

ELUCIDATING SELF-ASSEMBLY
AND ANTIMICROBIAL STRATEGIES
OF SYNTHETIC PEPTIDES:
AN IN SILICO INVESTIGATION

Irene Marzuoli

RANDALL CENTRE FOR CELL AND MOLECULAR BIOPHYSICS
KING'S COLLEGE LONDON



THIS DISSERTATION IS SUBMITTED FOR
THE DEGREE OF DOCTOR OF PHILOSOPHY

SEPTEMBER 2019

Declaration

This dissertation describes work I have carried out between October 2016 and September 2019 at the Randall Centre for King's College London, under the supervision of Professor Franca Fraternali (first supervisor) and Dr. Chris D. Lorenz (second supervisor).

This dissertation contains material appearing in the following articles:

- ...

In addition to the above, I have contributed to the following publications during the course of my PhD:

- ...

This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and acknowledgements. It has not been submitted in whole or in part for any degree or diploma at this or any other university.

Irene Marzuoli
September 2019

Acknowledgements

...

Summary

Elucidating self-assembly and antimicrobial strategies of
synthetic peptides an *in silico* investigation

Irene Marzuoli
King's College London

... ...

Contents

1	Introduction	11
1.1	Antimicrobial resistance	13
1.2	Alternative antibiotic strategies: antimicrobial peptides	19
1.3	Gene therapy	34
1.4	Delivery of therapeutic material	36
1.5	Closing the circle: an antimicrobial drug delivery vehicle	41
1.6	A computational approach to understand capzip	48
2	Methods	50
2.1	Algorithms for Molecular Dynamics	51
2.2	The force field problem	54
2.3	Beyond the classical framework	64
2.4	MD and experiments: ensembles versus averages	67
2.5	MD simulations: successes	69
3	Capzip simulations	82
3.1	Modelling the assembly	83
3.2	Simulations details	87
3.3	Analysis	94
3.4	Results: capsule in solution	98
3.5	Results: peptide-membrane interactions	113
3.6	Rest	120
3.7	Additional material	123
4	Lipid parametrisation	125
	Appendices	126
	A.1 On the derivation of the GP predictive distribution	126

List of Figures

1.1	Graphical abstract of introduction	12
1.2	Mechanisms of antimicrobial resistance to small drugs	15
1.3	Antimicrobial peptides	20
1.4	Principles of gene therapy	35
1.5	Materials for drug delivery vehicles	37
1.6	Cazip molecule	42
1.7	Experimental results on capzip	46
2.1	MD force field classifications	58
2.2	SIRAH force field amino acid and DNA description	61
2.3	MARTINI force field mapping for amino acids, lipids and water .	63
2.4	Disordered pore model	74
2.5	Polyarginine translocation	75
2.6	Peptide amphiphiles assembly through MARTINI simulations .	80
3.1	Hydrogen bonds between RRWTWE β -sheet	83
3.2	Building blocks of capzip assembly	84
3.3	Cohesion measures on the pentagonal subunit	86
3.4	Snapshot of simulations of capzip on model membranes	87
3.5	Atomistic run of buckyball in solution: final configuration	98
3.6	Structural measures on buckyball in solution	99
3.7	Correlation of motion between molecules of the buckyball	101
3.8	Branch pairing during simulations of the buckyball	101
3.9	Contacts between molecules during simulations of the buckyball .	103
3.10	Hydrogen bonds in the buckyball molecule	104
3.11	SASA per residue of a buckyball in solution	105
3.12	Non-bonded protein energy contribution to capsule structures .	108
3.13	Comparison of monolayer and bilayer structural properties	111

3.14 Snapshot of SIRAH monolayer simulation	112
3.15 SASA per residue of monolayer and bilater	112
3.16 ApL, D and S_{CD} for membrane simulations	115
3.17 Hexagonal order parameter of lipids in protein-lipid simulation	116
3.18 Snapshot from relevant membrane-peptide simulations	122

Chapter 1

Introduction

“**P**HILosophical introduction” to be finished/modified when the work is finished...

(...thus) This introduction is meant to give an overview of the many different challenges the fields of medicine and bioengineering have faced in recent years. These challenges have promoted the research on self-assembling antimicrobial peptides, despite they were not a primary source of interest in these fields, as other materials and concepts were deemed more suitable to solve the tasks coming along the way. It is therefore important to clarify the landscape of such other solutions and approaches to understand and value why a change in the research focus has come to age. Figure 1.1 provides a work flow of this introductory chapter to help the reader in identifying the sections of interest.

Motivations of the work: a graphical abstract

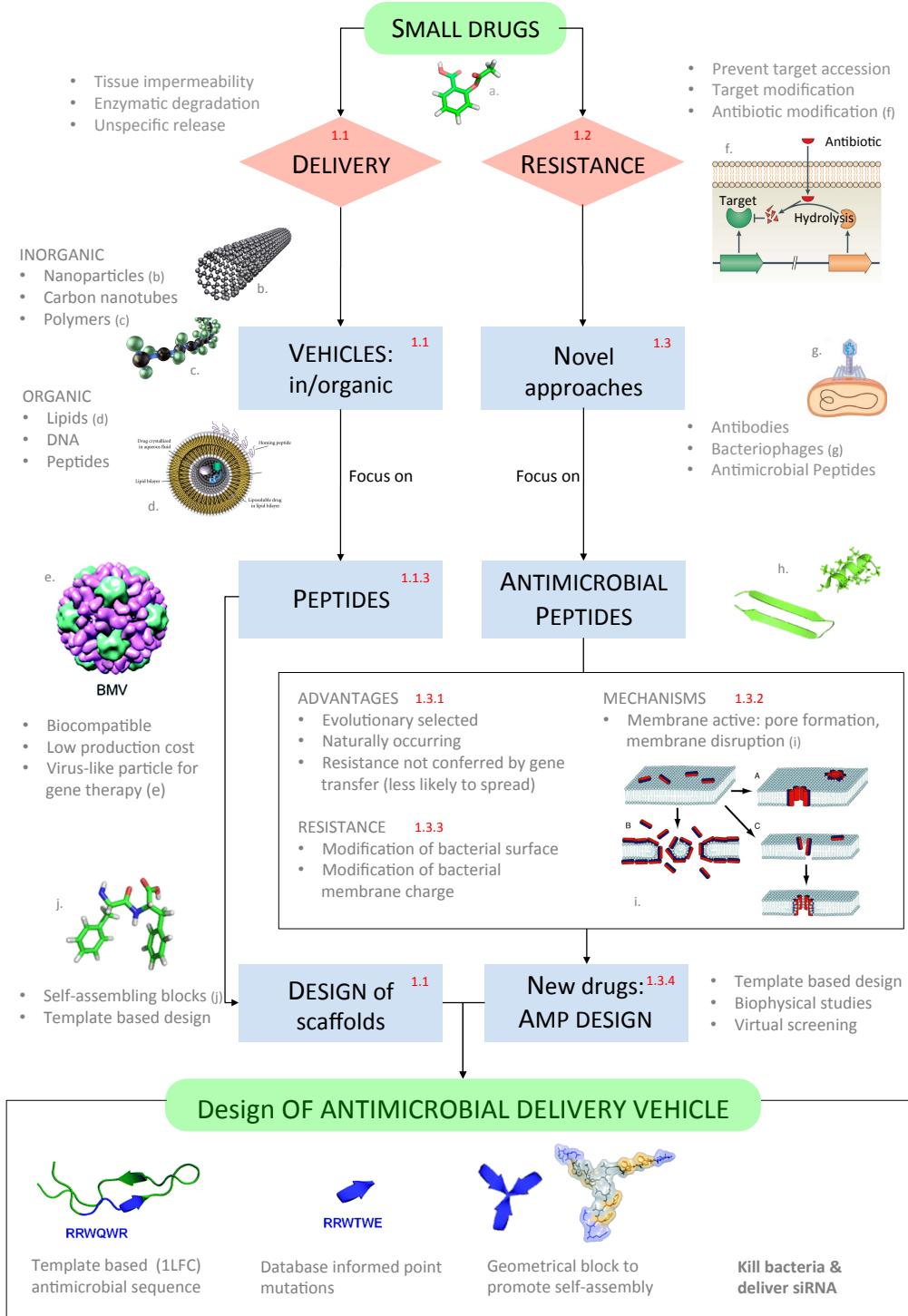


Figure 1.1: Figures a. (acetylsalicylic acid) and j. (diphenyl-alanine) in bond representation. Remaining figures adapted from: b. [-]; c. [?]; d. [?]; e. [?], f. [?]; g. [?]; h. [?]; i. [?]; k. [?]

1.1 Antimicrobial resistance

For most of the last century, the development of new drugs rotated around the paradigm that a drug is a small inorganic compound (of mass up to 900 Da), which intervenes on a specific target of a mammal or bacterial cell. Very often the targets of interest are (intracellular) proteins: out of the 695 small drugs approved by FDA (the American Food and Drug Administration agency) to target human molecules, 667 acts on proteins. Similarly, 189 of the 198 approved to treat pathogens have a protein as their target [?] (with all the caveats coming from the challenges of identifying an unambiguous target, especially when the drug binds to a protein complex or to a number of closely related gene products [?]).

In presenting the aforementioned figures, the data were naturally split among the drugs which act on human molecules, “repairing” some faulty process in the human body, or the ones active against bacteria, which “disrupt” the bacterium life cycle in order to kill or prevent the reproduction of the pathogen, commonly named as antibiotics. It appears evident that the pool of drugs available to the second purpose are in substantially lower number than the ones addressing human molecules. This comes from the nature of the action they perform: molecules targeting human proteins need to be highly specific to avoid interference with other proteins or with healthy cells, and in a sufficient number to address the variety of diseases affecting the human body. Antibiotic must be non-toxic for human cells as well, i.e. their target must not be shared between mammal and bacterial cells, but there is a less stringent requirement on their selectivity against different bacterial species. On the contrary, it is often useful to have a broad-spectrum compound. This cross-species efficacy and at the same time non-toxic property is obtained thanks to the evolutionary relationship among bacterial species, and between bacteria and humans: while the first are closely related, and therefore share homologous proteins with very similar structures, humans have less architectures in common with them, allowing for a resilience against bacteria-targeting drugs. To be precise, the set of bacterial species is very diverse and the cross-species effectiveness of some drugs does not extend to the whole bacterial population. This is actually demonstrated to be a positive feature, given the large amount of beneficial bacteria that live in symbiosis with the human body (especially

in the gut) and that must be preserved for an optimal wellness.

In the framework described above, it is understandable that first-time research on antibiotics was satisfied with the development of a handful of potent, broad-spectrum compounds. Penicillin, the first of them to be synthetically produced, was isolated from a mould in 1928 by Alexander Fleming. It acts inhibiting the formation of peptidoglycan cross-links in the bacterial cell wall, binding to the enzyme responsible for its catalysis, and thus preventing the wall complete formation [?] (for further details on bacterial cell membrane structure see Section 1.2.1). As foreseen from Fleming himself in his Nobel Prize acceptance speech, some species of bacteria quickly became immune to penicillin, and this was achieved in many ways: either by production of penicil-lase, an enzyme that degrades penicillin, or by subtle changes in the structure of the penicillin-binding proteins to prevent penicillin binding, or again by removal of the drug outside of the cell through specially re-purposed efflux pumps [?].

The mechanisms just outlined are not an exceptional characteristic of penicillin, and many drugs lost their effectiveness against some bacteria since their discovery till nowadays, urging the research of new ones on a constant basis. By now, a broad knowledge has been gathered on how bacteria escape the action of a drug: this understanding helps interpreting the pitfalls of existing drugs and identifying the characteristics sought in the developments of new compounds.

1.1.1 Mechanisms of antimicrobial resistance to small drugs

Antimicrobial resistance can manifest through many different mechanisms, which can be grouped in three main classes, in line with the three processes mentioned in the example of the penicillin resistant bacteria.

Prevention of access to target A first class of resistance mechanisms aim at minimising the intracellular concentration of the antibiotic preventing its penetration or maximising its efflux in the eventuality it has entered the cell (Figure 1.2(a)). Not all the molecules can enter the cell permeating the membrane, and this holds particularly for hydrophilic antibiotics tackling Gram-negative bacteria which are intrinsically less permeable than Gram-positive ones because of the additional presence of the outer membrane [?] (for fur-

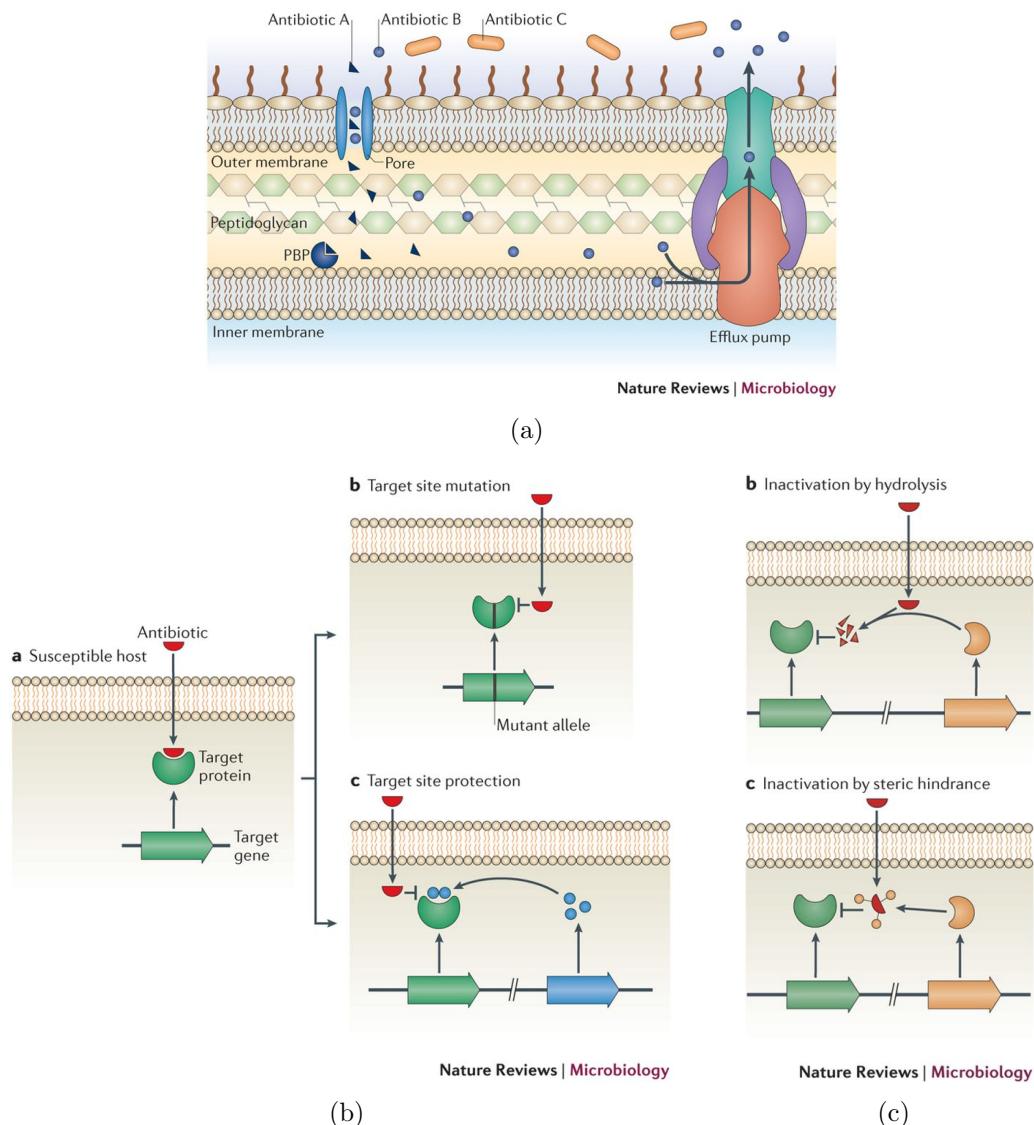


Figure 1.2: Mechanisms of antimicrobial resistance to small drugs. (a) Intrinsic mechanisms of resistance (removal of antibiotic B by efflux pump and inaccessibility of antibiotic C to the PBP target because of membrane impermeability). (b) Target site change via mutation or protection. (c) Direct interactions with antibiotics causing its disruption or structural modification. Reproduced from [?].

ther details on the bacterial membrane structure, see Section 1.2.1). These molecules must then be imported into the cell through outer-membrane porin proteins [? ?]. Resistance arise when porins are either replaced with more selective channels, which prevent the antibiotic penetration, or down regulated

so the internal concentration of the drug does not reach a critical concentration [?]. Porin-coding genes can also accumulate multiple mutations, to acquire the selectivity they lack in their wild type [?].

A complementary strategy to prevent drug influx is to employ bacterial efflux pumps. Some of them are denominated multidrug resistance (MDR) efflux pumps for their effectiveness in the task and are produced by many bacteria [? ?], but, additionally, can be transferred via plasmids to other bacterial species [?]. Indeed, bacteria are able to exchange genetic material with other individuals via small rings of DNA in a process called conjugation [?], so that the advantageous resistant genotypes can spread quickly across species. Over-expression of such efflux pump is observed in multidrug-resistant bacteria, triggered by exposition to the drug, and proceeding via mutation in the relative regulatory network, [?], or simply as a response to environmental signals [?].

Change or modification of the antibiotic target The second class of resistance mechanisms works modifying the antibiotic target: most antibiotics bind to their substrate with high affinity and specificity, thus small modifications in the target structure can disrupt an efficient binding, still allowing the target to maintain its normal function (Figure 1.2(b)).

Mutations of some residues in the binding pocket (upon mutation in the gene coding for it) or post-translational protection of the target via addition of chemical groups are equally wide spread mechanisms. Notable examples of the first include the development of methicillin and linezolid-resistant strains of *S. aureus* [? ?]. Again, it is interesting to notice that several of these mutations are acquired by horizontal gene transfer from other bacterial species, so that resistance development in one specie promotes quickly the insurgence of resistance in other ones. For the second, the most relevant mechanism of chemical group addition is methylation, which for example is very common when the drug target are rRNA subunits [?]).

Direct modification of antibiotics Finally, bacteria can destroy drugs, usually by hydrolysis, or modify them by transfer of a chemical group (Figure 1.2(c)).

The first example historically discovered of drug-degrading enzyme is penicillinase (a β -lactamase) which destroys penicillin [?]. Since then, thousands of similar enzymes have been identified that can modify antibiotics of different classes [? ?]: these enzymes co-evolve with newly developed drugs, to include in their spectrum of action new compounds of composition similar to the ones they were originally effective on [?].

Antibiotics constituted by large molecules with many exposed hydroxyl and amide groups are instead particularly susceptible to modification by addition of chemical groups. Many enzymes are responsible for this, and according to the chemical moiety added they are grouped in acetyltransferases, phosphotransferases and nucleotidyltransferases [?].

All together, the recent progress in understanding the mechanisms of antimicrobial resistance has helped in directing the development of new drugs, in particular the modification and the improvement of existing compounds to escape the resistance developed by bacteria. To ultimately employ the existing drugs at best, it is important to understand also the dynamics of AMR, beyond its molecular mechanisms, to devise the most effective strategies of drug administration.

1.1.2 Course of antimicrobial resistance

In the first stages of the insurgence of antimicrobial resistance (AMR) against a given drug, some strains of bacteria are not damaged by the standard doses of the drug as they came to possess some natural occurring mutations in their genome which promote an escape mechanism invalidating the drug effectiveness [? ?]. Usually, only a small population of bacteria is resistant in the first moments, however the resistant population will replicate faster than the peers of the same species because it is more fit in an environment challenged by the presence of the drug. It is noteworthy that this fitness might not be optimal in a natural drug-free environment - and indeed the wild population has not been selected for that genotype - but under the pressure derived from the treatment, other characteristics result more advantageous. In the short time scale it is usually sufficient to increase the doses of the drug to re-gain efficiency against the target, but resistant species can usually adapt to higher

doses of the same [?]. Moreover, high drug doses are not always applicable due to the severe side effects they are connected to.

As already mentioned, the spread of resistance between bacterial cells and even between species is very effective as bacteria exchange genetic material through conjugation. Therefore, despite AMR is an evolutionary mechanism, the fast pace at which bacteria replicates, their enormous population (in terms of individuals), and the relative easy horizontal gene transfer through conjugation, place the insurgence of resistance well within the human lifespan time scale [?].

It is then clear that this complex problem depends on many variables: the casual appearance of resistant individuals, the transfer of information between them, the relative gain in fitness that the mutation implies, but also the dosage and time line of the drug administration. Many mathematical models have been implemented to understand the issue [? ?], but it is known that some particular strategies of drug administration are worse than other, favouring the proliferation of so called “super” bugs. One example of bad administration strategy is the underdosage of antibiotics: a low drug load is likely to harm but not kill pathogens, and rather promote the fitness of resistant ones. In a sort of “gym” or “vaccination” process for bacteria, an underdosed drug would kill the weakest individuals but strengthen the resistant population, which would now be fitted for the challenges of higher doses [?]. Similarly, the abuse of antibiotics puts an high pressure on the pathogenic populations, which is desirable but at the same time can induce a faster emergence of escape mechanisms [?].

In this context it must be noticed that many drugs are bacteriostatic agent as opposed to bactericidal: i.e. they prevent the bacterium growth rather than kill it, as they are meant to control the bacteria reproduction and slow down the damage while host defence mechanisms eradicate them. Thus if an high dosage of a bactericidal agent may extinguish the bacterial population and eradicate the disease, for bacteriostatic drugs, once they are removed, bacteria start again the reproduction cycle.

It is then clear that the antibiotic landscape is a dynamic entity in which newly discovered drugs enter, while others exit after having been exploited for years, and must be monitored carefully and kept populated. The severity of the AMR issue is such that it has been raised to the status of national emergency

in several countries, including UK. Indeed, abuse or misuse of antibiotic can take many forms and strict regulations on the health, agricultural and food industry sector must be taken to prevent its loss of efficacy, as we are leaving the century in which antibiotics were discovered, to enter a phase in which we count the number of the ones loosing efficacy [?].

1.2 Alternative antibiotic strategies: antimicrobial peptides

In the landscape sketched above, it is evident that the development of novel drugs is of crucial importance. Even more beneficial would be to have at disposal a new paradigm for their design, in order to attack pathogens in a completely novel way, avoiding to target pathways which are known to lead easily to the development of antimicrobial resistance. Several novel materials have been developed for the task, not to rely on small molecules and exploit different mechanisms of action, for example antibodies, bacteriophages or antimicrobial peptides [?].

While the use of pathogen-specific antibodies relies on the mechanisms of the host immune system, bacteriophages therapy employs viruses which infect bacteria and archea rather than eukarya. But are antimicrobial peptides the focus of this thesis: indeed peptides can have an active role against bacteria when their sequence possesses specific characteristics. Such active sequences, capable of damaging and/or killing pathogens, are referred to as antimicrobial peptides. The following subsections will explore their characteristics, modes of action and the response of bacteria against them. It is crucial to understand the knowledge available versus the questions that are still open in order to direct the efforts of future research. This holds in particular when the investigation proceeds by the use of simplified models, as meaningful results can proceed only if such modelling is performed in a sensible and informed fashion.

1.2.1 Membrane active peptides

Antimicrobial peptides (AMPs) are naturally produced by eukarya, either as stand-alone sequences or embedded in larger proteins, as a first, weak, and broad-spectrum defence against bacteria [?]. This pool of molecules has been

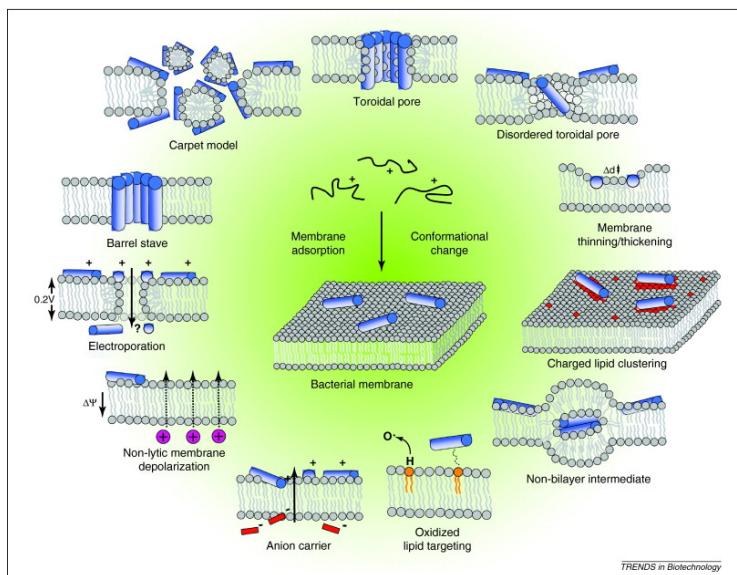


Figure 1.3: Events occurring at the bacterial cytoplasmic membrane following initial antimicrobial peptide (AMP) adsorption. Reproduced from [?].

selected though evolution to be active against pathogens, suggesting that they are weakly prone to provoke resistance reactions in the microbes they attack.

To exploit their potential and engineer AMP-like molecules, a careful characterisation and classification of such peptides must be done. This task has been carried on throughout the past decades but it is complex, so that up to date there are many peptides with ascertained antimicrobial activity for which the mode of action is still not fully understood [?]. However, some general characteristics of these sequences and some of the mechanisms they employ have emerged. Unsurprisingly, AMPs are heterogeneous in shape, targets and mode of action, to tackle the different challenges bacteria pose. Their size can vary between 6 and 59 amino acids [?]: despite being small with respect to the average size of a protein in the human body, these macromolecules are hundreds of times larger than small molecule drugs and as such they penetrate and act on bacteria differently with respect to small compounds.

The most common target of AMPs is the bacterial membrane. Many of them cause disruption of the physical integrity of the microbial membrane while others translocate into the cytoplasm to act on intracellular targets, and the combination of the two is not uncommon either [?] (Figure 1.3). In general, it is widely accepted that membrane interaction is a key factor for the

direct antimicrobial activity of AMPs [?].

As such, we propose a brief overview of the structure of this membrane [?], and of its differences with the one of mammalian cells.

Structure of bacterial membrane The determinant driving the interaction between AMPs and bacterial membranes is the positive charge that many AMPs present, opposed to the negative charge of the latter. It is striking that such simple mechanism, based on the presence of a certain number of negatively charged lipids, holds across many bacterial species despite the great variability found in their membrane composition. Indeed, based on the differences in their cell envelope structure, bacteria are classified into two macro families, Gram-positive and Gram-negative. In Gram-positive bacteria, the cytoplasmic membrane is surrounded by a thick peptidoglycan layer, while for Gram-negative bacteria this membrane (which assumes the name of internal one) is surrounded by a thin peptidoglycan layer as well as an outer membrane [?].

Starting from the inside and proceeding outwards, the cytoplasmic membranes of both Gram-positive and Gram-negative bacteria are rich in phospholipids like phosphatidylethanolamine, which is neutral, and phosphatidylglycerol, cardiolipin, and phosphatidylserine, which have negatively charged head-groups, highly attractive for positively charged AMPs. This is often sufficient to promote the preferential interaction between this membrane and peptides - provided they reach it. Perturbation of this membrane is highly disruptive as many functions are associated to it: as bacteria do not possess organelles, all the membrane related proteins reside and perform their function on the inner membrane.

In the case of Gram-negative bacteria, the inner membrane, together with the outer one, delimits the periplasm space, an aqueous cellular compartment, which allow to sequester harmful substances and to transport nutrients. Proceeding outward, inside the periplasm is situated the peptidoglycan cell wall. Repeated units of this substance, made of a disaccharide cross-linked by penta-peptide side chains (from which the name), constitute the rigid skeleton of Gram-negative bacterial cells. Damage in the peptidoglycan layer results usually in living but not viable cells, therefore it is fundamental for cell life. Grafted to the skeleton through Brauns lipoproteins is the outer membrane.

This membrane presents an asymmetric structure: phospholipids are present in the inner leaflet, while the outer one is composed of glycolipids, mainly lipopolysaccharides (LPS). This complex molecules consist of lipid A, which presents multiple fatty acids, and a polysaccharide, made of an inner core, covalently bond to the lipid, an outer core attached to the inner and finally a repetitive glycan polymer (O-antigen). The O-antigen is the molecule exposed by Gram-negative bacteria to the external environment and thus is the target of antibodies recognition.

Given the complexity of the Gram-negative cell envelope, and especially the presence of the LPS layer, these bacteria are particularly impermeable to hydrophilic molecule, which are usually imported within the cell through porins and similar transmembrane proteins.

For Gram-positive bacteria, the inner membrane is enveloped in a thick peptidoglycan layer. If its thickness in Gram-negative bacteria is reaches few nanometers, in Gram-positive ones it spans from 30 to 100 nm. Its composition is similar to the one described above, with some variations present among different bacteria, for example in the nature of the peptidic linker or in its precise position. This thick layer is threaded by long anionic polymers (the teichoic acids), mainly composed by glycerol phosphate, glucosyl phosphate, or ribitol phosphate repeats. Disseminated in this layer there are several surface proteins with various functions, among which adhesins, which attach to components of the host extracellular matrix.

Gram-positive are generally more permeable because they do not possess a double-membrane structure, nevertheless the peptidoglycan later they are coated with constitute a challenge for drug delivery.

Comparison with mammalian membrane The fact that AMPs tackle negatively charged membranes is crucial for their selectivity, i.e. the fact that they are harmless for the mammalian cells they are produced from [?]. Indeed, mammalian cells have a different membrane composition. They present a single membrane, which is very rich in protein (up to 50% of its volume) and in lipids, with a small percentage of carbohydrates, mainly embedded in glycoproteins, which promote the cell-cell recognition.

The lipidic component is abundant in zwitterionic phospholipids such as phosphatidylethanolamine, phosphatidylcholine, and sphingomyelin, providing

a neutral net charge [? ?]. Strictly speaking, some negatively charged lipids are present in a few mammal cell types, however they are located in the inner leaflet, while the zwitterionic phospholipids are more abundant in the outer leaflet, in an asymmetric composition [?]. This structure promotes weaker interactions between AMPs and mammalian cells with respect to bacterial ones, as the former is driven mainly by hydrophobic interactions, while the latter by electrostatic ones. Furthermore, the mammalian cell membrane has a high content of cholesterol [? ?], a sterol fat, which is proposed to stabilise the membrane regulating its fluidity across the differences of physiological temperatures, and it is also thought to favour a better accommodation of the perturbations caused by AMPs [?].

Another relevant difference between bacterial and mammal cells is that the first ones typically have a higher transmembrane potential - the difference of electrostatic potential between the inside and the outside environment: for bacteria it falls between -130 and -150 mV, while for mammalian cells between -90 and -110 mV [? ? ?]. Given that a potential generates an electric field across the membrane, the higher it is, the higher the electric field pointing from outside to inside the cell. A field in such direction pushes cationic compounds on the outside of the membrane toward the membrane itself. Therefore the stronger bacterial transmembrane potential may promote an enhanced - and thus disruptive - interaction of AMPs with the cell, contributing to the selectivity of AMPs between bacteria versus mammals [?].

1.2.2 Common mechanisms of action of AMPs

Investigating the perturbation and disruption of a bacterial membrane by antimicrobial peptides is a key point of this work, therefore it is important to highlight the mechanisms known so far through which AMPs reach this outcome. As already mentioned, many AMPs have a positive charge which facilitates the binding to the membrane via charge-charge recognition; accordingly, Arginine and Lysine residues are usually abundant in AMPs sequences. However, the disruptive action takes place through the interaction of the AMP with the hydrophobic core of the membrane, therefore their sequence contains also hydrophobic aromatic residues, especially Tryptophan, which favours the anchoring to the lipid core [?]. Overall, AMPs resort often to adopt an amphiphatic structure to segregate the hydrophilic from the hydrophobic amino

acids and thus act at the interface between membrane and solution. It is interesting to notice that some of them fold into the active structure only nearby the membrane, as they can expose their hydrophobic components to face its core, while in solution these ones are preferentially buried inside to be screened from the solvent [?]. Common folds adopted by AMPs are both α -helix or β -sheet rich structures. Amphiphatic α -helices present a charged side which is tailored to face towards the phospholipid head groups and an hydrophobic ones which is favourably buried into the acyl chains core, and a similar arrangement is found for structures rich in β -sheets include β -hairpins.

Membrane disruption Several models have been proposed to describe the exact mechanisms of AMPs penetration after they bind to the cytoplasmatic membrane, and how their combined action leads to membrane permeabilization (Figure 1.3) [? ?].

For example, for a single copy of a amphiphatic helical AMP, the proposed mechanism of action suggests that initially the peptide is attracted with its charged side to the membrane and lies parallel to its plane, with the hydrophobic side unfavourably exposed in solution. Then the helix rearranges to have the two faces in the respective favourable regions. Subsequently, the helix axis starts to form an angle with the membrane plane, and finally inserts deeper into the lipid core, often spanning the full membrane thickness. Similarly, for β -sheet rich structures, it is suggested that they insert within the membrane after a first flat approach. The final insertion arrangement depends on the peptide characteristics and length, the presence of kinks in its structure (in case of helices), and the interactions with other copies of the peptide.

The picture becomes more complex for oligomer-mediated insertion, i.e. when the action is triggered by the combined action of many copies of the peptide. At low peptide to lipid ratio, the favourable configuration is represented by peptides lying parallel to the membrane plane as described previously [?], but an increase in peptide concentration triggers the transition to an inserted state where the main axis of the AMP is perpendicular to the membrane. The organisation of AMPs inside the membrane core can assume different configurations, as described below.

The “barrel-stave” model proposes that AMPs insert perpendicularly into the bilayer. Recruitment of peptides in the same area results in the formation

of a transmembrane pore with a central lumen. The walls of the pore are constituted by the hydrophilic face of the peptides, while their hydrophobic side is interacting with the lipid tails around the pore. This model is adopted for example by the α -helical AMP alamethicin, which forms voltage-dependent ion channels by aggregation of four to six molecules [?].

In the “toroidal” pore model instead, the insertion of peptides forces the phospholipid to bend continuously from one leaflet to the other, resulting in a pore defined by both peptides and phospholipids head groups. The toroidal model differs from the barrel-stave model as the peptides are always associated with the lipid head groups even when they are perpendicularly inserted in the lipid bilayer. Toroidal pores are induced by α -helical magainins, protegrins and melittin [? ? ?], and lead to membrane perturbation which extends further away from the pore than in the barrel-stave case, as lipids must rearrange around them.

As a comparison, alamethicin induced barrel-stave pores have an inner and outer diameters of 1.8 nm and 4.0 nm respectively [?], while magainin-induced toroidal pores are larger and can vary in their size, with an inner diameter of 3.0-5.0 nm and an outer diameter of 7.0-8.4 nm, involving about 4 to 7 magainin monomers and about 90 lipid molecules [?].

Finally, in the “carpet” model, the accumulation of AMPs on the surface of the membrane, laying parallel to it, causes tension in the bilayer and the membrane is then disrupted by peptides in a detergent-like manner, leading to the formation of micelles. The critical threshold concentration triggers a cascade effect, in which formation of the first disruption allows the penetration of AMPs in the inner side of the bilayer. The cooperation between peptides on both sides of the lipid membrane enhances the AMP-induced curvature on the membrane causing accelerated disruption. The “carpet” model mechanism is observed for peptides presenting an α -helical structure (like melittin [?]) or several helices connected by short loops (ovispirin [?]).

The prevalence of examples with an helical structure for the above models derives from the fact that the understanding of how helical AMPs function is often easier than the one of β -sheet rich structures. Indeed, helices have a well defined fold (at least nearby the membrane environment), a compact structure, and often a clear segregation of complementary patches that can attract other copies of the peptide and thus promote the self-assembly process

necessary for the pore formation.

On the contrary, many β -sheet AMPs have a more flexible structure, and more diversified mechanisms of action [?]. AMPs rich in β -sheets can be divided into β -hairpins and peptides from the defensin family [?]. Many representative of the former class disrupt bacterial membranes via formation of toroidal pores: as an example, porcine peptide protegrin I triggers the toroidal pore formation assembling into a β -barrel structure when in contact with anionic membranes (while it folds into β -sheet aggregates on the surface of cholesterol containing membranes, thus acting selectivity on bacterial membranes only [?]).

In the case of defensins, many mechanisms are known according to the specific member of the family [?]. Some form transmembrane pores on planar bilayer when a physiologically relevant negative potential is applied to the membrane [?], while others form oligomers in phospholipid vesicles [?]. Although various descriptions of membrane damage have been reported, and include ion channels, transmembrane pores and extended rupture of the membrane, they are likely related, being a modulation of a similar acting principle.

Alternative mechanisms of action Finally, many non-lytic mechanisms are suggested for AMPs, especially for β -sheet structures: defensin A from *P. terramoviae* reduces the cytoplasmic potassium concentration [?], partially depolarising the inner membrane; tachyplesin from horseshoe crabs is able to bind to the minor groove of DNA, interfering DNAprotein interactions [?], and bovine lactoferricin can act synergistically with other antimicrobial agents by affecting the transmembrane potential and proton-motive force, resulting in inhibition of ATP-dependent multi-drug efflux pumps [?]. Moreover, after translocation within the cell, bovine lactoferricin can also inhibit DNA, RNA and protein synthesis. Section 1.5.1 will treat in detail the functioning of this AMP, distinguishing its role as membrane active peptide versus intra-cellular targeting compound: indeed, many works have focussed on lactoferricin antimicrobial processes versus locating the section of the sequence performing the membrane disruptive activity [? ?], to understand whether it retains the efficacy regardless of the fold.

These and similar strategies of investigations, conducted on several AMPs [?], provided the discovery of first minimal functioning antimicrobial blocks,

which promoted the understanding of how AMPs work in general, boosting the design of synthetic tailored AMPs from specific sequences.

1.2.3 Mechanisms of resistance to AMPs

Antimicrobial peptides are introduced here as a class of new drugs and a possible solution to the crisis of antimicrobial resistance. Any new drug entering the pool of the clinically approved compounds is (at least temporary) a solution to the problem of resistance to known antibiotics, but it must be clarified that bacteria can develop resistance to AMPs too. Never the less, the resistance to their action is generally not based on dedicated genes that are conferred by horizontal gene transfer, as in the case of many antibiotics resistance mechanisms [?]. Because of that, a certain increase of resistance after exposure to the drug is to be expected ('MIC creep'), but it is less likely to spread quickly to other species.

Some of the mechanisms of AMPs resistance are similar to the ones employed by bacteria to counteract small molecule drugs, for example over-expression of efflux pumps to dispose of AMPs, degradation of the peptide by extracellular enzymes and sequestration by the bacterial or biofilm matrix to prevent accession to the target [?].

Differently with respect to antibiotics hydrolysis, AMPs proteolitic degradation is operated by proteases, secreted on the extracellular side of the membrane specifically to destroy other proteins. Linear AMP are more prone to this type of degradation [?], as opposed to the ones presenting disulfide bonds [?], such as defensins, which nevertheless can be hydrolysed by more specific proteases enzymes [?].

Bacteria can also enhance their resistance to AMPs organising into specialised structures known as biofilms. These structures are formed by sessile bacteria adhering to a surface in an organized manner that allows the circulation of nutrients [?]. Biofilm bacteria secrete an extracellular matrix with adhesion and protection functions [?], which effectively repels and/or captures AMPs through exopolysaccharid or capsular polysaccharides. For example polysaccharide intercellular adhesin (PIA) produced by *S. aureus* and a variety of other bacteria is responsible for the resistance to both cationic AMPs (like HBD-3, LL-37) and the anionic one dermcidin [?]: deacetylation

of PIA increases its positive net charge, thus repelling more efficiently cationic AMPs, and increasing sequestration of the anionic ones at the same time [?].

It's worth noting that AMPs are nevertheless a promising alternatives in the treatment of biofilm-associated infections: indeed, in this type of infections (where bacteria are growing slowly) it is advantageous to have bactericidal agents such as AMPs, as opposed to bacteriostatic ones which target fast-growing bacteria only, as the majority of traditional antibiotics do [?].

But the most specific mechanism of AMPs resistance concerns modifications of the bacterial cell envelope: bacteria modify the characteristics of their surface to prevent the efficient binding of an AMP, even in the eventuality that the peptide reaches the bacterial envelope intact. The target of such modifications are different for Gram-positive and Gram-negative bacteria, according to their distinct cell envelopes. Gram-positive bacteria change the structure of their teichoic acids (TA): for example, D-Alanylation of TA observed in *S. Aureus* adds a positive charge to it, reducing the attraction of cationic AMPs and in turn increasing the cell wall density, so reducing the surface permeability [?]. Alternatively they can modify the bacterial peptidoglycan precursor, lipid II, which has a key role in the formation of the cell wall ad is often targeted by AMPs. For example, the replacement of its terminal D-alanine with D-lactate or D-serine [?] avoids the action of the glycopeptide vancomycin, as the functioning of this molecule proceeds by binding to the D-Ala-D-Ala terminal motif of the precursor.

In Gram-negative bacteria a positive charge can be added to lipopolysaccharides (LPS) by addition of amine-containing molecules [?] or by removing phosphate lipids (which have a negative charge) from lipid A [?], to obtain the same effects as for the increased charge in TA. They can also enhance the rigidity of the outer membrane to reduce permeability to AMPs via addition of extra acyl chains into lipid A [?]. Finally they can act at the cytoplasmic membrane level, as this is the final target of many antimicrobial peptides: in the eventuality that AMPs successfully pass the cell wall and reach this membrane, they are attracted to its surface by the negative charge of the lipids composing it, in particular phosphatidyl-glycerol (PG) and diphosphatidyl-glycerol (DPG, also called cardiolipin). Their negative charge can be masked by amino-acylation of the PG head group, so that the final compound repels AMPs through electrostatic interaction [?], or the overall rigidity of the cy-

toplasmic membrane can be enhanced, by an increase in saturated acyl chains which has been proven to confer resistance [?].

Finally, resistant bacteria often employ many of the aforementioned strategies at the same time, for example modification of the surface charge together with modification of other membrane components for a decreased recognition and augmented rigidity [?].

1.2.4 Principles of AMP design

The study and classification of AMPs provide knowledge on the characteristics a sequence must have to perform an antimicrobial function. As discussed in Section 1.2.2, there are some features which, comprehensively, help in discriminating AMPs against non antimicrobial peptides. The constantly increasing amount of data available is gathered in several curated databases [? ? ? ? ?], which catalogue AMPs (or subclasses of them, like membrane active, biofilm active or haemolytic peptides) based on such features to promote future data-driven prediction of the antimicrobial character of a sequence.

We recapitulate below a few key characteristics which are peculiar of AMP sequences. To be noticed that while some are easily retrievable from the sequence of the peptide, others imply direct experimental measures, and thus are difficult to implement into a simple sequence-based method to predict activity:

- **Structure:** both α -helical and β -sheet rich AMPs exist, as well as mixed structures. Short helix (~ 22 amino acids) and short β -sheet (~ 10 amino acids) are particularly common, acting through slightly different mechanisms (when known). When screening a potential AMPs, it must be considered that the fold close to a membrane environment might be different with respect to the one in solution.
- **Charge:** AMPs are charged moieties. Usually cationic (up to $\sim +10\text{ e}$), with fewer anionic examples (like dermcidin). Among cationic ones, not all the positive amino acids have equal role, for example Arginines are more effective than Lysins [?]. The potency, but also their haemolytic activity, are often directly related to the amount of charge [?].
- **Hydrophobicity:** AMPs contain also hydrophobic residues, usually with abundance of aromatic chains and specifically Tryptophan, as they

must insert and anchor into the lipid core of membranes, which is an hydrophobic environment.

- **Amphipathicity:** to host both the charged and hydrophobic residues, most AMPs organise themselves in an amphiphatic structure, i.e. the two types of amino acids side chains are located on the opposite side of the peptide.
- **Solubility** AMPs need good solubility to prevent aggregation in the aqueous environment they float in before arriving to the membrane. Aggregation would most likely impede their optimal interaction with the membrane.

A part from predictions regarding existing sequences, the knowledge of AMPs sequence-activity and structure-activity relationships is beneficial to find new, better performing ones. The design of new AMP sequences aims at producing peptides with improved characteristics:

- **specificity** against particular bacterial species;
- **stability** against the action of proteases, thus allowing a longer residence time in the body;
- **low cytotoxicity** at the therapeutic dose required (so an high therapeutic index).

The need for such improved peptides lies in the fact that their natural form constitutes a first broad spectrum defence our body employs against infectious bacteria and thus AMPs are often of mild potency. However, foreseeing their application as future drugs, it is desirable to tailor them to fulfil different criteria according to the infection to treat. At the present state of the art, a golden rule for the design of such sequences is still missing, however several methodological approaches to AMP design have been explored, and they can be grouped in three main lines: template based studies, biophysical studies and virtual screening [?].

Template based studies The main idea behind template based methods consists in modifying existing antimicrobial sequences in the direction of the

desired characteristics. The most widely explored templates are helical peptides, for their short sequences and because several of them (cecropin, magainin and protegrin) have been well characterised [?].

Ideally, an amino acid scanning of all the residues in an AMP provides information on the role of each of them, prompting at the most suitable mutations. High-throughput methods allow nowadays for such thorough investigation in the case of short AMPs [?], but similar, less resource consuming Alanine scannings can still point at the most important residues for the antimicrobial activity [?]. Alternatively, simpler approaches aim at designing peptides with enhanced charge and amphiphilicity, as these characteristics are deemed crucial for their effectiveness (see the paragraph above) [?].

However, all the above methods focus on single amino acids and can not take into account the interplay between residues, nor the three dimensional structure of the peptide. Without such information, it is difficult to extract general rules on why some mutations work better, and often the results of these studies give indeed enhanced AMPs, but cannot be generalised to other sequences. Therefore, recent works sought to integrate structural information on template based models, successfully designing peptides active against many bacterial lines at the same time [?] or enhancing the selectivity of some sequences for bacterial membranes [?].

A complementary approach to single point mutations on known peptides consists in designing minimal antimicrobial blocks: several investigations proved the importance of single residues and their intercalated pattern in natural and designed AMPs. For example, natural AMPs are rich in Tryptophan and Arginine residues [?], and synthetic ones have been produced with only Lysines-Leucine, or Arginine-Valine combinations to produce amphipathic helices [?]. An effort to extract principles from these examples is represented by text based models where amino acids constitute the letters and patterns occurring in natural AMPs are the grammar rules [?].

In general, the advantage of template based methods is in the reduced number of sequences to test, with decreased cost, as only a subspace of them is explored, namely the ones close to the original template.

Biophysical studies Biophysical studies aim at understanding the functioning of AMPs investigating their structure. Free energy perturbation, Molecular

Dynamics (MD) simulations and thermodynamics calculations can all provide knowledge on how the three dimensional arrangement of residues is important to allow their functional role. Contrary to sequence based methods, these techniques give an insight into the mechanism of action of an AMP but their drawback lies in the high computational cost. For example simulations can approach systems with a limited size, and only on short (microseconds) time scales, preventing the reproduction of phenomena of the order of millisecond (a detailed overview of the state of the art, advantages and drawback of MD simulations will be given in Chapter 2). For these reasons, such techniques have been applied to fewer systems in comparison to sequence based screenings, and only few mutations have been tested and compared *in silico*.

The strength of biophysical studies lay instead in the fact that they exploit the whole information available on a system (sequence, structure, chemistry), so they can single out the interactions that are crucial for a mechanism, clarifying whether they can be transferred to a different environment, or again they can discriminate cases in which similar sequences behave differently due to the environment around them. In this respect, they provide a generalisable knowledge applicable to different systems and thus to the design of novel AMPs at the atomistic level.

Examples of studies where MD simulations shed light on AMPs mechanism helping evaluating a mutant structure is the case of temporin [?] sequences, where a mutant with improved activity and decreased haemolytic activity was synthesised on a computational basis.

Virtual screening Contrary to biophysical assay, virtual screening methods are employed to analyse a large number of sequences, when an experimental or computational test of all of them would be prohibitive. The concept of these methods consists in the identifying descriptors which allow to predict the potency of the sequence: from the analysis of a database of AMP with known activity, a model is created and used to score novel synthetic sequences.

The recent evolution of machine learning (ML) techniques, artificial neural network in particular, gave a great impulse to virtual screening of AMPs (for a historically informed review see Table 1 in Ref. [?]). Machine learning appears particularly suitable to the task as the potency of AMPs is determined by the combination of many factors, the relative weight of which can be difficult to

identify. Moreover, these approaches can help in the identification of relevant features traditionally overlooked.

Practically, ML algorithms are trained on a set of AMPs labelled by their potency and characterised by different properties (features) along the line of the ones mentioned at the beginning of the section: sequence, partial charge, hydrophobicity, etc; but also experimental measures (pK measures, nuclear magnetic resonance data, octanol-water partition fraction and so on), together with theoretically computed features such as the van der Waals surface area. The more the input properties to consider, the more expensive is to train the model, but the higher the accuracy that can be meet. At the same time, the output descriptor is likely complex (i.e. combinations of many features) and thus of difficult interpretation. In principle, this is not a problem (provided there is no overfitting) as, rather than guide the design of AMPs from first principles, descriptors can be used to scores combinatorial sequences of the desired length to identify the best ones, but in practice, this is often difficult because of their exponentially growing number.

The second obstacle to ML procedures is given by the fact that the more features one wants to consider, the more sequences need to be given as input to the algorithm, i.e. need to be experimentally tested. Modern high-throughput synthesis methods, together with surrogate measures of bacterial killing are allowing it nowadays, as shown by Cherkasov et al. [?], who assessed the antimicrobial properties of thousands of 9 residues sequences and trained a neural network on the outcome, to then score novel sequences with good accuracy.

As already mentioned, peptide design has proven successful in producing sequences with improved potency or selectivity; however, it is still a case-dependent procedure, rather than a general, automated protocol easily applicable to enhance any sequence of choice.

1.2.5 Clinical applications

Antimicrobial peptides have been studies for many years, however the push to capitalise them to get compounds viable for the clinical stage has been delayed by many factors, including production costs, and lack of interest in the face of more potent small molecules which were deemed more economically advantageous by pharmaceutical companies. The constant creeping of AM

resistance though has focussed more effort on this class of compounds, mainly from small biopharmaceutical companies, and at present several of them are in clinical trials, in phase 1 or 2 [?].

The two major problems encountered so far for AMPs sequences in trial are the liability to proteolytic degradation, and the unknown toxicology profile when administered systemically [?]. For the last reason in particular, many of them in are trial for topical use against skin infections only, while they are deemed unsuitable for internal administration. Design of novel AMPs can be tailored to improve the liability to degradation, for example introducing D-amino acids, non natural amino acid analogues of opposite chirality, which, with appropriate formulations, are mimetic to the immune system [?]. Moreover, machine Learning protocols can help in pre-screening their toxicity through virtual screening methods.

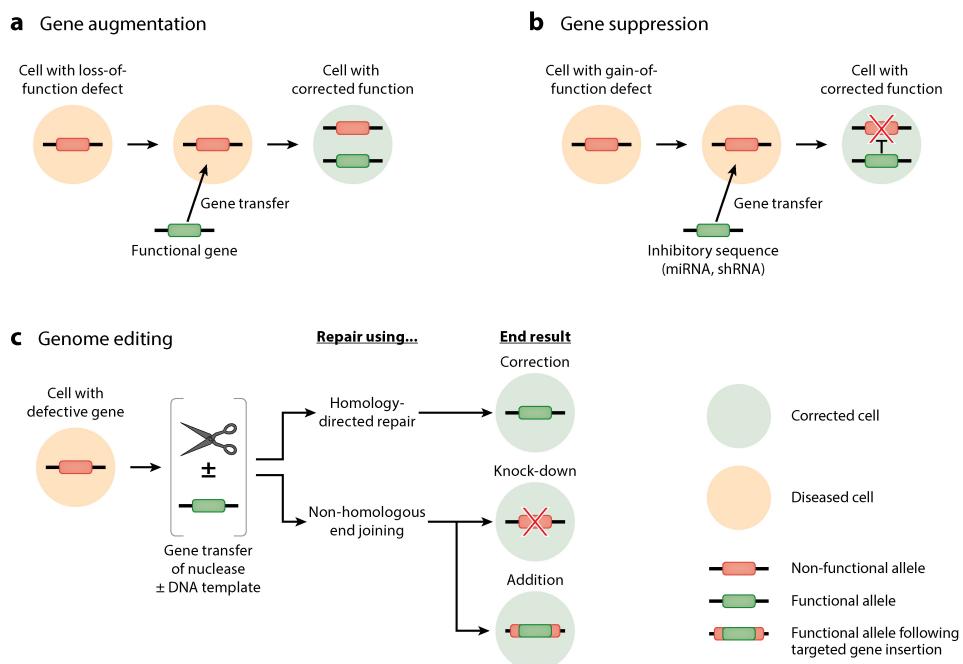
Overall, antimicrobial peptides remain a promising tool to counteract infections and, as their design is still - comparatively - in its infancy, there is room to explore novel applications and synthesise improved sequences apt to get to the clinical stage.

1.3 Gene therapy

Alongside the new compounds used to counteract bacterial infections, we want to bring the reader's attention to another class of therapies developed in the last decades for the treatment of non infectious diseases, which is relevant for the work of this thesis as well: gene therapy. In recent years it has greatly evolved and gained attention for the treatment of tumours, genetic diseases and complex acquired disorder [?].

The key concept is the delivery of genetic material to sick cells which possess a faulty copy of a gene, to influence its expression. Such fault can result in lack of synthesis of the protein of interest or in its misfold and/or malfunction. The correction can be performed in three different ways (Figure 1.4).

Augmentation gene therapy introduces an healthy gene copy to restore the normal functionalities of the protein of interest and thus of the cell: it usually consists in the delivery of a DNA strand, which in turn can be internalised in the genome and thus spread when the cell replicates, or not internalised and thus can influence the functionalities of the particular transfected cells only.



Anguela XM, High KA. 2019. Annu. Rev. Med. 70:273–88

Figure 1.4: Principles of gene therapy. Reproduced from [?].

Suppression gene therapy suppresses a detrimental gene, and this is particularly useful in the case of cancer, to impede cancer cells replication. Usually this strategy employs RNA interference, delivering miRNA (microRNA) or siRNA (small interfering RNA) strands which repress the transcription of the problematic RNA sequence. Finally, gene-editing, the most recent advance in the field, overlooks the possibility of correcting base pairs mutations to restore the original healthy sequence, and is often done through the functionalisation of the CRISPR-Cas9 technology, a mechanism found in prokaryotic organism as defence against viruses [?]. CRISPR (clustered regularly interspaced short palindromic repeats) is a library of DNA fragments from viruses that have previously infected the prokaryote, and the Cas9 enzyme (“CRISPR-associated protein 9”) uses these sequences to recognize and cleave strands of DNA complementary to the CRISPR sequence to blocks the reproduction of viruses if a following infection occurs. Research has been able to engineer the CRISPR-Cas9 technology to edit (rather than simply cleave) genes within eukaryotic organisms [?], thus performing a therapeutic role.

Despite the challenge posed by the development of genome editing tools,

and the risk associated to them (for example the possibility of deleterious insertional mutagenesis or deleterious immune responses), at present six gene therapies have received approval in the Western world [?], with many more undergoing regulatory review.

One of the main challenges in the development of such therapies lies in the identification of a suitable vector: delivery of free genome in solution results in poor internalisation and low therapeutic effect. Vectors instead allow the DNA/RNA to enter effectively into the cell: viruses can be used, modifying their genome to include the necessary sequence and remove the ones promoting viral replication [? ?]. But nowadays the outlook of gene therapy research lies not only in improving specific cargos to cure at the molecular lever more diseases, but also in the research of appropriate vectors with low toxicity, low induced immune response and high delivery efficiency. In that respect synthetic vectors started to be investigated for a virus-free delivery strategy. The system studied in this thesis proposes, among its other functions, to deliver genetic material into human cells.

1.4 Delivery of therapeutic material

The problem of gene delivery sets a parallel with the small drug one, introduced at the beginning of the chapter. Indeed small molecules needs delivering agents to be efficiently internalised in the cells, as well as genes do.

To reach the aimed organ, therapeutic molecules must be compatible with the different cellular environments they cross, but be preferentially retained, and act only on the ones they are designed for. This implies a subtle balance between a invasive activity on one side, and mimesis on the other, least the compound is recognised as dangerous and disposed by the immune and reticuloendothelial systems.

As en example, the trip of a standard, orally administered drug, passes through the digestive system, with its challenging acidic environment and limited permeation across the intestinal epithelium, and from there to the blood stream [? ?]. At this point, the drug is generally coated by a protein corona based on the molecule shape and charge [?]. The nature of this coating is difficult to predict and can disrupt or decrease significantly the efficacy of a compound as it modifies the way it is recognised and absorbed. From the blood

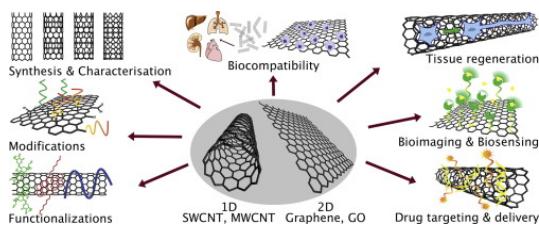


Figure 1.5: INCLUDE ONE EXAMPLE IMAGE (FROM PAPERS) FOR EACH MATERIAL (nanoparticles, carbon, polymers, lipids, DNA, proteins).

the drug diffuses in the tissues flanking the blood vessels naturally depleting its concentration downstream [?], so that regions further away in the line have less chances of getting a sizeable dose. Moreover, specific tissues are highly impermeable: the blood brain barrier (BBB) for example allows the passage only of small molecules with high lipid solubility [? ?]; while tumoural tissues are instead poorly vasculated, reducing the chances of delivery at their interior [?].

For all the above reasons, research has focussed on developing systems to assist the delivery of drugs. A mimetic carrier can not only improve delivery, but also be designed to selectively bind to particular tissues, or to trigger a drug release delayed time, or release it upon changes in environmental variables (for example pH) to reduce drug concentration in non targeted regions. A stand alone field of research has then focussed on the development of delivery vehicles irrespective from the quest for new drugs. The optimised products of the two separate efforts can then be paired according to the condition to address, to give a successful therapy.

At present, many molecules have been successfully employed to build drug vehicles, both organic and inorganic, to offer a range of different physico-chemical characteristics useful to target different cells [?] (Figure 1.5). We briefly list them to point out the variety and exotoxicity of structures which are useful, sometimes unexpectedly, to the medical world, and we then focus on a particular class, peptidic delivery vehicles: once more this class of molecule can offer a solution to a therapy-related task.

Inorganic materials for small drugs delivery Many inorganic compounds have been used to transport drugs or to enhance their biocompatibility. A first

class is constituted by metal nanoparticles, with golden nanoparticles covering a major role. Thanks to their metallic nature, these materials can be customised in shape and size (down to a 10 nm radius), and possess optical properties that allow to track them inside the body or to thermally stimulate them to trigger drug release [?]. They can also be coated with biologically active moieties to enhance their mimesis [?]. For these reasons, they are used to selectively treat tumoural cells, but up to now only a few golden nanoparticle based compounds have made to the clinical stage so far, as there are mixed evidence about their toxicity [?].

Similarly, materials made from carbon, especially carbon nanotubes, can be used for biomedical applications as they have a high loading efficiency thanks to the high surface area and easy interaction with biomolecules through van der Waals forces, π - π stacking or hydrophobic effect [?]. They can be conjugated to extra organic groups to increase their biocompatibility and have potential for targeted drug release upon change in environmental pH [?].

Finally polymers are a class of inorganic molecules already well validated as drug excipients. The most notable example is Polyethylene glycol (PEG): thanks to its high hydrophilicity, it is widely used to coat structures (e.g. inorganic nanoparticles, peptides) which in turn carry a drug [?]; or as a stand alone carrier system with high drug payload [?]. The great strength of polymers is their flexibility: as each of their constituent monomers can be either hydrophilic or hydrophobic, they can be engineered to assemble in many different structures, to swell slowly in water triggering a sustained drug release [?], or to undergo sol-gel phase transition upon specific changes in the environment [?].

Organic materials for delivery: lipids and DNA A somehow opposite approach for designing drug vehicles consists in using molecules similar to the ones present in the body, in an effort to exploit already available biocompatible materials and reduce toxicity [?]. In this category fall lipids, DNA and peptides.

A great variety of lipids is present in the cell membrane, and this is further enlarged by the production of synthetic ones. The components selected for drug delivery are usually taken from the biological lipidome, modifying the composition and possibly including synthetic molecules to tune the release

and robustness to degradation [?]. Being amphiphatic, lipids can encapsulate efficiently both hydrophobic or hydrophilic drugs, arranging respectively in micelles (monolayer spheres with the hydrophobic tails facing the interior) or in liposomes (bilayer spheres with a water filled core) [?], and by now, many lipids are approved as delivery agents for cancer and infection drugs [? ?].

DNA scaffolds are instead a more novel tool: DNA origami is nowadays an established technique to build three dimensional customised solids [?], and the nanometric knowledge about their constituents allows to fine tune them for a triggered release of the content [?]. First studies proved them successful in delivering anticancer agents [? ?], however they are very sensitive to different cellular environments which challenge their stability. This, united with high production costs and the relative young developments in their manipulation, prevented them to constitute a viable class of carriers so far, but at the same time holds promise for future improvements and applications.

Peptidic scaffolds Another widely used and trustworthy mimetic vehicle comes, quite surprisingly, from the world of pathogens: viruses have co-evolved with humans, to be able to penetrate into cells where they complete their reproductive cycle [?]. Therefore their capsid, the peptidic shell encapsulating the genome, is highly suitable for cell penetration. The first application sought historically was to employ genome free viruses to stimulate and train the natural immune response against the respective genome-loaded ones, creating viral vaccines - in a concept similar to the inoculation of dead bacteria to counteract the infections caused from them [?]. Later in the history, their potential as cargo carrier was pursued by first modifying the original genetic material to include sequences beneficial for the host cell, and inactivate the ones promoting the infectious duplication at the same time. In particular adeno-associated virus (AAV) has been widely studied [?] as it triggers a low immune response. To fully exploit the potential of a peptidic carrier many efforts have focussed on synthesising *in vitro* gene-free capsids, either as they appear in nature [?] or designing artificial building blocks, which assemble in so called Virus-Like particles (VLPs). This helps overcoming the reaction stimulated by specific viral capsids to which the immune system is sensible to. Similarly to other delivery vehicles, the surface of VLPs can be functionalised with additional molecules to improve the target selectivity and increase biocompat-

ibility, while the capsid peptidic scaffold grants robustness to the structure. Therefore, VLPs loaded with drugs can be tuned for an efficient intra cellular release [?].

A step further in engineering peptidic structures is represented by the design of self-assembling functional structures from first principles, exploiting the physico-chemical characteristics of peptides, regardless their resemblance of viral capsids. Indeed, self-assembling peptides can form nanostructures ranging from nanoparticles to nanotubes, nanofibers, nanorods and hydrogels [?]. Among their advantages, peptides present biocompatibility, a low production cost and a tunable bioactivity thanks to their chemical diversity, which helps in tailor the assembly toward the target of interest [?]. Moreover, the variety of amino acid available makes possible to load peptidic structures with both hydrophilic and hydrophobic drugs, according to their amino acid composition [?]. The peptidic self-assembly is modulated by the peptide length and its hydrophobic or hydrophilic character, given by its amino acid composition: simple phenylalanine dipeptides were designed with inspiration from a pathogenic pathway of molecular self-assembly and were shown to self-assemble in a multi-scale process producing nanotubes able to load drug molecules [?]. The relatively small diphenylalanine building block is non the less complex as it bears two charged termini (as the process is observed at neutral pH), and two aromatic hydrophobic rings, so that the dipeptide is driven towards assembly by the hydrophobic forces acting on the phenylalanine side chains and the complementary charges of the termini.

In a different approach, longer sequences can be employed to guide the formation of the local structure, as they organise spatially in well studied secondary structures with known interactions among themselves. This knowledge is possible as proteins are a fundamental component of the human body and as such an updated database of their structure is available (the Protein Data Bank [?]) and can be queried to understand how small peptides hierarchically assemble into larger units. Moreover, the vast literature on their interactions with membranes, cell receptors and in general biological components, can inspire the design of building blocks sensible to particular triggers within the body. From this background, the outlook of protein design often goes in the direction of surpassing natural limitations, synthesising exotic, non natural, geometries [? ?] for multifunctional materials.

1.5 Closing the circle: an antimicrobial drug delivery vehicle

Twice in this introduction peptide design has been brought to the reader's attention. First, it can produce antimicrobial peptides with improved potency or selectivity, or reduced toxicity. Second, design can engineer self-assembling building blocks for the formation of delivery scaffolds. As design is not bound to natural rules, it can foresee and imagine multifunctional materials which are not observed in nature. In particular, the introduction above poses the question of whether it is possible to engineer peptides able to perform both an antimicrobial and a delivery function at once (either of drugs or genetic material).

Such self-assembling, antimicrobial compounds would have a twofold interest for medical applications. First of all, self-assembly is functional to the antimicrobial activity: many AMP sequences have a weak potency, and only a high (critical) concentration can trigger the bactericidal mechanism. This is intuitive in the case of the carpet model strategy (see Section 1.2.2), where AMPs lay homogeneously on the surface of the bacterial membrane and breaks it upon sufficient coverage of its area. Also the barrel-stave and toroidal pore models rely on the mutual interactions between peptides to maintain the pore edges. Generally, as AMPs are positively charged, the localised presence of many copies of a sequence enhances the local electric field and charge imbalance, which are critical to the membrane stability. Second, in order for the assembly to be able to perform the additional delivery function, it must be able to either organise in a tailored structure (for example a capsule able to host a drug), rather than an amorphous aggregate, or to co-assemble with the cargo of interest.

Out of all the possible applications, the most promising is perhaps the use of such vehicles to deliver drugs to treat metabolic or genetic diseases: while the cargo tackles a defect of the host system, the vehicle can counteract the proliferation of bacteria. This is particularly important in situation where the host immune response is weakened and thus infections normally harmless can spread and cause damage. However, it must be noticed that the cargo is not bound to be a small molecule, as long as it can effectively co-assemble with the peptidic carrier. As mentioned in the previous section, gene therapy is also



Figure 1.6: Capzip molecule scheme and bond representation. [TO BE IMPROVED] Adapted from [?].

an actively expanding field which looks with interest at the development of vehicles for genetic material. Given that viruses have been the first choice for DNA/RNA delivery so far, peptidic carriers seem a natural evolution of them.

Given the above premises, it is evident the importance of pursuing the research on novel multifunctional peptidic materials. As mentioned when discussing AMPs design, to understand such systems, each of them must be characterised by itself, as a generalised knowledge is still lacking. With this aim, this thesis proposes to elucidate the behaviour of a specific synthetic self-assembling peptide, suitable for antimicrobial activity and gene delivery strategies. Its full characterisation will complete the knowledge on its mechanisms of action and complement the broader information already known on the class of such functional building blocks. This will be crucial to engineer new synthetic blocks with improved characteristics, either regard their antimicrobial activity, assembly performances, or tailored cargo delivery.

1.5.1 Capzip

The molecule capzip has been designed to perform the functions mentioned above at once. To recapitulate, the properties it possesses are:

1. assembly into nanoscale virus-like capsules with and without nucleic acids. This ensures that the vector can autonomously form and thus there is flexibility in the choice of the cargo;
2. antimicrobial activity of the molecule itself and of the capsule on a time scale useful for therapeutic applications;

3. promotion of gene transfer into mammalian cells when the peptide is co-assembled with the RNA strands, without causing cytotoxic and haemolytic effects.

Furthermore, the design effort aimed at building a template structure of minimal complexity, in order to reduce the synthesis effort to a short sequence. Arguably, short sequences are also more flexible in their assembly: it is thus important to explore them and prove whether even small blocks can form ordered structures.

Based on the above requirements, two design principles emerged: first the employment of a non-linear structure. There is indeed some evidence suggesting that non linear peptides are more prone to assemble in three dimensional structures, opposed to planar ones [?], and this holds in particular for short sequences which do not fold into a defined secondary structure. The second principle consists in the use of a template antimicrobial sequence which is short and has proved potency. Given that AMPs are usually anionic, the co-assembly with anionic RNA sequences is arguably inherited by consequence.

To satisfy the above guidelines, a short peptidic scaffold constituted by a β -Alanine and two Lysins has been engineered. Three identical copies of the antimicrobial sequence of choice are covalently bonded to the N-terminus of the scaffold sequence and to the nitrogen atom of the Lysin residues side chains (Figure 1.6). Regarding the antimicrobial sequence selected, it has been derived from the antimicrobial peptide bovine lactoferricin, which is in turn a portion of the Lactoferrin protein.

Lactoferrin Lactoferrin is an iron binding protein present in milk (in which it is most abundant, hence its name), saliva and other secretions, as well as in polymorphonuclear leukocytes. It works as an iron binder and provides a natural defence against bacteria and fungi [? ? ? ? ?], constituting a first defence for infants.

Lactoferrin contributes to bacterial suppression in several ways. At present, its known modes of action fall in three categories: first, thanks to its iron sequestering capabilities, it removes essential substrate required for bacterial growth [?]; second, it interacts with bacterial membranes and in particular binds to the lipopolysaccharides of bacterial walls, oxidising them and affecting the membrane permeability with consequent cell lysis [?]; finally it is

implicated in the stimulation of different immunological cells (killer cells [?], polymorphonuclear leukocytes, and macrophages [?]). The peptide fragment responsible for binding of lactoferrin to the bacterial membrane, named lactoferricin (Lfcin), has been identified near its N-terminus and found to have a more potent bactericidal effect than intact lactoferrin on a wide range of bacteria [? ? ? ?]. Similarly, a synthetic short peptide derived from a subsequence of human lactoferricin has been proven effective against bacteria as it depolarises the cytoplasmic membrane decreasing the pH gradient [?].

The bovine homolog of lactoferricin (LfcinB) has a higher bactericidal potency than human lactoferricin on several bacteria [?] and therefore has been more extensively studied. Its active core LFC is a 25-amino acid sequence which adopts a helical conformation in the full structure but, once isolated, crystallises in a β -hairpin with a disulfite bridge nearby the terminals which stabilises the fold, but was shown to be not essential for bactericidal activity [?]. Further experiments on LfcinB subsequences identified an even shorter antimicrobial core, constituted by the six amino acids RRWQWR [?]. This core presents a characteristic Tryptophan zipper motif WTW, which appears very often in nature in β -turn and β -sheet conformations, paired to another copy of the same motif, so that Tryptophan rings from facing strands are packed tightly against each other in an alternated way [?] (Figure –). In general, the six amino acid sequence contains both charged and hydrophobic residues, in line with the usual composition of antimicrobial peptides. Accordingly, its antimicrobial action is likely derived from the interaction with biological membranes through charge recognition first and aromatic rings insertion in a second moment.

To further elucidate this mechanism, several experimental investigations have been carried both on LfcinB and its subsequences. First, the structure of LfcinB in solution has been investigated by NMR (Nuclear Magnetic Resonance), resulting very flexible [?]. Then, the binding of its antimicrobial core to sodium dodecyl-sulfate micelles was studied [?], suggesting a favourable interaction of aromatic residues with the micelles surface. Similar experiments were performed on large unilamellar vesicles, constituted by lipids modelling biological membranes: ePE:ePC was chosen as a model of a mammal membrane, and ePE:ePG or ePC:ePG for a bacterial one. The experiments showed preferential binding to the latter ones, based on Tryptophan fluorescence [?],

suggesting a selective antimicrobial action on anionic membranes. Additional experiments have been performed on the full sequence or mutated subsequences [? ?] to investigate the binding to other different model membranes but, as the systems investigated are slightly different, as well as the experimental conditions, it is difficult to relate them and give a unified interpretation of the modes of action of lactoferricin derived peptides.

Finally, an alanine scanning has attempted to clarify the role of each amino acid in the antimicrobial activity of the LFC peptide [?]. The results suggested a binding function for the Tryptophan residues, in line with one of the roles Tryptophan can assume in antimicrobial peptides [?]. Other possible roles involve its propensity to form hydrogen bonds, in which case the residue would position itself at the interface between solution and membrane, rather than inside the latter (which happens instead when Tryptophan residues have a binding role).

The designed block From the active core of LFC (of sequence RRWQWR), a mutated sequence was obtained to comply the design criteria of a self-assembling building block. Two mutations were introduced to favour the assembly of arms belonging to different molecules in an antiparallel fashion. Specifically, given that the original RRWQWR sequence is found in a β -sheet (at least in the crystal lattice), the mutations aim at promoting a similar structure. Therefore, the Glutamine residue and the C-terminal Arginine of the lactoferrin motif were replaced with Threonine and Glutamic acid residues to have a self-complementary sequence RRWTWE: the pairing is promoted by the attraction of opposite charges at the ends of the sequence. Three copies of this sequence were thus covalently bonded to the scaffold described previously and shown in Figure 1.6, to obtain a self-assembling molecule in a three dimensional shape, hosting multiple copies of an actively antibacterial sequence.

1.5.2 A viable systems: experimental background and question

Many experiments have been performed to verify that capzip had the characteristics it was designed for. The set of experimental results obtained on the molecule has been published in Reference [?], while more recent results extend and consolidate the previous findings.

AFM/TEM	cryo-em
fluo hollow capsule	fluo RNA uptake

Figure 1.7: ... Reproduced from [?]

Experimental results First, the assembly ability have been tested: the peptide does not show assembly in pure water (as verified by Dynamic Light Scattering), while in biological buffer (MOPS, 150 mM) at physiological pH of 7.0 it forms capsules with dominating size range of 20-200 nm. This is confirmed by images of the capsules obtained with multiple techniques, namely transmission electron microscopy (TEM) (Figure 1.7), atomic force microscopy (AFM), and cryo-scanning electron microscopy (SEM). The fine structure of these assemblies appears irregular to the resolution power of such techniques. Some insight into the details of the assembly is given by Circular Dichroism (CD) spectra, which show a profile characteristic of β -turns and contain elements of a β -sheet structure and of indole rings, with minima at $\lambda \sim 200$ nm and 214 nm. Complementary evidence about the overall shape of the assembly was provided by the cross-sectional analysis of the assembled capsules by fluorescence microscopy using fluorescein to label capzip. The signal comes from the wall of the capsule only, showing an inner cavity (Figure 1.7). Finally, small angle X-ray scattering (SAXS) measurements were consistent with compact capsules interfacing with solvent.

The assembly process is also tested and monitored in combination with small interfering RNAs (siRNA): as mentioned in section 1.3, these sequences are a promising tools for RNA interference techniques which aim at inhibiting the expression of specific genes, however, they are easily degradable and thus difficult to deliver to the target cell without an appropriate vehicle. The co-assembly of a 21 base pairs duplex with the peptide shows the formation of structures similar to the ones formed by the stand alone peptide only: CD spectra highlight the helical signal from RNA together with the features proper of the peptide.

These co-assembled structures were tested for siRNA delivery in HeLa cells, showing that the presence of the peptide favours the internalisation of the genetic material with respect to the transfection results of a pure siRNA control. The delivery of fluorescent siRNA (Figure 1.7) showed that the internalisation

occurred within the first hours from the transfection in localised regions of the cytoplasm, suggesting an endocytic uptake. This distribution was stable over the first five hours of incubation after which the fluorescence signal decayed. Flow cytometry assays quantified the increase in siRNA uptake levels due to the presence of capzip, confirming that this molecule is competitive with other commercial transfection reagents (like Lipofectamine[®], unpublished results).

To further quantify the level of RNA internalisation, a mRNA knockdown experiment was performed on a HeLa cell line with two housekeeping genes, ACTB (β -actin, targeted) and GAPDH (reference) [?]. The silencing of β -actin mRNA was detected 22 ± 2 hours after transfection; and its knockdown “fitness” was expressed relative to cells treated with siRNA alone (background) and normalised against viable cell counts (Figure 1.7). Capzip fitness was lower than Lipofectamine[®] one, however cells treated with capzip remained viable after 24 or 48 hour, resulting in higher cell counts than the samples treated with the commercial reagent, suggesting that capzip has little cytotoxicity. The experiment above was performed at neutral to positive charge ratio close to one (where each siRNA molecule has a -42 e charge and capzip a +6 e charge), as test experiments performed at higher peptide-to-siRNA ratio showed no improved uptake.

Finally, the peptide does exert an antimicrobial function: the non-assembled peptide has shown to be effective against both Gram positive and negative bacteria (E. coli, P. aeruginosa and S. aureus), with no haemolytic effects and minimum inhibitory concentrations typical of other antimicrobial agents. On Supported Lipid Bilayer with negative total charge (mixed DLPC and DLPG, 3:1 ratio), the capsules create localized pores within minutes, as proven by AFM experiments repeated in time. The pore depth ranges between 1.4 and 2.2 nm, which is smaller than the radii of the capsules, however it is sufficient to disrupt the structure of the membrane. Finally, to prove the viability of capzip as antimicrobial agent *in vivo*, it was used to counteract methicillin-resistant S. aureus (MRSA) infections in G. mellonella larvae. The particular bacterial strain used was susceptible to vancomycin, which could be used as control: the larvae treated with capzip showed survival rates significantly higher than the untreated control, and comparable to those treated with high doses of vancomycin (unpublished results).

Open questions Despite the success of the experiments mentioned above, there is much information still to be uncovered on the precise mechanism of action of such peptide.

Specifically, both the assembly process and the antimicrobial mechanisms contain some unknown: regarding the former, it is important to understand which amino acids or sub-structures allow the pairing of molecules, whether such pairing is specific or not, how reversible it is, and how rigid the final structure is. Regarding the latter, it must be highlighted what molecules in the membrane the peptide binds to, and how this binding affects the full membrane structure. Finally, as there is evidence that the assembled molecule is a more powerful antimicrobial compound than the single molecule, it is interesting to understand whether any cooperative action is taking place or the enhanced antimicrobial power of the assembly is due only to the localised higher concentration.

Even if further experiments or future improvements in the techniques already employed might tackle some of the aspects above in a near future, arguably no experimental outcome can provide an atom-by-atom knowledge of the processes of interest in any time soon. Ideally though, one would like to track each of them, i.e. the processes happening in any the environments capzip has been exposed to (physiological solution, supported lipid bilayers, bacterial extracellular matrix, mammal cell membrane and cytoplasm) both in space and time with the finest level of details, and the impossibility of that leaves large gaps in the understanding of the system.

1.6 A computational approach to understand capzip

The gaps mentioned in the characterisation of the systems prompts for new investigations in order to complement the knowledge already provided.

Beside the quest to enrich the fundamental knowledge on self-assembling peptides and antimicrobial ones, the understanding of this very system is crucial for its further development. We outlined already in Section 1.2.4 how antimicrobial peptide design can proceed from already viable templates and empirical principles, when first principles are not available. Similar rules hold for designing self-assembling peptidic materials, to obtain tailored delivery vehicles (see Section 1.4). Therefore, a full knowledge of the interactions be-

tween peptides and between their assembled structures and the membrane, i.e. of the mechanism of its functions, will drive the engineering of new likewise peptides. A knowledge-driven design would hopefully provide new blocks suitable for a double action as the one capzip performs, and this in a shorter amount of time than a research based on less information or on a trial-and-error procedure of mutations in the chemical composition or in the architecture of the molecule. A few examples of possible knowledge-related improvements include the following:

- the knowledge of capzip binding mode to the bacterial membrane might suggest its suitability as a broad range spectrum compound or the possibility of tuning its action against specific pathogens;
- understanding the molecule-molecule interactions classifies the robustness of the assembled structure and the possibility of designing block which disrupts under particular chemical conditions only;
- querying the electrostatic profile of the assembled structure suggests which type of molecules, other than siRNA, could be efficiently co-assembled and thus delivered.

In recent years, computational techniques are stepping forward to complement incomplete experimental knowledge and complete the picture of how biological systems work. For this reason, it seems natural to query such techniques to study the capzip system as well. Zooming into the details of the interactions can be performed via a theoretical modelling of the system in time, and thus through the simulation of its evolution, starting from few basic principles and the knowledge of the chemical composition of its parts. The technique this work focusses on is Molecular Dynamics simulations, which aims at reproducing the behaviour of a system of atoms in a semi-classical description using basic physical laws, as it will be described in details in the next chapter.

Thus it is the aim of this thesis to prove that Molecular Dynamics simulations can clarify further details on the assembly mechanisms of capzip and on its interactions with biological membranes, in order to gather more information on the system and contribute in the future to the designed of new molecules with enhanced functional capabilities.

Chapter 2

Methods

MOLECULAR DYNAMICS simulations have been rightly defined as the ‘Computational Microscope’ [? ?] as they offer otherwise inaccessible insights into the molecular details underlying conformational changes of proteins and nucleic acids. Computational methods and tools based on MD are routinely applied in structural biology to quantitatively characterise the dynamics and thermodynamics of proteins and their complexes. The increasing computational power available, and the flexibility of the algorithms which can be implemented on different platforms, have made possible to access molecular dynamical properties inaccessible to experiments. These techniques and the associated force fields are also commonly used in the process of structure determination from NMR data and theoretical structure prediction [? ?].

Modelling and simulating a biological system consists in describing its components and their mutual interactions, and implementing the laws of physics in the attempt to reproduce its natural evolution. A quantum mechanic description would be the most accurate but expensive to achieve for large systems. To facilitate the task, several simplified models have been devised, each most suitable to investigate particular cases. In particular, a large class of models focuses on a classical mechanics description of the dynamics: by increasing the size of the systems and with longer time spans described, the classical approximations will become more accurate, and objectively the only possible representation computationally affordable.

We briefly present here the core theory and implementation of classical MD simulations, together with a discussion of their strengths, limitations and comparison with experiments. Understanding their methodology provides the in-

interpretative key with which simulations must be designed, run and interpreted in each specific case [?]. A review of relevant successes of MD simulations will complete the chapter.

2.1 Algorithms for Molecular Dynamics

In this section, the core algorithm behind MD simulations is presented, together with the corrections and refinements implemented on it, as it sets the ground for the approximations to follow: in classical MD, the system will be processed by an engine based on classical mechanics rules and finite steps approximations, and the forces acting on it can thus be modelled with a similar degree of approximation.

2.1.1 The Newton's law

In the aforementioned framework, Newton's second law of motion rules the dynamics, stating that the acceleration \mathbf{a} that a particle is subject to at time t , depends on the total force \mathbf{F} acting on the particle itself and on its mass m (bold denotes vectorial quantities):

$$\mathbf{F}(t) = m \cdot \mathbf{a}(t). \quad (2.1)$$

As the acceleration $\mathbf{a}(t)$ is the second derivative of the position $\mathbf{r}(t)$ with respect to time, given the initial position and velocity of the particle $(\mathbf{r}(t_0), \mathbf{v}(t_0))$, their temporal evolution can be computed integrating $\mathbf{a}(t) = \mathbf{F}(t)/m$ as follow:

$$\mathbf{v}(t) = \mathbf{v}(t_0) + \int_{t_0}^t \frac{\mathbf{F}(t')}{m} dt'; \quad (2.2)$$

$$\mathbf{r}(t) = \mathbf{r}(t_0) + \int_{t_0}^t \mathbf{v}(t') dt' + \int_{t_0}^t \int_{t_0}^{t'} \frac{\mathbf{F}(t')}{m} dt'' dt'. \quad (2.3)$$

In the case of complex biomolecular systems with many particles and multiple interactions acting between them, it is impossible to integrate analytically Equations 2.2-2.3, while a different and feasible approach consists in discretising them. The idea is to consider very short time steps of length Δt , so that in such intervals the forces can be considered constant, and the integration of Equations 2.2-2.3 becomes trivial. The simplest choice consists in considering

for the integration both the acceleration and the velocity at time t_0 , producing the Euler algorithm:

$$\mathbf{v}(t_0 + \Delta t) = \mathbf{v}(t_0) + \frac{\mathbf{F}(t)}{m} \Delta t; \quad (2.4)$$

$$\mathbf{r}(t_0 + \Delta t) = \mathbf{r}(t_0) + \mathbf{v}(t_0) \Delta t + \frac{\mathbf{F}(t)}{m} \Delta t^2. \quad (2.5)$$

This procedure contains approximation of the order of $(\Delta t)^2$ that will accumulate step after step, biasing greatly the outcome. However, a careful choice of the values to integrate allows to reduce considerably this approximations. For example, choosing the acceleration value at time t_0 but the velocity at time $t_0 + \Delta t/2$ decreases the error down to orders of $(\Delta t)^4$. This framework is at the basis of the so-called leap-frog algorithm, which is used in the vast majority of MD engines:

$$\mathbf{v}\left(t_0 + \frac{\Delta t}{2}\right) = \mathbf{v}\left(t - \frac{\Delta t}{2}\right) + \frac{\mathbf{F}(t)}{m} \Delta t; \quad (2.6)$$

$$\mathbf{r}(t_0 + \Delta t) = \mathbf{r}(t_0) + \mathbf{v}\left(t_0 + \frac{\Delta t}{2}\right) \Delta t. \quad (2.7)$$

This algorithm can thus “solve” every possible Newton equation, at the expenses of some precision.

Considering that MD deals with bonded atoms belonging to multi-atoms molecules, the length of these bonds must be kept within a physically meaningful value. The approximate procedure mentioned above may however give raise to unphysical configurations, even if the forces acting between single atoms tend to bring them in the correct geometry. On the other hand, if the initial configuration was far from equilibrium, very abrupt changes in the bond length can arise in the initial steps, which are unphysical as well. To correct this, constraint algorithms are applied after the update of atoms positions, to limit the change in bond length. The ones used in this work are LINCS (Linear Constraint Solver) [?] and SETTLE [?]: the first finds a solution iteratively within an approximation tolerance, as the problem of constraints is hardly solvable analytically; the second is an exact implementation of the solution for rigid bodies of three elements only, and as such is useful for the treatment of water molecules (which have three atoms in their atomistic description).

2.1.2 Thermostats and barostats

As an addition to the geometrical constraints algorithms mentioned above, other specific algorithms have been developed to more realistically reproduce the simulation's conditions of choice, for example Temperature and Pressure.

Most experiments are conducted under constant temperature, so that it is desirable to reproduce this condition in simulations. To set up a temperature, at the beginning of the simulations particles can be given random initial velocities sampled from the Maxwell-Boltzmann distribution, which describe velocities of atoms of a noble gas at temperature T . The velocity of each of them will be influenced by the specific interactions occurring in the system but, in a constant temperature environment, the total average kinetic energy (proportional to T) must remain constant. Even in absence of any dissipative term in the dynamics, the approximations performed in the MD algorithm make this quantity to drift away from its initial value, therefore to ensure temperature is maintained throughout the simulation, thermostat algorithms have been devised.

The principle behind a thermostat consists in rescaling the velocity of one or few selected particles at fixed interval of times, to restore the correct average kinetic energy. This fixed interval of time must not correspond to the timestep itself, as the goal of a thermostat is to maintain the average temperature, and not its value at all times, as fluctuations are expected in natural systems. Moreover, it is strongly recommended to couple solute and solvent to separate baths, to ensure that both maintain the correct temperature. Indeed, it is possible that the energy exchanged between solute and water (or other components) is not perfect, due to different conditions adopted for their simulations, like, for example, cut-offs or restraints. The most used thermostat algorithms are the Berendsen [?], Nosé-Hoover [? ?], Andersen [?] and velocity rescale [?], which differ in the quantities selected for the velocity rescaling.

Another macroscopic condition one wishes to maintain is either the volume or the pressure of the system. While maintaining the volume constant is easy (and, combined with constant temperature, gives the NVT ensemble), pressure regulation (i.e. maintaining a NPT ensemble) requires a barostat. Pressure is directly proportional to the average quantity of motion exchanged between the particles and the walls of the box they are confined to, which depends on the

frequency of collision and thus on the extent of the box. Barostats change the box size to regulate the pressure. It has to be noticed that most MD simulations are run under periodic boundary conditions, i.e. a particle which exits from the simulation box during a move is brought back on the opposite side of the box, leaving the box density constant. This mimics the presence of an infinite number of equivalent boxes one next to the other, and alleviates the finite size effects that arise when simulating small systems. In this scenario, particles are not bouncing on the box walls, rather a virtual pressure is computed from the velocities of the ones trespassing the box boundaries during a move. Similarly as for thermostats, the coupling frequency of barostats must be larger than the time step. Usually all the box dimensions are rescaled by the same amount. However, in the case of anisotropic systems like lipids, to maintain the correct pressure, the directions parallel to the membrane plane can be rescaled separately with respect to the one perpendicular to it. Also for pressure coupling several algorithms can be used: the Berendsen [?], Parrinello-Rahman [?], or Martyna-Tuckerman-Tobias-Klein (MTTK) [?] (to be used in conjunction with the Nosé-Hoover thermostat), are all barostats which differ in the way they approach the desired pressure (e.g. exponentially, in an oscillatory way, etc.).

2.2 The force field problem

Once the equations of motion and the control algorithms are set up, the next challenge is represented by modelling the forces and thus the potential energy function. Force fields for classical MD simulations usually rely on the breakdown of interactions into several independent, additive and derivable terms, identified on an empirical physical basis. We report here the functional form adopted for the GROMOS force field [? ?] as implemented in the GROMACS MD engine [? ? ?], as an example of the structure of a classical force field. Other force fields can have slightly different implementations, however the general classification of interactions and the type of functional forms used are similar.

Covalent (bonded) interactions Covalent interactions are modelled with potential energy terms representing bond stretching, angle bending, improper

and proper dihedral angles torsion. The functional form of the potential-energy for bonded interactions aims at a simplified, semi-classical description of the sub atomic motion of molecules, assuming harmonic-like vibrations around the equilibrium position of the bond, angle or dihedral in exam.

Specifically, in the GROMOS force field, a bond between atoms i and j is described by a fourth power potential (similar to a harmonic form, but computationally more efficient):

$$V_b(\mathbf{r}_{ij}) = \frac{1}{4} k_{ij}^b (|\mathbf{r}_{ij}|^2 - b_{ij}^2)^2 \quad (2.8)$$

where the force constant k_{ij}^b is given in kJ/mol/m² and b_{ij} is the equilibrium position of the bond between atom i and j .

The deviation of the angle between three atoms i , j and k from the preferred value θ_{ijk}^0 is implemented through a cosine based angle potential:

$$V_a(\theta_{ijk}) = \frac{1}{2} k_{ijk}^\theta (\cos(\theta_{ijk}) - \cos(\theta_{ijk}^0))^2 \quad (2.9)$$

$$\text{with: } \cos(\theta_{ijk}) = \frac{\mathbf{r}_{ij} \cdot \mathbf{r}_{kj}}{r_{ij} r_{kj}} \quad (2.10)$$

with k_{ijk}^θ in kJ/mol.

Improper dihedrals are used to ensure ring planarity and control the chirality of some tetrahedral centres. They are described through a harmonic potential:

$$V_{id}(\xi_{ijkl}) = \frac{1}{2} k_{ijkl}^\xi (\xi_{ijkl} - \xi_{ijkl}^0)^2 \quad (2.11)$$

where the ξ values are given in degrees and the force constant in kJ/mol/rad². By convention, the improper dihedral for a set of four atoms i , j , k and l , is taken as the angle between the plane defined by atoms (i, j, k) and the one defined by atoms (j, k, l) .

Finally, the last bonded interaction is represented by proper dihedrals, described though a periodic potential:

$$V_d(\phi_{ijkl}) = k_{ijkl}^\phi (1 + \cos(n \phi_{ijkl} - \phi_{ijkl}^0)) \quad (2.12)$$

following the convention that ϕ_{ijkl} is the angle between the (i, j, k) and $(j,$

k, l) planes, with i, j, k , and l four subsequent atoms (for example along a protein backbone). A value of zero for a proper dihedral corresponds to a *cis* configuration; n denotes the number of equally spaced minima available for the dihedral in a 360° turn. The constant k_{ijkl}^ϕ is expressed in kJ/mol.

It must be noticed that these potentials can not model the rupture of a bond: for this, more sophisticated descriptions are needed.

Non bonded interactions Non bonded interactions include the short range Pauli repulsion, the “mid”-range van der Waals attraction, and the long range electrostatic term.

The first two can be modelled together by a Lennard-Jones potential. Its functional form, describing the interaction between two neutral atoms at distance r , models the long range dispersion with a r^6 behaviour typical of the dipole-dipole interactions found in noble gases (London dispersion forces), while the Pauli term is represented by a r^{12} behaviour to ease the computation in relation with the previous one:

$$V_{LJ}(r) = 4\epsilon \left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^6 \right]. \quad (2.13)$$

Two parameters, ϵ and σ , tune the interaction strength and the equilibrium distance. They are fitted against experimental data and are specific of each pair of atoms species.

The Coulomb energy between two charges q_1 and q_2 at distance r is represented by the Coulomb law:

$$V_C(r) = \frac{1}{4\pi\epsilon_0} \frac{q_i q_j}{\epsilon_r r_{ij}} \quad (2.14)$$

with ϵ_0 the dielectric constant of vacuum and ϵ_r the relative dielectric constant, introduced to properly take into account the screening provided by the material surrounding the object, as polarisability is not included in this description.

The treatment of non-bonded interactions requires particular care because of their long range nature: in every point of the simulation box many forces from distant atoms are acting at the same time, making the prediction of the outcome difficult, moreover small shift in the parameters choice can give very different “macroscopic” results. The van der Waals forces decay fast,

therefore the tail of their functional can be cut after a threshold distance with little impact on the outcome; while Coulomb interactions, with their slower decay, must be taken into account throughout the whole simulation box. Many algorithms have been devised to efficiently compute them, like the Particle Mesh Ewald [?] or the Reaction Field [?] approaches.

Finally, all biomolecular force fields, and in particular their van der Waals interactions, are parametrised to describe systems at room temperature, therefore simulations performed at substantially different temperatures must be interpreted carefully.

2.2.1 Force fields: classifications

Many force fields for classical MD simulations adopt a functional form equal or similar to the one described above. Their difference lies in the number of degrees of freedom modelled, in a hierarchy of descriptions proceeding from detailed to coarse (Figure 2.1). Three possible classes of descriptions are:

- all-atoms force fields, where all the atoms are present in the description, represented as spheres of variable size according to their van der Waals radius (e.g. proportional to σ in a Lennard-Jones model). Examples of all-atoms force fields are AMBER [? ? ?], CHARMM [? ? ?] and OPLS all-atom [?].
 - united atoms force fields, similar to the previous ones but where non-polar hydrogens are incorporated in the heavy atom they are bonded to. The “united atom” is given a new σ parameter and increased mass according to how many hydrogen it includes. The GROMOS force field [? ?] follows this philosophy, and the OPLS force field has also a united atom version [?].
 - coarse-grain force fields, which group together in one unique bead few atoms, to reduce the number of variables to compute. The clustered atoms are such that their mutual distances are expected to vary little with respect to the ample movements of components of the system far away from each other (which will be grouped in different beads). The MARTINI [? ? ?] and SIRAH [? ?] force fields belong to this category.
-

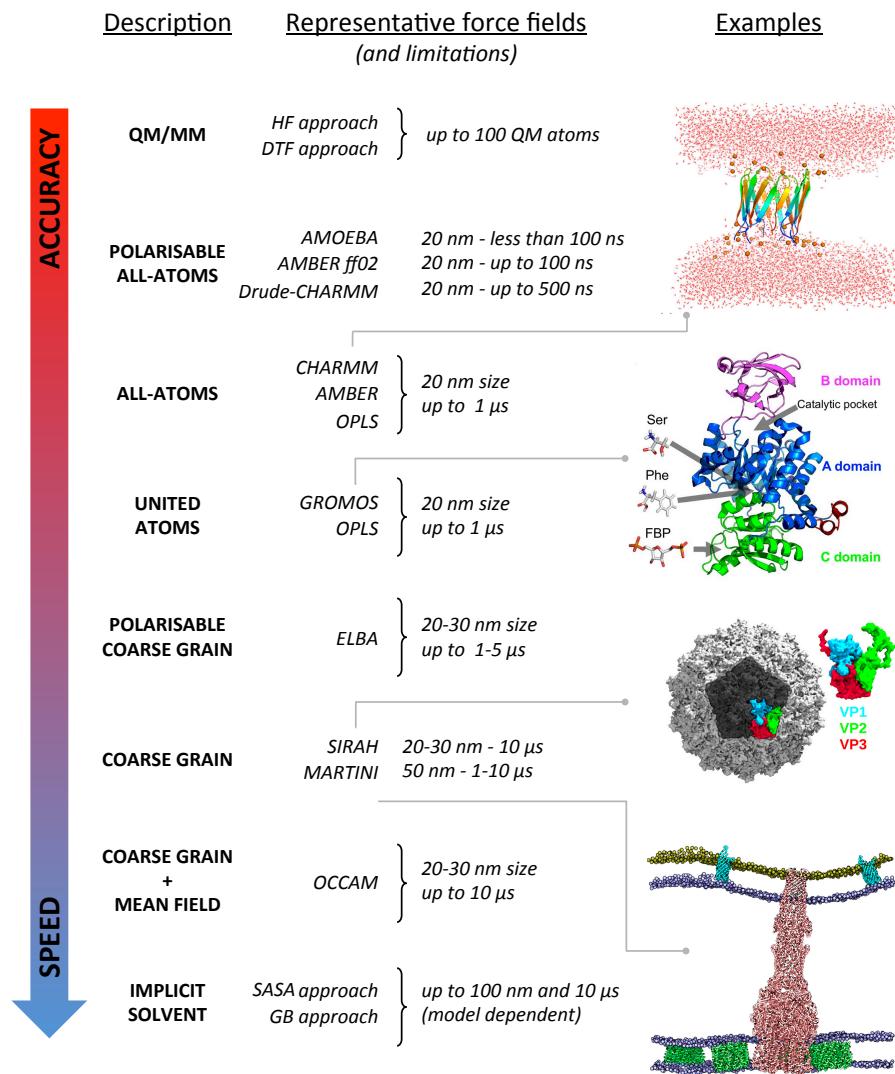


Figure 2.1: List of most popular simulation force field for biomolecules, ordered from detailed to coarse (references in Section 2.2.1). On the left, snapshot of notable systems simulated with the force fields CHARMM (adapted from Ref. [?]); GROMOS (adapted from Ref. [?]); SIRAH (adapted from Ref. [?]) and MARTINI (adapted from Ref. [?]).

We now give a more detailed insight in the characteristics and parametrisation strategies of the atomistic and coarse-grain force fields which will be employed in the work of this thesis.

2.2.2 The GROMOS force field

All-atoms and united atoms force fields are parametrised against first-principle or experimental values. While for the all-atom force fields AMBER and CHARMM the parametrisation is based on ab initio quantum mechanics calculations refined against experimental data [? ? ? ? ?], the united atom GROMOS force field relies on the reproduction of heat of vaporization of small molecules and free enthalpies of solvation of small compounds in different solvents, at physiological temperatures and pressures [? ? ?]. This procedure sets not only the constants of the bonded interactions, but also the partial charges of the atoms inside a molecule: as no electrons are included for the sake of efficiency, their redistribution across atoms which are bonded is modelled through fractional charges assigned to each atom (while the total charge of a molecule must sum to an integer). Moreover, it is assumed that the parametrisation performed for small moieties can be transferred to a larger compound including these moieties. This limits the number of chemical groups to be described in order to simulate biomolecules.

In every MD simulation, the description of water is crucial. Out of the many water model proposed, the GROMOS parametrisation has been performed with a flexible simple point charge (SPC [?]) model. This description represents water as a three atoms molecule, with a negative charge on the oxygen and a positive complementary one on the two hydrogen atoms, and allows flexible hydrogen-oxygen bonds. This model reproduces correctly the density and dielectric permittivity of water. To be noticed that, computationally wise, water molecules are the vast majority of the particles involved in a simulation and thus a significant fraction of the computer time is spent in updating their positions and calculating solvent-solvent interactions.

The improvement of computational techniques and reparametrisation strategies prompts the periodical release of newer versions of force fields. In the present work, we employed version 53a6 of the GROMOS force field [?] for the set of simulations involving peptidic assembly in solution, while we switched to 54a7 [?] and 54a8 [?] for the simulations involving biological

membranes. While it is advisable to always have a coherent set of parameters across simulations, to compare their outcome in a consistent manner, when extending the system simulated to include membranes, we deemed the newer parameter sets more suitable because of the improvements introduced in the phosphocholine head parametrisation (see Chapter 4 for a complete discussion on lipid parametrisation in GROMOS).

2.2.3 The SIRAH force field

The first coarse grain force field we introduce groups multiple atoms in one bead but aims at maintaining chemical and structural details of the biomolecules described. As such it sets itself between the atomistic GROMOS description, and the coarse grain MARTINI force field [? ? ?] which will be introduced in the following.

Two approaches are possible to develop a coarse-grain description: parameters can be fit directly to global quantities derived from experiments, in a top-down approach as performed in the atomistic GROMOS parametrisation; or coarse-grain simulations results can be fit to outcomes from atomistic ones, in a bottom-up approach.

SIRAH [? ?] is a top-down force field derived to fit structural properties of proteins. It aims at reducing the complexity of an atomistic description while still being able to reproduce the correct secondary structure of proteins across a wide variety of folds contained in the PDB, together with a correct representation of their dynamics.

To obtain this, it opts for a non-uniform granularity, i.e. according to the region of interest a different number of heavy atoms is grouped in a bead, from a minimum of two up to four. In the case of proteins, it maintains the backbone flexibility by grouping NH, C_αH and CO in three different beads, while side chains are represented with less details, generally grouping three atoms together. A schematic of the mapping for each amino acid is shown in Figure 2.2. Contrary to force fields where the amino acid backbone is mapped to one bead only, the SIRAH description allows to reproduce secondary structures without recurring to additional constraints. The dual granularity approach is based on physico-chemical intuition, and is more difficult to generalise than a uniform one. Nevertheless, the force field has been recently extended to lipids

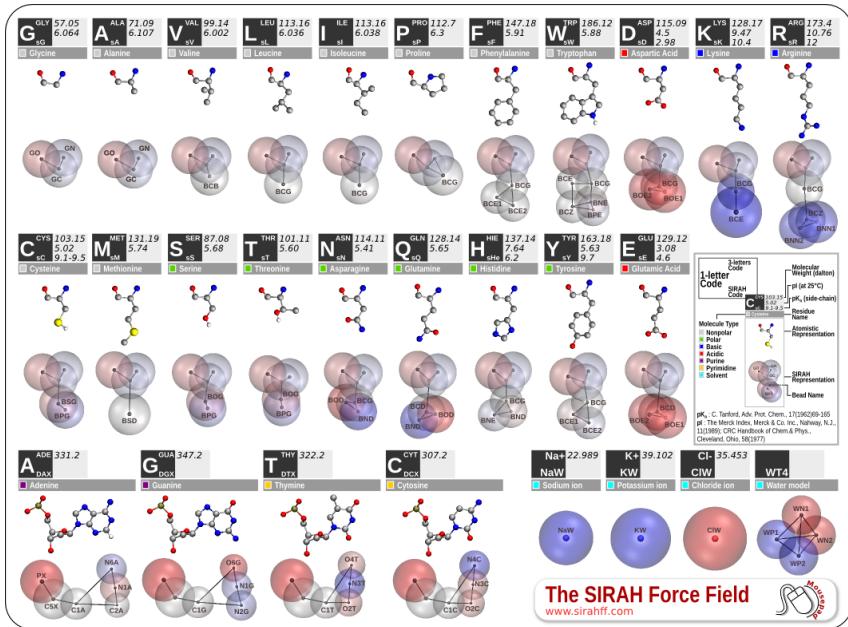


Figure 2.2: Description of amino acids and nucleic acids in the SIRAH force field. Reproduced from [?].

[?], while it comprised a parametrisation for DNA molecules since its infancy.

The modelling of water in a coarse-grain force field is also critical: usually, a few water molecules are grouped together in one bead. This has two implications: water particles are large and thus cannot solvate very narrow pockets; moreover, collapsing the molecules in one single point in space removes the separation of charges and the characteristic dipole every water molecule should have is lost. The dipole of water is responsible for hydrogen bonds formation and for the electrostatic screening observed in an aqueous solution. Such screening can be roughly modelled tuning the relative dielectric constant, but as this is a mean field approach, it cannot account for local effects. To partially obviate to that, SIRAH force field maps four waters to a tetrahedral molecule, with one bead on each vertex: all the bonds are rigid, and the structure serves the purpose of having a repartition of plus and minus charges, by assigning a positive charge to two vertices and the opposite charge to the other two, giving a polarisable structure. The geometrical arrangement reproduces the tetrahedral network of water molecules observed in its liquid state, which is characteristic of this fluid and tunes its remarkable properties.

Based on the above premises, SIRAH force field simulations of different

peptides and proteins in solution proved to match the relative NMR results, showing a good reproduction of secondary structures; simulations of lipids randomly oriented in water showed the formation of an organised bilayer, and the expected behaviour of a few transmembrane proteins in model membranes was correctly reproduced [? ?].

2.2.4 The MARTINI force field

The MARTINI force field is another popular coarse-grain description of biological molecules [? ? ?]: developed originally with a focus on lipids, much earlier than SIRAH, it has been then extended to include proteins, small ligands and DNA/RNA molecules.

MARTINI opts for a four-to-one approach, i.e. four heavy atoms are grouped in one bead, resulting in a uniform graining and a coarser description than the SIRAH one. The only exception to this scheme is represented by rings molecules, where a two-to-one approach is needed to maintain the circular topology (consequently, these beads have a reduced mass with respect to the others, all described with the same mass value).

The number of bead types has been kept to the minimum necessary to represent biological molecules. They are organised systematically in polar, non-polar, apolar, or charged, and each type has a number of subtypes with increasing polarity to differentiate the chemical nature of the underlying atomistic structures. This systematic parametrisation can be easily transferred to new compounds, without the need of introducing new bead types, analogously to what GROMOS does when parametrising small moieties.

Similarly to GROMOS, the MARTINI force field chooses a top-down approach to parametrise non-bonded interactions, tuning them against experimental partitioning free energies between polar and apolar phases, while bonded interactions are derived from reference all-atom information, in a bottom-up approach. Specifically, they are designed so that the results from coarse-grain simulations match the structural data of the underlying atomistic geometry, derived either from available structures or atomistic simulations. For example the distribution of the length of a bond in a molecule would be mapped to its coarse-grain version (i.e. computing the corresponding bead-to-bead distance) and compared with the respective one obtained from coarse-grain simulations.

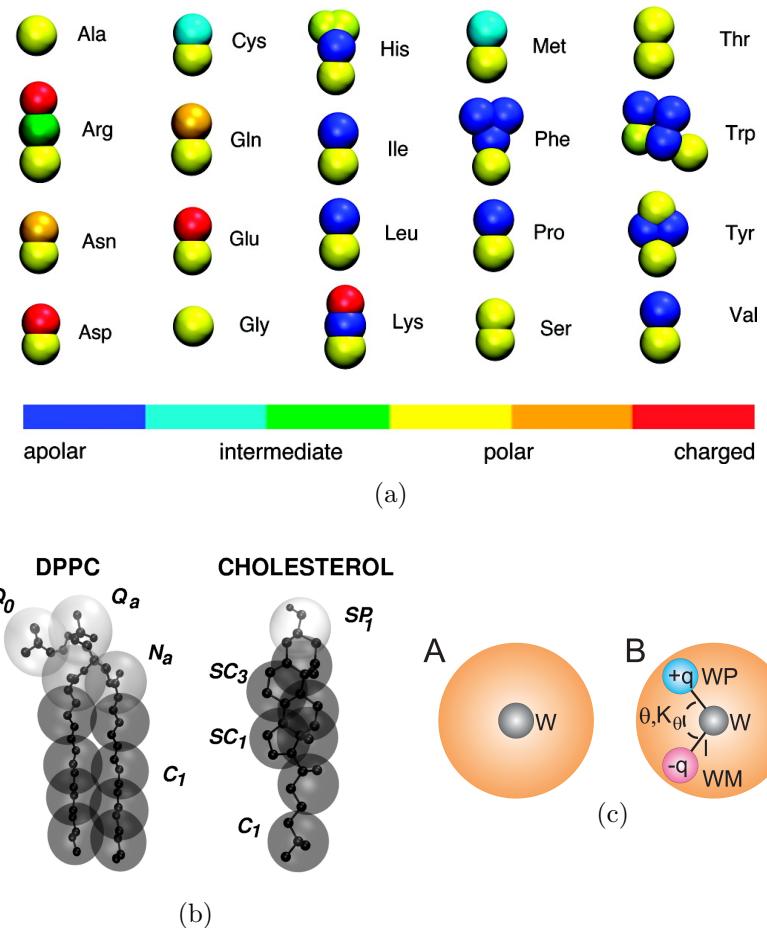


Figure 2.3: Description of amino acids (a), example lipids (b) and water models (c) in the MARTINI force field. In (c) the orange sphere represents the van der Waals radius of the central atom. Reproduced from [? ? ?]

The four-to-one mapping implies that the amino acid backbone is represented by one bead only (Figure 2.3(a)), preventing the description of directional bonds which are key to reproduce the secondary structure. The bonded parameters partially account for this, favouring for each residue type the backbone conformation in which it is most likely found (as computed from the Protein Data Bank - PDB [?]). When this is not sufficient, the protein can be constrained around a given structure through an elastic network model approach (ElNeDyn [?]). However, both the backbone parametrisation and the use of ElNeDyn imply that conformational changes in the structure are penalised and therefore not well sampled in MARTINI simulations.

The MARTINI force field provides two water models. The standard one

groups four water molecules in one bead only (Figure 2.3(c), left), loosing the polarisability typical of water molecules, the effect of which is partially restored with the use of a high dielectric constant. The polarisable water model [?] maps instead four water molecule to a three-beads molecule (Figure 2.3(c), right) with a positive and negative charge on two of them, which can account for the water dipole. This model allows to revert the dielectric constant back to a value closer to 1.

Overall, the MARTINI force field pushes the limits of simplification to enhance the simulations speed-up, with considerable gain in efficiency with respect to atomistic or even SIRAH simulations. Despite it can not capture some fine details of the systems studied, it has been successfully applied to describe the behaviour of many biological membranes [? ?], lipid self-assembly [?] and peptide-membrane binding [?]. The (re)introduction of a more detailed water model allowed the description of electroporation processes and the translocation of ions through bilayers [?].

2.2.5 Backmapping techniques

Coarse-grain descriptions are very effective in reproducing long time scales; however, to retrieve finer details after such extensive exploration, backmapping techniques have been designed to obtain atomistic configurations from the coarse-grain ones [?]. These backmapped structures can in turn be simulated at the atomistic level to explore the short time scale movements around such interesting conformation. The easy conversion between the two resolution, gave rise to many multiscale studies applied to biomolecular systems [?].

2.3 Beyond the classical framework

Without entering into the details, it is important to mention that the classical MD framework can benefit of additional terms aimed at improving its accuracy, and/or of specific techniques aimed at improving the sampling within accessible computer time.

Polarisability and quantum effects The first refinement possible is the introduction of polarisability, i.e. the displacement of electrons with respect to the nucleus, as a consequence of the surrounding electrostatic environment.

None of the force fields mentioned above accounts for that, because electrons and nucleus of an atom are modelled as a unique object. Specific force fields have been modelled to include this effect, either on top of atomistic descriptions, as in AMOEBA [? ?], Drude polarisable CHARMM [?] or AMBERff02 [?], or in combinations with a coarse-grain descriptions, as in the ELBA force field [?]. Polarisability does improve the accuracy of simulations (see Chapter 4, where it is discussed in the context of lipid tails parametrisation), but it can significantly slow down simulations. Moreover, for biological processes governed by quantum mechanics - such as photosynthesis, DNA mutation processes or some enzymatic activities - many semi-classical hybrid techniques have been developed [?]. They combine computational quantum mechanical modelling methods, such as Density Functional Theory (DFT) or Hartree-Fock computations (HF) [?], with classical Molecular Dynamics to gain the accuracy of a quantum description in the region of interest and the speed up of a classical one in the surrounding areas.

Reduction of the number of degrees of freedom At the other end of the spectrum, tackling instead efficiency issues, many models have been implemented to reduce the number of degrees of freedom to deal with, among which the coarse-grain force fields already mentioned. Another approach is constituted by the use of implicit solvent, where water is represented as a continuous medium, as opposed to explicit models which include all its particles [?]. Models of implicit solvent can be based on different assumptions: for example the solute-solvent interactions can be taken as proportional to the solvent accessible surface area (SASA) of every particle of solute [? ? ? ?], or instead can be derived from a solution of the Poisson-Boltzmann equation governing the charge density in a material, for example in the form of the Generalised Born equation [?] which is valid under particularly simple conditions. Another technique to alleviate the computational burden is constituted by hybrid particle-field algorithms. The idea is to treat non-bonded interactions through a mean field approach, where atoms/beads move in the field generated by the others. The field does not need to be updated at every time step, as it is a collective and thus slowly evolving variable; moreover, for each particle only the interaction with the field, and not with all the neighbouring particles needs to be computed, reducing the computations effort further. This

approach has been employed with a coarse-grain description of polymers and biological molecules in the OCCAM software [?].

Enhanced MD techniques Finally, computational strategies have been devised to bias the natural evolution of the system in order to enhance the sampling of many configurations, i.e. to speed up its pace. This is particularly necessary when studying large-scale conformational transitions. This scenario corresponds to transitions between states separated by an energy barrier, but also matches cases in which we want to describe more accurately the phase space of a rugged landscape. Indeed very often, in the impossibility of understanding from first principles which states of a system are the most energetically favoured, and can therefore influence its macroscopic behaviour, MD simulations are run to explore the broadest possible set of them. The interpretation of such sampled states may become difficult and requires sophisticated analysis techniques.

To be noticed that, as biomolecular systems have a complex landscape dense of energy barriers, a realistic, limited time simulation usually samples regions around the initial configuration only. In the simulations of proteins, this would often correspond to a structure derived from X-ray crystallography, which might not represent the native state of the protein in solution nor the functional form of interest. An alternative could be the use of different pre-modelled initial structures so to reduce the sampling of states far from the equilibrium.

The challenges mentioned above have promoted the development of enhanced MD techniques. As a non comprehensive list, we mention replica-exchange algorithms [?] which combine together multiple simulations held at different conditions, local potential-energy elevation (or metadynamics) [? ?] which avoids the re-sampling of already visited conformations adding an energy penalty to them, umbrella sampling [?] which reconstructs free-energy barriers from simulations held at specific values of the coordinate along which the barrier exists, or finally simply the use of higher temperature to overcome energy barriers [?]. To be noticed that a coarse-grain description of a system, by reducing the number of degrees of freedom, discards the high-frequency or less interesting ones, giving a smoother energy surface, so that the search is speed up both by the reduced computational load and by the fact that the

system is not trapped into local minima due to the landscape roughness.

2.4 MD and experiments: ensembles versus averages

The validation of MD simulations is performed by comparison with experiments: the properties obtained experimentally are computed from the MD trajectory as well, and the latter compared with the former. If these are correctly reproduced, it is usually assumed that the simulation is sampling the correct ensemble of states. This holds if the properties of the simulation are not drifting away, namely the system has reached equilibration and it is thus in a stationary state. Once the simulation has been validated, one can identify, from the conformations in the trajectory, the details of the processes responsible for the experimental outcome of interest, as such information is not accessible by the experiment itself.

The comparison however is not always easy, because of a fundamentally different focus that experiments and simulations have. The former measures often an average quantity in time and/or space, while simulations access the (hopefully complete) *ensemble* of states that the system can visits and their occupancy, which in turn determines the macroscopic behaviour measured in the experiments.

In the case of a protein in solution, the *ensemble* corresponds to the variety of shapes it adopts, which can differ significantly from the crystal-structures available. This flexibility is confirmed by small angle X-ray scattering (SAXS) and nuclear magnetic resonance (NMR) experiments, the results of which cannot be explained by a single-conformation scenario [? ? ?]. In this context, only simulations can help in deconvoluting the results to map them back to the conformations and their relative contribution responsible for the outcome, uncovering their relative importance.

However, it is still challenging to compare experimental and simulations derived ensembles. Following from the example above, many different combinations of structures can match a SAXS profile, so in each specific case it is important to understand which are the relevant properties playing a role in the measured ensemble before attempting comparisons.

In general, in the validation of MD outcomes, it is necessary to have a critical attitude both in the case of agreement and disagreement with the ex-

periment, and to interpret the result within the validity of the approximations performed by both of them [?]. Indeed, agreement may arise from either a simulation that reflects correctly the experimental system; but also when the property examined is insensitive to the details of the simulated trajectory, or it can be a result of compensation of errors, which is more likely to occur for systems with a high number of degrees of freedom (as biomolecular ones). Similarly, disagreement may hint at an error in the simulation (either in the model, the implementation, or simply the estimated simulation's convergence) or an error in the interpretation and/or conditions of experimental set-up (either in the result itself or its interpretation), so that both must be carefully checked to improve a convergence in the agreement. In this, some apparently negative results may suggest or stimulate new experimental settings to validate the hypotheses one was set to test [? ?].

Moreover, simulations still suffer from the limited computational time accessible to effectively simulate the system in study: most of the times, the experimental system is simply too large to be reproduced and the time scale of the process too long to be accurately sampled. Simulations are thus confined to explore a restricted space, implying that the initial conditions must be chosen carefully to optimise the search and avoid any bias which might persist for the whole length of the simulation. The use of enhanced MD techniques does increase the chances of sampling relevant states, however it introduces a bias which must be removed or properly accounted for in the interpretation of the results [? ? ? ? ?]. Finally, one should keep in mind that the force fields used are far from optimal, partly because they rely on approximate functional forms, and partly because it is difficult to find experimental observables measured with the desired resolution able to discriminate between sets of parameters.

Despite the challenges outlined, Molecular Dynamics simulations have played a crucial role in disclosing important details behind biophysical processes and in unravelling molecular details not accessible to experiments. The following section will highlight a few of the many successes of MD simulations.

2.5 MD simulations: successes

Consistently with the focus of this thesis, we will privilege examples of simulations investigating antimicrobial peptides as well as self-assembling ones, showing how computational techniques can help the design of novel molecules with improved specific characteristics.

2.5.1 Simulations of antimicrobial peptides

MD simulations of antimicrobial peptides are quite well documented since the first developments of the technique. Such peptides are a suitable system for a computational investigation as, in most of the cases, their mechanisms of action are not completely understood from the experimental information available (see Section 1.2.2). As experiments prove that even the mutation of one single residue in short AMPs can change remarkably the antimicrobial activity of the sequence (see Section 1.2.4), it is then clear that their action is governed by subtle atomic interactions, so that MD simulations, with their atomistic resolution, can help in understanding this aspect.

Systems As mentioned in the previous chapter, it has been proposed that most AMPs act through a process of attraction to the bacterial membrane, possible aggregation with other copies of the same sequence, insertion, and membrane lysis. The time scales of the overall process are accessible if using coarse-grain techniques, but not - or rarely - atomistic ones. For this level of description instead, the different steps are usually investigated separately, based on prior hypotheses: for example, the peptide can be positioned close to the membrane surface with an orientation known to promote binding (from experiments or based on energetic assumptions) [?], or again can be placed directly within the membrane core with different insertion depths, tilt angles and oligomerisation states to verify which configurations are the most disruptive ones [?]. In this case, the full insertion process can only be reconstructed from a “stepwise” knowledge combining the different states sampled and further exploring the intermediate regions if necessary. For these reasons, the choice of the conformation to simulate, i.e. the initial conditions in terms of the mutual position of peptide and membrane, is crucial, as it likely biases the simulation towards the sampling of a particular subset of configurations, and

this must be considered in the interpretation of the results. Recent advances are making possible the simulations of the full process even at the atomistic resolution for simple enough systems [? ?], as it will be shown in the following, nevertheless the “stepwise” approach is still common and the preferred one in case of complex AMP systems.

Model membranes The second important choice in the setup of a simulations of antimicrobial peptides concerns the model of the membrane to simulate. In an effort to keep complexity low, bacterial and mammal membranes can be modelled with a minimal number of lipids. Very often, models of bacterial membrane retain as only key characteristic an overall negative charge, with about 25% of the lipids being anionic (-1e charge) and the rest being zwitterionic, i.e. neutral but with positively and negatively charged regions separate in space (see Chapter 4) [? ? ? ?]. For a model mammal membrane instead, only zwitterionic lipids are employed, with the occasional inclusion of cholesterol, as it is deemed important in describing more realistically their behaviour. [? ? ? ? ?]. Because of their simplicity, very similar or identical systems are used also in experiments [? ? ?], making possible a direct comparison with simulations. Therefore, even if these simple membranes don’t model accurately the structure of the cellular envelope, simulations and experiments of these systems can provide a first explanation of some steps of the antimicrobial activity, with the two techniques complementing and validating each other.

For example, an *in silico* experiment simulated a dermicidin channel inserted into patches of phospholipids membranes with variable cholesterol content [?]. Nine membrane compositions were tested overall resulting in different membrane thickness, thus in a different orientation of the dermicidin channels inserted into it, with consequent variance in the conductance of the channel itself. This structure-function relationship shows the importance of an accurate membrane model to fully capture all the aspects of transmembrane protein activity as every change influences it. Notably, the simulations were performed with the coarse-grain model MARTINI, showing that a supra atomistic view retains enough details to investigate such systems.

Nevertheless, attempts to model more accurately cell membranes have been pursued. This can be performed at the atomistic level [?] but the task is

especially suited for a coarse-grain description, as the inclusion of all the elements of the cell membranes results in quite large systems for which atomistic computations started only in recent years to be affordable. Accordingly, coarse-grain (MARTINI) simulations have been incorporating more and more components into model membranes, describing the bacterial inner membrane, the bacterial wall, and finally the combination of the two [?] (see Figure 2.1, bottom). These large scale, coarse-grain simulations provide information on the mechanic characteristics of the system: for example, simulation of the outer membranes of Gram negative bacteria combined with the peptidoglycan layer (which, in bacteria, is positioned between the two membranes) elucidated how the distance between the two is variable, thanks to the presence of Braun's lipoproteins [?] which act as a bridge between them, and can bring them closer by bending and tilting. On the other hand, the permeability of membranes to ions and small compounds needs to be assessed at the atomistic level, and to access informative simulation time scales, smaller and simpler systems must be chosen for the task (e.g. the inner membrane only), often together with enhanced MD techniques such as metadynamics, umbrella sampling, and replica exchange umbrella sampling. [? ? ? ?].

Force field comparison Finally, simulations of the peptide interaction with a model membrane are clearly determined by the parametrisation of the force field employed for protein and lipids (and by their mutual consistency). There are multiple evidence suggesting that different force fields produce very different outcomes when simulating the same system, under the same conditions. This is also valid for simulations of pure lipid patches (see Chapter 4), resulting in incompatible values of area per lipid, organisation of the tails and energetic profiles across the membrane, and thus has an impact in the simulations of AMPs interacting with a membrane.

For example, Wang et al. [?] run simulations of the antimicrobial peptide melittin with different force fields, namely CHARMM27 and 36 (for protein and lipids respectively) [? ?], OPLS all atoms (for protein) and united atoms (for lipids) [?] and GROMOS 53a6 [?] (and the TIP3P water model for all of them [?]). Despite these parametrisation have similar values of partial charges on the different atoms, and similar bonded interactions at the protein level, the unfolding of melittin in the membrane was significantly different

among them, with the CHARMM force fields suggesting an almost completely folded state for melittin bound to a membrane, in line with the NMR available results, while the other force fields promoted partial unfolding. Most likely this can be attributed to the fact that lipid and protein parametrisations are obtained separately and might present some inconsistencies. The most evident example is the OPLS case, for which an all atom description of lipids is not available and a mixed description has been adopted. In the case of GROMOS, both components are parametrised at the united atom level, but their mutual consistency might be questioned as well, as fully explained in Chapter 4. In general, a united atom description is clearly less accurate than an all atom one, and in the case of lipids it has been postulated that it is not able to represent faithfully the dynamical processes happening in the hydrophobic tail region [?].

A similar investigation has been proposed by Bennett et al. [?], proving that the propensity of the synthetic AMP CM15 to form pores strongly depend on the force field used but also on some extent - at least at the time of the work - on the MD engine used (GROMACS compared to NAMD [?]). This should not come as a surprise because the membrane characteristics emerge from the collective behaviour of lipids, so that a small difference in the way their interactions are treated might be amplified resulting in different macroscopic outcome for the simulations [?].

A more systematic study on the topic has been performed by Sandoval-Perez et al. [?], focussing on the reproduction of membrane-protein interactions in different force fields (GROMOS 54a7 [?], CHARMM36, Amber14SB/Slipids [?] and Amber14SB/Lipid14 [?]). All of them were able to reproduce the overall positioning of transmembrane proteins in the case studies tested, together with their α -helical and β -sheet content, while they showed discrepancies in the insertion angle for a short helical peptide (sybII) spanning the membrane. The amino acid side chains insertion depth was also tested: the CHARMM force field suggested a deeper insertion for hydrophobic amino acids but the other parametrisations gave not so clear a distinction. In general, with the GROMOS force field, a higher energy is required to insert amino acids to the bilayer centre. Interestingly, all parametrisations gave a very broad minimum for the insertion of Tryptophan, comprising the phosphate region of the phospholipid membrane tested, but also part of the tail

and head regions. This is in line with the different interpretations given on the role of this amino acid for the action of antimicrobial peptides: either as an anchoring point positioned deeper in the hydrophobic region, or as a partner for hydrogen bonding with the hydrophilic heads. The subtle differences between parametrisations lead to the conclusion that for every particular system tested, the comparison with at least one experimentally measured quantity would be the only way to assess the simulation performance accurately.

Simulations of membrane-peptide interaction: examples Even in the context of a simplified model scenario and with the caveats coming from the chosen parametrisation, simulations of antibacterial peptides on a membrane have been successful in elucidating some of their mechanisms. The first important contribution consists in the introduction of the disordered toroidal pore concept: as explained in the previous chapter (Section 1.2.2), the models of membrane poration due to AMPs consist often in ordered structures (see Figure 1.3) where many peptides gather together to contour a pore, and they are either in contact with the hydrophobic tails of the lipids (barrel-stave model), or with their head, as lipid molecules bend around the pore to keep their tails screened from the outside environment (toroidal pore model). However, simulations of the short helical peptide magainin MG-H2 [?], among others, showed that a single copy of the helix, inserted at an angle with the normal to the membrane plane, was sufficient to displace the lipids around in a non organised manner and form a water-filled pore (Figure 2.4).

Regarding possible rearrangements of the antimicrobial peptide structure when interacting with a membrane, simulations of cathelicidin LL-37 on pure POPG (anionic) and POPC (zwitterionic) lipid patches showed that LL-37 has a propensity to bind to the former, as expected due to the opposite charge that membrane has with respect to the cationic peptide [?]. However the simulations highlighted also that, in contact with POPC, the helical secondary structure was lost, while the interaction with POPG preserved it, suggesting that the spatial arrangement of the residues, and not only the overall chemical character, is important for their action. Such type of information is hardly available to experiments or through a theoretical reasoning.

Further insights into the role of the secondary structure were obtained simulating the helical antimicrobial peptide CM15 nearby a POPC membrane,

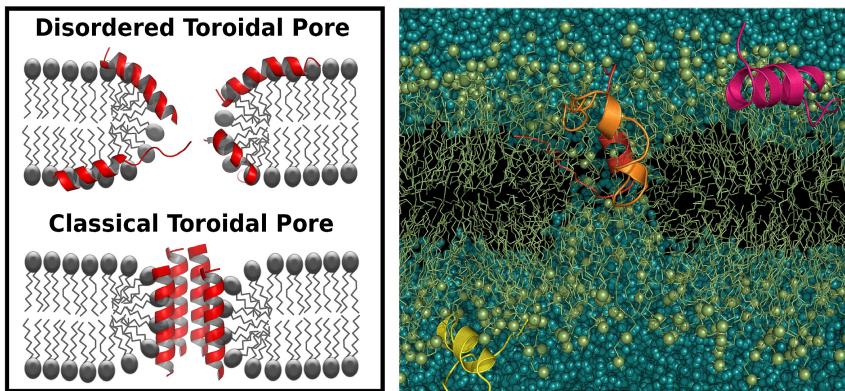


Figure 2.4: (Left) A cartoon image comparing the disordered toroidal pore state (lack of a well defined peptide orientation) to the traditional view. (Right) A snapshot of the disordered toroidal pore from simulations of melittin in DPPC. Peptides in cartoon, lipids in lines, water and ions in bead representation. Reproduced from [?].

starting from a fully structured helix or from a coil configuration: Wang et al. [?] proved that the interaction with the lipids is stronger when the peptide approaches the membrane in its disordered form rather than in a fully formed helix. This happens because of the larger flexibility of the coil arrangement which allows for more residues to come in contact with the membrane at once. Notable, the α -helical fold binds as well to the membrane, but only after a time span larger than 100 ns, which would have been inaccessible up to a few years ago, potentially leading to wrong conclusions.

The improvements in computational resources is slowly removing some of these obstacles, pushing the extent of simulation time to the microsecond timescale. In a recent example, the translocation of the helical PGLa peptide through the membrane has been observed as a rare event, dependent on the concentration of the peptide, on the multi microsecond timescale without the formation of an organised pore [?]. The *in silico* experiment still benefited of an enhanced sampling in the form of a higher temperature used for the simulations, but no pre-insertion of the peptide was performed. This study shed light on a possible mechanism of permeabilisation which is usually overlooked in favour of processes involving organised channels and pores. The fact that no organised neither disorganised pore is observed matches the experimental results which can not identify such structures for the peptide considered.

Similarly, simulations were able to shed light on the mechanism of translocation of Arginine-rich peptides, proposing a mechanism of action on an oth-

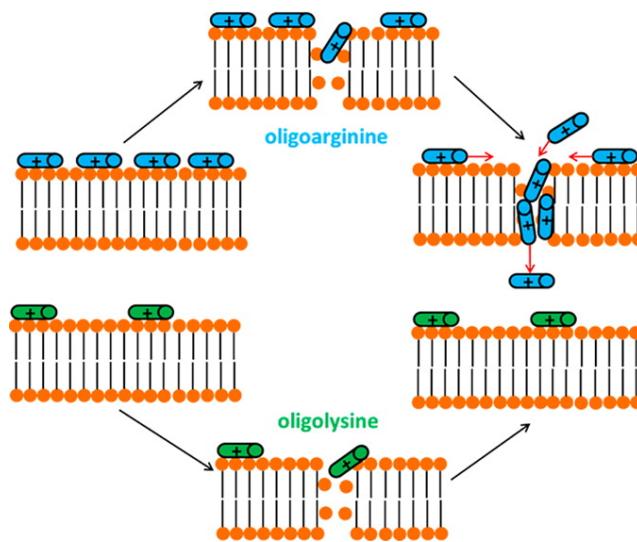


Figure 2.5: Cartoon representation of the different assembly mechanism that allow polyarginine but not polylysine to translocate through spontaneously formed pores. Reproduced from [?].

erwise puzzling problem [?]. These sequences have high positive charge, but despite this, possess a high propensity to penetrate membranes, overcoming the hydrophobic region represented by the lipid tails. Very similar peptides where the Arginines were swapped with Lysins showed no significant penetration. A commonly used explanation considers polyarginine translocation a quasi-equilibrium process, but this does not explain the selectivity against Lysins rich peptides. After extensive simulations of the two systems (multiple, hundreds of nanosecond long runs, with two different force fields), the proposed mechanism involves the spontaneous formation of thermal pores: in some of these rare events, the transient pore would be occupied by a peptide (a precursor), which slows down its dynamics and thus closure. In such situation, the translocation of other copies of the peptide is highly favoured if their concentration is sufficiently high. Indeed, other copies of the peptide are driven to aggregate with the precursor inside the membrane and are then pushed toward the opposite side as there is a lower charge density in that region. Differently from polyarginines, polylysins have a much lower aggregation propensity, so that the presence of a precursor peptide inside the membrane does not induce an enhanced insertion of further peptides.

The last two examples mentioned bring the attention on whether and for which peptides oligomerisation is necessary for an efficient antimicrobial ac-

tivity. MD simulations can offer insights on this aspect as well. Contrary to oligomerisation in solution, which can happen on shorter time scales, the spontaneous aggregation of peptides on a membrane surface requires a long time, as the structures must diffuse on the membrane to meet each other, and many other competing processes (such as insertion) are happening at the same time.

A recent example of how MD elucidated oligomerisation mechanisms comes from simulations of maculatin (an helical AMP), which showed that the pores it forms can include a variable number of helices and thus assume many different conformations [?]. The suggested process of pore formation proceeds via insertion of a single residue, closely followed by other ones which are able to penetrate the membrane thanks to the lipid defects already created by the first peptide.

Similar investigations can be carried on also for other cell penetrating peptides, which are not antimicrobial: as such, some of them aim at inserting within cells without necessarily causing poration. One example is constituted by the influenza fusion peptides, which have been extensively studied with a simulation set up similar to the one mentioned for AMPs: a few copies of the peptide were positioned on a model membrane and their oligomerisation and insertion processes were followed in time, showing the formation of aggregates of different sizes which perturbed locally the membrane [? ?].

To be noticed that, when investigating oligomerisation, the size of the system must necessarily be increased to include all the copies necessary to form the aggregates observed experimentally. As such, with the present accessible computer time, not all the systems can be investigated from unbiased initial conditions. In the case of protegrin, a β -hairpin antimicrobial peptide which has been long though to act through the formation of transmembrane β -barrels, many variables can influence the outcome of the unknown final structure. Even with the computer power available now, it is unlikely to sample all the possible conformations resulting in stable or transient β -barrel pores in simulations starting from a few peptides scattered on the surface, and this hinders the understanding of their relative importance. To overcome such problems, a semi-systematic investigation has been carried on by Lipkin et al. [?], simulating different assembly (see Figure 2.1, top). Microsecond long simulations discriminated which ones of these initial configuration formed stable pores for

the whole length of the simulations, and the ones which were disrupted. As in the previous example, several different possibilities were found stable in solution, suggesting that single AMPs might have multiple mechanisms of insertion into membranes.

Most of the examples above employ atomistic descriptions of the system. Similar investigations have been carried on also using the MARTINI force field, indeed the coarse-grain description does capture the pore-forming behaviour of some AMPs. As an example, simulations of maculatin and aurein on POPC membranes showed different propensities for pore formation versus aggregation, showing that the model retains enough details and chemical information to reproduce different membrane perturbing behaviours [?]. Nevertheless, the developers of the MARTINI model themselves pointed out how some aspects of pore formation might not be captured in a satisfactory way [?], for example the penetration of water can be misrepresented as can be intuitively expected from a model which clusters four water molecules together (indeed some pore conformations allow for the passage of fewer water molecule if not one at the time).

In general, the outlook of simulations of antimicrobial peptides interacting with membranes goes in the direction of reproducing longer time scales thanks to the enhanced computational power available, trying to match the experimental findings showing that many antimicrobial related processes happen at the microsecond scale or beyond. This enhanced power would also reduce the need to use biased initial conditions or higher temperatures to speed up the simulations. Moreover, gathering the contribution of the whole community, simulations will likely go in the direction of modelling more accurately the bacterial membrane, and while this is already at an advanced stage for coarse-grain simulations, it is still an ongoing process for atomistic ones. Finally, the force field issue must be solved in collaboration with experimentalists, finding new tests and experimental quantities to compare the computational outcome with and make the different parameters sets converge toward a similar description of the phenomena observed, which is consistent with the experimental results.

Simulation-aided AMPs design The role of simulations in aiding AMPs design has been briefly sketched in Section 1.2.4. As pointed out, MD simulations are hardly a tool to analyse large dataset, therefore a systematic analysis

can be performed for very small systems only, or, alternatively, the investigation can focus on a few selected sequences.

As already mentioned, when classifying AMPs, simulations can be helpful in integrating structural information which is otherwise lacking, when no crystal structure of the peptide is available. Such approach was followed by Liu et al. [?] to complement the chemical information available on a dataset of short AMPs, and the overall information was used to feed a predictor of AM activity of novel sequences. Preliminary results showed that such structural information of minimal AM sequences improved significantly the ability of the predictor to discriminate whether a new sequence was suitable for antimicrobial activity or not.

Another commonly followed approach consists in using simulations to elucidate the reasons why a particular mutation is important and effective in terms of increased activity or decreased toxicity. Indeed, for short sequences, such mutation screenings can be afforded experimentally, thus there is little need to predict whether they would be beneficial. Rather, once assessed they are, it is interesting to understand why: for example, simulations of ovispirin and a mutant peptide with reduced toxicity showed that the bend in the helix in the latter was responsible for mitigating the interaction with mammal membranes and thus reducing haemolysis [?]. Again, the mechanism of lipid disordering and insertion by indolicidin was assessed through MD, and the amino acids responsible for each of them separately were identified, so that mutants could be designed with either reduced toxicity or enhanced potency [?]. Finally, temporin and a derived sequence were investigate to discover that the mutant improved activity derived from a reduced aggregation propensity of the peptide in water, so that more copies were ready to bind to the membrane and thus disrupt it [?].

Many more examples can be listed, each with a slight different focus: the protocol of integrating simulations and design is usually customised according to the system in exam, as the field has not reached yet a systematic organisation. However, it is clear that simulations used in conjunction with experimental testing can be used to optimise already available AMP sequences, and thus contribute to device design rules for the creation of synthetic sequences with tailored properties.

2.5.2 Simulations of self-assembling peptides

Self-assembling peptides are another fascinating and challenging topic that MD simulations can help investigating. Simulating such systems implies different challenges with respect to the ones faced when simulating AMPs on membranes [? ?].

In theory, the set up of the system is quite straightforward: only the solvent characteristics and optionally the experimental salt concentration need to be matched, then a random initial configuration of the molecules - in the desired concentration - would allow the simulation of the process of interest. In reality, reproducing the experimental conditions often implies working with very large systems: with the level of dilution of the solutions employed in the experiments, a considerable volume needs to be simulated to host enough copies of the peptide to observe the assembly of large enough oligomers. However with this approach the time scale useful to witness a spontaneous assembly would greatly exceed the computational time available. For that reason, two main strategies have been adopted: coarse-grain simulations and the use of pre-assembled structures. Other routes include the choice of an implicit solvent model, or the use of other techniques such as Monte Carlo (a probabilistic exploration of the phase space rather than a dynamical algorithm) which can sometimes be less time consuming. Here, as we focus on MD simulations, we give a few examples of the first two strategies mentioned, which can be adopted in this framework.

Many studies have been performed with the coarse-grain MARTINI force field to witness assembly of surfactants [?], polymers [? ?] and lipids [? ?], and a few focussed on peptides as well [? ?]. For example, the assembly in water of peptide amphiphiles (PAs) into cylindrical fibers has been simulated at the coarse-grain level [?], showing a transition from small micelles to longer fibres (Figure 2.6). This example of minimal PA structures is particularly interesting for the study of AMPs as well, as it shares with them the amphiphatic character, so that having a general knowledge on how similar sequences assemble together would help in tuning their aggregation properties in water prior to the delivery to the membrane. In the work mentioned, pre-assembled fibres have been simulated as well at the atomistic level, to confirm their stability in solution.

The second approach consists in preparing the system in a pre assemble

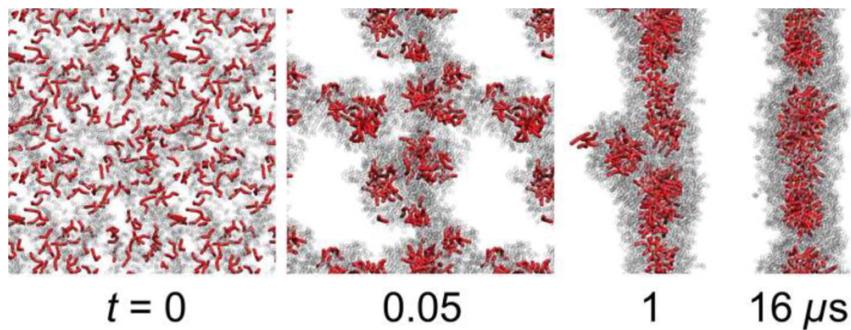


Figure 2.6: Process of peptide amphiphiles (PA) fiber formation assessed through MARTINI simulations. Hydrophobic tails in red, peptides in gray [?].

state that is somewhat suggested by experimental evidences and using MD simulations to verify whether the conformation is kept or it is disrupted, and which one out of many is the one most energetically favoured. It has been widely employed in cases where the final assembly was hypothesised to have a high degree of order, achievable only with a long sampling. This approach has been used to prove that a branched peptide can self-assemble in bilayers first, and then that a larger hypothetical structure assembled in the shape of a capsule was stable in the run time of the coarse-grain simulations employed. Specifically, the capsule has been build to match the peptide density on the self-assembled double layer and to respect the constraints derived from the presence of the curvature [?].

Similar approaches have been crucial in elucidating the assembly process of viral capsids: capsids are very large systems and the assembly of their protein subunits is mediated by energy barriers. For such reasons, already pre-assembled systems have been simulated to understand the interaction between the components and thus the first mechanisms of the assembly. This has been done recurring to ultra coarse-grain or elastic network models [?] first, and only in the most recent advances to atomistic simulations [? ?]. Additionally, smaller portions of a capsid can be simulated, to obtain a minimal information on the cohesion of its blocks [?].

The examples above show how Molecular Dynamics simulations have been employed for the investigation of many different systems during the years, adapting the resolution, set up and the techniques related to better query the systems of interest. Such overview suggests then that simulations would be a

suitable tool to investigate the system of interest of this work, namely the self-assembling antimicrobial peptide capzip. The two aspects of its behaviour will be studied separately, adopting the necessary approximations and strategies to make the simulations efficient and to query the related questions at each time.

The details of the systems simulated and the specific parameters used for each of the system simulated in the following chapters can be found in the relative sections, together with an extensive explanation of the motivation of the choices made.

Chapter 3

Capzip simulations

In Chapter 1 we introduced the molecule capzip and its properties, highlighting the unknowns of its mechanisms of action. In Chapter 2 we presented a review on Molecular Dynamics simulations, proving their successes in elucidating the behaviour of self-assembling and antimicrobial peptides in the past. Now, we employ this technique to understand better our system of interest. Given the exoticism of the unit, and the little atomistic information at disposal, modelling such peptide must proceed in a stepwise manner.

The first aim is to elucidate which structures it forms in solution and what interactions are keeping the molecules together. To understand the latter ones it is important to retain the highest level of detail possible, and for this reason we resorted to atomistic simulations first. This description has an high computational cost though, preventing the simulation of very large systems for a very long time, as it would be required to reproduce the natural assembly from a dispersed solution. Thus, we simulated increasingly complex pre-assembled blocks, verifying each time their behaviour in solution and inferring whether they are suitable to form a stable supramolecular structure. This approach, fully explained in Section 3.1, lead to the model of a minimal capsule, which has been subsequently investigated at coarser levels to explore its behaviour on longer time scales.

Of the pre-built structures a few selected ones were simulated in contact with model membranes to understand the determinants of their antimicrobial activity. Details on these simulations and the specific techniques employed to enhance the sampling are given in Section 3.2.

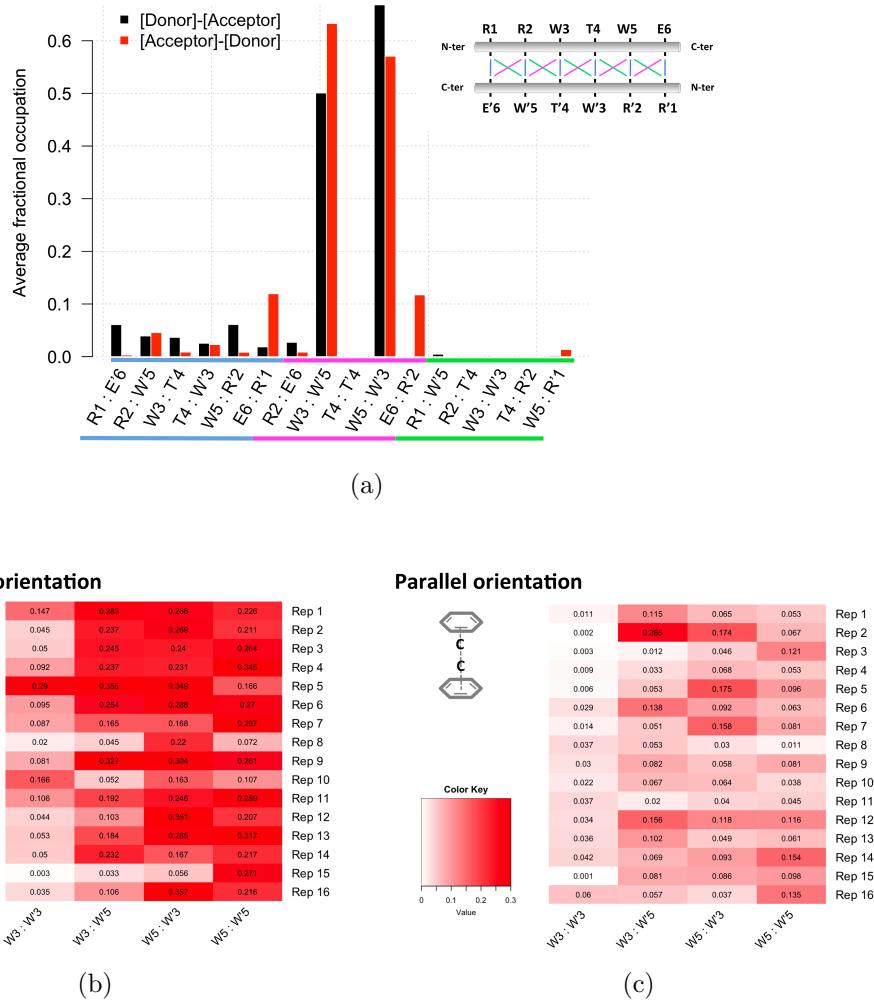


Figure 3.1: (a) Presence of backbone hydrogen bonds between amino acids in two facing antiparallel RRWTWE chains. The top right inset shows a scheme of the initial configuration. All the pairs highlighted in blue, green and pink are reported in the histogram; the same color code appears in the bar labels. Occupancy is averaged over 16 simulations of 20 ns. (b, c) For each replica and possible pair of facing Tryptophan residues in the β -sheet, the map gives the fraction of time for which a parallel or perpendicular π -stacking interaction has been observed. The second and fourth column correspond arrangement with the most populated hydrogen bonds.

3.1 Modelling the assembly

As previously mentioned, the antimicrobial sequence of capzip is designed with opposite charges at its extremes to favour an antiparallel β -sheets pairing with other copies of itself. MD simulations of two RRWTWE sequences paired in

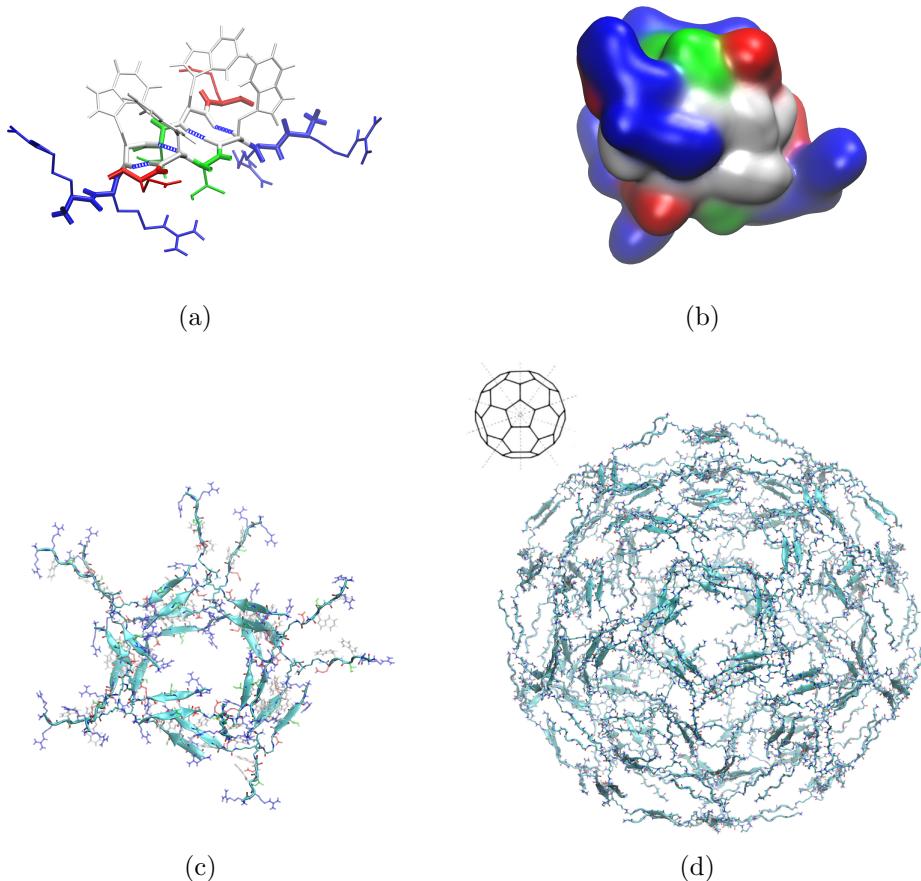


Figure 3.2: (a) Detail of β -sheet pairing with facing Tryptophan residues forming hydrogen bonds between their backbone atoms (bonds representation coloured by residue type, and hydrogen bond representation). (b) Two stacking β -sheets in surface representation, coloured by residue type. In white the partially buried hydrophobic patch. (c) A pentagonal subunit: ten antimicrobial molecules arranged in two stacking pentagons. Chains are paired in antiparallel β -sheets within each pentagon, and the two are interfacing with their Tryptophan residues in contact. (d) Atomistic structure of the buckyball simulated (bonds and cartoon representation) and geometrical model for comparison.

this fashion confirm that the assembly is stabilised by opposite charge interactions (with statistics gather over 16 replicas, each run for 20 ns). Moreover, backbone hydrogen bonds form between Tryptophan residues of facing strands, after a rearrangement of the mutual position of the backbones (Figure 3.1(a)). Finally, π -stacking contributes to the interaction as well, albeit in minor measure (Figure 3.1(b), (c)).

The favourable hydrophobic interactions between Tryptophan residues re-

sult in the creation of a hydrophobic patch which includes four of them (in white in Figure 3.2(a) on one side of the β -sheet plane. This creates an amphiphilic structure where the hydrophobic core is segregated from the remaining charged residues distributed at the other positions. The combination of two stacking β -sheets, paired to match their hydrophobic patches, constitutes an effective supramolecular assemblies to reduce solvent exposure of such residues.

This pairing strategy, however, needs to be applied in the context of full molecules assembly. The quasi three-fold symmetry of capzip suggests a regular geometric arrangement. The best examples of organised protein structures can be found in viral capsids, which are composed by the regular repetition of highly symmetric protein subunits. Inspired by this, we tested whether a geometrical organisation can represent a stable capsule, choosing as representative geometry a truncated icosahedron (buckyball).

Preliminary atomistic simulations (100 ns) were run on a pentagonal subunit formed by ten molecules arranged in two stacking pentagons (as in Figure 3.2(c)), proving the cohesion between molecules belonging to the subunit. Specifically, the number of contacts between backbone C_α s did not decrease in time but augmented slightly at the beginning, due to the compaction of the unpaired arms toward the core of the structure (Figure 3.3(a)). Moreover, for each pair of facing chains, it is computed the distance between their centres of mass. Figure 3.3(b) reports the variance of this distance over its average value, as a measure of the cohesion of the subunit, showing that in the majority of the cases less than 2% of variability is observed.

The pentagonal subunit respects the building principles of a) β -sheet pairing between antimicrobial sequences and b) double layer structure to screen the hydrophobic patches, and will constitute a face of the icosahedron. Each capzip molecule is centred in one vertex of the polygon, with the branches laying alongside the edges departing from it. On each edge two branches coming from opposite sides meet in an antiparallel fashion. When possible, a β -sheets with paired Tryptophan residues is organised (as in Figure 3.2(a)). The stacking pentagons interact through the hydrophobic patches of their β -sheet which are arranged in stacked positions.

The full truncated icosahedron was assembled from twelve pentagonal subunits (Figure 3.2(d)). As each subunit is formed by two stacked pentagons, the resulting structure has two concentric layers, for a total of 120 molecules,

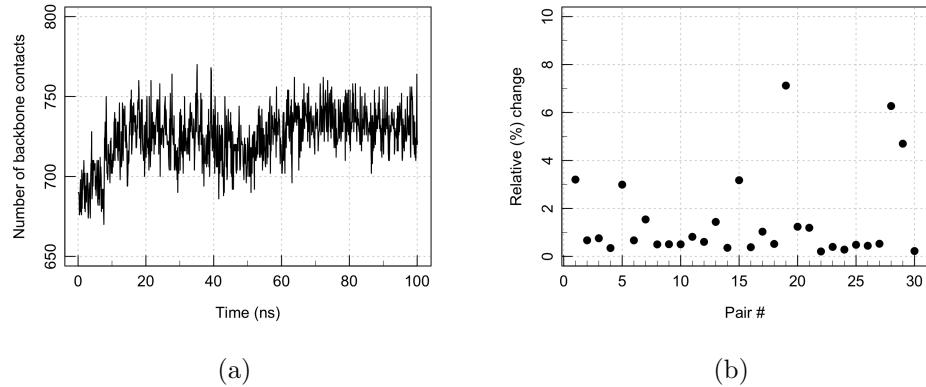


Figure 3.3: (a) Number of backbone contacts during a simulation of a pentagonal subunit. (b) Variability of the inter chain average distance between facing chains. The 30 pairs defined as facing are the chains belonging to the same β -sheet (5 for each of the two stacking pentagons), and for each stacking β -sheet the 4 possible inter-pentagon (inter-layer) pairs of chains.

and initial radius of 7.7 nm. This geometry represents a minimal model of the possible structures of capzip assembly in solution as, given the flexibility of the molecule, different geometries are possible, though proceeding from analogous interactions between the components. The final structure was simulated at atomistic and coarse-grain levels, respectively with the GROMOS 53a6 [?], SIRAH [?] and MARTINI [? ?] force fields (with both standard and polar water [?] to compare the two models). From the final configurations of the MARTINI coarse-grain model (standard water), atomistic coordinates were obtained and simulated, to be compared with the original atomistic dynamics. Moreover, additional simulations were run at all the coarse-grain levels on a structure made of one layer only (i.e. build from pentagonal block make by one pentagon only), to prove whether the bilayer structure was more energetically favoured.

A multiscale analysis is needed also to investigate the antimicrobial activity. Being highly costing to simulate a full truncated icosahedron on a bilayer patch at the atomistic level, the pentagonal subunit employed to build the complete structure (Figure 3.2(c)) was taken as representative of the latter. It was simulated close to the membrane plane, parallel to it, to avoid spending time in sampling non bound conformations (Figure 3.4(a)). This, together with a tailored use of an applied electric field (see Section 3.2), will speed up

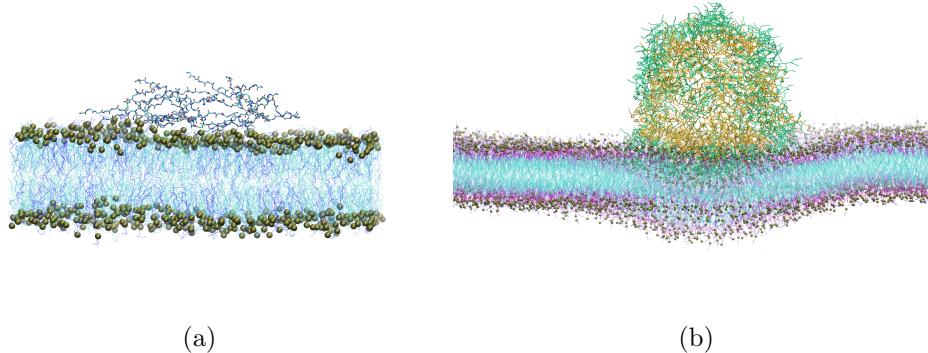


Figure 3.4: (a) Atomistic structure of a pentagonal subunit on a 740 lipid bacterial model membrane of composition DLPC:DLPG 3:1 (initial configuration). Peptide backbone in line and cartoon representation; lipid in cyan lines (DLPC) and blue ones (DLPG), all lipids phosphate in golden van der Waals beads. (b) coarse-grain (MARTINI) representation of the buckyball on a 2880 lipids bacterial model membrane (final configuration of the trajectory). Protein in bonds representation: green outer buckyball layer, yellow inner one. Lipid in line representation, coloured by bead type, and lipids phosphate in golden van der Waals beads.

simulations considerable.

To observe the natural binding of the peptide to the membrane, the process of the full buckyball approaching a model membrane was simulated with a MARTINI coarse-grain description (Figure 3.4(b)).

Two membrane patches were simulated for both resolutions, a model bacterial and a model mammalian membrane, to identify the different interactions with the peptide. The first one presents 25% of anionic lipids (DLPG), and the rest are zwitterionic (DLPC), while the second has only DLPC lipids. The choice of the bacterial model was dictated by the experiments performed on the capzip system, and the mammalian one employed the same zwitterionic lipid chosen in the bacterial to simplify the comparison.

3.2 Simulations details

Atomistic simulations All simulations were performed with the GROMACS software, version 5.5 and 2016 [? ? ?].

The atomistic coordinates for the peptidic supramolecular assemblies described in Section 3.1 were built combining GROMACS tools and the MOE

software [?]. Simulations were run with the GROMOS 53a6 force field [?]. Parameters for the central residue connected with the peptidic chains are computed with the ATB software [? ?]; the ones for the joining bonds are derived from tabulated values of analogous peptide bonds.

The systems were solvated with SPC water [?] and counter ions were added (Na^+ or Cl^-); further ions are introduced to reach the concentration of 150 mM (for the species with opposite charge with respect to the counter ions), to reproduce the experimental conditions. For simulations of a single β -sheets and of the pentagonal subunit, the systems were energy minimised with a steepest descent algorithm, then equilibrated in the NVT ensemble with decreasing positional restraints at increasing temperatures (100 K, 200 K, 250 K, 300 K and respectively 1000, 1000, 500, 250 $\text{kJ/mol}\cdot\text{nm}^2$ restraints, 100 ps each); then in the NPT ensemble, without restraints, at the same temperatures steps and for the same time. Production was followed for 100 ns.

For the truncated icosahedron structure the above equilibration was still insufficient. Due to the construction procedure, two thirds of the branches are not properly paired along the edges. Therefore, after an NVT equilibration as above, strong flat-bottom restraints ($1000 \text{ kJ/mol}\cdot\text{nm}^2$) were placed between the center of mass of imperfectly aligned branches throughout the NPT heating, to penalise their mutual separation with respect to their initial distance (100 ps runs at 100 K, 200 K and 250 K and 35 ns at 300 K). This was followed by a series of 10 ns runs at 300 K with decreasing restraints strength (750, 500 and 250 $\text{kJ/mol}\cdot\text{nm}^2$) and by a free production run (100 ns). Three different replicas were run, generated from the final configuration of the 300 K NPT run with $1000 \text{ kJ/mol}\cdot\text{nm}^2$ restraints.

Throughout all the simulations, the temperature was maintained by independently coupling the protein and the solvent (plus ions) to two external temperature baths using a velocity rescale thermostat [?] with coupling constant τ_T of 0.1 ps. The pressure was kept at 1 bar by Berendsen [?] or Parrinello-Rahman barostat [?] (for the equilibration phases and the production run respectively) using an isotropic coupling, with isothermal compressibility $4.5 \times 10^{-5} \text{ bar}^{-1}$ and coupling constant τ_P of 1 ps. Electrostatic interactions were treated using the smooth Particle Mesh Ewald (PME) algorithm [?], with a short-range cutoff of 0.9 nm. The van der Waals interactions were treated with a plain 0.9 nm cutoff. A SPC water model [?] is used. All atomistic dynamic

Capzip in solution simulations

System (Nr peptides)	Model	Time (ns)	Replicas
β -sheet (Fraction)	GR	100	32
Pentagonal subunit (10)	GR	100	1
Buckyball bilayer (120)	GR	100	3
Buckyball bilayer & monolayer (120)	SI	1000	3 & 1
Buckyball bilayer & monolayer (120)	MA	1000	3 & 2
Buckyball bilayer & monolayer (120)	MA_P	1000	2 & 1

Table 3.1: Table of simulations of capzip assembly in water. Model specifies the force field: GR = united atom GROMOS 53a6 [?], SI = coarse-grain SIRAH [?], MA = coarse-grain MARTINI [? ?], MA_P = coarse-grain MARTINI with polar water [?].

runs were performed using a 2 fs time step. An overview of simulations of peptide assembly in solution is given in Table 3.1.

The atomistic coordinates for the bacterial model membrane patch were built with the PACKMOL software [?], from pdb files of a single DLPC [?] and DLPG [?] molecule. Two patches were built, made respectively of 512 and 740 lipids, with composition DLPC:DLPG (3:1). The initial area per lipid was set to 7 nm², above the values found experimentally for either lipid species (0.608 ± 0.012 nm² for DLPC [?] and 0.656 ± 0.012 nm² for DLPG [?]). The correct area per lipid of the mixture was reached during a 400 ns equilibration, the final configuration of which was used for simulations with the peptide on the membrane. A similar procedure was held for DLPC, producing a patch of 748 lipids with an initial area per lipid of 0.656 ± 0.012 nm². This patch was used for control simulations against the 740 lipids bacterial one.

The initial configuration of atomistic simulations of the peptide on the membrane was generated from the equilibrated lipid bilayer (after 400 ns run) and the equilibrated pentagonal subunit (after 100 ns run with positional restraints on the C _{α}), placing the pentagon plane parallel to the membrane one and close to it (Figure 3.4). The inflategro [?] script was used to solve the partial overlap of the peptide side chains with the lipid molecules, removing the ones overlapping. The sizes of the two patches fit the pentagonal subunit

with respectively 3.5 nm and 5.4 nm distance between its periodic boundary images (along both x and y).

For simulations involving membranes, the version 54a7 of the GROMOS force field [?] was initially chosen, and it is thus the one selected for simulations involving the 512 lipid membrane patch, but upon further research the 54a8 [? ?] was deemed more suitable ad thus selected for the following tuns on the 740 lipids patches. Lipid parameters were taken from [?] for DLPC, while for DLPG they were built from the ones available in the literature for POPG [?].

The simulations set-up parameters are as above, except for the use of three thermal coupling groups (peptide, membrane, water plus ions), a semi-isotropic pressure coupling, and a larger cut off radius for both Coulomb and van der Waals interactions (1.2 nm). Additionally, for the 512 lipids membrane, a Reaction Field [?] was used instead of PME long range electrostatic treatment. Control simulations on membrane patches without peptide showed that the results in terms of area per lipid are compatible with the ones obtained from PME.

Each membrane patch was first equilibrated for 50 ps in NPT conditions at 50 K, then the temperature was gradually increased up to 300 K in 500 ps, and finally a 400 ns production was run. The final configuration was used as initial structure for the peptide-membrane simulations. A similar equilibration procedure was followed for peptide-membrane systems.

Additional simulations were performed applying an external electric field to the membrane, from the side hosting the peptide to the opposite one, to mimic the membrane potential and verify how the presence of the peptide affects the membrane response to external stimuli. In a first test performed on the 512 lipids bacterial patch, the field was increased by 20 mV/nm steps every 200 ns, until reaching the electroporation value, which resulted to be 130 mV/nm. For the bacterial membrane patches of both sizes, other simulations were performed starting from the unperturbed membrane configuration and the threshold field, in three replicates each. Three control runs were performed on the 512 patch without peptide, at the threshold value of the field, and one at the higher value of 140 mV/nm.

To be noticed that membrane field is estimated around 20 mV/nm across mammalian membranes and 35 mV/nm (from, respectively, a -70/-90 mV and

-130/-150 mV potential [?] and an estimate membrane thickness of 4 nm), but previous computational work explored the effects of fields up to 500 mV/nm [? ? ?], to witness poration within the simulations time, according to the resources available at the time.

An overview of simulations of peptide-membrane systems is given in Table 3.2 (control simulations on pure membrane are listed in Table ??).

SIRAH coarse-grain simulations SIRAH coarse-grain simulations were run with the SIRAH force field [?]. Peptide coordinates for the buckyball geometry were obtained from the atomistic ones using the converter distributed with force field. Parameters for the central residue were built from comparison with similar chemical moieties. All simulations were run adding Cl^- counter ions to balance the positive charges of the peptide and additional Na^+ and Cl^- ones to reach a 150 mM concentration (of the cationic species).

While for simulations of peptide assembly in solution we resorted to a multiscale procedure, comparing up to four force fields, for simulations on membrane we focussed only on atomistic and MARTINI simulations only. This has been performed in the interest of time, and following some concerns arisen in preliminary SIRAH runs of the capsule on a membrane, which showed an unstable behaviour (contraction and expansion of the membrane). Given that there are little benchmarks so far on the performances of SIRAH for lipids, and none for systems as large as the one studied here, we are continuing pursuing the investigation at an exploratory level, working in contact with the developers of the force field to help elucidating its limitations and work out the best equilibration procedure to treat large membrane patches.

For SIRAH simulations, the temperature coupling was performed with a velocity rescale thermostat [?] and coupling constant τ_T of 0.1 ps, and the pressure coupling at 1 bar pressure, with 4.5×10^{-5} bar $^{-1}$ isothermal compressibility, using a Parrinello-Rahman barostat [?] with a τ_P of 6 ps. Electrostatic interactions were treated using the PME algorithm [?], with a short-range cutoff of 1.2 nm and relative dielectric constant of 1. The van der Waals interactions are treated with a plain 1.2 nm cutoff.

After energy minimization, a 4 ns NVT equilibration was run at 300 K, followed by a 10 ns NPT run, both with positional restraints (1000 kJ/mol·nm 2) on the solute. Two 10 ns run (NPT ensemble, 300 K) were then performed

Capzip on membrane simulations					
	Peptides	Lipids	Model	Time (μ s)	Replicas
Bacterial	10	512	GR	500	2
	10	740	GR	500	1
	120	2880	MA	10000	2
	120	2880	MA_P	10000	1
Mammalian	10	740	GR	500	1
	120	2880	MA	10000	1
	120	2880	MA_P	10000	1
Electroporation simulations					
	Peptides	Lipids	Model	Time (ns)	Rep. E (mV/nm)
Bacterial	10	512	GR	$75^P, 20^P, 71^P$	3 130
	10	740	GR	$60^P, 50^P, 70^P$	3 130
	120	2880	MA_P	500	1 20
	120	2880	MA_P	168^P	1 40
Mamm.	10	740	GR	$20^P, 28^P, 39^P$	3 130
	10	2888	MA_P	500	1 20
	10	2888	MA_P	??	1 40

Table 3.2: Table of simulations of peptide-membrane complexes. All the mentioned ones run at 150 mM concentration of NaCl. Models: GR = united atom GROMOS, MA = coarse-grain MARTINI, MA_P = coarse-grain MARTINI with polar water. Superscript P denotes poration. For electroporation simulations on pure membranes, see SI Table 3.6. A number of 10 peptides denotes the pentagonal subunit, 120 the buckyball bilayer.

with backbone restraints of 1000 and 100 kJ/mol·nm², respectively. Similar to the procedure adopted for the atomistic simulations, during the latter, flat bottom positional restraints (1000 kJ/mol·nm²) were enforced on the unpaired branches during the NPT. Three additional 10 ns equilibrations were run at 300 K, with no backbone restraints and decreasing flat bottom ones (respectively 750, 500 and 250 kJ/mol·nm²). Finally the production run was carried

on for 1 μ s. All runs were performed with a 20 fs time step.

MARTINI coarse-grain simulations For MARTINI [? ?] coarse-grain simulations, peptide coordinates and parameters were obtained from the atomistic ones using martinize.py [?], and pycgtool.py [?] for the central residue. Parameters for the joining bonds were derived from tabulated values of analogous ones.

The bacterial and mammalian model membranes, hosting 2880 lipids each, were built with insane.py [?], with composition DLPC:DLPG 3:1 and pure DLPC respectively. The simulations parameters used for lipids are consistent with Ref. [?]. The peptide-membrane systems were built placing the buckyball at a minimum distance of 1 nm from the membrane surface.

For either the simulations in solution and with the membrane, counter ions were added for the standard MARTINI simulations, while for the ones run with Polar MARTINI, additional ones to reach a 150 mM concentration were inserted.

For simulations performed with the standard water model, the temperature coupling was performed with a velocity rescale thermostat [?] with a coupling constant τ_T of 1 ps. An isotropic or semi-isotropic pressure coupling was applied (for peptide in solution or peptide on membrane simulations) at 1 bar pressure, with 4.6×10^{-5} bar $^{-1}$ isothermal compressibility, using a Berendsen [?] or Parrinello-Rahman barostat [?] (equilibration and production phase respectively) with a τ_P of 12 ps. An isotropic or semi-isotropic coupling has been used for simulation without and with membranes respectively. Coulomb interactions were treated with a Reaction Field scheme [?] and cut off radius of 1.1 nm, van der Waals interaction with a cut off scheme and the same cut off radius. The relative dielectric constant is set to 15. Simulations performed with the polar water model were run with the parameters above, except the relative dielectric constant set to 2.5, and the choice of a PME scheme for the long range Coulomb interaction (1.2 nm cut off radius).

For simulations of the capsule in solution, for both water models, after energy minimization four 10 ns equilibration runs (NPT ensemble, 300 K, 10 fs time step) were performed with flat bottom positional restraints on the unpaired branches (respectively at 750, 500 and 250 kJ/mol·nm 2) - as done for the other force fields. Finally the production run was carried on for 1 μ s.

All runs were performed with a 20 fs time step.

It is interesting to notice that the MARTINI force field with the standard water model produce very similar results even without such refined equilibration procedure. However, we choose to follow it to have results more consistent with the other parametrisations and be certain that the differences observed are due only to the force field and not the equilibration procedure.

The membranes used in the simulations are equilibrated for 1 μ s and the final configuration used to build the peptide-membrane system, together with the equilibrated capsule structure (before production). The full system is then after energy minimised and equilibrated for 500 ps. Production is followed for 10 μ s for simulations with the standard water, and 1 μ s with polar water.

The adoption of the polar water model allows to perform electroporation experiments also with the MARTINI force field (while the standard water, not bearing a dipole, is unable to screen the externally applied electric field, resulting in unphysical effects). We thus resorted to a procedure similar to the one employed for atomistic simulations, testing an external electric field of magnitude 20 mV/nm and 40 mV/nm, finding that the latter is sufficient to trigger electroporation in model bacterial membranes in less than 500 ns (while the former is not). It is likely that longer simulations would allow to observe this behaviour even with lower values of the force field, however, in the interest of time, we selected this value for further investigation.

Specifically, we selected a configuration from the early stages of poration (where the pore showed a diameter of 2 nm) and continue the simulations switching to an isotropic pressure coupling. Indeed,

3.3 Analysis

Simulations in solution Several structural analysis were performed on the outcome of the simulations of the buckyball in solution. The Radius of gyration (R_g), the Root Mean Square Deviation with respect to the initial configuration (RMSD) were computed with the GROMACS software. To get the average distribution of the mass of the capsule around its center, the Radial Distribution Function (RDF) of the protein masses around their center of mass was computed (with GROMACS). The profile could be fitted with a Gaussian function (with the R [?] software), so that the position of its maximum

can be taken as an average radius of the capsule, and its Full Width at Half Maximum (FWHM) as an estimate of the thickness of the bilayer.

The dynamical character of the structure was assessed computing the correlation of motion between the molecules. The central atom (or bead) from which the branches depart was taken as reference. For all the pairs i and j of such reference positions, the covariance of motion $\sigma^2(i, j)$ was computed (with GROMACS), where the covariance is the sum of the components along each axis: $\sigma^2(i, j) = \sigma^2(i, j)_x + \sigma^2(i, j)_y + \sigma^2(i, j)_z$. Then for each pair, this measure was normalised as:

$$\text{corr}(i, j) = \frac{\sigma^2(i, j)}{\sqrt{\sigma^2(i, i) \cdot \sigma^2(j, j)}}. \quad (3.1)$$

Another structural measure performed concerns the pairing of the branches. Two branches are defined as paired if their center of mass is closer than a cut off distance of 1.2 nm. This simple measure discards any more precise information on the orientation of the chains with respect to each other, and aims at checking whether the network of molecules present in an ideal buckyball structure is maintained. In the ideal buckyball, contacts within the same layer sum up to 90 for each layer (for a total of 180 in a bilayer). This measure can be easily applied to any description (atomistic or coarse-grain) without disagreement in the interpretation. The computational pipeline combines GROMACS tools and a post-processing in R language.

To characterise the chemical determinants that promote the assembly, we investigate the interactions between amino acids of different types, computing the number of contacts between backbone and side chains of single amino acids, filtered for the ones present a least 50% of the simulation time, and classified them by amino acid type. We define contacts between amino acids backbones if the $C\alpha$ of two residues (or the corresponding coarse-grain bead) are closer than a cutoff distance of 0.6 nm; and between side chains if selected reference atoms in the side chain are closer than the same cut off; finally mixed ones if the proximity is between a $C\alpha$ and the side chain reference atom. As side chains reference, we took the heavy atom or bead farthest away from the backbone (respectively for GROMOS, SIRAH and MARTINI: CZ/SC2/BCZ for Arg, CZ2/SC4/BNE for Trp, OG1/SC1/BPG for Thr and CD/SC1/BCD for Glu). The functions to perform the analysis were built on the ones implemented in

the MDAnalysis software (see Appendix –).

Further investigation has been carried on for atomistic simulations computing the hydrogen bonds between amino acids and grouping them by amino acid type and by region of occurrence (e.g. between two backbones, side chains or connecting a backbone atom and a side chain one).

Finally, the Solvent Accessible Surface Area was computed for each amino acid (with GROMACS). Its value was averaged in time and over all the residues of the same type. It was normalised over its reference value for each residue type X, obtained as the measure of the SASA of X from a Gly-X-Gly tripeptide. The resulting measure (named Q_{SASA}) takes into account the size of the side chain of each amino acid, giving a measure of exposure which can be compared between different residues. This normalisation is somewhat inappropriate for coarse-grain models, however we employed it as it provides nevertheless a coarse regularisation for size effects.

Simulations on a membrane To assess the influence of the peptide or of the presence of an electric field on the membrane, the area per lipid (ApL) is monitored, i.e. how much space each lipid molecule has on the membrane plane. In the case of approximately flat membranes aligned to the xy plane, this can be assessed from the product of the lateral dimensions of the simulation box divided by the number of lipids in one leaflet. For highly curved membranes, as the ones obtained under the effect of a strong electric field, this measure does not reflect the true spacing between lipids. In that case the true ApL was computed through the algorithm developed by Braun et al. [?]: it first identifies the undulating reference surface $u(\mathbf{r})$ of the membrane fitting a Fourier series to a reference atom per lipid (presently the Phosphorus), then it computes the true ApL based on the interpolated surface $u(\mathbf{r})$ as:

$$ApL = \frac{1}{N} \int_{x_{box}} \int_{y_{box}} \left[\sqrt{1 + (\nabla u(\mathbf{r}))^2} \right] dx dy \quad (3.2)$$

where the gradient $\nabla u(\mathbf{r})$ is zero for a perfectly flat surface.

For atomistic simulations, the average order of lipid tails has been assessed computing the deuterium order parameters S_{CD} of the acyl chains for each lipid bilayer leaflet separately. S_{CD} measures the orientation with respect to the bilayer normal of a carbon-hydrogen bond in a given position along the

chain for each lipid in the bilayer. Their spread is evaluated according to the ensemble average:

$$S_{CD} = \frac{1}{2} \langle 3 \cos^2 \theta - 1 \rangle. \quad (3.3)$$

As the GROMOS force field employs a united-atom representation, the tetrahedral positions of the hydrogens are constructed based on the neighbouring carbons positions.

To evaluate the mobility of lipids on the membrane plane, for each simulation we extracted the trajectory of the phosphorus atom of every lipid, in the top and bottom leaflet separately, removing the collective motion of the leaflet with respect to the simulation box frame. Each trajectory was used to compute the Mean Square Displacement (MSD) of the lipid as a function of time. These curves were averaged over all the lipids of a given species belonging to the leaflet of interest, and, for a given interval of time, on all the possible time windows of that length fitting within the simulation time analysed. The diffusion coefficient D was obtained from a linear fit of the average MSD profile, following Einstein equation in two dimensions:

$$\langle (x - x_0)^2 \rangle = 4Dt. \quad (3.4)$$

The fit was performed in the regions which showed a linear dependence. This implied discarding the first interval of the profile, where the behaviour is not linear, and the last 50 ns, where the poorer statistics leads to more noisy data. The exact amount of the discarded initial region depends on the system studied, e.g. whether the protein is present on the membrane or not.

For atomistic simulations of the protein on the membrane, the network of hydrogen bonds between the protein and lipids has been computed, classifying the interactions by the protein residue type and the lipid species they occur between. Finally, another useful information was given by the number of lipids of a species within a threshold distance from the protein, in time. The aforementioned diffusion constant can also be computed selectively on the lipids which, at the initial time, are within this threshold distance or, conversely, which are further away from it.

To quantify whether a lipid patch is undergoing transition to a gel phase, the regular packing of the lipids' chains was quantified by the hexagonal order parameter S_6 [?]. Specifically, a chain was represented by its position on the

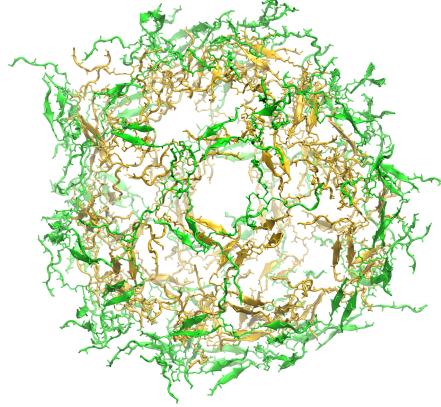


Figure 3.5: Final snapshots (100 ns) from the simulation of a buckyball in solution (replica 3): bonds and cartoon representation, backbone only, green external layer and yellow internal one.

xy plane (parallel to the membrane surface), computed as the average x and y position of its carbon atoms. For each chain j it was computed the set of neighbouring chains $\{n\}$ as the ones within a 0.65 nm radius from j . Then S_6 is defined as:

$$S_{6,j} = \frac{1}{6} \left| \sum_{k \in \{n\}} e^{6i\theta_{jk}} \right| \quad (3.5)$$

with θ_{jk} being the angle between the vector connecting j and k , and the x axis (and i is the imaginary unit). A chain is in gel phase if it has an hexagonal order parameter larger than 0.72 [?].

3.4 Results: capsule in solution

We list here the results for the capsule in solution first: starting from the atomistic simulations, we then proceed to compare the outcome with the ones obtained by different coarse-grain models.

3.4.1 Atomistic simulations

We discuss first the results from atomistic simulations, color coded as black or in shades of grey in the following sections. However, all the plots present back to back the results from coarse-grain runs as well, anticipating the discussion of Section 3.4.2, as this would make easier the comparison later on.

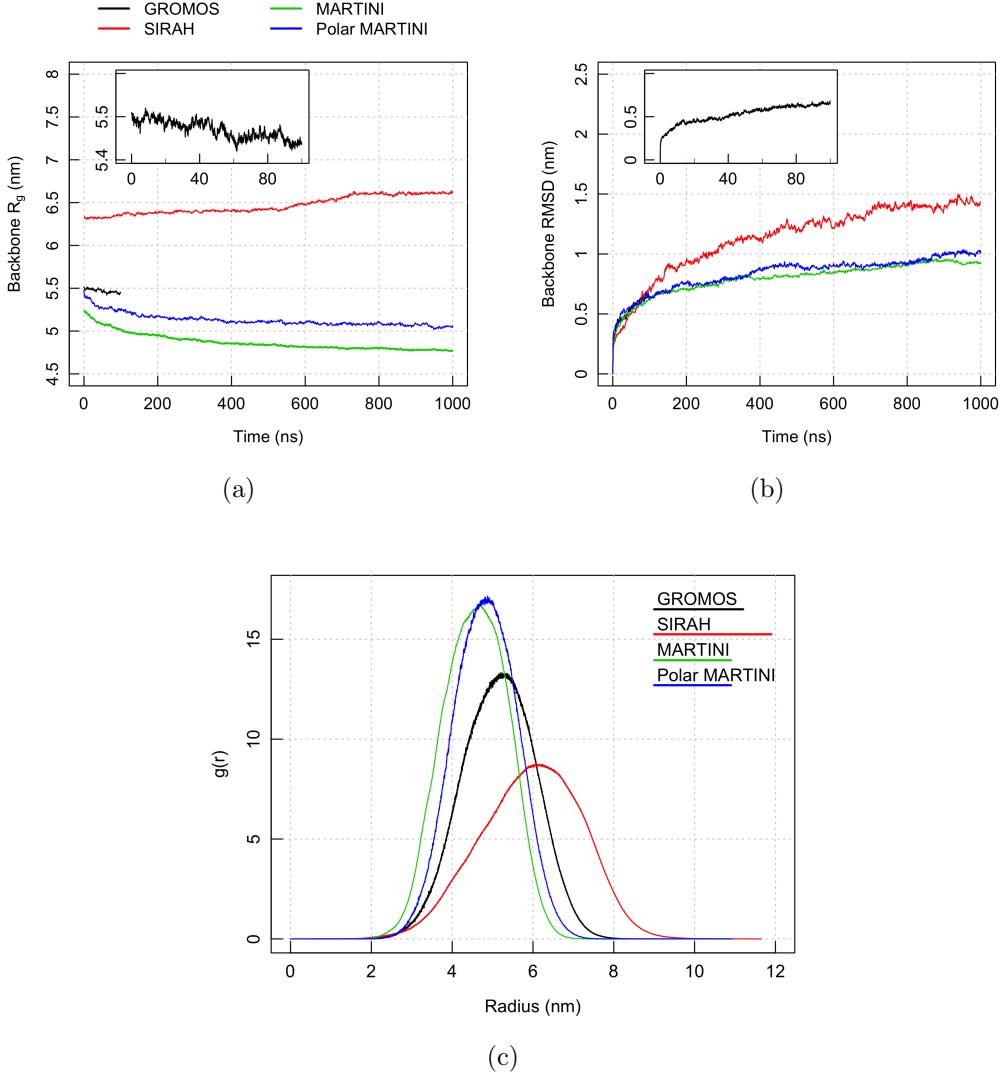


Figure 3.6: (a) R_g and (b) RMSD computed respectively on the Protein and its backbone. Results are displayed for simulations performed in GROMOS (100 ns), SIRAH, MARTINI and MARTINI with polar water (all 1 μ s). Inset: zoom on the GROMOS values. (c) RDF of Protein masses around their center of mass, computed on both layers, displayed for the same simulation set up as in (a,b). For each label of the legend, the bar has length of the respective RDF FWHM (thickness estimate). All results are showed for Replica 1 of each simulation setup.

Global capsule structure Atomistic simulations of the buckyball in solution show a consistently equilibrated structure across the three replicas (Figure 3.5 and SI movie –). This is proven by both the stable value of the protein R_g and the almost plateauing backbone RMSD (Figure 3.6(a) and (b)). It is interesting to notice that previous simulations performed with a shorter equilibration, without the phase employing flat bottom restraints, resulted in the immediate disruption of several connections in the buckyball network, with resulting larger R_g . This suggests that the structural pairing present in the structure can form only when the chains are in close contact. For this to happen, many conformations need to be sampled, and this is compatible with the long time of assembly observed experimentally (up to 7-10 hours). The RDF of the protein masses around the buckyball centre of mass shows a Gaussian profile (Figure 3.6(c)). The fact that no masses are observed nearby the origin means that the molecules do not collapse to the center and the central cavity is maintained. To be noticed that, given the way RDF is computed and normalised, a uniformly full object would display a flat distribution. A fit of the RDF to a Gaussian curve returns a mean value of 5.1 nm and a FWHM of 2.2 nm, which gives an estimate of the bilayer thickness. A similar computation is repeated for the inner and outer layer separately, providing 1.1 nm of distance between the two distributions means. This interlayer distance is compatible with the distance between the backbones of stacking β -sheet in structures like densely packed amyloids with cross- β sheet quaternary structure (1.0 nm [?]).

This thickness value hints at the fact that the two layers are closely packed. This is confirmed by the analysis of the pairwise correlations of motion: molecules at the same polar coordinates (i.e. stacking radially one on the other) have a positive correlation and so move coherently, while the ones at opposite poles, as well the ensemble of all possible pairs, do not show particular correlation (Figure 3.7).

Contacts between molecules The measure of paired branches in the buckyball network, shown in Figure 3.8, reports an average (in time) of 240 pairings between arms belonging to the same layer (summing over inner and outer), four pairing per molecule, and around 60 only for inter-layer ones. This first value is larger than the one predicted for a perfect buckyball arrangement (which

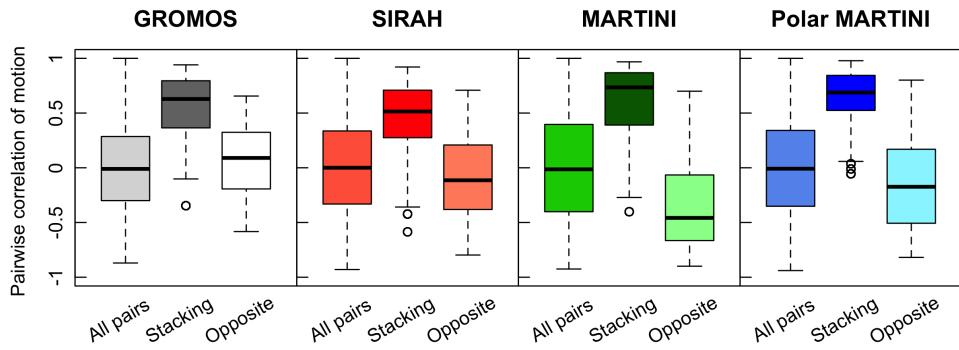


Figure 3.7: Distribution of the correlation of motion between different molecules in the buckyball simulations. Black band: median of the distribution; box: first and third quartiles; whiskers: maximum and minimum, outliers excluded (hollow dots). Results are shown for Replica 1 of each simulation setup.

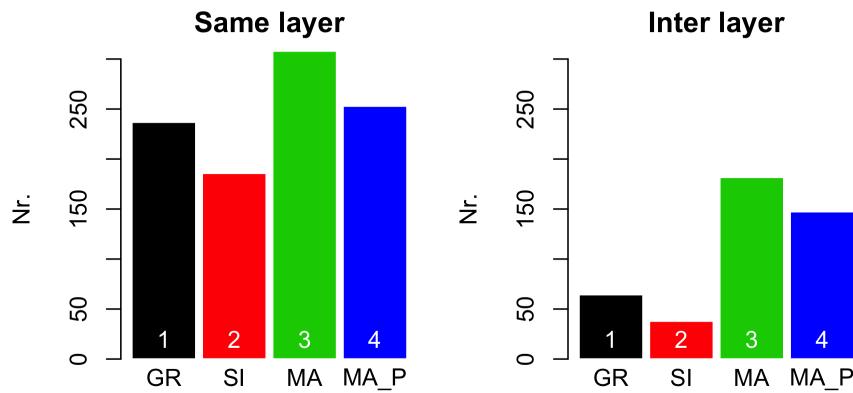


Figure 3.8: Number of paired branches within the same layer and between layers, as defined in the main text. Results are shown for Replica 1 of each simulation setup.

would be three), likely due to the fact that the structure contracts slightly with respect to the initial size and some branches deviate from the original position, locating themselves in the proximity of multiple neighbours. On the contrary, few inter layer contacts are observed within 1.2 nm distance cut off because of the steric hindrance of the side chains which are located in between the two layers, as they keep the average position of inter-layer backbone stretches at a distance greater than the cut off chosen.

To look more into the details of the interactions between arms of the molecules, we computed the contacts between backbone and/or side chains of amino acids, normalised by the number of molecules present, and classified the ones which survive more than 50% of the simulation time by residue type (Figure 3.9). The number of backbone contacts per molecule is around 3 for Threonine, Glutamic acid, 4 for Arginine, and between 6 and 7 for Tryptophan residues. As in each molecule there are 3 Threonine and 3 Glutamic acid but 6 Arginine and 6 Tryptophan, this proves that on average each residue is well paired with another one, except for Arginines: only two thirds of them are paired on average, likely because of their terminal position. At the same time however the bar plot shows also that there is no rigid arrangement between branches. Indeed, for example, Tryptophan residues are not chiefly paired to Tryptophan ones, as the optimal arrangement would be, suggesting flexibility in the structure. Nevertheless, the analysis highlights that Trp residues are key to form contacts with the neighbours and (from the central column) that their cation- π interaction with Arginine through their side chains is an important element of the structure.

Hydrogen bonds interaction Some of these contacts are mediated by an hydrogen bond interaction: the number of them present during the simulations is computed, and divided by the number of molecules. We classify the hydrogen bonds occurring between backbones and/or side chains of the amino acids, breaking them by amino acid type. Tryptophan contributes to a large number of backbone hydrogen bonds (Figure 3.10 top), especially with other Tryptophan residues, consistently with what found in the analysis of contacts carried on above. Arginin side chains are the most prone to establish H-bonds as a donor with many different amino acid side chains, but especially with Glutamic acid as expected from the facing positions they occupy in the

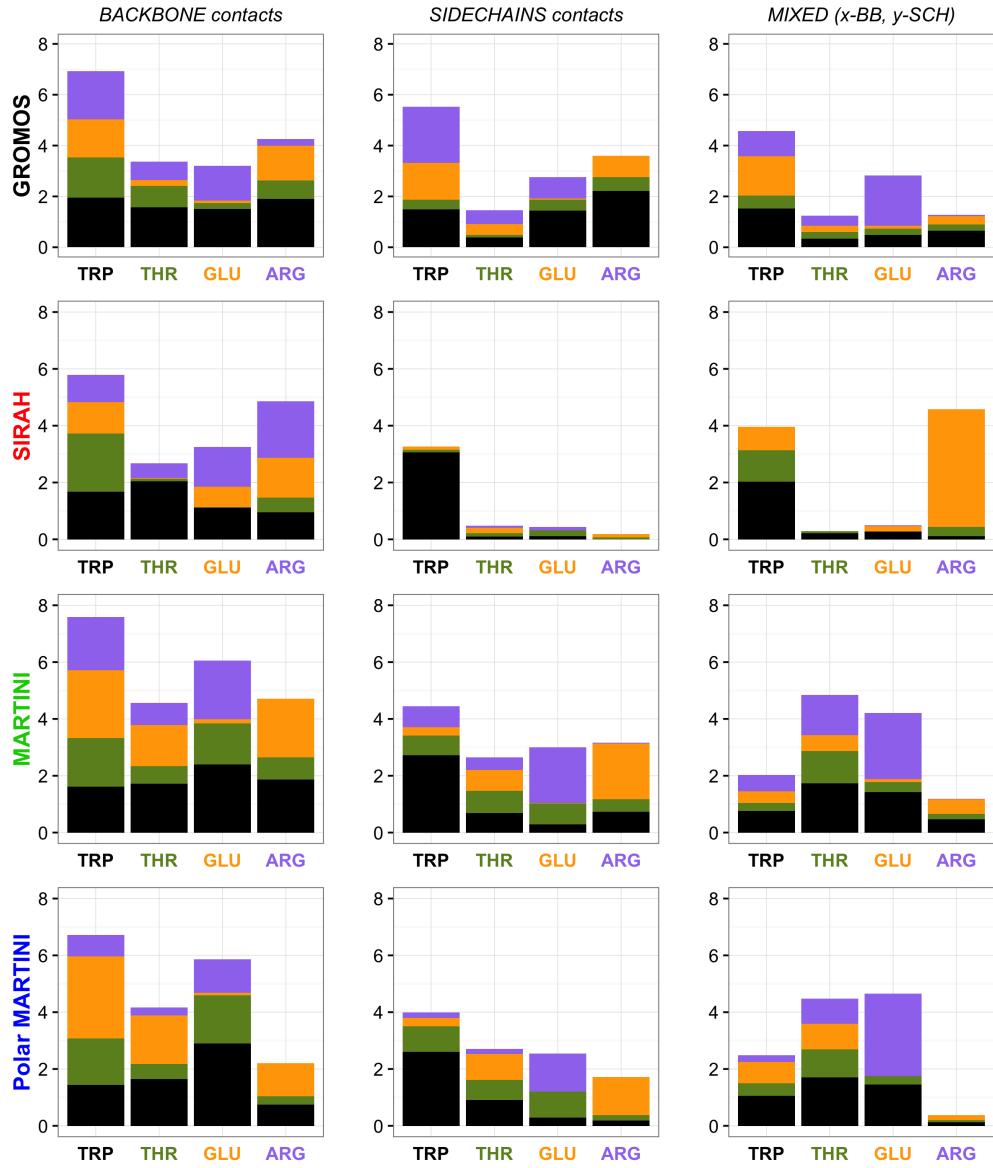


Figure 3.9: Contacts with persistence greater than 50% between C_{α} s (Backbone contacts), between side chains (Sidechain contacts) and mixed (in which case, for each bar, the residue on the x -axis contributes to the contact with its backbone, and the bar is split by the identity of the partner engaging its side chain). The simulation setup is reported along the y -axis. Results are shown for Replica 1 of each simulation setup.

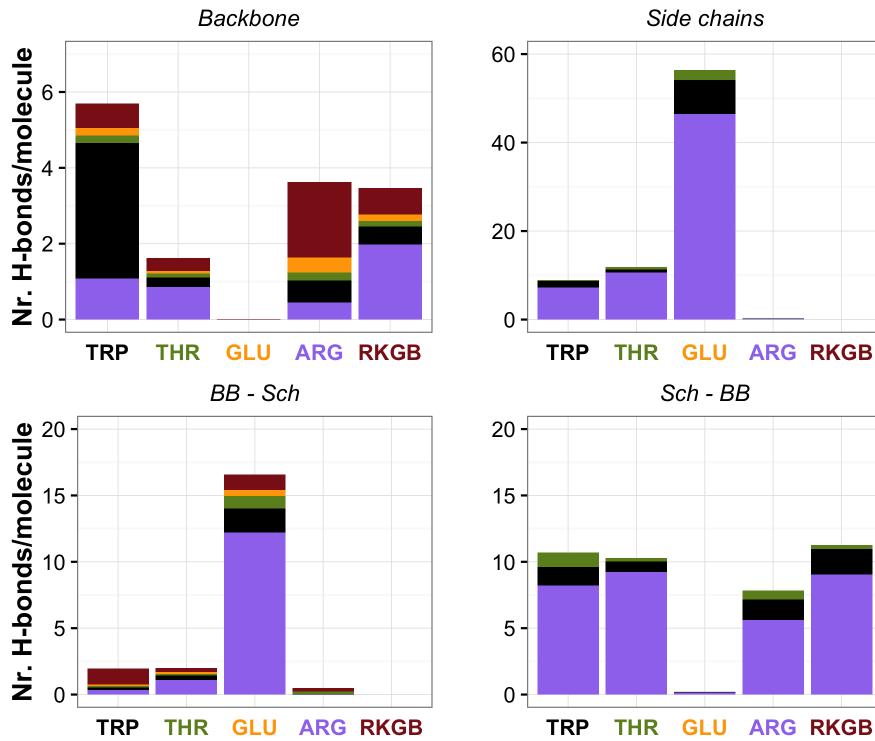


Figure 3.10: Average number of hydrogen bonds per molecule occurring between amino acids, including the central scaffold RKGB, in a 100 ns atomistic simulation of the buckyball in solution. Result are shown for Replica 1. For each bar, the residue on the *x*-axis is the acceptor, and the bar is split by the identity of the donors. In the case of Backbone - Side chain and Side chain - backbone, the first mentioned correspond to the acceptor (and thus the residue on the *x*-axis).

molecules arrangement.

Chemical characteristics of the surface: SASA Finally, it is important to understand what residues are exposed at the surface of the structure, especially in view of future applications: in order to make the peptide co-assemble with other products, the two must have a compatible chemical character. To understand what surface the peptide exposes to the solution, we compute the average Solvent Accessible Surface Area (SASA) per molecule and break it down by amino acid type. Figure 3.11 shows that half of the accessible surface is represented by the charged residues Arginine, while Tryptophan contribute to less than one quarter to it, despite having bulky side chains. For atomistic simulations, we can compare the SASA per each residue X, with the reference SASA computed on simulations of a Gly-X-Gly tripeptide (computation per-

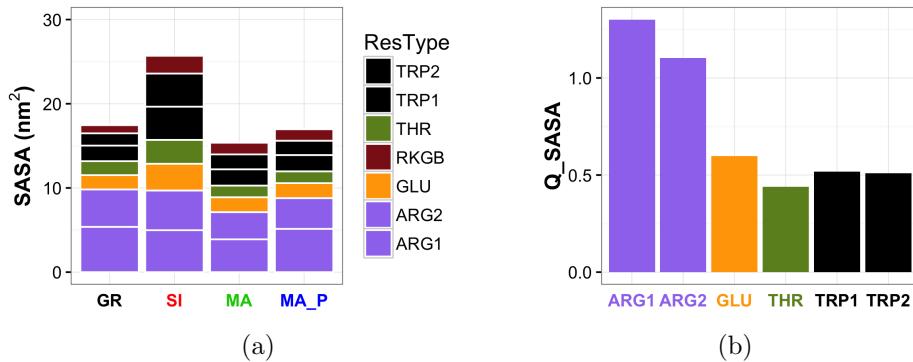


Figure 3.11: (a) Break down of the Solvent Accessible Surface Area (SASA) per molecule by residue types. Results are shown for Replica 1 of each simulation setup. (b) Normalised SASA over the reference SASA computed for each amino acid type X as the value in a Gly-X-Gly tripeptide.

formed with the software POPS [?]). The resulting ratio Q_{SASA} is greater than one for both the Arginines: while for the terminal one (ARG1) it is suspected, the fact that also the second has $Q_{SASA} > 1$ proves that these residues are highly exposed in solution. On the contrary, all the other residues have values around 0.5: for Tryptophan this is due to their propensity to be buried inside the structure, while for Glutamic acid and especially Threonine, despite having charged or polar side chains, they are shielded from the solvent by the large Trp side chains nearby them.

3.4.2 Multiscale comparison of model capsule

We performed a multiscale analysis of the capsule structure simulating it with different coarse-grain force fields, with a twofold aim: first we wanted to simulate the assembly for a longer time, to observe how its structural properties are maintained on the medium time scale (of the order of the microsecond). Second, we believe that proving the stability of the capsule with different descriptions strengthens the evidence that the assembly proposed is indeed a favourable arrangement of the molecules in solution.

As mentioned before, to this aim we compared simulations run with the SIRAH, MARTINI force fields and MARTINI used in conjunction with polar water. The investigation is also useful to compare how the different descriptions differ and to infer the advantages of each model. We first comment on the quantities already analysed at the atomistic level (if applicable), and then

extend the analysis to simulations of a monolayer capsule, which has been modelled to prove the greater stability of the bilayer one.

Force field comparison As foreseeable, the structures obtained with coarse-grain force fields are slightly different with respect to the atomistic one and among themselves. The SIRAH simulated bilayer capsule has a more open structure with respect to the atomistic one, with a skewed and broader radial density profile, while MARTINI with the standard water model provides a more compact configuration (respectively 6.0 nm and 4.6 nm average radius, with 2.9 nm and 1.9 nm average thickness - see Figure 3.6(a) and (c)). MARTINI with polar water provides a slightly larger structure with respect to standard MARTINI (4.8 nm average radius), with a comparable thickness. This is likely due to the poorer properties of solvation of standard MARTINI water, which cause the protein beads to preferentially interact between themselves rather than with the solvent.

The correlation of motion between molecules for all the coarse-grain force field is similar to the results of the atomistic one, with a slight anticorrelation between molecules at the opposite poles (Figure 3.7). This is due to the contraction or expansion happening at the beginning at the simulation, when the capsule adjusts to the equilibrium size, which depend to some extent on the force field. These effects though are slightly more pronounced for the MARTINI force fields, probably due to the greater cohesion between the beads which makes them moving coherently.

The number of chain paired in the SIRAH simulations is fewer than in the atomistic ones (Figure 3.8), in line with a more expanded structure, while MARTINI simulations propose a higher number, consistently with the reduced size of the capsule. In particular, the inter layer contacts are significantly higher (more than the double), as would be expected when the two layers are closer and more strongly bound, as suggested by the values of the stacking molecules correlation. Finally, MARTINI suggest a more compact structure than Polar MARTINI, privileging contacts between residues rather than contacts with water molecules.

Breaking down the contact analysis by residue type (Figure 3.9), coarse-grain simulations shows necessarily a different organisation, due to the fact that different parametrisation results in different proportions between the side

chains volumes of different amino acids. Never the less, in all the representations, Tryptophan has a prominent role in establishing contacts with its neighbours at the backbone level, while different force field suggest different role of the side chains. Quite surprisingly, the SIRAH force field does not promote interactions between side chains apart from the Tryptophan ones with themselves. This seems due to the more expanded structure of the capsule and the larger solvation of the amino acids, which are more exposed in solution (see next paragraph discussing SASA values). Moreover, none of the coarse-grain descriptions seem to capture the preferential cation- π interaction between Arginine and Tryptophan.

Finally, the values of SASA cannot be compared between force fields, because of the different dimensions and number of atoms/beads employed. However, it is interesting to notice that consistently across force fields, Arginine constitute around half of the exposed surface, with the exception of SIRAH, where the more open structure results in other residues to be more exposed as well (Figure 3.11(a)).

Overall, these results suggest that every coarse-grain force field has a particular propensity for an equilibrium distance between peptidic components due to the different property of solvation of the water model chosen. To better understand this, we computed the Coulomb and Lennard-Jones contribution to the energy due to peptide-peptide interactions or peptide-solvent ones. Figure 3.12(a) plots these value for the second half of the simulated time for each coarse-grain force field (to be noticed that for the SIRAH description the Protein-Protein terms include both the short range interactions and the 1-4 interactions, i.e. the ones computed between atoms separated by three bonds. These interactions are not part of the force field for MARTINI models, so only the short range ones are accounted in those cases).

A great difference can be seen among force field, as expected. Looking at the Coulomb component, it is clear that the mean field approach adopted by MARTINI, which consists in adopting a high relative dielectric constant ($\epsilon = 15$) to compensate the absence of water dipoles, decreases massively the contribution of the Protein-Protein electrostatics to the overall energy with respect to the two other models. Indeed this approach decreases the interactions also on the short range, while in reality the screening effect due to water can be seen only on long distances (of about 1 nm in a 0.1 M salt

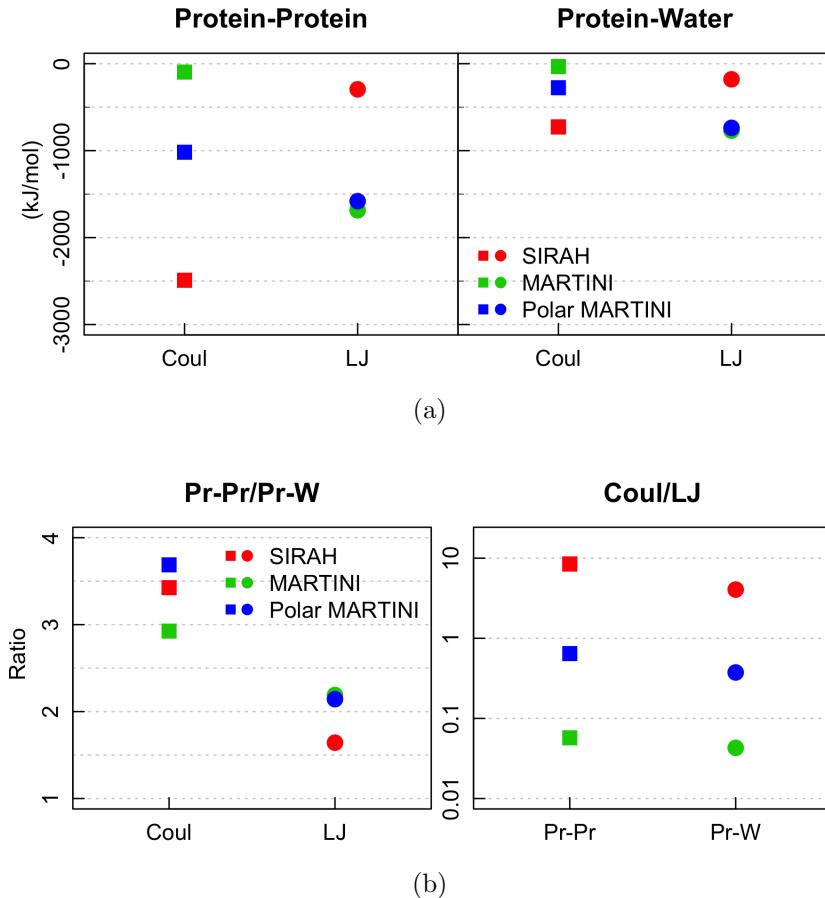


Figure 3.12: (a) Protein-Protein and Protein-Water non-bonded interactions, normalised per molecule. Values obtained as average on the second half of the trajectory of Replica 1 for each setup. (b) Ratio between the Protein-Protein and Protein-Water interactions for each force field, for Coulomb and Lennard-Jones respectively; or between Coulomb and Lennard/Jones, for Protein-Protein and Protein-Water interactions separately (note the log scale on y). Values computed as for plot a).

solution), and, due to the $1/r$ behaviour of the Coulomb interaction, these short scale contributions are clearly important. The partial reversion to $\epsilon = 2.5$, together with an introduction of the water dipole, increases the amount of the Coulomb contribution by 10-fold (in absolute value). This is higher than the ratio of the ϵ between the two models (equal to 6), likely due to the fact that the different screening allows a rearrangement of the charges which brings opposite ones closer to each other.

SIRAH simulations instead, run at $\epsilon = 1$, present Coulomb components 2.5 times larger than the ones of Polar MARTINI, suggesting that the two water models, both representing the separation of water charges in an approximate way, give a similar energy contribution despite being different.

This difference in Coulomb energies affects both the Protein-Protein and Protein-Water interactions consistently, however it can have consequences on the dynamics because it changes their proportion with respect to the Lennard-Jones contributions. These interactions are not changed between the two MARTINI models, making their role predominant in standard MARTINI, while competing with the Coulomb contribution in Polar MARTINI (still being larger overall). On the contrary, SIRAH parametrisation opts for a smaller role of Lennard-Jones with respect to the Coulomb contribution. Therefore, protein electrostatics have an increasing contribution in MARTINI, Polar MARTINI and SIRAH respectively, both in absolute terms and with respect to the Lennard-Jones contribution.

This can partially explain the differences in sizes observed across the models: despite the Protein-Protein Coulomb energy is more negative in SIRAH, because of the way it is computed, with a smaller dielectric constant these contributions are less screened, thus the many positive amino acids composing (which are more abundant than the negative ones) the capsule can repel each other more effectively.

Finally, to understand the ratio between the Protein-Protein and Protein-Water components, for both Coulomb and Lennard-Jones, is computed (Figure 3.12(b)). All the three force fields present values in the same range. The ratio of Lennard-Jones interactions is slightly lower in SIRAH than in the two MARTINI models, and the ratio of Coulomb interactions is between 3 and 4 for all the models. Interestingly, Polar MARTINI has a greater Protein-Protein component with respect to standard MARTINI (in absolute value,

which means a more negative contribution). From this can be deduced that the slightest more open structure observed in Polar MARTINI is likely not due to the more favourable solvation properties of the polar water model, but from a different balance of electrostatic versus Lennard-Jones, as explained before.

An interesting follow up on this topic would be investigating whether a tuning of the dielectric constant used in either cases can produce more consistent results between the two water model. However, the choice of the constant was optimised to reproduce at best the properties of bulk water and solvation free energies of ions for both cases. This then raises the question whether the protein parametrisation performs equally good in conjunction with both model, or which one is more compatible.

Bilayer versus monolayer capsules To prove also on large scale objects that the bilayer structure is indeed essential to grant a structure which does not disassemble or change shape, we perform simulations of a monolayer capsule (specifically taking the external layer of the capsule already simulated) in the three coarse-grain force fields employed so far. We focus here on the results obtained with the SIRAH and Polar MARTINI models, as the ones observed with standard MARTINI resemble closely the Polar version.

The RMSD with respect to the initial configuration of the production (Figure 3.13(a), top) shows that the monolayer undergoes a larger conformational change than the bilayer in the SIRAH force field. This effect is not traceable in Polar MARTINI within the same plot. However, if we consider the RMSD with respect to the initial structure built (the geometrically regular polyhedra as in Figure 3.2(d)), it is clear that for Polar MARTINI the discrepancy between monolayer and bilayer is even more pronounced (Figure 3.13(a), bottom). This is likely due to the fact that rearrangements are more favoured in short time during the more coarser simulations.

This larger change of the monolayer is due to a larger contraction of the structure, which collapses within its center (Figure 3.13(b)). This is evident especially for Polar MARTINI, while for the $1 \mu\text{s}$ SIRAH simulations, together with the contraction it is observed that Tryptophan side chains have a tendency to become closer, modifying locally the structure and causing a local puckering (see Figure 3.14). Likely this mechanism is present also in the MARTINI ones, but is it overshadowed by the general contraction, which brings the chains close

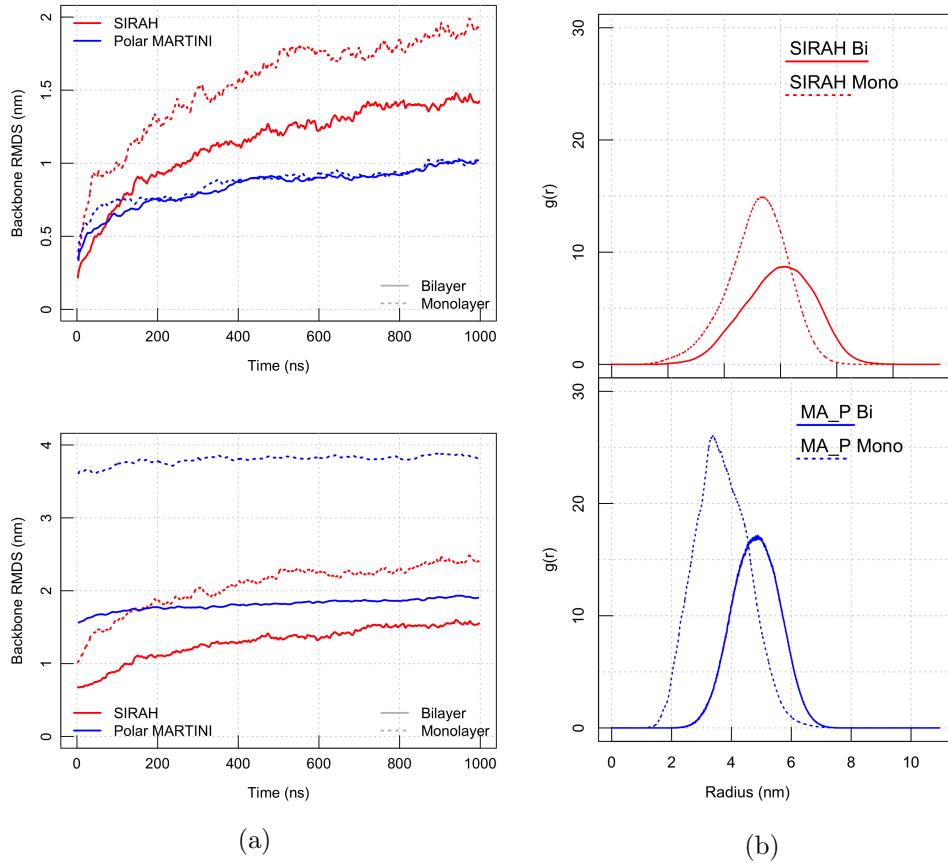


Figure 3.13: (a) RMSD of the monolayer and bilayer structures for SIRAH and Polar MARTINI force field, with respect to the initial structure of the production (top), and initial geometrical configuration (external layer of Figure 3.2(d)). (b) RDF of Protein masses around their center of mass. For each label of the legend, the bar has length of the respective RDF FWHM (thickness estimate). Results are shown for Replica 1 of each simulation setup.

enough to screen the hydrophobic components without a puckering of the arms as the one observed in SIRAH.

Finally, the SASA of each residue type computed on the initial configuration is higher for the monolayer than the bilayer, while this difference partially levels out during the simulations, due to the rearrangements mentioned above (Figure 3.15).

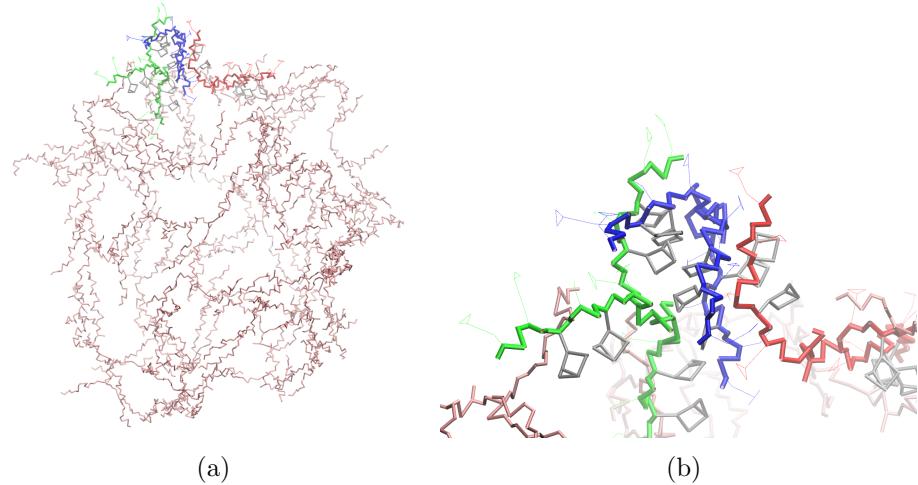


Figure 3.14: [REDO] (a) Final configuration of a SIRAH monolayer simulations (Replica 1), with highlighted a protruding puckered structure recurrently occurring several positions of the capsule. (b) Detail of the puckered structure. Representations: backbone of the full structure in pink bonds; backbone of highlighted triskelion molecules in green, blue and red bonds; side chains of selected molecules in green, blue and red lines, excluded Tryptophan, in silver bonds.

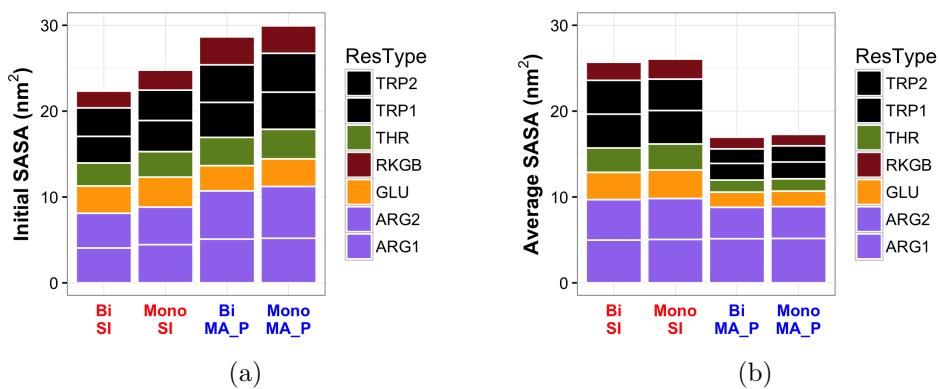


Figure 3.15: Break down of the Solvent Accessible Surface Area (SASA) per molecule by residue types for simulations of the bilayer and monolayer structure. Results are shown for Replica 1 of each simulation setup. (a) SASA compute from the initial configuration; (b) from the average over the production run.

3.4.3 Discussion

All the findings above strengthen the hypothesis that a favourable packing must allow for the screening of hydrophobic residues, arranging multiple copies of them close enough to interact together. For this reason, the bilayer structure is more prone to maintain its original configuration.

Based on this, we deemed that the capsules observed experimentally must have a non-monolayer structure: a single layer would not provide enough structural stability and, moreover, is not compatible with the thickness observed in some of the TEM images collected experimentally.

The multiscale investigation of this system provided both information on how it is structured, and a useful comparison between force fields, proving how the same system can be described in different ways from different models. This limited comparison might still be useful especially in the light of the comparatively new SIRAH force field, which has been less assessed than the MARTINI description.

3.5 Results: peptide-membrane interactions

We now discuss the results obtained from atomistic simulations of the pentagonal peptidic subunit in contact with a model bacterial or mammal membrane, both under standard simulation conditions and with an applied external electric field. In this we elucidate the local effect of the peptide on the lipid organisation within the bilayer. We then assess the process of membrane binding through coarse grain simulations, without and with electric field, comparing the MARTINI and Polar MARTINI force fields for the simulation without an external potential, and necessarily resorting to Polar MARTINI for the ones involving the electric contribution.

3.5.1 Atomistic simulations of the bacterial model membrane

Influence of the peptide on the 512-lipids patch For the model of bacterial membrane chosen (composition DLPC/DLPG 3:1), the ApL obtained from simulations was, according to the setup, $0.5796 \pm 0.0052 \text{ nm}^2$ or $0.5686 \pm 0.0036 \text{ nm}^2$ (respectively 512 lipids with GROMOS 54a7 force field or 740 lipids with GROMOS 54a8 force field - Figure 3.16(a), lines 1 and 5). The

differences can be attributed to the force field, however extensive tests on many phosphocholine lipids (see Chapter 4) suggests that the trend is generally the opposite, with 54a8 giving larger values than 54a7. Therefore we attribute the discrepancy partly to reproducibility issues, partly to the different sizes of the patches which has a great effect, as it will be shown in the following.

Regarding the smaller patch, the presence of the pentagonal peptide sub-unit (made of ten antimicrobial molecules, bearing a $+60\text{e}$ charge) made the ApL decrease by 9% with respect to the apo value (Figure 3.16(a), line 3). A similar effect (qualitatively and quantitatively) was observed applying an electric field of 70 mV/nm intensity or combining the presence of the peptide with a weak field of 20 mV/nm, comparable with physiological values (Figure 3.16(a), lines 2 and 4). Likely, they give comparable perturbations as the high protein charge produces a similar field which interacts strongly with the anionic lipid species.

The fact that all the different perturbations lead to a similar reduction of ApL also from the quantitative point of view suggests that the membrane is approaching the maximum packing allowed. Consistently, the diffusion coefficient D had a 6-fold reduction for both species of lipids involved for any of the perturbations mentioned and the order parameter S_{CD}^{avg} increased slightly (Figure 3.16(b)) and (c)). The variability on this last quantity is high, showing that some regions reached a S_{CD}^{avg} value larger than 0.3, which is often associated to transition to the gel phase [?]. Computing the hexagonal order parameter S_6 for the bottom leaflet of the system in presence of the peptidic subunit shows that only a few lipids were in the gel phase ($S_6 \geq 0.72$), without the appearance of ordered clusters (Figure 3.17), however none of them has this property in the apo simulations.

Influence of the peptide on the 740-lipids patch It is remarkable that when the same perturbations were applied to a larger membrane, ApL and tail order parameter S_{CD}^{avg} were equal within the error between the different conditions tested, while the diffusion coefficient D was still significantly affected, being reduced by 2.5 times (Figure 3.16(a)-(c), lines 5 to 8).

This suggests that the simulation conditions can affect remarkably the outcome and the possible interpretation: we reckon that the larger impact on the smaller membrane is due to the fact that local perturbations are effectively

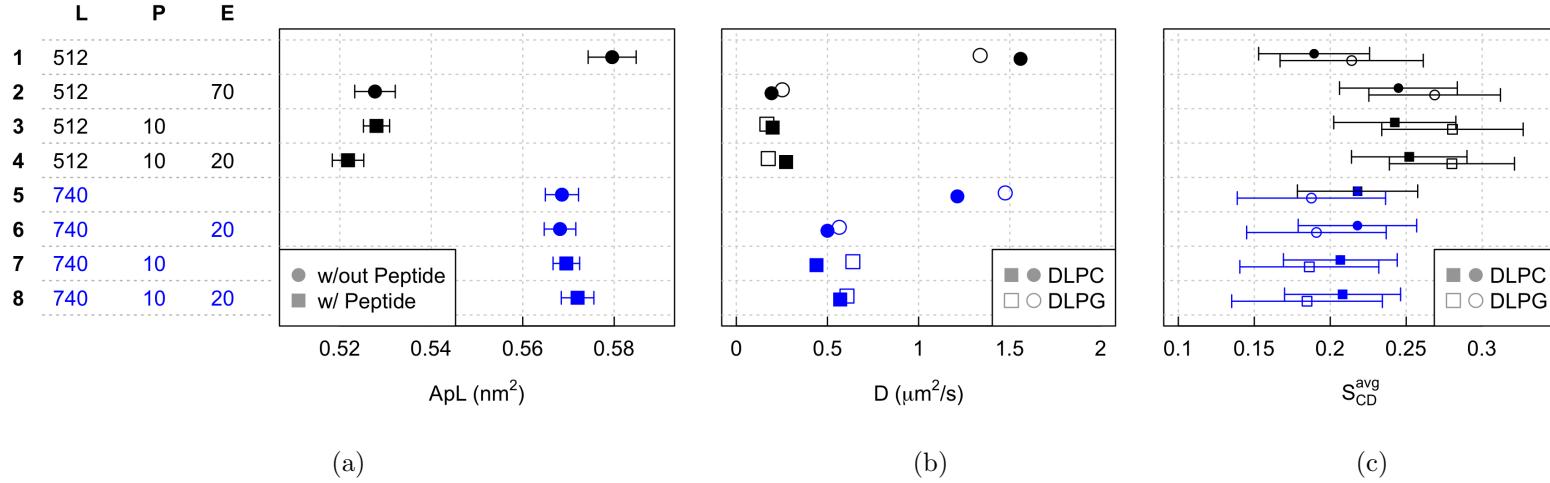


Figure 3.16: (a) Area per lipid (ApL); (b) lateral lipid diffusion coefficient (D), (c) average tail order S_{CD} from simulations of model bacterial membrane patches simulated under different conditions. On the left of (a) and (c) a schematic of the system setup shows the number of lipids L, number of peptides P, and electric field E applied in the z direction (in mV/nm) in each case. Black points refer to simulations on 512 lipids patches, blue ones on 740 lipids patches; squares denotes presence of peptide molecules; in (b) full symbol are for DLPC and hollow for DLPG. Bars denote the standard error; for the diffusion coefficient, these are smaller than the point size. When more replicas are run, results are shown from replica 1. Analysis performed discarding the first 200 ns of simulations.

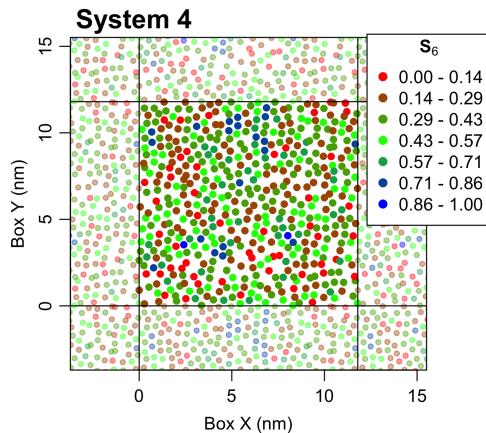


Figure 3.17: Hexagonal order parameter S_6 for the bottom leaflet lipid acyl chains, computed on the last frame of a simulation of a 512 bacterial patch with 10 peptide molecules (line 3 in 3.16). Chain plotted by the average xy position of their carbon atoms, and colour coded by the S_6 value; chains of the periodic images shown faded out; boundaries of the simulation box in solid black lines.

communicated to other parts of the membrane through periodic boundary conditions, while this is less effective in a larger patch. Alternatively, this can also hint at the fact that the 54a8 force field proposes a more rigid structure of the membrane with respect to GROMOS 54a7. However, results from simulations of the membrane alone presents similar values of both ApL and D (with a slight reduction in the ApL when going to the larger membrane). We reserve further tests on this subject for simulations under an applied electric field, as it will speed up the appearance of any differences between the two setups.

Locally reduced mobility The common effect observed on both membrane sizes is a diminished mobility of the lipids. To understand the local determinants of this outcome, we computed the diffusion coefficient for lipids based on their distance from the peptide at the initial frame of the production run (Table 3.3). All lipids are slowed down in their diffusion, with the ones closer to the protein being more affected. To be noticed that values of the diffusion coefficient smaller than $1 \mu\text{m}/\text{s}^2$ (as the ones observed for the lipids within 1 nm from the peptide) are usually associated to the gel phase [?].

This effect is due to electrostatic and hydrogen bond interactions between the lipids and the peptide. We computed their number in the large membrane

DLPC			DLPG		Any		
	Region	Nr.	D ($\mu\text{m}^2/\text{s}$)	Nr.	D ($\mu\text{m}^2/\text{s}$)	Nr.	D ($\mu\text{m}^2/\text{s}$)
Peptide	$d < 1$	93	0.248(3)	38	0.212(3)	131	0.239(3)
	$d < 2$	152	0.278(3)	52	0.288(5)	204	0.280(3)
	$d < 3$	239	0.353(3)	70	0.318(3)	309	0.344(3)
	$d > 3$	329	0.408(3)	98	0.638(3)	427	0.465(3)
Apo	All	568	1.213(5)	169	1.475(8)	737	1.278(6)

Table 3.3: Diffusion coefficients of lipids in a 740 lipids bacterial patch, simulated in presence of the pentagonal peptide subunit. The values are computed for groups of lipids which, at the initial time, were within 1 nm, 2 nm, 3 nm from the peptide or further away. Values are split for the two lipid species involved. The values for the patch simulated without the peptide are reported as reference. Error from linear fit in parenthesis.

E=0	Total	$\tau \geq 40\%$	E=20	Total	$\tau \geq 40\%$
All	2041	62	All	92	75
	R W	R W		R W	R W
DLPC	668 73	25 8	DLPC	33 8	26 8
DLPG	928 114	13 3	DLPG	27 5	19 5

Table 3.4: Number of hydrogen bonds between the pentagonal peptide sub-unit and lipids in simulations of the 740 lipid bacterial patch with the sub-unit and no electric field (left) or a field with value 20 mV/nm (right). Values are divided by the total number counted and the number of those present for more than 40% of the simulated time. The total peptide-lipid hydrogen bonds and the one between each lipid species and either Arginine or Tryptophan residues are given.

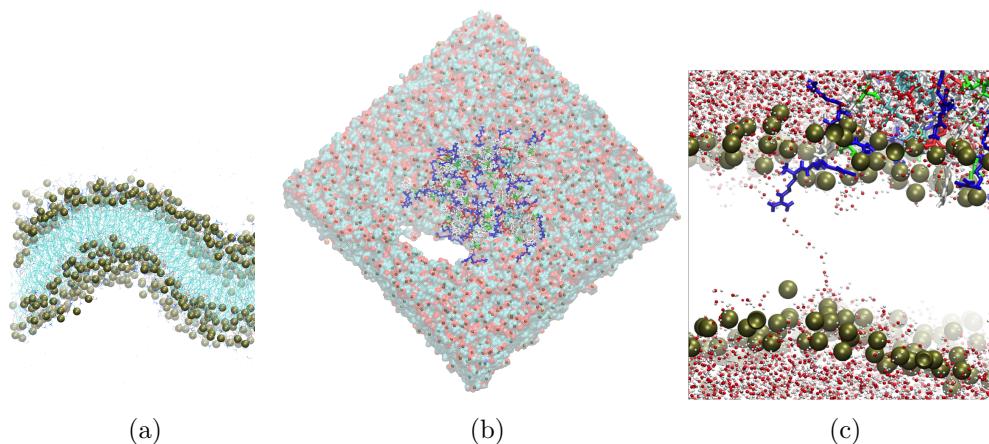


Figure 3.18: Panel a: Membrane deformation due to an electric field applied to the membrane (512 lipids, $E = -130 \text{ mV/nm}$). Panel b: pore formation due to the action of the peptide. Panel c: pore precursor due to Arginine insertion and water penetration.

simulations, monitoring the ones which persisted more than 40% of the time (Table 3.4), singling out the ones involving Arginine and Tryptophan residues. Arginine promotes long life bonds, and the number of them is equally divided between DLPC and DLPG residues, despite these last are way less in number. Tryptophan instead promotes about 2 times more bonds, but very few persist for more than 40% of the simulation time. These interactions, together with the electrostatic coupling between positive residues and DLPG, bond some lipids to remain close to the peptide, slowing down their dynamics which in turn influence the one of the molecule nearby.

We think that this reduced “fluidity” is an important perturbation of the membrane structure, as it allows for less rearrangements in case of further external stimuli.

Electroporation results As it is not possible to witness penetration of the peptide through the membrane within the available simulation time, we opted to perform electroporation simulations. An electric field of increasing intensity was applied along the negative z direction perpendicular to the membrane, with the peptide on the positive z side, to model the field generated by the transmembrane potential.

The field was increased of 20 mV/nm every 200 ns (or 10 mV/nm when approaching the poration threshold), showing that the critical value of 130

mV/nm triggered poration in presence of the peptide. This was confirmed by three replicas run from the initial unperturbed membrane configuration with a 130 mV/nm field, all resulting in poration after 20 ns, 75 ns and 71 ns respectively (Figure 3.18(b)). Similarly, for simulations with the same field on a larger patch (740 lipids) poration occurs at 60 ns, 50 ns and 70 ns respectively. As a control, a 512 lipids bacterial membrane is simulated under the same conditions, for 600 ns in three replicates: despite a reduced diffusion coefficient and the appearance of curved regions, no poration was observed (Figure 3.18(a)). The appearance of a curvature made necessary to compute the area per lipid taking it into account as explained in Section 3.3. The three replicas run at $E = 130$ mV/nm provided values of 0.520 nm^2 , 0.514 nm^2 and 0.550 nm^2 , and their discrepancy is due to the different shapes the membrane adopts during the run (because of that the values were computed over the last 10 ns only). Indeed, once a deformation appears it can be quickly enhanced by the electric field and small casual variations in its initial phases can bring to very different outcomes.

The electroporation threshold for the membrane alone was set at the higher value of 140 mV/nm: out of three simulations run with such value of the electric field, two resulted in disruption around time equal to 160 ns and –, while another presented a curved but still intact membrane after 200 ns.

This procedure of verifying the conditions promoting poration was performed on the 512 lipids patch. The poration mechanism, both with and without peptide, is preceded by a high deformation (curvature) of the membrane. However this does not happen with the large patch, for which simulations in presence of the peptide and an electric field of 130 mV/nm lead to poration, but this was initiated in a flatter membrane conformation (Figure 3.18(b)). To investigate whether this is due to the different size, or rather the force field used (54a7 versus 54a8), we run an additional control simulation of the 512 lipids patch with the 54a8 force field and the electric field set at the electroporation threshold of 140 mV/nm, as this would promote a quick insurgence of the curvature, if any. The patch developed a curved shape similar to the outcome from the corresponding 54a7 simulations. The membrane was not electroporated in the 200 ns run, as was observed in one out of the three replicas of the 54a7 simulations. Remarkably, the ApL values of these two simulations are quite similar (0.554 versus 0.556 nm^2 for 54a7 and 54a8 respectively, computed

taking into account the curvature), despite the differences outlined generally between the two force fields (see Chapter 4). However, from the other electroporation simulations it emerges that there is a high variability in them, so that the consistency of the two values might be fortuitous.

This proves that the size of the patch has the largest influence in the outcome of the simulations.

MEASURE CURVATURE?

Regardless the shape the membrane assumes before disruption, the peptide speeds up the collapse mechanism as the charged Arginine residues insert into the membrane core interacting with the negatively charged phosphate of the lipids, and promote the penetration of water molecule (Figure 3.18(c)). Additionally, the peptide enhances the invagination of the membrane and these two aspects together (stronger invagination and Arginine insertion) allow pore formation at a value of the electric field lower than the one necessary for electroporation.

This findings, together with the previous results, suggests that the rigidification process (measured by the reduction of the diffusion coefficient), regardless the membrane compactness (measured by the area per lipid), is a proxy for poration as it makes the lipids less able to accommodate perturbations and to seal a forming water channel.

3.5.2 Atomistic simulations of the mammalian model membrane

After the investigation of a model bacterial membrane, we focussed on a mammalian one, modelled as a DLPC patch. The investigation proceeded directly with simulations of a large patch

say 130 rupture apl (larger, error in dlpg) hbond with protein curvature in peptide one?

hb prot (590, 28) hb arg (461, 17) hb trp (57, 5)

0.495(1) diffusion

3.5.3 Coarse grain simulations of the buckyball on model membranes

Standard MARTINI simulations Coarse grain simulations allow to model the behaviour of the full capsule interacting with the membrane. At first, sim-

ulations with the standard MARTINI model were run, due to their higher computational speed (the full system comprising a 2880 lipids patch and the capsule measures approximately 30 nm along each side of the simulation box).

Coarse-grain simulations of the full buckyball on a DLPC:DLPG membrane (3:1 ratio) confirms the binding of the peptide on the latter, driven from charge-charge recognition: in both the replicas run, the peptide approaches the membrane after about 2 μ s, remaining bound up to the 10 μ s simulated. Post-binding, the capsule diffuses on the membrane and produces an increasingly high curvature on it, in a process which tends to maximize the contact area (see SI Movie –). No poration is observed, probably due to the force field characteristics which stabilise the structure of both the membrane and the peptide assembly. Additionally, longer time scales might be needed to observe poration. Instead, we opted for an electroporation experiment, in line with what done in the atomistic case (see next paragraph).

A quantitative analysis on the membrane patch shows that the lipids diffusion coefficient is decreased after the binding, by 9% for DLPC and by 15% for DLPG (see SI Table 3.5 and SI Figure ??). Moreover, as DLPG is negatively charged, it is recruited around the peptide and remain bound to it, diffusing much slower than the surrounding lipids.

prot diffusion on bact membrane after binding martini non polar counterions 28.4(2) martini polar mix 177(2) martini polar E20 53.3(6)

A control simulations with a pure DLPC membrane shows no binding of the peptide: the distance of the icosahedron from the membrane is on average 3 nm and never smaller than 1 nm. This does not exclude a binding on longer time scales, but hints at the fact that, not being driven by the opposite charge recognition, the peptide-membrane interaction is weaker and less likely to happen.

3.6 Rest

	MARTINI		Polar MARTINI			
	DLPC	DLPG	DLPC	DLPG		
Bacterial	Pre	310.0(4)	302.6(5)	E 0	307(2)	305(2)
	Post	267.6(2)	245.5(2)	E-20	264.7(7)	268.6(5)
	Protein, post					
Mamm.	DLPC		DLPC			
	Any	308.0(3)	E 0	349.5(3)		
			E-20	278.4(7)		

Table 3.5: Analysis of lipid diffusion in a coarse-grain simulation of the buckyball binding to a model bacterial membrane (DLPC:DLPG 3:1). Data for Replica 1, binding time $2.14 \mu\text{s}$. Protein lateral diffusion after binding to the membrane $16.50(2) \mu\text{m}^2/\text{s}$. ^a Number of lipids within 2 nm distance from the buckyball in the last $5 \mu\text{s}$ of simulations. ^b In a simplistic model in which the dumping of lipid lateral diffusion coefficient is due to some lipids bound to the buckyball, and thus they diffuse at its pace, while the other are free to move as before, it is valid the following equation: $D_{post} = D_{pre} \cdot (N_{lipid} - N_{bound}) + D_{protein} \cdot N_{bound}$.

Electroporation simulations					
	Lipids	Model	Time (ns)	Rep.	E (mV/nm)
Bacterial	512	GR (54a7)	400	1	0, 70, 120
	512	GR (54a7)	600	3	130
	512	GR (54a7)	150 ^{EP} , 154 ^{EP} , 200	3	140
	512	GR (54a8)	200	1	140
	740	GR (54a8)	400 (350, finish & redo prop)	1	0
	740	GR (54a8)	400	1	20
	720	MA	1000	1	0
Mamm.	2880	MA_P	500	1	20, 40
	748	GR (54a8)	400	1	0
	748	GR (54a8)	20 ^{EP} , 28 ^{EP} , 39 ^{EP}	3	130
	722 (54a8)	MA	1000	1	0
	2888	MA_P	500	1	20, 40

Table 3.6: Table of control simulations of membranes complexes. All run at 150 mM concentration of NaCl. Models: GR = united atom GROMOS, MA = coarse-grain MARTINI, MA_P = coarse-grain MARTINI with polar water. Superscript EP refers to the rupture of the membrane.

3.7 Additional material

No peptide		Bacterial	Mamm.
E = 0	512	54a7 (1)	54a7 no salt (1)
	740	54a8 (1)	54a8 (1R)
E = 130	512	54a7 (3)	–
	740	–	54a8 (1)
+ Bact/512/E-140: 54a8 (1R)			
Peptide		Bacterial	Mamm.
E = 0	512	54a7 (1)	–
	740	54a8 (1)	54a8 (1R)
E = 130	512	54a7 (3)	–
	740	54a8 (3)	54a8 (3)

Table 3.7: Table of the simulations run for different systems. Electric field expressed in mV/nm. For each cell, the syntax reports: Nr. of lipids *Force field* (Nr. of replicas).

Chapter 4

Lipid parametrisation

Appendices

A.1 On the derivation of the GP predictive distribution

This appendix gives a sketch of the procedure by which Eq. (??) is obtained, which substantially relies on the properties of multivariate Gaussian distributions. For full details on this one can consult the excellent Refs. [?] and [?].