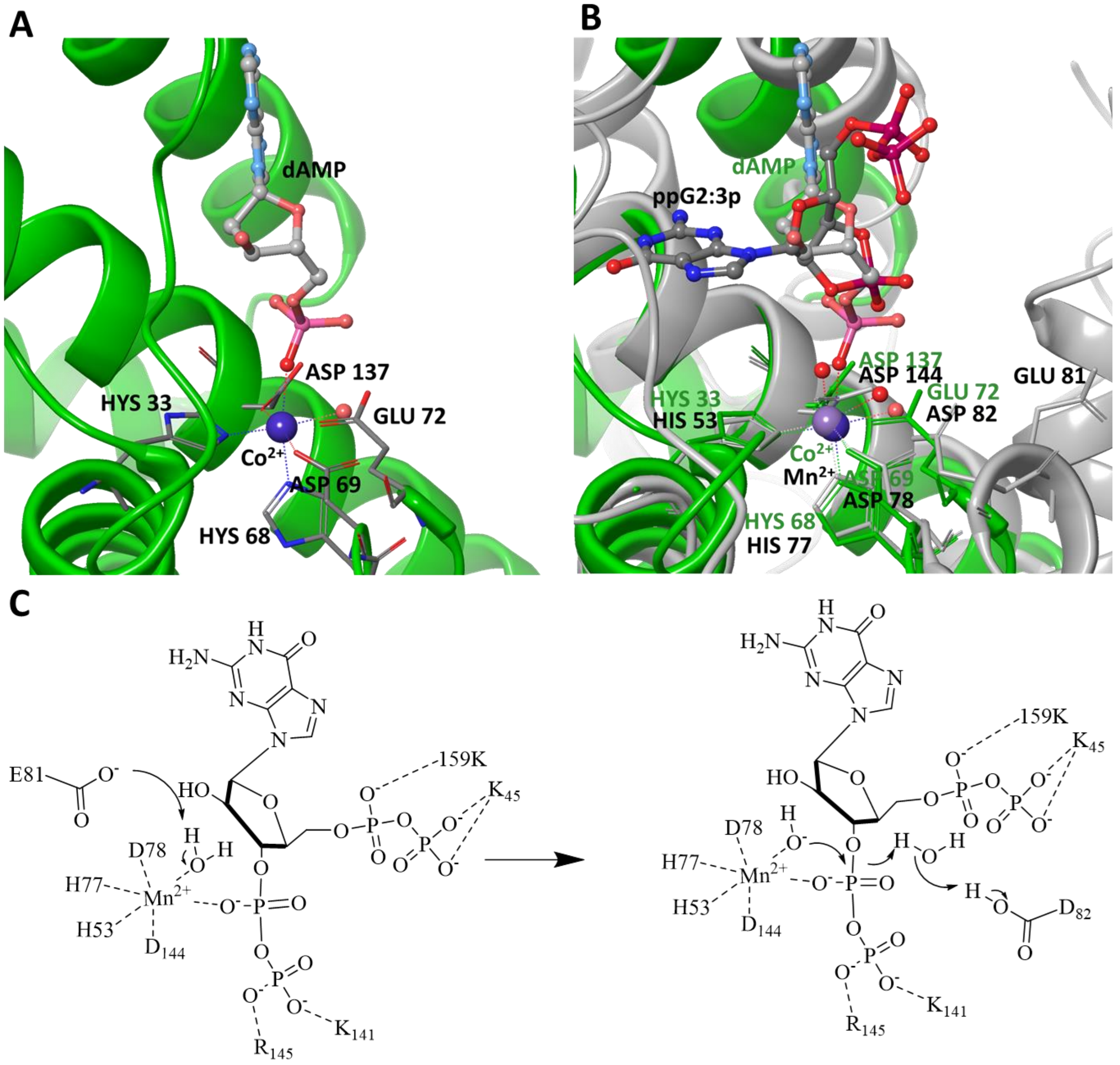


Figure 14: A) dAMP into the active site of 5'-deoxyribonucleotidase YfbR from Escherichia coli. dAMP is represented in ball and stick the residues involved in the coordination to the Co^{2+} (blue ball) are represented in lines and the protein is represented in green ribbon. B) Superposition of Rel_{Seq} chain B (white) and 5'-deoxyribonucleotidase YfbR from Escherichia coli (green). Superposition was performed using the residues involved in metal coordination and the metals themselves (RMSD 0.42 Å). C) Hypothesized mechanism of hydrolysis of Rel_{Seq}.

1.4.2 SAS

Small alarmone synthetase (SAS) enzymes contain only the SYNTH domain.[46] The most famous and better characterized enzymes of this class are the one called RelP (where one of the most studied is the one from *S. aureus* discussed above) and RelQ, that show a sequence identity of about 30%.[23] Enzymes belonging to these two classes do not present any regulatory domain and for this reason, their expression is regulated at transcriptional level, where ethanol stress, alkaline shock and cell-wall targeting antibiotics exposure can positively modulate the transcription.[68–71] Considering the structures of RelP in *Staphilococcus aureus* and RelQ in *Bacillus subtilis* as examples of these two categories of SAS, the two enzymes are very similar in both tertiary structure (Figure 15) and in their ability of homotetramerize[72] (in Figure 15d RelP homotetramer is shown). The main difference of the two enzymes consists in the production of (p)ppGpp. Actually RelP synthesises more alarmone than RelQ due to the more rigid G-loop (loop where are located amino acids involved in GDP/GTP binding) that facilitate GDP coordination.[72] Another interesting difference between the two enzymes is the presence of a (p)ppGpp allosteric recognition motif in RelQ that lacks in RelP.[72]

Of the other enzymes belonging to SAS class, the most studied have been RelV[73] from Vibrio cholerae, RelS from *Corynebacterium glutamicum*[39] and RelZ (MS_RHII-RSD) from *Mycobacterium smegmatis*.[37] In particular, RelZ is one of the most interesting enzyme belonging to this family due to its double ability of synthesising (p)ppGpp, like for the other SAS, and repairing the 'R-loops' which are RNA-DNA hybrid able to interfere with DNA repair, replication and transcription, thus compromising genome integrity and function.[74–76] Actually, this enzyme presents the SYNTH domain fused to a RnaseHIIs domain (no 3D structure has been provided).

RelP from *Staphylococcus aureus* (already discussed above Figure 11a) has been studied in our group in parallel with Rel_{Seq}. This enzyme, already described in paragraph 1.4.1.1.1, homodimerize thanks to a binding surface coordinated by Fe^{3+} and homo-tetramerizes due to a second binding surface formed by helices 5 and 6 (Figure 15d) to increase its enzymatic activity.

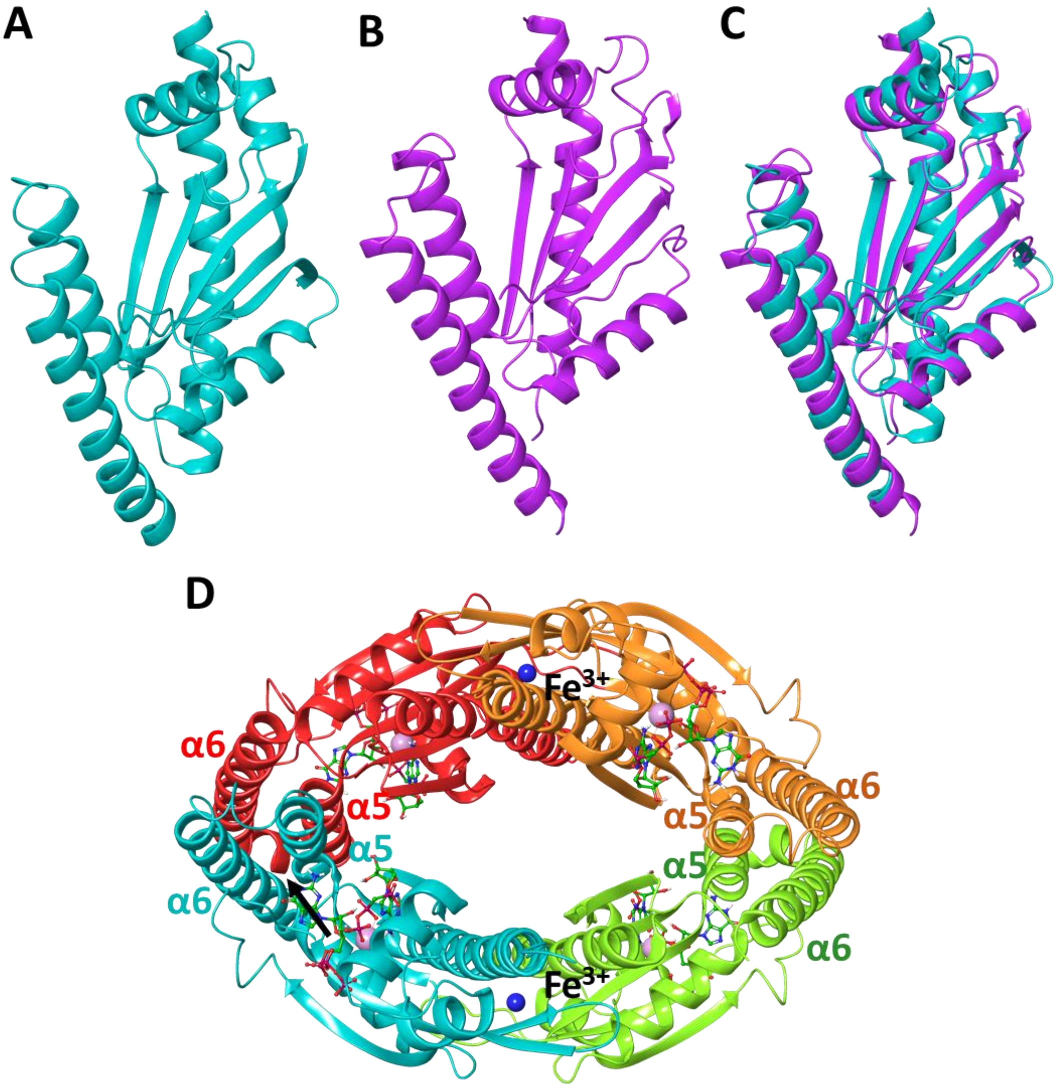


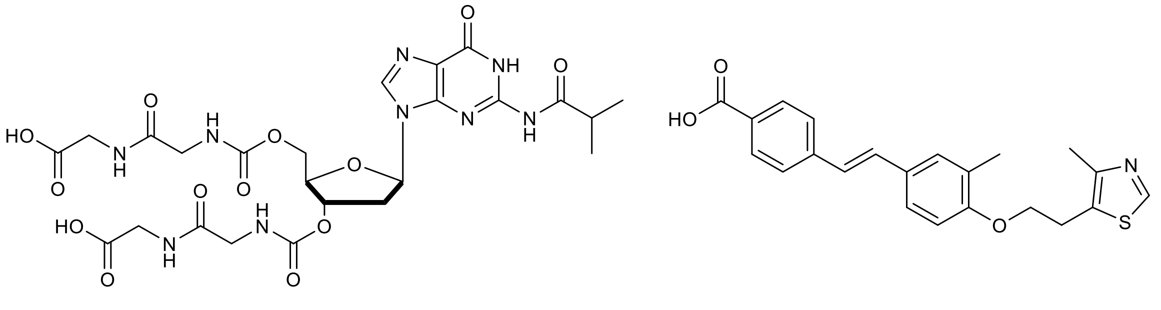
Figure 15: A) 3D representation of RelP from S. aureus (PDB entry: 5DEC). The protein is shown in cyan ribbons B) 3D representation of RelQ from B. subtilis (PDB entry:6FGJ) The protein is shown in purple ribbons C) superposition of RelP from S.aureus (cyan) and RelQ from B.subtilis D) 3D conformation of homotetrameric RelP (PDB entry:6EWZ). The binding is coordinated by Fe^{3+} (blue balls) and the two terminal α-helices. GTP and the modified non-hydrolysable ATP are represented in ball and stick green carbons, Mg^{2+} is rendered in pink balls.

1.4.3 SAH

Small alarmone hydrolases (SAHs) contain only the HYD domain. As described for RelH of *Corynebacterium glutamicum,*[77] their hydrolysis activity is Mn^{2+} and pH dependent. We know so far that this subfamily is the only one expressed also in eukaryotes such as *Drosophila melanogaster* and humans.[50,78] In both organisms, these SAHs called MESH1 can hydrolyse both (p)ppGpp and ppApp (adenosine-5'diphosphate-3'-diphosphate). In *D. melanogaster* MESH1 seems to have a role in starvation response since the lack of this enzyme induces growth reduction and impaired revival of amino acid depletion.[78] Otherwise, in humans, its role is the dephosphorylation of NADPH into NADH and inorganic phosphate for the cellular ferroptosis control.[79]

1.5 RSH inhibitor examples

The first attempts to inhibit (p)ppGpp synthesis with small molecules dates back to 2010-2013 when Wexselblatt and co-workers synthesised a small family of (p)ppGpp analogues[80–82] of which the main representative is Relacin[81] (Figure 16). This compound was able to inhibit the enzymatic activity of RelA and the Rel protein of *D. radiodurans* with a low mM IC50, and also to inhibit (p)ppGpp synthesis in a cell assay on the gram positive *B. subtilis* (but not on the gram negative *E. coli*). [81]
Later on, an auxotrophy-based high-throughput screening on *B. subtilis* only led to aspecific inhibition of the SR[83] while a wider structure–activity relationship (SAR) study on (p)ppGpp analogues failed to identify compounds more potent than Relacin.[84]
More recently, a high throughput screening of a GSK library of 2M compounds on Rel$_{Mtb}$ (*M. tuberculosis*) in a fluorescence polarization assay singled out compound X9 (Figure 16) as Rel$_{Mtb}$ inhibitor (IC50 in the low μM) and enhancer of the killing activity of antibiotic isoniazid.[5]
In 2021 Legèr and collegues[85] found out that NirD, the small subunit of nitrite reductase, inhibits (p)ppGpp synthesis and the activation of the stringent response by binding to the active site of RelA in *E. coli* both *in vitro* and *in vivo*.



Relacin **X$_9$**

Figure 16: 2D representation of Relacin and compound X9

CHAPTER 2: METHODS

In this chapter the computational methods used in this thesis are provided. Thechnique such as fragments-based drug design, docking simulations, molecular dynamics simulations and the PAINS (pan-assay interference compounds) filter used to analysed the results are described.

2.1 Fragment Based Drug Design

Fragments based drug design (FBDD) workflows are multi-step processes starting with target selection followed by an initial screen of the fragment library using biophysical techniques or computational approaches. The FBDD starts with the identification of fragments or low molecular weight compounds, "Rule of 3" (Ro3) compliant, soluble in dimethyl sulfoxide (DMSO) or phosphate buffered saline (PBS), that generally bind with weak affinity to the target of interest. These small molecules tend to be more polar and more soluble than larger druglike molecules and are therefore thought to translate into compounds with favourable physicochemical properties. After a validation of a potential hit, an iterative cycle of fragment modification can occur leading to the identification of a lead compound. Sometime different hits can be identified and linked together to increase affinity to the target. The lead compound is then tested both computationally and *in vitro* to ensure its binding affinity and then other modification that can affect solubility, rigidity or dimension are performed. If the final compound shows the desirable effect it can proceed to the *in vivo* phases.

2.2 Docking Simulations

The interaction between biological systems is essential for the activation or inhibition of biological processes. The characterization of the recognition process between proteins or a protein and its ligand can help in understanding how biological mechanisms, including diseases, occur and so develop drugs that can stop or intensify a biological process.[86,87] Molecular docking approach is a helpful computational technique able to predict the way a molecule binds to a protein using 3D coordinates. The atomistic coordinates of the protein can be taken from crystallographic or Nuclear Magnetic Resonance (NMR) experiments or from homology models while, on the other hand, ligand coordinates can be generated using computational tools.[88] The molecular docking procedure generates the so-called docking poses, a set of ligand-protein complexes, that correspond to local minima of the complex. The procedure tries to find out the native binding mode of the ligand generating complexes with the lowest free energy. The docking method is characterized by two steps:

1) sampling different conformations of the ligand into the active pocket of the protein and 2) associating a score value (scoring function) to each conformation to rank complexes formed.

In general, the first step uses algorithms that accommodate the molecule into the pocket changing orientation and conformation of ligand, protein or both (depending on the method used) into the binding site. While the scoring step is based on quantitative methods (scoring function) that evaluate binding affinity between two items in order to rank binding poses.

The molecular docking procedure not only is a useful technique able to predict the correct binding mode of the natural ligand, but it can also find out, *via* virtual screening (VS) campaigns, new ligands or it can be used to optimize a lead compound already identified.

If the binding mode of a protein is unknown, molecular docking simulations can be used to identify key residues involved in the ligand binding. Moreover, mutation analysis can also help in understanding the nature of drug-resistance or drug inefficacy in the patient. Finally, docking simulations can highlight possible off target effects evaluating the binding affinity of a molecule in different targets.

2.2.1 Sampling

Conformational rearrangements occur during the binding process of a ligand and its targets. Evaluating all the possible binding modes, including molecular flexibility, is time expensive and difficult in a computational point of view. Therefore, docking algorithms use three main strategies to solve this problem: i) rigid bodies, where both ligand and target are treated as rigid bodies and only the six rotational and translational degrees of freedom are explored; ii) rigid target and flexible ligand, where the conformational degrees of freedom of the ligand are also explored; iii) fully or partially flexible target and flexible ligand, in which also the protein conformational degrees of freedom are explored.[89] Nowadays, to balance speed and accuracy, the most used docking method is the one that treats the protein as a rigid body and the ligand as flexible.[90]

Modification of the structural parameters of the ligand, comprising rotational, translational and conformational degrees of freedom, are performed for the sampling that can be performed by using two main methodologies: systematic and stochastic. In the first method, all the free energy landscape is explored, by varying gradually each structural parameter, until the global minimum is found.[91] As easily understandable, this method is highly time consuming. The stochastic method, instead, provides a random structural parameter change at each step allowing the generation of a wide variety of different solutions. Different can be the stochastic algorithm used (Monte Carlo, Genetic algorithms, Tabu Search, Swarm Optimization) to accept or reject the proposed solutions according to probabilistic criterion, that reduces the computational cost. The main drawback of this second method is that the global minimum free energy conformation can be not found. Thus multiple run are recommended to increase the chance of an optimal solution.[89]

The conformational changes that a protein can face during ligand binding, including rearrangement of the secondary and tertiary structures, are harder to evaluate and more time consuming. Five are the method that can deal with these issues:[89,91,100,92–99]

i) Soft docking, in which van der Waals repulsive contributions are softened leading the overlap of small atoms to better accommodate the ligand into the binding pocket. The method is fast but can be used only when small local receptor motions occur.

ii) Side-chain flexibility, where different conformations of the side chains are sampled maintaining the backbone fixed. Also in this case, the method can be used only for local motions of the target.

iii) Molecular relaxation, Monte Carlo or Molecular Dynamics (MD) minimization are performed once the docking is performed to optimize structure and evaluate stability.

iv) Ensemble docking, where different conformations of the target, generated by NMR, crystallographic experiments or from computational models (molecular modelling or MD). The method is promising but lacks of a protocol that *a priori* helps in the selection of an optimal subset of target structures.[98] Big rearrangements make this method fail.

v) Collective degrees of freedom approaches, that include all the target flexibility considering only the dominant motion modes by reducing high-dimensional conformational landscape. This method can be used only after normal mode or principal component analysis (PCA) therefore it is limited due to the high computational cost required.

2.2.2 Scoring functions

The scoring functions are mathematical equations consisting in terms representing physical properties of the interacting molecules. They are used to estimate ligand-protein theoretical affinity.

Scoring functions can be classified into different types:[101] empirical, force-field based and knowledge based.

In the empirical functions the binding free energy is provided by summing weighted values of unrelated variables.[89] The scoring function is the sum of terms describing the complex binding (hydrogen-bond, hydrophobic and hydrophilic interactions, desolvation and entropic effects). All these terms are weighted by proper coefficients optimized to reproduce experimentally determined affinity data of complexes. This scoring function is fast, but the accuracy depends on the training test. Force field functions calculate binding energy via classical force field using the sum of bonded (bond stretch, angle bending and torsional energies), non-bonded (electrostatic and van der Waals), desolvation and entropic energies terms.[90] These kind of scoring functions cost a lot in term of time and usually overestimate charged atoms interactions.[89]

The Knowledge-based scoring functions base their equations on statistical observations of protein-ligand contacts found in large 3D databases (i.e., PDB). This method assumes that contacts that occur more frequently are energetically more favourable than other interactions.[90] Therefore, the score obtains a big improvement if a favourable (the most recurrent) interaction is present and a little improvement in the presence of a rarer one. These scoring functions are very fast, but their reliability depends on the training set diversity.

Every scoring function has limitation, therefore a consensus scoring approach, that provide the usage of different docking programs and scoring functions, is needed to improve the docking results accuracy.

2.2.3. GLIDE Docking

Different are the software packages and scoring functions used nowadays. In this thesis the software Glide[93,102,103] (Grid-based Ligand Docking with Energetics) and GlideScore scoring function were used.

Glide is a docking program used to predict the binding pose of molecules into proteins binding sites ranking them by scoring function. It uses the systematic conformational search as sampling method and a mixture of the empirical and force-field based terms for the scoring-function. Glide can use different docking protocols with three different accuracies:[93,102,103] i) the Standard-Precision (SP) method able to identify a wide variety of binders reducing the false negative due to the employ of a soft scoring function; ii) high-throughput virtual screening (HTVS) method that uses the same algorithm and the same scoring function of SP, but it reduces both the number of intermediate conformations and the thoroughness of the sampling and the final torsional refinement; iii) the Extra-Precision (XP) method that starts with the SP sampling and then exploits its own procedure. This last method is used to minimize false positive, that could pass SP approach, by using a *harder* scoring-function that includes additional penalties if the ligand shape is not completely complementary to the receptor.

Glide uses a series of hierarchical filters to optimize binding pose to better accommodate the ligand into the pocket of the protein. First the torsional angle space of the ligand is explored, and several conformations are generated. The lower energy conformations are then screened into the active pocket of the protein to identify the correct positioning and orientation, if possible. The protein binding site is defined as a grid of boxes, of 1 $Å^3$ side dimensions, in which Coulomb/van der Waals (vdW) properties of the protein are assigned. The ligand placement is then validated conferring a score (derived from a discretized version of ChemScore function[104]), that calculates steric clashes penalties, hydrophobic interactions, hydrogen-bonds (H-bond) and metal-ligation interactions. Subsequently the best poses are minimized in the protein grid using a molecular mechanics scoring function and a multi-drug strategy. If the conformation is still valid, the dimension of the boxes forming the protein grid, that includes coulomb and vdW parameters of the protein, is decreased in the area where ligand and protein interact to improve accuracy of the method. Then the poses with the lowest (best) energy are minimized *via* Monte Carlo approach. Finally, poses are rescored by using GlideScore (GScore), the score given by Glide procedure, that is based on ChemScore and includes terms such as penalties for electrostatic mismatches, steric-clash term, amide twist penalties, excluded volume penalties, rewards etc.

GScore equation can be summarized by equation 2.1:

$$GScore = o.o5 * vdW + 0.15 * Coul + Lipo + Hbond + Metal + Rewards + RotB + Site \quad (2.1)$$

where vdW is van der Waals energy, Coul is Coulomb energy, Lipo is lipophilic term, HBond is hydrogen-bonding term, Metal is metal-binding term, Rewards represents the rewards and penalties for hydrophobic enclosure, buried polar groups, amide twists etc. RotB is the penalty awarded for freezing rotatable bonds, Site is the term for polar interactions in the active site. SP/HTVS and XP equations are slightly different and they are reported in [102,103].

2.3 Molecular Dynamics Simulation

MD simulations are techniques that try to understand the macroscopic properties of a protein studying its microscopic behaviour. For example, it can figure out mechanisms involved in protein conformational changes, or it can help in calculating the binding free energy changes of a particular drug candidate. The first time MD was applied was in 1950s and early 1960s on liquids, while the first run on a macromolecule was performed by Martin Karplus group in 1977.[105,106] Statistical mechanics provided the mathematical expression that relates properties of motion and distribution of atoms of the N body systems to macroscopic thus finding out the connection between microscopic and macroscopic properties.[107] Therefore, statistical mechanics is the science branch that study the macroscopic systems from a molecular point of view.

2.3.1 Equation of Motion

The equation of motion, Newton's second law, is the mathematical expedient on which the MD simulation method is based on.

$$F_i = m_i a_i \tag{2.2}$$

F represents the force acting on a particle i, m_i is its mass and a_i is the acceleration of the particle i.

Equation 2.2 can be written as:

$$F_i(t) = m_i \frac{d^2 r_i(t)}{dt^2} \tag{2.3}$$

where $r_i(t)$ is the position vector of atom i at time t, m_i is the mass of the atom i and $F_i(t)$ is the force acting on it

By using equation 2.3, we can obtain from the trajectory the position, the velocity and the acceleration of the particles once the initial structure of the system, a set of initial velocities consistent with the simulation, temperature and a potential energy function E_{tot} for which $F_i = -\nabla_i E_{tot}(r_i, \dots, r_N)$ are provided.

Potential energy derivative can be related to the position changes as a function of time by using Newton's equation. Initial distribution of velocity in MD simulation is usually set randomly with the magnitude conforming to the required temperature and corrected to maintain the overall momentum (*P*) equal to zero by solving equation 2.4.

$$P = \sum_{i=1}^{N} m_i \boldsymbol{v}_{\boldsymbol{i}} \ = 0 \tag{2.4}$$

Most of the times the velocities v_i are set randomly from a Maxwell-Boltzmann distribution at a given temperature (equation 2.5):

$$p(\boldsymbol{v}_{ix}) = \sqrt{\frac{m_i}{2\pi k_{B\,T}}} \quad exp\left[-\frac{1}{2}\,\frac{m_i \boldsymbol{v}^2{}_{ix}}{2k_B T}\right] \tag{2.5}$$

where *p(**v**$_{ix}$)* is the probability of an atom *i* with mass m_i, moving in the *x* direction at a temperature *T* with velocity *v*.

2.3.2 Integrator

What we use to accelerate the atoms in the direction of the force applied are algorithms called integrators.[108] The equation of motion cannot be solved analytically, therefore, numerical methods have been developed for integrating the equations of motion. The integrator assumed that velocities, positions and acceleration can be approximated by a Taylor series. Moreover, the new position of the atom at time *t* + *dt* can be determined by Verlet algorithm using position and acceleration of the same atom at time t and position of the same atom at time *t* - *dt* as follow:

$$\boldsymbol{r}(t + dt) = \boldsymbol{r}(t) + \boldsymbol{v}(t)dt + \frac{1}{2}\boldsymbol{a}(t)dt^2 \tag{2.6}$$

$$\boldsymbol{r}(t - dt) = \boldsymbol{r}(t) - \boldsymbol{v}(t)dt + \frac{1}{2}\boldsymbol{a}(t)dt^2 \tag{2.7}$$

From these two equations the equation 2.8 can be obtained:

$$\boldsymbol{r}(t + dt) = 2\boldsymbol{r}(t) - \boldsymbol{r}(t - dt) + \boldsymbol{a}(t)dt^2 \tag{2.8}$$

where ***r*** is the position, ***v*** is the velocity (first derivative), ***a*** is the acceleration (second derivative with respect to time). Unfortunately, the algorithm used in this method is not self-starting due to the need of the estimation of the initial position and the results are not very precise.[108]

To solve these drawbacks the velocity Velvet algorithm, that uses the velocity to yield velocity, position and acceleration of the atoms is more frequently used:

$$\boldsymbol{v}(t + dt) = \boldsymbol{v}(t) + \frac{1}{2}[\boldsymbol{a}(t) + \boldsymbol{a}(t + dt)]dt \tag{2.9}$$

Some other integrator uses algorithms including higher order terms such as the Beeman's algorithm that uses the leapfrog algorithm.[108]

2.3.3 Force Field

Ab initio, semi-empirical quantum chemistry calculations and empirical methods are the way energy is calculated. Despite the great accuracy of the *ab initio* description using quantum mechanical calculations, computational power has prevented the use of this method for systems with more than few hundreds of atoms. That is why molecular mechanics (MM) are used. MM uses a set of parameters, named force field, to calculate potential energy. Potential energy calculated with this approach consists in bonding (bond lengths, angles and dihedral angles) and non-bonding (vdW and electrostatic interactions) terms given by equation 2.10

$$U = \sum\nolimits_{bonds}\frac{1}{2}K_r(r - r_{eq})^2 + \sum\nolimits_{angles}\frac{1}{2}K_\theta(\theta - \theta_{eq})^2 + \sum\nolimits_{dihedrals}\frac{V_n}{2}\left[1 + cos\left(n\emptyset - \Upsilon_{eq}\right)\right] +$$
$$\sum\nolimits_{i<j}4\epsilon_{ij}\left[\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}}\right] + \sum\nolimits_{i<j}\frac{1}{4\pi\varepsilon}\frac{q_iq_j}{r_{ij}} \qquad (2.10)$$

where U is the total potential energy and K_r, K_θ, and V_n are the force constants for bond-stretching, angle bending and dihedral angle deformations respectively. Similarly, r, θ, and φ represent the bond-length, valence and dihedral angle values respectively; r_{eq}, θ_{eq}, and γ_{eq} are the equilibrium values of bond-length, angles and phase angle respectively; ϵ_{ij} is the depth of the potential well, A_{ij} and B_{ij} are the finite distances at which the inter-particle potential is zero; ε is the dielectric constant; q_i and q_j are charges of atoms i and j, and r_{ij} is the distance between them.

The choice of the appropriate force field with the adequate parameters is of critical importance for the reliability of the simulation. Many force fields have been developed to be used in MD simulations. Several of them share analogues mathematical forms with different method to optimize the parameters. Among the most popular we can cite AMBER,[109] OPLS-AA,[110] CHARMM[111] and GROMOS.[112] All of them are all-atoms force field, which means that parameters for each atom are provided, except for GROMOS which is an united-atom force field, i.e. non-polar hydrogen parameters are not provided. Many optimized variants of each force field have been provided in the years especially for OPLS-AA[113–115] and AMBER.[109,116–118] Despite many tests were published to evaluate their reliability,[119–121] it is not possible to define the best force field to use. The use of one or another is dependent on the system investigated and the information we want to obtain from the simulations.

Most of the force fields have been developed to parametrize proteins and nucleic acids. However, to deal with drug discovery field, many generalized force fields were also developed to parametrize organic molecules. Examples are MMFF,[122] CHARM General force field (CGenFF)[123] and the General AMBER Force Field (GAFF).[124] OPLS-AA force fields can also be included in the list, since it parametrizes both proteins and organic moieties.

Nowadays, most of MD simulations are performed in explicit solvent therefore many parametrized models are available. In particular TIP3P,[125] TIP4P, [125] TIP5P,[126] SCP[127] and OPC[128] are the most used. The main difference between these solvation models consists in the number of interaction points used to represent water molecules. TIP3P and SCP use three interaction points (the three water atoms), TIP4P and OPC use four interaction points (adding a dummy atom with negative charge) and TIP5P uses five interaction points (using two dummy atoms with negative charges). More are the interaction sites more improved is the electrostatic distribution around the molecule higher is the computational cost in time. Thus, the three-site points are the most used solvation models. To further decrease computational costs, implicit solvation models can be also used. In this case the solvent is represented as a potential. This model is efficient with huge systems but less accurate than the explicit ones.

2.3.3.1 Interactions

Due to the simplification of energy calculation, MM method can be used on systems of thousands of atoms. However, the calculation of energy in big system is time consuming, therefore non-bonded interactions of atoms divided by a distance greater than a cut-off are ignored or scaled by a factor. The particle mesh Ewald (PME) is the most used method to deal with long range electrostatics where the potential energy is solved by using an approximate method.

PME is applicable if the dimension of the system is ‘infinite’. Therefore, MD simulations uses the expedient of the periodic boundary conditions (PBC), that tries to minimize the boundary effect of a finite system as the one used. This technique works duplicating the box containing the system replicating the simulation motion observed in the first box. Therefore, if an atom gets close to the boundary or goes out from the box, the same atom with the same velocity and direction appears in the opposite side of the box maintaining the number of atoms unchanged.

2.3.4 Thermodynamic Ensemble

Conservation of the energy is implied in Newtonian dynamics. Therefore, MD simulations would provide different configurations distributed according to the microcanonical ensemble environment NVE (constant number of particles N, volume V and energy E). However, this ensemble does not represent the realistic experimental or physiological conditions. More realistic conditions can be simulated by using the canonical ensemble (NVT) where number of molecules, volume and temperature are maintained constant, or isothermal-isobaric ensemble (NPT) where the constant pressure replaces the constant volume of canonical ensemble. Temperature and pressure can be maintained constant using thermostat and barostat, respectively, along the simulation.

The thermostat keeps the average simulation temperature close to the desired one. Different strategies were developed to respect this aim and the most common strategy is to modify or rescale properly the atomic velocities of the particles. The instantaneous temperature T(t), which can vary due to interconversion between kinetic and potential terms of the total energy, is related to kinetic energy *via* velocities of particles by using equation 2.11:

$$\frac{1}{2}N_f k_b T(t) = \frac{1}{2}\sum_i m_i\, v_i^2(t) \tag{2.11}$$

where N_f is the number of degrees of freedom, k_b is the Boltzmann's constant, m_i is the mass of atom i and v_i is the velocity of atom i. On the other hand, the barostat changes the volume of the system, rescaling the atomistic positions, to adjust the pressure.

The most used algorithms are the Langevin[129] thermostat, Berendsen[130] and Nosè-Hoover[131] velocity-rescaling. Berendsen[130] and Parrinello-Rahman[132,133] are instead the most used barostat.

2.3.5 MD limitations

The first limitation is obviously connected to the classical nature of the Newton aquation of motion that can only provide dynamic evolutions of the system protons without considering electron motion. Therefore, enzymatic or chemical reaction cannot be studied unless combined techniques (i.e. QM/MM) are used.

The second limitation is related to the force field. The parameters used to rich the desirable energy derived from the optimization of parameters fitted with the data obtained by experimental procedures or quantum mechanical calculations of little molecules or fragments that are then used as building blocks. Therefore, the parameters obtained by big molecule are not provided. Furthermore, force fields are specialized for different molecules (amino acids, nucleic acids, sugar) therefore also the choice of the correct one can be considered a limitation.

Finally, the most severe limitation is connected to the timescale. Actually, some motions can be seen after millisecond time simulations or more and nowadays this time exploration costs a lot of time especially if the timestep (the period that passed from the registration of a motion to another) is little as for the normal MD simulation (about 2 femtoseconds = $2 \cdot 10^{-15}$ s).

2.4 PAINS filter

Pan-assay interference compounds (PAINS)[134] are compounds that give false positive results in a large variety of high-throughput screening (HTS) campaigns. This attitude comes from their ability of binding non-specifically to proteins. In 2010 Baell and Holloway[134], after analysing results of six HTS campaigns, identified a great number of PAINS and listed them into three filter databases based on the resulted obtained in their study: filter A that contains the PAINS families with 150 or more members, filter B containing PAINS families with a number of members between 15 and 149, filter C with PAINS family whose members are less than 15.

They first labelled as problematic the structures that hit (give positive results) in at least two HTS campaigns. Subsequently, they grouped the compounds in families depending on the structure composition. If in a family of compounds the problematic members consisted in more than the 30% of the total number of compounds, the family was included in the PAINS database. The filters they built were first tested on the same compounds used in the six HTS campaigns previously analysed. From this test, some groups, such as 2-alkenylfurans, that were wrongly supposed to be problematic due to the linkage to the real problematic families were deleted from the PAINS databases (the filters). Finally the three filters previously listed were published and some years later implemented in many software, such as Canvas[135,136] that was used in the *in silico* procedures of this thesis.

CHAPTER 3: Rel$_{Seq}$ DYNAMIC BEHAVIOUR

Molecular Dynamics (MD) simulations were run using AMBER 18 package.[137] Staring from the X-ray structure of Rel$_{Seq}$ (chain A) in the synthetase 'on' conformation, four systems were prepared. By removing the GDP from the synthetase site the apo form was studied and compared to the bound state. By removing the Mn^{2+} from the hydrolase site its effect on the enzyme stability was also explored.

3.1 Systems preparation and MD setup

The crystallographic structure of hydrolase OFF/ synthetase ON conformation of the long RSH from *Streptococcus dysgalactiae subsp. equisimilis* (Rel$_{Seq}$, PDB[138] entry:1VJ7, chain A, residues 1-385)[62] was prepared using the 'Protein Preparation Wizard' tool of the Schrodinger suite (2018-v3).[139] According to Epik[140] calculation, performed at pH=7, the GDP was considered fully deprotonated (total formal charge -3). All the crystallographic water molecules were deleted and the structures of the two missed loops (K110-N123 and K153-D158) located in the HYD domain were built using Prime.[141,142] The longest K110-N123 loop was further refined by the 'Refine Loops' utility available in Prime[141,143] (OPLS3e[114] force field, VSGB[144] solvation model, 'ultra extended' sampling option and default parameters) and the model with the lowest Prime energy was selected for the refinement step of the Protein Preparation tool. The H-bond network of the protein was optimized and the protonation states of residues were determined according to PROPKA at pH=7. Finally the protein was relaxed by running a restrained minimization, (OPLS_2005[113], converge for heavy atoms to RMSD of 0.3 Å).

To investigate the dynamical behaviour of Rel$_{Seq}$ and to elucidate how it can be affected by the presence of the natural ligand (GDP) into the SYNTH site, and by the Mn^{2+} cofactor into the HYD domain, four systems were built using Rel$_{Seq}$ chain by deleting the ligand and/or the ion from the protein previously prepared:

- *"holo"* system, including Rel$_{Seq}$ bound to both GDP and Mn^{2+};
- *"holo no Mn",* including Rel$_{Seq}$ bound only to GDP;
- *"apo",* including Rel$_{Seq}$ bound only to Mn^{2+};
- *"apo no Mn",* including Rel$_{Seq}$ alone.

The systems were solvated in a TIP3P[125] cubic water box (*holo* and *holo no Mn*: 27904 and 23841 water atoms, *apo* and *apo no Mn*: 27912 and 23851 water atoms for no-refined and refined loop structures, see below) and submitted to three geometrical optimization with the deepest descent algorithm (first only water, then only the protein and finally all the system) for a total of 6000 cycles by using AMBER 18. Then, 200 ps of equilibration step were computed in NVT ensemble restraining the protein atoms (harmonic constant of 10 kcal/mol·Å^{-2}) and gradually increasing the temperature to 300 K (from 10 K to 150 K within the first 100 ps and from 150 K to 300 K in the last 100 ps). The time step (dt) was set to 0.5 fs and the collision frequency (Langevin thermostat[129]) to 1 ps^{-1}. A second equilibration step of 200 ps, with no restraints, was performed (NVT, T=300 K, dt=0.5 fs and Langevin[129] thermostat). The last equilibration step consisted in a 100 ps long NPT simulation (T= 300 K, P= 1 atm Langevin isotropic coupling with a pressure relaxation time of 2 ps). This step improves the total energy of the system (from -215862.6726 kcal/mol to around -220000 kcal/mol) and stabilize the density at 1.006 g/cm^3.

Per each system, production runs consisted of three MD simulations, each one 500 ns long, in NVT ensemble (Langevin thermostat, T = 300 K, dt = 2 fs, collision frequency = 1 ps^{-1}) using the amber ff14SB force field for protein,[117] GDP amber parameters downloaded from amber website (http://research.bmh.manchester.ac.uk/bryce/amber/) and Lennard-Jones (LJ) 12-6 non-bonded model for the Mn^{2+}.[145] The three replicas differ for the starting velocities, chosen randomly by the program (based on date and time), and for the starting 3D atomistic coordinate of gap K110-N123 (before the refinement and after it).

3.2 ANALYSES

A meta-trajectory was obtained by concatenating the three runs of each system and analysed with CPPTRAJ,[146] an analysis tool implemented in AMBER , and the plots were generated using Jupyter notebook and specific python libraries, i.e. pandas,[147] scipy,[148] matplotlib[149] libraries.

The presence of Mn^{2+} does not affect the behaviour and the stability of both the *holo* and *apo* systems and similar results were observed. in the following sections only the results of the *holo* and *apo* systems in the presence of Mn^{2+} are provided (see appendix section for the results of the corresponding system without Mn^{2+}).

3.2.1 Protein analysis

RMSD

The Root Mean Square Deviation (RMSD) of atom positions calculates the difference in the 3D atomistic coordinates (deviation) of the protein in each frame compared to a reference structure.

The equation 3.1 was used to calculate the RMSD:

$$RMSD = \sqrt{\frac{\sum_{i=1}^{N} \delta_i^2}{N}} \tag{3.1}$$

where δ_i is the distance between current and reference position of atom I and N is the total number of atoms included into the calculation. Generally, the reference position is the starting structure or the mean one calculated over the simulation.

Figure 17 shows the RMSD computed for Cα atoms of the protein in *holo* and *apo* forms *versus* the total simulation time with respect to an average structure calculated by CPPTRAJ on the meta-trajectory and used as reference. The two meta-trajectories converged into stable conformations.

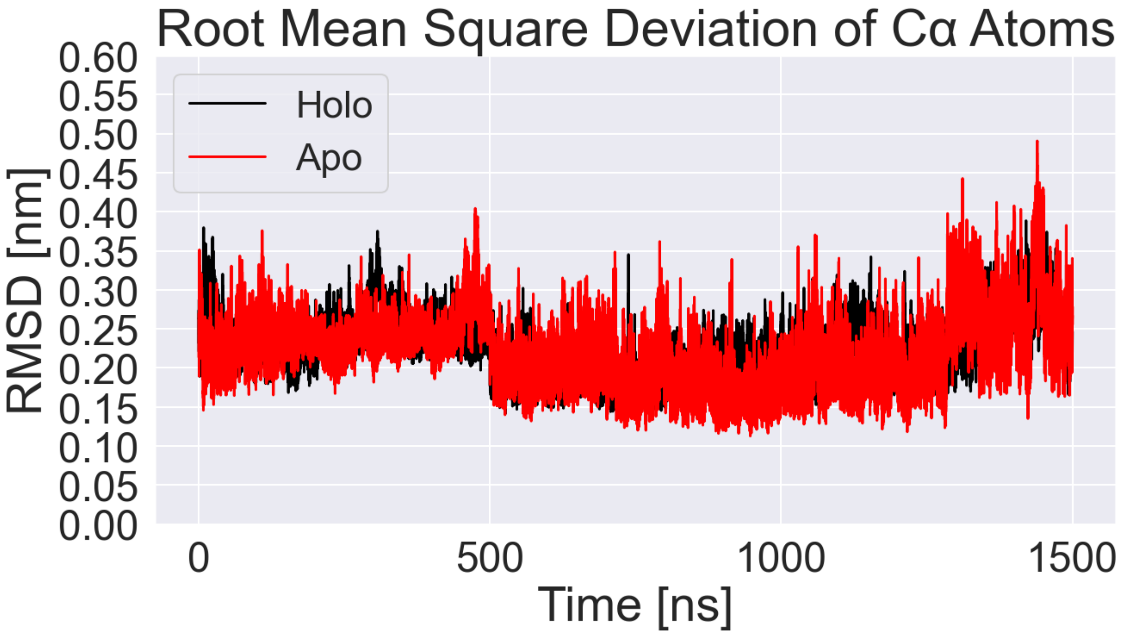


Figure 17: *RMSD calculated on Cα atoms of holo (black) and apo (red) Rel$_{Seq}$.*

RMSDs of Cα atoms of single domains were also calculated to evaluate which domain caused the highest deviation in the protein compared to the average structure and, as shown in Figure 18, the HYD domain is characterized by higher RMSD values compared to the SYNTH domain.

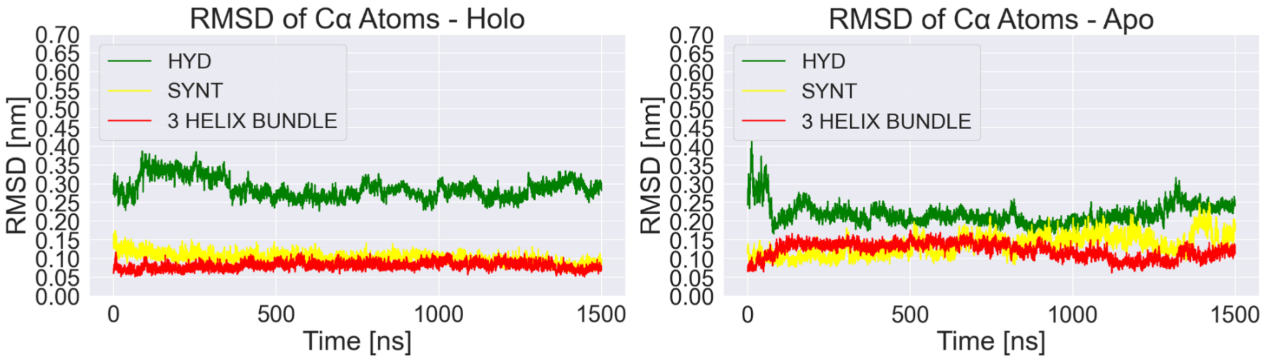


Figure 18: RMSD calculated on Cα atoms of single domains of holo system (left) and apo system (right). HYD domain is shown in green, 3 helix bundle is shown in red, SYNTH domain is shown in yellow.

RMSF

The Root Mean Square Fluctuation (RMSF) was calculated to evaluate residues fluctuation during the simulation by using equation 3.2:

$$RMSF = \sqrt{\frac{\sum_{i=1}^{T}\left|r_i(t)-\bar{r}_t^2\right|}{T}} \tag{3.2}$$

where T is the total number of frames, $r_i(t)$ is the position of the ith residues at ith frame and $\bar{r}_t$ is the average position of ith residues in the trajectory.

The fluctuation profile of Cα atoms of the enzyme was computed by calculating the RMSF considering the same averaged 3D atomistic coordinates used for the RMSD analysis as reference structure. As expected, the most flexible residues during the simulations are those located in loop regions and particularly the ones located in the loop built by Prime (loops K110-N123 and K153-D158) (Figure 19). Comparing the two meta-trajectories, similar RMSF profiles were observed except in the case of the G-loop (SYNTH site, residues Y299-S310), where residues involved in GDP binding are located. In fact, in this region the fluctuation is higher in the *apo* system, where the GDP is missing, than in the *holo* one where the presence of GDP stabilizes the loop and the near residues.

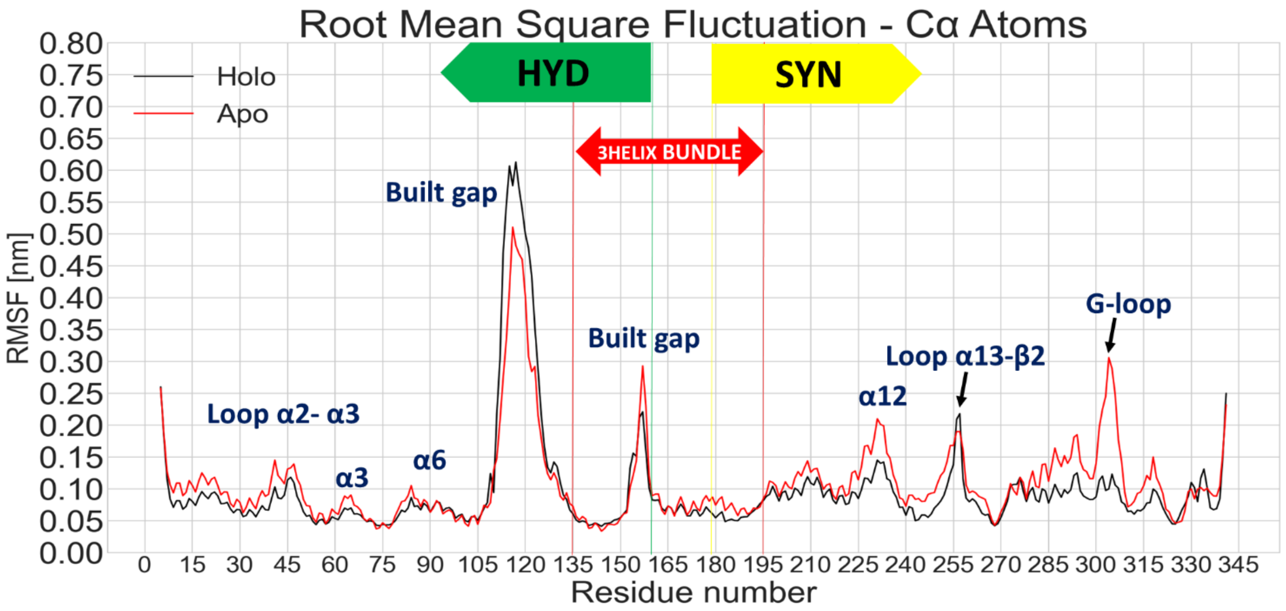


Figure 19: RMSF of holo (black) and apo (red) systems

Radius of Gyration

The Radius of gyration (RoG)[150] is the distribution of atoms around its centre of mass and it is calculated to evaluate the compactness and the folding state of proteins. First, the coordinates of the centre of mass R_c are determined by using equation 3.3

$$\sum m_i(r_i - R_c) = 0 \tag{3.3}$$

where m_i is the mass of the i^th^ atom, r_i its coordinate and R_c represents the coordinates of the centre of mass.

Considering the atoms as a ball with radius R, the RoG is then calculated by equation 3.4:

$$R_g^2 = \sum_{i=1}^{N} \frac{(r_i - R_c)^2}{N} + \frac{3}{5}R^2 \tag{3.4}$$

where N is the number of atoms excluding hydrogen atoms in a protein and $\frac{3}{5}R^2$ is the radius square of a ball with radius R and a uniform density.

RoG was calculated to evaluate how the ligand and the ion could affect protein folding. As shown in Figure 20 and Figure 21, the average RoG calculated on Cα atoms of the entire protein (*holo* = 22.8 $Å^2$, *apo* = 22.6 $Å^2$), HYD domain (*holo* = 15.5 $Å^2$, *apo* = 15.6 $Å^2$) and of the SYNTH domain alone (*holo* = 16.5 $Å^2$, *apo* = 16.6 $Å^2$) were very similar highlighting that the ligand had no influence on the folding of the enzyme.

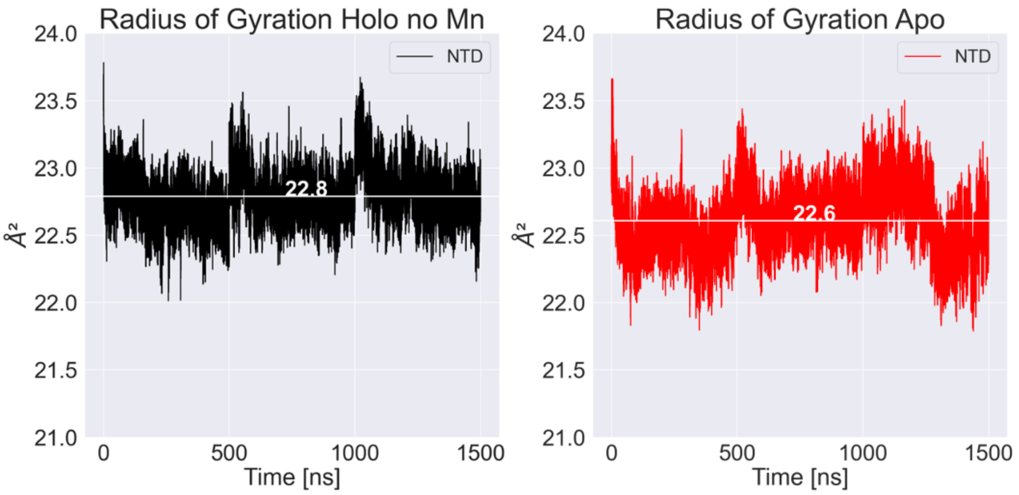


Figure 20: Radius of gyration values for backbone atoms of the entire systems (NTD stands for N-termina domain). Holo system is represented in black, apo system is represented in red. The average value of the RoG is represented in white lines and numbers.

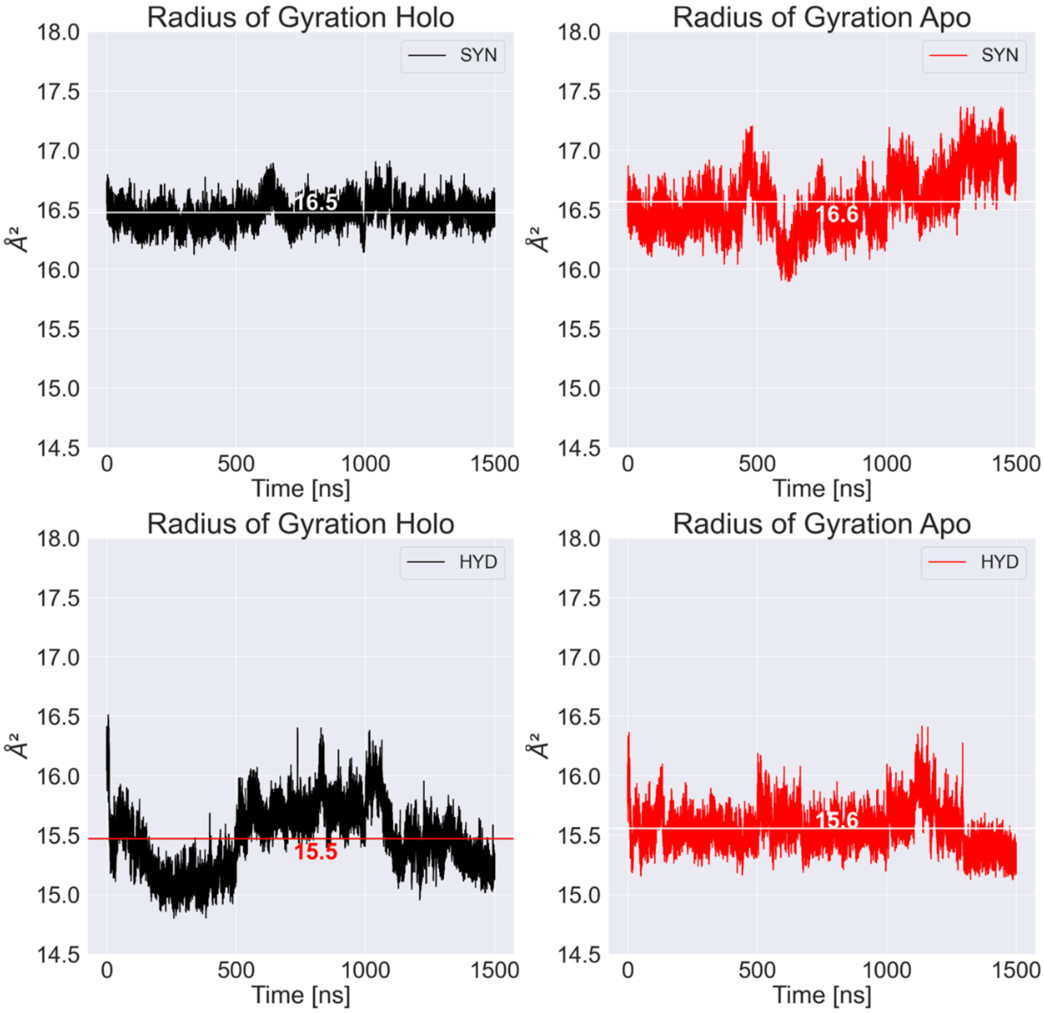


Figure 21: Graphics representing the radius of gyration values for backbone atoms of the SYNTH (top panel) and HYD (bottom panel) domains. Holo system is represented in black apo system is represented in red. The average value of the RoG is shown in white lines and numbers for the graphics representing SYNTH domain of both holo and apo systems and for the graphic representing the HYD domain of apo system while it is shown in red line and number for the graphic representing the HYD domain of holo system .

Cluster analysis

Cluster analysis is a technique that groups data in different data sets, called clusters, in a way that objects belonging to the same cluster are more similar to each other than to those included in the other clusters. In this case, this method was used to isolate the most representative conformations explored by the protein, or by the ligand, during the simulation, in order to evaluate potential conformational changes during the MD. Among the several algorithms that can be used to perform cluster analysis, in this thesis the hierarchical agglomerative one was chosen,[151] using the average linkage[151] method. The analyses were computed considering Cα atoms of HYD domain (residues 176-341) with a RMSD threshold of 1.2 Å. This means that a new cluster was generated when the RMSD of Cα atoms was higher than 1.2 Å.

This analysis was performed to better understand the conformational changes that occurred in the protein. We tried to figure out if the 3 helix bundle and part of the SYNTH domain rearranged to switch OFF the SYNTH activity, inducing the activation of the HYD domain, as already proposed by Hogg and collaborators.[62]

According to the results, the 3D atomistic coordinates of the structured regions of the three domains remained unchanged (Figure 22). However, the HYD domain showed the highest number of clusters within the three domains with the smallest percentage of frames belonging to the three main clusters (less than 5% each Table 1). The rearrangements of this domain are connected principally to the high flexibility of the big loop that was built and refined and to the loop in the 3-helix bundle that was built (included in the calculations because part of the HYD domain). If the cluster analysis is carried out excluding these two loops (106-135 and 153-158), the number of clusters is reduced to one main cluster and (92% of structure, Table 1).

The central 3 helix bundle is the most stable domain. In fact, there is one main cluster for the *holo* system and three main clusters for the *apo* (Table 1). The main difference among the clusters is connected to the conformation of the loop connecting helix 8 and helix 9 that we built.

The most interesting difference between the two systems was observed in the SYNTH domain. Although the number of clusters is not so different, the main cluster of *holo* system comprises a higher number of frames compared to the *apo* one (table 1). Furthermore, as shown by Figure 22, the flexibility of the G-loop is much higher in the *apo* system than in the *holo* one, as already showed by the RMSF plot (Figure 19).

In conclusion, the cluster analysis pointed out the significant conformational stability of HYD and 3 helix bundle domains, and confirmed that the presence of GDP is essential to maintain the G-loop in a stable orientation.

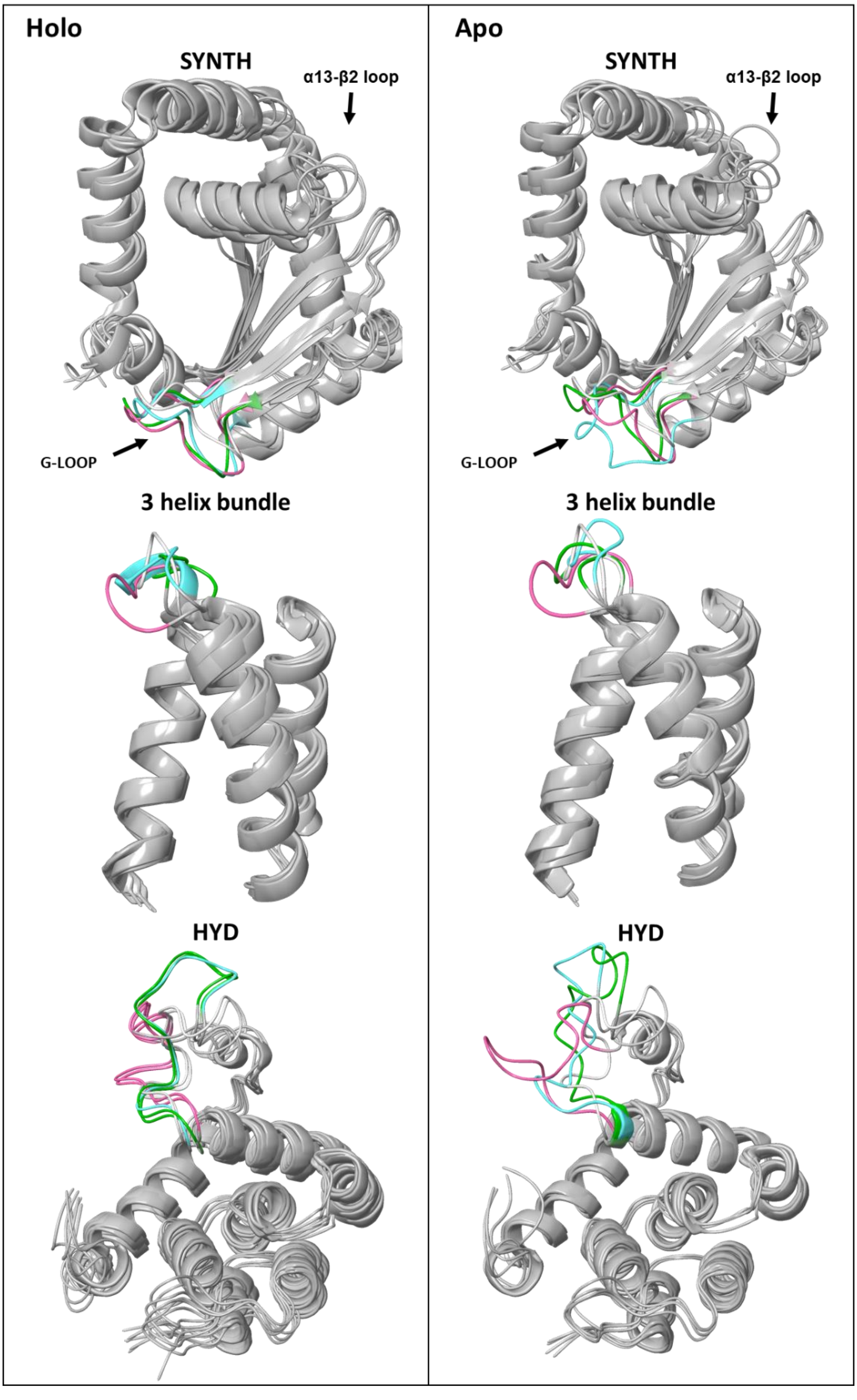


Figure 22: 3D representation of the conformations isolated from the three main clusters. The starting X-ray structure is shown in white, the medoid structures are shown in grey ribbons. The most variable regions are colored according to a cluster color code: the main cluster is shown in green, cluster 2 is shown in purple and cluster 3 is shown in light blue.

Table 1: Number of clusters and percentage of frames belonging to the three main clusters

System	Domain	Total number of clusters	% frames in cluster 1	% frames in cluster 2	% frames in cluster 3
Holo	**HYD**	475	4	4	3
	HYD no loops	3	92	5	3
	3 helix bundle	6	75	13	7
	SYNTH	6	64	33	1
Apo	**HYD**	318	8	7	7
	HYD no loops	7	88	4	3
	3 helix bundle	5	49	30	18
	SYNTH	9	51	34	10

Solvent Accessible Surface area

The Solvent Accessible Surface Area (SASA) of the SYNTH domain was calculated by using the Linear Combinations of Pairwise Overlaps (LCPO) algorithm[152] in order to investigate if the SYNTH pocket of the *apo* system showed a wider 'open' conformation to better accommodate the ligands. The total SYNTH surface area calculated on the prepared X-ray structure is 8169.02 $\AA^2$ and as shown in Figure 7 SASA of both *holo* and *apo* systems increased during the simulations. However, the surface area average of *holo* system (9113.9 $\AA^2$) is lower than that calculated for the *apo* one (9244.4 $\AA^2$), confirming in part that the motility of the loop seen in the cluster analysis could lead to the opening of the pocket. To further investigate this aspect, the SASAs of the residues locate within 5 Å distance from GDP and the residues of the G-loop were calculated (Figure 23). As expected, in both analyses the SASA of *apo* system is higher and more susceptible to oscillation than that of the *holo* one.

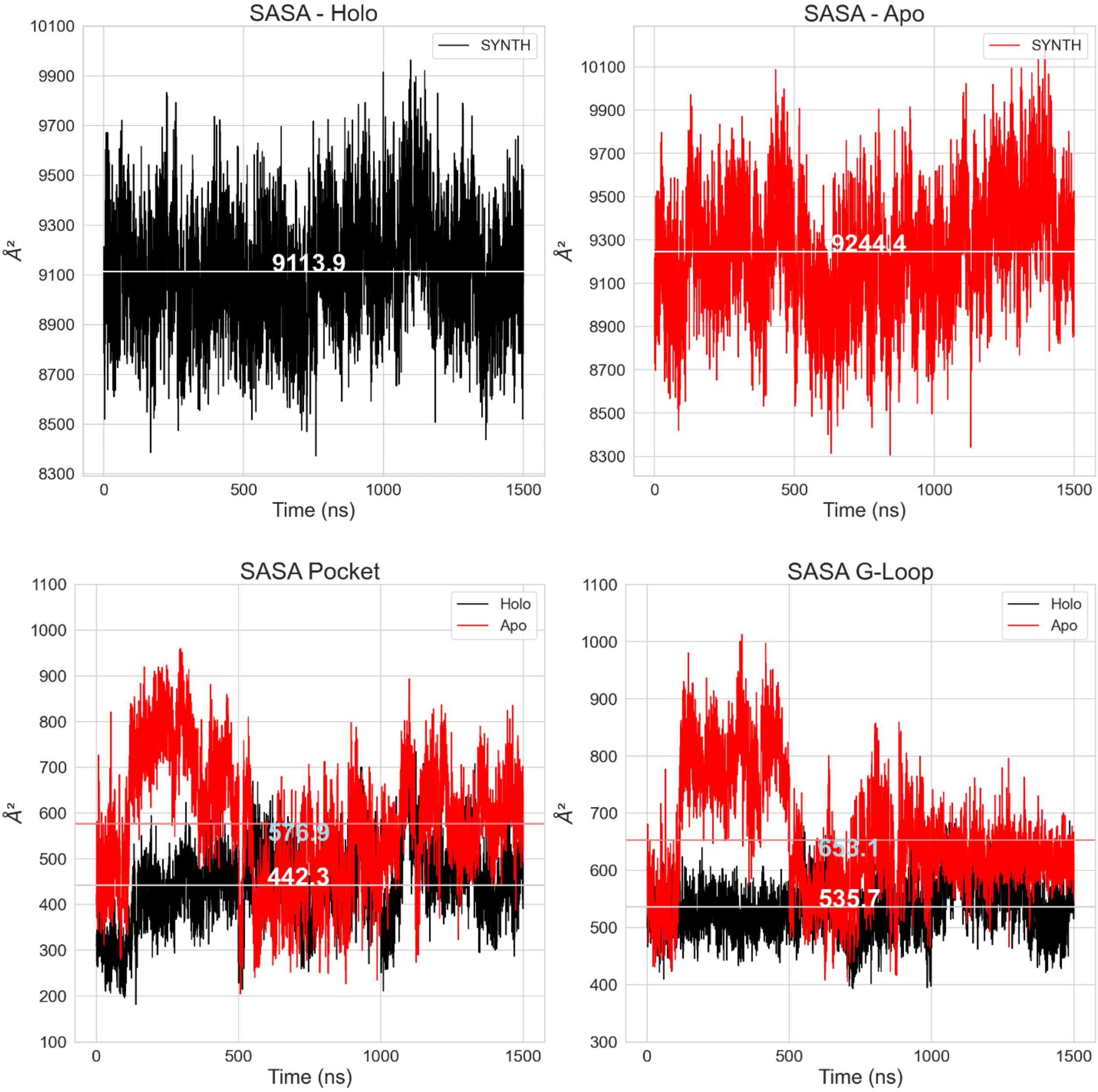


Figure 23: In the graphs is expressed the total Surface Area (in Å^2) vs time (in ns) of systems holo (top left), apo (top right), SASA of residues within 5 Å from GDP (bottom left) and SASA of G-loop residues *(bottom right). Holo SASAs are shown in black, apo SASAs are shown in red. Average SASAs of holo system is shown in white, average SASAs of apo system is shown in salmon line and light grey number.*

Principal component analysis (PCA)

Essential dynamics was carried out to further elucidate the principal motions of the protein in the four systems. First, a correlation matrix was calculated according to equation 3.5:

$$AvgCorr(a,b) = \frac{\sum V_a(i)\cdot V_b(i)}{N} \tag{3.5}$$

Where the average correlation between the vectors V_a and V_b, representing two motion vectors of residue a and residue b, is calculated as the average of the dot product of those vectors over all *N* frames.

Looking at the correlation matrices showed in Figure 24, a positive correlation can be observed between residues located in the same domain, while some residues belonging to different domains are anticorrelated. Moreover, few residues of the 3 helix bundle domain correlate or anticorrelate either with the HYD domain or with the SYNTH one. Moreover, *apo* system shows a higher correlation, and anticorrelation, compared to the *holo* form. These data suggest that, in little part, GDP influences the flexibility of the system. In detail, in apo system residues located in helix 12 (residue between 217-231) and loops surrounding beta sheet 2 and beta sheet 2 itself (residue between 265-271), showed more compactness and anticorrelation with residues located in alpha 2 (22-38) and alpha 3 - alpha 4 (50-75) compared to the same residues in *holo* system.

Moreover, residues located in the loops we built (105-135 and 153-158) show two correlation trends different for *holo* and *apo* systems highlighting the probable bias caused by the high flexibility of these loops.

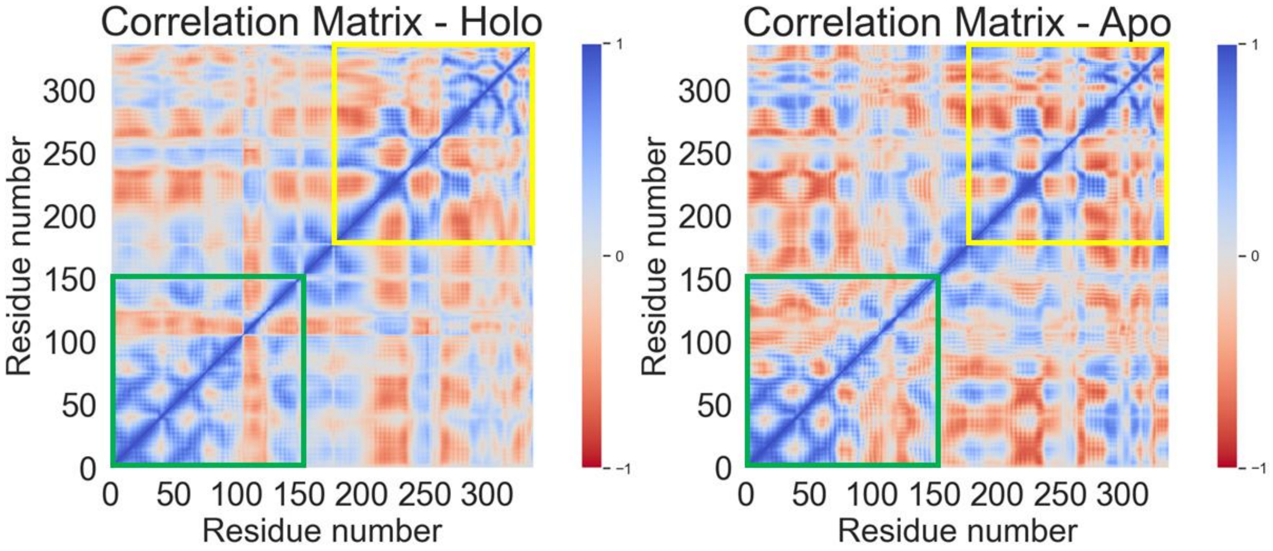


Figure 24: Correlation matrices computed for each meta-trajectory. In x and y axes are shown the Rel$_{Seq}$ residues numbered from 1 instead from 5. Residues that correlate are shown in blue, residues that anticorrelate are shown in red, residues that do not present any correlation are shown in white. The HYD domain is defined by a green box, the SYNTH domain defined by a yellow box,

Starting from the diagonalization of the covariance matrix of dimension 3N X 3N where N is the number of residues of the protein studied (equation 3.6)

$$R^T CR = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_{3N}) \tag{3.6}$$

where $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_{3N}$ are the eigenvectors, R is an orthonormal transformation matrix, R^T is the transpose of R and C is the covariance matrix with elements C_{ij} for coordinates i and j (equation 3.7),

$$C_{ij} = \langle (x_i - \langle x_i \rangle)(x_j - \langle x_j \rangle) \rangle \tag{3.7}$$

where x_i and x_j are the mass-weighted Cartesian coordinates of a system of N-particles i and j have values range from 1 to 3N, and $\langle x \rangle$ is the average over all the structures sampled during the simulation. the principal components (PC) were computed for single domains and PC analysis (PCA)[153] was performed. PCA is a technique performed to reduce the number of dimensions needed to describe data, in this case protein motions, defined as eigenvectors. The eigenvectors are computed via a decomposition process (singular value decomposition) performed on covariance or correlation matrices that filters the motions from the largest to the smallest spatial scale, that is determined by the corresponding eigenvalues. In this way, lower-dimensional data, namely PCs, are generated trying to preserve all the possible data variance. So, PCA can be defined as an orthogonal linear transformation that converts the data into a new simplified coordinate system. According to this method, the biggest scalar projection of the data resides in the first coordinate (the first PC), the second one in the second coordinate (the second PC) and so on. These data represent a qualitative investigation useful to describe the motions of the protein observed into different conditions (*apo* and *holo*) and considering the simulated time-window.

To exclude the biases caused by the high flexibility of the loops we built, PCs of HYD domain were calculated excluding residues located in these loops. The first three components represent about the 43% of the entire motion of the domain (Figure 25) and they consist principally in the motion of the catalytic loop (residues H40-Y50), involved in the formation of the catalytic HYD site, and of the residues located near the loop we built in the 3 helix bundle (figure 26), for both *apo* and *holo* systems. The main motions of the central 3 helix bundle (data not shown) are exclusively due to the loops where the gap was built. On the other hand, the first three principal components of the SYNTH domain represent about 65% of the entire motion of the domain (Figure 27). In particular, for the *holo* system they consist of a little twist of the domain, that induces a transition of the G-loop towards to the centre of the active site, with an important motion of the loop between helix 13 and beta sheet 2

(Figure 28). The same behaviour can be seen for the *apo* system with an increased motion of the G-loop (Figure 28).

The scatter plots (Figure 29) of HYD PCs showed superimposable motions between the two systems, while, the scatter plots of the first three PCs of the SYNTH domain motions shown, as expected, a higher range of exploration in the *apo* compared to the *holo* systems (Figure 29).

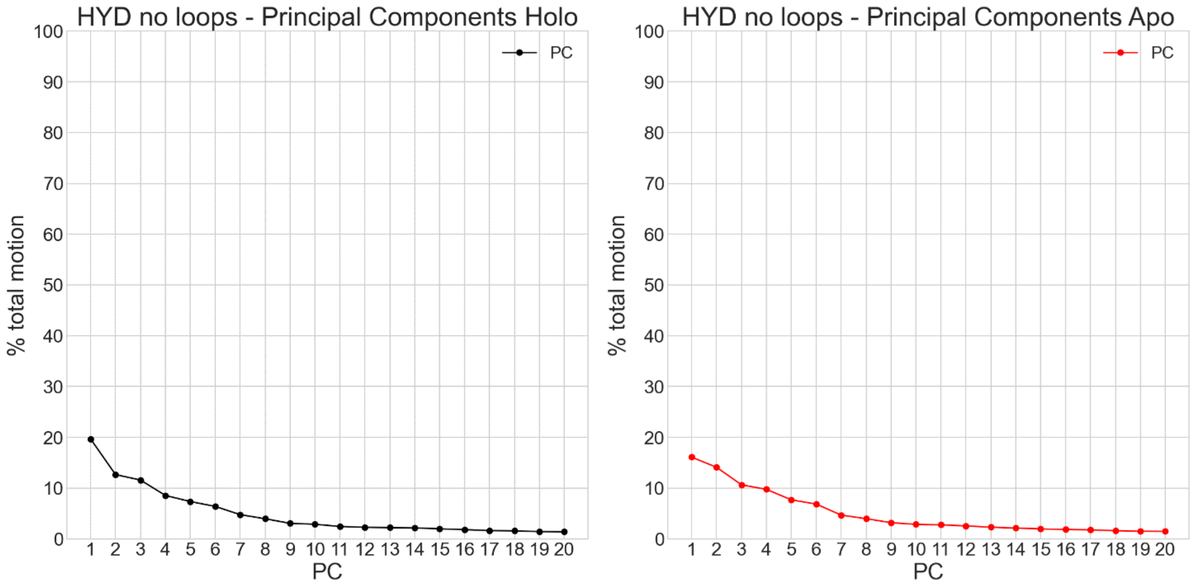


Figure 25 : Weight of 20 Principal Components motion on the overall motion in percentage of the HYD domain. Holo system PCs are shown in black in the left panel. Apo system PCs are shown in red in the right

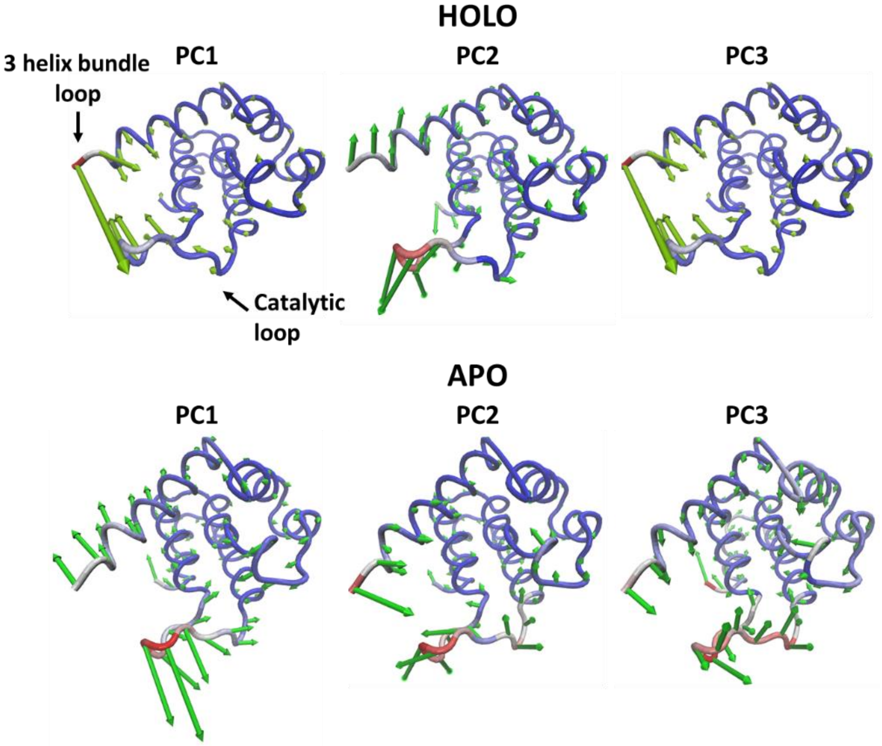


Figure 26: 3D representation of the first three principal components motion of HYD domain of holo (top panel) and Apo (bottom panel) systems. The eigenvectors are shown by green arrows, the eigenvalues are

shown by the length of the arrow and by the colour code: red high eigenvalue, white middle eigenvalue, blue low eigenvalue.

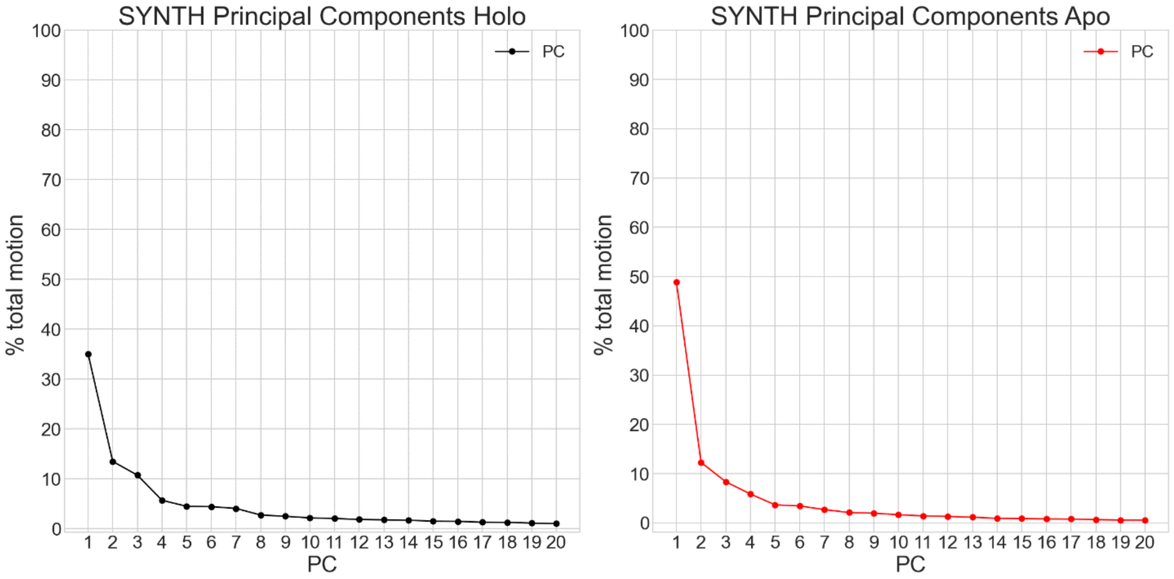


Figure 27: Weight of 20 Principal Components motion on the overall motion in percentage of the SYNTH domain. Holo system PCs are shown in black in the left panel. Apo system PCs are shown in red in the right panel

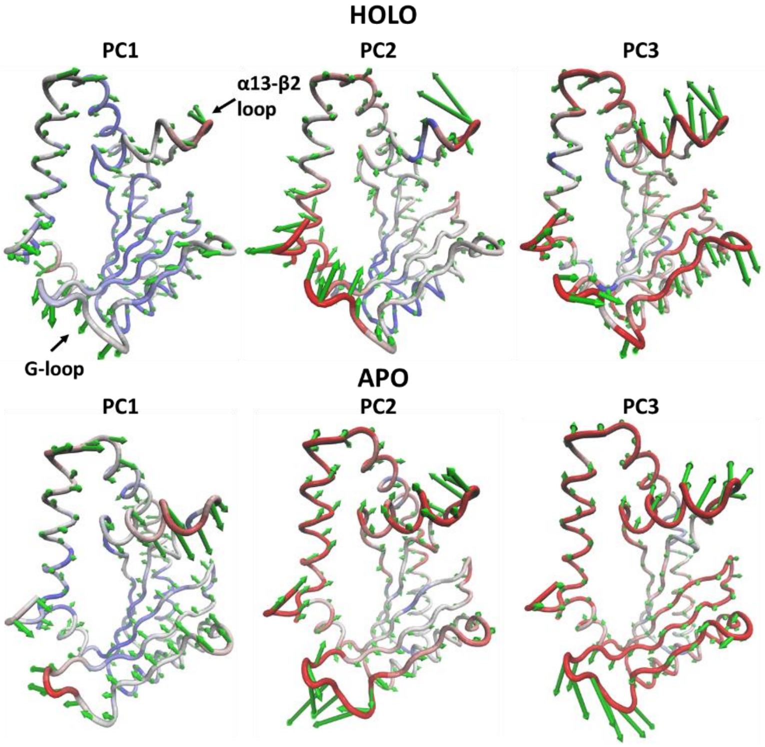


Figure 28: 3D representation of the first three principal components motion of SYNTH domain of holo (top panel) and Apo (bottom panel) systems. The eigenvectors are shown by green arrows, the eigenvalues are shown by the length of the arrow and by the colour code: red high eigenvalue, white middle eigenvalue, blue low eigenvalue.

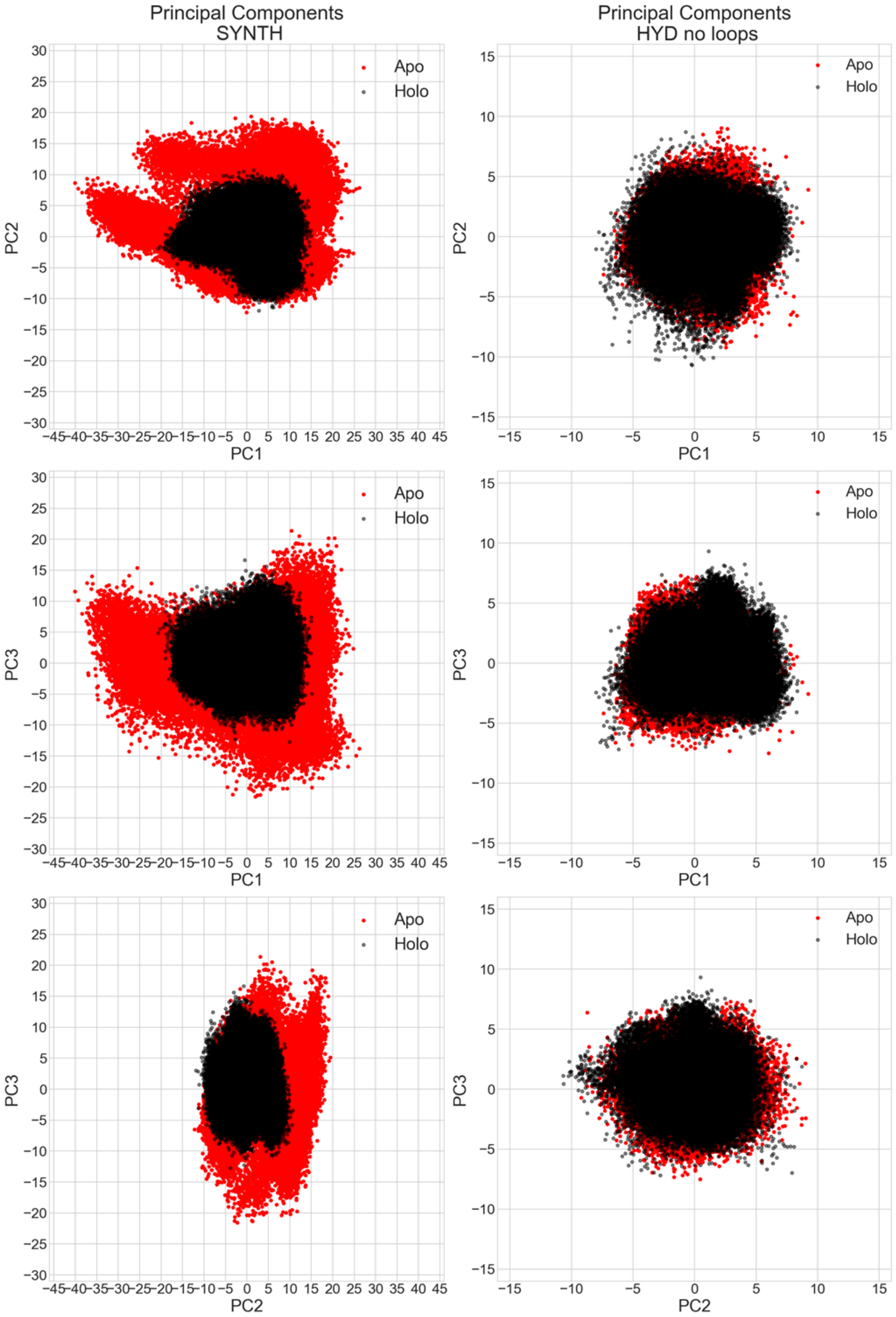


Figure 29: 2D scatter plots representing PC1 vs PC2 (top panel), PC1 vs PC3 (middle panel), PC2 vs PC3 (bottom panel) of SYNTH (left) and HYD (right) domains. Holo system is represented in black, Apo system is represented in red

RMSIP

To investigate if the eigenvectors of different simulations overlap, evaluating both conformational subspace sampled and similarity of the essential subspace explored, the root mean square inner product (RMSIP)[154] were calculated by solving equation 3.8:

$$RMSIP = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N}\left(n_i \cdot v_j\right)^2} \tag{3.8}$$

where N is the number of eigenvectors we want to compare and n_i and v_j are the two eigenvectors of the two simulations, or simulation windows, we want to compare.

RMSIP was evaluated between replicas of the same system and between meta-trajectories of the *holo* and *apo* system.

Comparing the two meta-trajectories, the RMSIP (Table 2) calculated on the first three eigenvectors of SYNTH and HYD domains (excluding the residues forming the loops we built) are 0.6 and 0.9, respectively, defining these motions overlapped (generally a value ≥ 0.6 defines a good overlap of two eigenvectors).[154] Similar overlap can be seen calculating the RMSIP of the tree main eigenvectors of single replicas (Table 2). Actually, comparing all the replicas of the same system two by two, the RMSIP average calculated for the SYNTH domain are 0.7 and 0.6 for *holo* and *apo* systems, respectively, while the lowest RMSIP value calculated for the HYD domain in both *holo* and *apo* systems is 0.8. These data highlight that the three replicas sample similar conformational subspace.

Table 2: RMSIP of the first three eigenvectors of the two metatrajectories and the single runs

Domain	**System**	**Comparison**	**RMSIP**
SYNTH		**Holo vs Apo**	0.9
	Holo	**run 1 vs run 2**	0.6
		run 1 vs run 3	0.7
		run 2 vs run 3	0.7
	Apo	**run 1 vs run 2**	0.6
		run 1 vs run 3	0.6
		run 2 vs run 3	0.6

Domain	**System**	**Comparison**	**RMSIP**
HYD		**Holo vs Apo**	0.6
	Holo	**run 1 vs run 2**	0.9
		run 1 vs run 3	0.8
		run 2 vs run 3	0.9
	Apo	**run 1 vs run 2**	0.9
		run 1 vs run 3	0.8
		run 2 vs run 3	0.9

3.2.2 GDP analysis

Cluster analysis

A second set of cluster analyses was performed on the GDP of the systems containing the ligand (Figure 30 and Table 3). The analyses were carried out considering the heavy atoms of the ligand (a new cluster is created if the RMSD of the new structure is > 1 Å), using the same clustering method of protein cluster analyses (hierarchical agglomerative, average linkage).

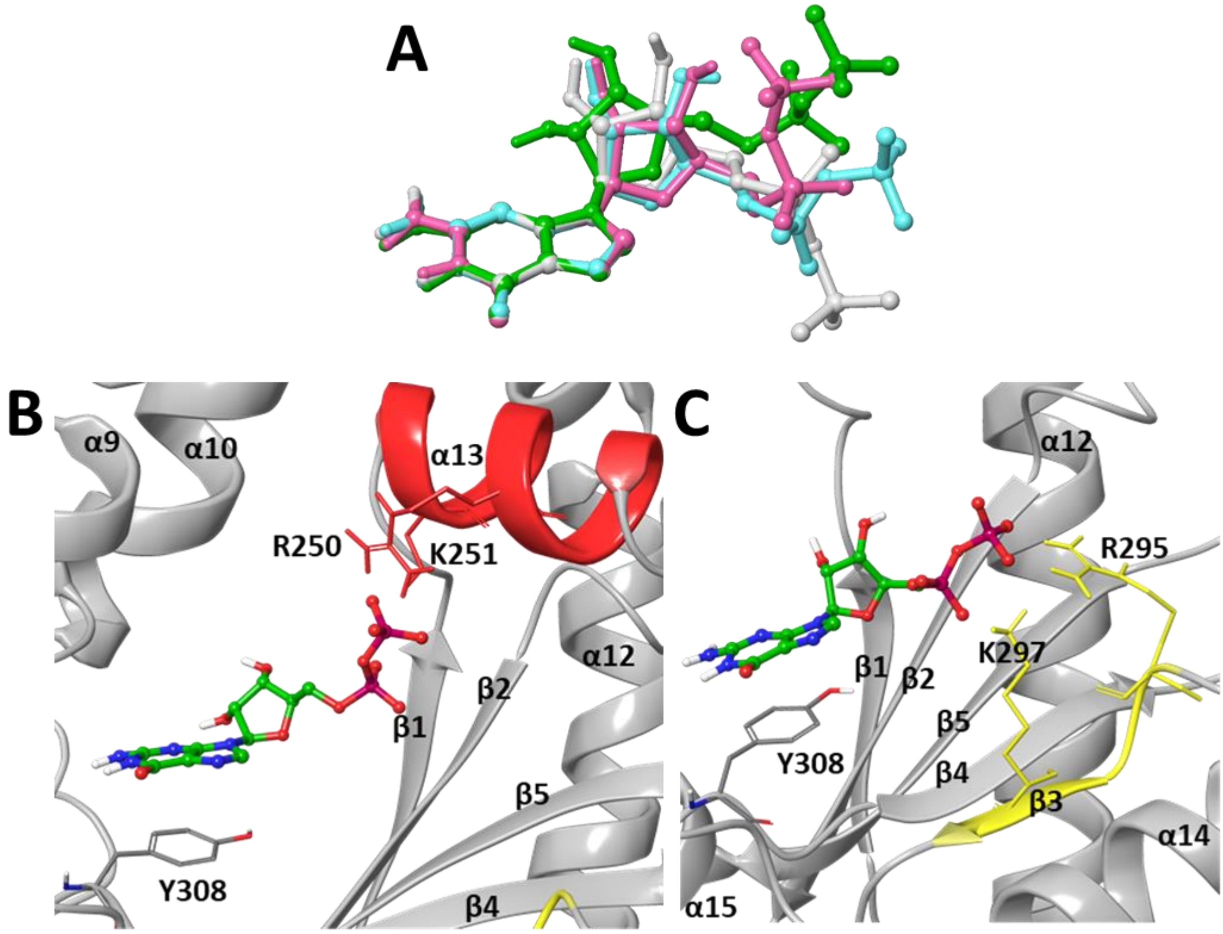


*Figure 30**:** A) Cluster analysis of holo system. The X-ray structure is shown in white, the main cluster is shown in green, the second cluster is shown in light-blue, the third cluster is shown in purple B) 3D representation of the main GDP binding mode (main cluster) into the SYNTH site of Rel_{Seq}* C) *3D representation of the alternative GDP binding mode (third cluster) into the SYNTH site of Rel_{Seq}. For both boxes b and c, GDP is shown in green carbon balls and sticks, Y308 is shown in grey carbon sticks, residues locate in helix 13 are shown in red sticks, residues locate in the beta sheet 3 are shown in yellow sticks, the protein is shown in grey ribbons, helix 13 is shown in red ribbon and beta sheet 3 is shown in yellow ribbons*

Table 3: Number of clusters and percentage of frames belonging to the first three clusters resulting from GDP cluster analysis

Cluster	Total number of clusters	% frames in cluster 1	% frames in cluster 2	% frames in cluster 3
GDP	18	26	20	15

As shown in Figure 30, according to the cluster analysis, whereas the guanine ring of GDP maintains the same interactions in the pocket during the simulations, phosphate groups rotate changing their interaction network and interacting with the residues located in helix 13 or with the ones located in β3 sheet (Figure 30b and c).

GDP interaction analysis

GDP-protein interactions were monitored during the simulation. The distances between GDP guanine ring atoms and the atoms of residues involved in the X-ray binding mode were monitored. As shown in Figure 31, the guanine ring maintained a distance within 4.5 Å to Y308, N306 and A335 residues, underlying the stability of the interactions between the ligand and these residues. On the other hand, a less stable interaction can be seen by the fluctuation of the distance between the ligand and K304. The phosphate groups interact with residues K243, R250, K251 (located in the α13), R295 and K297 (located in β3) as shown by Figure 32. The interaction network changes are probably due to the absence of the ATP (the second ligand essential for the reaction) and Mg^{2+} that could lock phosphate groups in an established position.

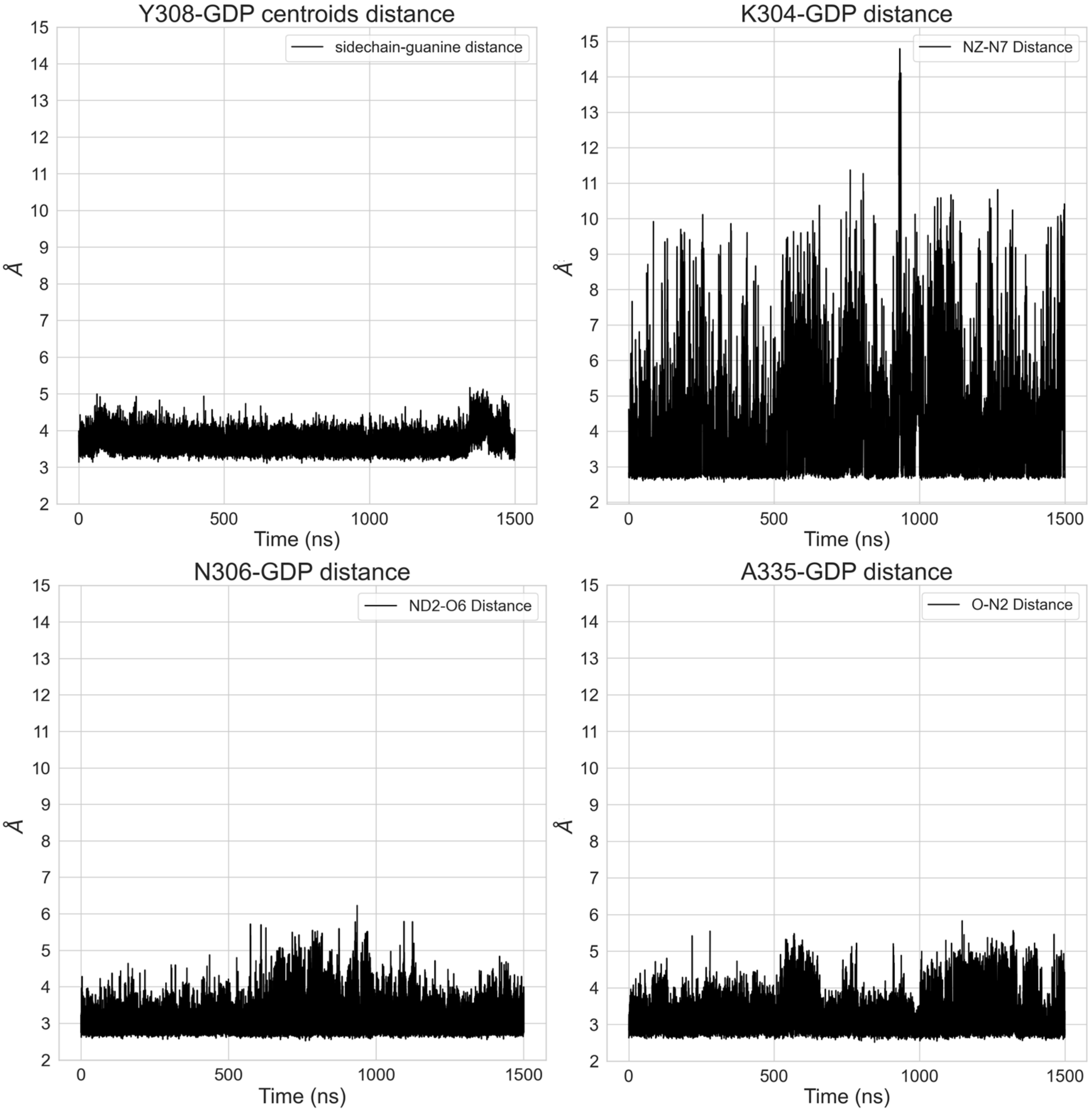


Figure 31:Distances calculated during simulation time between centroid of guanine ring and the centroid of Y308 side chain (top left), distances calculated during simulation time between atom N7 of GDP and NZ of K304 (top right), distances calculated during simulation time between atom O6 of GDP and ND2 of N306 (bottom left) and distances calculated during simulation time between atom N2 of GDP and O of A335 (bottom right)

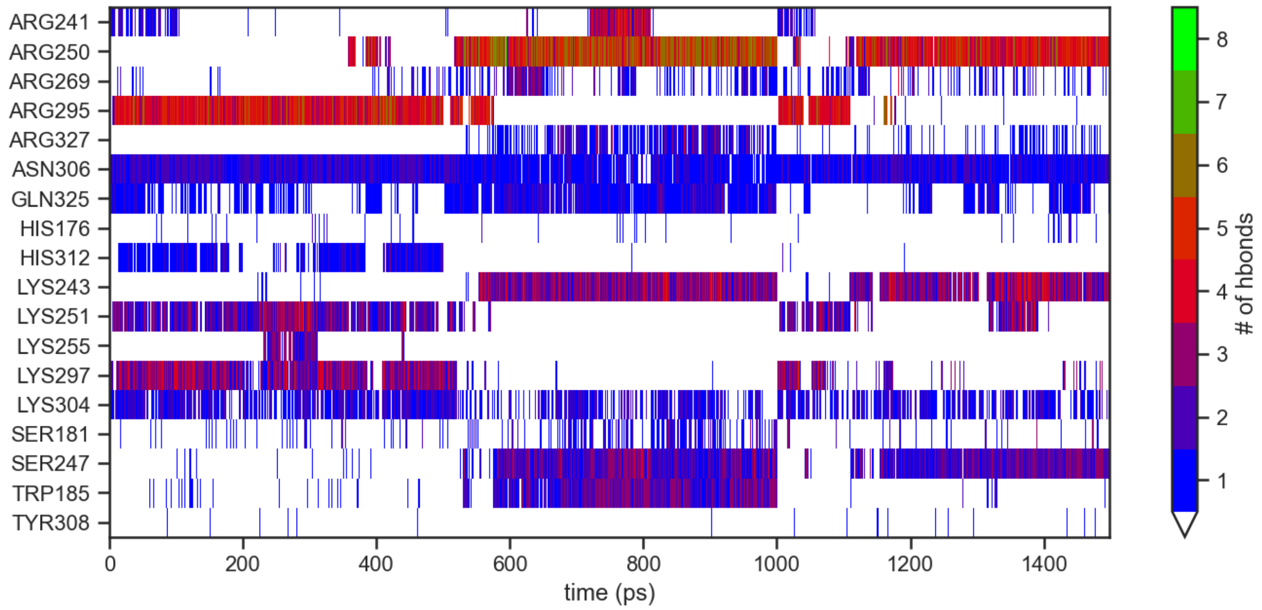


Figure 32: Time series of H-bonds and salt bridges formed by GDP during the MD simulations. H-bonds analysis was carried out by MDTraj[155] *according to the Baker-Hubbard*[156] *criterion, with a cutoff distance (A--D) of 3.5 Å and angle (D-HA) ≥ 120°*

3.3 Identification of potential SYNTH allosteric sites

In order to evaluate the allosteric communication between HYD and SYNTH pockets and to identify possible allosteric sites in the SYNTH site, Ohm software was used.[157]

According to the methodology implemented in Ohm, first, the average atom-contacts matrix is calculated from a 3D protein structure using equation 3.9

$$C_{ij} = \sum_{a,b} H\left(r_0 - \left|\overrightarrow{r_a^i} - \overrightarrow{r_b^j}\right|\right) \tag{3.9}$$

where C_{ij} is the number of atom contacts between residue i and residue j, a and b are the atoms of residues *i* and j, respectively, r_0 is the distance cutoff (default is 3.4), $\overrightarrow{r_a^i}$ and $\overrightarrow{r_b^j}$ are the position of atom *a* in residue *i* and the position of atom *b* in residue j, respectively, H is the Heaviside step function. If residue *i* and residue *j* are two consecutive residues, *a* and *b* cannot be backbone atoms simultaneously.

Then, the number of contacts between *i* and *j* is divided first by the number of atoms in residue i and then by the number of atoms in residue j (equation 3.10)

$$N_{ij} = \frac{C_{ij}}{C_i}, N_{ji} = \frac{C_{ij}}{C_j} \tag{3.10}$$

where N_{ij} is the average number of atom contacts of residue *i* with respect to residue *j*, N_{ji} is the average number of atom contacts of residue j with respect to residue i, C_{ij} is the number of contacts between residues i and j, C_i is the number of atoms in residue i and C_j is the number of atoms in residue j. If the number of residues belonging to i and j is different, N_{ij} and N_{ji} are different. From this matrix, the perturbation propagation probability matrix is calculated by equation 3.11

$$P_{ij} = 1 - p_{ij} = 1 - e^{-\alpha \cdot N_{ij}}, P_{ji} = 1 - p_{ji} = 1 - e^{-\alpha \cdot N_{ji}} \quad (3.11)$$

where P_{ij} is the probability that residue i propagates its perturbation to residue j, p_{ij} is the probability that residue i does not propagate its perturbation to residue j and α is a parameter, defined by the user, to amplify or reduce the probability of perturbation (default is 3.0).

The algorithm of perturbation propagation works by defining four vectors V, W, B and T. Vectors V, W and T are vectors of size N, where N is the number of residues in the target. On the other hand, vector B shows a different dimension since it collects all the neighbours of each residue in order to consider in B_i element all residues in contact with residue i. V elements can present a value of 0 or 1 if the residue connected to that element does not undergo to conformational changes or if it does, respectively. W and T vectors are instead built with all values equal to 0.

All the residues in the active site have the values of their element in vector W and V are set to 1. Then, taking as example the residue "n" in the active site, all the neighbours of this residues, based on B_n, are identified. If residues "m" is neighbour of "n", a random number "r", in the range between 0 and 1, is generated and if r < P_{nm}, V_m value is set to 1, otherwise to 0. Once V_m is assigned, W_m is set to 1. This process is performed for all the residues until all the elements in W have a value of 1. Then, the value of T_i is summed by 1 if V_i has a value of 1. When all the T elements are summed, V and W are cleared and the process is repeated 10000 times. Finally, T is normalized. The value of T_i at the end of the process, is the allosteric coupling intensity (ACI) of residue i with respect to residue n (a residue located in the active site). To identify the allosteric pathway, the propagation algorithm is adjusted: a stack S is constructed to collect each pathway that goes from the active site to the allosteric one. If n is a residue in the active site and m is a residue in the allosteric one, the propagation process is performed and, if the end of the path is m, the path is saved in S. The process is repeated 10000 times and all the paths in S are statistically evaluated. The most often identified path corresponds to the most probable allosteric pathway. The importance of a residue in the pathway is also calculated. If $\{p_i\}$ is the set of all the pathways containing residue a, and p_i represents the importance of the allosteric pathway i, pa is used to identify the importance of residue a in that pathway. The value of p_a is initially set to 0 and then it is updated for each pathway in $\{p_i\}$ collection by using equation 3.12

$$p_a = p_a + p_i - p_a * p_i \quad (3.12)$$

When the importance of all pathways is substituted into the equation, p_a, and so the importance of residue a, is defined.

Ohm can also calculate the allosteric hotspots. First, a distance matrix M(*i,j*) is generated by calculating all the distances between *i* and *j* residue pairs. A residue is set as neighbour of another residue if the distance between one of the atoms of the first residue and one of the atoms of the second one is maximum 4.5 Å. All the neighbours of residue *i*, G(*i*), are so identified. A vector D (direction) is then initialized with size N (number of residues). A value of -1 is assigned to D(*i*) if the considered residue presents the highest ACI with respect to its neighbours. Accordingly, all the residues that have a direction value equal to -1 represent allosteric hotspots.

Ohm was used for both *holo* and *apo* systems to identify allosteric sites and to define the allosteric pathway(s) connecting HYD to SYNTH sites. For what concern the allosteric sites calculation, as shown by Figure 33, the residues involved in GDP interaction and from those important for (p)ppGpp synthesis (i.e., R250, K251, R269, K304, N306, Y308, H312, E323, Q325, R327 and A335), were defined as the active site in SYNTH domain. As a result, the HYD pocket was identified as a potential allosteric site, as already known from published literature[62] and a little site behind the SYNTH pocket (between helix 12 and helix 14) was identified as a secondary potential allosteric site. Even though more analysis should be performed to confirm this data, both sites can be considered as allosteric sites to potentially block the synthetic activity of the target.

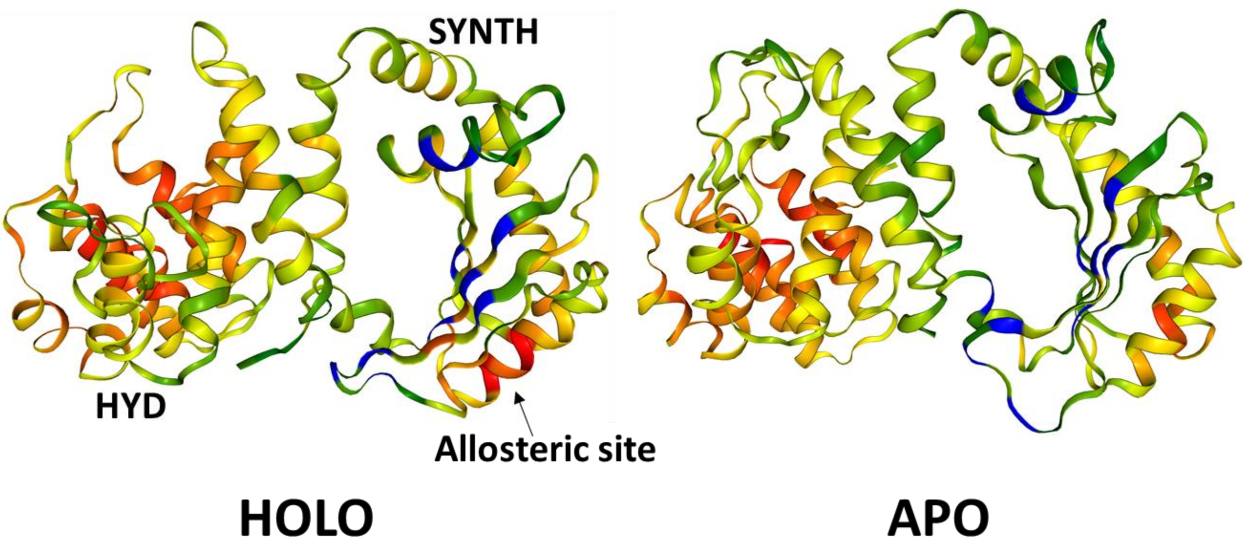


Figure 33:Allosteric hotspot identified by Ohm for holo (left) and apo (right) systems. Value of allosteric coupling intensity is rendered by a colour scale from blue (low value) to red (high value).

The pathways connecting SYNTH and HYD pockets were also evaluated considering residues T151 and Y308 for HYD and SYNTH sites, respectively. These two residues were chosen due to their involvement in the interaction with ligands.

In Figure 34 the most representative pathways connecting residues Y308 and T151 are shown. For both *holo* and *apo* systems, T151 is connected to Y308 by the propagation of the perturbation of the residues R150, E186 and W185. The rest of perturbation is instead different between the two systems due to the presence of GDP (residue 344) in *holo* system. W185 points towards guanine ring of GDP and probably it influences the position of the GDP into the SYTH site. Therefore, it is not so strange that this residue is present in both allosteric pathways.

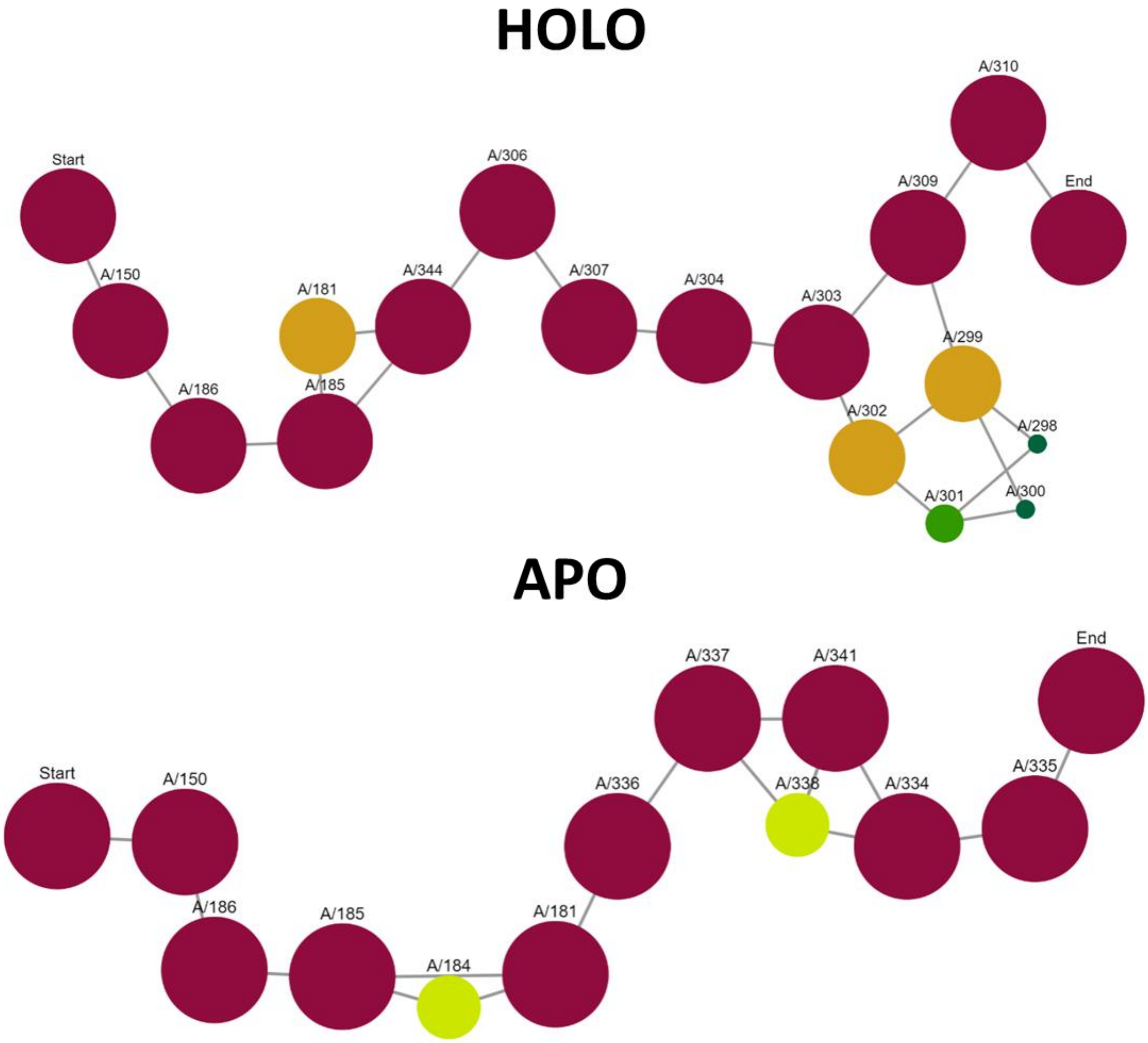


Figure 34: allosteric pathway connecting residue Y151 (Start) to residue Y308 (End) for holo (top panel) and apo (bottom panel) systems. The dimension of the spheres represents the importance of the residue for the allosteric connection.

3.4 Conclusion

The dynamic behaviour of Rel$_{Seq}$ was studied by performing MD simulations on the prepared chain A in its SYNTH *holo* and *apo* forms. As a result, the only appreciable difference among the two systems consisted in the SYNTH domain motions. Particularly, the G-loop involved in GDP interaction is more flexible in *apo* system due to the absence of the GDP that, if present, stabilizes its behaviour. Furthermore, GDP interactions were monitored during the simulations figuring out that the guanine ring maintained the same interactions shown in the X-ray structure, while the phosphate groups rotate to interact with residues located in helix 13 or β3 sheet.

Y308 is the only residue interacting with GDP that inhibits the synthetic activity of the enzyme if mutated. Therefore, the π-π stacking with Y308, that is maintained for the entire simulation, was defined as the key interaction to discriminate from potential ligand to non-ligand and this criterion was applied in the VS campaign described in chapter 4.

The HYD pocket was identified as an allosteric site of the SYNTH pocket, such as another site located between helix 12 and helix 14. Finally, residues R150, E186 and W185 represent the most important allosteric pathway connecting HYD to SYNTH sites.

3.4 Appendix

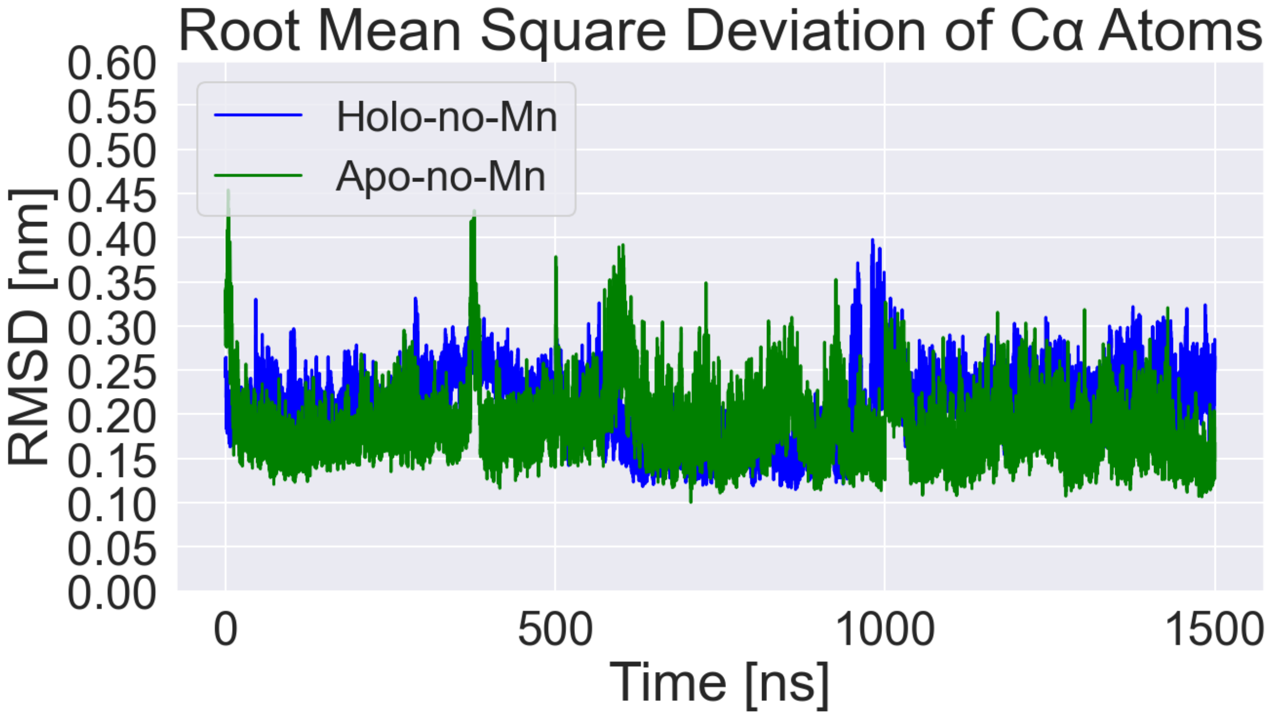


Figure 35: RMSD of Cα atoms of the entire holo (blue) and Apo (green) Rel_{Seq} without Mn^{2+}. The two systems reached a stable conformation during the MD syulations

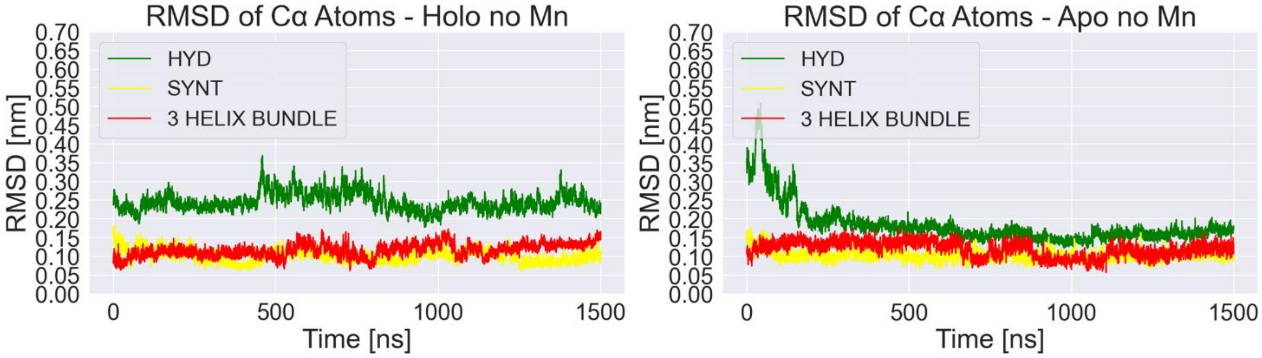


Figure 36: RMSD calculated on Cα atoms of single domains of holo no Mn (left) and apo no Mn (right) systems. HYD domain is shown in green, 3 helix bundle is shown in red, SYNTH domain is shown in yellow. The HYD domain is the one that caused the most part of the deviation also for the two systems lacking the ion.

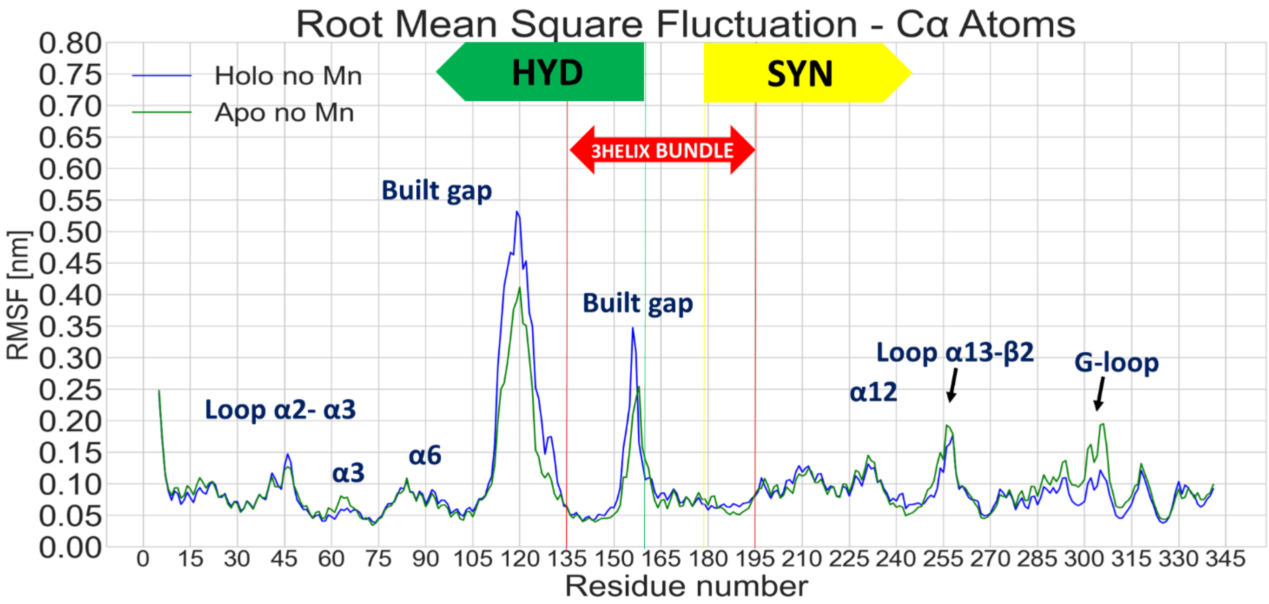


Figure 37: RMSF calculated on Cα atoms of holo no Mn system (blue) and apo no Mn system (green). The trend of the two systems is similar with the residues located in loop 110-123 that fluctuate the most. SYNTH domain fluctuate more in apo no Mn system than in holo no Mn one with residues located in the loop between α13 and β2 and the one located in the G-loop that fluctuate more.

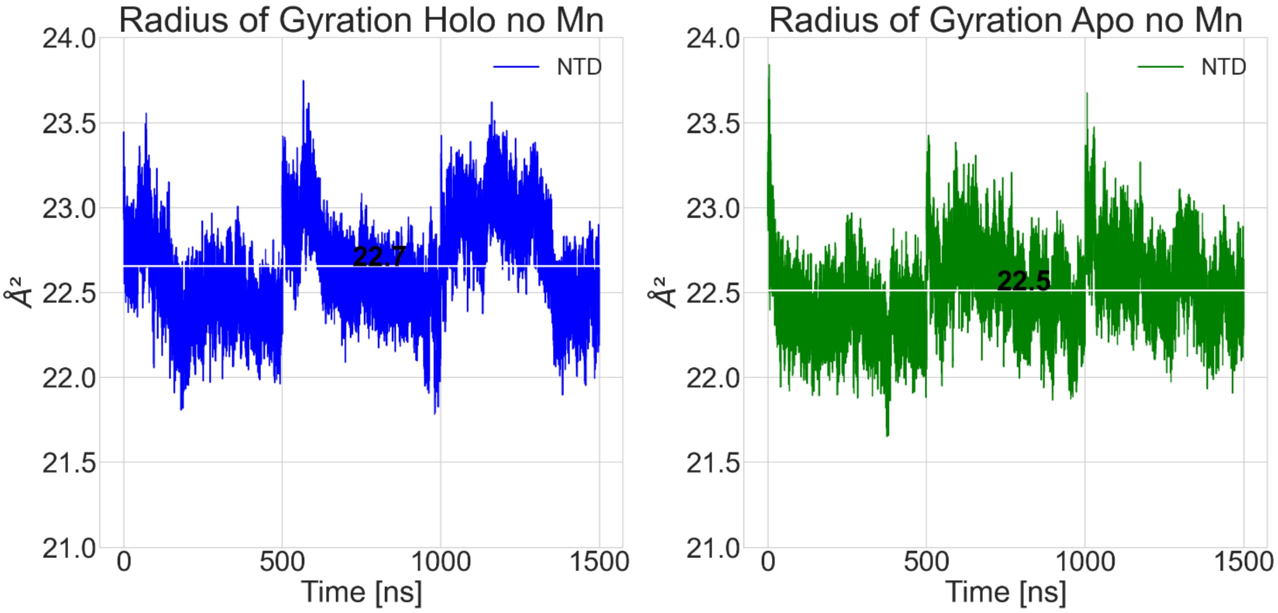


Figure 38: Radius of gyration values for backbone atoms of the entire systems. Holo no Mn system is represented in blue (left panel), apo no Mn system is represented in green (right panel). The average value of the RoG is represented in white lines and black numbers. SASA behaviour is similar among system with similar average

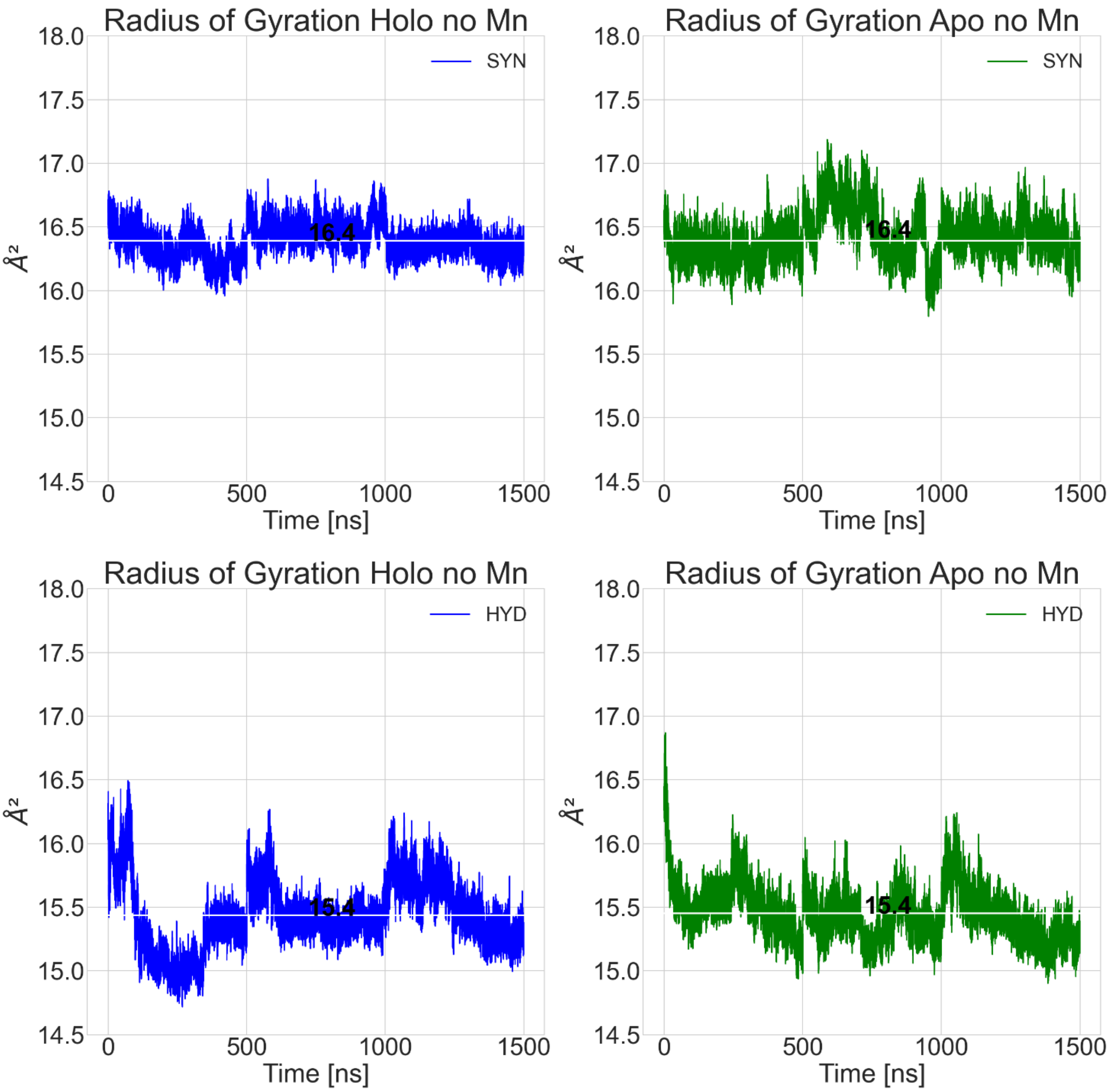


Figure 39: Graphics representing the radius of gyration values for backbone atoms of the SYNTH (top panel) and HYD (bottom panel) domains. Holo no Mn system is represented in blue, apo no Mn system is represented in green. The average value of the RoG is represented in white lines and black numbers. SASA behaviour is similar among system with identical average

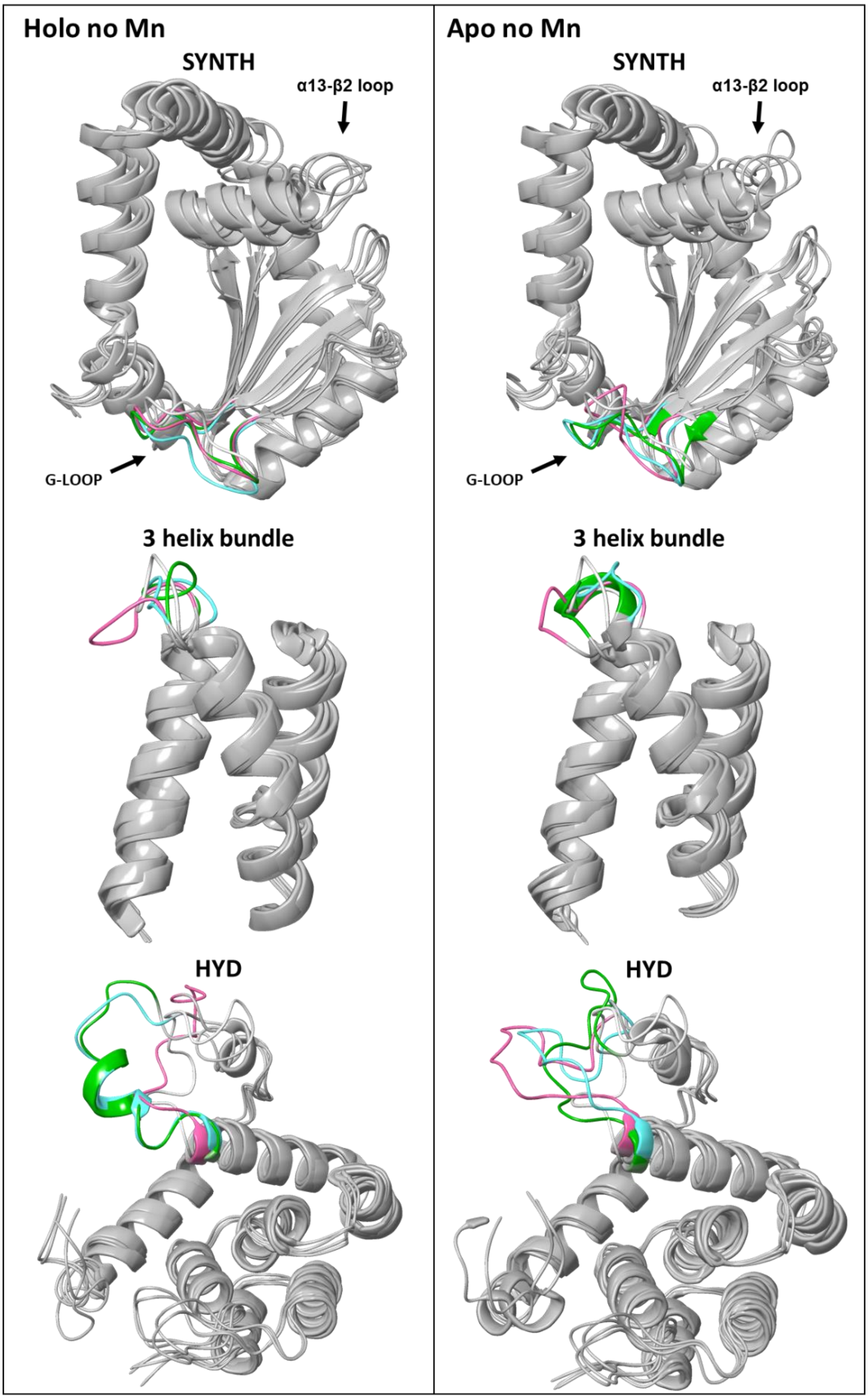


Figure 40: Cluster analysis performed on cα atoms of single domains. The protein is shown in grey ribbon, x-ray is shown in white, the main cluster is shown in green, cluster 2 is shown in purple and cluster 3 is shown in light blue. SYNTH domain showed the highest difference in 3D conformation between clusters with the loop between α13 and β2 the one that changed most. G-loop of apo system is more flexible in apo system than in holo one. 3 helix bundle and HYD domain cluster differences are due to the change in 3D conformation of the gaps we built

Table 4: Number of clusters and percentage of frames belonging to the first three clusters

System	Domain	Total number of clusters	% frames in cluster 1	% frames in cluster 2	% frames in cluster 3
Holo no Mn	**HYD**	422	10	8	5
	HYD no loops	10	62	24	6
	3 helix bundle	7	49	28	17
	SYNTH	32	16	16	15
Apo no Mn	**HYD**	373	10	6	4
	HYD no loops	10	79	12	4
	3 helix bundle	8	40	24	24
	SYNTH	139	23	5	5

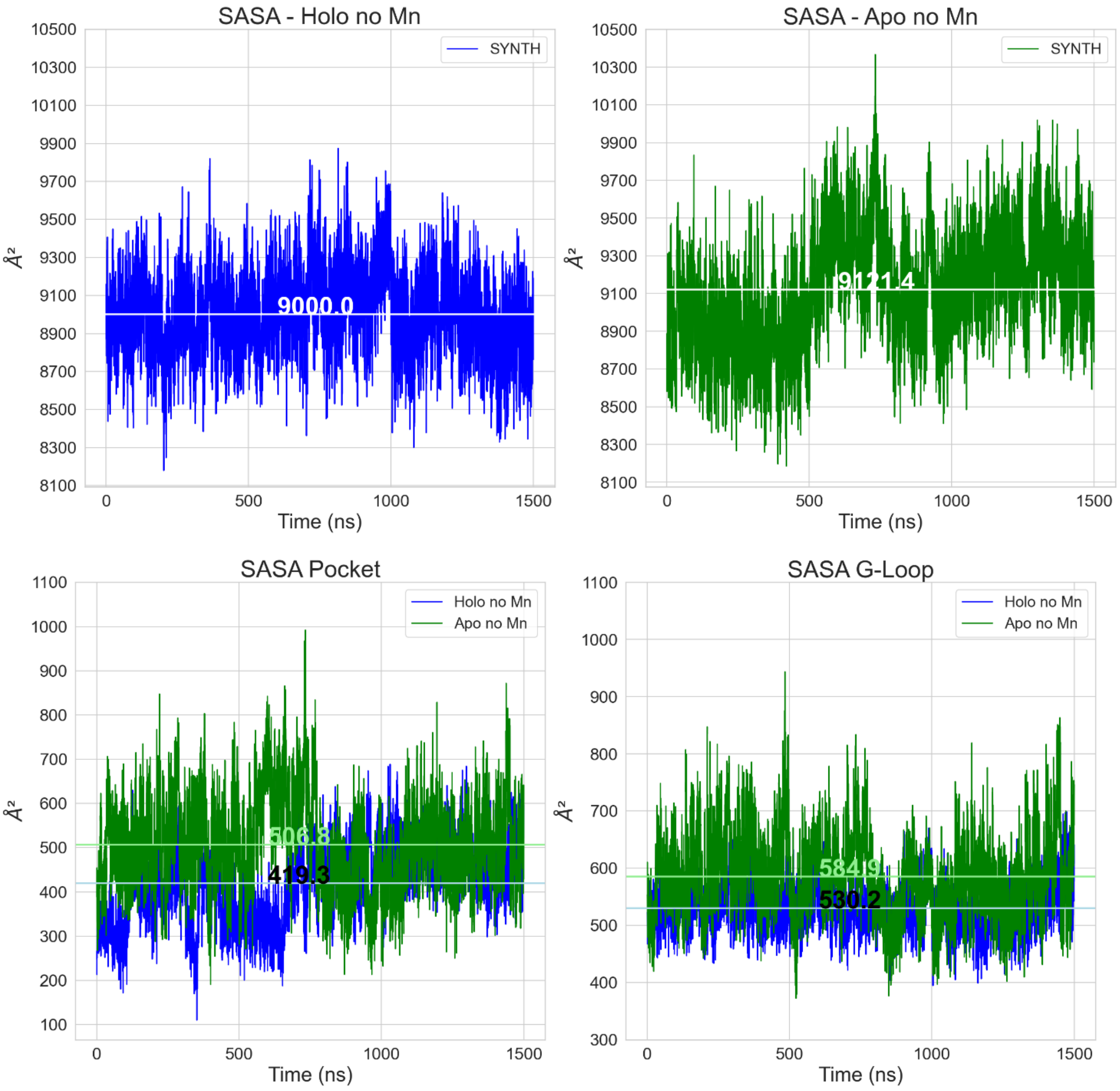


Figure 41: In the graphs is expressed the total Surface Area (in Å^2) vs time (in ns) of systems holo no Mn (top left), apo no Mn (top right), SASA of residues within 5 Å from GDP (bottom left) and SASA of G-loop residues (bottom right). *Holo no Mn SASAs are shown in blue, apo no Mn SASAs are shown in green. Average SASAs of holo no Mn system is shown in white line and number for the graphic representing the SYNTH domain and in light blue line and black number for the other two graphics, SASAs of apo no Mn system is shown in white line and number for the graphic representing the SYNTH domain and in salmon line and light grey number number for the other two graphics.*

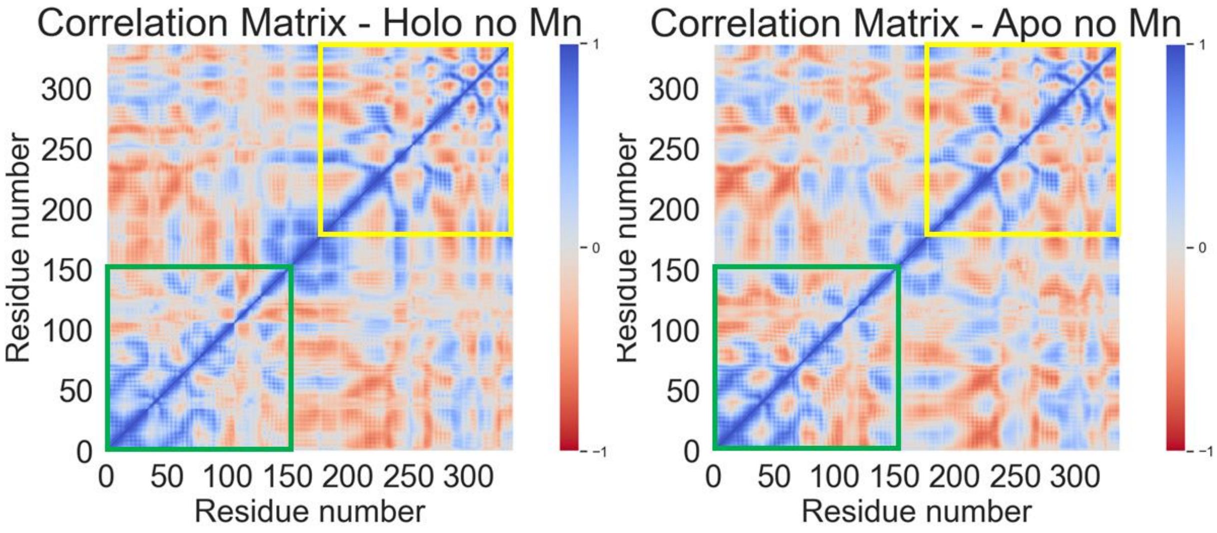


Figure 42: Correlation matrices of holo no Mn (left) and apo no Mn (right) systems. The HYD site is defined by a green box, the SYNTH site is defined by a yellow box, The two systems shows similar correlations trend, with apo no Mn system showing greater correlation values compared to the holo no Mn one, with residues located in the same domain having correlated motions while some residues located in two different domains having anticorrelated motion. As for the systems containing the ion, residues located in helix 12 (residue between 217-231) and loops of beta sheet 2 and beta sheet 2 itself (residue between 265-271), showed more compactness and anticorrelation with residues located in in alpha 2 (22-38) and alpha 3 - alpha 4 (50-75).

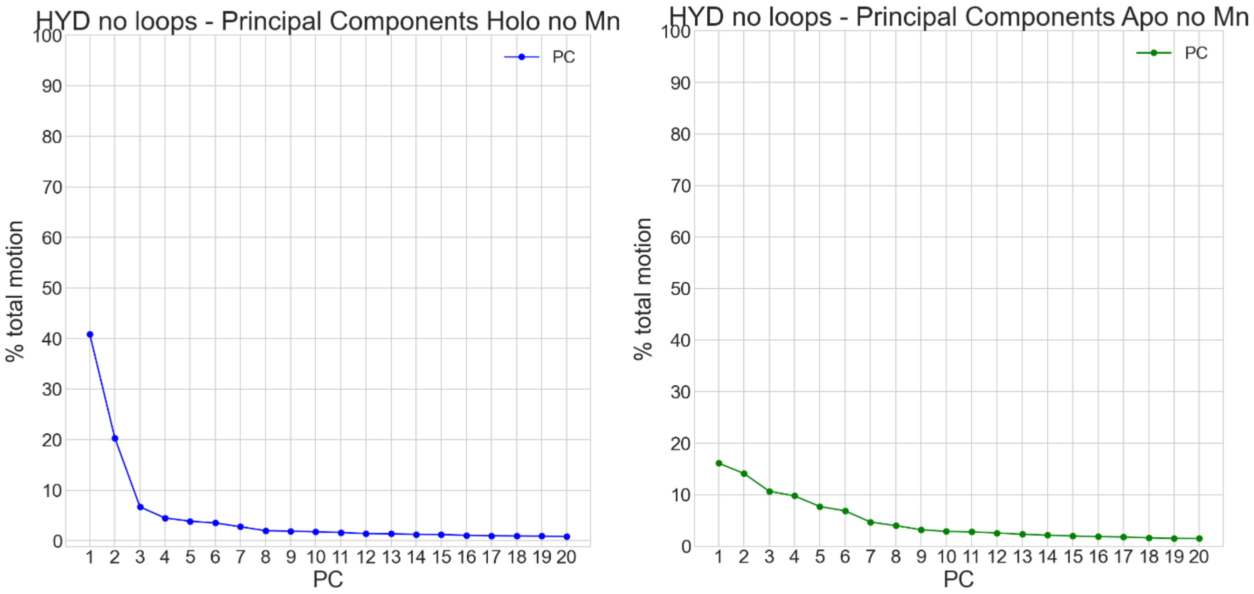


Figure 43: Weight of 20 Principal Components motion on the overall motion in percentage of the HYD domain. Holo no Mn system PCs are shown in blue in the left panel. Apo no Mn system PCs are shown in green in the right panel. The first three PC of the holo no Mn system describe the 68% of the total motion. The first three PC of the apo no Mn system describe the 49% of the total motion

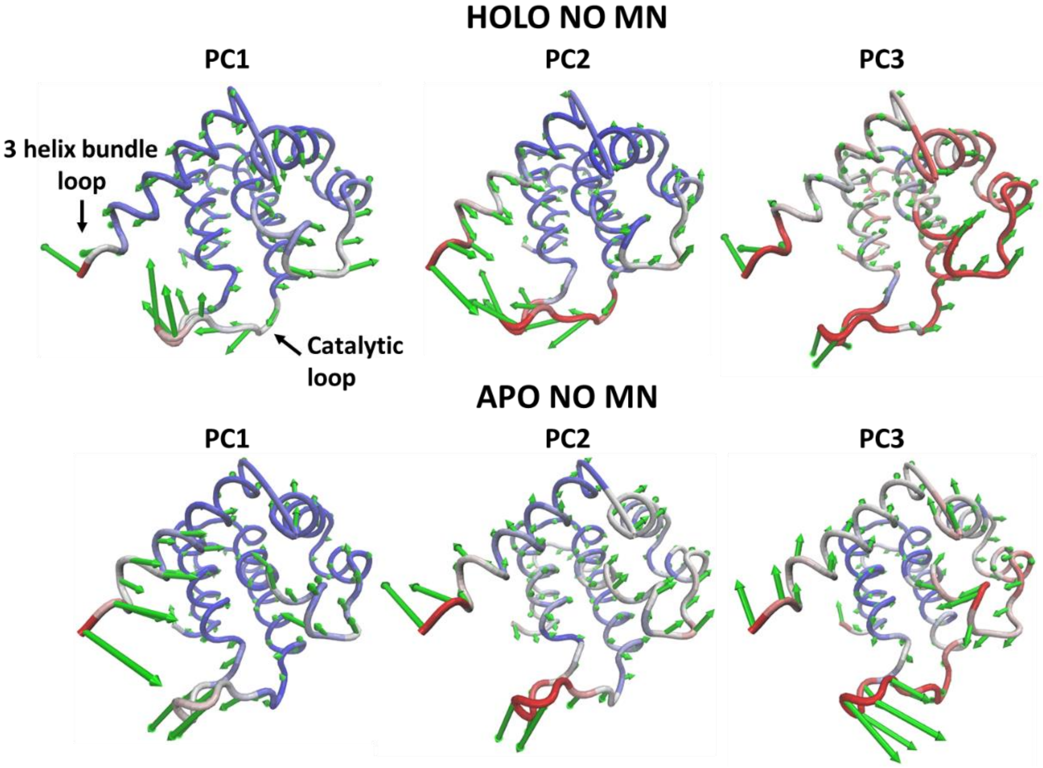


Figure 44: 3D representation of the first three principal components of HYD domain of holo no Mn (top panel) and apo no Mn (bottom panel) systems. The eigenvectors are shown by green arrows, the eigenvalues are shown by the length of the arrow and by the colour code: red high eigenvalue, white middle eigenvalue, blue low eigenvalue. The three main PCs involve principally motions of the catalytic loop and residues located near rhe 3 helix bundle loop we built.

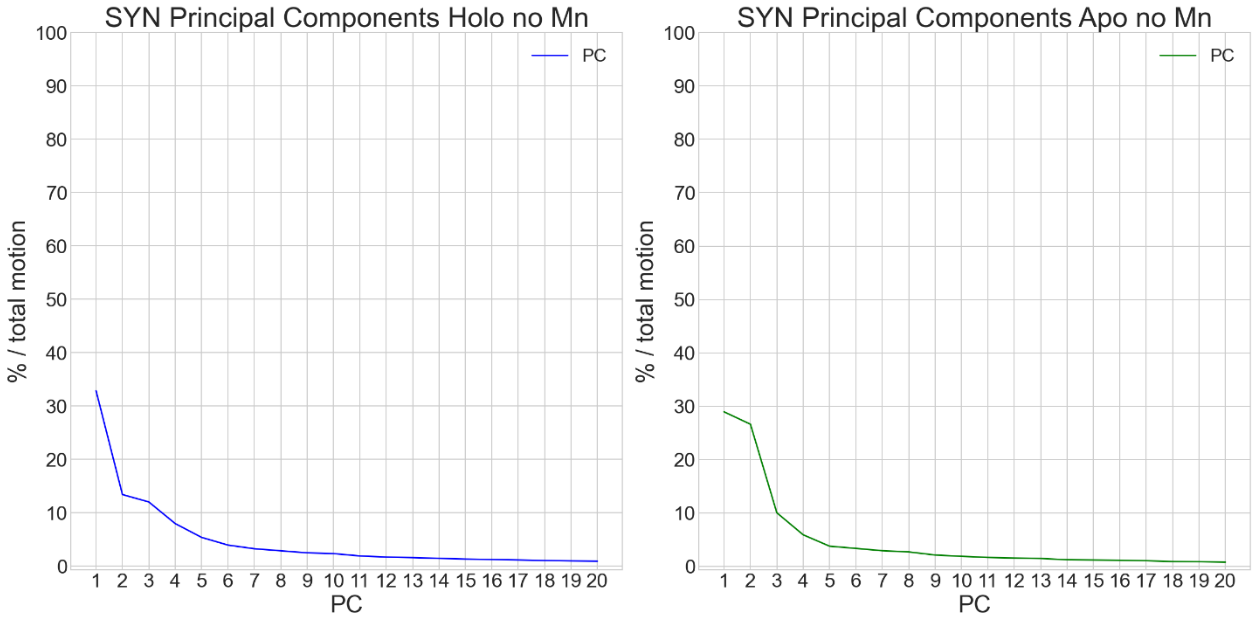


Figure 45: Weight of 20 Principal Components motion on the overall motion in percentage of the SYNTH domain. Holo no Mn system PCs are shown in blue in the left panel. Apo no Mn system PCs are shown in green in the right panel. The first three PC of the holo no Mn system describe the 58% of the total motion. The first three PC of the apo no Mn system describe the 66% of the total motion

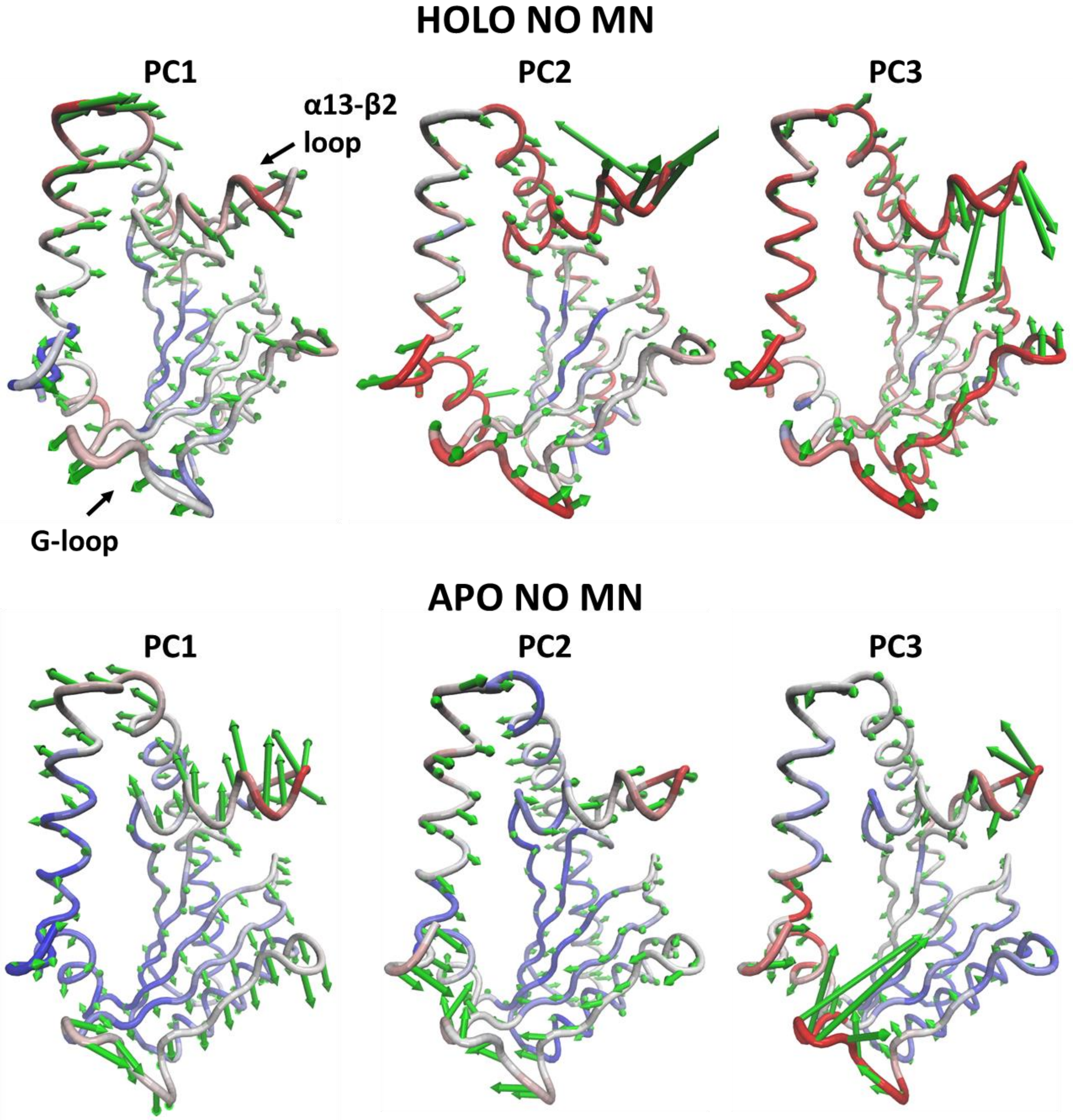


Figure 46: 3D representation of the first three principal components of SYNTH domain of holo no Mn (top panel) and Apo no Mn (bottom panel) systems. The eigenvectors are shown by green arrows, the eigenvalues are shown by the length of the arrow and by the colour code: red high eigenvalue, white middle eigenvalue, blue low eigenvalue. The three main PCs involve a twist of the SYNTH site wih motion of the G-loop and loop between α13 and β2.

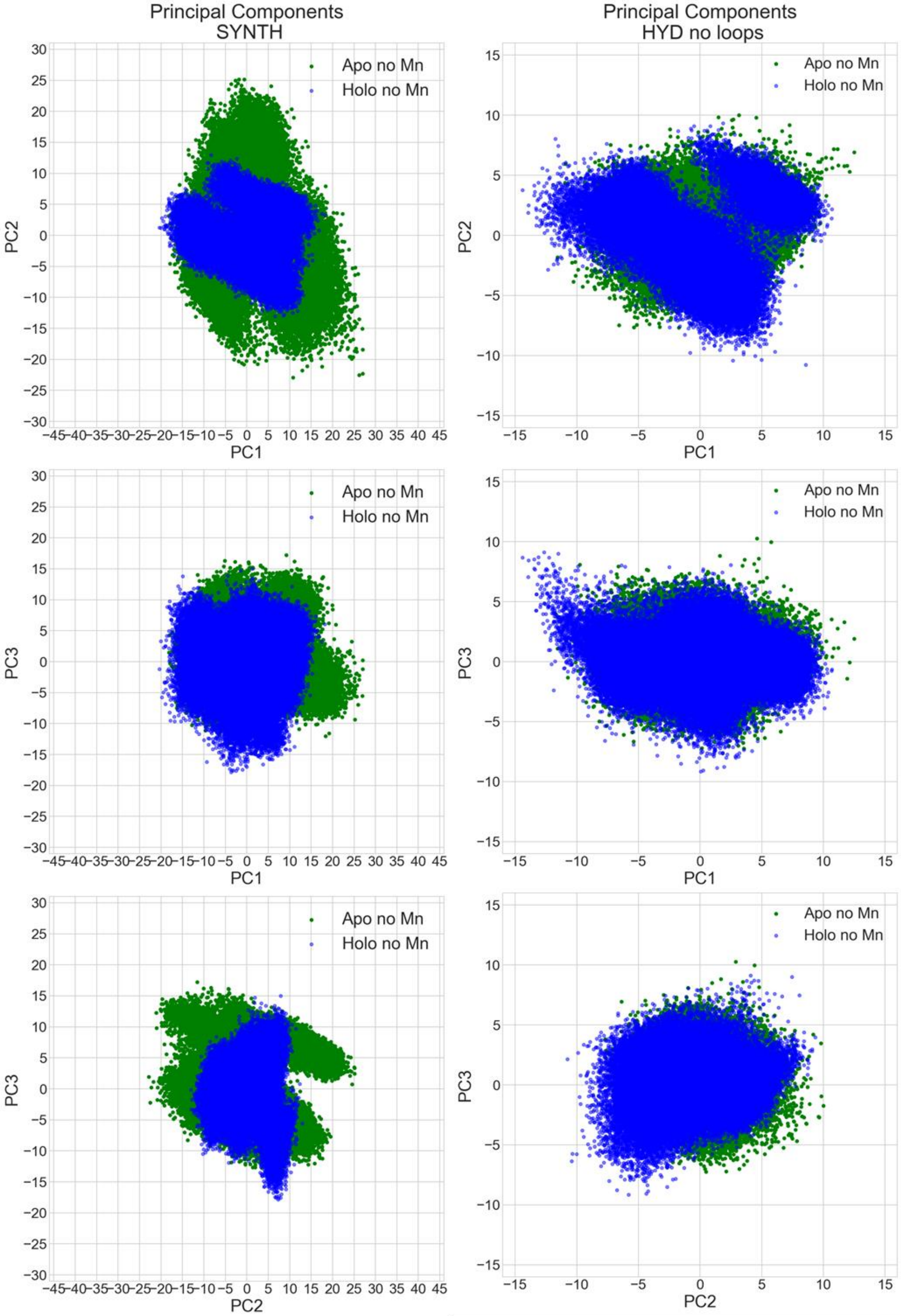


Figure 47: Scatter plots representing PC1 vs PC2 (top panel), PC1 vs PC3 (middle panel), PC2 vs PC3 (bottom panel) of SYNTH (left) and HYD domain (right). Holo no Mn system is represented in blue, Apo no Mn system is represented in green. For the SYNTH domain (left panels), PCs explore a higher range of values in the apo no Mn system compared to the holo no Mn one, on the other hand in the HYD domain (right panels) the exploration is similar between systems.

Table 5: RMSIP of the first three eigenvectors of the two metatrajectories and the single runs

Domain	System	Comparison	RMSIP	Domain	System	Comparison	RMSIP
SYNTH		Holo no Mn vs Apo no Mn	0.7	HYD		Holo no Mn vs Apo no Mn	0.9
	Holo no Mn	run 1 vs run 2	0.6		Holo no Mn	run 1 vs run 2	0.7
		run 1 vs run 3	0.6			run 1 vs run 3	0.7
		run 2 vs run 3	0.6			run 2 vs run 3	0.8
	Apo no Mn	run 1 vs run 2	0.8		Apo no Mn	run 1 vs run 2	0.9
		run 1 vs run 3	0.8			run 1 vs run 3	0.9
		run 2 vs run 3	0.7			run 2 vs run 3	0.8

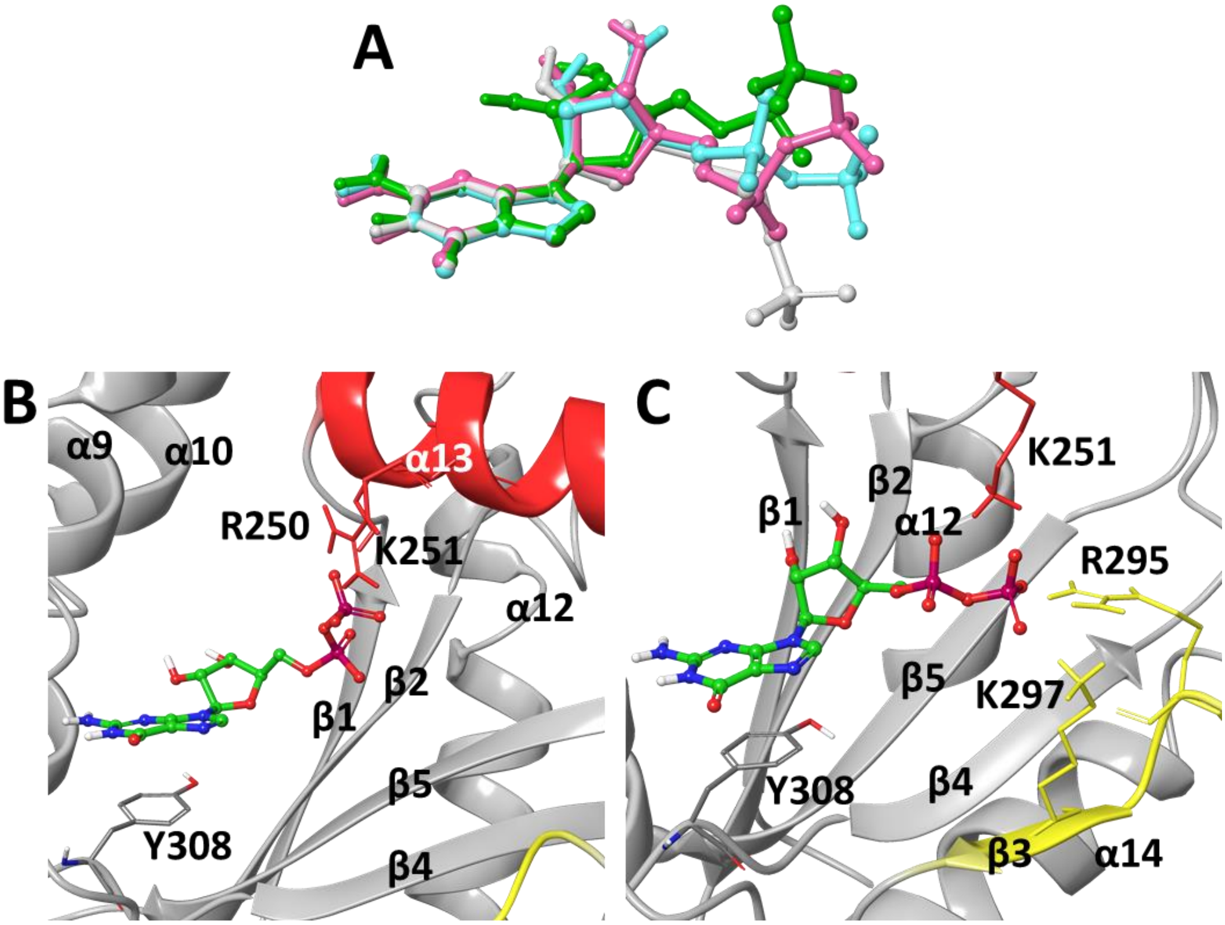


Figure 48: A) Cluster analysis of holo no Mn system. The X-ray structure is shown in white, the main cluster is shown in green, the second cluster is shown in light-blue, third cluster is shown in purple B) 3D representation of the main GDP binding mode (main cluster) into the SYNTH site of Rel_{Seq} C) 3D representation of the alternative GDP binding mode (third cluster) into the SYNTH site of Rel_{Seq}. For both boxes b and c, GDP is shown in green carbon balls and sticks, Y308 is shown in grey carbon sticks, residues locate in helix 13 are shown in red sticks, residues locate in the beta sheet 3 are shown in yellow sticks, the protein is shown in grey ribbons, helix 13 is shown in red ribbon and beta sheet 3 is shown in yellow ribbons

Table 6: Number of clusters and percentage of frames belonging to the first three cluster resulting from GDP cluster analysis performed on holo no Mn system.

Cluster	Total number of clusters	% frames in cluster 1	% frames in cluster 2	% frames in cluster 3
GDP	26	19	16	13

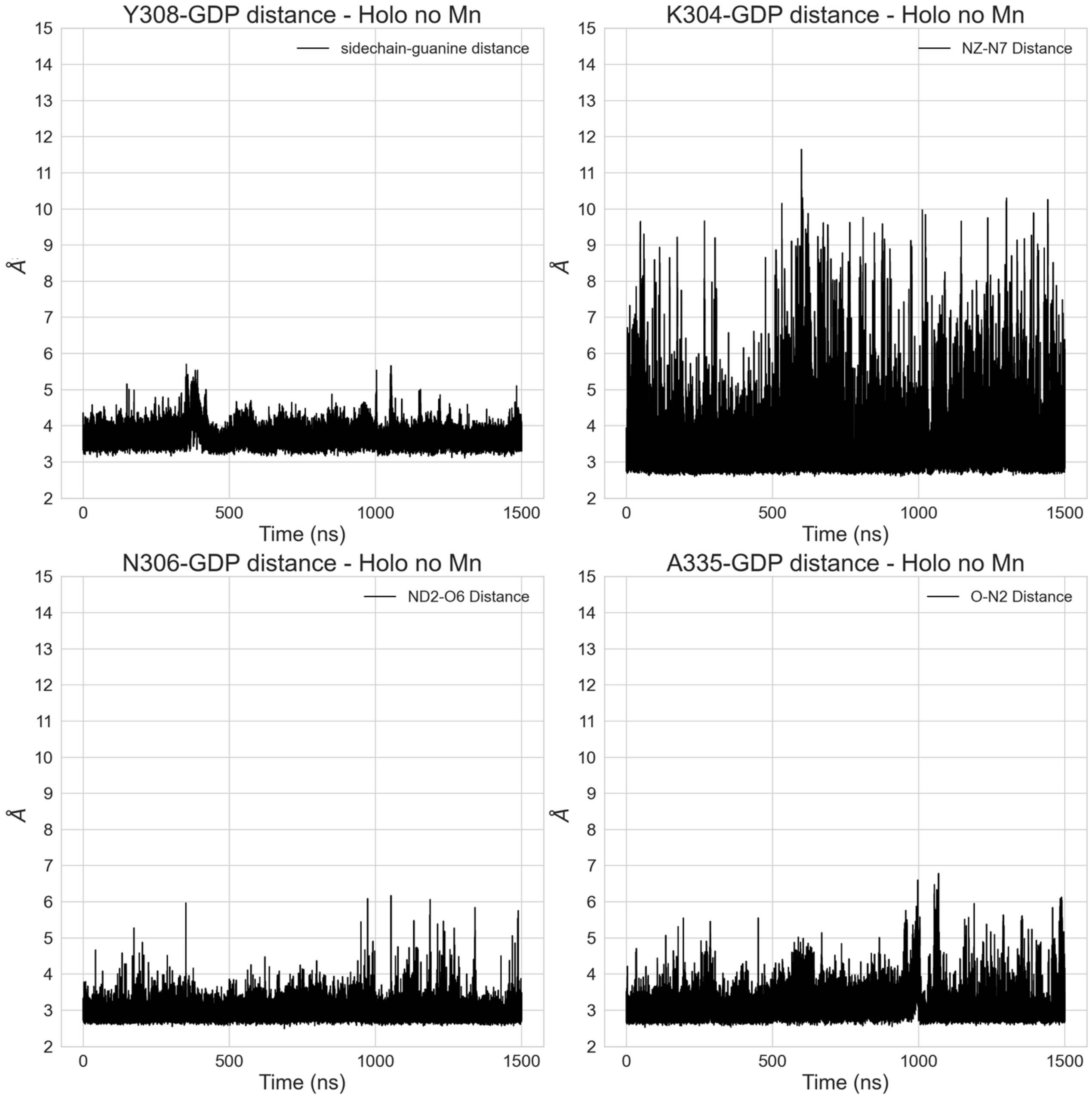


Figure 49: Distances calculated during simulation time between centroid of guanine ring and the centroid of Y308 side chain (top left), distances calculated during simulation time between atom N7 of GDP and NZ of K304 (top right),), distances calculated during simulation time between atom O6 of GDP and ND2 of N306 (bottom left) and distances calculated during simulation time between atom N2 of GDP and O of A335 (bottom right). The four interactions formed during X-ray crystallization are maintained during all the simulation with the one formed between GDP N7 atom and K304 side chain being the less stable.

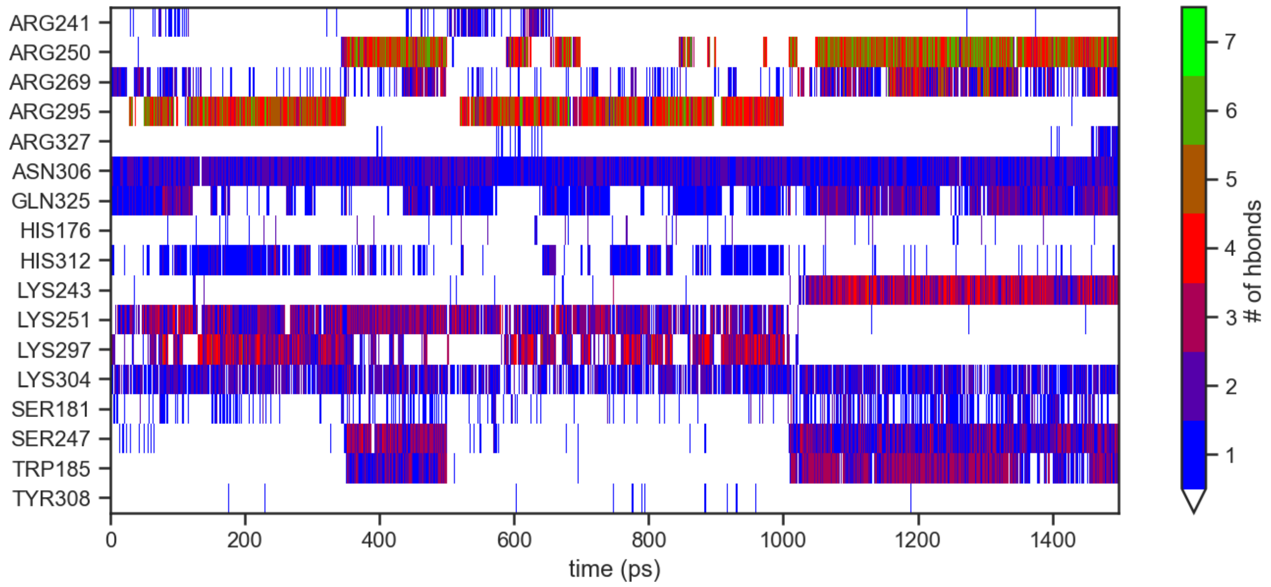


Figure 50: Time series of H-bonds and salt bridges formed by GDP during the MD simulations. GDP interacts principally with residues located in α13 (K243, R250, K251) or with the one located in β3 R295, K297. H-bonds analysis was carried out by MDTraj according to the Baker-Hubbard criterion, with a cutoff distance (A--D) of 3.5 Å and angle (D-HA) ≥ 120°.

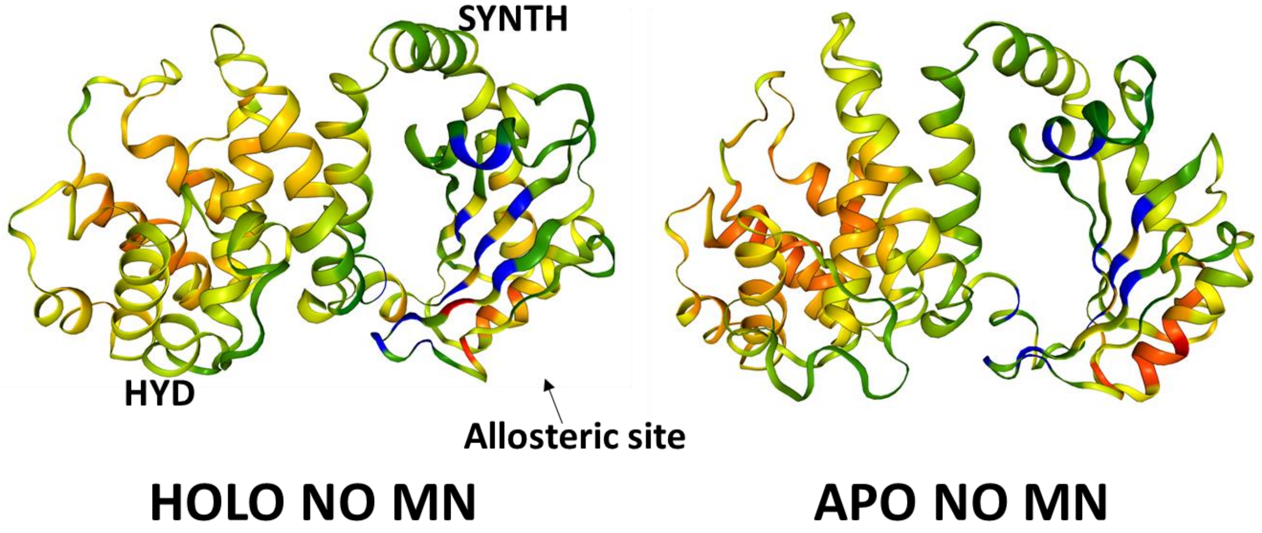


Figure 51:Allosteric hotspot identified by Ohm for holo no Mn (left) and apo No Mn (right) systems. Value of allosteric coupling intensity is rendered by a colour scale from blue (low value) to red (high value). Selecting the amino acids located in the SYNTH site and listed in paragraph 3.3, the allosteric sites of the SYNTH pocket identified for these two systems are the same found for the holo and apo systems i.e. the HYD site and a pocket located between helix 12 and helix 14.

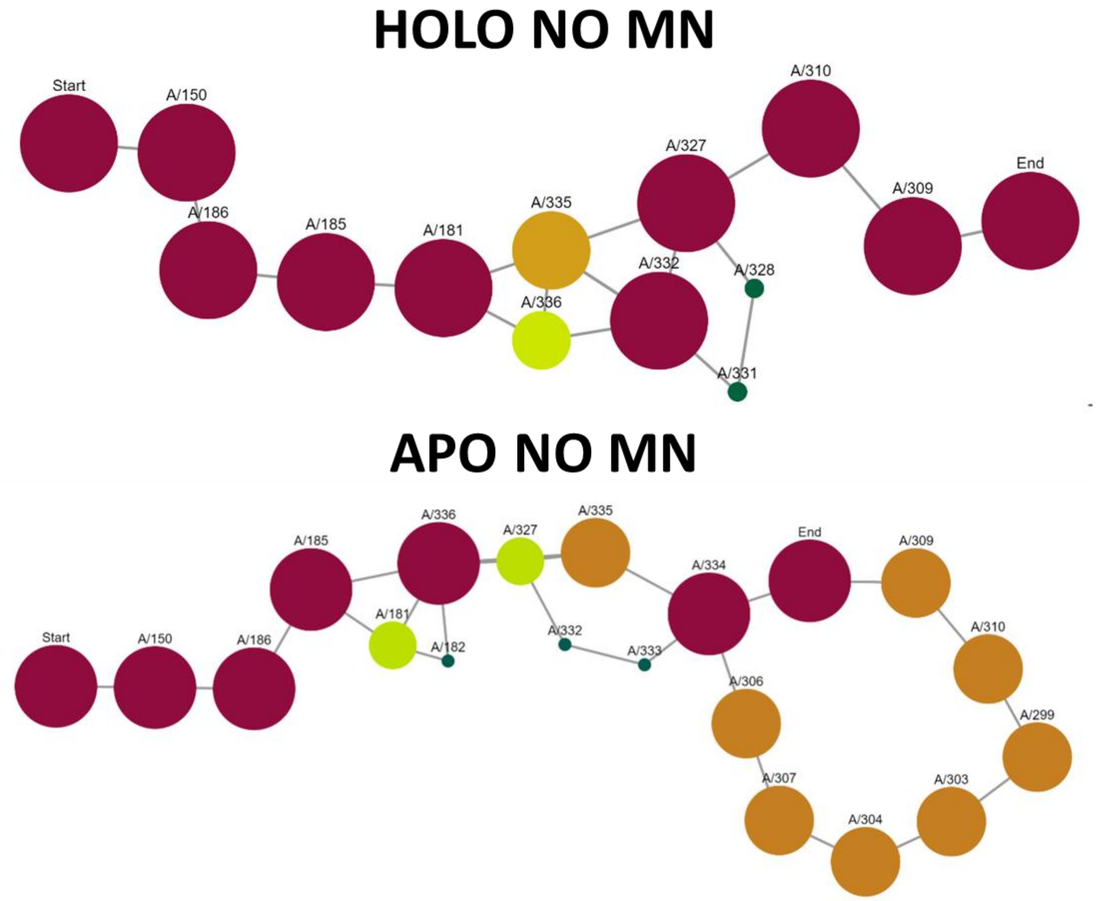


Figure 52: allosteric pathway connecting residue Y151 (Start) to residue Y308 (End) for holo no Mn (top panel) and apo no Mn (bottom panel) systems. The dimension of the spheres represents the importance of the residue for the allosteric connection. The pathways connecting T151 to Y308 start, as for holo and apo systems, with the perturbation of residues R150, E186 and W185, pointing out the importance of these residues for the communication between the HYD and SYNTH pockets.

CHAPTER 4: FRAGMENT BASED VIRTUAL SCREENING INTO THE SYNTHETASE SITE

The fragment-based drug design (FBDD) workflow applied for the Virtual Sreening (VS) of prepared libraries and based on docking calculation is shown in Figure 53.

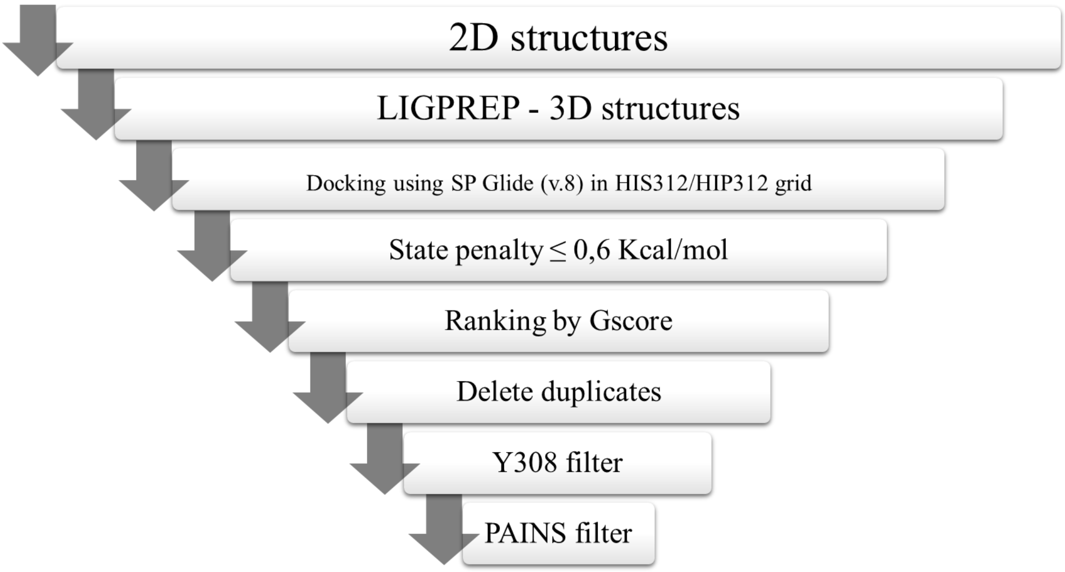


Figure 53: VS workflow

Ligprep[139] was used to convert fragments to their 3D structures and to generate tautomeric and ionization states then the fragments were evaluated into the synthetase site of the Rel$_{Seq}$. Less stable tautomeric and ionization forms, duplicates and PAINS were deleted. Finally poses that do not form an aromatic contact with residue Y308 were excluded from the calculations. From the resulting fragments, the most representative chemotypes were used for a library expansion to expand the chemical space explored. Then the most promising fragments selected from VS and library expansion workflow were used in MD simulations and thermal shift assays (TSA). The chemotype resulted the most selective for the SYNTH domain was then used to design new compounds that were tested in both MD simulations and TSA.

4.1 Fragment libraries and their preparation

FBDD is a valuable alternative method in drug discovery compared to the traditional high-throughput screening (HTS). The main difference between HTS and FBDD consists in the composition and the size of the libraries. Whereas the former identifies mature chemical starting point for hit-to-lead, the second incrementally constructs the lead, by starting from less potent items with equivalent or better ligand efficiency (LE) which express the docking score weighed by the heavy atoms. These second structures can be combined if they bind to different part of the target pocket, and they can be easily modified in order to improve binding affinity and adjust ADME properties. Generally, the number of false negative structures is higher for HTS strategies compared to FBDD ones. For these reasons fragments libraries were preferred to big databases.

Seven fragment libraries from six vendors were selected for the VS campaign in order to maximize the chemical space explored: the 'Maybridge Ro3 Diversity Set' (https://www.maybridge.com), the 'Asinex-Fragments-21872 library' (http://www.asinex.com), two libraries from Life chemicals, the 'Fragment Libraries with Experimental Solubility Data' (https://lifechemicals.com), the 'OTAVA Solubility Fragment Library' (https://www.otavachemicals.com), the 'FragmentLibrary_sdf_13808' from CHEMBRIDGE (https://www.chembridge.com) and the 'Preplated Fragment-Based Library' from SPECS (https://www.specs.net).

Libraries were downloaded from January to February 2019. A brief description of each library is reported below.

MAYBRIDGE Ro3 Diversity Set:

Fragments contained in this library respect the Ro3, their solubility was experimentally measured in PBS buffer (1 mM), they were optimized for surface plasmon resonance (SPR), they are PAINS free and clean from toxic and reactive groups. Moreover, fragments in this database present an exceptional diversity (Tanimoto similarity index of 0.66 based on standard Daylight fingerprinting[158]) and pharmacophoric enrichment.

ASINEX Asinex-Fragments-21872:

The database presents high diversity of scaffolds and synthetic handles moreover, together with the classic fragments used for fragment screening, saturated fused ring, spiro, bridged systems and macrocycles with a tendency towards multiple chiral centres are added to the database due to their use in marketed drug.

Life Chemical Fragment Library with Experimental Solubility libraries:

Both libraries contain fragments with the following specifications:

- Average molecular weight (MW) 240 Da and logP 1.2
- Guaranteed solubility of all compounds in DMSO at high concentration (200 mM)
- 77 % of the library are soluble in water phosphate buffer at 1 mM, and 60 % at 5 mM

Fragments show characteristics in the range listed below:

Table 7: Specification of fragments belonging to Life Chemical databases

Parameter	Range
MW	100 to 300
ClogP	–2 to 3
TPSA	< 100 Å^2
H-Acceptors	≤ 4
H-Donors	≤ 3
Rotatable Bonds	≤ 3

OTAVA Solubility Fragment Library:

It consists of 1021 low molecular weight fragments with Ro3 compliance and assumed solubility in both DMSO (200 mM) and PBS (1 mM). Moreover, the use of different filters removes from the library:

- compounds containing any atom different to O, N, C, H, Br, I, Cl, F, S, or P
- compounds that do not contain at least one aliphatic or aromatic ring.
- compounds containing more than 4 halogen atoms
- compounds containing reactive functional groups bearing the risk of covalent binding to the target protein
- Reactive molecules, PAINS, redox-active molecules, aggregator compounds were removed from the library

Finally, compounds present the following characteristics:

Table 8: Specification of fragments belonging to OTAVA library

Parameter	Value	Average
MW	< 300	197.0
CLogP	< 3	1.52
Number of Rotatable Bonds	≤ 3	1.6
Number of H-Donors	≤ 3	1.04
Number of H-Acceptors	≤ 4	2.56
PSA	< 80	49.9
Number of Rings	≥1	1.83
Experimentally Assured Aqueous Solubility	≥ 1 mM	
Solubility in DMSO	≥ 200 mM	
Diversity based on fingerprint distances	0.89	
Number of clusters	302	
Number of singletons	180	
Sum of Halogen Atoms	≤ 4	

Chembridge FragmentLibrary_sdf_13808 library:

Approximately half of the compounds meet the criteria for the High Solubility Subset (HSS). These fragments have a minimum DMSO solubility of 200mM and minimum solubility in PBS (pH 7.4) of 200µM (with many fragments soluble at 1mM in PBS). The remaining fragments have DMSO solubility of less than 200mM or PBS solubility of less than 200 µM.

Fragments comply with the Ro3 parameters (MW ≤ 300, H-bond donors ≤ 3, H-bond acceptors ≤ 3, cLogP ≤ 3) along with rotatable bond count and calculated TPSA (topological polar surface area) limits

Finally, they present these characteristics:

Table 9: Specification of fragments belonging to Chembridge library

Parameter	ChemBridge Cutoff Values	Average Property Values
MW	150-300	225.06
H-donors	≤ 3	1.22
H-acceptors	≤ 3	2.33
cLogP	≤ 3	1.27
Rotatable Bonds	≤ 3	2.01
cLogSw	≥ -2.50	-1.68
TPSA	≤ 120	49.66

SPECS Preplated Fragment-Based Library:

This library of 4532 fragments was thought to perform High throughput screening *in vitro*. Fragments contained in this library respect the Ro3 and are soluble in DMSO.

Fragments of libraries were prepared and converted into their three dimensional (3D) structure by using the 'Ligprep' utility.[139] Tautomers and at most 32 stereoisomers (default number used when fragment stereochemistry is unknown) per ligand were generated by the program, using OPLS_2005 force field, and protonation states were computed by using Epik tool at pH 7. During this process, the energy of each generated state is calculated. The energy of the ground-state, the structure with the lowest energy, is set to zero while to the other states is assigned an increasing energy depending on the decreasing stability of the form at the selected pH. This energy value, called state penalty, is expressed in kcal/mol. The compounds were then energy minimized by using 'MacroModel',[159] implemented with truncated Newton conjugated gradient[160] (TNCG) method. The final number of fragments per library is reported in Table 10. The energy minimized structures were selected for docking calculations.

Table 10: Libraries used for the VS campaign

Name of the library	Number of 2D molecules	Number of 3D molecules
***Maybridge* Rule of 3**	2500	3036
***Asinex* Fragments**	21872	47892
***Life chemicals* Fragment Libraries with Experimental Solubility Data I**	11667	20204
***Life chemicals* Fragment Libraries with Experimental Solubility Data II**	2921	6299
***OTAVA* Solubility fragment library**	1021	1606
***Chembridge* Fragment library**	13808	28037
***SPECS* Preplated fragment-based library**	4532	7892
LIBRARIES COMBINED	58,321	114,966

4.2 Docking calculations

VS campaigns were performed in the X-ray structure of Rel$_{Seq}$ chain A (PDB[138] entry:1VJ7, chain A, residues 1-385).[62] The experimental structure of SYNTH domain well overlaid to the main clusters identified during MD simulations (RMSDs of X-ray vs Clusters Cα atoms are between 1.7 Å and 2.1 Å). The protein was prepared as described in paragraph 3.1.

Due to the presence of a histidine (H312) in the active site that interacts with the β-phosphate group of the GDP (Figure 10), two models were generated: one with the neutral 'HIS312' and one with the charged 'HIP312' form.

4.2.1 Docking set up

Both grid models (HIS312 and HIP312) were generated using GLIDE[103] and the OPLS_2005[113] force field with the default parameters. A cubic region of 24.5Å centreed on GDP molecule was used with a cubic inner box of 10 Å.

The docking protocol was set up and validated using GDP as ligand. The procedure was carried out by using GLIDE v8.0,[103] the standard precision (SP) method and the OPLS_2005 force field. The protein was considered as a rigid body while the ligand was set free to move (no ring sampling was performed and the option 'use enhanced sampling' that adds variations on the input structure to the conformational search was increases by three times). Five poses were saved out of ten post-minimized structures, and no Epik state penalty was added to the docking score. The GDP binding mode observed in the crystallized structure, and discussed in paragraph 1.4.1.1.1, was well reproduced by the docking poses for both HIS312 and HIP312 models. The RMSDs calculated between the heavy atoms of GDP X-ray structure and GDP best poses were 0.78 Å and 1.23 Å for HIS312 and HIP312 grids, respectively (Figure 544).

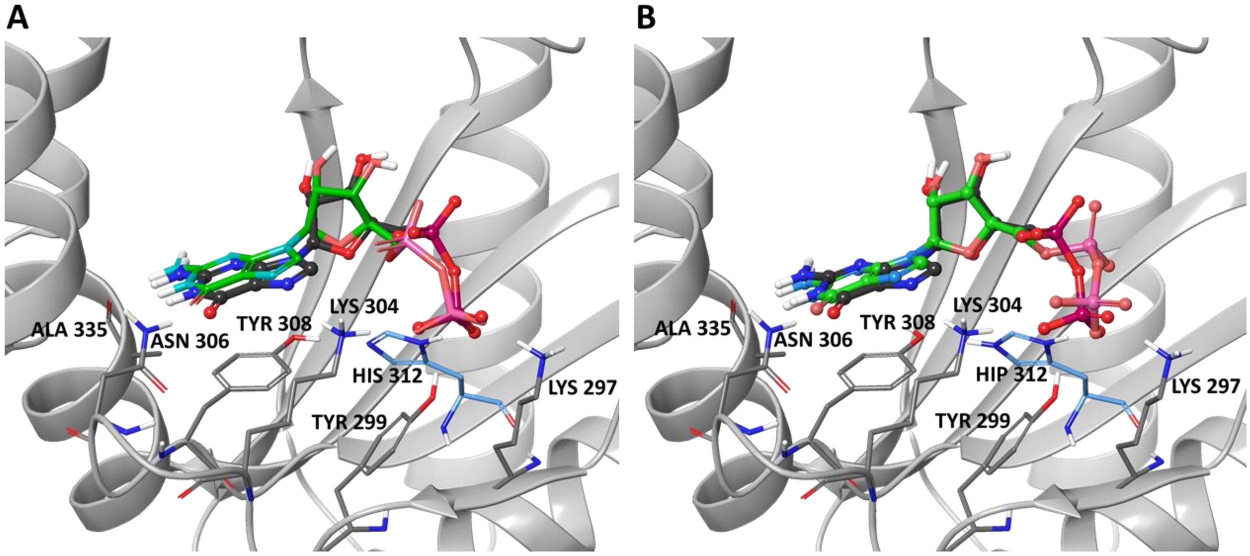


Figure 54: Docking best pose of GDP into HIP312 (A) and HIS312 (B) models overlaid to the X-ray structure (balls and sticks in dark grey carbons). Residues involved in GDP binding mode are shown.

4.2.2 Docking results

A total of 114,966 fragments were evaluated into the SYNTH pocket of both HIS312 and HIP312 models by means of docking simulations using the protocol validated for the X-ray ligand GDP (see paragraph 4.2.1). A single docking pose per fragment was saved for post-docking analysis. Docking poses were first ranked by Glide score and then filtered by 'state penalty' property (state penalty value ≤ 0.6 kcal/mol) to remove less stable and less populated ionization and tautomeric forms. This filtering reduces the number of poses to 86,300 and 84,676 for HIS312 and HIP312 grids, respectively. The duplicates were removed by using the 'Filter duplicates' tool of Maestro (based on

SMARTs), leading to a final number of fragments to analyse of 74,226 and 72,617 for HIS312 and HIP312 models, respectively. Finally, only the fragments presenting an aromatic atom in contact with Y308 side chain, (i.e. with an aromatic atom within 5 Å from any of the heavy atoms of Y308 aromatic ring) were considered.

Indeed, Y308 is a key residue for the synthetic activity of the enzyme and its mutation into asparagine or serine inhibits the synthetic activity.[62] In the X-ray complex Y308 forms a π-π interaction with the guanine of GDP that is also stable during the MD simulations.

The poses that passed the Y308 interaction filter were 30,982 for the HIS312 and 31,851 for the HIP312 grid. In the last step, the PAINS filters were used to remove potential fragments as they represent poor choices for drug development. Three filters (namely PAINS1, PAINS2 and PAINS3),[134] as implemented in Canvas[135,136] suite, were applied. The number of fragments at the end of the workflow was 30,126 and 30,960 for HIS312 and HIP312 grids, respectively. The top 1% of these fragments (ca. 300 fragments) was visually inspected and recurrent scaffolds were identified. In particular, three scaffolds, indole, benzimidazole and aminobenzoic acids, took our attention due to their enrichment in the top 1% (Table 11 and Table 12). Enrichment factor (EF), that compares the number of fragments belonging to a chemotype presented in the top 1% to random selection, was calculated using equation 4.1:

$$EF = \frac{\frac{Scaffold_{subset}}{N_{subset}}}{\frac{Scaffold_{total}}{N_{total}}} \tag{4.1}$$

where Scaffold$_{subset}$ is the number of fragments containing the scaffold type in the subset (top 1%), N$_{subset}$ is the total number of molecules in the subset, Scaffold$_{total}$ is the total number of fragments containing the scaffold type at the end of the VS workflow, N$_{total}$ is the total number of molecules at the end of the VS workflow

Table 11: Enrichment factors of indole, benzimidazole and aminobenzoic acid fragments identified in HIP312 model

HIP	Scaffold/Tot	Scaffold/Tot 1%	EF
Benzimidazole	314/30960=0.001= **1.0%**	19/310=0.061= **6.1%**	0.061/0.001= **6.1**
Aminobenzoic acids	84/30960=0.003= **0.3%**	20/310=0.065= **6.5%**	0.065/0.003= **21.7**
Indole	261/30960=0.008= **0.8%**	10/310=0.032= **3.2%**	0.032/0.008= **4.0**

Table 12: Enrichment factors of indole, benzimidazole and aminobenzoic acid fragments identified in HIS312 model

HIS	**Scaffold/Tot**	**Scaffold/Tot 1%**	**EF**
Benzimidazole	320/30126=0.011= **1.1%**	50/301=0.166= **16.6%**	0.166/0.011= **15.1**
Aminobenzoic acids	58/30126=0.002= **0.2%**	2/301=0.007= **0.7%**	0.007/0.002= **3.5**
Indole	222/30126=0.007= **0.7%**	6/301=0.020= **2.0%**	0.020/0.007= 3.**0**

If we compare the EF values between the two grids, aminobenzoic acids were principally found in the HIP312 model ($EF_{HIP312} = 21.7 > EF_{HIS312} = 3.5$), benzimidazole in the HIS312 model ($EF_{HIS312} = 15.1 > EF_{HIP312} = 6.1$) and indole equally in both grids ($EF_{HIP312} = 4.0$, $EF_{HIP312} = 3.0$).

For each chemotype, the most interesting fragments, B1, B2, A2-6, and I1, I2, I4 (Figure 55), were selected for further studies. Among the top 1% fragments we also selected three singletons, BO1, BT1 and TP1 (Figure 55), due to their LE (Table 13) and binding mode (figure 56). For each fragment, ten docking poses were analysed considering an aromatic filter for the interaction with Y308 as implemented in Maestro (i.e. face to face π-π stacking for aromatic rings centroids distance ≤ 4.4 Å and angle between planes ≤ 30°, face to edge for aromatic rings centroids distance ≤ 5.5 Å and angle between planes ≥ 60°). All the compounds showed high stability of the π-π interaction (Table 13 and figure 56), so they were selected for purchasing.

As expected, all these fragments provided good LE with B1, A4, I1 and BO1 resulting the best within benzimidazole, aminobenzoic acid, indole and singleton structures.

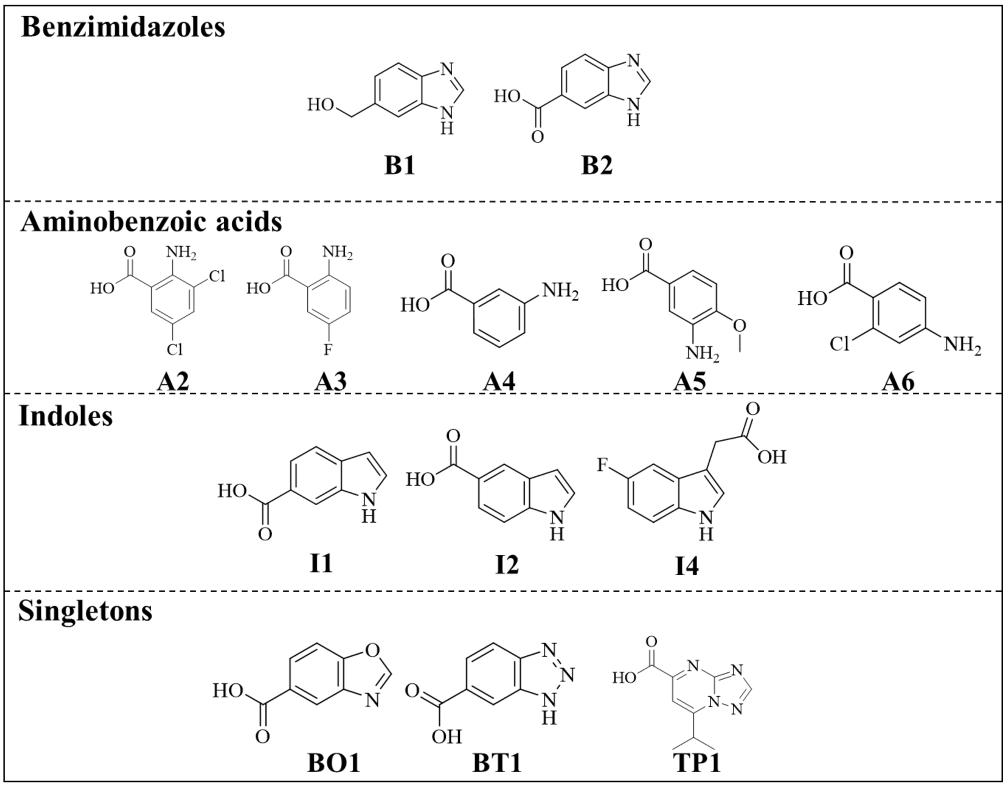


Figure 55: 2D structures of fragments selected from the VS procedure for purchasing

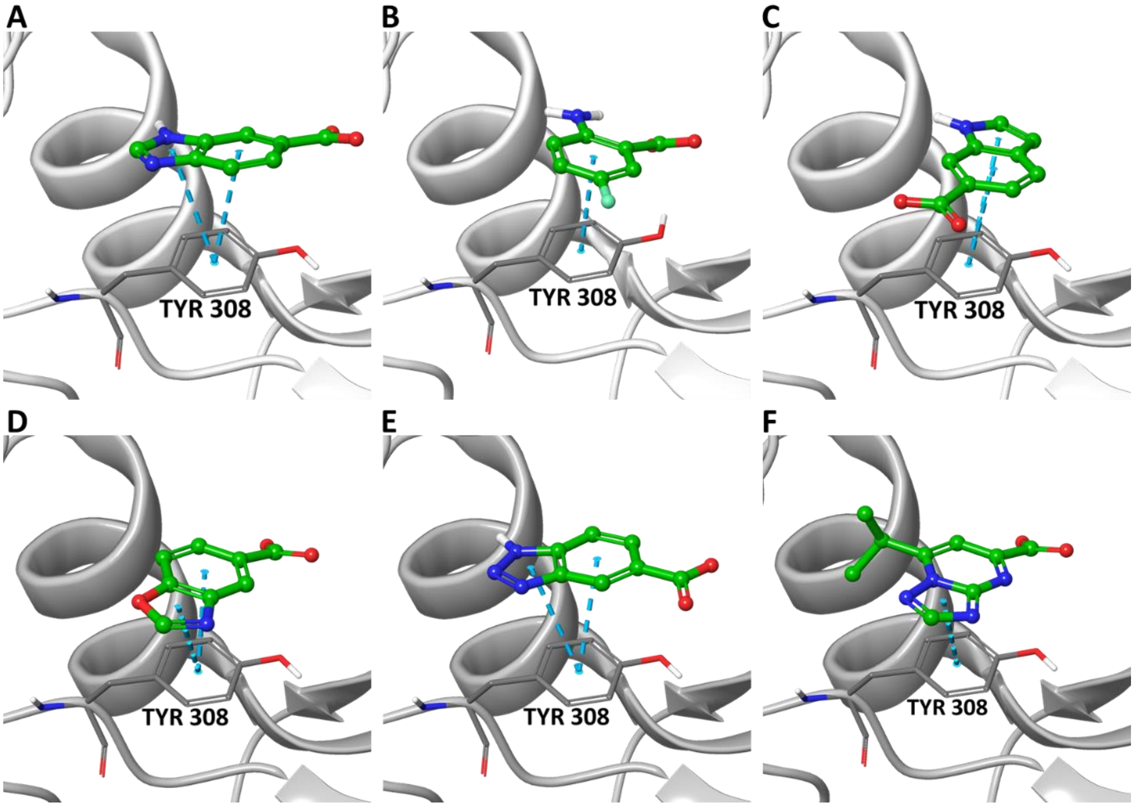


Figure 56: Best poses of A) B2 in HIS312 model, B) A3 in HIP312 model, C) I1 in HIS312 model, D) BO1, E) BT1, F) TP1 in HIS312 model. The π-π interaction with Y308 is shown as cyan dotted line Y308 is shown in grey carbons sticks.

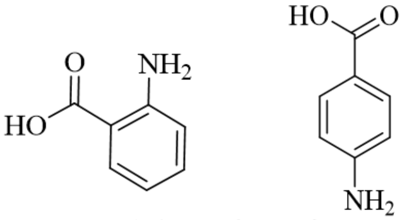
Table 13: Docking results: number of poses forming π-π with Y308 in HIP312 (left) and HIS312 (right) models of fragments selected from VS. Docking score of the best pose and ligand efficiency values were also reported.

Fragment	**# of poses forming π-π stacking with Y308**	**Docking score**	**L. E.**
B1	9/10	-6.496	-0.591
B2	6/10	-7.186	-0.599
A2	9/10	-6.529	-0.544
A3	6/10	-7.010	-0.637
A4	7/10	-6.499	-0.650
A5	6/10	-7.101	-0.592
A6	9/10	-6.155	-0.560
I1	4/10	-6.942	-0.579
I2	5/10	-6.536	-0.545
I4	7/10	-6.634	-0.474
BO1	9/10	-7.251	-0.604
BT1	8/10	-7.083	-0.590
TP1	6/10	-8.424	-0.562

Fragment	**# of poses forming π-π stacking with Y308**	**Docking score**	**L. E.**
B1	9/10	-7.499	-0.682
B2	6/10	-7.081	-0.590
A2	7/10	-5.526	-0.460
A3	7/10	-5.917	-0.538
A4	7/10	-5.752	-0.575
A5	4/10	-6.125	-0.510
A6	8/10	-6.037	-0.549
I1	5/10	-6.052	-0.504
I2	5/10	-6.41	-0.534
I4	6/10	-6.691	-0.478
BO1	9/10	-6.317	-0.526
BT1	5/10	-6.808	-0.567
TP1	7/10	-7.317	-0.488

4.3 Library expansion

Starting from fragments B1, B2, A4, I1 and I4 (Figure 55) selected from the VS campaign previously discussed and fragments A1 and 4-aminobenzoic acid (Figure 57), used to select the three aminobenzoic scaffolds, we performed a library expansion on the PubChem[161] database (using a Tanimoto index[162] ≥ 90%) in order to maximize the exploration of the chemical space.



A1 **4-aminobenzoic**

Figure 57: 2D structures of the fragments added for the library expansion

The structures retrieved from PubChem were collected into three datasets according to the chemotypes (indoles, benzimidazoles and aminobenzoic acids), and, after the removal of duplicates, the 3D structures were generated using Ligprep and filtered according to the Epik state penalty value (≤ 0.6 kcal/mol) (Table 144). The same docking protocol used in the VS calculation was applied to evaluate the binding into the SYNTH active site of Rel$_{Seq}$ (Figure 58).

Table 14 Fragment datasets obtained by using PubChem database

query fragment	**n° of downloaded fragements ***	**n° of fragments per datasets**	**N° of unique fragments (Duplicate filter)**	**N° of 3D structures (Ligprep)**	**State penalty filter < 0.6 kcal/mol**
B1	527	2120	2049	7791	4023
B2	1593				
A1	890	2219	1612	2003	1658
A4	662				
4-aminobenzoic	667				
I1	1426	3684	3291	5920	5137
I4	2258				

** 2D tanimoto similarity ≥ 90% respect to starting fragment*

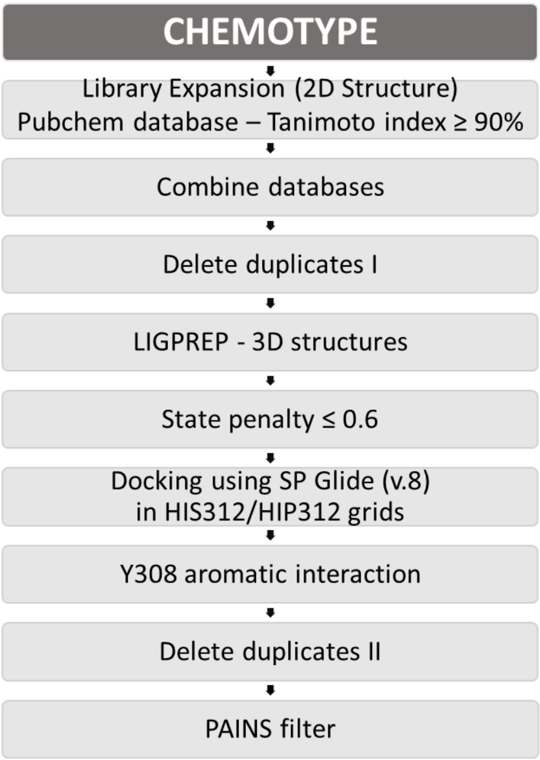


Figure 58: Library expansion workflow

For each dataset, docking results in both HIS312 and HIP312 grids were analysed and filtered saving the fragments able to form a contact with the side chain of Y308 (one pose per ligand). A second step of duplicate removal was also performed to delete duplicates generated by Ligprep, and the tautomer with the best docking score was saved. Finally, PAINS were excluded from the analysis (Table 15 and Table 16).

Table 15: Number of docking poses passing the filters for the HIP312 model

Library set	**N° of fragments docked in HIS312 grid**	**N° of fragments after 'Y308 aromatic interaction' filter**	**N° of unique fragments after 'delete duplicate II' filter**	**N° of fragments after 'PAINS' filter**
Benzimidazole	4023	2040	1386	1371
Aminobenzoic acids	1629	1186	829	732
Indoles	5128	2007	1430	1406

Table 16: Number of poses passing the filters for the HIS312 model

Library set	N° of fragments docked in HIS312 grid	N° of fragments after 'Y308 aromatic interaction' filter	N° of unique fragments after 'delete duplicate II' filter	N° of fragments after 'PAINS' filter
Benzimidazole	4023	1832	1201	1186
Aminobenzoic acids	1629	778	572	502
Indoles	5128	1638	1171	1144

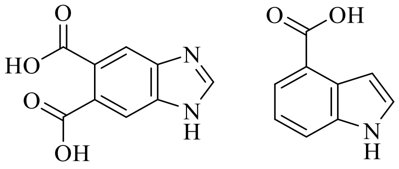
Two compounds (Figure 59), one benzimidazole B4 and one indole I3, with improved or comparable ligand efficiency to the original VS fragments were found (Table 17). B4 includes a second carboxylic group on the ring that favours the interaction with the enzyme (Figure 60c) and improved the LE; I3 has the carboxylic group in a different position compared to the other indoles.

To complete the study, B4 and I3 were evaluated into both grid models analysing 10 poses (Table 17). B4 showed a good LE and increases the number of interactions into the pocket (Table 17 and Figure 60c), but the π-π stacking with Y308 is less stable compared to B1 and B2 (Table 17). I3 reproduced the π-π stacking with Y308 in most of the saved poses with an improved LE compared to I1 and I4 (Table 17 and Figure 61d).

Both fragments were added to the list of fragments to be purchased.

Additional fragments (B3, A1, A7, I5, I5COOH and I6, Figure 62) available in the laboratory where also studied in both models of Rel$_{Seq}$ and added to the purchasing list. Indeed, docking results (Figure 61 and Table 17) into the SYNTH site, highlight that they bind to the pocket forming the desired interaction (the ring stacking with Y308) with good LE values. In particular, B3 showed the best LE of the whole set.

The final list of fragments, selected for MD simulations and to purchase to perform thermal shift assays (TSA), are shown in Figure 62.



B4 **I3**

Figure 59: *2D representation of fragments selected from the library expansion*

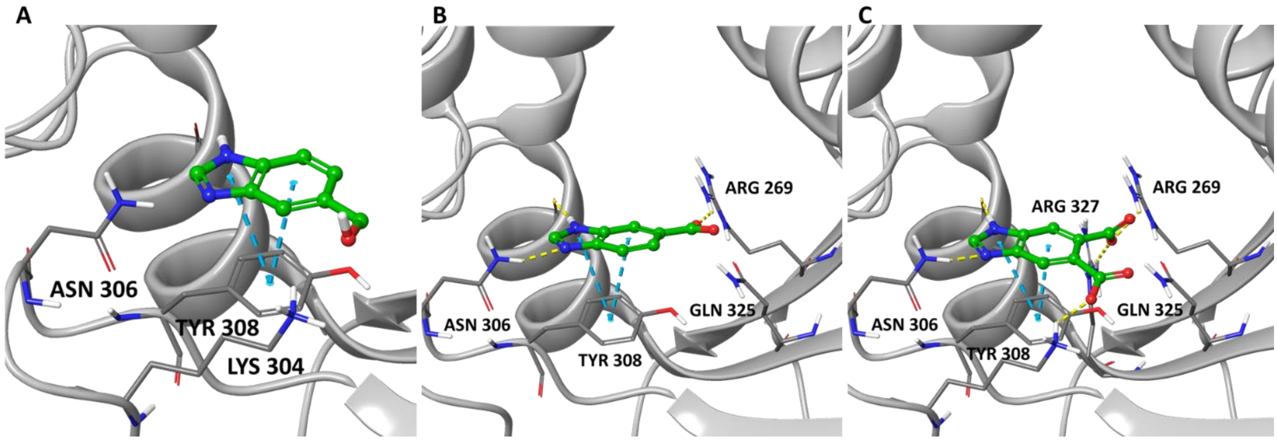


Figure 60: Complete binding mode of A) B1, B) B2 and C) B4 best ranked poses. The fragments are shown in green carbons balls and sticks, the π-π interaction with Y308 is shown in dotted cyan line, salt bridges are shown in purple dotted lines, H-bonds are shown in yellow dotted lines residues involved in the interaction with the fragments are shown in grey carbons sticks, the protein is shown in grey ribbons.

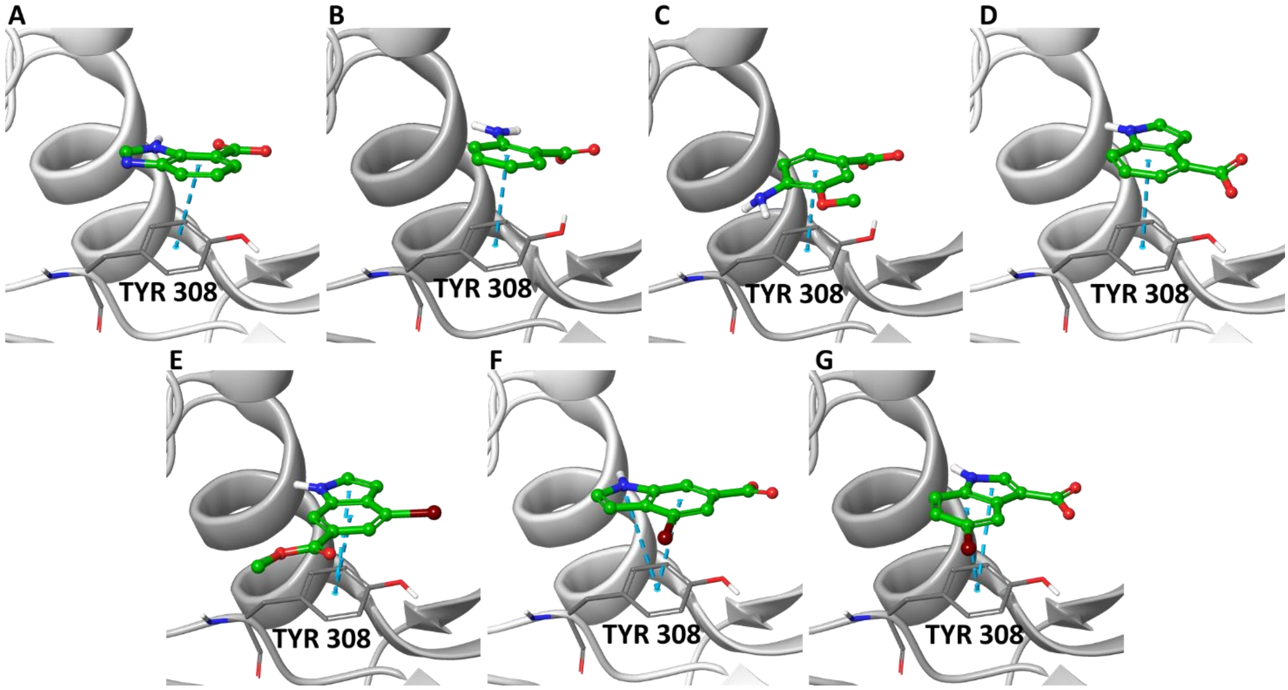


Figure 61: Best poses of A) B3, B) A1, C) A7 D) I3 E) I5 F) I5COOH and G) I6. The fragments are shown in green carbons balls and sticks, the π-π interaction with Y308 is shown in dotted cyan line, Y308 is shown in grey carbons sticks, the protein is shown in grey ribbons.

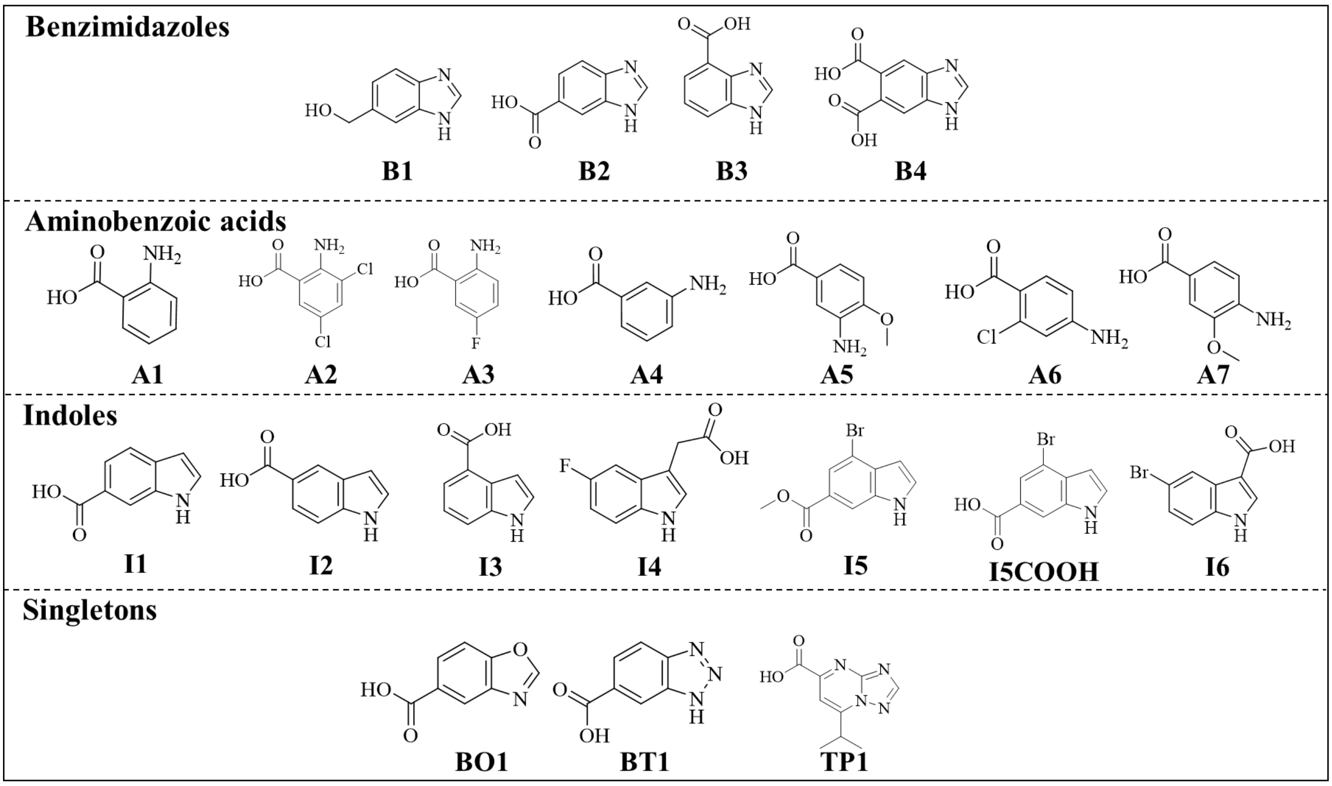


Figure 62: Complete list of fragments selected for purchasing

Table 17: Docking results: number of poses forming π-π with Y308 in HIP312 (left) and HIS312 (right) models of fragments selected from VS. Docking score of the best pose and ligand efficiency values were also reported

Fragment	# of poses forming π-π stacking with Y308	Docking score	L.E.
B3	6/10	-7.319	-0.610
B4	5/10	-8.992	-0.599
A1	9/10	-6.429	-0.643
A7	7/10	-6.621	-0.552
I3	7/10	-6.778	-0.565
I5	7/10	-5.089	-0.363
I5COOH	5/10	-6.743	-0.519
I6	6/10	-6.085	-0.468

Fragment	# of poses forming π-π stacking with Y308	Docking score	L.E.
B3	9/10	-7.413	-0.618
B4	3/10	-8.387	-0.495
A1	8/10	-5.753	-0.575
A7	9/10	-6.21	-0.468
I3	7/10	-6.617	-0.551
I5	7/10	-5.282	-0.377
I5COOH	4/10	-5.884	-0.453
I6	4/10	-5.887	-0.453

All the fragments were also docked into the SYNTH site of the chimera our research group built and published, already discussed in chapter 1.4.1.1.1. The dimension of the site is too big and the fragments too small to find out stable conformations.

4.4 Thermal shift assays

Thermal shift is an experimental technique in which thermal denaturation temperature is monitored following the increase in fluorescence reported by a protein-bond dye.[163] In particular, an environmentally sensitive hydrophobic dye (e.g. SYPRO orange), upon thermal denaturation binds to the hydrophobic regions that progressively become exposed, with an increase of fluorescence emission. The binding of small molecules (e.g. fragments) to the protein can induce conformational changes affecting melting temperature, therefore this technique allows the screening of several compounds with limited consumption of protein.

The robustness of the technique was validated comparing dissociation constant (K_D) of Rel$_{Seq}$ natural substrates GDP (0.26 ± 0.06 mM) and ATP (0.49 ± 0.09 mM) calculated by TSA with values obtained by the most used tryptophan assay ($K_D^{GDP} = 0.15 \pm 0.01$ mM $K_D^{ATP} = 0.39 \pm 0.04$ mM).

We, therefore, evaluated the affinity of the 21 fragments selected from the *in silico* study for three Rel$_{Seq}$ constructs, consisting of Rel$_{Seq}$ residues 1-385 (Rel$_{Seq}$), residues 1-224 (Rel$_{Seq}$ HYD) and residues 79-385 (Rel$_{Seq}$ SYNTH*),* by titration of the protein in TSA. The results (Table 18) show that twelve fragments, A1, A3, A4, A5, A6, A7, I1, I2, I3, I4, BO1 and TP1, can bind selectively the SYNTH domain with a mM K_D, three items, B2, I5 and I5COOH, can bind HYD domain with milli or sub mM K_D and finally B3 and I6 can bind both domains with a mM K_D.

More interesting, aminobenzoic acid fragments are completely selective for the SYNTH domain, therefore, they were used for the design of new potential ligands of this site (discussed in paragraph 4.9). The benzoxazole and triazole-pyrimidine structures seem to be also selective but further investigations are needed. On the other hand, benzimidazole and indole can be used to design ligands for both domains depending on how chemical space is modified.

Table 18: K_D values calculated by performing thermal shift assays using three constructs of Rel_{Seq} (1-385 Rel_{Seq}, 79-385 Rel_{Seq} SYNTH, 1-224 Rel_{Seq} HYD).

Fragment	**K_d Rel_{Seq} (mM)**	**K_d Rel_{Seq} SYNTH (mM)**	**K_d Rel_{Seq} HYD (mM)**
B1	*no binding*	*no binding*	*no binding*
B2	*double effect*	*no binding*	1.9 ± 0.7
B3	3.4 ± 0.3	4.3 ± 0.4	3.3 ± 0.9
B4	*no binding*	*no binding*	*no binding*
A1	1.2 ± 0.3	10.8 ± 2.2	*no binding*
A2	*no binding*	*no binding*	*no binding*
A3	1.5 ± 0.1	5.5 ± 0.9	*no binding*
A4	6.6 ± 1.2	9.8 ± 2.8	*no binding*
A5	1.1 ± 0.2	2.2 ± 0.4	*no binding*
A6	4.3 ± 1.1	6.5 ± 1.2	*no binding*
A7	4.0 ± 0.9	4.3 ± 0.5	*no binding*
I1	6.5 ± 1.1	9.6 ± 1.5	*no binding*
I2	2.5 ± 0.6	5.5 ± 0.9	*no binding*
I3	4.0 ± 0.5	9.9 ± 4.5	*no binding*
I4	3.2 ± 0.7	17.8 ± 7.0	*no binding*
I5	0.6 ± 0.1	*no binding*	0.25 ± 0.04
I5COOH	10.0 ± 2.0	*no binding*	9.4 ± 1.2
I6	5.7 ± 1.3	7.7 ± 0.5	21.7 ± 14.1
BO1	2.2 ± 0.3	2.7 ± 0.6	*no binding*
BT1	*no binding*	*no binding*	*no binding*
TP1	3.4 ± 0.8	8.3 ± 1.6	*no binding*

4.5 Molecular dynamics simulations

The stability of the docking poses for the fragments shown in Figure 62 was explored by running MD simulation with Desmond (100ns, NPT, water TIP3P, force field OPLS3e, dt = 2 fs).

For indole, benzimidazole and singleton fragments, the best pose in the HIS312 model was used as input structure while for aminobenzoic acid fragments the best pose in the HIP312 model was used. We decided to use HIP312 models for aminobenzoic acids due to the higher EF calculated in this grid compared to the HIS312 one (Table 11 and Table 12).

During the simulations, the π-π stacking with Y308, formed if two aromatic rings stacked face-to-face with distance within 4.4 Å and an angle between planes < of 30° or if they are stacked face-to-edge with centroid distance within 5.5Å and an angle between planes > of 60°, was monitored using the simulation interaction diagram (SID) tool by Desmond (Figure 63).

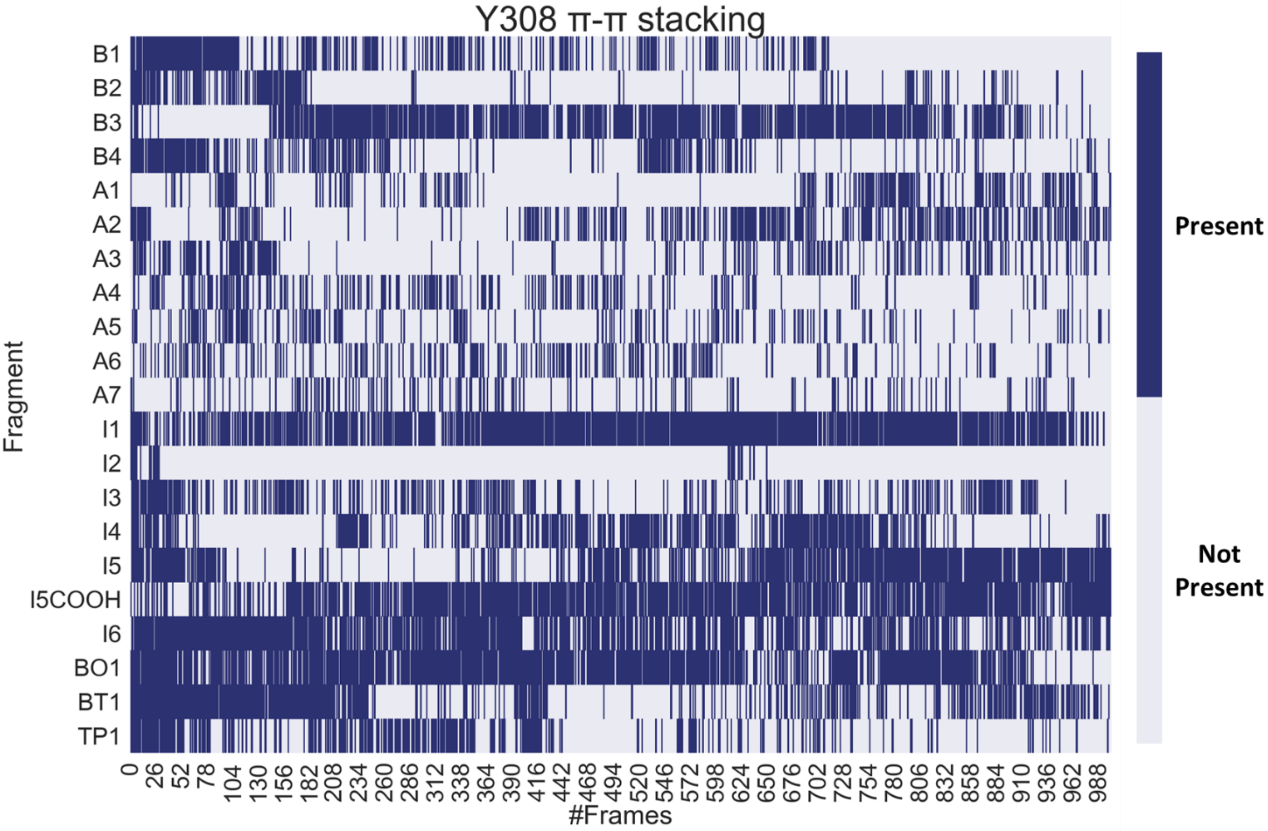


Figure 63: Time series of the π-π stacking with Y308 for the selected fragments. In blue are highlighted the frames where the interaction is present. Every fragment, except for B1 and I2, form and break the interaction for the entire run.

Fragments I1 (frame forming the π-π = 83%), BO1 (77%), I5COOH (77%), I6 (69%), B3 (64%) maintained the π-π interaction for more than the 60% of the simulation time, highlighting the high stability of these fragments for the SYNTH site. Fragments I5 (55%), BT1 (48%), I4 (41%), I3 (35%), A2 (34%), TP1 (33%), B1 (32%), B4 (28%), A4 (23%), A6 (23%), A1 (23%), A3 (22%), A5 (21%), maintained the interaction for at least the 20% of the simulations while it is rarely found for B2, A7 and I2 (18%, 16% and 3%, respectively). All the fragments formed and broke the interaction for all the simulation (Figure 63), except for fragments B2 and I2 that lost the interaction after 70 ns and 60 ns, respectively. To better understand the fragments behaviour into the pocket, the distances between the centroid of Y308 side chain and centroids of fragments rings were monitored during the simulations (Figure 64). As expected, all the fragments, with the exception of B1 and I2 that left the pocket (Figure 65), maintained a contact with the residues within a distance of 5.5 Å (compatible with a π-π stacking interactions[164]). Only A1 deviates from this value (average distance of 6.36 ± 1.34 Å) and, together with I4 (average distance of 5.52 ± 1.15 Å), showed higher fluctuations (I2 average distance 7.03 ± 2.28 Å, B1 average distance 13.87 ± 15.68 Å).

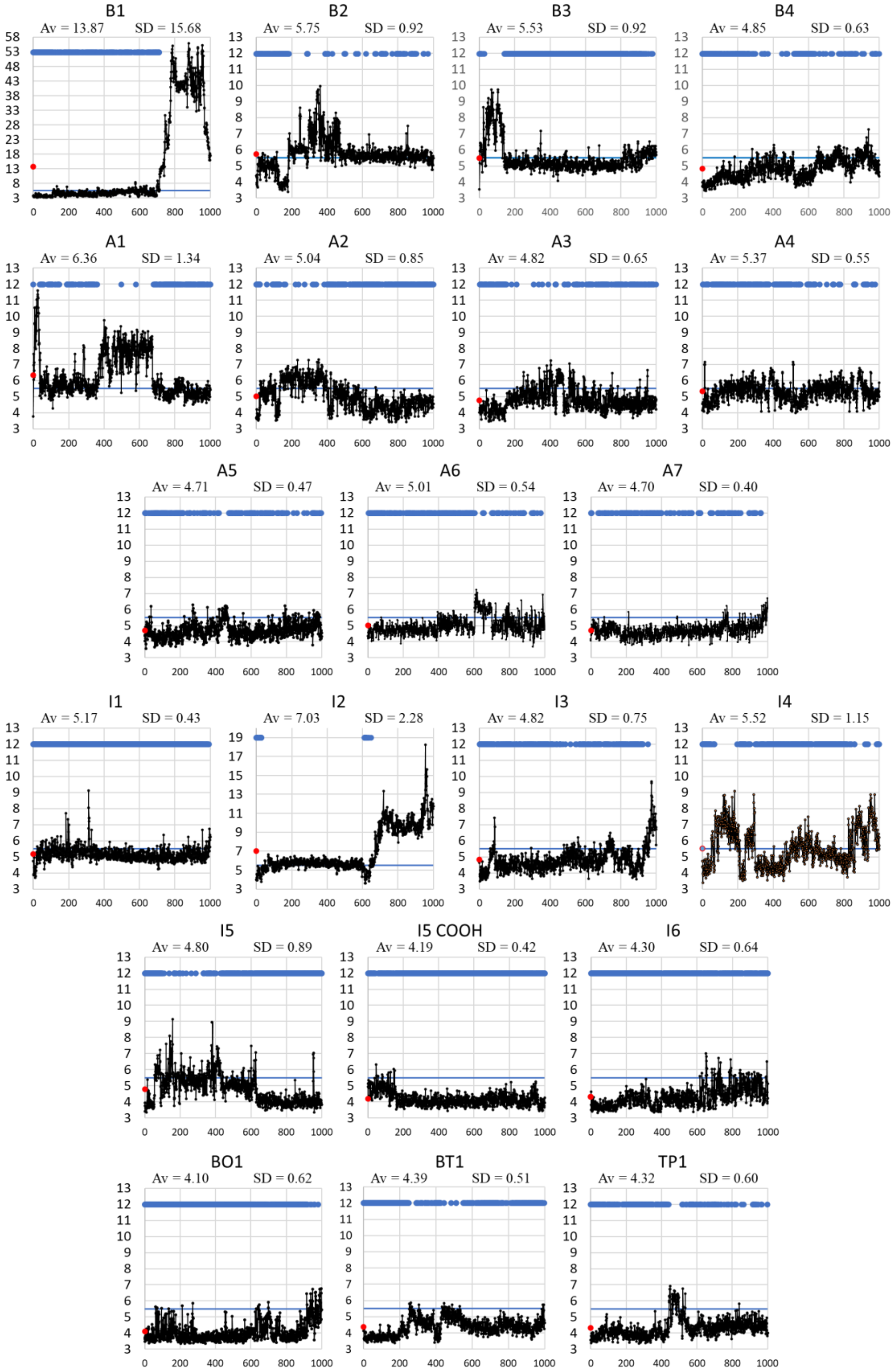


Figure 64: Distances between the centroids of Y308 side chain and of the fragment rings monitored during MD simulations (in black). The average (Av) and standard deviation (SD) of distances are also reported. The light-blue line defines the maximum distance at which a π-π can be formed. The presence of the π-π stacking calculated by Desmond is reported in cyan.

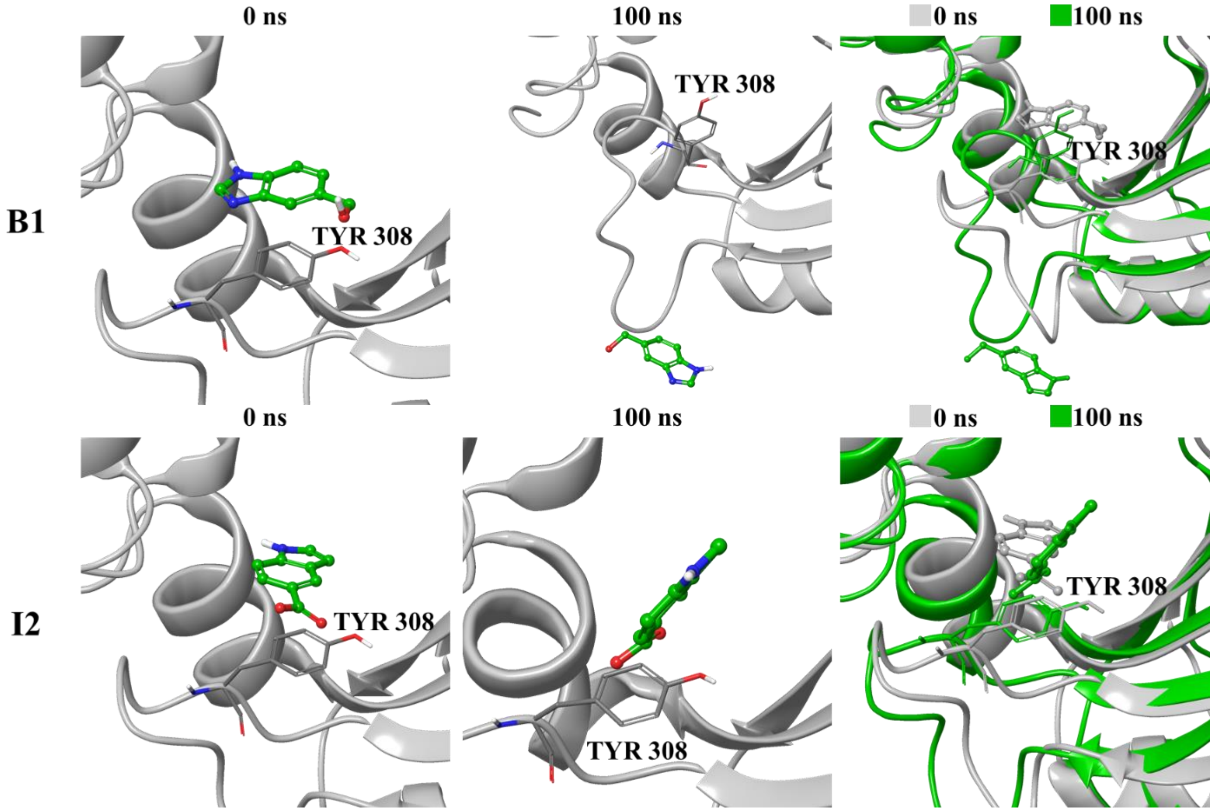


Figure 65; First frame (left). last frame (middle) and superposition of the two frames (right) of the MD simulations of B1 (top) and I2 (bottom). For boxes on the left and middle the protein is shown in grey ribbons, residue Y308 is shown in sticks grey carbon atoms and the ligands are shown in balls and sticks green carbon atoms. For the boxes on the right, the protein is shown in grey ribbons for frames at 0 ns and in green ribbons for the frames at 100 ns, the same colour code of the protein was used for residue Y308 in sticks and for the ligands in balls and sticks.

Comparing the experimental TS data for SYNTH domain, fourteen fragments showed *in silico-in vitro* agreement. In detail, fragments B3, A1, A3, A4, A5, A6, A7, I1, I3, I4, I6, BO1 and TP1 maintained the stacking with Y308 during the simulations and bind to the SYNTH domain in the TS experiment, while fragment B1 lost the interaction and does not bind to the SYNTH domain.

4.6 SiteMap calculation

Rel$_{Seq}$ might present different binding pockets or binding surfaces that can be bound by compounds to modulate its activity in place of the HYD and the SYNTH pockets.

Several are the tools available to define these binding regions and we used SiteMap[165] available in Schrodinger suite.

By using a sphere of 1 Å size, the program defines the SASA of the protein discriminating hydrophilic and hydrophobic points.

A hydrophilic point is assigned if:

Gridphilic = vdW energy + oriented-dipole energy < threshold (usually -8 kcal/mol) (4.2)

while a hydrophobic point is assigned if

Grid phobic = vdW energy – 0.30 * oriented-dipole energy < threshold (the least restrictive is -0.75 kcal/mol) (4.3)

where vdW stand for van der Waals.

The philic map is further divided into H-bond acceptors, H-bond donors and metallic regions.

Once the points are assigned, a binding grid is created and the SiteScore is calculated by using the following equation:

$$\text{SiteScore} = 0.0733\sqrt{n} + 0.6688e - 0.20p \qquad (4.4)$$

where n is the number of site points assigned for the binding grid (capped at 100), e is the enclosure score, and p is the hydrophilic score capped at 1 to limit its impact in highly polarized pocket. SiteScore of 0.8 was found to be the best threshold to define druggable (≥ 0.8) and not druggable (< 0.8) sites.

The four binding sites, identified for the HIS312 model, are shown in Figure 66.

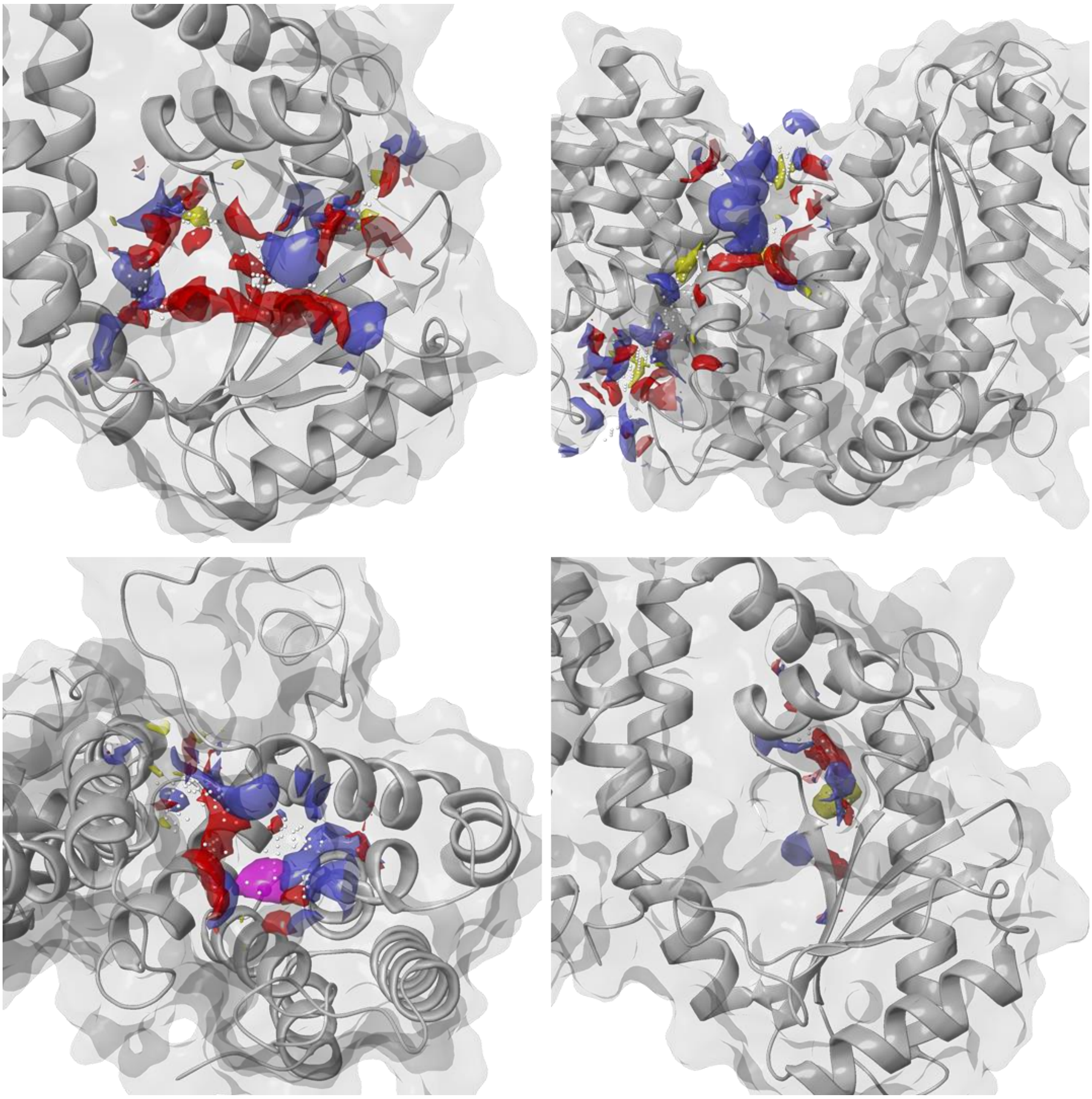


Figure 66: Binding sites identified by SiteMap. The top left figure shows the SYNTH site, top right figure shows a binding surface found between helix 3, helix 8 and helix 10, the bottom left figure shows the HYD site and the bottom right figure shows a pocket between helices 11, helices 12, helices 13 and beta sheet 1. Regions where H-bond donors are present are shown in red, regions where H-bond acceptor are present are shown in blue, regions where the metal is present is shown in purple, hydrophobic regions are shown in yellow, the site point that are assigned to regions are shown in white.

The SYNTH site was identified as the most druggable site of the enzyme with scores of 1.053. The HYD site, was also identified as potential pocket (third in ranking) also in the 'hydrolase off' conformation (SiteScore = 0.950). Two new regions, different from the known binding pockets were identified by the program. A binding surface between helix 3, helix 8 and helix 10 with SiteScore of 0.966 (second in the ranking) and a little pocket behind the SYNTH site with SiteScore of 0.834

(fourth and last druggable site). The first is a binding surface that, if Rel$_{Seq}$ shows the same behavior of the long RSH from *S. subtilis,*[65] should be involved in a potential homodimerization of the long RSH enzymes, while the second is a pocket between helices 11, helices 12, helices 13 and beta sheet 1 (Figure 8a behind the SYNTH site) that could be used for docking simulations.

SiteMap was used also on the second model used for MD simulations (discussed in chapter 3) where the gap K110-N123 was refined. Results showed the same regions already obtained in the first model. The only difference consists in the second ranked site identified where the HYD site is fused to the binding surface between helix 3, helix 8 and helix 10 (SiteScore = 0.972 and figure 67).

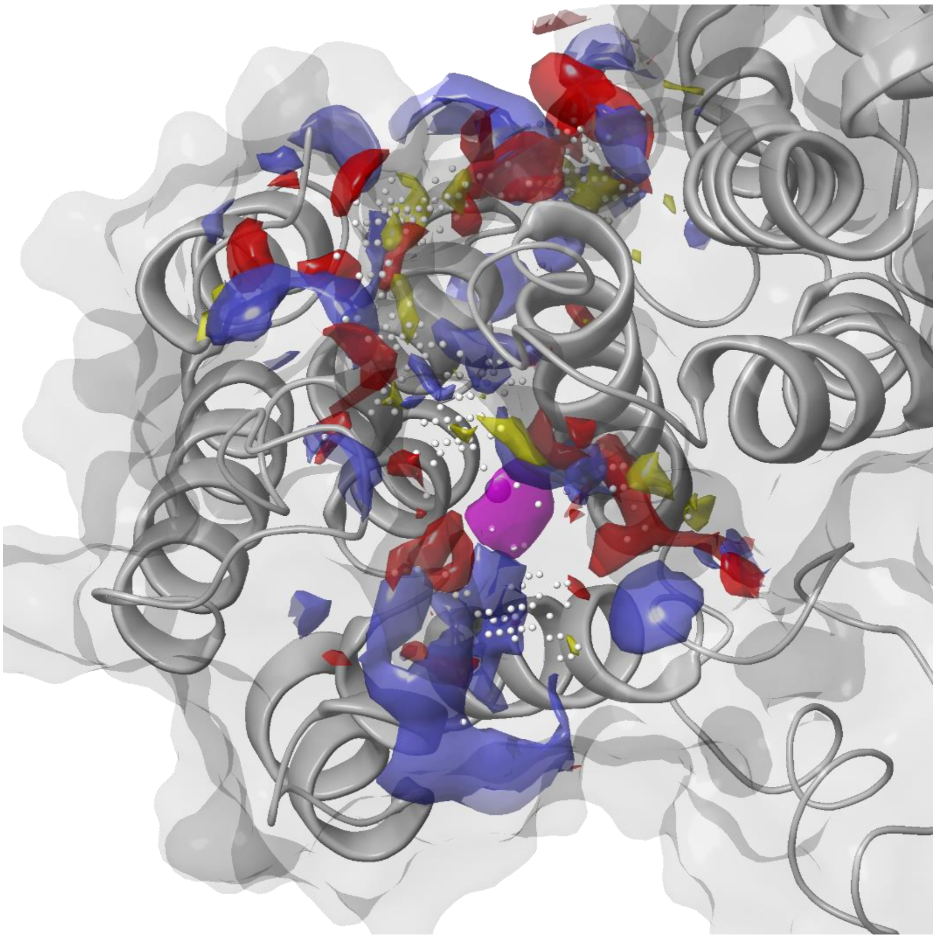


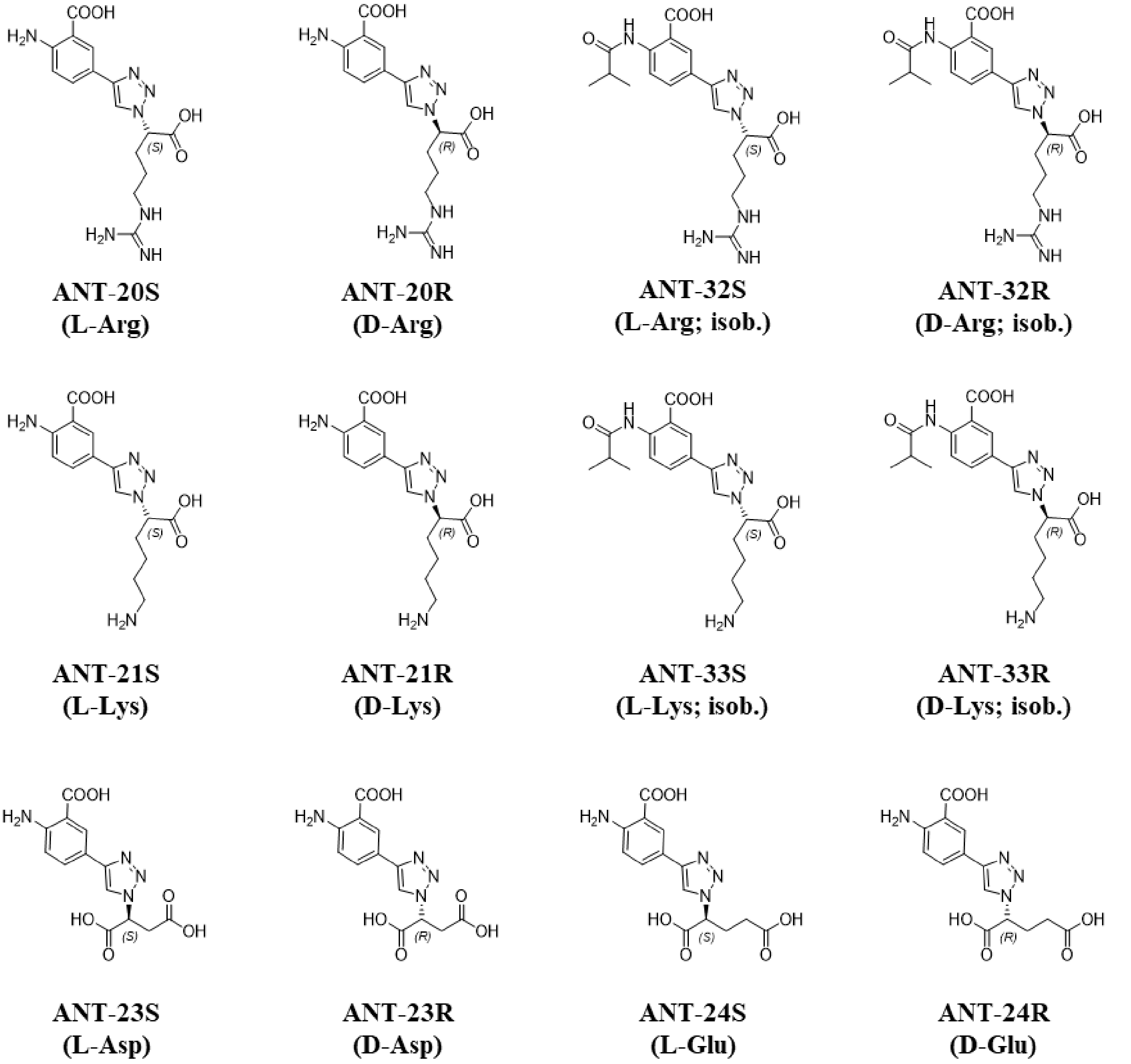
Figure 67: 3D representation of the only site that differs between refined and not refined systems. Regions where H-bond donors are present are shown in red, regions where H-bond acceptor are present are shown in blue, regions with the metal is shown in purple, hydrophobic regions are shown in yellow, the site point that are assigned to regions are shown in white.

The GDP pocket of the SYNTH site was used for the FBDD, but we do not exclude that in the future also the other sites identified can be used for VS campaigns.

4.7 The ANT library: Design of 2-aminobenzoic acid compounds

As previously discussed, aminobenzoic acid is the scaffold selective for the SYNTH site. Therefore, we decided to develop a small library of compounds starting from the anthranilic acid (2-aminobenzoic acid, A1 in Figure 62) scaffold.

Thus, starting from the binding mode of fragment A1 and analysing the way the best docking pose bound to the SYNTH pocket, different moieties, named ANT-derivatives, were designed (Figure 68) with the aim of interacting with amino acids involved in the X-ray GDP binding mode (Figure 10).



ANT-20S (L-Arg) **ANT-20R (D-Arg)** **ANT-32S (L-Arg; isob.)** **ANT-32R (D-Arg; isob.)**

ANT-21S (L-Lys) **ANT-21R (D-Lys)** **ANT-33S (L-Lys; isob.)** **ANT-33R (D-Lys; isob.)**

ANT-23S (L-Asp) **ANT-23R (D-Asp)** **ANT-24S (L-Glu)** **ANT-24R (D-Glu)**

Figure 68: 2D libraries of ANT-derivatives designed to fit Rel$_{\text{Seq}}$ SYNTH site.

ANT-derivatives share a common scaffold, shown in Figure 69, where the anthranilic acid ring is connected to an amino acid by a triazole used as a linker. Moreover, the amine group of some ANT-derivatives (ANT-32 and ANT-33) is capped by an isobutyric moiety.

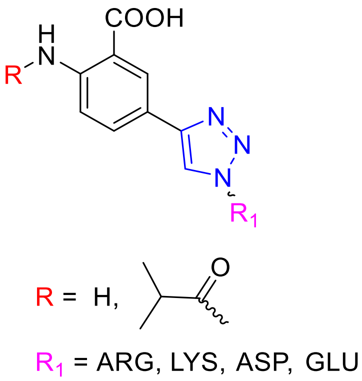


Figure 69: ANT-derivatives scaffold

The anthranilic acid ring, should stabilize the ligand into the purine ring pocket by forming the π-π stacking with Y308 side chain, the triazole ring was added as a spacer and to increase solubility. Asp and Glu residues of ANT-23 and ANT-24 were chosen to replace phosphate groups. Arg and Lys of ANT-20, ANT-21, ANT-32 and ANT-33 were selected to interact with residues D264 and E323, catalytical amino acids involved in Mg^{2+} coordination. Finally, in ANT-32R&S and ANT-33R&S the isobutyric capping group of Relacin (Figure 16) was used.

The binding ability of these compounds was tested both *in vitro* by performing TSA, once they were synthesised by our research group, and *in silico* by performing molecular docking and MD simulations.

4.8 Thermal shift assay

ANT-derivatives binding affinity to both SYNTH and HYD domains was tested by TSA on the three constructs we already used for fragments. As shown by Table 19, all ANT compounds showed a low-mM K_D for the enzyme containing only the SYNTH domain, improved compared to the GDP one (1.79 ± 0.10 mM), except for ANT-21S probably for its low purity. ANT derivatives resulted selective for the SYNTH domain, with the exception of ANT-20R&S that can also bind to the enzyme containing only the HYD domain.

ANT-20R and ANT-23S are the best tested compounds with micromolar K_D values (0.16 ± 0.02 mM for both compounds).

Table 19: K_D values of ANT-derivatives calculated by performing thermal shift assays using three constructs of Rel_{Seq} (1-385 Rel_{Seq}, 79-385 Rel_{Seq} SYNTH, 1-224 Rel_{Seq} HYD).

Amino acid	**Isomer**	**K_d Rel$_{Seq}$ (mM)**	**K_d Rel$_{Seq}$ SYNTH (mM)**	**K_d Rel$_{Seq}$ HYD (mM)**	**Compound**
Arg	S	1.11 ± 0.22	0.86 ± 0.41	2.09 ± 0.39	***ANT-20S***
	R	1.28 ± 0.31	0.16 ± 0.02	1.60 ± 0.49	***ANT-20R***
	S + cap	1.54 ± 0.36	0.31 ± 0.04	*no binding*	***ANT-32S***
	R + cap	1.28 ± 0.31	0.52 ± 0.10	*no binding*	***ANT-32R***
Lys	S	3.05 ± 0.42*	nd	nd	***ANT-21S****
	R	0.13 ± 0.02	0.56 ± 0.11	*no binding*	***ANT-21R***
	S + cap	2.67 ± 0.47	0.21 ± 0.04	*no binding*	***ANT-33S***
	R + cap	1.26 ± 0.31	0.31 ± 0.03	*no binding*	***ANT-33R***
Asp	S	0.31 ± 0.06	0.16 ± 0.02	*no binding*	***ANT-23S***
	R	3.54 ± 0.99	0.86 ± 0.31	*no binding*	***ANT-23R***
Glu	S	4.36 ± 2.36	0.98 ± 0.12	*no binding*	***ANT-24S***
	R	2.71 ± 0.48	0.21 ± 0.02	*no binding*	***ANT-24R***

* *low compound purity*

4.9 Docking of the ANT library into the SYNTH site

ANT-derivatives are bigger items compared to the fragments used in the VS, therefore the grid boxes for both HIS312 and HIP312 model were generated with increased dimensions of the inner and outer boxes (14 Å and 34 Å side, respectively). Furthermore, new docking protocols, for both HIS312 and HIP312 models, were validate for the GDP molecule using, the default options and the OPLS3e force field. No Epik state penalty was added to the score. A slight difference discriminate HIS312 grid protocol from the HIP312 one: only the HIP312 model needed an enhanced sampling option of the conformational sampling by two times to reproduce the GDP X-ray binding mode (Figure 70). We observed an improvement in the RMSDs results respect to the OPLS_2005 protocol: RMSDs calculated between GDP X-ray structure and GDP best poses in HIS312 and HIP312 grids are 0.55 Å and 1.18 Å, respectively.

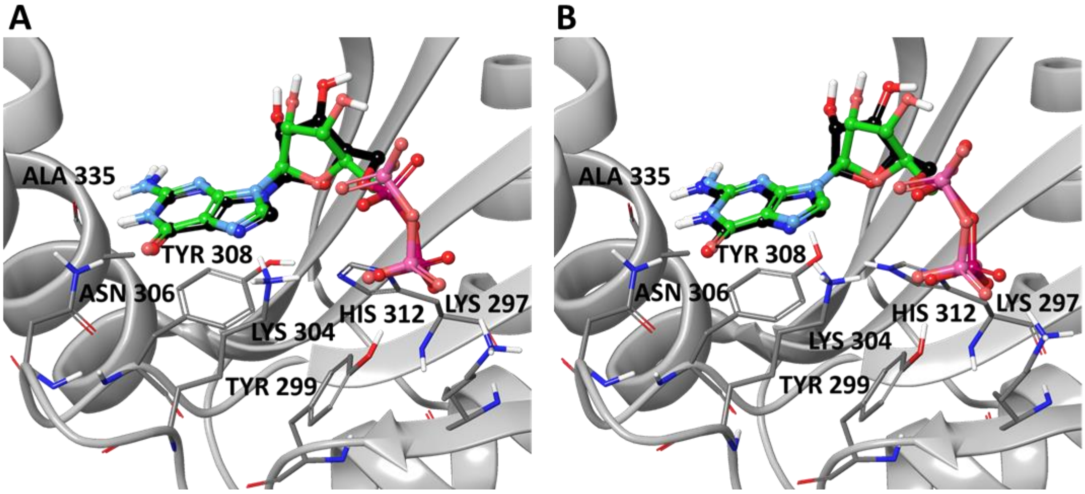


Figure 70: Docking best pose of GDP into HIS312 (A) and HIP312 (B) models, using the new OPLS3e docking protocol, overlaid to the X-ray structure (balls and sticks in dark grey carbons). Residues involved in GDP binding mode were shown.

Molecular docking simulations were performed into both HIS312 and HIP312 models (Table 20, Figure 71 and Figure 72) saving ten poses.

Table 20: Scores of best poses and number of poses forming π-π stacking with Y308 of ANT derivatives. Green refers to HIP Blu refers to HIS

Compound	Docking score	Ligand efficiency	π-π Y308	Docking Score	Ligand Efficiency	π-π Y308
ANT-20S	-6.989	-0.269	10/10	-7.382	-0.284	9/10
ANT-20R	-7.118	-0.274	8/8	-7.729	-0.297	10/10
ANT-21S	-6.804	-0.284	10/10	-6.873	-0.286	9/10
ANT-21R	-6.671	-0.278	7/~~8~~	-6.857	-0.286	7/10
ANT-23S	-7.885	-0.343	7/10	-7.046	-0.306	5/10
ANT-23R	-7.387	-0.321	7/10	-6.979	-0.303	4/10
ANT-24S	-7.853	-0.327	6/10	-6.971	-0.29	5/10
ANT-24R	-7.621	-0.318	7/10	-7.031	-0.293	7/10
ANT-32S	-7.261	-0.234	6/10	-7.826	-0.252	4/10
ANT-32R	-7.427	-0.24	8/10	-8.003	-0.258	6/10
ANT-33S	-6.276	-0.216	0/10	-7.341	-0.253	2/10
ANT-33R	-6.849	-0.236	0/10	-7.465	-0.257	3/10

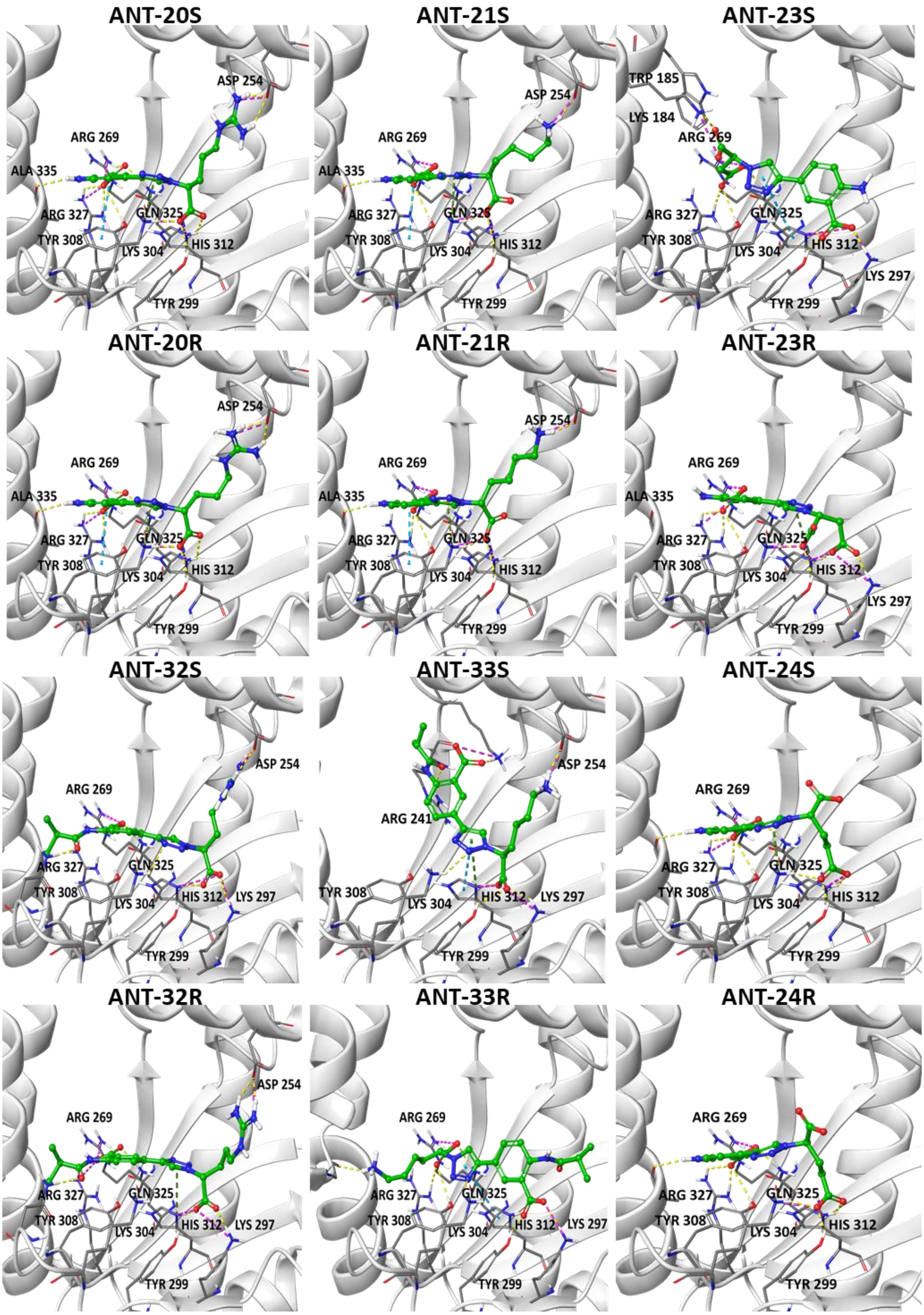


Figure 71: *3D representation of the main cluster poses of ANT derivative moieties in HIP312 protein model. ANT-derivatives were rendered in green carbons balls and sticks, residues involved in the interactions were rendered in grey carbon sticks*

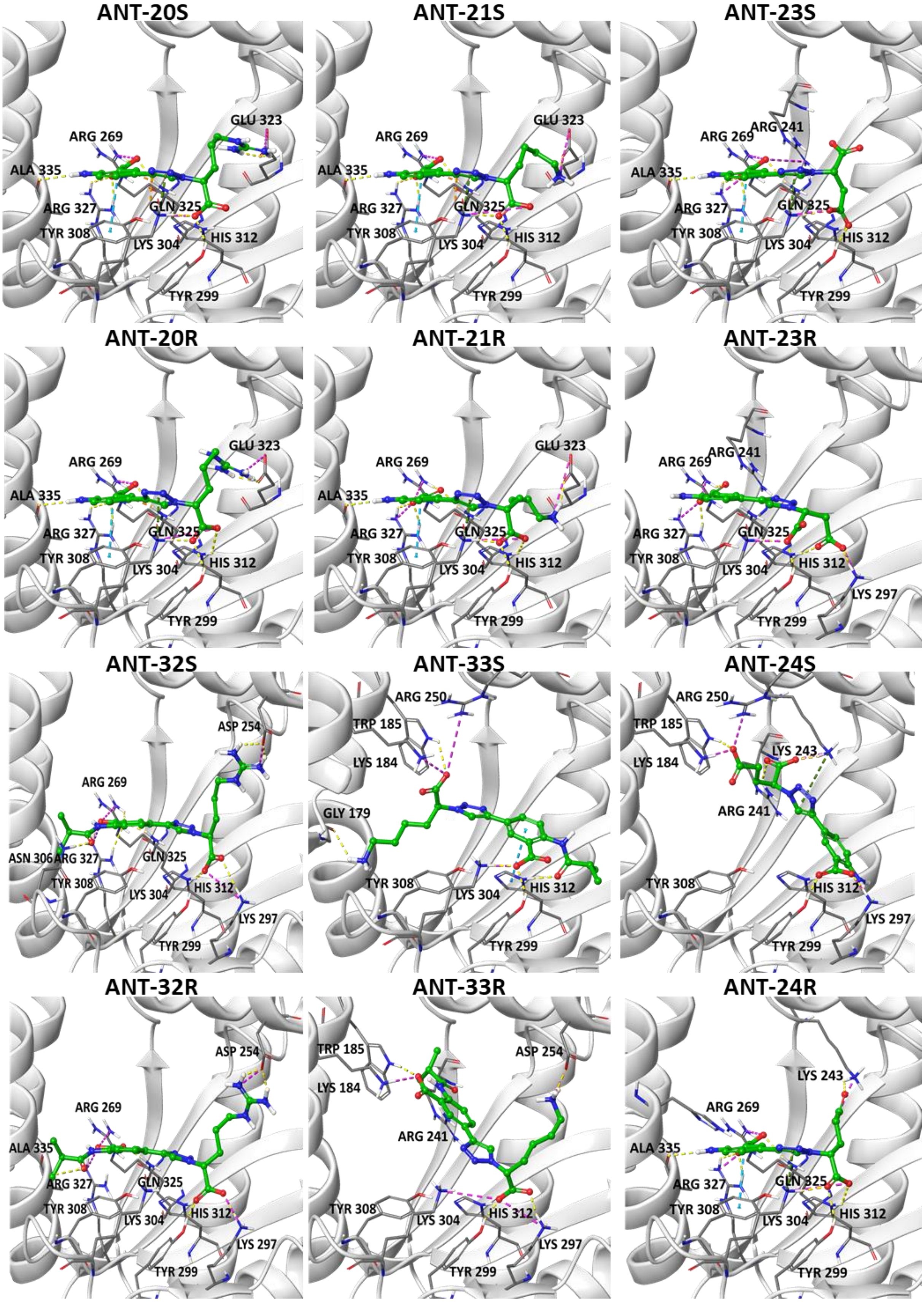


Figure 72: 3D representation of the main cluster poses of ANT derivative moieties in HIS312 protein model. ANT-derivatives were rendered in green carbons balls and sticks, residues involved in the interactions were rendered in grey carbon sticks

The anthranilic acid ring formed the desired π-π stacking with Y308. The ring is further stabilized into the pocket by forming salt bridges/H-bonds with side chains of R269, Q325 and R327. The triazole ring not only is needed as a spacer and to increase solubility, but it can also interact with positively charged or aromatic amino acids (particularly with K304 and H312) forming π-cationic or π-π stacking, respectively, potentially increasing affinity to the SYNTH pocket. The C-terminus carboxylic groups of the amino acids and Asp or Glu side chains interacted with positively charged residues of the pocket, while Arg or Lys formed salt bridge with residue E323 or D254 (Figure 72). Unfortunately, D264 side chain is oriented out of the SYNTH catalytic site, therefore compounds docked into the site cannot form any interaction with this residue.

Considering the results into both HIS312 and HIP312 models (Table 20 and Figures 71 and 72), ANT-20 (both enantiomers), ANT-21 (both enantiomers), ANT-24R and ANT-32R are the compounds that most preserved the key interactions.

These data do not fully respect the TS experimental data: whereas ANT-20R is the best compound also in docking simulations, ANT-23S desired binding mode is not repeated among the saved poses. ANT-20R shows similar results in both HIS312 and HIP312 grids. On the other hand, ANT-23S presents an alternative binding mode in HIP312 grid where anthranilic acid ring and amino acid tail switched their position. The disagreement between *in silico* and *in vitro* data can be caused by the high number of positive charged residues, present in the SYNTH site, with which the negative charges of ANT-23S can interact with. On the other hand, the limited number of negatively charged residues located in one of the sides of the pocket forces the ANT-20R to maintain a stable pose. Therefore, MD simulations were carried out to ensure binding mode stability.

4.10 Molecular dynamics simulations

To ensure binding mode stability, MD simulations using Desmond (100ns, NPT, TIP3P water box, force field OPLS3e, dt = 2 fs) were performed on the best ranked poses. We used HIS312 model to better compare data with those obtained with the TS experiments performed at pH=8.

The π-π stacking with Y308 was monitored (Figure 73) and used to interpret the TS data. The interactions formed by the amino acid were also evaluated and are shown and discussed in the appendix (paragraph 4.12).

SID tool by Desmond was used to evaluate the interactions.

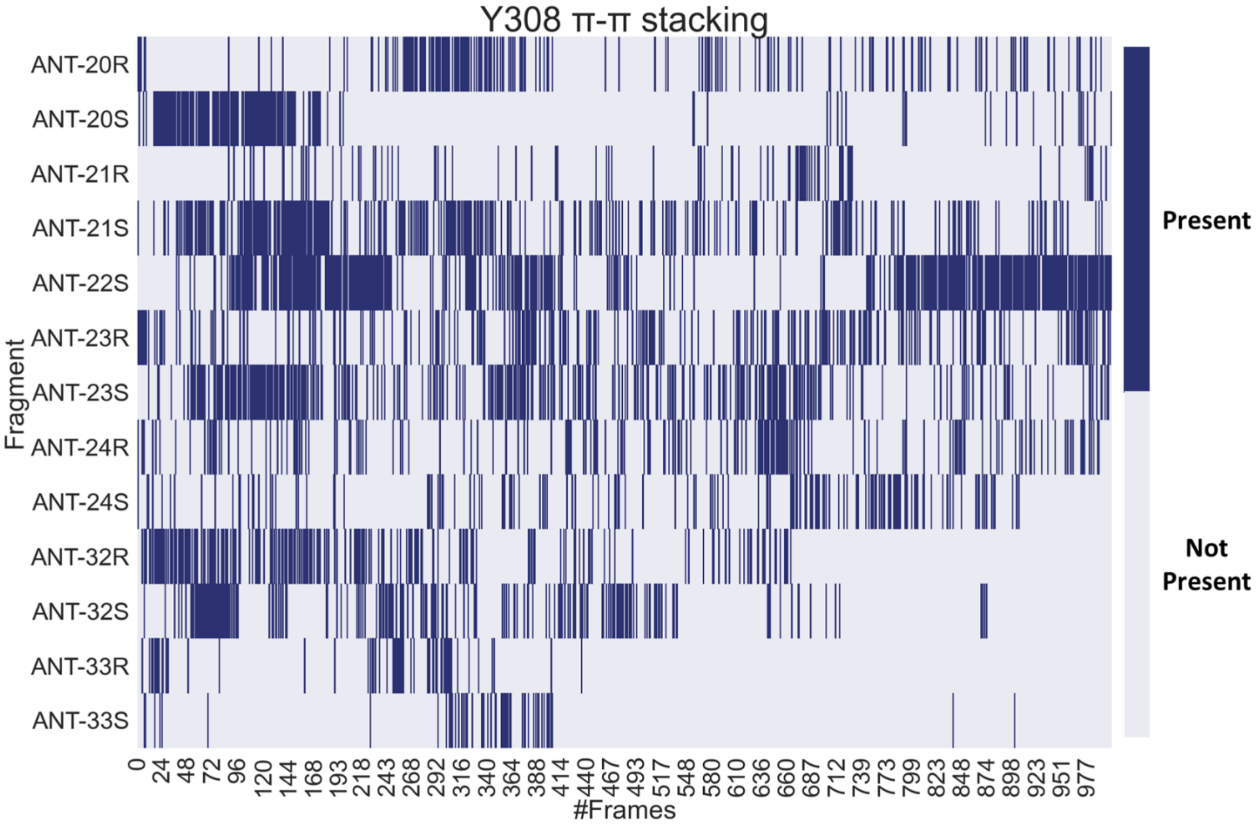


Figure 73: Time series of the π-π stacking with Y308 for ANT derivatives. In blue are highlighted the frames where the interaction is present. Every Compound, except for ANT-32R and ANT-33R, maintains the interaction during all the molecular dynamics simulations

ANT-21R (frame forming the π-π = 59%), ANT-23S (35%), ANT-21S (33%), formed the interaction with Y308 for more than 30% of the simulation time and are the most stable fragments. The other ANT derivatives ANT-32R (21%), ANT-23R (20%), ANT-20R (20%) ANT-32S (19%), ANT-20R (19%), ANT-20S (16%), ANT-24S (15%), ANT-33R (6%) and ANT-33S (5%) are less stable but maintained the interaction for the entire simulations except for the capped ANT-derivatives that lost the interaction in a period of time between 45 ns and 90 ns.

The distances between the centroids of anthranilic acid ring and Y308 side chain were also monitored (Figure 74). The non-capped compounds maintained this distance below or around 5.5 Å resulting in a very stable interaction (distance average values < 5.40 Å) while for the capped one the distances increased with ANT-32S and ANT-33R exceeding the limit of 5.5 Å (5.88 Å and 6.05 Å, respectively). The data here presented suggested that the capping group considerably reduces the stability of the stacking leading ANT-32S and ANT-33R in losing the desired π-π stacking. However, ANT-32R and ANT-33S, such as all the non-capped compounds, can be considered good ligand candidates.

Considering the two compounds with the best K_D in the SYNTH domain (Table 19), ANT-20R showed a high number of frames where the interaction is formed (20%) maintaining the same trend already shown in molecular docking simulations. On the other hand, ANT-23S represents one of the most promising ligands in this *in silico* procedure, maintaining the interaction for 35% of the entire simulations, in contrast to the data previously discussed.

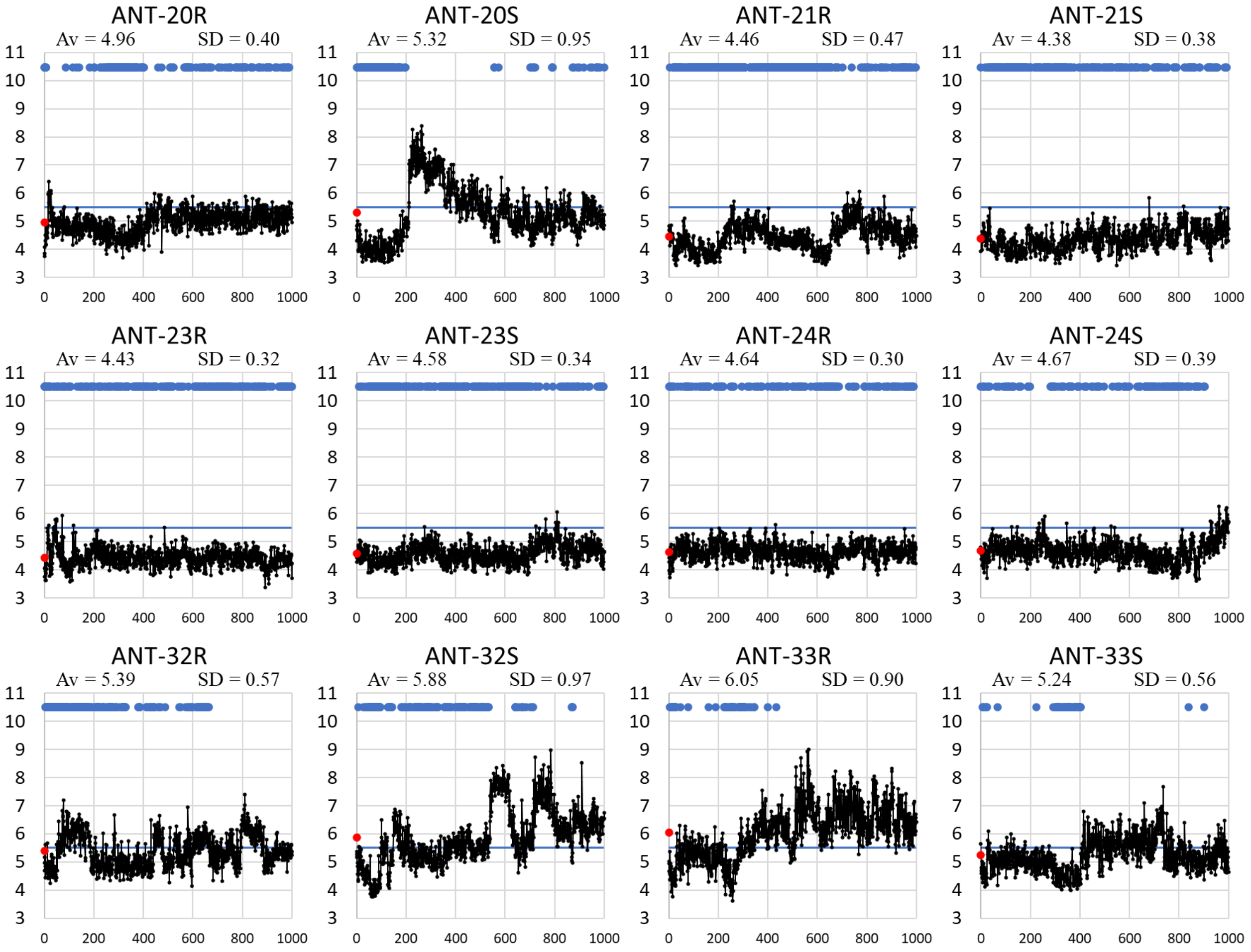


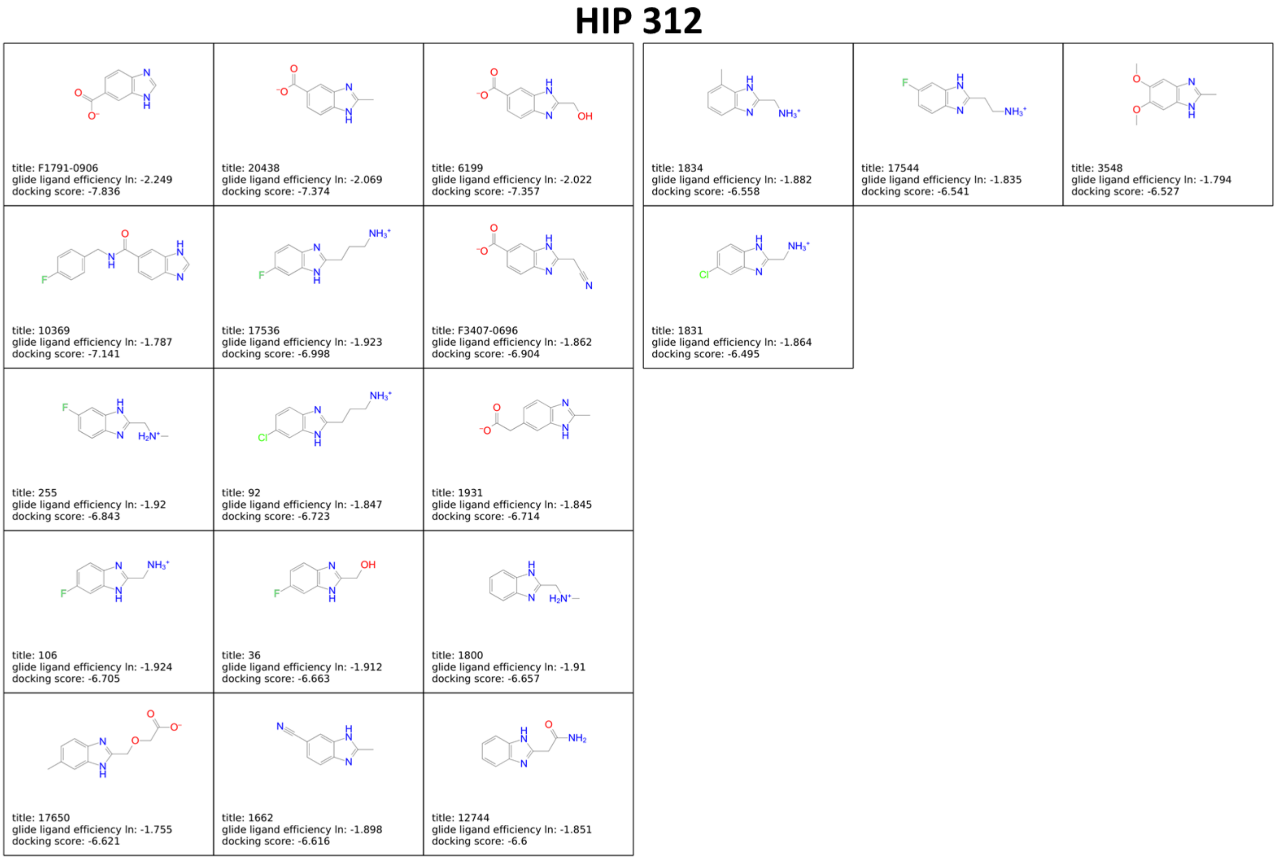
Figure 74: Time series of the π-π stacking with Y308 for designed ANT derivatives. In blue are highlighted the frames where the interaction is present. Every compound keeps itself close to Y308 side chain.

All the ANT-derivatives, with the exception of ANT-21S (lack of *in vitro* data), ANT-32R and ANT-33S (lost the key interaction), showed agreement between data obtained by MD simulations and the ones obtained by TS experiments. As expected, these compounds can be used for further modification to improve binding affinity.

4.11 Conclusion

FBDD starting from a VS into the SYNTH site of the Rel$_{Seq}$, identified twelve fragments, based on benzimidazole, aminobenzoic acid, indole, benzoxazole, benzotriazole and triazolo-pyrimidine scaffolds, that selectively bind to the domain. Among these scaffolds, the aminobenzoic one is the most promising due to its selectivity tested in TSA. Therefore, starting from the 2-aminobenzoic acid scaffold (anthranilic acid), a set of compounds, named ANT-derivatives, was designed and as expected, they were able to bind the SYNTH domain with sub mM affinity.

4.13 Appendix



title: F1791-0906
glide ligand efficiency ln: -2.249
docking score: -7.836

title: 20438
glide ligand efficiency ln: -2.069
docking score: -7.374

title: 6199
glide ligand efficiency ln: -2.022
docking score: -7.357

title: 1834
glide ligand efficiency ln: -1.882
docking score: -6.558

title: 17544
glide ligand efficiency ln: -1.835
docking score: -6.541

title: 3548
glide ligand efficiency ln: -1.794
docking score: -6.527

title: 10369
glide ligand efficiency ln: -1.787
docking score: -7.141

title: 17536
glide ligand efficiency ln: -1.923
docking score: -6.998

title: F3407-0696
glide ligand efficiency ln: -1.862
docking score: -6.904

title: 1831
glide ligand efficiency ln: -1.864
docking score: -6.495

title: 255
glide ligand efficiency ln: -1.92
docking score: -6.843

title: 92
glide ligand efficiency ln: -1.847
docking score: -6.723

title: 1931
glide ligand efficiency ln: -1.845
docking score: -6.714

title: 106
glide ligand efficiency ln: -1.924
docking score: -6.705

title: 36
glide ligand efficiency ln: -1.912
docking score: -6.663

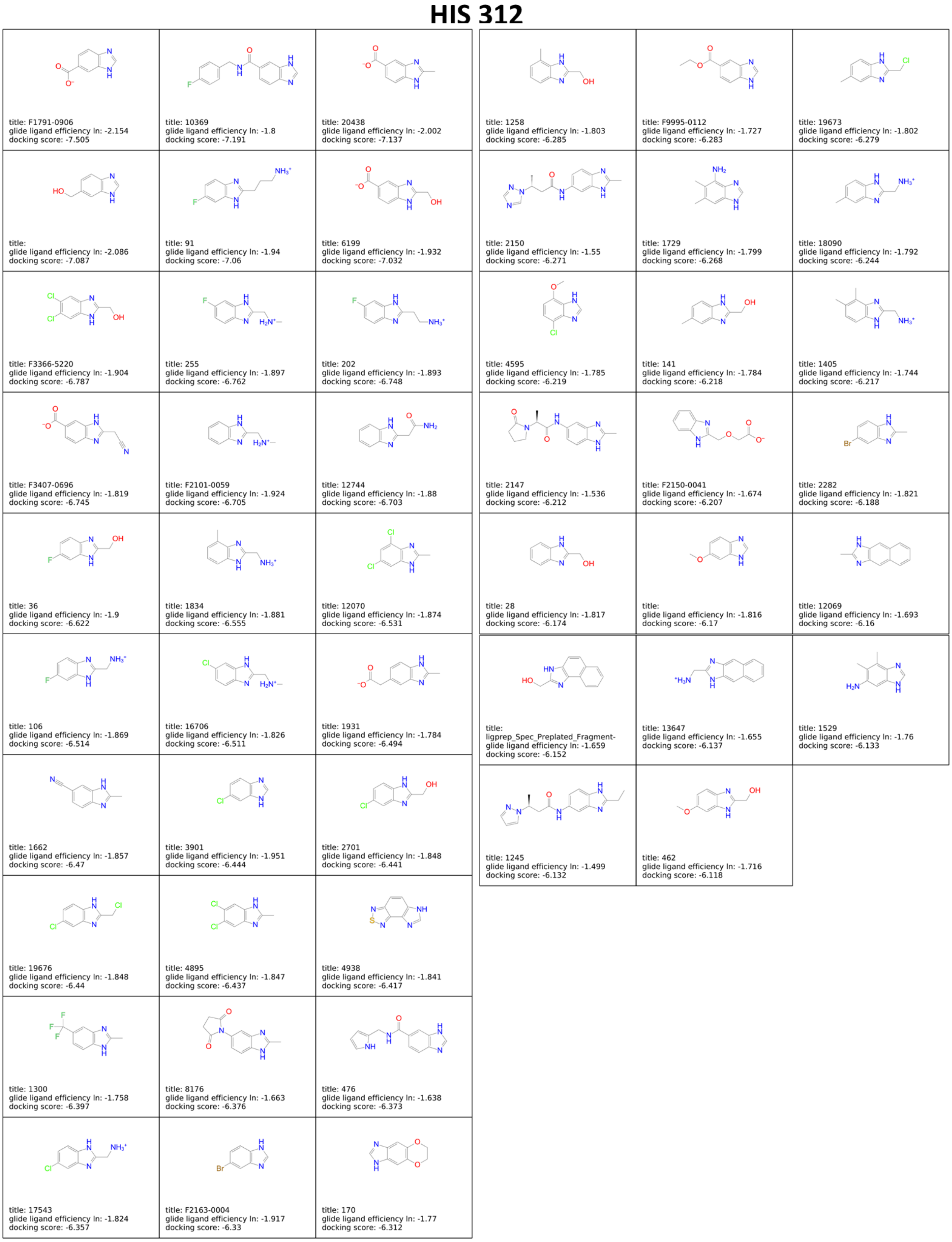
title: 1800
glide ligand efficiency ln: -1.91
docking score: -6.657

title: 17650
glide ligand efficiency ln: -1.755
docking score: -6.621

title: 1662
glide ligand efficiency ln: -1.898
docking score: -6.616

title: 12744
glide ligand efficiency ln: -1.851
docking score: -6.6

Figure 75: Benzimidazole fragments present in the top 1% of the VS performed into HIP312 grid



HIS 312

title	glide ligand efficiency ln	docking score
F1791-0906	-2.154	-7.505
10369	-1.8	-7.191
20438	-2.002	-7.137
	-2.086	-7.087
91	-1.94	-7.06
6199	-1.932	-7.032
F3366-5220	-1.904	-6.787
255	-1.897	-6.762
202	-1.893	-6.748
F3407-0696	-1.819	-6.745
F2101-0059	-1.924	-6.705
12744	-1.88	-6.703
36	-1.9	-6.622
1834	-1.881	-6.555
12070	-1.874	-6.531
106	-1.869	-6.514
16706	-1.826	-6.511
1931	-1.784	-6.494
1662	-1.857	-6.47
3901	-1.951	-6.444
2701	-1.848	-6.441
19676	-1.848	-6.44
4895	-1.847	-6.437
4938	-1.841	-6.417
1300	-1.758	-6.397
8176	-1.663	-6.376
476	-1.638	-6.373
17543	-1.824	-6.357
F2163-0004	-1.917	-6.33
170	-1.77	-6.312
1258	-1.803	-6.285
F9995-0112	-1.727	-6.283
19673	-1.802	-6.279
2150	-1.55	-6.271
1729	-1.799	-6.268
18090	-1.792	-6.244
4595	-1.785	-6.219
141	-1.784	-6.218
1405	-1.744	-6.217
2147	-1.536	-6.212
F2150-0041	-1.674	-6.207
2282	-1.821	-6.188
28	-1.817	-6.174
	-1.816	-6.17
12069	-1.693	-6.16
ligprep_Spec_Preplated_Fragment-	-1.659	-6.152
13647	-1.655	-6.137
1529	-1.76	-6.133
1245	-1.499	-6.132
462	-1.716	-6.118

Figure 76: Benzimidazole fragments present in the top 1% of the VS performed into the HIS312 grid

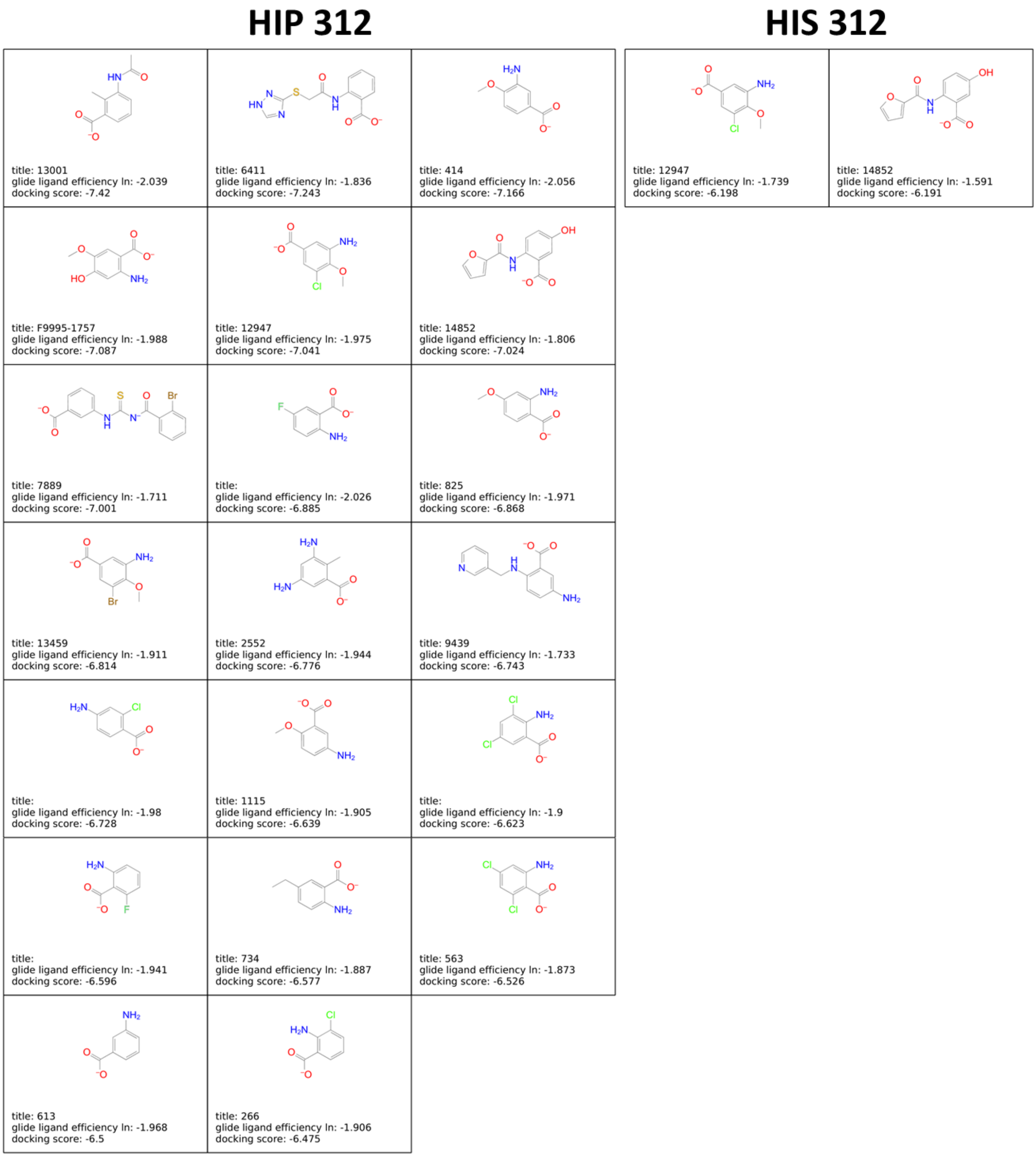

HIP 312
HIS 312
title: 13001
glide ligand efficiency ln: -2.039
docking score: -7.42
title: 6411
glide ligand efficiency ln: -1.836
docking score: -7.243
title: 414
glide ligand efficiency ln: -2.056
docking score: -7.166
title: 12947
glide ligand efficiency ln: -1.739
docking score: -6.198
title: 14852
glide ligand efficiency ln: -1.591
docking score: -6.191
title: F9995-1757
glide ligand efficiency ln: -1.988
docking score: -7.087
title: 12947
glide ligand efficiency ln: -1.975
docking score: -7.041
title: 14852
glide ligand efficiency ln: -1.806
docking score: -7.024
title: 7889
glide ligand efficiency ln: -1.711
docking score: -7.001
title:
glide ligand efficiency ln: -2.026
docking score: -6.885
title: 825
glide ligand efficiency ln: -1.971
docking score: -6.868
title: 13459
glide ligand efficiency ln: -1.911
docking score: -6.814
title: 2552
glide ligand efficiency ln: -1.944
docking score: -6.776
title: 9439
glide ligand efficiency ln: -1.733
docking score: -6.743
title:
glide ligand efficiency ln: -1.98
docking score: -6.728
title: 1115
glide ligand efficiency ln: -1.905
docking score: -6.639
title:
glide ligand efficiency ln: -1.9
docking score: -6.623
title:
glide ligand efficiency ln: -1.941
docking score: -6.596
title: 734
glide ligand efficiency ln: -1.887
docking score: -6.577
title: 563
glide ligand efficiency ln: -1.873
docking score: -6.526
title: 613
glide ligand efficiency ln: -1.968
docking score: -6.5
title: 266
glide ligand efficiency ln: -1.906
docking score: -6.475

Figure 77: Aminobenzoic acid-fragments present in the top 1% of the VS performed into the HIP312 (left) and HIS312 (right) grids

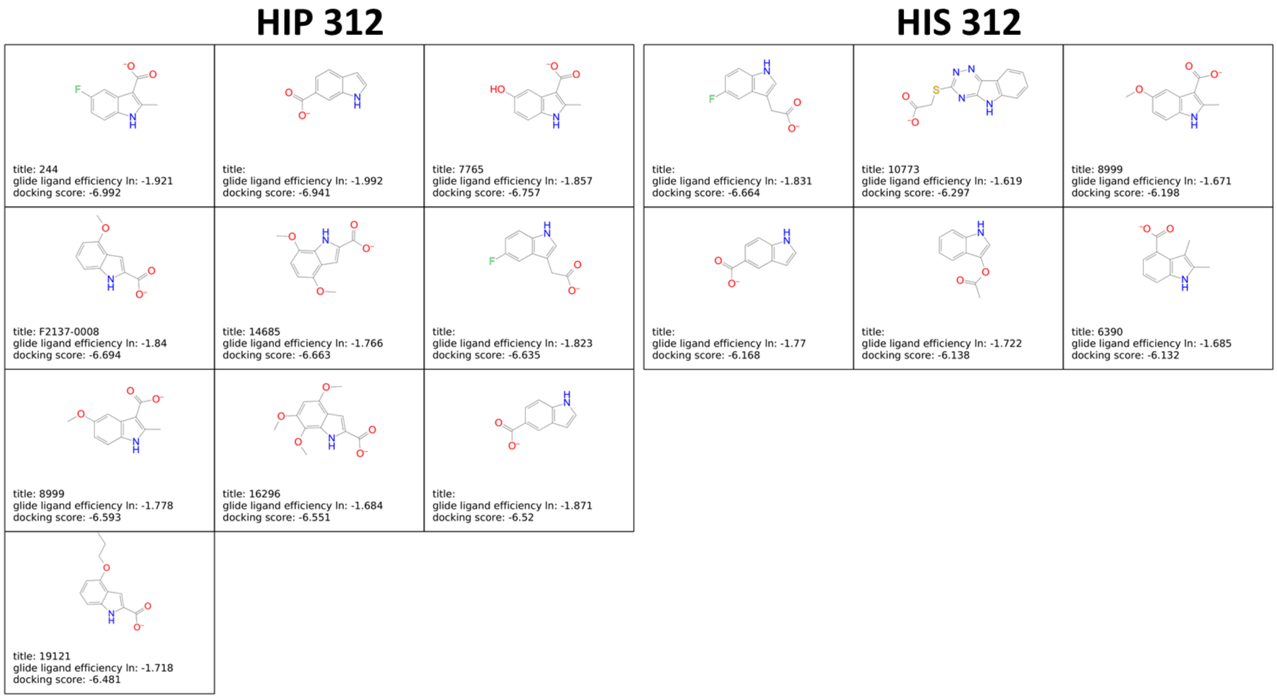

HIP 312
HIS 312
title: 244
glide ligand efficiency ln: -1.921
docking score: -6.992
title:
glide ligand efficiency ln: -1.992
docking score: -6.941
title: 7765
glide ligand efficiency ln: -1.857
docking score: -6.757
title: F2137-0008
glide ligand efficiency ln: -1.84
docking score: -6.694
title: 14685
glide ligand efficiency ln: -1.766
docking score: -6.663
title:
glide ligand efficiency ln: -1.823
docking score: -6.635
title: 8999
glide ligand efficiency ln: -1.778
docking score: -6.593
title: 16296
glide ligand efficiency ln: -1.684
docking score: -6.551
title:
glide ligand efficiency ln: -1.871
docking score: -6.52
title: 19121
glide ligand efficiency ln: -1.718
docking score: -6.481
title:
glide ligand efficiency ln: -1.831
docking score: -6.664
title: 10773
glide ligand efficiency ln: -1.619
docking score: -6.297
title: 8999
glide ligand efficiency ln: -1.671
docking score: -6.198
title:
glide ligand efficiency ln: -1.77
docking score: -6.168
title:
glide ligand efficiency ln: -1.722
docking score: -6.138
title: 6390
glide ligand efficiency ln: -1.685
docking score: -6.132

Figure 78: Indole- fragments present in the top 1% of the VS performed into HIP312 (left) and HIS312 (right) grids

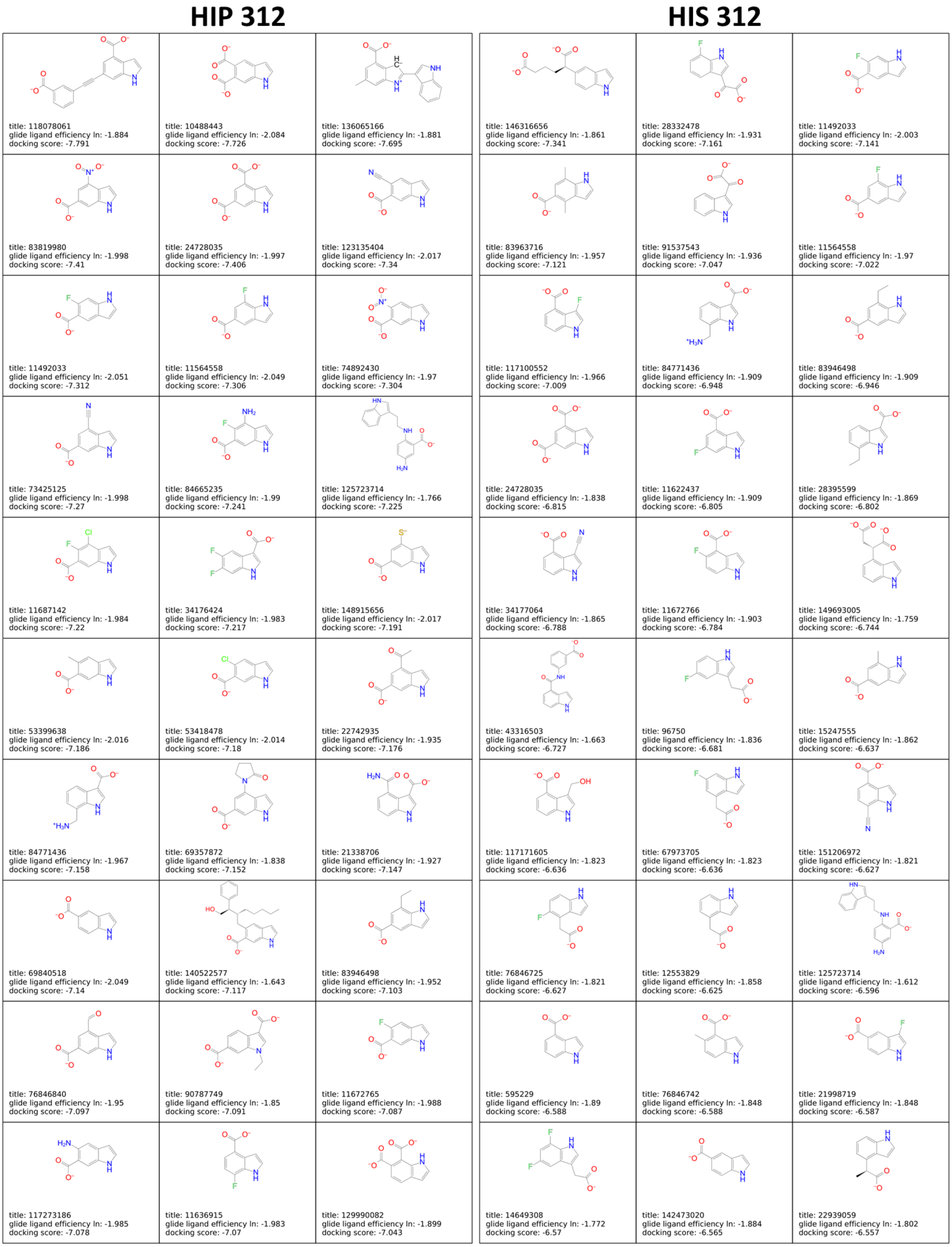

HIP 312
HIS 312
title: 118078061 glide ligand efficiency ln: -1.884 docking score: -7.791
title: 10488443 glide ligand efficiency ln: -2.084 docking score: -7.726
title: 136065166 glide ligand efficiency ln: -1.881 docking score: -7.695
title: 146316656 glide ligand efficiency ln: -1.861 docking score: -7.341
title: 28332478 glide ligand efficiency ln: -1.931 docking score: -7.161
title: 11492033 glide ligand efficiency ln: -2.003 docking score: -7.141
title: 83819980 glide ligand efficiency ln: -1.998 docking score: -7.41
title: 24728035 glide ligand efficiency ln: -1.997 docking score: -7.406
title: 123135404 glide ligand efficiency ln: -2.017 docking score: -7.34
title: 83963716 glide ligand efficiency ln: -1.957 docking score: -7.121
title: 91537543 glide ligand efficiency ln: -1.936 docking score: -7.047
title: 11564558 glide ligand efficiency ln: -1.97 docking score: -7.022
title: 11492033 glide ligand efficiency ln: -2.051 docking score: -7.312
title: 11564558 glide ligand efficiency ln: -2.049 docking score: -7.306
title: 74892430 glide ligand efficiency ln: -1.97 docking score: -7.304
title: 117100552 glide ligand efficiency ln: -1.966 docking score: -7.009
title: 84771436 glide ligand efficiency ln: -1.909 docking score: -6.948
title: 83946498 glide ligand efficiency ln: -1.909 docking score: -6.946
title: 73425125 glide ligand efficiency ln: -1.998 docking score: -7.27
title: 84665235 glide ligand efficiency ln: -1.99 docking score: -7.241
title: 125723714 glide ligand efficiency ln: -1.766 docking score: -7.225
title: 24728035 glide ligand efficiency ln: -1.838 docking score: -6.815
title: 11622437 glide ligand efficiency ln: -1.909 docking score: -6.805
title: 28395599 glide ligand efficiency ln: -1.869 docking score: -6.802
title: 11687142 glide ligand efficiency ln: -1.984 docking score: -7.22
title: 34176424 glide ligand efficiency ln: -1.983 docking score: -7.217
title: 148915656 glide ligand efficiency ln: -2.017 docking score: -7.191
title: 34177064 glide ligand efficiency ln: -1.865 docking score: -6.788
title: 11672766 glide ligand efficiency ln: -1.903 docking score: -6.784
title: 149693005 glide ligand efficiency ln: -1.759 docking score: -6.744
title: 53399638 glide ligand efficiency ln: -2.016 docking score: -7.186
title: 53418478 glide ligand efficiency ln: -2.014 docking score: -7.18
title: 22742935 glide ligand efficiency ln: -1.935 docking score: -7.176
title: 43316503 glide ligand efficiency ln: -1.663 docking score: -6.727
title: 96750 glide ligand efficiency ln: -1.836 docking score: -6.681
title: 15247555 glide ligand efficiency ln: -1.862 docking score: -6.637
title: 84771436 glide ligand efficiency ln: -1.967 docking score: -7.158
title: 69357872 glide ligand efficiency ln: -1.838 docking score: -7.152
title: 21338706 glide ligand efficiency ln: -1.927 docking score: -7.147
title: 117171605 glide ligand efficiency ln: -1.823 docking score: -6.636
title: 67973705 glide ligand efficiency ln: -1.823 docking score: -6.636
title: 151206972 glide ligand efficiency ln: -1.821 docking score: -6.627
title: 69840518 glide ligand efficiency ln: -2.049 docking score: -7.14
title: 140522577 glide ligand efficiency ln: -1.643 docking score: -7.117
title: 83946498 glide ligand efficiency ln: -1.952 docking score: -7.103
title: 76846725 glide ligand efficiency ln: -1.821 docking score: -6.627
title: 12553829 glide ligand efficiency ln: -1.858 docking score: -6.625
title: 125723714 glide ligand efficiency ln: -1.612 docking score: -6.596
title: 76846840 glide ligand efficiency ln: -1.95 docking score: -7.097
title: 90787749 glide ligand efficiency ln: -1.85 docking score: -7.091
title: 11672765 glide ligand efficiency ln: -1.988 docking score: -7.087
title: 595229 glide ligand efficiency ln: -1.89 docking score: -6.588
title: 76846742 glide ligand efficiency ln: -1.848 docking score: -6.588
title: 21998719 glide ligand efficiency ln: -1.848 docking score: -6.587
title: 117273186 glide ligand efficiency ln: -1.985 docking score: -7.078
title: 11636915 glide ligand efficiency ln: -1.983 docking score: -7.07
title: 129990082 glide ligand efficiency ln: -1.899 docking score: -7.043
title: 14649308 glide ligand efficiency ln: -1.772 docking score: -6.57
title: 142473020 glide ligand efficiency ln: -1.884 docking score: -6.565
title: 22939059 glide ligand efficiency ln: -1.802 docking score: -6.557

Figure 79: Top 30 indole fragments resulting from the library expansion performed into HIP312 (left) and HIS312 (right) grids

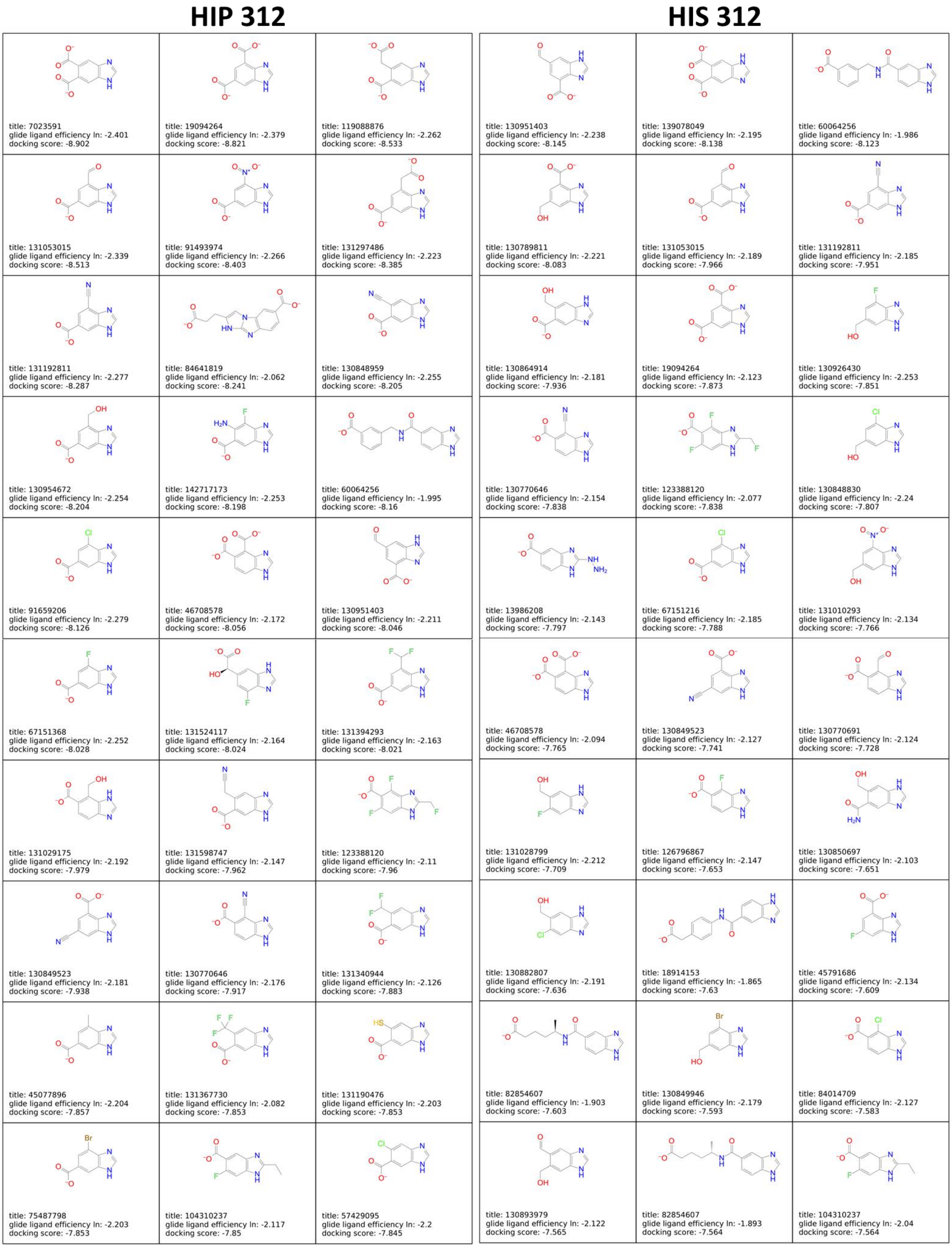
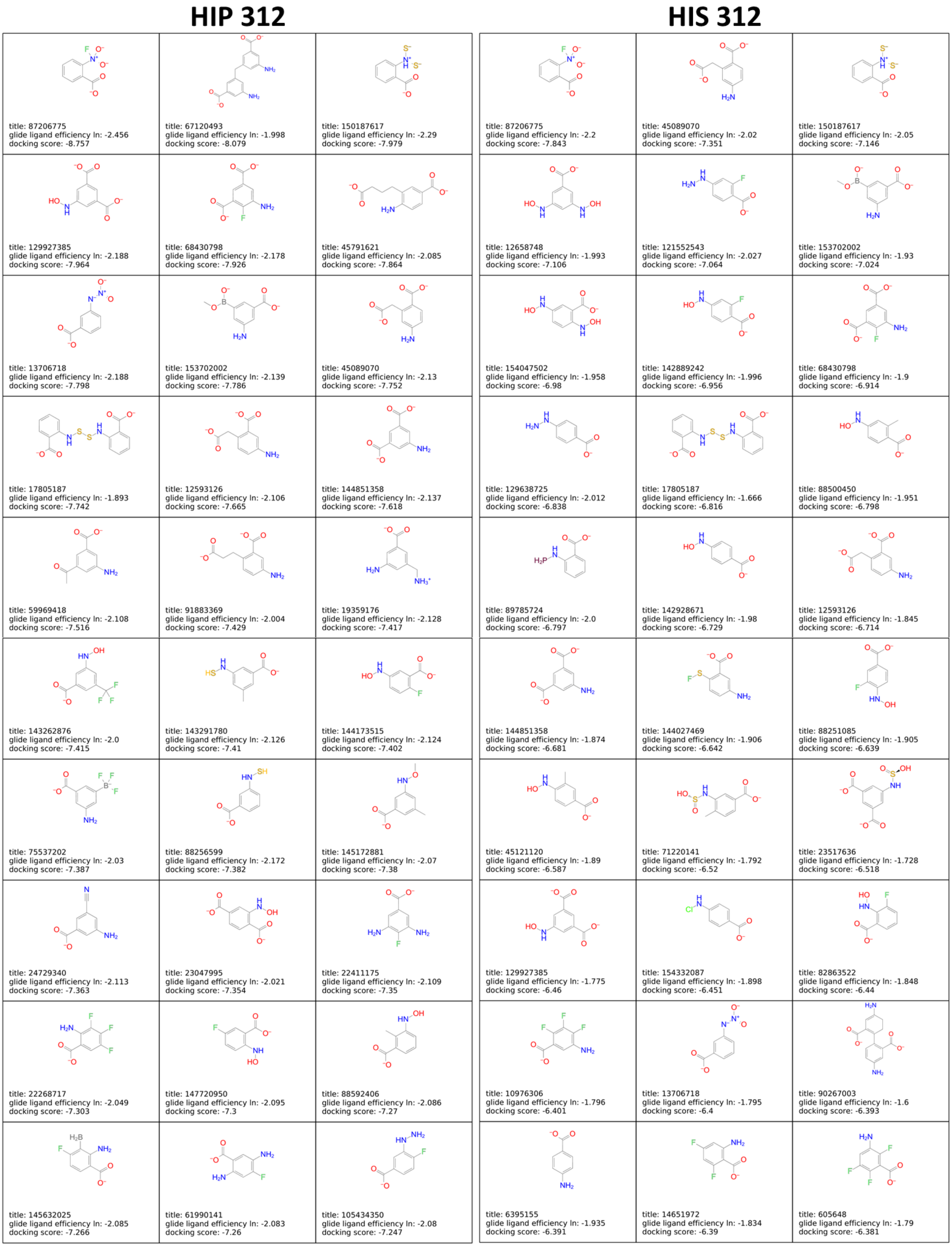


Figure 80: Top 30 benzimidazole fragments resulting from library expansion performed into HIP312 (left) and HIS312 (right) grids



HIP 312

title: 87206775
glide ligand efficiency ln: -2.456
docking score: -8.757

title: 67120493
glide ligand efficiency ln: -1.998
docking score: -8.079

title: 150187617
glide ligand efficiency ln: -2.29
docking score: -7.979

title: 129927385
glide ligand efficiency ln: -2.188
docking score: -7.964

title: 68430798
glide ligand efficiency ln: -2.178
docking score: -7.926

title: 45791621
glide ligand efficiency ln: -2.085
docking score: -7.864

title: 13706718
glide ligand efficiency ln: -2.188
docking score: -7.798

title: 153702002
glide ligand efficiency ln: -2.139
docking score: -7.786

title: 45089070
glide ligand efficiency ln: -2.13
docking score: -7.752

title: 17805187
glide ligand efficiency ln: -1.893
docking score: -7.742

title: 12593126
glide ligand efficiency ln: -2.106
docking score: -7.665

title: 144851358
glide ligand efficiency ln: -2.137
docking score: -7.618

title: 59969418
glide ligand efficiency ln: -2.108
docking score: -7.516

title: 91883369
glide ligand efficiency ln: -2.004
docking score: -7.429

title: 19359176
glide ligand efficiency ln: -2.128
docking score: -7.417

title: 143262876
glide ligand efficiency ln: -2.0
docking score: -7.415

title: 143291780
glide ligand efficiency ln: -2.126
docking score: -7.41

title: 144173515
glide ligand efficiency ln: -2.124
docking score: -7.402

title: 75537202
glide ligand efficiency ln: -2.03
docking score: -7.387

title: 88256599
glide ligand efficiency ln: -2.172
docking score: -7.382

title: 145172881
glide ligand efficiency ln: -2.07
docking score: -7.38

title: 24729340
glide ligand efficiency ln: -2.113
docking score: -7.363

title: 23047995
glide ligand efficiency ln: -2.021
docking score: -7.354

title: 22411175
glide ligand efficiency ln: -2.109
docking score: -7.35

title: 22268717
glide ligand efficiency ln: -2.049
docking score: -7.303

title: 147720950
glide ligand efficiency ln: -2.095
docking score: -7.3

title: 88592406
glide ligand efficiency ln: -2.086
docking score: -7.27

title: 145632025
glide ligand efficiency ln: -2.085
docking score: -7.266

title: 61990141
glide ligand efficiency ln: -2.083
docking score: -7.26

title: 105434350
glide ligand efficiency ln: -2.08
docking score: -7.247

HIS 312

title: 87206775
glide ligand efficiency ln: -2.2
docking score: -7.843

title: 45089070
glide ligand efficiency ln: -2.02
docking score: -7.351

title: 150187617
glide ligand efficiency ln: -2.05
docking score: -7.146

title: 12658748
glide ligand efficiency ln: -1.993
docking score: -7.106

title: 121552543
glide ligand efficiency ln: -2.027
docking score: -7.064

title: 153702002
glide ligand efficiency ln: -1.93
docking score: -7.024

title: 154047502
glide ligand efficiency ln: -1.958
docking score: -6.98

title: 142889242
glide ligand efficiency ln: -1.996
docking score: -6.956

title: 68430798
glide ligand efficiency ln: -1.9
docking score: -6.914

title: 129638725
glide ligand efficiency ln: -2.012
docking score: -6.838

title: 17805187
glide ligand efficiency ln: -1.666
docking score: -6.816

title: 88500450
glide ligand efficiency ln: -1.951
docking score: -6.798

title: 89785724
glide ligand efficiency ln: -2.0
docking score: -6.797

title: 142928671
glide ligand efficiency ln: -1.98
docking score: -6.729

title: 12593126
glide ligand efficiency ln: -1.845
docking score: -6.714

title: 144851358
glide ligand efficiency ln: -1.874
docking score: -6.681

title: 144027469
glide ligand efficiency ln: -1.906
docking score: -6.642

title: 88251085
glide ligand efficiency ln: -1.905
docking score: -6.639

title: 45121120
glide ligand efficiency ln: -1.89
docking score: -6.587

title: 71220141
glide ligand efficiency ln: -1.792
docking score: -6.52

title: 23517636
glide ligand efficiency ln: -1.728
docking score: -6.518

title: 129927385
glide ligand efficiency ln: -1.775
docking score: -6.46

title: 154332087
glide ligand efficiency ln: -1.898
docking score: -6.451

title: 82863522
glide ligand efficiency ln: -1.848
docking score: -6.44

title: 10976306
glide ligand efficiency ln: -1.796
docking score: -6.401

title: 13706718
glide ligand efficiency ln: -1.795
docking score: -6.4

title: 90267003
glide ligand efficiency ln: -1.6
docking score: -6.393

title: 6395155
glide ligand efficiency ln: -1.935
docking score: -6.391

title: 14651972
glide ligand efficiency ln: -1.834
docking score: -6.39

title: 605648
glide ligand efficiency ln: -1.79
docking score: -6.381

Figure 81: Top 30 aminobenzoic acid fragments resulting from library expansion performed into HIP312 (left) and HIS312 (right) grids

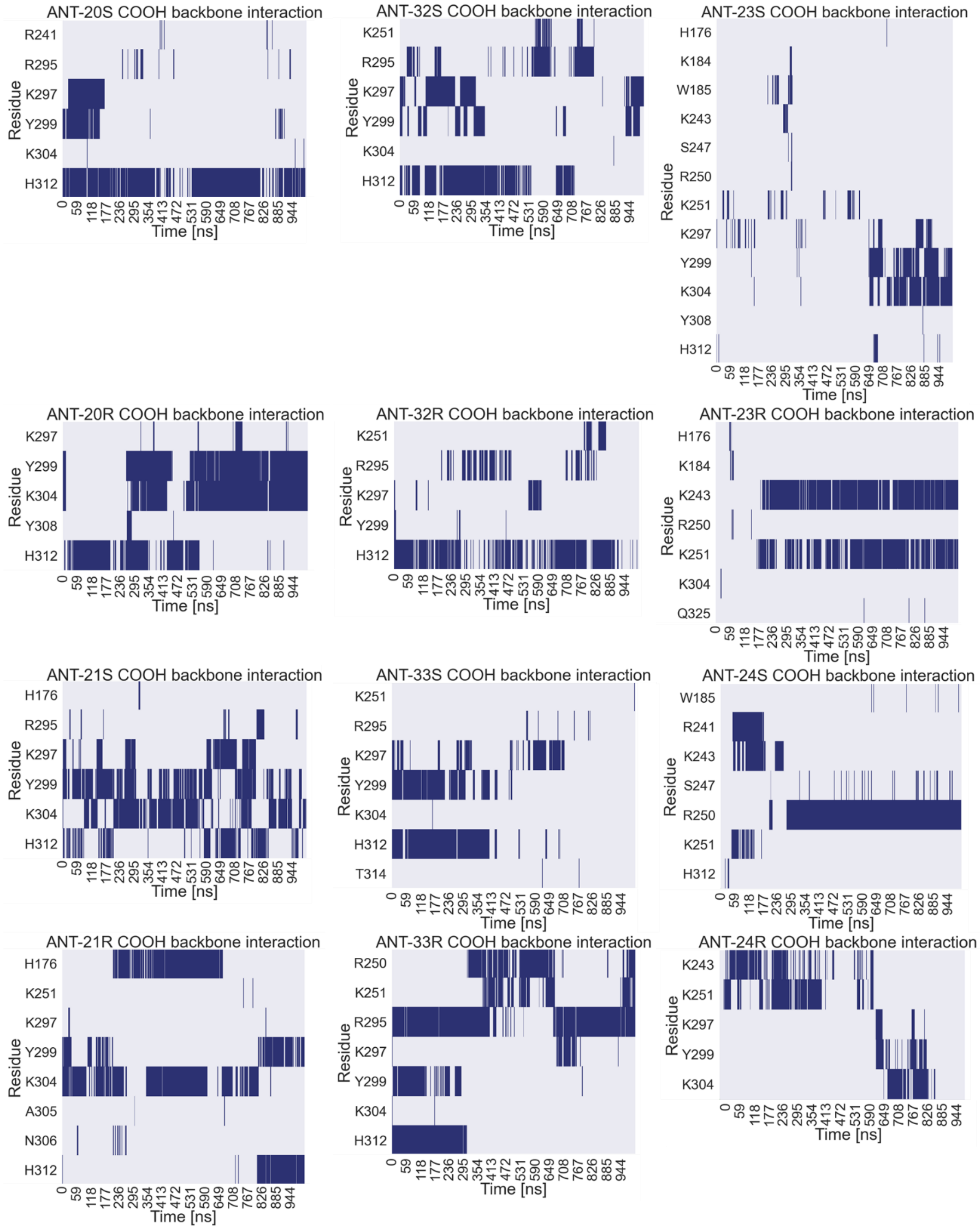


Figure 82: Time series of the salt bridge or H-bond formed between ANT derivatives amino acid backbone and Rel$_{Seq}$ SYNTH site residues. In blue are highlighted the frames where the interaction is present ANT-20R, ANT-20S, ANT-21R, ANT-21S, ANT-32R and ANT-32S formed stable interactions with the amino acids involved in X-ray GDP phosphate groups binding mode, ANT-23R and ANT-24S interacted with the residues located in α13 and ANT-33R preferred residues located in β3.

.

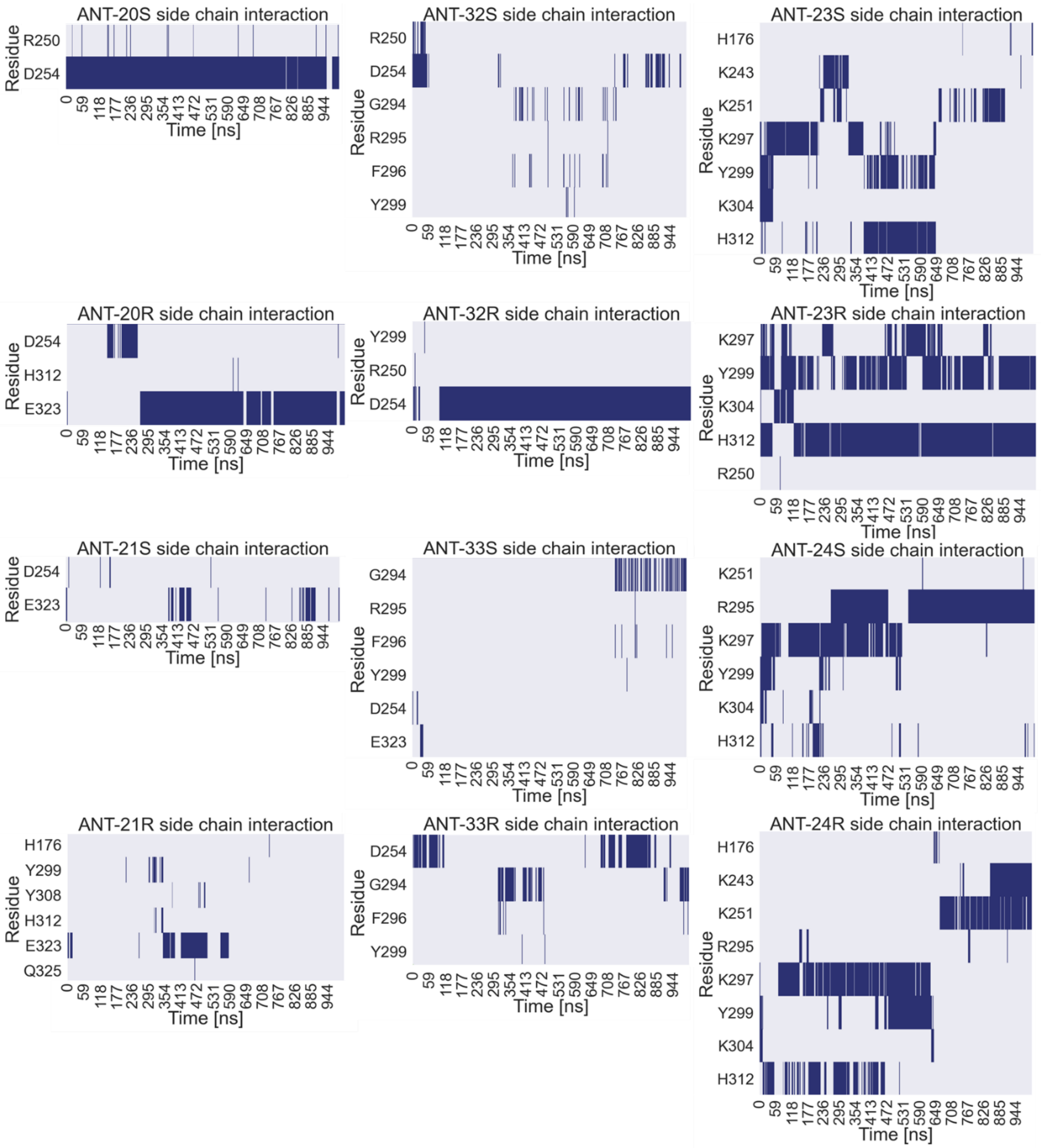


Figure 83: Time series of the salt bridge or H-bond formed between ANT derivatives amino acid side chain and Rel$_{Seq}$ SYNTH site residues. In blue are highlighted the frames where the interaction is present. ANT-20R and ANT-21S formed salt bridge with E323, ANT-20S, ANT-32R and ANT-33R interacted with residue D254, ANT-23R maintained stable interactions with residues involved in X-ray GDP phosphate groups, ANT-24S instead preferred the one located in β3 while ANT-23S and ANT-24R first formed salt bridge and H-bond with the residues involved in X-ray GDP phosphate groups binding mode and then moved toward the residues located in alpha 13.

CHAPTER 5: STUDY INTO THE HYDROLASE SITE

Docking and MD simulations into the HYD site of Rel$_{Seq}$ “HYD on” conformation (chain B) were performed on the 21 fragments selected from the *in silico* study described in chapter 4 to better understand the selectivity of these fragments. In addition, a VS campaign, using a workflow similar to the one used for the SYNTH site, was performed to identify chemotype able to bind to the HYD site.

5.1 Docking calculations

Molecular docking simulations were performed in the X-ray structure of Rel$_{Seq}$ chain B (PDB entry:1VJ7, residues 1-385)[62] cantered on the 5'-diphosphate 2':3'-cyclic monophosphate (ppG2’:3’p) cocrystalized in the HYD site.

5.1.1 System preparation

In the X-ray structure of Rel$_{Seq}$ chain B (PDB entry 1VJ7), the HYD active site includes a Mn^{2+} ion and the non-natural ligand ppG2’:3’p. The structure was prepared by using ‘Protein Preparation Wizard’ tool of Schrodinger suite. Crystallized water molecules were deleted except for the two water molecules coordinating the Mn^{2+} (WAT 2009 and WAT 2178). The gap Y113-M131 of the HYD domain and the two gaps R211-A216 and D254-Q261 in the SYNTH one were built by Prime tool. Residue protonation states were evaluated by using PROPKA at pH 7. According to Epik results (pH=7) ppG2’:3’p was considered fully deprotonated (total formal charge of -3). Finally, a restrained minimization was performed (OPLS_2005, converge for heavy atoms to RMSD of 0.3Å).
ppG2’:3’p binding mode was already described in chapter 1 (paragraph 1.4.1.1.2) and, as already discussed, it differs from the natural ligand for the presence of a mono phosphate cyclic group that replaces the pyrophosphate one in position 3’. Thus ppG2’:3’p coordinates the Mn^{2+} by a water bridge instead of a direct coordination by using an oxygen atom of the α phosphate group (Figure 12). For this reason, two grid models were generated depending on the presence (WATER grid) or absence (NO WATER grid) of the water molecules coordinating the Mn^{2+}.

5.1.2 Docking set up

Both grid models (WATER and NO WATER) were generated by using GLIDE and the OPLS_2005 force field with default parameters. A cubic region 23.9 Å centred on ppG2':3'p molecule was used with an inner box side set to 10 Å.

The docking protocol was set up on the crystalized ligand into the HYD binding site of the WATER grid model using GLIDE v8.0 in the SP method and OPLS_2005 force field. The protein was considered as a rigid body while the ligand was set free to move (no ring sampling was performed and the option 'use enhanced sampling' was increase by two times). Five poses were saved out of ten post-minimized structures, and no Epik state penalty was added to the docking score. Three of the resulting docking poses, including the best one, well reproduces ppG2':3'p binding mode. The RMSD calculated between the heavy atoms of ppG2':3'p X-ray structure and ppG2':3'p best pose was 0.40 Å, thus indicating a good overlap (Figure 84a). The same validated docking protocol was also tested on the NO WATER model: four out of five poses, including the best one, maintained the same X-ray interaction for guanine ring and 5' phosphate groups, while the 3' phosphate group coordinate Mn^{2+} directly instead of using a water bridge (Figure 84b). Due to this ligand shift toward Mn^{2+}, the RMSD between the ppG2':3'p best pose and ppG2':3'p X-ray structures is higher (1.14 Å) compared to the WATER grid result.

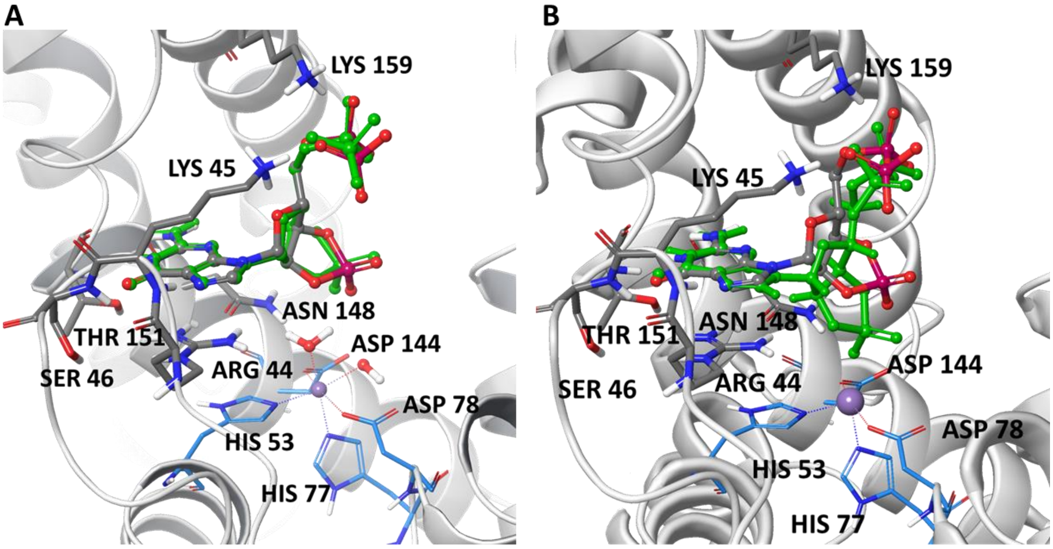


Figure 84: A) Superposition of ppG2':3'p X-Ray (grey) and the docking best pose into WATER grid (green). The RMSD calculated on ppG2':3'p heavy atoms is 0.40 Å B) Superposition of X-Ray (grey) and docking best pose into NO WATER grid (green) ppG2':3'p. The RMSD, calculated on ppG2':3'p heavy atoms, is 1.14Å

5.2 Docking results

Molecular docking simulations were performed into both WATER and NO WATER models (Table 21, Table 22, Figure 85 and Figure 86) saving ten poses for each fragment.

All the poses showed a π-cationic interaction with R44 (in Figure 85 and Figure 86 are shown one representative pose per each chemotype) and H-bonds with K45 backbone and T151 backbone or side chain. The binding mode of each fragment is similar into the two grids except for fragments A5, A6, B4, I2 and I5 where the rotation of the fragments into the site produces two different binding poses (Figure 85c and Figure 86c for a comparison). Docking scores and LE of the best poses are reported in Table 21 for both grids. B1, A1, A3 and A4 are the fragments with the highest LE. In the top-ranked poses none of the fragments coordinates the Mn^{2+} neither directly nor indirectly.

MD simulations were run to evaluate pose stability comparing the results to TSA data on fragments affinity for the HYD construct (Table 18).

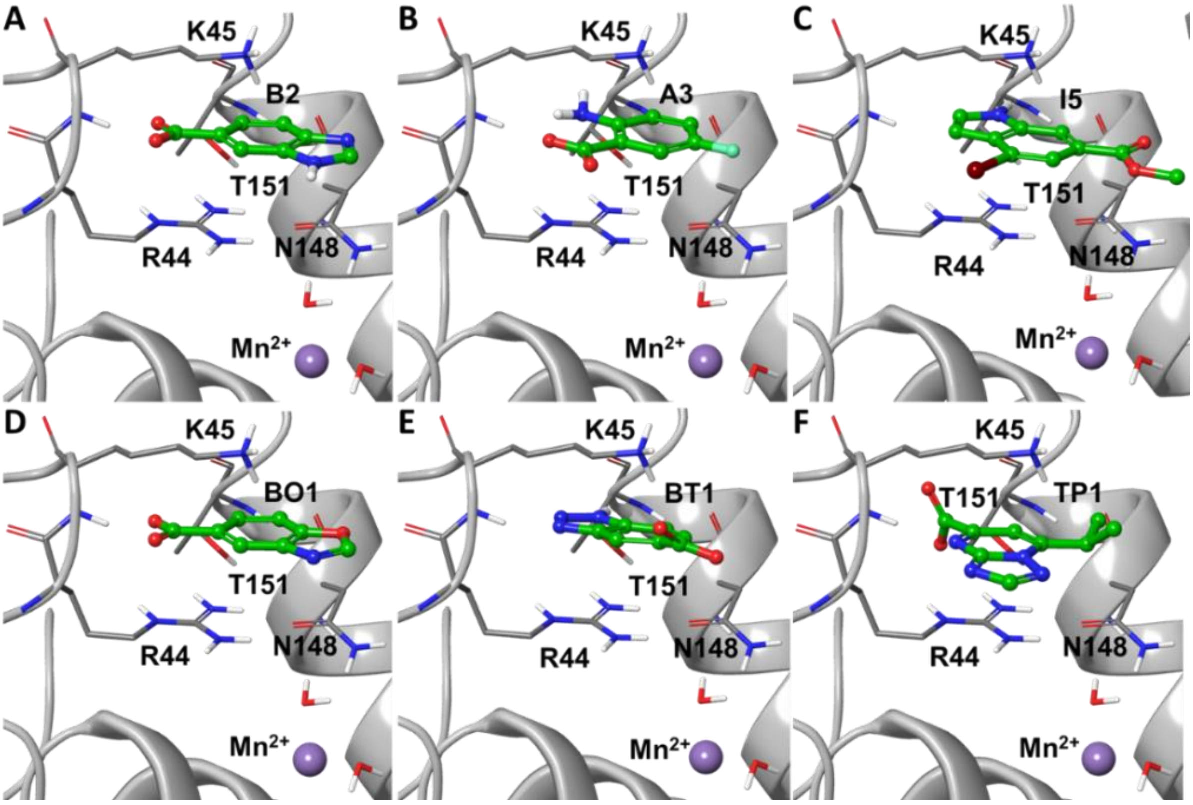


Figure 85: 3D representation of the best poses of fragments. A) A3, B) B2 C) I5 D) BO1 E) BT1 and F) TP1 docked into the HYD site of WATER grid. Fragments are shown in green ball and sticks, protein is shown in grey ribbon.

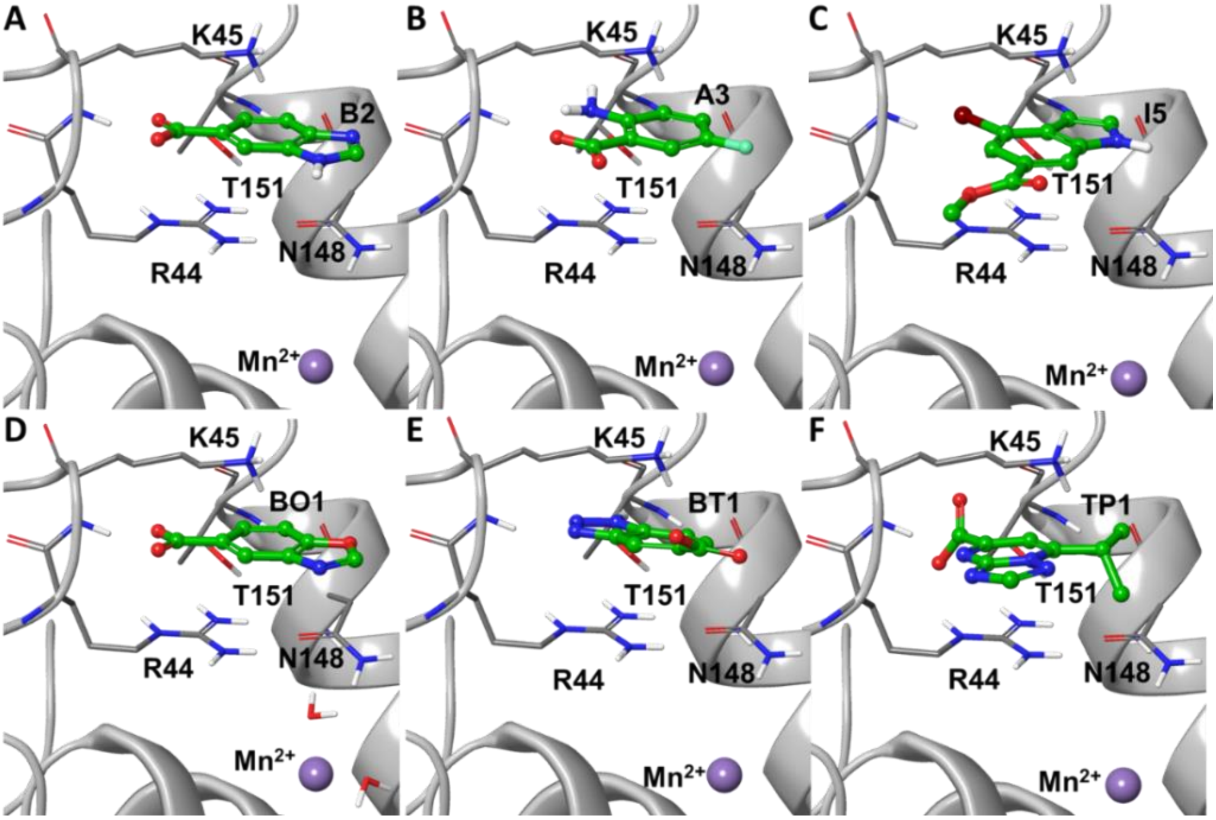


Figure 86: 3D representation of the best poses of fragments A) A3, B) B2 C) I5 D) BO1 E) BT1 and F) TP1 docked into the HYD site of NO WATER grid. Fragments are shown in green ball and sticks, protein is shown in grey ribbon.

Table 21: Docking score and ligand efficiency of the best poses of fragments docked into the HYD pocket of the WATER grid (blue) and NO WATER grid (orange)

COMPOUND	DOCKING SCORE	LE
B1	-7.376	-0.671
B2	-6.769	-0.564
B3	-6.644	-0.554
B4	-7.983	-0.532
A1	-6.506	-0.651
A2	-6.235	-0.520
A3	-6.753	-0.614
A4	-6.702	-0.670
A5	-6.76	-0.563
A6	-6.766	-0.615
A7	-6.627	-0.552
I1	-6.827	-0.569
I2	-6.704	-0.559
I3	-6.749	-0.562
I4	-7.341	-0.524
I5	-5.984	-0.427
I5COOH	-6.358	-0.489
I6	-7.112	-0.547
BO1	-6.764	-0.564
BT1	-6.724	-0.560
TP1	-7.835	-0.522

COMPOUND	DOCKING SCORE	LE
B1	-6.839	-0.622
B2	-6.851	-0.571
B3	-6.577	-0.548
B4	-7.156	-0.477
A1	-6.481	-0.648
A2	-6.5	-0.542
A3	-6.799	-0.618
A4	-6.656	-0.666
A5	-6.422	-0.535
A6	-7.257	-0.660
A7	-6.826	-0.569
I1	-6.774	-0.565
I2	-6.717	-0.560
I3	-6.77	-0.564
I4	-7.343	-0.524
I5	-5.799	-0.414
I5COOH	-6.348	-0.488
I6	-7.171	-0.552
BO1	-6.759	-0.563
BT1	-6.712	-0.559
TP1	-7.47	-0.498

5.3 Molecular dynamics simulations

Starting from the best poses obtained in the WATER model grid, 100 ns of MD simulations were performed using Desmond (NPT, TIP3P water, force field OPLS3e, dt = 2 fs, T = 300K, P = 1.01325 bar).

To test the ability of fragments to bind to the site several interactions were monitored during the simulations using SID tool (π-cationic interaction with R44, H-bond with N148, H-bond with T151, coordination to Mn^{2+}). The π-cationic interaction with R44 (considered formed if the charged nitrogen of guanidinium group is within 6 Å from aromatic ring centroid) and the H-bond (H--A distance ≤ 2.8 Å, DH-A angle ≥ 120°, XA-H angle ≥ 90°) with T151 (both backbone and side chain), two of the residues that inhibit hydrolytic activity if mutated,[62] are shown in Figure 87 and Figure 88, respectively.

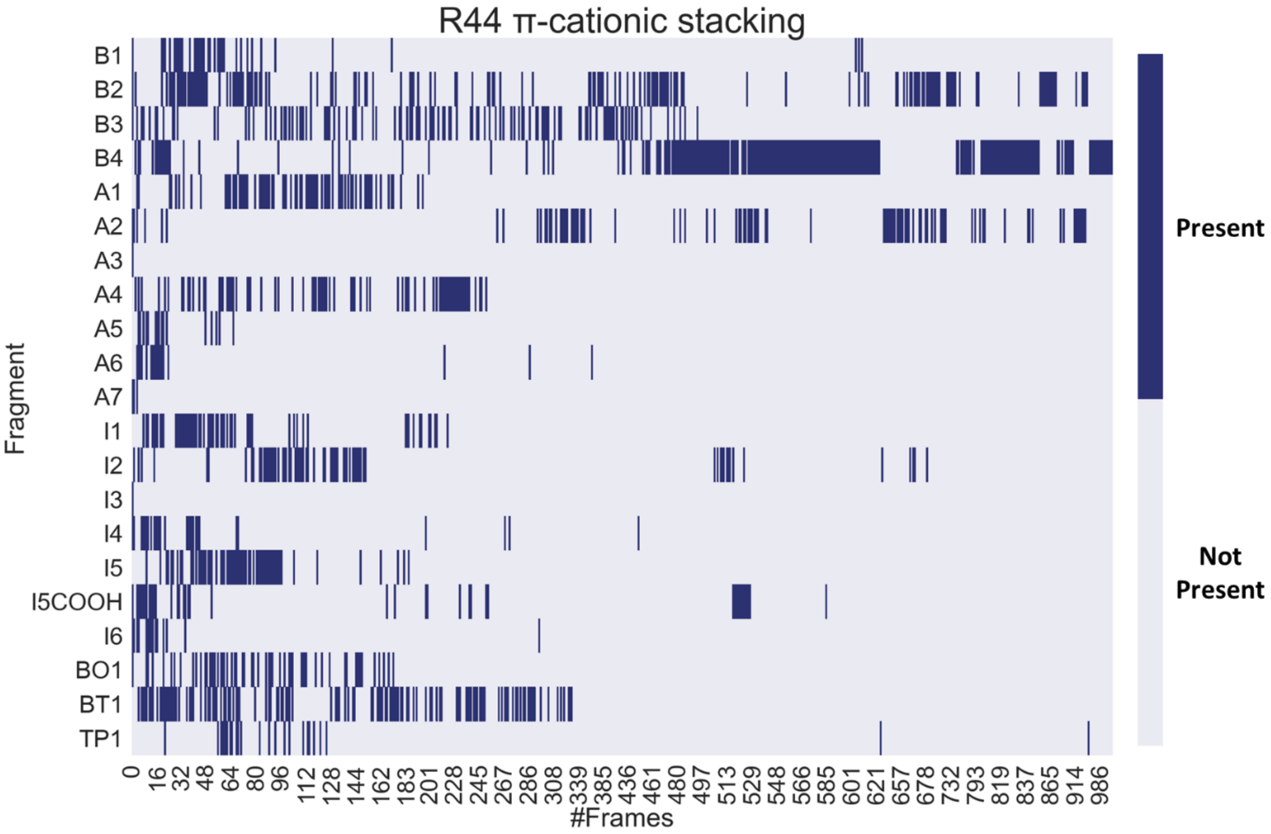


Figure 87: Time series of the π-cationic interaction between the selected fragments and R44. The frames where the interaction is formed are marked in blue while frames where the interaction is not present are marked in grey. Only fragments B2, B4, A2 and TP1 maintain the stacking during all the simulations. The graphic was produced by using Jupyter-lab

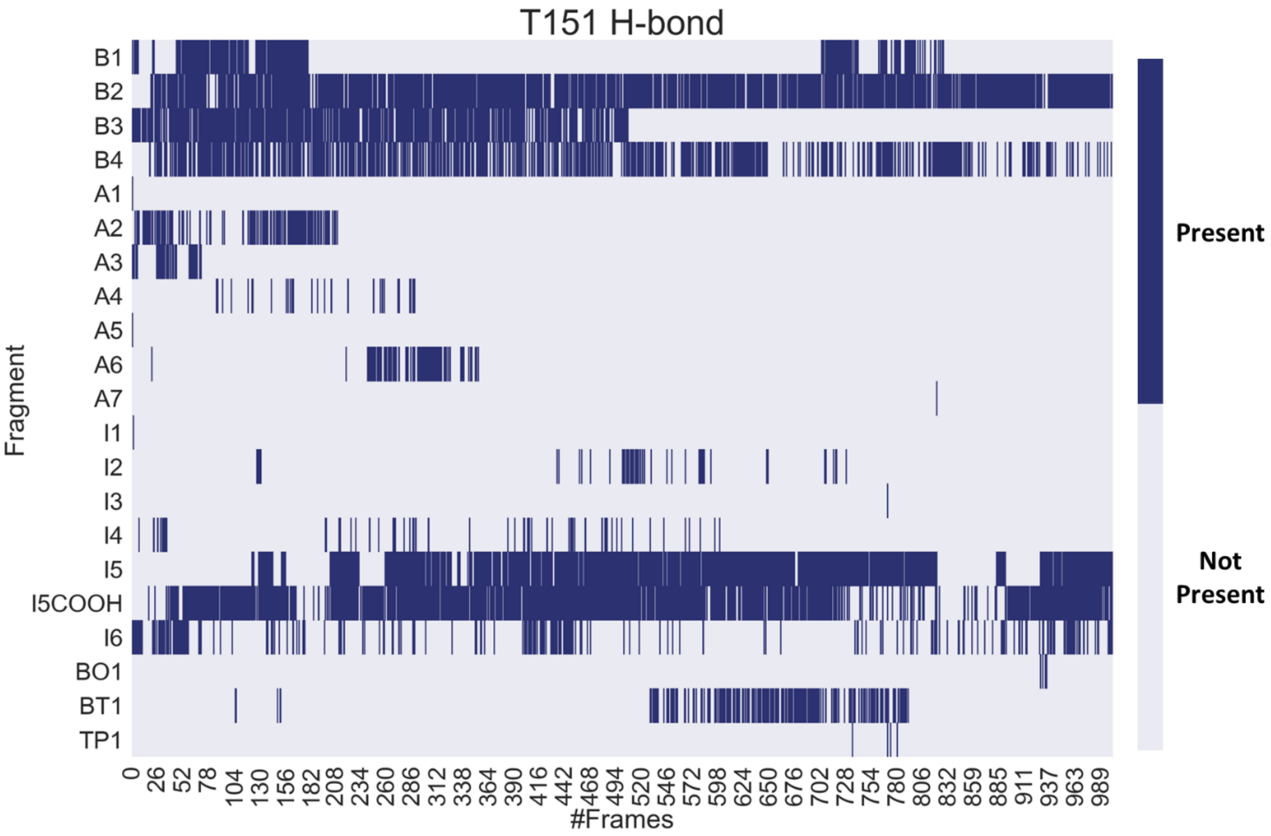


Figure 88: Time series of the H-bond between the selected fragments and T151. The frames where the interaction is formed are marked in blue while frames where the interaction is not present are marked in grey. Fragments B2, B4, I5, I5COOH and I6 maintain the stacking during all the simulations. The graphic was produced by using Jupyter-lab

All the fragments except for B2, B4 and A2, lost the interaction with R44 during the simulations (Figure 87), The H-bond with T151 and the fragments is mainly maintained for B2 (95%), I5COOH (88%), I5 (66%) and I6 (23%). All these fragments showed a K_D value in the millimolar range in the TS experiments with the HYD domain (see Table 18). The distances between I6 NH and the T151 backbone or side chain was monitored (figure 89) if we consider the H-bond formed when the distance between T151 OH or C=O and NH of I6 is equal to or less than 2.8 Å, the percentage of frames in which the interaction is formed, and so the stability of the interaction, increases from 23% (SID calculation) to 32% and with a threshold of 3.5 Å, the percentage increases up to 78%.

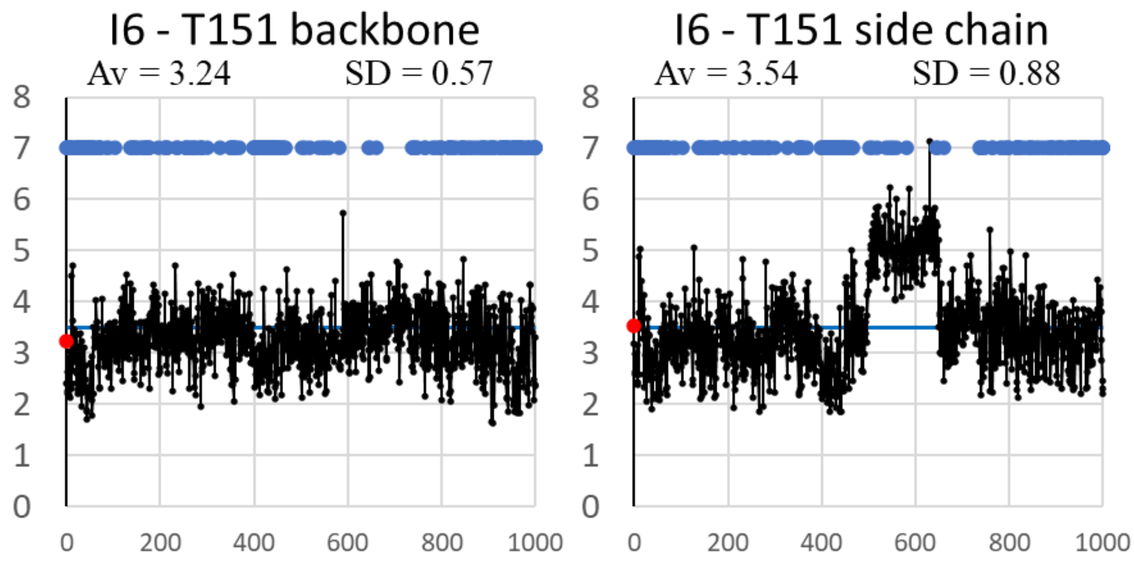


Figure 89: Distances between the centroids of T151 backbone (left) or side chain (right) and the NH of I6 indolic ring monitored during MD simulations (in black). The average (Av) and standard deviation (SD) of distances are also reported. The light-blue line define the maximum contact distance for an H-bond (3.5 Å). The presence of the H-bonds calculated by Desmond is reported in cyan.

The interaction between T151 and B3 is lost, even if this fragment has K_D comparable to B2 in TS assays, and, on the contrary, B4, that does not bind to the HD domain *in vitro,* maintained this interaction. Summing up, four out of five fragments that bind to the HYD domain *in vitro,* maintain the interaction with T151 during the simulations.

Among all the interactions analysed (π-cationic interaction with R44, H-bond with N148 data not shown, H-bond with T151, coordination to Mn^{2+} data not shown) the H-bond with the side chain or the backbone of residue T151 is quite in agreement with TS results and was used as filter for the VS campaign.

5.4 Fragment-based virtual screening into HYD site

The same seven libraries described in paragraph 4.1 were used to perform a VS campaign into the HYD site considering both WATER and NO WATER grid models. The workflow used (Figure 90) is similar to the one used for the SYNTH site VS (Figure 53) with the exception of the interaction filter. The Y308 aromatic interaction filter was replaced with the T151 H-bond filter (H--A distance ≤ 2.8 Å, DH-A angle ≥ 120°, XA-H angle ≥ 90°).

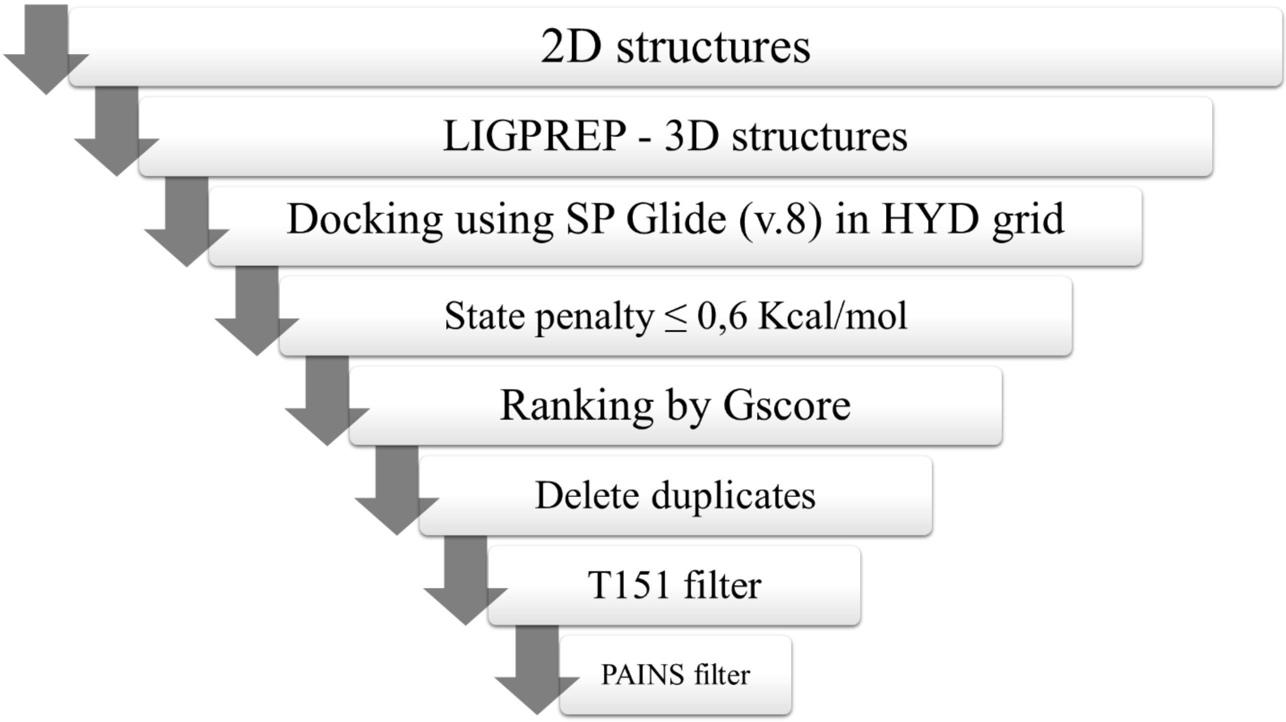


Figure 90: Workflow used for the VS into the HYD site

The 114,966 fragments were evaluated into the HYD site of both WATER and NO WATER grids by performing molecular docking simulations. Also in this screening, a single docking pose was saved for post-docking analysis. The fragments docked were filtered by state penalty (< 0.6 kcal/mol) and ranked by Glide score. The filter reduced the number of poses to 109,370 and 71,764 for WATER and NO WATER grids, respectively. Duplicates were then removed by using ‘Filter duplicates’ tool in Maestro (based on SMARTS) leading to a total of 83,153 fragments for WATER grid and 62,575 fragments for NO WATER grid to analyse. Considering the data obtained by MD simulations and discussed in the previous paragraph, the only poses inspected were the ones forming a H-bond (H--A distance ≤ 2.8 Å, DH-A angle ≥ 120°, XA-H angle ≥ 90°) with T151. The number of fragments was reduced to 2,146 and 2,899 for WATER and NO WATER grids, respectively. Finally, PAINS structures were deleted and 4,751 fragments, 2,041 and 2,710 for WATER and NO WATER grids respectively, were analysed.

None of the fragments selected for the SYNTH site (Figure 62) was found in the top 1% of the poses that passed the filters. Only two fragments passed the VS workflow: B1 and B2 of which only B2 binds to the HYD domain *in vitro*.

The EFs of fragments containing indole, benzimidazole and aminobenzoic acid scaffolds in the top 1% (20 for WATER and 27 for NO WATER grids, respectively) of the poses analysed were calculated to elucidate scaffold selectivity. As shown in Table 22 and Table 23, where EFs were calculated using equation 4.1 already used for the VS into SYNTH site, indole is the only chemotype that enrichs in both WATER and NO WATER models, benzimidazole enrichs only in the NO WATER grid while aminobenzoic acid structures are less populated in the top 1% of both grids compared to random selection (value < 1).

Table 22: Enrichment factor calculated on WATER model of fragment composed of scaffolds selected in SYNTH VS run

Chemotype	% Scaffold/TOT	% Scaffold/TOT 1%	EF
	(2041)	(20)	
Indole	30/2041 = 0.010 = 1.5%	1/ 20 = 0.014 = 5.0%	0.05/0.015 = 3.3
Benzimidazole	52/2041 = 0.025 = 2.5%	0/20 = 0 = 0	0/0.025 = 0
Aminobenzoic acids	85/2041 = 0.042 = 4.2%	0/20 = 0 = 0%	0/0.042 = 0

Table 23: Enrichment factor calculated on NO WATER model of fragment composed of scaffolds selected in SYNTH VS run

Chemotype	% Scaffold/TOT	% Scaffold/TOT 1%	EF
	(2710)	(27)	
Indole	83/2710 = 0.031 = 3.1%	2/27 = 0.074 = 7.4%	0.074/0.031 = 2.4
Benzimidazole	94/2710 = 0.035 = 3.5%	1/27 = 0.037 = 3.7 %	0.037/0.035 = 1.1
Aminobenzoic acids	113/2710 = 0.042 = 4.2%	1/27 = 0.037 = 3.7%	0.037/0.042 = 0.9

These data agree with TS experiments where three indoles, two benzimidazoles and none of the aminobenzoic acids bound to the HYD site.

Best ranked fragments belonging to benzimidazole, anthranilic acid and indole families were reported in Figure 91. The three fragments found in the NO WATER grid and the indole fragment found in the WATER one were found in the top 1% of the respective grids, while the aminobenzoic acid and the benzimidazole fragments found in the WATER grid were not found in the top 1%.

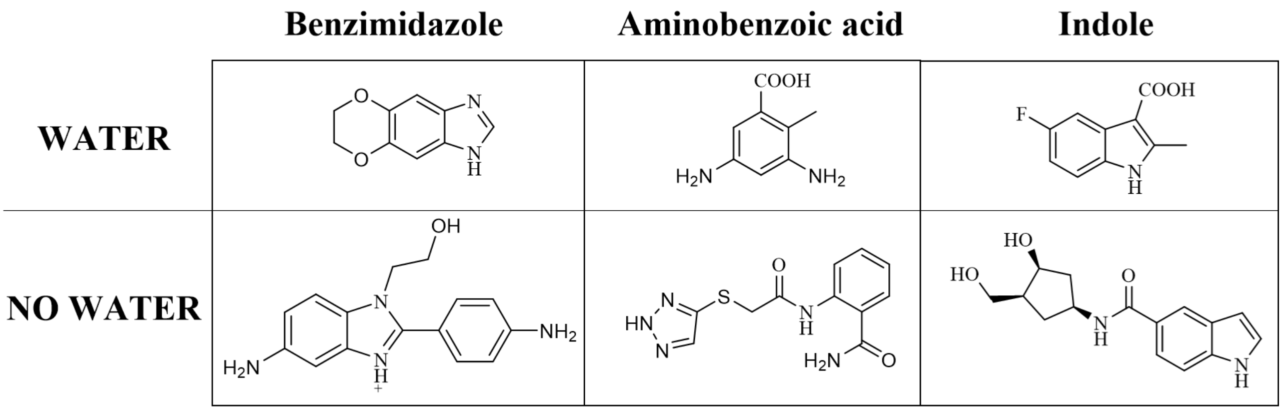


Figure 91: top ranked fragments belonging to benzimidazole, aminobenzoic acid and indole families.

Considering the top 1% of the fragments resulted from the VS (about 20), two chemotypes were identified as the most representative: the indole and the isatin (best ranked fragment is shown in Figure 92).

Isatin was only found in NO WATER grid, but its EF is very high: 100. Therefore, this chemotype, such as indole, can be used for the design of new compounds for binding to the HYD site.

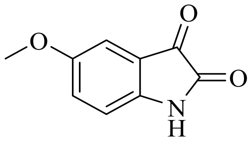


Figure 92: Best ranked fragment of isatin family found in WATER grid

5.5 Conclusion

The twenty-one fragments selected for the SYNTH site were analysed into the HYD site of chain B to define their selectivity and to compare the data with the one obtained by TSA experiments. From the MD simulations, that followed the docking simulations, we found the H-bond with T151 to be formed by fragments that bind to the HYD domain in TSA experiments.

These data lead to a VS performed into the HYD pocket using the same libraries used into the SYNTH site selecting only the fragments interacting with T151. Results showed that only B1 and B2 survived the filters used, and indole and isatin were the most promising chemotype for the developing of potential ligands into the HYD site.

5.6 Appendix

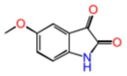
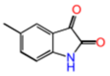
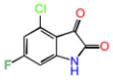
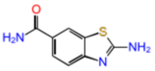
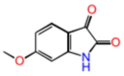
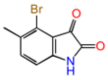
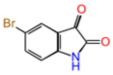
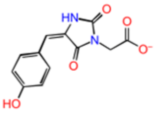
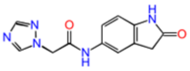
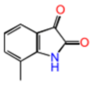
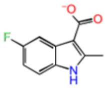
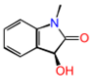
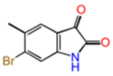
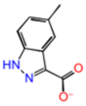
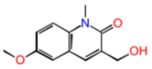
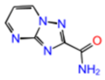
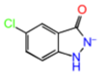
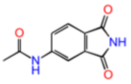
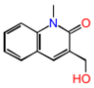
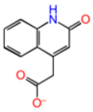
 title: F0896-0193 glide ligand efficiency: -0.621 docking score: -8.07	 title: F0266-0774 glide ligand efficiency: -0.665 docking score: -7.983	 title: F2147-1128 glide ligand efficiency: -0.611 docking score: -7.94
 title: ChemBridge fragment library ligprep.sdf:1560 glide ligand efficiency: -0.607 docking score: -7.894	 title: ligprep Spec Preplated Fragment-Based Library.sdf:3652 glide ligand efficiency: -0.605 docking score: -7.87	 title: ligprep Spec Preplated Fragment-Based Library.sdf:1005 glide ligand efficiency: -0.598 docking score: -7.768
 title: ChemBridge fragment library ligprep.sdf:2808 glide ligand efficiency: -0.643 docking score: -7.72	 title: asinex fragments ligprep.sdf: glide ligand efficiency: -0.4 docking score: -7.605	 title: asinex fragments ligprep.sdf: glide ligand efficiency: -0.399 docking score: -7.589
 title: ligprep Spec Preplated Fragment-Based Library.sdf:4316 glide ligand efficiency: -0.629 docking score: -7.548	 title: ligprep OTAVA Solubility Fragment Library.sdf:244 glide ligand efficiency: -0.539 docking score: -7.543	 title: ligprep Spec Preplated Fragment-Based Library.sdf:3436 glide ligand efficiency: -0.622 docking score: -7.468
 title: ligprep Spec Preplated Fragment-Based Library.sdf:1003 glide ligand efficiency: -0.572 docking score: -7.431	 title: ChemBridge fragment library ligprep.sdf:322 glide ligand efficiency: -0.571 docking score: -7.418	 title: F2185-0126 glide ligand efficiency: -0.463 docking score: -7.413
 title: asinex fragments ligprep.sdf: glide ligand efficiency: -0.617 docking score: -7.409	 title: F1607-0002 glide ligand efficiency: -0.672 docking score: -7.39	 title: F1105-0116 glide ligand efficiency: -0.492 docking score: -7.376
 title: F3308-0232 glide ligand efficiency: -0.525 docking score: -7.351	 title: F2124-0025 glide ligand efficiency: -0.487 docking score: -7.304	

Figure 93: Top 1% of fragments resulted from the VS workflow into the WATER grid

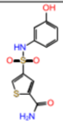
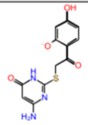
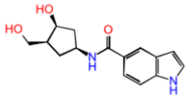
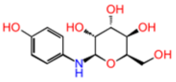
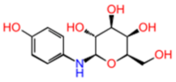
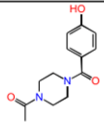
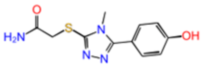
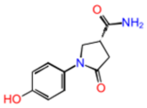
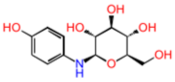
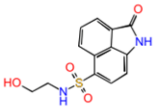
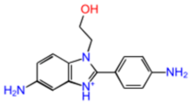
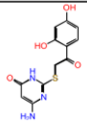
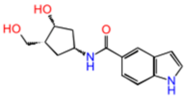
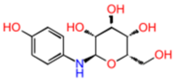
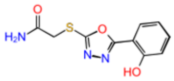
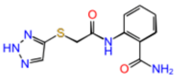
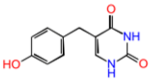
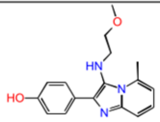
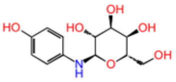
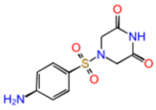
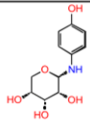
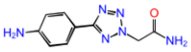
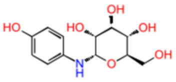
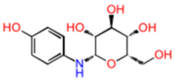
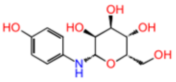
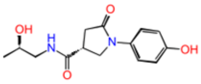
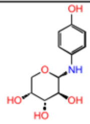
 title: asinex fragments ligprep.sdf: glide ligand efficiency: -0.476 docking score: -9.041	 title: asinex fragments ligprep.sdf: glide ligand efficiency: -0.451 docking score: -9.019	 title: F6127-0222 glide ligand efficiency: -0.438 docking score: -8.755
 title: asinex fragments ligprep.sdf: glide ligand efficiency: -0.459 docking score: -8.726	 title: asinex fragments ligprep.sdf: glide ligand efficiency: -0.459 docking score: -8.726	 title: F1929-0265 glide ligand efficiency: -0.485 docking score: -8.723
 title: asinex fragments ligprep.sdf: glide ligand efficiency: -0.483 docking score: -8.696	 title: asinex fragments ligprep.sdf: glide ligand efficiency: -0.542 docking score: -8.675	 title: asinex fragments ligprep.sdf: glide ligand efficiency: -0.451 docking score: -8.567
 title: asinex fragments ligprep.sdf: glide ligand efficiency: -0.428 docking score: -8.552	 title: asinex fragments ligprep.sdf: glide ligand efficiency: -0.427 docking score: -8.54	 title: asinex fragments ligprep.sdf: glide ligand efficiency: -0.424 docking score: -8.482
 title: F6127-0222 glide ligand efficiency: -0.423 docking score: -8.453	 title: asinex fragments ligprep.sdf: glide ligand efficiency: -0.445 docking score: -8.45	 title: asinex fragments ligprep.sdf: glide ligand efficiency: -0.492 docking score: -8.371
 title: asinex fragments ligprep.sdf: glide ligand efficiency: -0.44 docking score: -8.363	 title: ChemBridge fragment library ligprep.sdf:3746 glide ligand efficiency: -0.521 docking score: -8.343	 title: asinex fragments ligprep.sdf: glide ligand efficiency: -0.379 docking score: -8.342
 title: asinex fragments ligprep.sdf: glide ligand efficiency: -0.438 docking score: -8.322	 title: asinex fragments ligprep.sdf: glide ligand efficiency: -0.462 docking score: -8.311	 title: asinex fragments ligprep.sdf: glide ligand efficiency: -0.488 docking score: -8.297
 title: asinex fragments ligprep.sdf: glide ligand efficiency: -0.516 docking score: -8.255	 title: asinex fragments ligprep.sdf: glide ligand efficiency: -0.433 docking score: -8.234	 title: asinex fragments ligprep.sdf: glide ligand efficiency: -0.433 docking score: -8.226
 title: asinex fragments ligprep.sdf: glide ligand efficiency: -0.432 docking score: -8.214	 title: asinex fragments ligprep.sdf: glide ligand efficiency: -0.411 docking score: -8.21	 title: asinex fragments ligprep.sdf: glide ligand efficiency: -0.481 docking score: -8.18

Figure 94: Top 1% of fragments resulted from the VS workflow into the WATER grid

CHAPTER 6: CONCLUSIONS

The increasing rate of bacterial cells able to survive antibiotic treatments is of critical interest. In this context, one of the mechanisms involved in their tolerance is the formation of persister, a phenotype that reduces the biochemical processes of bacterial cells to drive them in a dormant state. This state prevents the antibiotic to recognize and kill bacteria. The mechanisms involved in the formation of persisters cells have not been completely clarified, yet. However, our efforts have been focused on the upstream of the stringent response, that is activated by stress conditions such as nutrient starvation or antibiotic treatments. This signalling cascade consists of the hyper activation of RSH enzymes that synthesise the alarmone (p)ppGpp allowing its accumulation in the cell and inducing pleiotropic effects that provoke the metabolic slowdown.

On this basis, this PhD thesis was focused on the study of RSH proteins and on the identification of potential inhibitors able to block (p)ppGpp synthesis.

To reach this goal, the first ever crystalized RSH, the long RSH from S*treptococcus dysgalactiae subsp. equisimilis* (Rel_{Seq}) that contains both HYD and SYNTH domains, was studied.

First, the dynamic behaviour of Rel_{Seq} was investigated evaluating if it can be affected by GDP and Mn^{2+}. Thus, MD simulations of the “SYNTH on” conformation of Rel_{Seq} were performed. Four systems were simulated with or without the ligand and/or the coenzyme.

As a result, the behaviour of the enzyme, particularly for the HYD domain, was not affected by the presence or the absence of the metal, leading us to hypothesize that the role of Mn^{2+} is not connected to protein stability or motion. On the other hand, GDP can influence the flexibility of the SYNTH domain. In fact, in the systems where GDP is present, the residues located in the SYNTH domain, in particular those forming the G-loop, fluctuate less and the general motions of the domain are more limited.

GDP interaction network was also evaluated during the simulations. The guanine ring maintains the X-ray interactions while phosphate groups rotate to interact with residues located in different part of the pocket. This analysis highlighted the crucial role of the residues involved in the interaction with the GDP guanine ring in stabilizing the ligand in the pocket. Furthermore, the π-π stacking between residue Y308 and the guanine ring of GDP was identified as the most stable interaction during MD simulations and, moreover, Y308 is the only residue, among the ones involved in guanine ring interactions, that inhibits the synthetic activity of the enzyme if mutated. Therefore, the interaction with this residue was considered the best criterion to discriminate between potential ligands and no ligands in the VS campaign performed for the SYNTH site.

The VS, as first step of the FBDD, was performed using seven fragment libraries that were filtered excluding the less stable tautomeric and ionization forms of fragments, possible duplicates, PAINS and fragments that did not form the desired π-π stacking with residue Y308. Among the resulting fragments, benzimidazole, aminobenzoic acid and indole were the most represented chemotypes. Therefore, a library expansion was performed to further explore the chemical space of these scaffolds and two more fragments were identified (B4 and I3). Finally, twenty-one fragments were selected for purchasing and were further tested by MD simulations and TSA. As a result, eighteen fragments containing one of the three main chemotypes were selected from VS, library expansion and in house already purchased items and three singletons were added from the VS campaign. The TS data, obtained from three truncated version of Rel$_{Seq}$ containing both domains, only the HYD domain or only the SYNTH domain, showed that twelve fragments bind selectively to the SYNTH domain (A1, A3, A4, A5, A6, A7, I1, I2, I3, I4, BO1 and TP1), three bind selectively to the HYD domain (B2, I5 and I5COOH) and two can bind both domains (B3 and I6). More interesting, aminobenzoic acid is the only chemotype completely selective for the SYNTH domain. From a computational point of view, all the fragments, except for B1 and I2, conserved the desired π-π stacking with residue Y308 for all the MD simulations, and therefore can be considered as potential ligands. Comparing data of MD simulations and TSA, fourteen fragments showed an *in silico* - *in vitro* agreement: B3, A1, A3, A4, A5, A6, A7, I1, I3, I4, I6, BO1 and TP1 resulted active on the SYNTH domain, while B1 did not maintained the interaction during the simulations and do not bind the domain in TSA.

Considering the selectivity of the aminobenzoic acid fragments, new compounds were generated starting from the anthranilic acid (2-aminobenzoic acid) and for this they were called ANT-derivatives. These compounds were docked and tested via both TSA and MD simulations. As expected, the ANT-derivatives bind to the SYNTH domain *in vitro* while two of the capped ANT-derivatives (ANT-32S and ANT-33R) lose the π-π stacking with residue Y308 during MD simulations. Therefore, ANT-32S and ANT-33R are the only two ANT-derivatives that did not show agreement between TS and MD data.

The selectivity of the fragments for the SYNTH site was also evaluated *in silico* by performing docking and MD simulations into the HYD site of the Rel$_{Seq}$ "HYD on" conformation. Comparing TS and MD data, four out of five fragments that bind to the HYD domain in vitro, formed H-bond with residue T151 (B2, I5COOH, I5, I6). Therefore, the interaction with T151 was considered the best criterion to discriminate potential ligand to not ligand in the following VS into the HYD site. As a result, indole and isatin were identified as the scaffolds that can be used for the design of HYD ligands.

In conclusion, this work led us to identify fragments and compounds that can potentially bind to the SYNTH site of Rel$_{Seq}$ and act as inhibitors of this enzyme, paving the way to novel pharmacological perspectives in the field of bacteria persistence. However, an experimental validation of their epitope mapping and binding site is needed and will be performed in the next future by X-ray crystallization and STD-NMR (Saturation-Transfer Difference). Moreover, the identification of scaffolds in principle suitable for the HYD site lays the foundation for the design of new ligands selective for this domain and opens a new scenario for the treatment of infectious diseases.

REFERENCES

1. 2020 antibacterial agents in clinical and preclinical development: an overview and analysis. https://www.who.int/publications/i/item/9789240021303.

2. Pacios, O. *et al.* (p)ppGpp and its role in bacterial persistence: New challenges. *Antimicrob. Agents Chemother.* **64**, (2020).

3. Mulcahy, L. R., Burns, J. L., Lory, S. & Lewis, K. Emergence of Pseudomonas aeruginosa strains producing high levels of persister cells in patients with cystic fibrosis. *J. Bacteriol.* **192**, 6191–6199 (2010).

4. Fauvart, M., de Groote, V. N. & Michiels, J. Role of persister cells in chronic infections: Clinical relevance and perspectives on anti-persister therapies. *Journal of Medical Microbiology* vol. 60 699–709 (2011).

5. Dutta, N. K. *et al. Inhibiting the stringent response blocks Mycobacterium tuberculosis entry into quiescence and reduces persistence*. http://advances.sciencemag.org/ (2019).

6. Hobbs, J. K. & Boraston, A. B. (p)ppGpp and the Stringent Response: An Emerging Threat to Antibiotic Therapy. *ACS Infect. Dis.* **5**, 1505–1517 (2019).

7. Balaban, N. Q. *et al.* Definitions and guidelines for research on antibiotic persistence. *Nat. Rev. Microbiol.* **17**, 441–448 (2019).

8. Bigger, J. W. Treatment of Staphylococcal infections with penicillin by intermittent sterilisation. *Lancet* **244**, 497–500 (1944).

9. Moyed, H. S. & Bertrand, K. P. hipA, a newly recognized gene of Escherichia coli K-12 that affects frequency of persistence after inhibition of murein synthesis. *J. Bacteriol.* **155**, 768–775 (1983).

10. Korch, S. B., Henderson, T. A. & Hill, T. M. Characterization of the hipA7 allele of Escherichia coli and evidence that high persistence is governed by (p)ppGpp synthesis. *Mol. Microbiol.* **50**, 1199–1213 (2003).

11. Balaban, N. Q., Merrin, J., Chait, R., Kowalik, L. & Leibler, S. Bacterial persistence as a phenotypic switch. *Science (80-.).* **305**, 1622–1625 (2004).

12. Smith, E. E. *et al. Genetic adaptation by Pseudomonas aeruginosa to the airways of cystic fibrosis patients*. www.pnas.orgcgidoi10.1073pnas.0602138103 (2006).

13. Fisher, R. A., Gollan, B. & Helaine, S. Persistent bacterial infections and persister cells. *Nature Reviews Microbiology* vol. 15 453–464 (2017).

14. Gefen, O. & Balaban, N. Q. The importance of being persistent: Heterogeneity of bacterial populations under antibiotic stress: Review article. *FEMS Microbiology Reviews* vol. 33 704–717 (2009).

15. Wakamoto, Y. *et al.* Dynamic persistence of antibiotic-stressed mycobacteria. *Science (80-.).* **339**, 91–95 (2013).

16. Maisonneuve, E., Castro-Camargo, M. & Gerdes, K. Erratum: (p)ppGpp Controls Bacterial Persistence by Stochastic Induction of Toxin-Antitoxin Activity (Cell (2013) 154(5) (1140–1150)(S0092867413009586)(10.1016/j.cell.2013.07.048)). *Cell* vol. 172 1135 (2018).

17. Shan, Y. *et al.* ATP-Dependent persister formation in Escherichia coli. *MBio* **8**, (2017).

18. Conlon, B. P. *et al.* Persister formation in Staphylococcus aureus is associated with ATP depletion. *Nat. Microbiol.* **1**, (2016).

19. Hood, R. D., Higgins, S. A., Flamholz, A., Nichols, R. J. & Savage, D. F. The stringent response regulates adaptation to darkness in the cyanobacterium Synechococcus elongatus. *Proc. Natl. Acad. Sci. U. S. A.* **113**, E4867–E4876 (2016).

20. Gallant, J., Palmer, L. & Pao, C. C. Anomalous synthesis of ppGpp in growing cells. *Cell* **11**, 181–185 (1977).

21. Wells, D. H. & Gaynor, E. C. Helicobacter pylori initiates the stringent response upon nutrient and pH downshift. *J. Bacteriol.* **188**, 3726–3729 (2006).

22. Glass, T. L. & Holmes, \w Michael. *Synthesis of Guanosine Tetra-and Pentaphosphates by the Obligately Anaerobic Bacterium Bacteroides thetaiotaomicron in Response to Molecular Oxygen*. http://jb.asm.org/.

23. Irving, S. E. & Corrigan, R. M. Triggering the stringent response: Signals responsible for activating (p)ppGpp synthesis in bacteria. *Microbiology (United Kingdom)* vol. 164 268–276 (2018).

24. Ross, W., Vrentas, C. E., Sanchez-Vazquez, P., Gaal, T. & Gourse, R. L. The Magic Spot: A ppGpp Binding Site on E. coli RNA Polymerase Responsible for Regulation of Transcription Initiation. *Mol. Cell* **50**, 420–429 (2013).

25. Cashel, M. & Gallant, J. Two compounds implicated in the function of the RC gene of escherichia coli. *Nature* **221**, 838–841 (1969).

26. Hauryliuk, V., Atkinson, G. C., Murakami, K. S., Tenson, T. & Gerdes, K. Recent functional insights into the role of (p)ppGpp in bacterial physiology. *Nature Reviews Microbiology* vol. 13 298–309 (2015).

27. Costanzo, A. & Ades, S. E. Growth phase-dependent regulation of the extracytoplasmic stress factor, σE, by guanosine 3′,5′-bispyrophosphate (ppGpp). *J. Bacteriol.* **188**, 4627–4634 (2006).

28. Gaca, A. O., Colomer-Winter, C. & Lemos, J. A. Many means to a common end: The intricacies of (p)ppGpp metabolism and its control of bacterial homeostasis. *Journal of Bacteriology* vol. 197 1146–1156 (2015).

29. Shimada, T., Yoshida, H. & Ishihama, A. Involvement of cyclic AMP receptor protein in regulation of the rmf gene encoding the ribosome modulation factor in escherichia coli. *J. Bacteriol.* **195**, 2212–2219 (2013).

30. Song, S. & Wood, T. K. ppGpp ribosome dimerization model for bacterial persister formation and resuscitation. *Biochem. Biophys. Res. Commun.* **523**, 281–286 (2020).

31. Corrigan, R. M., Bellows, L. E., Wood, A. & Gründling, A. ppGpp negatively impacts ribosome assembly affecting growth and antimicrobial tolerance in Gram-positive bacteria. *Proc. Natl. Acad. Sci.* **113**, E1710–E1719 (2016).

32. Wood, A., Irving, S. E., Bennison, D. J. & Corrigan, R. M. The (p)ppGpp-binding GTPase Era promotes rRNA processing and cold adaptation in Staphylococcus aureus. *PLOS Genet.* **15**, e1008346 (2019).

33. Libby, E. A., Reuveni, S. & Dworkin, J. Multisite phosphorylation drives phenotypic variation in (p)ppGpp synthetase-dependent antibiotic tolerance. *Nat. Commun.* **10**, 1–10 (2019).

34. Tagami, K. *et al.* Expression of a small (p)ppGpp synthetase, YwaC, in the (p)ppGpp ^{0} mutant of *Bacillus subtilis* triggers YvyD-dependent dimerization of ribosome. *Microbiologyopen* **1**, 115–134 (2012).

35. Nishino, T., Gallant, J., Shalit, P., Palmer, L. & Wehr, T. Regulatory nucleotides involved in the rel function of Bacillus subtilis. *J. Bacteriol.* **140**, 671–679 (1979).

36. Oki, T., Yoshimoto, A., Ogasawara, T., Sato, S. & Takamatsu, A. Occurrence of pppApp-synthesizing activity in actinomycetes and isolation of purine nucleotide pyrophosphotransferase. *Arch. Microbiol.* **107**, 183–187 (1976).

37. Petchiappan, A., Naik, S. Y. & Chatterji, D. RelZ-mediated stress response in mycobacterium smegmatis: PGPP synthesis and its regulation. *J. Bacteriol.* **202**, (2020).

38. Yang, N. *et al.* The Ps and Qs of alarmone synthesis in staphylococcus aureus. *PLoS One* **14**, e0213630

(2019).

39. Ruwe, M., Kalinowski, J. & Persicke, M. Identification and Functional Characterization of Small Alarmone Synthetases in Corynebacterium glutamicum. *Front. Microbiol.* **8**, 1601 (2017).

40. Sajish, M., Kalayil, S., Verma, S. K., Nandicoori, V. K. & Prakash, B. The Significance of EXDD and RXKD Motif Conservation in Rel Proteins. *J. Biol. Chem.* **284**, 9115–9123 (2009).

41. Yang, J. *et al.* Systemic characterization of pppGpp, ppGpp and pGpp targets in Bacillus reveals NahA converts (p)ppGpp to pGpp to regulate alarmone composition and signaling. *bioRxiv* 2020.03.23.003749 (2020) doi:10.1101/2020.03.23.003749.

42. Zhang, Y., Zborníková, E., Rejman, D. & Gerdes, K. Novel (p)ppGpp binding and metabolizing proteins of Escherichia coli. *MBio* **9**, (2018).

43. Gao, A., Vasilyev, N., Kaushik, A., Duan, W. & Serganov, A. Principles of RNA and nucleotide discrimination by the RNA processing enzyme RppH. *Nucleic Acids Res.* **48**, 3776–3788 (2020).

44. Ooga, T. *et al.* Degradation of ppGpp by nudix pyrophosphatase modulates the transition of growth phase in the bacterium Thermus thermophilus. *J. Biol. Chem.* **284**, 15549–15556 (2009).

45. Gaca, A. O. *et al.* From (p)ppGpp to (pp)pGpp: Characterization of regulatory effects of pGpp synthesized by the small alarmone synthetase of Enterococcus faecalis. *J. Bacteriol.* **197**, 2908–2919 (2015).

46. Atkinson, G. C., Tenson, T. & Hauryliuk, V. The RelA/SpoT Homolog (RSH) superfamily: Distribution and functional evolution of ppgpp synthetases and hydrolases across the tree of life. *PLoS One* **6**, (2011).

47. Manav, M. C. *et al.* Structural basis for (p)ppGpp synthesis by the Staphylococcus aureus small alarmone synthetase RelP. *J. Biol. Chem.* **293**, 3254–3264 (2018).

48. Irving, S. E., Choudhury, N. R. & Corrigan, R. M. The stringent response and physiological roles of (pp)pGpp in bacteria. *Nature Reviews Microbiology* 1–16 (2020) doi:10.1038/s41579-020-00470-y.

49. Ronneau, S. & Hallez, R. Make and break the alarmone: Regulation of (p)ppGpp synthetase/hydrolase enzymes in bacteria. *FEMS Microbiology Reviews* vol. 43 389–400 (2019).

50. Jimmy, S. *et al.* A widespread toxin−antitoxin system exploiting growth control via alarmone signaling. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 10500–10510 (2020).

51. Aravind, L. & Koonin, E. V. The HD domain defines a new superfamily of metal-dependent phosphohydrolases. *Trends Biochem. Sci.* **23**, 469–472 (1998).

52. Tamman, H. *et al.* A nucleotide-switch mechanism mediates opposing catalytic activities of Rel enzymes. *Nat. Chem. Biol.* **16**, 834–840 (2020).

53. Takada, H. *et al.* Ribosome association primes the stringent factor Rel for recruitment of 1 deacylated tRNA to ribosomal A-site 2 3. doi:10.1101/2020.01.17.910273.

54. Arenz, S. *et al.* The stringent factor RelA adopts an open conformation on the ribosome to stimulate ppGpp synthesis. *Nucleic Acids Res.* **44**, 6471–6481 (2016).

55. Brown, A., Fernández, I. S., Gordiyenko, Y. & Ramakrishnan, V. Ribosome-dependent activation of stringent control. *Nature* **534**, 277–280 (2016).

56. Loveland, A. B. *et al.* Ribosome.RelA structures reveal the mechanism of stringent response activation. *Elife* **5**, (2016).

57. Gratani, F. L. *et al.* Regulation of the opposing (p)ppGpp synthetase and hydrolase activities in a bifunctional RelA/SpoT homologue from Staphylococcus aureus. *PLOS Genet.* **14**, e1007514 (2018).

58. Battesti, A. & Bouveret, E. Acyl carrier protein/SpoT interaction, the switch linking SpoT-dependent

stress response to fatty acid metabolism. *Mol. Microbiol.* **62**, 1048–1063 (2006).

59. Germain, E. *et al.* YtfK activates the stringent response by triggering the alarmone synthetase SpoT in Escherichia coli. *Nat. Commun.* **10**, 1–12 (2019).

60. Jenvert, R. M. K. & Schiavone, L. H. Characterization of the tRNA and ribosome-dependent pppGpp-synthesis by recombinant stringent factor from Escherichia coli. *FEBS J.* **272**, 685–695 (2005).

61. Wendrich, T. M., Blaha, G., Wilson, D. N., Marahiel, M. A. & Nierhaus, K. H. Dissection of the mechanism for the stringent factor RelA. *Mol. Cell* **10**, 779–788 (2002).

62. Hogg, T., Mechold, U., Malke, H., Cashel, M. & Hilgenfeld, R. Conformational antagonism between opposing active sites in a bifunctional RelA/SpoT homolog modulates (p)ppGpp metabolism during the stringent response [corrected]. *Cell* **117**, 57–68 (2004).

63. Mechold, U., Murphy, H., Brown, L. & Cashel, M. Intramolecular regulation of the opposing (p)ppGpp catalytic activities of RelSeq, the Rel/Spo enzyme from streptococcus equisimilis. *J. Bacteriol.* **184**, 2878–2888 (2002).

64. Singal, B. *et al.* Crystallographic and solution structure of the N-terminal domain of the Rel protein from *Mycobacterium tuberculosis*. *FEBS Lett.* **591**, 2323–2337 (2017).

65. Pausch, P. *et al.* Structural Basis for Regulation of the Opposing (p)ppGpp Synthetase and Hydrolase within the Stringent Response Orchestrator Rel. *Cell Rep.* **32**, 108157 (2020).

66. Civera, M. & Sattin, S. Homology Model of a Catalytically Competent Bifunctional Rel Protein. *Front. Mol. Biosci.* **0**, 3 (2021).

67. Zimmerman, M. D., Proudfoot, M., Yakunin, A. & Minor, W. Structural insight into the mechanism of substrate specificity and catalytic activity of an HD domain phosphohydrolase: the 5′-deoxyribonucleotidase YfbR from Escherichia coli. *J. Mol. Biol.* **378**, 215 (2008).

68. Abranches, J. *et al.* The Molecular Alarmone (p)ppGpp Mediates Stress Responses, Vancomycin Tolerance, and Virulence in Enterococcus faecalis. *J. Bacteriol.* **191**, 2248–2256 (2009).

69. Nanamiya, H. *et al.* Identification and functional analysis of novel (p)ppGpp synthetase genes in Bacillus subtilis. *Mol. Microbiol.* **67**, 291–304 (2008).

70. Geiger, T., Kästle, B., Gratani, F. L., Goerke, C. & Wolz, C. Two small (p)ppGpp synthases in staphylococcus aureus mediate tolerance against cell envelope stress conditions. *J. Bacteriol.* **196**, 894–902 (2014).

71. Pando, J. M. *et al.* Ethanol-induced stress response of Staphylococcus aureus. *Can. J. Microbiol.* **63**, 745–757 (2017).

72. Steinchen, W. *et al.* Structural and mechanistic divergence of the small (p)ppGpp synthetases RelP and RelQ. *Sci. Rep.* **8**, (2018).

73. Das, B., Pal, R. R., Bag, S. & Bhadra, R. K. Stringent response in Vibrio cholerae: Genetic analysis of spoT gene function and identification of a novel (p)ppGpp synthetase gene. *Mol. Microbiol.* **72**, 380–398 (2009).

74. Krishnan, S., Petchiappan, A., Singh, A., Bhatt, A. & Chatterji, D. R-loop induced stress response by second (p)ppGpp synthetase in Mycobacterium smegmatis: functional and domain interdependence. *Mol. Microbiol.* **102**, 168–182 (2016).

75. Murdeshwar, M. S. & Chatterji, D. MS_RHII-RSD, a dual-function RNase HII-(p)ppGpp synthetase from Mycobacterium smegmatis. *J. Bacteriol.* **194**, 4003–4014 (2012).

76. García-Muse, T. & Aguilera, A. R Loops: From Physiological to Pathological Roles. *Cell* **179**, 604–618 (2019).

77. Ruwe, M., Rückert, C., Kalinowski, J. & Persicke, M. Functional Characterization of a Small Alarmone

Hydrolase in Corynebacterium glutamicum. *Front. Microbiol.* **9**, 916 (2018).

78. Sun, D. *et al.* A metazoan ortholog of SpoT hydrolyses ppGpp and functions in starvation responses. *Nat. Struct. Mol. Biol.* **17**, 1188–1194 (2010).

79. Ding, C. K. C. *et al.* MESH1 is a cytosolic NADPH phosphatase that regulates ferroptosis. *Nat. Metab.* **2**, 270–277 (2020).

80. Wexselblatt, E. *et al.* ppGpp analogues inhibit synthetase activity of Rel proteins from Gram-negative and Gram-positive bacteria. *Bioorg. Med. Chem.* **18**, 4485–4497 (2010).

81. Wexselblatt, E. *et al.* Relacin, a Novel Antibacterial Agent Targeting the Stringent Response. *PLoS Pathog.* **8**, e1002925 (2012).

82. Wexselblatt, E., Kaspy, I., Glaser, G., Katzhendler, J. & Yavin, E. Design, synthesis and structure–activity relationship of novel Relacin analogs as inhibitors of Rel proteins. *Eur. J. Med. Chem.* **70**, 497–504 (2013).

83. Andresen, L. *et al.* Auxotrophy-based High Throughput Screening assay for the identification of Bacillus subtilis stringent response inhibitors. *Sci. Rep.* **6**, 1–8 (2016).

84. Beljantseva, J. *et al.* Molecular mutagenesis of ppGpp: Turning a RelA activator into an inhibitor. *Sci. Rep.* **7**, 1–10 (2017).

85. Léger, L., Byrne, D., Guiraud, P., Germain, E. & Maisonneuve, E. Nird curtails the stringent response by inhibiting rela activity in escherichia coli. *Elife* **10**, (2021).

86. M, X. & MA, L. Induced fit docking, and the use of QM/MM methods in docking. *Drug Discov. Today. Technol.* **10**, e411-8 (2013).

87. Agarwal, S. & Mehrotra, R. An overview of Molecular Docking. *JSM Chem* **4**, 1024 (2016).

88. Korb, O., Stützle, T. & Exner, T. E. Empirical Scoring Functions for Advanced Protein−Ligand Docking with PLANTS. *J. Chem. Inf. Model.* **49**, 84–96 (2009).

89. Guedes, I. A., de Magalhães, C. S. & Dardenne, L. E. Receptor–ligand molecular docking. *Biophys. Rev. 2013 61* **6**, 75–87 (2013).

90. Meng, X.-Y., Zhang, H.-X., Mezei, M. & Cui, M. Molecular docking: a powerful approach for structure-based drug discovery. *Curr. Comput. Aided. Drug Des.* **7**, 146–57 (2011).

91. Ferreira, L. G., Santos, R. N. Dos, Oliva, G. & Andricopulo, A. D. Molecular Docking and Structure-Based Drug Design Strategies. *Mol. 2015, Vol. 20, Pages 13384-13421* **20**, 13384–13421 (2015).

92. Audie, J. & Swanson, J. Recent work in the development and application of protein–peptide docking. *http://dx.doi.org/10.4155/fmc.12.99* **4**, 1619–1644 (2012).

93. Friesner, R. A. *et al.* Extra Precision Glide: Docking and Scoring Incorporating a Model of Hydrophobic Enclosure for Protein-Ligand Complexes. (2006) doi:10.1021/jm051256o.

94. M, T., AS, M. & AM, B. Information-driven modeling of protein-peptide complexes. *Methods Mol. Biol.* **1268**, 221–239 (2015).

95. S, V., DR, H. & D, K. Sampling and scoring: a marriage made in heaven. *Proteins* **81**, 1874–1884 (2013).

96. Changeux, J.-P. & Edelstein, S. Conformational selection or induced fit? 50 years of debate resolved. *F1000 Biol. Rep.* **3**, 19 (2011).

97. Teodoro, M. & Kavraki, L. Conformational Flexibility Models for the Receptor in Structure Based Drug Design. *Curr. Pharm. Des.* **9**, 1635–1648 (2005).

98. Yuriev, E., Holien, J. & Ramsland, P. A. Improvements, trends, and new ideas in molecular docking: 2012–2013 in review. *J. Mol. Recognit.* **28**, 581–604 (2015).

99. D, R. *et al.* Kemp elimination catalysts by computational enzyme design. *Nature* **453**, 190–195 (2008).

100. Cai, J. *et al.* Peptide deformylase is a potential target for anti-Helicobacter pylori drugs: Reverse docking, enzymatic assay, and X-ray crystallography validation. *Protein Sci.* **15**, 2071 (2006).

101. Kitchen, D. B., Decornez, H., Furr, J. R. & Bajorath, J. Docking and scoring in virtual screening for drug discovery: methods and applications. *Nat. Rev. Drug Discov. 2004 311* **3**, 935–949 (2004).

102. Thomas A. Halgren, *,† *et al.* Glide: A New Approach for Rapid, Accurate Docking and Scoring. 2. Enrichment Factors in Database Screening. *J. Med. Chem.* **47**, 1750–1759 (2004).

103. Friesner, R. A. *et al.* Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy. *J. Med. Chem.* **47**, 1739–1749 (2004).

104. Eldridge, M. D., Murray, C. W., Auton, T. R., Paolini, G. V & Mee, R. P. *Empirical scoring functions: I. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes*. *Journal of Computer-Aided Molecular Design* vol. 11 (1997).

105. JA, M., BR, G. & M, K. Dynamics of folded proteins. *Nature* **267**, 585–590 (1977).

106. Wieczorek, G. & Niedzialek, D. Molecular Dynamics. *eLS* 1–18 (2020) doi:10.1002/9780470015902.A0003048.PUB3.

107. and, S. A. A. & McCammon*, J. A. Molecular Dynamics: Survey of Methods for Simulating the Activity of Proteins. *Chem. Rev.* **106**, 1589–1615 (2006).

108. Leach, A. R. *Molecular Modelling PRINCIPLE AND APLICATIONS*. (Prentice Hall).

109. V, H. *et al.* Comparison of multiple Amber force fields and development of improved protein backbone parameters. *Proteins* **65**, 712–725 (2006).

110. William L. Jorgensen, *, David S. Maxwell, and & Tirado-Rives, J. Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids. *J. Am. Chem. Soc.* **118**, 11225–11236 (1996).

111. A. D. MacKerell, J. *et al.* All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins †. *J. Phys. Chem. B* **102**, 3586–3616 (1998).

112. C, O., A, V., AE, M. & WF, van G. A biomolecular force field based on the free enthalpy of hydration and solvation: the GROMOS force-field parameter sets 53A5 and 53A6. *J. Comput. Chem.* **25**, 1656–1676 (2004).

113. Banks, J. L. *et al.* Integrated Modeling Program, Applied Chemical Theory (IMPACT). *Journal of Computational Chemistry* vol. 26 1752–1780 (2005).

114. Roos, K. *et al.* OPLS3e: Extending Force Field Coverage for Drug-Like Small Molecules. *J. Chem. Theory Comput.* **15**, 1863–1874 (2019).

115. Lu, C. *et al.* OPLS4: Improving Force Field Accuracy on Challenging Regimes of Chemical Space. *J. Chem. Theory Comput.* **17**, 4291–4300 (2021).

116. Lindorff-Larsen, K. *et al.* Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins* **78**, 1950 (2010).

117. Maier, J. A. *et al.* ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *J. Chem. Theory Comput.* **11**, 3696–3713 (2015).

118. Tian, C. *et al.* ff19SB: Amino-Acid-Specific Protein Backbone Parameters Trained against Quantum Mechanics Energy Surfaces in Solution. *J. Chem. Theory Comput.* **16**, 528–552 (2019).

119. Lindorff-Larsen, K. *et al.* Systematic Validation of Protein Force Fields against Experimental Data. *PLoS One* **7**, e32131 (2012).

120. Martín-García, F., Papaleo, E., Gomez-Puertas, P., Boomsma, W. & Lindorff-Larsen, K. Comparing

Molecular Dynamics Force Fields in the Essential Subspace. *PLoS One* **10**, e0121114 (2015).

121. Maffucci, I. & Contini, A. An Updated Test of AMBER Force Fields and Implicit Solvent Models in Predicting the Secondary Structure of Helical, β-Hairpin, and Intrinsically Disordered Peptides. *J. Chem. Theory Comput.* **12**, 714–727 (2016).

122. Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94 - Halgren - 1996 - Journal of Computational Chemistry - Wiley Online Library. https://onlinelibrary.wiley.com/doi/10.1002/%28SICI%291096-987X%28199604%2917%3A5/6%3C490%3A%3AAID-JCC1%3E3.0.CO%3B2-P.

123. Vanommeslaeghe, K. *et al.* CHARMM General Force Field (CGenFF): A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields. *J. Comput. Chem.* **31**, 671 (2010).

124. Wang, J., Wolf, R. M., Caldwell, J. W., Kollman, P. A. & Case, D. A. Development and testing of a general Amber force field. *J. Comput. Chem.* **25**, 1157–1174 (2004).

125. Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W. & Klein, M. L. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **79**, 926–935 (1983).

126. Mahoney, M. W. & Jorgensen, W. L. A five-site model for liquid water and the reproduction of the density anomaly by rigid, nonpolarizable potential functions. *J. Chem. Phys.* **112**, 8910 (2000).

127. Berendsen, H. J. C., Postma, J. P. M., Gunsteren, W. F. van & Hermans, J. Interaction Models for Water in Relation to Protein Hydration. 331–342 (1981) doi:10.1007/978-94-015-7658-1_21.

128. Izadi, S., Anandakrishnan, R. & Onufriev, A. V. Building water models: A different approach. *J. Phys. Chem. Lett.* **5**, 3863–3871 (2014).

129. Grest, G. S. & Kremer', K. *Molecular dynamics simulation for polymers in the presence of a heat bath*. vol. 33 (1986).

130. Berendsen, H. J. C., Postma, J. P. M., Van Gunsteren, W. F., Dinola, A. & Haak, J. R. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* **81**, 3684–3690 (1984).

131. Nosé, S. A unified formulation of the constant temperature molecular dynamics methods. *J. Chem. Phys.* **81**, 511 (1998).

132. Parrinello, M. & Rahman, A. Polymorphic transitions in single crystals: A new molecular dynamics method. *J. Appl. Phys.* **52**, 7182 (1998).

133. Bussi, G., Donadio, D. & Parrinello, M. Canonical sampling through velocity rescaling. *J. Chem. Phys.* **126**, 014101 (2007).

134. Baell, J. B. & Holloway, G. A. New substructure filters for removal of pan assay interference compounds (PAINS) from screening libraries and for their exclusion in bioassays. *J. Med. Chem.* **53**, 2719–2740 (2010).

135. Sastry, M., Lowrie, J. F., Dixon, S. L. & Sherman, W. Large-Scale Systematic Analysis of 2D Fingerprint Methods and Parameters to Improve Virtual Screening Enrichments. (2010) doi:10.1021/ci100062n.

136. Duan, J., Dixon, S. L., Lowrie, J. F. & Sherman, W. Analysis and comparison of 2D fingerprints: Insights into database screening performance using eight fingerprint methods. *J. Mol. Graph. Model.* **29**, 157–170 (2010).

137. D.A. Case, I.Y. Ben-Shalom, S.R. Brozell, D.S. Cerutti, T.E. Cheatham, III, V.W.D. Cruzeiro, T. A. D. *et al.* AMBER 2018, University of California, San Francisco.

138. Berman, H. M. *et al.* The Protein Data Bank. *Nucleic Acids Res.* **28**, 235–42 (2000).

139. Schrödinger Release 2018-3: LigPrep, Schrödinger, LLC, New York, NY, 2018. (2018).

140. Shelley, J. C. *et al.* Epik: a software program for pK(a) prediction and protonation state generation for drug-like molecules. *J. Comput. Aided. Mol. Des.* **21**, 681–91 (2007).

141. Jacobson, M. P. *et al.* A hierarchical approach to all-atom protein loop prediction. *Proteins Struct. Funct. Bioinforma.* **55**, 351–367 (2004).

142. Zhu, K., Pincus, D. L., Zhao, S. & Friesner, R. A. Long Loop Prediction Using the Protein Local Optimization Program. (2006) doi:10.1002/prot.21040.

143. Sellers, B. D., Zhu, K., Zhao, S., Friesner, R. A. & Jacobson, M. P. Toward better refinement of comparative models: Predicting loops in inexact environments. *Proteins* **72**, 959 (2008).

144. Li, J. *et al.* The VSGB 2.0 Model: A Next Generation Energy Model for High Resolution Protein Structure Modeling. doi:10.1002/prot.23106.

145. Stote, R. H. & Karplus, M. Zinc binding in proteins and solution: A simple but accurate nonbonded representation. *Proteins Struct. Funct. Bioinforma.* **23**, 12–31 (1995).

146. Roe, D. R. & Cheatham, T. E. PTRAJ and CPPTRAJ: Software for processing and analysis of molecular dynamics trajectory data. *J. Chem. Theory Comput.* **9**, 3084–3095 (2013).

147. Mckinney, W. Data Structures for Statistical Computing in Python. (2010).

148. Virtanen, P. *et al.* SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods 2020 173* **17**, 261–272 (2020).

149. Hunter, J. D. Matplotlib: A 2D graphics environment. *Comput. Sci. Eng.* **9**, 90–95 (2007).

150. Lobanov, M. Y., Bogatyreva, N. S. & Galzitskaya, O. V. Radius of gyration as an indicator of protein structure compactness. *Mol. Biol. 2008 424* **42**, 623–628 (2008).

151. Xu, N., Finkelman, R. B., Dai, S., Xu, C. & Peng, M. Average Linkage Hierarchical Clustering Algorithm for Determining the Relationships between Elements in Coal. (2021) doi:10.1021/acsomega.0c05758.

152. Weiser, J., Shenkin, P. S. & Clark Still, W. Approximate Atomic Surfaces from Linear Combinations of Pairwise ( ) Overlaps LCPO¨JORG. *J. Comput. Chem.* **20**, 217230 (1999).

153. Tipping, M. E. & Bishop, C. M. Mixtures of probabilistic principal component analysers. *Neural Comput.* **11**, 443–482.

154. Amadei, A., Ceruso, M. A. & Nola, A. Di. On the Convergence of the Conformational Coordinates Basis Set Obtained by the Essential Dynamics Analysis of Proteins' Molecular Dynamics Simulations. *Proteins* **36**, 419–424 (1999).

155. McGibbon, R. T. *et al.* MDTraj: A Modern Open Library for the Analysis of Molecular Dynamics Trajectories. *Biophys. J.* **109**, 1528–1532 (2015).

156. Baker, E. N. & Hubbard, R. E. Hydrogen bonding in globular proteins. *Prog. Biophys. Mol. Biol.* **44**, 97–179 (1984).

157. Wang, J. *et al.* Mapping allosteric communications within individual proteins. *Nat. Commun. 2020 111* **11**, 1–13 (2020).

158. Daylight Theory Manual.

159. Schrödinger Release 2018-3: MacroModel, Schrödinger, LLC, New York, NY, 2018.

160. Ponder, J. W. & Richards, F. M. An efficient newton-like method for molecular mechanics energy minimization of large molecules. *J. Comput. Chem.* **8**, 1016–1024 (1987).

161. Kim, S. *et al.* PubChem in 2021: new data content and improved web interfaces. *Nucleic Acids Res.* **49**, D1388–D1395 (2020).

162. Lipkus, A. H. A proof of the triangle inequality for the Tanimoto distance. *J. Math. Chem.* **26**, 263–265 (1999).

163. Dart, M. L. *et al.* Homogeneous Assay for Target Engagement Utilizing Bioluminescent Thermal Shift. *ACS Med. Chem. Lett.* **9**, 546–551 (2018).

164. Piovesan, D., Minervini, G. & Tosatto, S. C. E. The RING 2.0 web server for high quality residue interaction networks. *Nucleic Acids Res.* **44**, 367–374 (2016).

165. Halgren, T. A. Identifying and Characterizing Binding Sites and Assessing Druggability. *J. Chem. Inf. Model.* **49**, 377–389 (2009).