Integrating Data-Driven Approach and Mechanistic Explainability to Unveil the Molecular Mechanism of TAT-RasGAP$_{317-326}$ in Anticancer and Antibacterial Domains.

Doctoral Dissertation submitted to the
Faculty of Informatics of the Università della Svizzera Italiana
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy

presented by

Filip Stojceski

under the supervision of

Prof. Andrea Danani

March 2025

Dissertation Committee

Prof. **Andrea Danani**, Research Advisor, USI-SUPSI, IDSIA, Switzerland
Prof. **Igor Pivkin**, Università della Svizzera Italiana USI, Switzerland
Prof. **Rolf Krause**, Università della Svizzera Italiana USI, Switzerland
Prof. **Marco Agostino Deriu**, Politecnico di Torino, Italy
Dr. **Athanasios Kalogeras**, Research Director, ISI, Research Centre ATHENA, Greece

Dissertation accepted on 06th May 2025

Research Advisor Prof. Andrea Danani	PhD Program Directors Prof. Walter Binder/ Prof. Silvia Santini,

I certify that except where due acknowledgement has been given, the work presented in this thesis is that of the author alone; the work has not been submitted previously, in whole or in part, to qualify for any other academic award; and the content of the thesis is the result of work which has been carried out since the official commencement date of the approved research program.

Filip Stojceski
Lugano, 06th May 2025

“My brain is only a receiver,
in the Universe there is a core
from which we obtain knowledge,
strength and inspiration.
I have not penetrated
into the secrets of this core,
but I know that it exists.”

- Nikola Tesla -

Abstract

This thesis explores the multifaceted realm of cell-penetrating peptides (CPPs), highlighting their transformative potential in biomedical applications through the integration of genetic algorithms (GA), machine learning (ML), and Molecular Dynamics (MD) simulations. These methodologies are employed to design *de novo* CPP sequences and explore their multifunctional capabilities, paving the way for innovative therapeutic strategies. CPPs, typically composed of 5 to 30 amino acids and characterized by a net positive charge at physiological pH, have emerged as powerful tools for facilitating efficient intracellular delivery of therapeutic agents. Their ability to traverse cellular membranes while maintaining low cytotoxicity makes them indispensable in drug delivery and targeted therapy.

Central to our exploration is the unveiling of TAT-RasGAP$_{317-326}$ bioactive capabilities, a chimeric CPP-based construct exhibiting both antitumoral and antibacterial activities. This novel peptide demonstrates its efficacy in selectively killing cancer cells through intricate mechanisms that do not conform to established programmed cell death pathways. The antitumoral effect starts with the translocation of the peptide into the cytosol, binding lipids like phosphatidylinositol bisphosphate (PIP2) and phosphatidylserine (PS), which are enriched in the inner leaflet of the membrane. These interactions induce membrane permeabilization and disruption, ultimately leading to cell death. This unique mode of action underscores the peptide's potential as a targeted anticancer agent with minimal off-target effects.

Beyond its antitumoral properties, TAT-RasGAP$_{317-326}$ also demonstrates promising antibacterial capabilities, offering a potential solution to the growing challenge of antibiotic resistance. MD simulations provided detailed insights into the peptide's interactions with the E. coli Bam protein, a critical component of the bacterial outer membrane assembly machinery. These computational investigations uncovered the mechanism by which TAT-RasGAP$_{317-326}$ inhibits the Bam protein machinery, providing valuable insights into its mode of action and highlighting its potential applications in addressing antibiotic-resistant bacterial strains, such as E. coli. Furthermore, the study proposes novel mutations within the E. coli Bam protein that could modulate its susceptibility to TAT-RasGAP$_{317-326}$, offering a pathway to enhance or fine-tune the peptide's antibacterial efficacy.

This holistic approach underscores the potential of TAT-RasGAP$_{317-326}$ as a versatile therapeutic agent, capable of addressing both cancer and bacterial infections, while contributing to the broader field of peptide-based drug development. Through this work, we aim to advance the frontiers of biomedical science, offering new tools to combat some of the most pressing health challenges of our time.

Acknowledgements

First and foremost, I extend my deepest gratitude to my scientific advisor, **Prof. Andrea Danani**, for granting me the extraordinary opportunity to pursue this PhD and for his unwavering guidance and steadfast support throughout this journey.

I express my profound gratitude to **Dr. Gianvito Grasso** for his exceptional mentorship and the wealth of knowledge he generously shared during my PhD. His insights into molecular modeling and peptide design were instrumental in navigating the complexities of this research, and his patience in fostering my growth as a scientist remains deeply appreciated.

To **Dr. Gabriele Maroni** and **Dr. Dario Piga**, I owe immense thanks for their indispensable contributions to the machine learning aspects of this thesis. Their expertise, collaborative spirit, and willingness to tackle challenges alongside me were pivotal in overcoming technical hurdles and advancing the computational framework of this work.

I am deeply grateful to **Prof. Christian Widmann** and **Dr. Nicolas Jacquier**, along with their dedicated research groups, for providing the experimental data that anchored this study in real-world biological relevance.

To my girlfriend, family and friends, whose encouragement, patience, and emotional support carried me through the highs and lows of this journey, thank you. Your belief in my work, even when deadlines loomed and challenges seemed insurmountable, provided the resilience needed to bring this thesis to an end.

Contents

1. Introduction

Cancer and antibiotic resistance represent two of the most pressing global health challenges of the 21st century, contributing to significant mortality, morbidity, and economic burden worldwide[1–4]. Cancer, a complex and heterogeneous disease characterized by the uncontrolled growth and spread of abnormal cells, remains a leading cause of death globally. According to the International Agency for Research on Cancer (IARC) of the World Health Organization (WHO), cancer accounted for approximately 10 million deaths in 2022, with lung, colorectal, stomach, liver, and breast cancers leading the statistics[2]. Based on population growth, aging, and lifestyle changes, and assuming stable overall cancer rates, the IARC estimates that more than 35 million new cancer cases will occur by 2050, leading to a 77% increase from the 20 million cases diagnosticated in 2022[2,5]. Therefore, the incidence of cancer is projected to sharply rise in the coming decades, emphasizing the urgent need for innovative and effective therapeutic approaches.

Current cancer therapies, including chemotherapy, radiotherapy, and immunotherapy, while effective in many cases, are fraught with challenges. Chemotherapy and radiotherapy often lack specificity, indiscriminately targeting both cancerous and healthy cells, leading to severe side effects and reduced quality of life for patients[6–9]. Additionally, the emergence of drug-resistant cancer cells further complicates treatment, rendering many traditional chemotherapeutic agents ineffective[10]. Immunotherapy, though revolutionary, faces limitations such as immune-related adverse events, high costs, and variable patient responses[11]. These challenges underscore the need for novel, targeted therapies that can overcome resistance, minimize side effects, and improve patient outcomes. Bioactive peptides (BPs) have emerged as a promising class of therapeutic agents in the fight against cancer[12,13]. These small, naturally occurring or synthetic peptides exhibit diverse biological activities, including the ability to selectively target cancer cells while sparing normal tissues[14]. As a result, BPs are gaining attention as a viable alternative to conventional cancer treatments.

Simultaneously, bacterial resistance to antibiotics has emerged as another critical global threat, with significant implications for public health[3,4]. Recent studies estimate that antibiotic-resistant infections caused at least 4.95 million deaths globally and 541,000 deaths in WHO European region in 2019[3,4], a number expected to rise to 8,2-10 million by 2050 if no significant interventions are made[15,16]. The six leading resistant pathogens, namely, *Escherichia coli*, *Staphylococcus aureus*, *Klebsiella pneumoniae*, *Streptococcus pneumoniae*, *Acinetobacter baumannii*, and *Pseudomonas aeruginosa* are of significant concern due to their widespread prevalence and remarkable capacity to resist conventional antibiotics[3,4]. The growing prevalence of multidrug-resistant bacteria undermines decades of progress in treating infectious diseases and significantly increases healthcare costs, hospital stays, and mortality rates.

The problem of bacterial resistance to antibiotics further highlights the need for innovative therapeutic solutions. The overuse and misuse of antibiotics in healthcare, agriculture, and livestock farming have accelerated the development of resistant strains, rendering many first-line antibiotics ineffective[17–19]. BPs also hold great promise in addressing this global crisis. Their unique modes of action, such as disrupting bacterial membranes or targeting intracellular processes, make it difficult for bacteria to develop resistance[20,21]. The ability of BPs to act on

multidrug-resistant bacteria positions them as a critical tool in the fight against antibiotic resistance.

In conclusion, the rising global burden of cancer and antibiotic resistance underscores the urgent need for novel therapeutic strategies. BPs, with their diverse mechanisms of action and potential to overcome current therapeutic limitations, offer a promising solution to these complex challenges. This thesis investigates the mechanism by which a novel BP addresses both cancer and antibiotic-resistant bacteria. The objective is to contribute to the development of targeted and effective treatments for these dual health crises.

2. Biological Background

2.1. Bioactive Peptides

BPs are short organic biopolymers formed by amino acids joined by covalent bonds known as peptide bonds[22]. Many of them are derived from food sources and have demonstrated positive health effects. While some BPs are naturally present in their free form, the majority are embedded within the structure of parent proteins and are predominantly released through enzymatic processes[23]. In this connection, many physiological and functional properties of proteins are thought to be associated with biologically active peptides[22,23]. They exhibit drug- or hormone-like activities and can be classified by their mode of action into antitumoral, antimicrobial, anti-thrombotic, antihypertensive, immunomodulatory, and antioxidative types[22,24] (Figure 1). In detail, the activity of the peptides is determined by their amino acid composition and sequence. Despite the correlation between peptide structure and functional properties is not fully understood, many BPs share some common characteristics, including a peptide length of 2–20 amino acid residues and the presence of hydrophobic amino acids along with proline, arginine, or lysine groups[22–24].

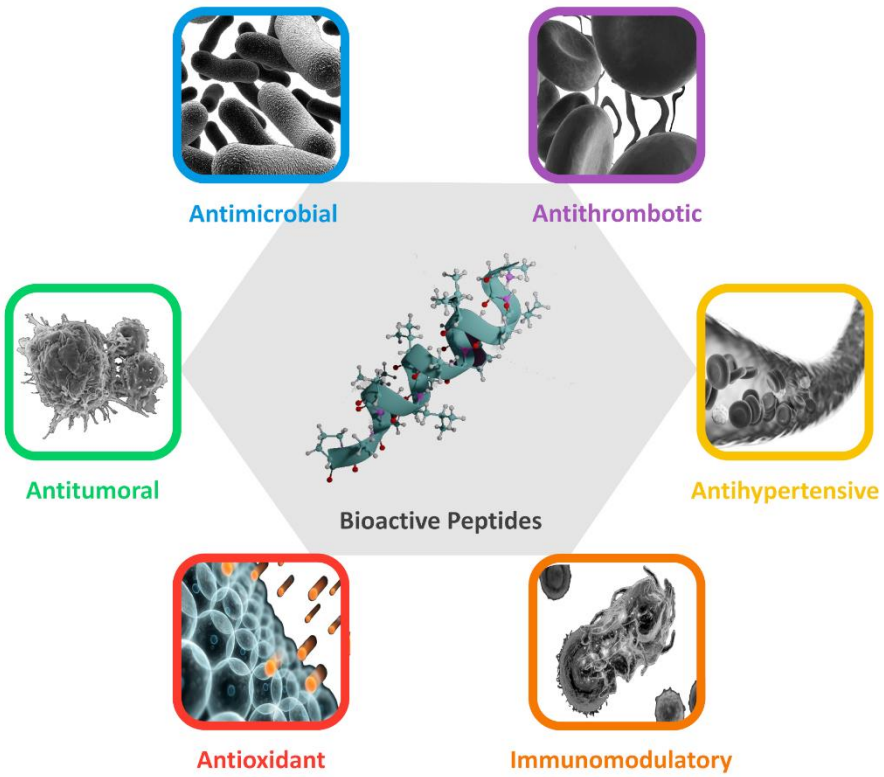


Figure 1: Qualitative illustration of the bioactive peptide's (BPs) possible mode of action, including antitumoral, antimicrobial, anti-thrombotic, antihypertensive, immunomodulatory, and antioxidative activities.

One of the most interesting feature of BPs is their ability to resist the action of digestive peptidases[25,26]. A recent study revealed that peptides resistant to in vitro gastrointestinal

digestion tend to have lower molecular size and hydrophobicity, more positive net charge at intestinal pH, the absence of C-terminal leucine, branched-chain aliphatic N-terminal residues, higher histidine and proline, compared to unstable peptides[25]. This resistance can be further enhanced by using D-peptides instead of L-peptides, as D-peptides are less susceptible to enzymatic degradation in the intestinal lumen, which improves their stability and bioavailability[27]. A typical example of D-peptides with remarkable properties are cell-penetrating peptides (CPPs)[28]. Furthermore, their D-configuration not only enhances their resistance to enzymatic degradation but also extends their stability in biological environments, making them highly effective tools in drug delivery and biomedical applications. For instance, CPPs are often conjugated with orally administered insulin to improve its absorption and bioavailability[29]. The ability of CPPs to penetrate cell membranes facilitates the uptake of insulin across the intestinal epithelium, enhancing its therapeutic efficacy and enabling more efficient delivery compared to traditional oral insulin formulations[30]. It is evident that BPs are integral in augmenting numerous physiological functions and demonstrate significant therapeutic potential, especially when their stability and bioavailability are enhanced. Among these, CPPs stand out as a powerful subclass, offering remarkable properties for drug delivery and biomedical applications.

2.2. Cell-Penetrating Peptides

In the modern era of nanotechnology and medicine, macromolecular entities such as peptides, proteins, and nucleic acids represent a promising prospect to overcome the limitations of conventional therapeutics[31]. However, the presence of cell and tissue barriers, together with the low membrane translocation of macromolecules, often hampers systemic drug distribution. Within this context, several peptides with membrane penetrating function[32–34] could support the intracellular delivery of hydrophilic macromolecules to eukaryotic cells. A plethora of small peptides with membrane translocation capacities has been identified in literature[35–38]. They are termed as CPPs (Figure 2), which are short peptides that can successfully deliver proteins, peptides, nanoparticles, small drugs, siRNAs, and DNA across the lipid bilayer of the cell membrane into different cell types[39]. From a physicochemical standpoint, CPPs typically consist of 4–40 amino acids and possess cationic and amphipathic character[40]. Most CPPs are positively charged at physiological pH, with arginine and lysine residues playing a key role in facilitating membrane penetration and the internalization process. From a structural perspective, certain CPPs exhibit an α-helix or β-sheet configuration, while others do not possess a defined conformation[40]. The considerable heterogeneity of these peptides makes it challenging to establish a precise definition of CPPs within a structural and physicochemical framework.

CPPs have demonstrated their usefulness in several fields such as cancer treatment[41–43], antibacterial therapy[44–46], drug delivery [47–50], *in vivo* imaging and diagnostic [51–54], gene therapy[55–59], and radiotherapy[60–62]. CPPs enter cells in a noninvasive manner, preserving the integrity of the cellular membranes and are considered highly efficient and safe[32,63]. Thus, they provide new avenues for research and applications in life sciences[64]. The mode of CPP cellular entry is still debated and, until recently, no proteins had been identified that regulate this process. Currently, the most accredited hypotheses suggest that CPPs enter cells through two non-mutually

exclusive mechanisms[65,66]: endocytosis and direct translocation. The CPP entry process starts after the initial electrostatic interactions between the positively charged amino acids of the CPPs and the negatively charged components of the cell membrane[31,32,39,67–70]. Interaction with acid sphingomyelinase[71] and glycosaminoglycans[72,73], local membrane deformation[74], as well as calcium fluxes[75] have been suggested to play a role in CPP internalization. However, the precise entry mechanisms are still debated and not fully understood at the molecular level.

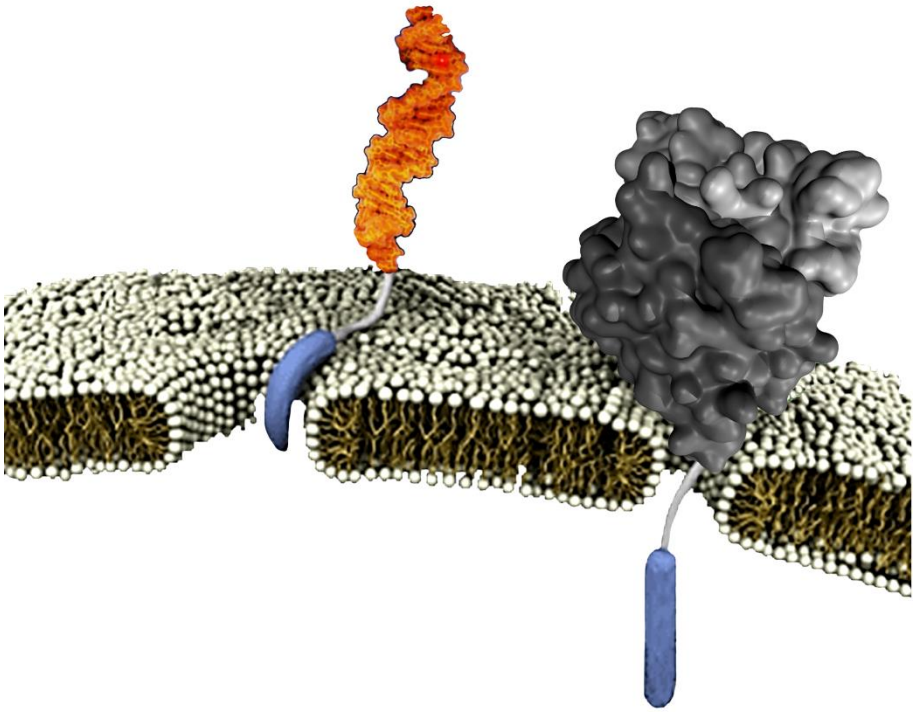


Figure 2: Qualitative representation cell-penetrating peptides (CPPs) crossing the plasmatic membrane with linked RNA fragment and protein cargos.

Rationally designed synthetic or naturally derived CPPs may have intrinsic biological functions, including antimicrobial, anticancer, anti-inflammatory, and immunomodulatory properties[76]. For instance, CPPs with innate immunomodulatory properties can be used to harmonize various pathological conditions[76]. Another crucial challenge is the development of novel CPPs with antimicrobial peptide (AMP) functions, which is urgently needed to combat the rise of drug-resistant microbial strains[77]. In this sense, the better comprehension of the CPPs penetration mechanisms could also improve their bioactive abilities, as for example, antibacterial and antitumoral effects[78,79]. Understanding the intricate relationship between the antitumoral and antibacterial capabilities of CPP-based constructs and their membrane-penetrating function presents an exciting avenue for future research and development[80,81]. Therefore, investigating the synergistic effects of CPPs, CPP-based constructs, and CPP-conjugated sequences in the fields of antibiotic development and cancer treatment could lead to the discovery of innovative therapeutic strategies.

2.3. Anti-Microbial and Anti-Cancer Peptides

The exploration of the triad of antitumoral efficacy, antibacterial activity, and membrane-penetrating proficiency opens new horizons for the development of multifaceted therapy approaches for the BPs[78,79]. In this connection, the continual advancements in biology and biomedicine have unveiled a plethora of BPs derived from animal, plant, and micro-organism origins[82]. These peptides demonstrate noteworthy properties, including immunomodulatory, anti-inflammatory, antimicrobial, and anticancer activities[83,84]. Their multifunctionality extends across various biological activities, bridging their immunomodulatory and anti-inflammatory roles with potent antimicrobial and anticancer effects. In detail, multiple experimental investigations on the antimicrobial capabilities of several BPs have demonstrated their efficacy against a spectrum of bacteria, fungi, and viruses[85–87]. AMPs typically comprise fewer than 100 amino acids, with the majority of AMPs falling within the range of 20 to 30 amino acids[88]. At physiological pH, these peptides commonly have amphipathic characteristics, concurrently displaying cationic charge sections balanced with hydrophobic sections[83,89].

These peptides offer several advantages in comparison to traditional antibiotics, encompassing their multi-hit mode of action, potency, and rapid onset of action with low levels of resistance[90]. AMPs primarily target highly conserved structures, such as the phospholipid membrane through pore formation or essential components like peptidoglycans in Gram-negative and Gram-positive bacteria, and glucan in fungal cell walls[87]. Within this context, a recent study identified a novel subset of AMPs with significant antifungal properties[91]. These peptides have demonstrated a strong ability to inhibit the metabolic activity and growth of several medically significant fungal species. Moreover, AMPs with enhanced antifungal properties show remarkable efficacy against biofilms, effectively destabilizing microbial communities, which usually are composed of both bacteria and fungi[87,91]. Beyond these actions, AMPs can also act intracellularly, disrupting processes such as protein biosynthesis or DNA replication. Their intracellular capabilities become even more pronounced during viral infections, where they can interfere with multiple stages of the viral life cycle, from viral receptor-cell interactions to the replication process[87,92]. A qualitative representation of AMPs mechanisms of action is shown in Figure 3. In recent literature, it was introduced new approach for enhancing the antibiotic activity of AMPs by conjugation with a cationic CPP[93,94]. Therefore, the combination of AMPs and CPPs can yield bioactive agents that benefit from the combined advantages of each peptide. CPPs not only facilitate access to plasmatic membranes or previously inaccessible locations but also may interact synergistically with bacterial membranes to enhance bactericidal efficacy of the AMPs[95].

From another point of view, several scientific investigations have consistently substantiated the antitumor efficacy inherent in specific AMPs, thus categorizing them as anticancer peptides (ACPs)[96,97]. ACPs are low molecular weight peptides ranging between 10-60 amino acids with a α-helical or β-sheet structures, or a combination of both[84,98]. Many studies have demonstrated that certain cationic ACPs possess a selective toxicity towards cancer cells, causing minimal or no harm to normal cells[96,99]. Ideally, anticancer therapies should be designed to effectively target and destroy a wide variety of cancer cell types while minimizing collateral damage to healthy cells. One of the key distinguishing features between cancerous and healthy cells lies in the

composition and structure of their cell membranes, which presents an opportunity for selective therapeutic targeting[100]. Notably, ACPs can engage in electrostatic interactions with the negatively charged cell membrane of cancer cells, thereby inducing necrosis via the destruction of the plasmatic membrane[101,102]. This receptor-independent mechanism of action offers the potential to overcome the challenge of drug resistance, a common obstacle associated with many modern anticancer therapies[103]. Indeed, one significant advantage of ACPs lies in their reduced likelihood to induce drug resistance, a characteristic that distinguish them from conventional chemotherapeutic agents[98,104]. In addition, ACPs can also exert their anticancer activity by other well-known mechanisms, including promoting pore formation, induction of apoptosis, inhibiting the angiogenesis pathway, or recruiting and activating immune cells[102,105]. A qualitative representation of ACPs mechanisms of action is shown in Figure 3.

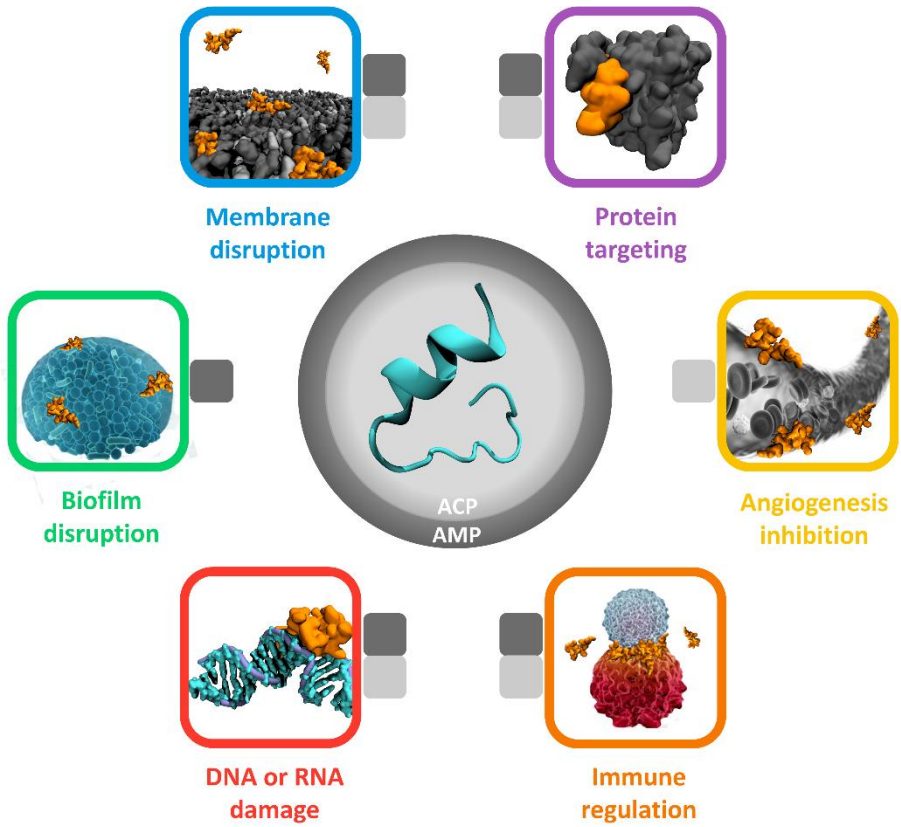


Figure 3: Schematic representation of the ACPs (light grey) and AMPs (dark grey) mechanisms of action. It is important to note that both peptides share similar mechanisms of action in killing bacteria or cancer cells, with the key differences lying in their ability to disrupt biofilms and inhibit angiogenesis.

However, the efficacy of ACPs alone can be limited in some cases due to challenges such as enzymatic degradation and low bioavailability. To address these issues, the development of chimeric peptides, which combine diverse functional characteristics into a single conjugated BP, holds significant promise. Such innovative approaches could enhance stability, target specificity, and therapeutic efficiency, paving the way for more effective cancer treatments in the future. An excellent example of the potential of chimeric peptides in cancer therapy is demonstrated by a study on magainin II (MG2), an AMP with inherent antitumoral capabilities[106]. While MG2

exhibited cytotoxic effects on tumor cells, its activity was limited to high concentrations due to inefficiencies in cell membrane binding and intracellular entry. To overcome these limitations, MG2 was conjugated to a CPP moiety (Antp), resulting in the chimeric peptide MG2A[106]. The fusion increased the antitumoral activity of MG2, with MG2A showing IC50 values for tumor cells that were at least 30 times lower than those of the unconjugated peptide. This study highlights the promising therapeutic potential of combining AMP and CPP functionalities into a single chimeric peptide for targeted cancer therapy. Another compelling example of a chimeric peptide with multivalent therapeutic potential is TAT-RasGAP$_{317\text{-}326}$[107].

2.4. TAT-RasGAP$_{317\text{-}326}$

In the exploration of CPPs conjugated multi-bioactive peptides, the spotlight falls on the promising therapeutic candidate: TAT-RasGAP$_{317\text{-}326}$[107]. This novel peptide combines remarkable membrane-penetrating capabilities with a dual impact, showcasing both antitumoral and antibacterial effects. In detail, TAT-RasGAP$_{317\text{-}326}$ is a CPP composed of a cell-penetrating sequence (G^{48}RKKRRQRRR^{57}) derived from the human immunodeficiency virus (HIV) trans-activator of transcription (TAT) protein, and a ten amino acid sequence (W^{317}MWVTNLRTD^{326}) derived from the Src homology 3 (SH3) domain of p120 RasGAP (Ras GTPase activating protein)[108]. This chimeric peptide has several anticancer properties, including inhibition of metastatic progression and tumor cell sensitization to anticancer therapies[109–111] and a broad antimicrobial activities toward both gram-negative and gram-positive bacteria[112,113].

The first reported anticancer activity of TAT-RasGAP$_{317\text{-}326}$ was its ability to sensitize tumor cells, but not normal cells, to genotoxin-induced death[108] and to radiotherapy[114]. By leveraging the TAT CPP moiety, TAT- RasGAP$_{317\text{-}326}$ can efficiently penetrate cell membranes and deliver its bioactive component into tumor cells. This chimeric design not only enhances its intracellular delivery but also significantly boosts its therapeutic efficacy, making it a powerful candidate for targeted cancer treatments. Based on these findings, the peptide was tested in preclinical animal models and was shown to sensitize tumor xenografts in mice to cisplatin, doxorubicin, and ionizing radiation with no apparent toxicity to healthy tissues[114,115]. TAT-RasGAP$_{317\text{-}326}$ can also directly kill a subset of cancer cell lines in a manner that is distinct from apoptosis, necroptosis, parthanatos, pyroptosis, and autophagy[107]. Therefore, this peptide exhibits cytotoxic properties that may pose significant challenges for cancer cells in developing resistance via alterations within established regulated cell death pathways. Remarkably, altering just a single amino acid in the RasGAP moiety (W317A) completely abolishes its anticancer and antimicrobial efficacy, further underscoring the vital role of its precise peptide sequence[107,112,116].

To further expand on this note, in vitro experiments demonstrated that TAT- RasGAP$_{317\text{-}326}$, but not its mutated or truncated variants, effectively killed a range of bacteria, including *Escherichia coli*, *Acinetobacter baumannii*, *Staphylococcus aureus*, and *Pseudomonas aeruginosa*[112,113]. Subsequent in vivo studies showed that TAT-RasGAP$_{317–326}$ effectively protects mice from lethal E. coli-induced peritonitis when administered locally at the first stages of infection[112]. Beyond its bactericidal effects, TAT-RasGAP$_{317–326}$ also targets bacterial communities, demonstrating potent activity against biofilm formation of *A. baumannii* and *P. aeruginosa*[113]. It is important

to mention that TAT-RasGAP$_{317–326}$ is more effective than conventional antibiotics in eradicating biofilms and it also inhibits biofilm formation by *S. aureus*[113]. These findings highlight the potential of TAT-RasGAP$_{317–326}$ as a treatment for biofilms and support the need for further research into the development of AMPs for addressing biofilm-associated infections.
Building on these findings, the multifunctional properties of TAT-RasGAP$_{317-326}$ further underscore its promise in treating complex infections and cancer. By combining membrane penetration, antitumoral activity, and antibacterial effects, this peptide demonstrates a unique ability to target multiple therapeutic areas simultaneously. The seamless integration of these three effects in TAT-RasGAP$_{317-326}$ not only showcases its potential as a dual-action therapeutic agent but also opens doors to innovative strategies for addressing complex medical challenges, paving the way for advanced and targeted anticancer and antibacterial therapies.

2.5. TAT-RasGAP$_{317-326}$ Targets in Cancer Cell

TAT-RasGAP$_{317-326}$ appears to possess killing properties that cancer cells may find challenging to counteract by developing resistance through modifications within established regulated cell death pathways. The mechanisms by which TAT-RasGAP$_{317-326}$ induces cell death remain not fully understood to this date. Recent studies have shown that TAT-RasGAP$_{317-326}$ enters the cytoplasm by directly translocating across the plasma membrane in a manner dependent on membrane potential[117]. The same study demonstrated that depolarization reduces the peptide's uptake by approximately 100-fold and completely inhibits its ability to kill cells[117]. This means that the access into the cytoplasm is needed to kill cells.
Preliminary in vitro analysis (Figure 4) indicates that TAT-RasGAP$_{317-326}$ and control peptides do not bind to liposomes composed of lipids commonly found on both sides of the plasma membrane, including phosphatidylcholine (PC), phosphatidylethanolamine (PE), or cholesterol[118]. Conversely, TAT-RasGAP$_{317-326}$ exhibited stronger binding to inner-leaflet-enriched phosphatidylinositol (4,5) bisphosphate (PI[4,5]P2), phosphatidylserine (PS), and phosphatidic acid (PA) compared to the W317A mutant[118]. The W317A mutant exhibits significantly reduced interactions with specific phospholipids compared to the wild-type peptide, indicating that the RasGAP moiety of TAT-RasGAP$_{317-326}$ plays a greater role than the TAT segment in mediating peptide-lipid binding[118].

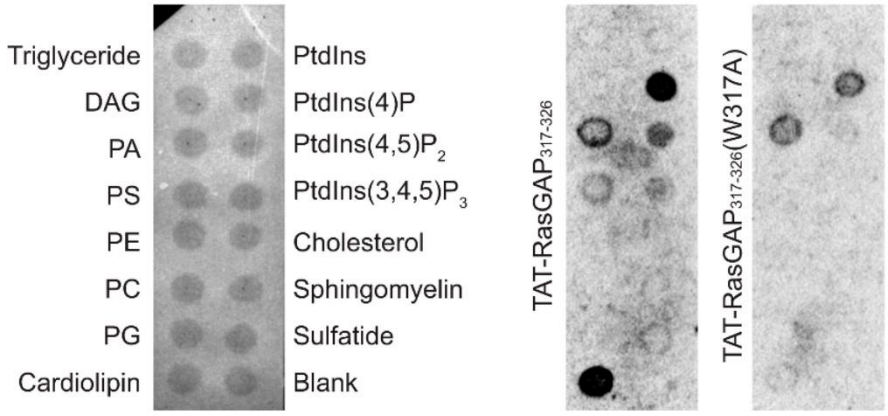


Figure 4: Membranes spotted with the indicated lipids were incubated with 0.5 mg/mL TAT-RasGAP$_{317-326}$ and W317A mutant, and TAT for 1 h, and the interaction with the lipid was revealed by immunoblotting using an anti-TAT antibody. Blots are representative of two independent experiments[118].

The reasons behind the distinct binding behavior of the wild-type peptide and the W317A mutant remain still unknown. However, understanding at molecular level this single-point mutation can provide crucial insights into the mechanism by which TAT-RasGAP$_{317-326}$ induces cell death. By further investigating how this mutation alters peptide-lipids interactions and downstream effects, it may be possible to unravel the pathway involved in the peptide's cytotoxic activity.

2.6. TAT-RasGAP$_{317-326}$ Targets in Bacteria

TAT-RasGAP$_{317-326}$ is a peptide known for its ability to kill various bacterial species, including *Escherichia coli* (*E. coli*)[119]. Recent research has elucidated that this peptide exerts its bactericidal effects by targeting the β-barrel assembly machinery (Bam) complex, specifically the BamA subunit, which is essential for the proper assembly and insertion of outer membrane proteins in Gram-negative bacteria[120]. Disruption of BamA function compromises the structural stability of the outer membrane, rendering *E. coli* vulnerable to environmental stressors and antimicrobial agents[121]. The BamA protein features extracellular loops that are critical for its function and interaction with antimicrobial agents. Notably, preliminary analyses (Figure 5) have demonstrated that alterations in these loops, particularly the deletion of negatively charged residues, confer resistance to TAT-RasGAP$_{317-326}$ in *E. coli*. In detail, the minimal inhibitory concentration (MIC) of the peptide for the bacterial strains carrying the D498N, D498K, and D497N D498K D500N mutations nearly doubled, conferring resistance to these *E. coli* strains[120].

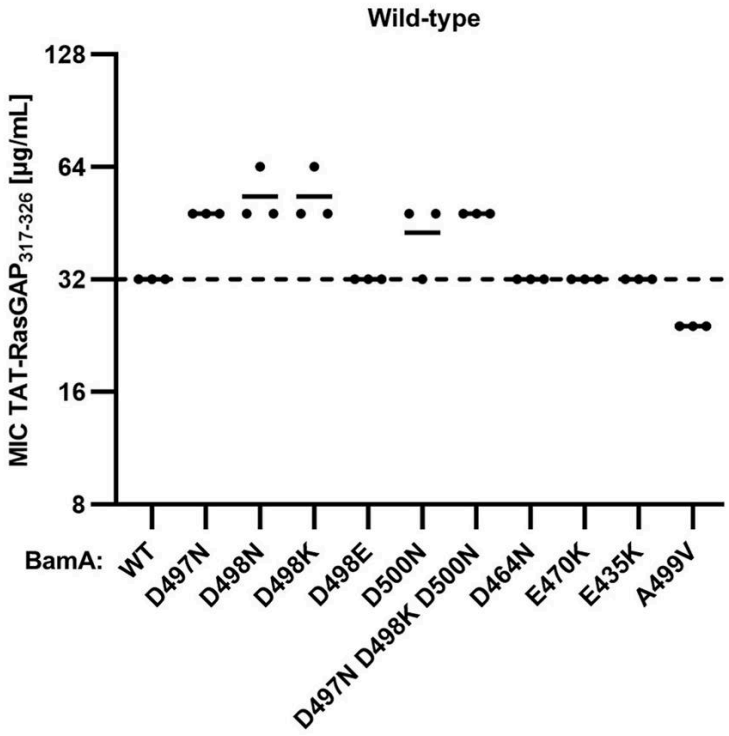


Figure 5: Removal of negative charges at the BamA's surface, but not on membrane-embedded domains causes an increased resistance to TAT-RasGAP317-326. Indicated point mutants were designed by CRMAGE and the minimal inhibitory concentrations (MIC) of TAT-RasGAP$_{317-326}$ towards them was measured in triplicate. The dashed line corresponds to the average MIC of the wild type. Instead, continuous lines represent the average MIC for the three replicates[120].

The molecular mechanisms responsible for the inhibition of BamA by TAT-RasGAP$_{317-326}$, as well as the resistance conferred by the loss of specific negative charges, remain still unknown. Understanding the specific interactions between TAT-RasGAP$_{317-326}$ and the BamA subunit not only sheds light on the peptide's mechanism of action but also provides valuable insights into potential strategies for overcoming gram-negative bacterial resistance. Investigations at a molecular level into these interactions could inform the design of novel antimicrobial agents targeting the Bam complex, therefore enhancing the efficacy of treatments against Gram-negative bacterial infections.

3. Aim of the Work

In the expansive landscape of BPs, CPPs stand out as pivotal players with multifaceted roles, particularly in the realms of antibacterial and antitumoral activities. Their intrinsic ability to traverse cellular membranes positions CPPs as an indispensable tool in drug delivery and therapeutic development. Understanding the correlation between the penetrative ability and the CPPs sequence represents a challenging task.

Recognizing the significance of studying CPPs sequences, this thesis aims to integrate machine learning (ML) and genetic algorithms (GA) to design novel CPP sequences. Central to this endeavor is the focus on the *de novo* design of CPPs, utilizing innovative computational approaches to transform non-CPP sequences into novel peptides with augmented cell-penetrating capabilities. The overarching goal is to harness the potential of CPPs, which exhibit diverse biological activities, including antitumoral and antibacterial effects[95,122]. By leveraging ML algorithms, we aim to identify sequence motifs and structural features critical for CPP functionality, facilitating the rational design of peptides with improved cell-penetrating capabilities.

Alongside the development of the CPP optimization algorithm, this thesis also seeks to perform an in-depth molecular dynamics (MD) analysis of the anticancer potential of the chimeric CPP TAT-RasGAP$_{317-326}$. This investigation aims to elucidate the dynamic behavior of TAT-RasGAP$_{317-326}$ within a cellular-like plasmatic membrane, shedding light on its interactions, stability, and structural changes over time. A key focus of this study is to explore the antitumoral mechanism of TAT-RasGAP$_{317-326}$ by examining the effects of a specific mutation, W317A, on its structure and behavior. By comparing the native peptide with its W317A mutant, the analysis will provide critical insights into the role of this residue in mediating the peptide's anticancer activity. Understanding the interaction of TAT-RasGAP$_{317-326}$ and its W317A variant with cellular-like membranes at a molecular level is essential for unraveling its antitumoral mechanisms and advancing its potential therapeutic applications.

Furthermore, the final branch of this research involves an in-depth MD exploration of TAT-RasGAP$_{317-326}$ concerning its antibacterial activity against *E. Coli*, specifically through interactions with the BamA protein subunit. This investigation aims to uncover the molecular reasons underpinning the antibacterial properties of TAT-RasGAP$_{317-326}$ and its potential as a dual-action therapeutic agent. Understanding the intricate mechanism of interactions with the BamA protein subunit not only provides insights into the antibacterial efficacy of TAT-RasGAP$_{317-326}$ but also holds the promise of paving the way to the development of new antibacterial therapeutics.

The underlying rationale for this multifaceted thesis stems from the versatility of CPPs, which play a pivotal role in various biological activities, including antitumoral and antibacterial activities. CPPs possess inherent cell-penetrating capabilities, making them crucial components for drug delivery and therapeutic development. The thesis aims to bridge the gap between ML-driven peptide design and the MD exploration of a promising bioactive peptide, such as TAT-RasGAP$_{317-326}$. By doing so, it aspires to contribute to the development of novel CPPs with multi-bioactive functionalities, advancing the understanding and potential applications of these peptides in biomedicine.

4. Material and Methods

4.1. Introduction to Molecular Dynamics

MD broadly refers to computational simulation techniques that study the evolution of physical and chemical systems at atomistic and molecular levels by integrating Newton's equations of motion[123]. These methods enable the exploration of particle trajectories and interactions, which define the type of MD approach. In detail, MD can be categorized into three main types:

- Classical molecular dynamics[124].
- Semi-classical molecular dynamics.
- Quantum molecular dynamics (e.g., ab initio or Car-Parrinello methods[125]).

Quantum molecular dynamics, which solves the Schrödinger equation, provides the most accurate representation of biological systems by incorporating quantum mechanical principles. However, the Schrödinger equation can only be solved exactly for a limited number of particles. Finding approximate solutions for complex systems requires computational resources that are often prohibitively high.
As molecular systems studied through MD become increasingly complex, encompassing larger numbers of atoms and interactions, classical MD emerges as a highly versatile and computationally efficient approach. Classical MD extends classical molecular mechanics (MM) to dynamic simulations by applying Newton's laws of motion. In this framework, molecules are computationally modeled using parameters that represent their physical properties in accordance with classical physics.
This approach allows for the simulation of systems containing hundreds of thousands of atoms, enabling the investigation of a wide range of molecular mechanisms. With classical MD tools, researchers can study phenomena such as protein folding and unfolding, receptor-ligand docking, drug delivery mechanisms, protein free energy landscapes, polymer aggregation, protein-membrane interaction, multiscale modeling, and much more. As a result, classical MD remains an invaluable tool for exploring the complex behavior of biomolecular and chemical systems at an atomistic level.

4.2. Molecular mechanics

Molecular mechanics describes molecular systems using classical mechanics. The potential energy of all particles is estimated as a function of the nuclear coordinates, if the Born–Oppenheimer approximation is valid. Functional form and parameter sets used to calculate the potential energy for each type of the system’s atom are collected in Force Fields. In general, MM is based on a simple parametrization model: atoms are represented by balls, which have mass as the real element mass, and bonds are represented by springs (using the Hook law) with an equilibrium distance equal to calculated bond length or equal to experimental data.

Molecular mechanics (MM) describes molecular systems using classical mechanics, offering a simplified yet effective framework to study the structure and behavior of molecules. In MM, the potential energy of all particles is calculated as a function of their nuclear coordinates, if the Born-Oppenheimer approximation is valid. Furthermore, the potential energy function, which governs these calculations, is a mathematical expression that accounts for various energy contributions, including bond stretching, angle bending, torsional rotations, and nonbonded interactions, including van der Waals forces and electrostatic interactions.

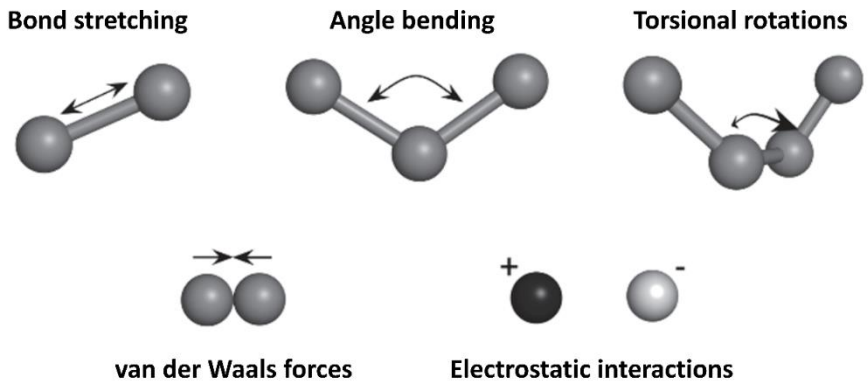


Figure 6: Qualitative representation of the energetic contributions in the potential energy function, including bond stretching, angle bending, torsional rotations, and nonbonded interactions, including van der Waals forces and electrostatic interactions.

The functional form of the potential energy equation and the associated parameter sets required for its computation are collected in the so-called force fields. Force fields are tailored to specific types of systems, such as proteins, nucleic acids, lipids, or small organic molecules, and contain pre-defined parameters for atomic masses, bond lengths, bond angles, and other molecular features. These parameters are derived from experimental data or quantum mechanical calculations, ensuring an accurate representation of molecular behavior within the classical mechanics framework.

In general, MM employs a straightforward parametrization model, where atoms are depicted as rigid spheres with masses corresponding to their actual atomic masses. Bonds between atoms are represented as springs (Hooke's Law), with an equilibrium distance that matches either calculated bond lengths or experimental measurements. Beyond bonded interactions, MM also accounts for nonbonded interactions, which play a critical role in molecular behavior. These include van der Waals forces, which describe dispersion and repulsion effects between atoms, and Coulombic forces, which account for electrostatic interactions between charged or partially charged particles. The sum of these energy components provides the total potential energy of the system, which determines the molecular conformation and interactions under study.

This approach makes MM a powerful tool for exploring molecular systems, from small organic compounds to large biomolecules, enabling simulations of equilibrium states, conformational changes, and intermolecular interactions with high computational efficiency.

4.2.1. The potential energy function

As mentioned previously, potential energy function is defined in MM to modulate atom interactions. The potential energy is the sum of two macro-energy contributions, the potential energy of bonded and non-bonded interactions:

$$E(r^N) = E_{bonded}(r^N) + E_{non-bonded}(r^N) \quad (2.0)$$

where $E(r^N)$ denotes the potential energy as a function of the position (r) of the N particles. These 2 categories are further subdivided into:

$$E_{bonded}(r^N) = E_{bonds}(r^N) + E_{angles}(r^N) + E_{dihedrals}(r^N) \quad (2.1)$$

$$E_{non-bonded}(r^N) = E_{Van\ der\ Waals}(r^N) + E_{electrostatic}(r^N) \quad (2.2)$$

The covalent bond stretching contribution of expression (2.1) can be expressed by a harmonic potential (Figure 7-A):

$$E_{bonds}(r_{ij}) = \frac{1}{2}k_{ij}(r_{ij} - r_{0,ij})^2 \quad (2.3)$$

where k_{ij} represents the bond stiffness and $r_{0,ij}$ represents the equilibrium distance.
The angle vibration between 3 atoms i-j-k of expression (2.1) is also represented by a harmonic potential (Figure 7-B):

$$E_{angles}(\theta_{ijk}) = \frac{1}{2}k_{ijk}(\theta_{ijk} - \theta_{0,ijk})^2 \quad (2.4)$$

where $\theta_{0,ijk}$ represents the angle equilibrium distance between the three atoms and k_{ijk} is the angle stiffness.
The third term of the expression (2.1) inserts an energy contribution that takes into account the torsional rotations between 4 atoms (Figure 7-C):

$$E_{dihedrals}(\phi_{ijkl}) = k_{ijkl}(1 + \cos(n\phi_{ijkl} - \phi_{0,ijkl})) \quad (2.5)$$

where k_{ijkl} is the stiffness of the dihedral angle, n is called multiplicity, which gives the number of minimum points in the function, as the angle can rotate through 360°, and $\phi_{0,ijkl}$ determines where the torsional phase crosses its minimum value.

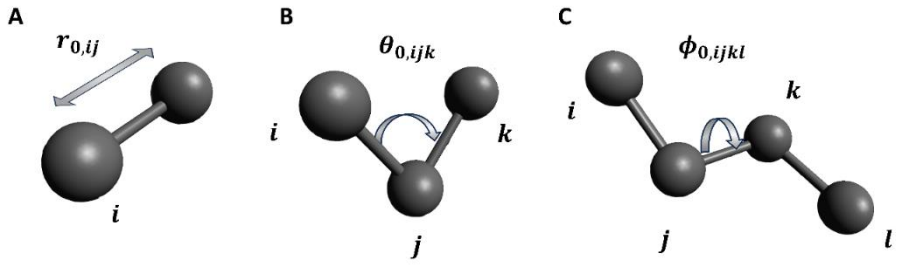


Figure 7: Qualitative representation of bonded parameters: A) bond between atoms i-j with equilibrium distance of $r_{0,ij}$; B) angle between atoms i-j-k with equilibrium angle of $\theta_{0,ijk}$; C) torsional angle between atoms i-j-k-l with equilibrium torsion angle of $\phi_{0,ijkl}$.

Non-bonded interactions are modeled as functions of the inverse of the distance between atoms, meaning they are not dependent on any specific bonding relationships between the atoms. In various force fields, non-bonded interactions are broadly categorized into two main types: van der Waals interactions and electrostatic interactions (Figure 8).

Highly electronegative atoms can draw the electronic cloud toward themselves, creating an uneven distribution of charge within the molecule. This results in regions of partial positive and negative charges, which can be represented computationally by assigning atomistic partial charges to the respective atoms. Electrostatic interactions, both within a single molecule (intramolecular) and between separate molecules (intermolecular), can be calculated using Coulomb's law (2.6). This fundamental law describes the force between two charges as directly proportional to the product of their magnitudes and inversely proportional to the square of the distance between them:

$$E_{electrostaic}(r_{ij}) = \frac{q_i q_j}{4\pi\varepsilon_0 r_{ij}} \tag{2.6}$$

where ε_0 is the electrical permittivity in the vacuum, r_{ij} is the distance between the charge q_i of the atom i and the charge q_j of the atom j.

In addition to electrostatic interactions, van der Waals interactions also play a significant role in defining the chemical and physical properties of materials. Although relatively weak compared to covalent bonds or ionic interactions, van der Waals forces are crucial in determining the stability and behavior of molecular systems, particularly in non-covalent assemblies. Van der Waals forces are broadly categorized into three types based on the underlying mechanisms, including the London dispersion, Keesom and Debye forces.

Despite their relative weakness in comparison to the electrostatic interactions, van der Waals interactions collectively have a significant impact, particularly in systems with many interacting particles, such as molecular crystals, biomolecules, and polymers. The representation of Van der Waals forces in MD is expressed by the Lennard-Jones function[126]:

$$E_{Van\ der\ Waals}(r_{ij}) = 4\varepsilon\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right] \tag{2.7}$$

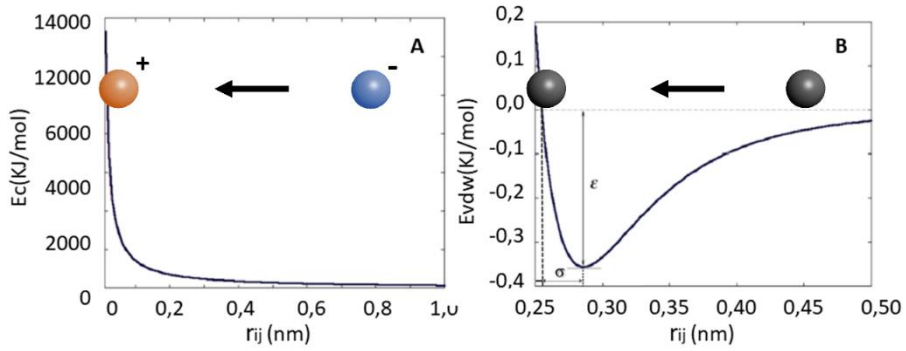


Figure 8: Representation of non-bonded interactions: A) Represents the Coulomb energy in function of distance r_{ij} between two atoms with the same charge. B) Represents the Lennard-Jones energy as a function of distance r_{ij}; where σ represents the collision diameter and the ε represents the well depth of the potential energy.

Van der Waals interactions, along with Coulomb interactions and covalent forces, are essential components in molecular simulations, ensuring a comprehensive representation of molecular systems and their behavior.

4.2.2. Truncating the potential: cut-off radius

The calculation of non-bonded interactions, such as van der Waals and electrostatic forces, is computationally demanding, especially in large molecular systems where the number of pairwise interactions scales quadratically with the number of particles. To optimize these calculations, various methods are employed, including the use of cut-off radii and advanced algorithms like Particle-Mesh Ewald (PME) for electrostatic interactions.

Applying a cut-off radius simplifies the computation of non-bonded interactions by truncating them beyond a predefined distance. Within this radius, interactions are explicitly calculated, while those beyond are ignored, significantly reducing the computational effort. The cut-off radius must be carefully chosen to balance computational efficiency and accuracy, as too small a radius can omit important long-range interactions, while a very large radius negates the computational advantage.

For long-range electrostatic interactions, truncation alone is insufficient because electrostatic forces decay slowly with distance $\frac{1}{r}$ and have a substantial influence even at large separation. The PME method offers an efficient solution by combining real-space and reciprocal-space calculations.

In the PME method[127]:

1. The system's charges are distributed onto a 3D grid with specified spacing, approximating the continuous charge distribution.
2. The charge distribution on the grid is transformed into the Fourier domain using a 3D Fast Fourier Transform (FFT).
3. Electrostatic potentials are calculated in the reciprocal space by solving the Poisson equation.
4. The results are transformed back into real space using an inverse FFT.

5. The forces acting on individual atoms are obtained through interpolation of the grid-based potentials.

This approach separates the electrostatic interactions into short-range and long-range components. Short-range interactions are computed directly in real space using a cut-off radius, while the long-range interactions are efficiently handled in reciprocal space. PME scales computationally with $N\log(N)$, where N is the number of particles, making it substantially faster than direct pairwise calculations (N^2) for large systems.

4.2.3. Periodic boundary condition

The classical way to minimize edge effects in a finite system, such as MD simulation box, is to apply periodic boundary conditions (PBCs). PBCs is a widely used technique in MD simulations to minimize edge effects and approximate an infinite system. In computational studies, the need to simulate a finite system of particles often introduces challenges at the boundaries, such as surface artifacts or unrealistic interactions. PBCs address these issues by replicating the simulation box in all spatial directions, creating a theoretically infinite lattice that ensures continuity of interactions across the boundaries of the simulated system (Figure 9).

In a typical molecular dynamics setup, the molecular system of interest is solvated with water and neutralized with ions. This solvated system is confined within a simulation box, which can have various geometries depending on the study's requirements, including *cubic*, *rhombic dodecahedron*, and *truncated octahedron*. The simulation box is surrounded by translated copies of itself in the x, y, z directions, creating a seamless lattice. When a particle exits one side of the box, it re-enters on the opposite side with the same velocity, preserving the system's continuity and avoiding edge effects. While PBCs effectively mitigate boundary-related issues, they can introduce artifacts unique to this approach. For instance, particles in the central box can interact with their periodic images, potentially distorting the system's dynamics. To minimize PBC-related artifacts, the dimensions of the simulation box must be carefully chosen. The box should be large enough to prevent interactions between a molecule and its periodic images.

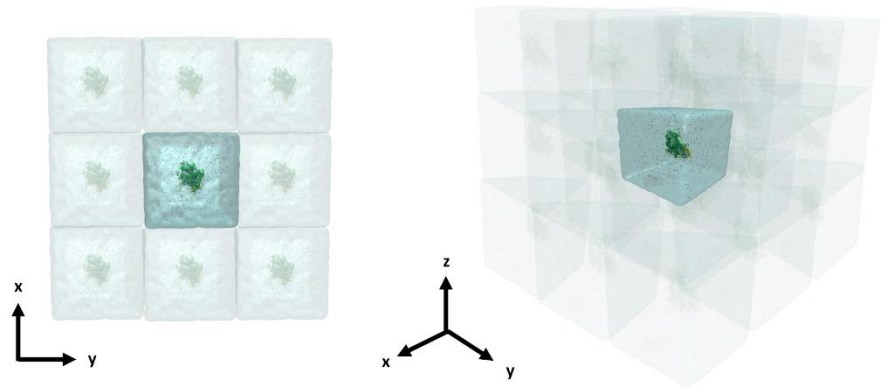


Figure 9: Scheme of system in periodic boundary condition (PBC) in 2D (on the left) and 3D (on the right). The box in the center is the original box, surrounded by copy of itself in the x, y, z axis.

PBC and the Particle-Mesh Ewald (PME) Algorithm

As discussed in the previous chapter, PBCs are integral to the PME method for efficiently calculating long-range electrostatic interactions. The replication of the simulation box ensures that the Fourier-based PME calculations accurately account for interactions across periodic boundaries. However, for the PME algorithm to function correctly, the system's net charge must be neutral. If the system contains charged particles, the total charge must be neutralized to zero coulombs by adding counter-ions. This prevents the introduction of infinite charges during the application of PBCs, maintaining the physical validity of the simulation.

Practical Considerations for PBCs

- Box Geometry: The choice of box geometry should balance computational efficiency and accuracy. Non-cubic shapes like dodecahedral or octahedral boxes can reduce the number of solvent molecules required, lowering computational costs.
- Box Size: The dimensions of the simulation box must ensure that a molecule does not interact with its periodic image. A commonly used guideline is to maintain a minimum distance of 1.0–1.5 nm between the solute and the box edges, but the choice is still case dependent.
- Neutralization: Systems with charged particles require the addition of counter-ions to achieve overall charge neutrality, ensuring accurate calculations of electrostatic forces.

4.2.4. Cut-off Restrictions

As previously mentioned, cut-off radii are used to truncate non-bonded interactions at a defined distance threshold, simplifying computational efforts by ignoring interactions beyond this range. When PBCs are employed, the minimum image convention is introduced to ensure that the interactions are calculated correctly within the periodic framework:

$$R_C < \frac{1}{2}\min(\|a\|, \|b\|, \|c\|) \tag{2.8}$$

where R_C is the cut-off radii and $\min(\|a\|, \|b\|, \|c\|)$ is the shortest box vector. The minimum image convention requires that the cut-off radius must not exceed half the shortest box vector, for not having more than one image within the cut-off radius. This rule may generally not be enough, in fact, extending the concept of minimum image convention we should have a length of the box vector that exceeds at least the length of the studied molecule projected along that direction, plus 2 times the radius of cut-off.

4.3. Molecular dynamics

MD is a computational technique that through the integration of Newton's equations solves the dynamics of a molecular system. More generally MD is a theoretical/computational algorithm

which can calculate average proprieties of a system by sequentially sampling microstates ensemble in time. Generic scheme of the MD operation is shown in Figure 10:

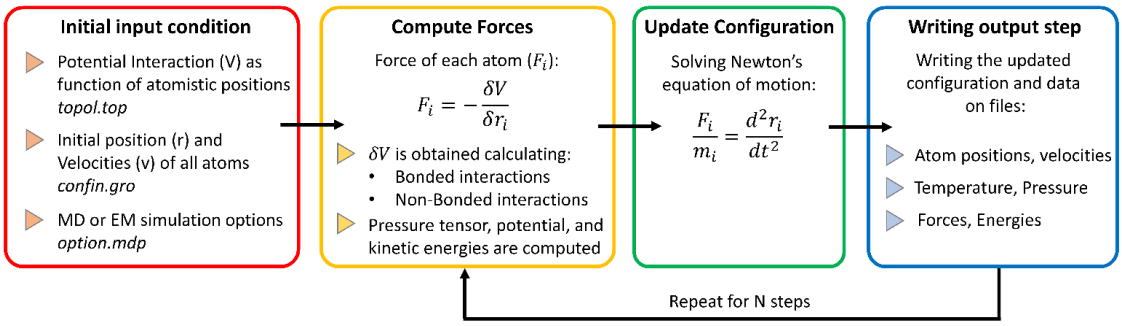


Figure 10: General scheme of MD algorithm. From the initial conditions the potential energies are calculated. Forces acting on the atoms are derived from potential energy function and through the integration of Newton's equations, new positions and velocities of the system are predicted. The cycle is repeated for the N steps set at the start.

4.3.1. Statistical ensembles

In statistical mechanics and thermodynamics, statistical ensembles represent the set of all possible microstates of a system that correspond to the same macroscopic or thermodynamic state. These ensembles allow us to connect the microscopic description of a system, defined by individual particle properties, to observable thermodynamic quantities.
A microscopic state is fully characterized by the positions and momenta of all particles within a system. These properties define a phase space, which is a multidimensional space where every point represents a unique microstate of the system. The collection of points in the phase space that correspond to the same macroscopic thermodynamic state forms a statistical ensemble. In 1902 Willard Gibbs defined 3 important thermodynamic ensembles, *Micro-canonical*, *Canonical* and *Grand-canonical*, to whom we also consider a fourth case that is the *Isothermal-isobaric* ensemble[128] (Figure 11):

- *Micro-canonical Ensemble* (**NVE**): Represents an isolated system, which cannot exchange either energy or matter with a reservoir. The system has a constant number of particles (N), volume (V), and total energy (E).
- *Canonical Ensemble* (**NVT**): Represents a closed system that can exchange energy, but not matter, with a reservoir. The system has a constant number of particles (N), volume (V), and temperature (T). This ensemble is often used for systems maintained at a constant temperature, typically achieved using thermal reservoirs or thermostats in simulations.
- *Grand-canonical Ensemble* (**µVT**): Represents an open system capable of exchanging both energy and matter with a reservoir. The system has a constant chemical potential (µ), volume (V), and temperature (T).
- *Isothermal-Isobaric Ensemble* (**NPT**): Represents a closed system that can exchange energy with its surroundings, and where the system's volume can fluctuate to maintain constant pressure. The system has a constant number of particles (N), pressure (P), and temperature

(T). This ensemble is widely used to simulate biological and chemical systems under experimental conditions where both temperature and pressure are controlled.

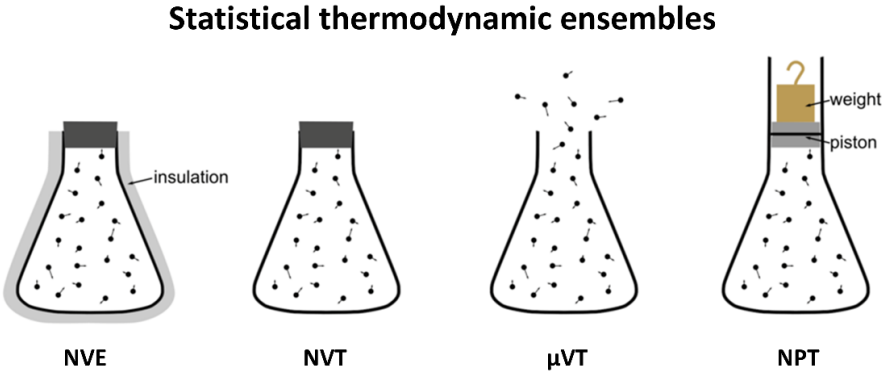


Figure 11: Qualitative representation of the 4 statistical thermodynamic ensembles: Micro-canonical (NVE), canonical (NVT), grand-canonical (µVT), isothermal-isobaric (NPT).

Statistical ensembles provide a critical framework for bridging microscopic and macroscopic states, linking the phase space of a system to observable thermodynamic properties such as temperature, pressure, and free energy. The phase space, a multidimensional representation of the positions and momenta of all particles in a system, serves as the foundation for defining ensembles. In computational studies, ensemble generation is essential for approximating equilibrium states.

Two primary methods are employed to generate representative statistical ensembles: MD simulations and Monte Carlo simulations. Both techniques sample microstates from the phase space, enabling the calculation of ensemble averages for macroscopic properties. The ensemble average of propriety A can be expressed in the following way:

$$\langle A \rangle_{ensamble} = \int \int A(p^N, q^N)\, \rho(p^N, q^N) dp^N dq^N \tag{2.9}$$

where $A(p^N, q^N)$ is the observable of interest, p is the position of the N atoms, q is the momentum of the N atoms and $\rho(p^N, q^N)$ is called the ensemble density of probability function:

$$\rho(p^N, q^N) = \frac{1}{Q} exp\left[-\frac{H(p^N, q^N)}{k_B T}\right] \tag{2.10}$$

where k_B is the Boltzmann's constant, T is the absolute temperature, $H(p^N, q^N)$ is the Hamiltonian and Q is the expression of the canonical discrete partition function:

$$Q = \sum_i exp[-\beta E_i] = \sum_i exp\left[-\frac{1}{k_B T} E_i\right] \tag{2.11}$$

where k_B is the Boltzmann's constant, T is the absolute temperature and E_i is the total energy of the system in the i-microstates. The canonical partition function describes the thermodynamic and statistical properties of a system by normalizing sum of Boltzmann's factors over all microstates. Discrete partition function is dimensionless and can be further extended to the continuous by replacing the summation with an integral operator:

$$Q = \int \int exp\left[-\frac{H(p^N,q^N)}{k_B T}\right] dp^N dq^N \tag{2.12}$$

Equation 2.12 is crucial to have the connection between the thermodynamic microstates' variable, which we can't measure, and the thermodynamic macroscopic state, which we can measure. Unfortunately, the calculation of this integral is extraordinarily complex because it considers all the possible microstates of the system. A way to simplify the problem is to use the ergodic hypothesis.

4.3.2. Ergodic hypothesis

The ergodic hypothesis plays a pivotal role in simplifying the calculation of macroscopic properties in statistical mechanics and thermodynamics. Solving the integral presented in equation (2.12) to determine ensemble averages can be prohibitively complex due to the high-dimensional nature of phase space. The ergodic hypothesis provides a practical means to overcome this challenge by establishing a critical equivalence between time averages and ensemble averages.

In its essence, the ergodic hypothesis posits that over sufficiently long periods, a system will explore all accessible microstates in its phase space that are consistent with the system's total energy. This implies that the time spent by a particle in a specific region of phase space is proportional to the volume of that region. Consequently, every microstate with the same energy is equally probable over long timescales.

The usefulness of the ergodic hypothesis lies in its ability to bridge temporal and statistical perspectives. Instead of calculating macroscopic quantities as ensemble averages over an exhaustive set of microstates, we can evaluate them as time averages of a property during a sufficiently long simulation:

$$\langle A \rangle_{ensemble} = \langle A \rangle_{time} \tag{2.13}$$

Therefore, we can compute the ensemble average of propriety A using the time average of the same propriety, which is more practical in computational simulations. Specifically, the time average of a property A can be expressed as:

$$\langle A \rangle_{time} = \lim_{\tau \to \infty} \frac{1}{\tau} \int_{t=0}^{\tau} A\big(p^N(t), q^N(t)\big) dt \approx \frac{1}{M} \sum_{t=0}^{M} A(p^N, q^N) \tag{2.14}$$

where t is the simulation time, M is the number of time steps and $A(p^N, q^N)$ is the instantaneous value of the propriety A.

4.3.3. Temperature coupling

Maintaining a constant temperature in MD simulations is crucial for accurately representing thermodynamic ensembles, such as NVT or NPT ensembles. To achieve this, MD software incorporates temperature coupling schemes that weakly couple the system to an external thermal bath. These algorithms control the temperature by rescaling the velocities of the particles in the system, thereby adjusting the kinetic energy. Several widely used temperature coupling schemes include the *Berendsen*[129], *v-rescale*[130], *Nosé-Hoover*[131], and *Andersen*[132] algorithms.

The *Berendsen* thermostat employs a weak coupling approach to maintain the system temperature T near a target value T_0. It achieves this by correcting the temperature deviation through an exponential decay controlled by a user-defined time constant tau:

$$\frac{dT}{dt} = \frac{T_0 - T}{\tau} \tag{2.15}$$

This correction ensures that deviations from the desired temperature decay smoothly over time. The temperature control is enforced by rescaling the velocities of the particles at each time step using a scaling factor λ:

$$\lambda = \sqrt[2]{1 + \frac{n_{TC}\Delta t}{\tau_T}\left(\frac{T_0}{T(t-1/2\Delta t)} - 1\right)} \tag{2.16}$$

where n_{TC} is the time step, λ is the scaling factor that is limited in the range 0.8< λ <1.25 for stability issues and τ_T is a time constant close to τ in equation 2.15, which can be joined by the relation:

$$\tau = \frac{2C_V\tau_T}{N_{df}k} \tag{2.17}$$

where C_V is the total heat capacity of the system, N_{df} is the total degrees of freedom of the system and k is the constant of Boltzmann. The Berendsen thermostat is computationally efficient and effectively reduces temperature fluctuations. However, it does not produce a rigorous canonical ensemble because it lacks stochastic elements to correctly sample phase space.

The *v-rescale* algorithm builds upon the *Berendsen* thermostat by introducing a stochastic term that ensures proper sampling of the canonical ensemble. The dynamics of the kinetic energy K are governed by the following equation:

$$dK = (K_0 - K)\frac{dt}{\tau_T} + 2\sqrt{\frac{KK_0}{N_{df}}}\frac{dW}{\sqrt{\tau_T}} \quad (2.18)$$

where K is the kinetic energy and dW is a Wiener process. The inclusion of the stochastic term enhances the algorithm's ability to maintain a Maxwell-Boltzmann distribution of velocities, a critical feature for achieving a true canonical ensemble.
The v-rescale thermostat combines the strengths of the Berendsen scheme with improved accuracy in sampling the kinetic energy distribution:

1. **No Oscillations**: The rescaling of velocities avoids oscillatory behavior, ensuring smooth convergence to the target temperature.
2. **First-Order Decay**: Deviations from the target temperature decay in a controlled, first-order manner, enhancing stability.
3. **Canonical Ensemble**: The stochastic term ensures correct energy fluctuations, producing a proper canonical ensemble.

Temperature coupling schemes like *Berendsen* and *v-rescale* are widely used in MD simulations to study systems in equilibrium conditions. The choice of thermostat depends on the specific requirements of the simulation, such as for example whether strict adherence to the canonical ensemble is necessary (production runs). While *Berendsen* is suitable for equilibration runs, *v-rescale* is mandatory for production runs where accurate thermodynamic properties are required. However, the *v-rescale* thermostat is nowadays employed to both equilibrate the system and perform production simulation[133].

4.3.4. Pressure coupling

Maintaining constant pressure in MD simulations is mandatory for correctly reproducing the NPT thermodynamic ensemble. In MD, several algorithms can simulate pressure coupling with an external pressure bath, as for the temperature coupling discussed previously. The most commonly used pressure barostats include the *Berendsen*[129] and *Parrinello-Rahman* schemes[134]. Recently, the stochastic cell rescaling (*c-rescale*)[135] barostat has emerged as a robust alternative, particularly for maintaining accuracy in pressure control during MD simulations.

The *Berendsen* barostat operates on a weak coupling scheme, effectively controlling the system's average pressure. While it is well-suited for system equilibration under pressure coupling, it does not yield the correct NPT ensemble, as it artificially suppresses pressure fluctuations. Consequently, it is not recommended for production MD runs but is an efficient option for the initial equilibration phase. The *Berendsen* algorithm maintains the pressure P_0 by rescaling the box vectors at each pressure coupling step n_{PC} using a scaling matrix μ in accordance with:

$$\frac{dP}{dt} = \frac{P_0 - P}{\tau_P} \tag{2.19}$$

The scaling matrix μ is defined as follows:

$$\mu_{ij} = \delta_{ij} - \frac{n_{PC}\Delta t}{3\tau_P}\beta_{ij}\{P_{0ij} - P_{ij}\} \tag{2.20}$$

where δ_{ij} is the identity matrix, τ_P is the time constant, β_{ij} is the isothermal compressibility of the system, for water at 1 atm and 300K $\beta_{ij} = 4{,}6\ x\ 10^{-5} bar^{-1}$. Box-scaling can be performed in all directions (isotropic) or independently along different axes(anisotropic). For anisotropic scaling, it may be necessary to adjust τ_P or reduce n_{PC} to ensure stability and avoid errors due to constraints.

On the other hand, the *Parrinello-Rahman* barostat is designed to simulate the true NPT ensemble and allows for natural pressure fluctuations. It is ideal for MD production runs but is less suited for equilibration because large oscillations in box vectors can destabilize the simulation. This algorithm rescales the box vectors, represented by the matrix b, based on the following equation:

$$\frac{db^2}{dt^2} = VW^{-1}b'^{-1}(P - P_0) \tag{2.21}$$

where V is the box volume, W^{-1} is the matrix with the coupling strength parameters and P is the pressure. In addition to box rescaling, the atomic equations of motion are also rescaled:

$$\frac{d^2 r_i}{dt^2} = \frac{F_i}{m_i} - \left\{b^{-1}\left[b\frac{db'}{dt} + \frac{db}{dt}b'\right]b'^{-1}\right\}\frac{dr_i}{dt} \tag{2.22}$$

The coupling strength matrix W^{-1} is determined by:

$$W^{-1}{}_{ij} = \frac{4\pi^2\beta_{ij}}{3{\tau_P}^2 L} \tag{2.23}$$

where β_{ij} is the isothermal compressibility, τ_P is the pressure time constant and L is the largest box matrix element.

The *c-rescale* barostat is a first-order barostat that samples the correct volume fluctuations by including a suitable noise term. This barostat addresses the limitations of both the *Berendsen* and *Parrinello-Rahman* methods. It incorporates stochastic fluctuations to accurately reproduce the true NPT ensemble while maintaining robust stability during simulations. *C-rescale* can be straightforwardly implemented in the existing MD codes and can be used effectively in both equilibration and production phases[135].

4.3.5. Energy minimization

Energy minimization (EM) is a pivotal process in MD simulations used to optimize the atomic structure of a system by finding a local minimum on the potential energy surface (PES). The PES represents the multidimensional landscape of potential energy as a function of atomic positions. The goal of EM is to locate a local stationary point where the potential energy is minimized, and the net interatomic forces are ideally close to zero. However, due to the inherent nature of most optimization algorithms, which are unidirectional and follow downhill trajectories, only the nearest local minimum relative to the starting point can be identified.
By reducing the potential energy and minimizing the interatomic interaction forces, EM prepares the system for further simulations, such as MD or Monte Carlo simulations, by ensuring that the initial configuration is energetically stable. It is important to mention that this stabilization is essential for obtaining reliable results in subsequent steps. Minimization algorithms are broadly classified into two categories:

- **Derivative methods**: These algorithms utilize the gradient of the potential energy to guide the search for the minimum. Common examples include the *steepest descent*, *conjugate gradient*, and *Newton-Raphson* methods.
- **Non-derivative methods**: These rely solely on the evaluation of energy values without using gradients. The *simplex* method is a notable example.

The steepest descent algorithm is the most robust and commonly used method for energy minimization in MD simulations, despite not being the most efficient for locating minima (Figure 12). Its simplicity and ease of implementation make it an attractive choice for many applications. The algorithm operates by calculating force F from the negative gradient of the potential energy V, and updating atomic positions iteratively according to the following equation:

$$r_{n+1} = r_n + \frac{F_n}{\max(|F_n|)} h_n \tag{2.24}$$

where r_{n+1} is the new position, r_n is the actual position, F_n is the force and h_n is the maximum displacement initially selected (for example 0.01 nm). If the $V_{n+1} < V_n$ the new positions are accepted and $h_{n+1} = 1.2h_n$, but if the $V_{n+1} > V_n$ the new positions are not accepted and $h_n = 0.2h_n$. The code stops when the maximum number of force evaluation steps has been reached, or when the maximum of the absolute value of force is less than a certain ε value, which is considered acceptable between 1 and 10.

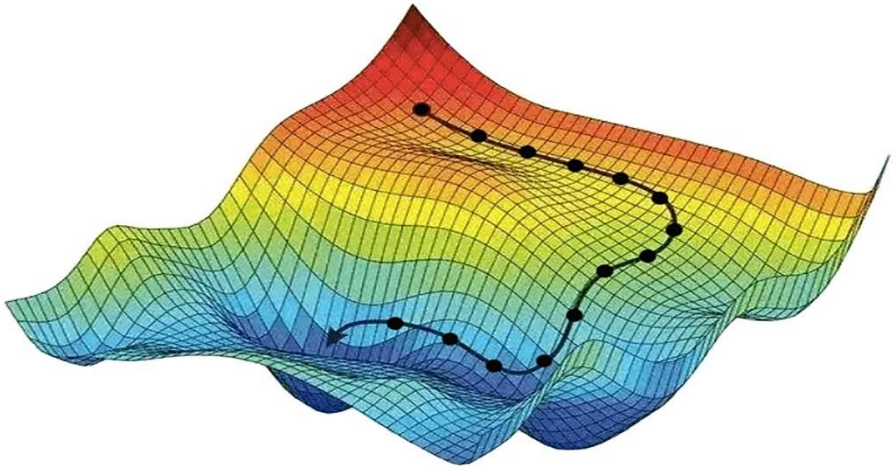


Figure 12: Potential energy surface (PES) with illustrated an example of Energy minimization (EM) sequence by the steepest descent algorithm. The starting point of the system is in a high energy state (red) and gradually transitions step by step toward a low-energy state (blue).

4.3.6. Leapfrog integrator

The leapfrog algorithm is a numerical method used in MD that by integrating the differential equations of motion produces an output trajectory. The leapfrog integration method (Figure 13) uses the position r of atoms at moment t and the velocity v at a previous time $t = t - \frac{1}{2}\Delta t$ updating positions and velocities using the forces on atoms F at time t, through the following relations:

$$v\left(t + \frac{1}{2}\Delta t\right) = v\left(t - \frac{1}{2}\Delta t\right) + \frac{\Delta t}{m}F(t) \tag{2.25}$$

$$r(t + \Delta t) = r(t) + \Delta t v\left(t + \frac{1}{2}\Delta t\right) \tag{2.26}$$

The leapfrog algorithm is time reversible of third order in r and produces trajectories whit the following position update equation:

$$r(t + \Delta t) = 2r(t) - r(t - \Delta t) + \frac{1}{m}F(t)\Delta t^2 + O(\Delta t^4) \tag{2.27}$$

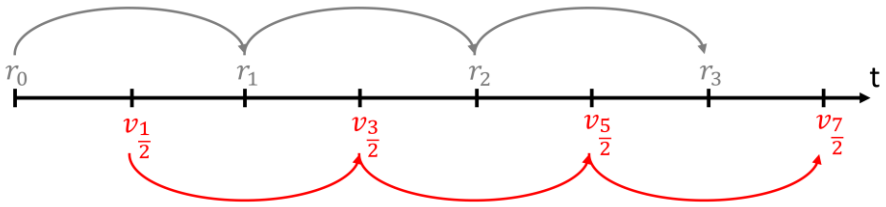


Figure 13: Scheme of the leapfrog algorithm, where r is the position, v is the velocity, and t is time. This method is called leapfrog because positions and velocities are leaping like frog over a Δt time difference.

4.3.7. Constraint algorithm

During MD simulations, it is essential to maintain the relative positions of atoms within a specific oscillation range, as determined by the chemical and physical properties of the system. The relative positions are described by the system's topology, which includes the constraints imposed by the bonded terms. Generally, in MD the covalent distance between heavy atoms and hydrogen atoms are constrained for enhancing the computational efficiency by reducing the vibrational component along some degrees of freedom. To ensure that these distances remain within physically meaningful limits, constraint algorithms are employed. Two widely used algorithms for enforcing bond constraints in MD simulations are the SHAKE[136] and the LINCS[137] algorithms.

The LINCS (Linear Constraint Solver) algorithm is a non-iterative method designed to efficiently restore bond lengths to their correct covalent distances after an unconstrained position update. This algorithm is particularly advantageous due to its speed and robustness compared to iterative methods like SHAKE. The LINCS algorithm operates in two primary steps:

- **Initial Projection**: The algorithm projects the new, unconstrained bond lengths onto the old, constrained ones. In this step, the deviations of the bonds from their ideal lengths are initially set to zero.
- **Bond Distance Correction**: A correction is applied to account for bond-length distortions caused by unconstrained rotations or updates during the simulation.

4.4. Machine learning

4.4.1. LightGBM Algorithm

Light Gradient Boosting Machine (LightGBM)[138] is an advanced implementation of the gradient boosting framework. This ML algorithm is highly praised for its efficiency and performance, especially in scenarios involving large datasets and the need for fast computation. The fundamental concept of LightGBM, as with other gradient boosting methods, is to build an ensemble of weak prediction models, which are in most cases simple decision trees, sequentially. Each tree in the sequence attempts to correct the residuals or errors of the

preceding trees. The final predictive model is an aggregation of these weaker decision trees. The model's output for a given input x is:

$$F(x) = \sum_{m=1}^{M} \beta_m f_m(x)$$

where M represents the total number of trees, $f_m(x)$ is the output of the m^{th} decision tree and β_m is the corresponding weight, often set to $\beta_m = 1/M$. In each iteration of training, the model focuses on correcting its previous mistakes. At each iteration, the negative gradient of the loss function with respect to the model's predictions is computed (pseudo-residuals) and quantifies the direction and magnitude of the error. This gradient information is then used to guide the construction of the next tree, specifically aiming to reduce the residual error. The update formula can be expressed as:

$$F_{t+1}(x) = F_t(x) + \alpha \cdot f_{t+1}(x)$$

$F_{t+1}(x)$ is the updated model after iteration $t+1$, $F_t(x)$ is the model from the previous iteration t, $f_{t+1}(x)$ is the newly added tree, and α is the learning rate, which controls how strongly each new tree influences the final model.

LightGBM introduces two innovative approaches to improve efficiency: Gradient-based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB). GOSS improves the training process by prioritizing training instances that have larger gradients, meaning it focuses more on the harder-to-predict instances. On the other hand, EFB efficiently manages sparse data by bundling together exclusive features, which are those features that are rarely non-zero at the same time. These techniques allow LightGBM to handle large volumes of data with higher speed and lower memory usage compared to traditional gradient boosting methods. Moreover, LightGBM supports categorical features natively and is capable of handling missing data, making it a versatile tool in various machine learning tasks, from classification to regression. Despite its efficiency, it retains a high degree of accuracy, making it a favored choice among data scientists and researchers.

4.4.2. Global feature attribution methods: Mean Decrease in Impurity

In ML, quantifying the contribution of each input variable in predicting the response is called feature importance attribution. In most implementations of tree-based ensemble models, such as Random Forest (RF) and Gradient Boosting (GB), feature importance is typically computed using the Mean Decrease in Impurity (MDI) method proposed by Breiman[139].

MDI is a method specifically tailored for tree-based models. A decision tree is a predictive model organized as a tree structure T, mapping input features $x \in \mathcal{X}$ to output targets $y \in \mathcal{Y}$. Typically, the input space $\mathcal{X}$ is represented as a P-dimensional real space $\mathbb{R}^P$, with P denoting the number of input features. In classification tasks, $\mathcal{Y}$ comprises a set of discrete labels $\{1,2,\dots,C\}$, where C indicates the number of distinct, mutually exclusive classes. The decision

tree for classification is known as a classification tree. Noteworthy, in binary classification where $C = 2$, the typical assumption is $y \in \{0,1\}$.

The decision tree uses the root node to represent the entire input space $\mathcal{X}$. Each internal node t corresponds to a specific subset of $\mathcal{X}$, and the branches from these nodes result from binary decisions or splits $s_t = (x_k < c)$, further dividing the portion of input space associated with node t into two child nodes t_L and t_R. Nodes without children, situated at the ends of the tree, are known as terminal or leaf nodes. A test instance x is classified by traversing from the root down to a leaf node, where in classification tasks, each leaf node assigns a prediction $\hat{y}$ based on the majority class of the training samples that end there. This structure enables the tree to adaptively map various regions of $\mathcal{X}$ to the appropriate output predictions, mirroring the training data's underlying distribution.

During the training phase, the tree structure is developed starting from a dataset encapsulated at the root node. Nodes are recursively created via a greedy procedure that groups samples with identical labels values. At each node t, the algorithm selects the optimal split $s_t = s^*$ that maximizes the reduction in impurity:

$$\Delta i(s_t) = i(t) - w_{t_L} i(t_L) - w_{t_R} i(t_R)$$

Here, $i(t)$ represents an impurity measure (e.g., the Gini index, or the Shannon Entropy for classification), and $w_{t_L} = N_{t_L}/N_t$ and $w_{t_R} = N_{t_R}/N_t$ denote the proportions of samples at node t moving to t_L and t_R respectively. The effectiveness of each split, quantified by $\Delta i(s_t)$, signifies how well it purifies the node. Tree construction typically ceases when the nodes are pure or when no further informative splits can be made.

Practical implementations may also include constraints like maximum tree depth or minimum sample count per leaf to prevent overfitting. In an ensemble approach, multiple decision trees (N_T in total) are built, and the importance I_j of each input feature x_j in predicting the target is evaluated by averaging the weighted decreases in impurity from all nodes where x_j influences the split, across all trees:

$$I^{MDI}{}_j = \frac{1}{N_T} \sum_{T} \sum_{t \in T : v(s_t) = x_j} w_t \Delta i(s_t)$$

where w_t is the proportion N_t/N of samples reaching node t, and $v(s_t)$ indicates the feature used in the split.

This metric highlights the feature's overall impact on the model's accuracy, particularly noting the significance of features used in early splits due to their broader effect on the input space. This emphasizes their key role in the model's decision-making process.

4.4.3. Local feature attribution methods: Shapley Additive explanations (SHAP)

Shapley Additive explanations (SHAP), implemented in the Python SHAP library, offer practical solutions by estimating the values based on available data and optimizing the computation for

complex models like ensemble trees. SHAP optimizes calculations for ensemble models, reducing computational complexity from $O(TL2^P)$ to $O(TLD^2)$, where T is the number of trees, L is the maximum number of leaves, and D is the maximum depth.
Shapley Values represent local feature attributions, reflecting each feature's influence on the prediction at the level of an individual observation x. However, aggregating Shapley Values across a dataset provides a global view of feature importance via a matrix Φ, where each row corresponds to an observation, and each column to a feature:

$$\Phi = \left(\phi_j^{(i)}\right)_{i=1,\dots,n\ ,j=1,\dots,P}$$

Feature importance can be derived by averaging the absolute Shapley Values per feature:

$$I^{SHAP}{}_j = \frac{1}{n}\sum_{i=1}^{n}\left|\phi_j^{(i)}\right|$$

Like many other interpretation methods, the Shapley value approach encounters challenges when dealing with correlated features[140]. To simulate the absence of a feature from a coalition, the method involves marginalizing that feature, typically by sampling from the feature's marginal distribution. This approach works well for independent features. However, when features are dependent, this sampling process may introduce unrealistic feature values for the given instance.
One possible solution is to use the PartitionExplainer from the SHAP library, which requires as argument a hierarchical clustering of the input features based on correlation. The PartitionExplainer computes Shapley values recursively through the hierarchy of features that defines feature coalitions, considering groups feature together and allocating credit based on how one group would perform as a whole. This means that if the clustering provided to the PartitionExplainer groups these correlated features together, the method effectively manages feature correlations. Essentially, the total credit given to a group of closely related features remains consistent and does not fluctuate based on changes in their correlation during the explanation process. This ensures that the explanation reflects their combined influence accurately, without being affected by any alterations to their interdependencies.

4.4.4. Model Reliability: Local Outlier Factor

Among different possibilities to assess the reliability of the model, Local Outlier Factor (LOF)[141] unsupervised ML for novelty detection is one of the most commonly used algorithms.
Let $\mathcal{D}_T = \left\{\left(x^{(i)}, y^{(i)}\right)\right\}_{i=1}^{N}$ be the N-dimensional set used for training, with $x^{(i)}$ and $y^{(i)}$ being the feature and label, respectively, associated to the i-th sample. Let $x^{(N+1)}$ be the feature associated with a new sample. We assess the reliability of the model by measuring the level of novelty of $x^{(N+1)}$ with respect to features used for training, collected in the set $\mathcal{X}_{\mathrm{T}} = \left\{\mathrm{x}^{(\mathrm{i})}\right\}_{i=1}^{N}$.

In particular, the reliability of the model associated with the new feature $x^{(N+1)}$ decreases as the deviation of $x^{(N+1)}$ with respect to the training dataset $\mathcal{X}_{\mathrm{T}}$ increases.
The LOF algorithm measures the local density deviation of a given data point $x^{(N+1)}$ with respect to its neighbors, by comparing the density of areas surrounding $x^{(N+1)}$ to the densities of the areas surrounding its neighbors. A high LOF value indicates that the sample is far from its neighbors (i.e., an outlier or a novel sample with respect to $\mathcal{X}_{\mathrm{T}}$). Thus, it deviates from the training data, and this can suggest that the model's reliability is low for this sample. Conversely, a low LOF value suggests the sample is similar to its neighbors, suggesting high model reliability for this sample. The main steps of the LOF algorithms are reported below, assuming the features are already normalized:

- Let us define the augmented dataset $\widetilde{\mathcal{X}} = \mathcal{X}_{\mathcal{T}} \cup x^{(N+1)}$. For each pair of points $\mathrm{x}, \mathrm{q} \in \widetilde{\mathcal{X}}$ compute the Euclidean distance between x and q. Let us denote as $\mathrm{d}(\mathrm{x}, \mathrm{q})$ such a distance.
- For a positive integer k such that $0 < \mathrm{k} \leq \mathrm{N}$ and each point $\mathrm{x} \in \widetilde{\mathcal{X}}$, compute the so-called k-distance (denoted as $\mathrm{dist_k(x)}$), being the distance of x with its k-th nearest point.
- For each point $\mathrm{x} \in \widetilde{\mathcal{X}}$, compute the k-nearest neighbors of x (denoted as $N_k(x)$) defined as the set of samples of $\widetilde{\mathcal{X}}$ (excluding x) with distance smaller than $\mathrm{dist_k(x)}$. Basically, $N_k(x)$ includes all the points of $\widetilde{\mathcal{X}}$ that lie in the circle centered at x and of radius $\mathrm{dist_k(x)}$. Denote as $|N_k(x)|$ the cardinality of $N_k(x)$, which can be larger or equal than k.
- For a pair of points $\mathrm{x}, \mathrm{q} \in \widetilde{\mathcal{X}}$, define the reachability distance of x from q defined as:

$$\mathrm{rdist_k(x, q)} = max\{d(x, q), dist_k(q)\}$$

- For each point $\mathrm{x} \in \widetilde{\mathcal{X}}$, compute the local reachability density of x defined as:

$$\mathrm{lrd_k(x)} = \frac{|N_k(x)|}{\sum_{q \in N_k(x)} rdist_k(x, q)}$$

  The local reachability density is thus given by the inverse of the average reachability distance of the point x from its neighbors. Low values of local reachability density thus mean that the point x is "far" from its neighbors.
- For each point $x \in \widetilde{\mathcal{X}}$, the $lrd_k(x)$ is then compared with the local reachability density of its neighbors to compute the value of the local outlier factor for x, specifically:

$$\mathrm{LOF_k}(x) = \frac{\sum_{q \in N_k(x)} lrd_k(q)}{|N_k(x)| lrd_k(x)}$$

Note that a value of $\mathrm{LOF_k}(x)$ of approximately 1 indicates that the "local density" of the point x is comparable to one of its neighbors. Thus, x is not an outlier. On the other hand, values of $\mathrm{LOF_k}(x)$ drastically larger than 1 indicate that x is an outlier.
It is worth pointing out that if we are only interested in detecting if a new point $x^{(N+1)}$ is a novelty, there is no need to compute the LOF value for all points in the dataset $\widetilde{\mathcal{X}}$. Indeed, it is

enough to compute the k-nearest neighbors of $x^{(N+1)}$, along with the local reachability density of its neighbors.

4.4.5. Performance evaluations

In the field of machine learning and classification tasks, evaluating the performance of a model is critical to understanding its effectiveness. One way to assess a classifier's performance is by examining its predictions in relation to the actual class labels. This evaluation involves categorizing the outcomes of the classifier into four key metrics, which are defined based on the comparison of predicted labels and true labels. These metrics are:

- True positive (TP): The number of positive class members, which are properly predicted by the classifier and are labeled as positive class.
- True negative (TN): The number of negative class members, which are properly predicted by the classifier and are labeled as negative classes.
- False positive (FP): The number of negative class members, which are falsely predicted by the classifier and are labeled as positive class.
- False negative (FN): The number of positive class members, which are falsely predicted by the classifier and are labeled as negative class.

Threshold-dependent and threshold-independent metrics are often used to evaluate binary classification performance. The threshold-dependent metrics are Sensitivity (SNV), Specificity (SPC), Accuracy (ACC), and the Matthews correlation coefficient (MCC), which are defined as follows:

SNV is defined as a probability that a classifier truly predicts the result as positive when the corresponding sample is truly positive. The SNV is also called the true positive rate, and it is calculated as follows:

$$SNV = \frac{TP}{TP + FN}$$

SPC is defined as the probability that a classifier truly predicts the result as negative when the corresponding sample is truly negative. The SPC is also called the true negative rate, and it is calculated as follows:

$$SPC = \frac{TN}{TN + FP}$$

ACC is a measure of the overall correctness of a model and is calculated as the ratio of correctly predicted instances to the total instances. It provides a general assessment of the model's effectiveness.

$$ACC = \frac{TP + TN}{TP + TN + FP + FN}$$

MCC is defined as the correlation coefficient between the predicted result and the corresponding ground truth. It has a value between +1 and −1. If MCC=+1 it means that the classifier predicts the result truly. If MCC=0 it means that the classifier cannot predict the result better than a random manner. If MCC=−1 it means that there is a full contradiction between the predicted result and the corresponding ground truth. The MCC scale is calculated as follows:

$$MCC = \frac{\mathrm{TP} \cdot \mathrm{TN} - \mathrm{FP} \cdot \mathrm{FN}}{\sqrt{(\mathrm{TP} + \mathrm{FP}) \cdot (\mathrm{TP} + \mathrm{FN}) \cdot (\mathrm{TN} + \mathrm{FP}) \cdot (\mathrm{TN} + \mathrm{FN})}}$$

Additionally, the Area Under the Receiver Operating Characteristic Curve (AUC) serves as a threshold-independent metric to evaluate model performance.

4.4.6. Genetic algorithm

Genetic algorithm[142,143] (GA) is an optimization technique to address the complexity of the parameters space in the optimization process. GAs, inspired by the principles of natural selection and genetics, are heuristic optimization algorithms widely utilized for solving complex combinatorial and numerical problems. It is designed to minimize or maximize a given fitness function $J(\theta)$ by iteratively improving a population of potential solutions. The main steps involved in a generic genetic algorithm are outlined below:

1. **Initialization**: Generate an initial population of P_{init} individuals, where each individual represents a potential solution to the optimization problem. These initial solutions are typically created randomly within the allowed parameter space, forming the starting point of the algorithm. Each individual encodes the design parameters θ in a format appropriate for the problem, such as binary strings, real-valued vectors, or sequences.
2. **Fitness Evaluation and Selection**: Evaluate the fitness $J(\theta)$ of each individual in the population based on the problem's objective function. Select N_{sel} individuals with the highest fitness (for maximization problems) or the lowest fitness (for minimization problems) to serve as parents for the next generation. Various selection methods, such as roulette wheel selection, tournament selection, or rank-based selection, can be employed to choose the parents.
3. **Crossover (Recombination)**: Generate offspring by performing crossover or recombination on pairs of selected parents. This operation involves combining the genetic information (the encoded design parameters θ) of two parents to produce one or more offspring. A common crossover technique is single-point or multi-point crossover, where portions of one parent's genetic material are exchanged with the other parent's material. Alternatively, uniform crossover may be used, where each element of the offspring's genetic code is randomly inherited from one of the two parents with equal probability.

4. **Mutation**: Apply mutation to a subset of the offspring. Mutation introduces small, random modifications to an individual's genetic representation, such as flipping bits in binary strings, adding noise to numerical values, or altering specific components of the solution. The purpose of mutation is to maintain diversity in the population and prevent premature convergence to suboptimal solutions. The mutation rate (e.g., 10%) determines the likelihood of mutation occurring for each element of the genetic code.
5. **Replacement**: Form the next generation by replacing the old population with the newly generated offspring and a subset of surviving parents. The population size is generally kept constant. Various replacement strategies can be employed, such as elitism, where a fixed number of the best-performing individuals are retained from the previous generation, or generational replacement, where only offspring are used to form the new population.
6. **Termination Criterion**: Repeat the fitness evaluation, selection, crossover, mutation, and replacement steps for a predetermined number of generations or until a termination criterion is met. Termination criteria can include reaching a target fitness value, exceeding a predefined computational budget (e.g., number of generations), or detecting a lack of improvement in fitness over successive generations.
7. **Result Extraction**: Once the termination criterion is met, the individual with the best fitness in the final population is selected as the optimized solution to the problem. This individual represents the genetic algorithm's approximation of the global optimum for the fitness function $J(\theta)$.

GA are versatile and can be tailored to a wide range of optimization problems by adjusting the encoding scheme, selection mechanism, crossover and mutation operators, and termination conditions. A schematic representation of the GA is provided in Figure 14.

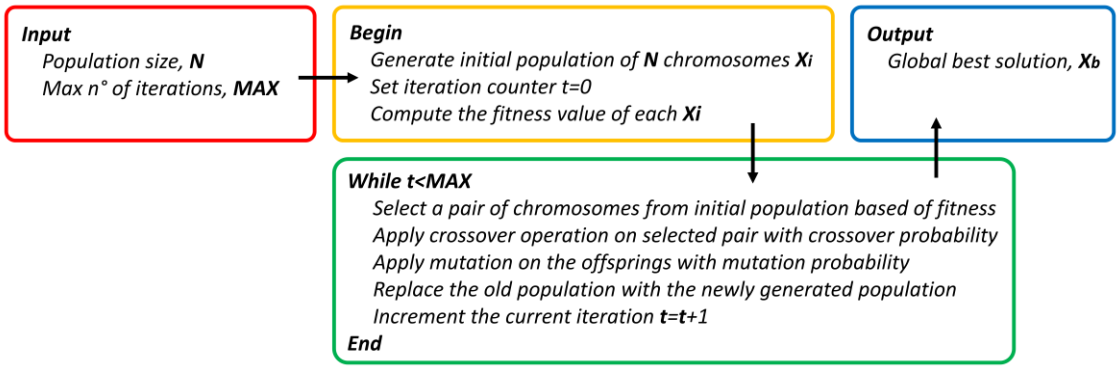


Figure 14: Schematic representation of the Genetic Algorithm sequential operations.

5. Toward the Rational Design of Penetrating Peptides by Machine Learning

In this chapter, it is described the rational *de novo* CPP design employing a synergy of ML and GA. The convergence of ML techniques and GA serves as a catalyst for the design of CPPs with enhanced penetrability, setting the stage for subsequent chapters to unveil the molecular mechanism of a specific multi-active CPP. In this connection, predicting peptide penetration into biological membranes using computational methods has proven to be both an efficient and cost-effective approach for screening expansive chemical libraries[144–151].

Among the numerous techniques used to enhance this predictive performance, one promising methodology is the integration of ML models with sequence-derived descriptors. Early breakthroughs in the field include the CPPred-RF[147], which utilized the random forest (RF) algorithm integrated with several sequence-based descriptors. Building on this foundation, Qiang et al. took a similar approach and developed the CPPred-FL, which leveraged multiple RF models trained on various attributes, ranging from amino acid composition to specific position information[152]. Soon after, Fu et al. proposed a unique support vector machine (SVM) algorithm, which predicts CPPs based entirely on the amino acid composition[153].

Further advancements led to the creation of the KELM-CPPpred, a pioneering ML framework[145] that also relied heavily on amino acid composition for CPP prediction. The combination of both sequence and structure-based descriptors marked a turning point in the field, significantly improving CPP prediction accuracy. For instance, Manavalan et al. presented a framework that successfully integrated multiple algorithms, including RF, SVM, extremely randomized trees (ERT), and k-nearest neighbour (K-NN) algorithms[144]. By leveraging diverse sets of data, such as amino acid composition and physicochemical properties, their approach achieved greater precision in distinguishing CPPs from non-CPPs.

Moreover, CellPPD-Mod[154] was introduced as a powerful computational tool that uses the RF method to discriminate between CPPs and non-CPPs based on molecular descriptors, fingerprints, and sequence features up to 25 residues in length. More recently, the ML-based framework BChemRF-CPPred further expanded predictive capabilities by combining techniques such as artificial neural networks (ANN), SVM, and Gaussian classifiers[155]. Another notable contribution by Manavalan et al., called MLCPP2.0[156], encompassed a stacking ensemble of models to augment the prediction of CPP uptake efficiency.

Innovations in deep learning have also made a significant impact. Zhang et al. proposed a Siamese neural network augmented by a contrastive learning optimization module for the discrimination between CPPs and non-CPPs[157]. This novel integrated approach has demonstrated exceptional performance and represents a reliable method for differentiating CPPs from non-CPPs. In the same vein, the SVM-based in silico tool CellPPD stands as a practical application allowing users to generate all possible mutations and predict cell penetration[158].

Despite significant advancements in the field, the importance of interpretable and explainable ML models cannot be overstated. These models not only offer pivotal insights into the performance of CPPs but also reveal the underlying patterns and key factors influencing their penetration capabilities. Their ability to explain their predictions in understandable terms goes

a long way in promoting the understanding of CPPs, ultimately paving the way for the development of better drug delivery systems.

Here, we present LightCPPgen, an interpretable computational pipeline for virtual screening and generation of CPPs. To achieve interpretability and explainability, the ML model is designed to provide clear and tangible rationales for its predictions. The generative component of LightCPPgen employs a constrained optimization approach powered by a GA. This algorithm optimizes non-penetrating peptides by maximizing their similarity to an input sequence while ensuring that the generated sequences meet specific penetration criteria. The constraint ensures that the generated peptides are predicted to exhibit cell-penetrating properties, as determined by the ML model. Crucially, the pipeline also offers an interpretable explanation for why each generated sequence is classified as cell-penetrating, providing valuable insights into the predictive process.

5.1. Material and Methods

5.1.1. Dataset Description

In this study, we employed the layer 1 training and testing datasets of MLCPP2.0[156]. The training dataset was derived combing six existing datasets from C2Pred[148], CellPPD[158], CPPred-RF[147], KELM-CPPpred[145], MLCPP[144], and BChemRF-CPPred[155]. Initially, positive, and negative samples (CPPs and non-CPPs) were independently grouped. Subsequently, the CD-HIT[159] algorithm was applied to the positive samples, excluding CPPs that shared more than 85.0% sequence identity with other CPPs. This refinement process resulted in a set of 573 CPPs. Analogously, non-CPPs were processed, excluding sequences that shared >85.0% sequence identity with other non-CPPs or CPPs. To achieve a balanced training set, 573 non-CPPs were randomly selected. The positive samples of the test dataset were collected from the CPP entries of the CellPPD[158], KELM-CPPpred[145], MLCPP[144], and BChemRF-CPPred[155] independent datasets, the CPPsite 2.0[160], and Basith et al. CPP methods evaluation[161]. To ensure dissimilarity with the training dataset, sequences sharing high similarity were excluded using a CD-HIT[159] cutoff of 90.0%, yielding a total of 157 CPPs. Instead, the negative samples for the test dataset were collected from existing methods, excluding sequences with more than 70% sequence identity with the training samples, resulting in 2184 non-CPPs.

5.1.2. LightGBM and Cost-sensitive learning

A Cost-sensitive learning is a paradigm in ML where different types of prediction errors incur different costs[162]. Cost-sensitive learning differs from standard learning approaches by focusing on minimizing a cost function that accounts for the real-world impact and severity of different types of errors. This method is especially relevant in scenarios where the consequences of errors are asymmetrical, for instance, misclassifying a positive example as negative may carry a significantly different cost compared to misclassifying a negative example as positive.

In the CPP field, cost-sensitive learning is employed to prioritize minimizing false positives when designing novel CPPs. Balancing the reduction of false positives with the risk of increasing false

negatives is the key to ensure that the model remains effective in identifying true CPPs without missing potential discoveries. To address this point, a class weighting mechanism is incorporated into the cost function to control how the model learns during training. For instance, in the case of binary classification tasks where the log-likelihood function is used as the cost function:

$$L = \sum_{i=1}^{n} \left[y^{(i)} \ln\left(P\left(y^{(i)}\right)\right) + \left(1 - y^{(i)}\right) \ln\left(1 - P\left(y^{(i)}\right)\right) \right]$$

where $P\left(y^{(i)}\right)$ denotes the predicted probability that $y^{(i)}$ equals 1, the cost function can be modified by introducing asymmetric costs C_{FP} and C_{FN} for false positives and false negatives, respectively:

$$L = \sum_{i=1}^{n} \left[C_{FP} y^{(i)} \ln\left(P\left(y^{(i)}\right)\right) + C_{FN}\left(1 - y^{(i)}\right) \ln\left(1 - P\left(y^{(i)}\right)\right) \right].$$

In LightGBM, the modification of the cost function can be achieved by turning the hyperparameter called `scale_pos_weight`. This hyperparameter, typically used to adjust for class imbalance by giving greater weight to the positive class, can be set to a value less than the default value of 1.0 to prioritize reducing false positives in binary classification tasks. This unique approach shifts the model's focus, making it more conservative in predicting positives, thereby reducing the occurrence of false positives.

5.1.3. Feature engineering

In alignment with established literature[155,156,163], our investigation focuses on two principal categories of molecular descriptors. The first category includes descriptors based on the molecular structure, which capture the physicochemical properties of peptide. These descriptors are generated using the RDKit[164], Biopython[165] libraries and homemade scripts.

Several structure-based descriptors, including molecular weight (MW), the number of rotatable bonds (NRB), topological polar surface area (tPSA), fraction of sp3-hybridized carbon atoms (Fsp3), octanol-water partition coefficient (cLogP), number of aromatic rings (NAR), number of hydrogen bond donors (HBD) and acceptors (HBA), number of primary amino groups (NPA), number of guanidinium groups (NG), and the net charge (NetC) was calculated using RDKit library. Additionally, we calculated the isoelectric point (IsoP) and hydrophobicity using the Python Peptides library, and aromaticity with the Biopython library[165]. Finally, we computed the length of the peptide sequence.

The second category includes sequence-based descriptors, directly derived from the peptide's primary amino acid sequence, employing the iFeatureOmega library[166]. Among these, various descriptors was calculated, such as amino acid composition (AAC), enhanced amino acid composition (EAAC), composition of k-spaced amino acid pairs (CKSAAP), di-peptide composition (DPC), tri-peptide composition (TPC), dipeptide deviation from expected mean (DDE), grouped amino acid composition (GAAC), enhanced GAAC (EGAAC), composition of k-

spaced amino acid group pairs (CKSAAGP), grouped di-peptide composition (GDPC), grouped tri-peptide composition (GTPC), Moran correlation, Geary correlation, normalized Moreau-Broto autocorrelation (NMBroto), the composition, transition, and distribution set of features (CTDC, CTDT, CTDD), conjoint triad (CTriad), k-spaced conjoint triad (KSCTriad), sequence-order-coupling number (SOCNumber), quasi-sequence order (QSOrder), pseudo-amino acid composition (PAAC), amphiphilic pseudo-amino acid composition (APAAC), adaptive skip dipeptide composition (ASDC), auto covariance (AC), cross covariance (CC), auto-cross covariance (ACC), AAindex (AAINDEX), BLOSUM62, and Z-scale (ZSCALE).
A comprehensive set of 13,833 descriptors was obtained to fully explore the feature space.

5.1.4. Feature selection

The feature selection procedure was utilized to reduce the initial set of 13,833 descriptors to a smaller set of 20 features. The initial extensive feature set allows for a deep and potentially novel exploration of correlations. However, it also presents several challenges, including:

1. *Computational Load*: Computing a vast array of features, along with training and predicting with the model, imposes significant demands on computational speed and memory resources.
2. *Risk of Overfitting*: A large feature set increases the risk of the model capturing noise instead of the true underlying patterns, which can result in poor generalization to new data.
3. *Complex Interpretability*: A large number of features can make it challenging to discern the key factors driving the model's predictions, complicating interpretation.

In this sense, feature selection emerges as a crucial step in this research to solve the above-mentioned challenges.
This study employed a hybrid feature selection approach that combines MDI feature selection (an embedded method) with greedy forward search (a wrapper method). Initially, a LightGBM model is trained using the entire feature set P, and the MDI feature importance I_j computed internally by the model for each feature $j \in P$ is used to prune the feature set to $\tilde{P} = \{j \in P : I_j > 0\}$. This initial preprocessing step substantially reduces the size of the feature pool, allowing for a more efficient implementation of the subsequent greedy forward search.
Finally, from the reduced feature set $\tilde{P}$, the greedy forward search method is used to find the final feature subset $S^* = 20$ as follows:

1. Start with an empty feature set $S = \emptyset$.
2. Repeat {
    a. Generate candidate subsets $S_j = S \cup \{j\}$ for each $j \notin S$.
    b. For each $j \notin S$ train a model using S_j and estimate its generalization error.
    c. Update S to the best S_j found in step b. }

3. Output the best feature subset S^* evaluated during the entire search procedure, or until a stopping criterion is met (e.g., maximum number of features to select).

5.1.5. Validation strategy

To evaluate the out-of-sample performance of the model during hyperparameter tuning and feature subset selection, we employed stratified k-fold cross-validation. This method divides the training set into k smaller subsets (or folds) using stratified sampling, ensuring that the class distribution within each fold reflects the overall class proportions of the dataset. During the process, each fold is used as a validation set exactly once, while the model is trained on the combined data from the remaining $k-1$ folds. The overall performance is then determined by averaging the results across all k folds.
For the final performance evaluation and comparison with models reported in the literature, we used the independent test dataset from MLCPP 2.0, assessing model performance through metrics such as accuracy ACC, SPC, SVN, MCC, and AUC.

5.1.6. Genetic algorithm details

Starting from a non-CPP sequence, LightCPPgen aims at generating a new penetrating amino acid sequence satisfying the following key features:

- It should be "close" to the original non-penetrating peptide to preserve as much as possible its biological and physicochemical properties (i.e., structural stability and solubility). Closeness is measured according to the similarity score obtained from the global pairwise sequence alignment with the BLOSUM62 substitution matrix, using Biopython library[165]. From a bioactivity perspective, minimizing the number of mutations helps reduce the risk of unwanted side effects and facilitates synthesis.
- It should achieve the highest possible probability of cell-penetration while minimizing the number of mutations introduced. The focus of LightCPPgen is to address the balance between reducing the number of mutations while simultaneously enhancing the likelihood of cell penetration.
- It should exhibit a low level of non-novelty relative to the dataset used for training the LightGBM surrogate model. This constraint minimizes exploration into regions with lower model confidence, reducing the likelihood of unreliable predictions with the aim of enhancing the success rates in experimental validation. The LightGBM model acts as a surrogate during the design process, predicting the penetration probability of a given peptide. Consequently, ensuring the model's high reliability is essential for the success of the design process.

According to the above considerations, the following performance metric is adopted as a criterion to guide the design, and thus used as a fitness function to be minimized by a genetic algorithm:

$$\mathrm{J}(\theta) = \mathrm{w}_1 \mathrm{d}_{sim(\theta,\theta_0)} + \mathrm{w}_2 \max\left(0, \left(1 - \mathrm{p}(\mathrm{y} = 1|\theta)\right)^2 - 0.2^2\right) + \mathrm{w}_3 \max\left(0, \mathrm{LOF}_{\mathrm{k}(\theta)}{}^2 - 1.5^2\right)$$

In the definition of the fitness function $\mathrm{J}(\theta)$:

- θ represents the vector of design parameters, which is the ordered sequence of amino acids that define the peptide. Each element of this vector corresponds to an amino acid, represented by a single Latin letter according to the one-letter code system (with an alphabet of 20 characters, one per each natural amino acid), turning the design into a combinatorial optimization problem. The length of the vector θ corresponds to the length of the amino acid sequence, which is typically predetermined by the user.
- θ_0 is a given vector representing the amino acid sequence of the initial non-penetrating peptide.
- $\mathrm{d}_{sim(\theta,\theta_0)}$ is the similarity distance between the candidate and the initial peptide.
- $\mathrm{p}(\mathrm{y} = 1|\theta)$ is the probability of a candidate peptide (characterized by the amino acid sequence θ) to be penetrating.
- $\mathrm{LOF}_{\mathrm{k}(\theta)}$ is the value of the Local Outlier Factor for the candidate peptide.
- $\mathrm{w}_1, \mathrm{w}_2$, and w_3 represent the weights by which it is possible to modify the impact of one or more factors in the fitness function $\mathrm{J}(\theta)$. In this specific application, they are all set to 1.

According to the above definition of the fitness function $\mathrm{J}(\theta)$, the similarity distance $\mathrm{d}_{sim(\theta,\theta_0)}$ between the candidate and initial peptide is minimized. Furthermore, polynomial barrier functions are used to penalize $\mathrm{LOF}_{\mathrm{k}(\theta)}$ novelty values larger than 1.5 and the probability of penetrability lower than 0.8. The value of 1.5 for the $\mathrm{LOF}_{\mathrm{k}(\theta)}$ was chosen to prioritize the reliability of predictions by penalizing solutions that lie outside the feature space of the training dataset. Ensuring a low $\mathrm{LOF}_{\mathrm{k}(\theta)}$ value helps maintain the model's predictive reliability, as higher LOF values indicate that the solution deviates from the data distribution used during training. Similarly, the threshold of 0.8 for the penetrability score was selected to emphasize the generation of highly penetrating solutions. By penalizing solutions with a predicted penetrability score below 0.8, we ensure that the optimization process prioritizes candidates with a stronger likelihood of exhibiting cell-penetrating properties.
Users can adjust the weights ($\mathrm{w}_1, \mathrm{w}_2$, and w_3) assigned to the similarity distance $\mathrm{d}_{sim(\theta,\theta_0)}$, penetrability score $\mathrm{p}(\mathrm{y} = 1|\theta)$, and $\mathrm{LOF}_{\mathrm{k}(\theta)}$ components of the fitness function based on their specific optimization requirements. Increasing a weight above 1 enhances the influence of that term in the fitness function, while reducing it below 1 decreases its impact.

The genetic optimization algorithm above described is used to minimize the fitness function $J(\theta)$ over the design parameters θ. A schematic depiction of the CPP design algorithm is shown in Figure 15.

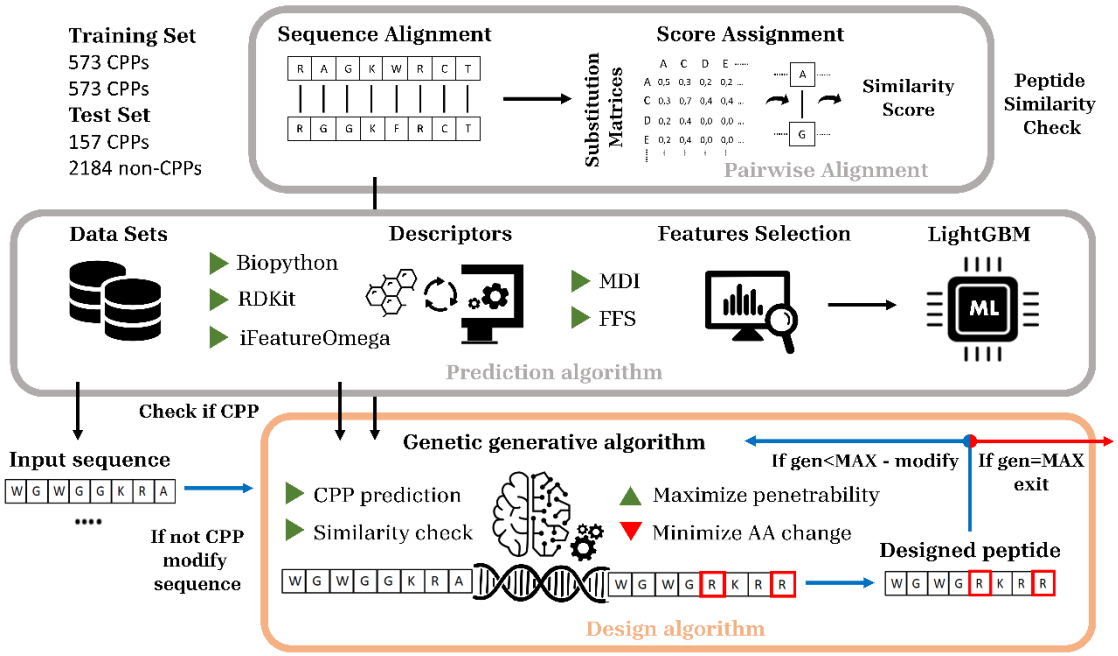


Figure 15: Schematic representation of the overall hybrid framework combining Genetic algorithm with Machine Learning predictor and Pairwise alignment algorithm. RDKit: Open-Source Cheminformatics Software; iFeatureOmega: Feature engineering tool; MDI: Mean Decrease in Impurity; FFS: Forward feature selection.

5.2. Results

In this section, we present the results of the de novo CPP design using a synergistic approach that combines ML with GA. Specifically, we detail the performance of the LightGBM-based predictor (Sections 5.2.1 and 5.2.2), the explainability and interpretability of the model (Section 5.2.3), and the design algorithm (Section 5.2.4). To evaluate the effectiveness of the design algorithm, we utilized non-CPP samples from the independent MLCPP2.0 dataset.

5.2.1. Modeling and feature selection results

This work introduces LightCPP, a LightGBM-based binary classification model designed to distinguish between CPP and non-CPP sequences. The numerical results presented in this section were obtained by evaluating the algorithm's performance on Layer 1 of the independent dataset from the publicly available MLCPP 2.0 dataset. Following the presentation of results from the feature selection phase, details of the training process are provided at the end of this section.

The initial LightGBM model was trained using a complete feature set consisting of 13,833 variables. Utilizing the feature importance scores derived from the MDI method, irrelevant

features were identified and removed, reducing the feature set to 375 variables. This significant reduction facilitated the execution of forward search within a practical timeframe. To benchmark model performance before and after sequential feature selection, the intermediate model, referred to as LightCPP (375 features), achieved performance metrics of 0.930, 0.962, 0.690, 0.981, and 0.687 for AUC, ACC, SNV, SPC, and MCC, respectively.

Subsequently, a greedy forward search was performed to identify an optimal subset of features, using MCC as the evaluation metric. The final model, LightCPP (20 features), achieved performance metrics of 0.909, 0.960, 0.690, 0.979, and 0.677 for AUC, ACC, SNV, SPC, and MCC, respectively. Compared to the intermediate model, there was a slight decline in metrics of 2.26%, 0.21%, 0.0%, 0.20%, and 1.46% for AUC, ACC, SNV, SPC, and MCC, respectively. Despite this minor reduction in performance, the final model highlights the effectiveness of the feature selection algorithm by achieving competitive metrics with a significantly reduced feature set. This process resulted in a final subset of 20 features, detailed below:

1. **CTDC_charge.G1**: Presence of the amino acids of the first group (i.e., arginine and lysine) along the peptide sequence computed through the composition transition distribution (CTD) descriptor. This feature provides an indication of the amount of positively charged residues along the peptide chain.
2. **CTDD_normwaalsvolume.2.residue100**: CTDD (CTD-Distribution) descriptor following the Normalized van der Waals volume attribute of the second group of amino acids (i.e., asparagine, valine, glutamic acid, glutamine, isoleucine, leucine). In detail, it corresponds to a fraction of the entire sequence, from the location of the first residue of the group mentioned above, until 100% of occurrences are contained.
3. **APAAC_Pc2.Hydrophobicity.1**: Amphiphilic pseudo amino acid composition[167] (APAAC) characterizes the hydrophobicity-hydrophilicity balance and the sequence-order correlation of the adjacent residues considering their hydrophobicity property.
4. **PAAC_Xc1.V**: Pseudo-Amino Acid Composition (PAAC) captures information about the physicochemical properties (hydrophobicity, hydrophilicity, and side chain mass) of the valine amino acids content and its distributions in protein sequences.
5. **CTDD_hydrophobicity_ARGP820101.1.residue50**: CTDD (CTD-Distribution) descriptor following the Hydrophobicity attribute of the first group of amino acids (i.e., glutamine, serine, threonine, asparagine, glycine, aspartic acid, and glutamic acid), according to the Argos'[168] (ARGP820101) AAindex. In detail, it corresponds to a fraction of the entire sequence, from the location of the first residue of the group mentioned above, until 50% of occurrences are contained.
6. **CTDD_hydrophobicity_FASG890101.2.residue50**: CTDD (CTD-Distribution) descriptor following the Hydrophobicity attribute of the second group of amino acids (i.e., asparagine, threonine, proline, glycine) in the CTD descriptors, according to the Fasman'[169] (FASG890101) AAindex. In detail, it corresponds to a fraction of the entire sequence, from the location of the first residue of the group mentioned above, until 50% of occurrences are contained.

7. **KSCTriad_g5.g5.gap1**: Conjoint k-spaced Triad (KSCTriad) calculates the numbers of three amino acids of the g5 group (arginine and lysine) that are separated by 1 residue (gap1).
8. **APAAC_Pc1.A**: Amphiphilic pseudo amino acid composition[167] (APAAC) characterizes the hydrophobicity-hydrophilicity balance and the sequence-order correlation of the peptide. In this case, it refers specifically to the alanine amino acid, meaning that its content and distribution along the peptide sequence are important to predict the penetration capability.
9. **DDE_LA**: Dipeptide deviation from expected mean[170] (DDE) of the leucine-alanine dipeptide inside the entire peptide sequence. This descriptor is constructed from three parameters, namely, the dipeptide composition (Dc), the theoretical mean (Tm), and the theoretical variance (Tv). In detail, Tm and Tv depend on the amino acid types and the peptide length, while Dc depends on the "LA" dipeptide occurrences and the peptide length.
10. **DDE_LP**: Dipeptide deviation from expected mean (DDE) of the leucine-proline dipeptide within the entire peptide sequence.
11. **ASDC_DM**: Adaptive skip dipeptide composition[171] (ASDC) of aspartic acid and methionine. ASDC is a modified dipeptide composition descriptor, which also considers the correlation information present between adjacent and non-adjacent aspartic acid and methionine residues.
12. **CKSAAP_IR.gap2**: Composition of k-spaced Amino Acid Pairs (CKSAAP) of the amino acid pair isoleucine-arginine separated by 2 residues (gap2).
13. **QSOrder_Grantham.Xr.T**: Quasi-sequence-order (QSOrder) descriptor that characterizes the sequence order and the spatial relationships of the threonine utilizing the Grantham's[172] distance matrix.
14. **CKSAAP_LL.gap4**: Composition of k-spaced Amino Acid Pairs (CKSAAP) of the amino acid pair leucine-leucine separated by 4 residues (gap4).
15. **CKSAAGP_positivecharger.uncharger.gap1**: Composition of k-Spaced Amino Acid Group Pairs[166] (CKSAAGP) of the amino acids belonging to the positive charged (lysine, arginine, histidine) and un-charged (serine, threonine, cysteine, proline, asparagine, glutamine) groups separated by 1 residues (gap1).
16. **Zscale_p1.z5**: Zscale descriptor developed by Sandberg et al. in 1998[173] that characterizes the amino acids at position 1 (p1) with the fifth physicochemical scale (z5).
17. **CKSAAGP_uncharger.aromatic.gap9**: Composition of k-Spaced Amino Acid Group Pairs[166] (CKSAAGP) of the amino acids belonging to the un-charged (serine, threonine, cysteine, proline, asparagine, glutamine) and aromatic (phenylalanine, tyrosine, tryptophan) groups separated by 9 residues (gap9).
18. **PAAC_Xc1.G**: Pseudo-Amino Acid Composition (PAAC) captures information about the physicochemical properties (hydrophobicity, hydrophilicity, and side chain mass) of the glycine amino acids and its distributions in protein sequences.
19. **ASDC_AS**: Adaptive skip dipeptide composition[171] (ASDC) of alanine and serine. ASDC is a modified dipeptide composition descriptor, which also considers the correlation information present between adjacent and non-adjacent alanine and serine residues.
20. **CKSAAP_TR.gap4**: Composition of k-spaced Amino Acid Pairs (CKSAAP) of the amino acid pair threonine-arginine separated by 4 residues (gap4).

5.2.2. Comparison with the state-of-the-art predictors

The LightCPP model metrics were evaluated by comparing them with state-of-the-art models reported in the literature. The comparison specifically focused on models that utilize the independent dataset from MLCPP 2.0 as a test set, including C2Pred, BChemRF-CPPred, MCLPP, MLCPP 2.0 (Layer1), and SiameseCPP. In detail, the comparison was made using Layer1 of the independent dataset from MLCPP 2.0, referencing the performance metrics listed in Table 2 from Zhang et al. 2023[157] and Table 10 from Manavalan et al.[156]. The numerical results are summarized in Table 1, where the best and second-best performances for each metric across all models are highlighted in bold and underlined, respectively.

Table 1: Performance detail of our proposed LightCPP in comparison with C2Pred, BChemRF-CPPred, MLCPP, MLCPP2.0, and SiameseCPP on the Layer1 of the independent dataset from MLCPP 2.0.

Model	*AUC*	*ACC*	*SNV*	*SPC*	*MCC*
C2Pred	0.867	0.781	0.790	0.781	0.326
BChemRF-CPPred	0.914	0.893	0.745	0.903	0.467
MLCCP	0.920	0.892	<u>0.809</u>	0.898	0.497
MLCCP2.0 (Layer1)	<u>0.928</u>	0.934	**0.847**	0.940	0.624
SiameseCPP	-	0.959	0.624	**0.983**	0.652
LightCPP (375 features)	**0.930**	**0.962**	0.690	<u>0.981</u>	**0.687**
LightCPP (20 features)	0.909	<u>0.960</u>	0.690	0.979	<u>0.677</u>

The data in Table 1 reveal that no single model consistently outperforms the others across all evaluation metrics. However, the models introduced in this section demonstrate the best performance for AUC, ACC, and MCC (LightCPP with 375 features) and the second-best performance for ACC and MCC (LightCPP with 20 features). The models most comparable to the proposed ones are MLCPP 2.0 (Layer1) and SiameseCPP, which need a more detailed comparative analysis.

When comparing the proposed models with MLCPP 2.0 (Layer1), the AUC values are similarly high across all classifiers, exceeding 0.9, with the 20-feature model performing approximately 2% lower than the others. For ACC, the proposed models outperform MLCPP 2.0 (Layer1) by nearly 3%. However, these high ACC values may indicate a bias toward predicting the majority class (non-CPPs), which aligns with the training strategy of penalizing false positives more heavily than false negatives. In terms of SPC and SNV, the LightCPP models excel at minimizing false positives, while MLCPP 2.0 (Layer1) performs better at identifying positive cases, reducing false negatives. Additionally, the proposed models achieve higher MCC values, reflecting superior overall performance across all quadrants of the confusion matrix.

In comparison with SiameseCPP, the proposed models show comparable ACC scores. However, in terms of SNV, the proposed models perform nearly 10% better, demonstrating a stronger ability to correctly identify CPPs. While SiameseCPP achieves slightly higher SPC, the difference

is minimal. Finally, the MCC values indicate that the LightCPP models provide more balanced performance, effectively managing both positive and negative classifications.

5.2.3. Explainability of LightCPP

Understanding the importance and significance of the underlying features driving ML-based predictions is essential for their effective application in CPP design. The 20 features that govern the predictive performance of LightCPP (20 features, as listed in Table 1 and detailed in Section 5.2.1) provide valuable insights into the model's predictive capabilities. The feature importance and impact, computed with SHAP feature importance on the independent set of MLCPP 2.0 are shown in Figure 16.

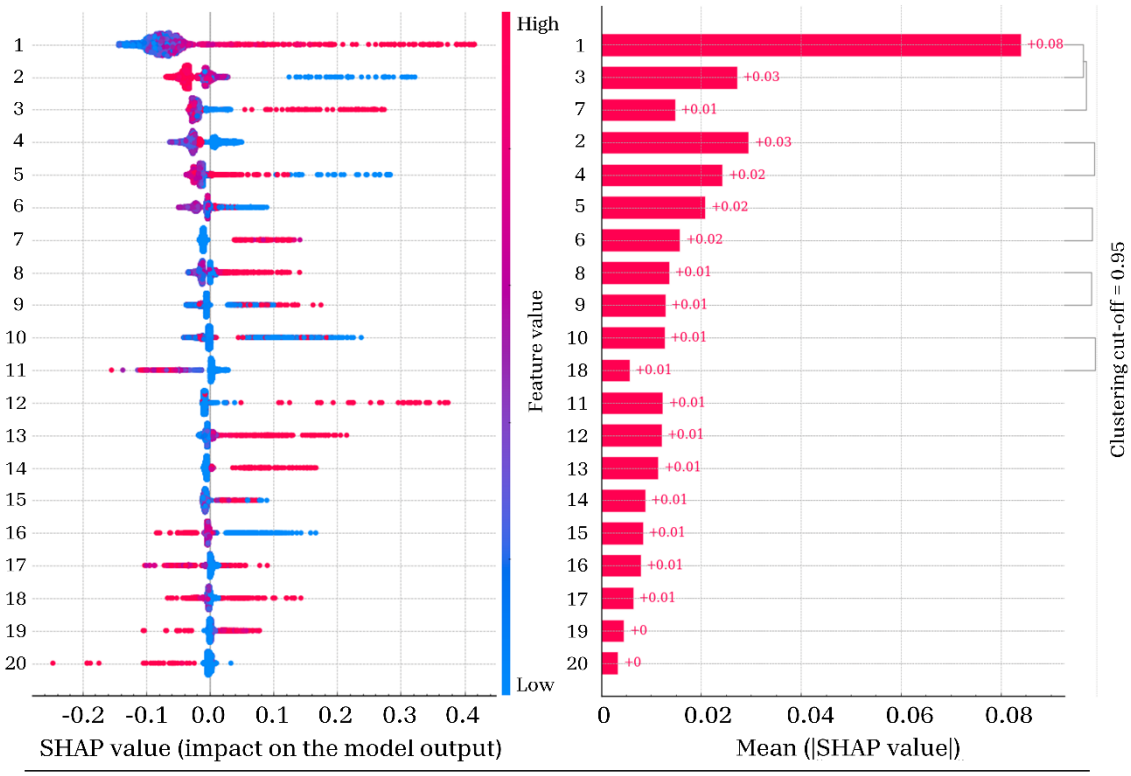


Figure 16: Feature importance and impact as computed with SHAP on the independent set of MLCPP 2.0. The right bar plot ranks the variables by their average impact on model prediction. The left dot plot shows each data point with the signed contribution of each feature. Blue dots indicate low variable values, while red dots indicate high values.

The CTDC charge feature (CTDC_charge.G1) highlighted the critical role of positively charged amino acids in predicting CPP behavior. High positive charge values were consistently associated with CPP classification, while low values indicated non-penetrating peptides. Another pivotal feature influencing CPP prediction was the distribution of specific amino acids, including asparagine, valine, glutamic acid, glutamine, isoleucine, and leucine, along the peptide sequence (CTDD_normwaalsvolume.2.residue100). Low values of this feature are associated with CPP classification, suggesting either the absence of these amino acids or their concentration at the beginning of the sequence. Conversely, high values led to non-CPP classification, which is indicative of the presence of these residues near the C-terminal tail.

The amphiphilic pseudo amino acid composition (APAAC), which accounts for the balance between hydrophobicity and hydrophilicity as well as sequence-order correlation based on adjacent residues' hydrophobicity (APAAC_Pc2.Hydrophobicity.1), revealed that both high and low hydrophobicity levels significantly impact cell penetration. This underscores the essential role of hydrophobic distribution in mediating peptide-cell interactions. Similarly, the Conjoint k-spaced Triad (KSCTriad) method, which evaluates arginine and lysine patterns within the peptide sequence (KSCTriad_g5.g5.g5.gap1), emerged as another key feature. Frequent occurrences of arginine and lysine triplets separated by a single residue were strongly associated with CPP activity, highlighting the significance of these patterns in facilitating penetration.

The alanine APAAC also proved to be informative for distinguishing CPPs from non-CPPs. High alanine APAAC values were indicative of CPP activity, suggesting that a higher and more uniform distribution of alanine enhances membrane interaction and uptake. Conversely, lower alanine APAAC values reflected a sparse distribution and an unfavorable hydrophobic-hydrophilic balance, which is correlated with non-CPP classification. Additionally, the valine presence (PAAC_Xc1.V) played a crucial role in CPP classification. Low valine levels were associated with CPP classification, suggesting that a minimal or localized presence of valine favors penetration. On the other hand, high concentrations of valine, especially if unevenly distributed, hindered membrane penetration, highlighting valine's complex influence on CPP functionality.

Three additional features significantly contributed to CPP prediction: CKSAAP_IR.gap2, CKSAAP_LL.gap4, and CKSAAP_TR.gap4. These features correspond to amino acid pairs separated by specific gaps. For the CKSAAP_IR.gap2 feature, representing isoleucine-arginine pairs separated by two residues, higher occurrences correlated with CPP classification, while lower occurrences aligned with non-CPP classification, emphasizing the importance of this specific pair for penetration capability. Similarly, the CKSAAP_LL.gap4 feature, representing leucine-leucine pairs separated by four residues, showed that high values aligned with CPP predictions, whereas low values indicated non-CPP predictions. In contrast, CKSAAP_TR.gap4, which relates to threonine-arginine pairs separated by four residues, was associated with non-CPP predictions when occurring at high frequencies, suggesting that such spatial arrangements may inhibit cell entry. Minimal occurrences of this specific pair had no significant impact on predictions.

These observed physicochemical patterns, including amino acid gaps of varying lengths, highlight the intricate interplay between charge distribution, hydrophobicity, and spatial arrangement within peptide sequences that collectively influence membrane penetration capability. The feature importance and impact for the training set are presented in Figure S1 of

the Supporting information section 1. Notably, the feature importance analysis demonstrated remarkable consistency between the training and testing sets, with only slight variations in the order of features. This consistency underscores the robustness and relevance of the identified features across diverse peptide sequences.

5.2.4. Design Algorithm Analysis

The effectiveness of the design algorithm was evaluated using all the non-CPP sequences from the MLCPP2.0 testing set as input for the optimization loop. During the design process, the LightCPP model (based on 20 features) served as a surrogate predictor, estimating the penetration probability of each peptide generated by the optimization algorithm. Furthermore, the population and generation parameters were set to 500 and 50, respectively (see Supporting information section 2 for further details). The closeness of the designed peptide sequence to the original one is measured according to the similarity score obtained from the global pairwise sequence alignment with the BLOSUM62 substitution matrix[174]. The main evaluation criteria focus on the amino acid pairs and types modified between the original and optimized sequences. The design of new penetrating sequences is accomplished by substituting certain amino acids with others, aiming to maximize penetrability while maintaining similarity to the original peptide. The substitution patterns are illustrated in the heatmap matrix in Figure 17, where each entry represents the frequency of amino acid pair substitutions between the original and optimized sequences. Additionally, Figure 17-right displays the percentages of each type of amino acid deleted, while Figure 17-top shows the percentages of each type of amino acid inserted. These figures were derived by summing the columns and rows of the substitution matrix, respectively.

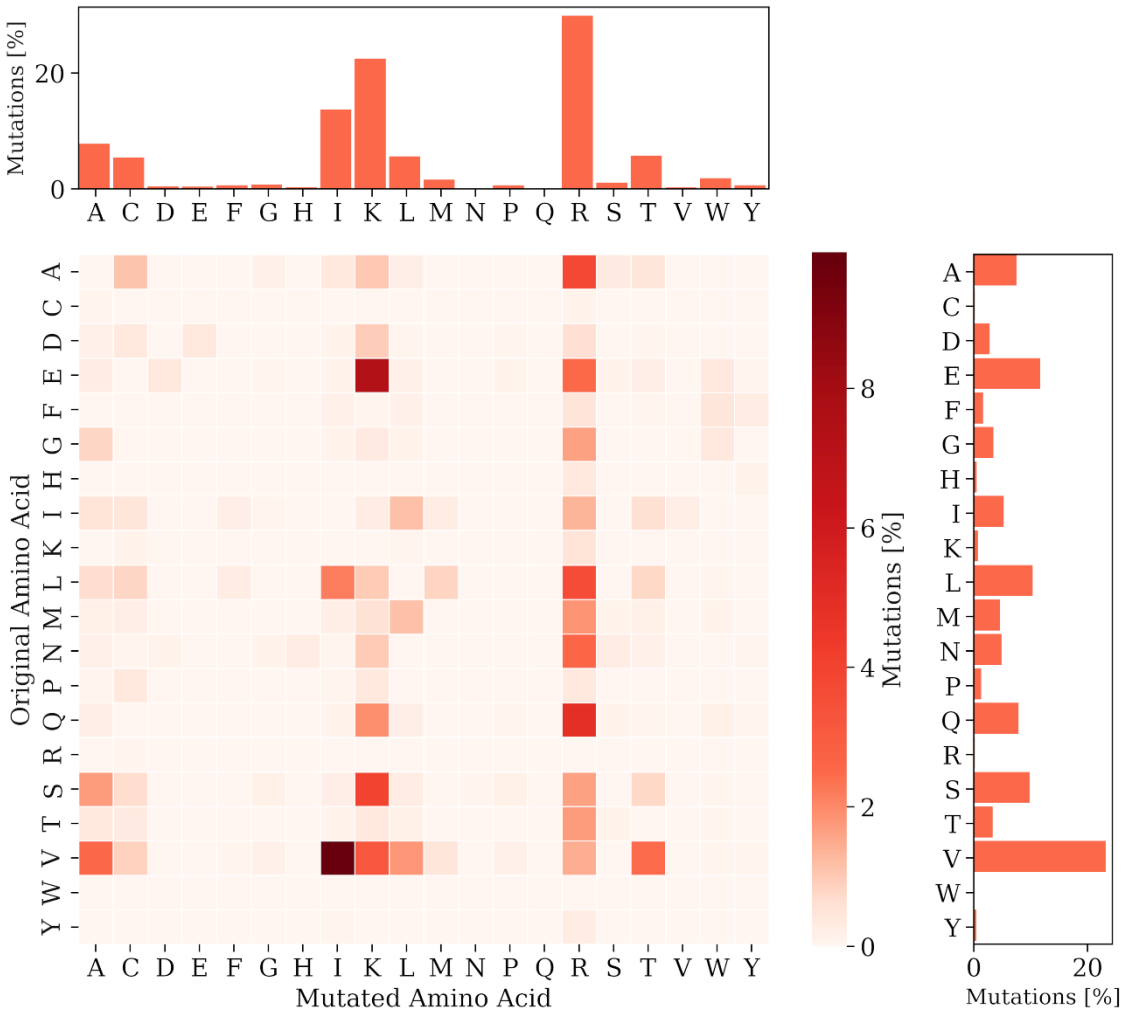


Figure 17: The heatmap shows the original replaced amino acids in the matrix's rows and the newly inserted amino acids in the matrix's columns, with the results of the substituted amino acid pairs shown as percentages. Additionally, the top figure shows the percentages of each newly inserted type of amino acid, while the right figure displays the percentages of each deleted type of amino acid.

Notably, cationic amino acids (R and K) were the most frequently inserted into the newly designed CPPs, with arginine being favored over lysine, underscoring its critical role in peptide penetrability. This importance is further emphasized by the fact that arginine was almost never removed from the original sequences (Figure 17-right). Conversely, anionic amino acids (D and E) were rarely inserted into the newly designed CPPs. This trend aligns with the key role of positive charge in CPP functionality, as identified previously by the ML model (Figure 16). To emphasize this point, aspartic and glutamic acids were often removed from the original sequences, especially for the glutamic acid (Figure 17-right). Together, these observations highlight the promotion of more positively charged sequences, consistent with the predictive importance of positive charge in our model.

Positively charged amino acids are crucial for the cell penetration ability of peptides, but hydrophobic amino acids and their distribution along the sequence also play an important role. Indeed, hydrophobic amino acids such as alanine, cysteine, isoleucine, and leucine (A, C, I, and

L) were frequently incorporated into the optimized CPP sequences (Figure 17-top). It is worth noting that cysteine was moderately used as a substitute and was almost never deleted from the original sequences, suggesting its potential key role in enhancing the penetration capability of CPPs.

Valine emerged as the most commonly substituted amino acid in the newly designed sequences, not only within the hydrophobic amino acid group but across all amino acid types. Feature importance and impact analyses (Figure 16) indicate that a high content of valine in the sequence correlates with non-CPP predictions (PAAC_Xc1.V). Therefore, it is reasonable to mention that the optimization algorithm reduces valine content to shift the prediction towards the CPP class. Figure 17 further shows that valine is predominantly replaced by isoleucine, a pattern that warrants further investigation and will be explored in more detail in the discussion section.

Another notable finding is the consistent addition and removal of alanine, isoleucine, and leucine in the designed CPPs (Figure 17). This behavior exhibited by the optimization algorithm highlights not only the importance of these specific amino acids but also suggests that their distribution within the sequence is critical for the CPP's penetration ability. These findings emphasize the complexity and specificity involved in CPP design, leading to multiple optimization strategies aimed at enhancing the peptide's penetration potential.

The SHAP analysis was also performed on a single randomly selected peptide (IF**SN**TAL**V**NC**M**R**Q**TLQDTGHNP) before and after the optimization process (IF**KR**TALINC**R**R**R**TLQDTGHNP) (Figure 18). Additional examples of SHAP analysis conducted on randomly selected peptides can be found in Figure S4, Figure S5, and Figure S6 of the Supporting Information section 3. Notably, the substituted amino acids after the optimization process caused several features to shift from non-CPP to CPP prediction, including CTDC_charge.G1, CKSAAP_IR.gap2, APAAC_Pc2.Hydrophobicity.1, and PAAC_Xc1.V, among others.

The optimization algorithm focused on enhancing the peptide's positive charge by incorporating 3 arginine and 1 lysine residues, replacing serine, asparagine, methionine, and glutamine. These substitutions led to a significant increase in the CTDC_charge.G1 feature, a key determinant for cell penetration, reinforcing the crucial role of positively charged amino acids in CPP functionality. The increase in positive charge aligns with the known importance of these residues in facilitating cellular uptake.

Additionally, the algorithm strategically reduced the valine content (PAAC_Xc1.V), which is associated with non-CPP predictions, as previously noted. In its place, valine was substituted with isoleucine, a shift observed that reinforces the transition toward a CPP classification (as observed in Figure 16). This substitution also led to an increase in the CKSAAP_IR.gap2 value, indicating a more favorable spatial arrangement of isoleucine and arginine residues, which is linked to enhanced penetration potential.

Furthermore, the algorithm paid particular attention to balancing the hydrophobicity-hydrophilicity of the peptide, as evidenced by the changes in the APAAC_Pc2.Hydrophobicity.1 feature (Figure 18). By fine-tuning this balance, the algorithm ensured that the resulting peptide would possess a more favorable hydrophobicity-hydrophilicity pattern for efficient cellular uptake.

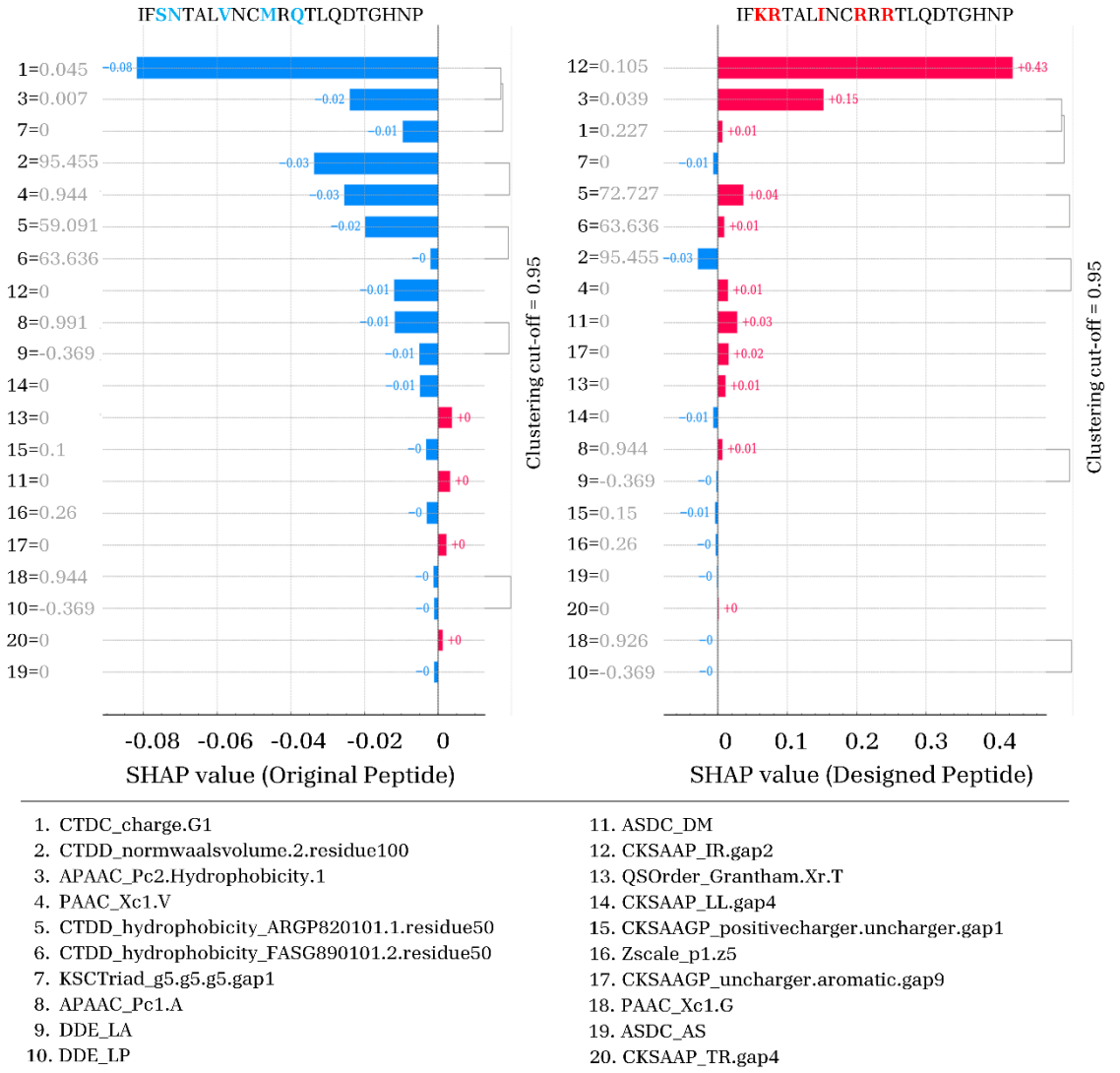


Figure 18: SHAP values for a randomly selected peptide (IFSNTALVNCMRQTLQDTGHNP) are displayed before (on the left) and after (on the right) the optimization. The original deleted amino acid is highlighted in bold blue (top-left), while the newly inserted amino acids are highlighted in bold red (top-right). The values for each feature before and after the optimization process are shown in grey.

The overall pattern of these adjustments, as reflected in the SHAP value analysis, demonstrates the algorithm's capability to make precise, targeted changes that optimize the peptide's characteristics for improved CPP functionality. This process illustrates the intricate relationship between amino acid composition, their distribution within the sequence, and the biophysical properties that enable efficient cell penetration. Through these adjustments, the optimization algorithm effectively tailors peptides for enhanced performance in penetrating cellular membranes, showcasing the value of strategic design in CPP development.

5.3. Discussion

ML methods have become essential tools in predicting and analyzing a broad range of biological processes. With the growing availability of extensive datasets on CPPs, these techniques have proven valuable in predicting and classifying peptides capable of translocating across cellular membranes[144,145,147–149,152,155–157]. However, despite the development of several ML models for forecasting cellular uptake and membrane translocation of CPPs, several challenges remain in the field. To move forward in the design of novel CPPs, it is crucial to develop interpretable models that provide a deeper understanding of the factors influencing CPP performance. Moreover, while ML predictions can differentiate between CPPs and non-CPPs, they do not support the design of new CPPs starting from a non-penetrating sequence.

The first breakthrough in the CPP design field was the development of CellPPD[158,163], which allows users to generate all possible single-substitution mutants of a given peptide sequence. Furthermore, it enables an iterative generation of additional mutants, such as second- or third-round variants, based on the selected analogues. Further progress was made by Tran et al., which utilizes a recurrent neural network (RNN) to generate CPP sequences[175,176]. However, their framework differs from LightCPPgen, as it generates CPPs starting from a random sequence prefix and extends it to a fixed length (e.g., 20 residues), rather than beginning with a non-penetrating peptide[175].

The LightCPPgen framework enables targeted modifications with minimal amino acid changes, aiming to overcome the limitations (e.g., trial-and-error approaches) of traditional experimental design methods. While similar approaches have been used in the design of AMPs[177,178], this is the first instance where an ML model built on a GA is used to design CPPs from non-penetrating sequences. This innovative approach focuses on optimizing the candidate peptides' penetrability score while retaining as much as possible of the original peptide sequence through the similarity distance. By maximizing similarity with the original non-penetrating macromolecule, the modified peptides are likely to retain the original biological and physicochemical properties that are crucial for their intended function, while also gaining the ability to cross cellular membranes. This minimally invasive modification strategy helps avoid potential negative effects of more radical sequence alterations, reducing the risk of compromising peptide stability and functionality. Additionally, this strategy enhances the integration of the newly designed peptides into existing therapeutic frameworks, increasing their clinical potential.

The lightGBM-based CPP predictor embedded in our optimization algorithm demonstrated high accuracy and efficiency in distinguishing CPPs from non-CPPs, with performance metrics (Table 1) comparable to those of existing models like MLCPP2.0[156] and SiameseCPP[157], while utilizing just 20 features (section 5.2.1). Feature importance analysis revealed that the net positive charge of CPPs is a key predictor of their penetrability, in line with previous studies showing that positive charge enhances cellular translocation (CTDC_charge.G1, as shown in Figure 16)[179–182]. Moreover, the positive charge of CPPs is primarily involved into the membrane translocation via various pathways, including direct penetration, macropinocytosis, and clathrin-mediated endocytosis[117,183–186]. Despite the net positive charge is of paramount importance, its

distribution within the peptide sequence is also crucial[181,187], as indicated by the KSCTriad_g5.g5.g5.gap1 feature.

In addition, peptides with certain amino acids, such as asparagine, valine, glutamic acid, leucine, and isoleucine, at the C-terminal end were more likely classified as non-CPPs (CTDD_normwaalsvolume.2.residue100, as shown in Figure 16). The presence of valine, leucine, and isoleucine highlights the critical role of hydrophobic residues within the peptide sequence in facilitating its penetration ability. Indeed, the presence of hydrophobic amino acids in CPPs was found to play significant roles in the penetration process (APAAC_Pc2.Hydrophobicity.1, APAAC_Pc1.A and CKSAAP_LL.gap4 features in Figure 16). This supports recent findings that incorporating hydrophobic moieties into arginine-rich CPPs enhances their cell penetration ability[188]. This is further supported by the CKSAAP_IR.gap2 feature (Figure 16), which captures specific patterns involving isoleucine-arginine pairs spaced by 2 residues. In summary, the interplay between hydrophobic and cationic amino acids with specific patterns results to be a crucial characteristic for enhancing CPPs' ability to cross cellular membranes, which in agreement with recent literature[188,189].

By combining an ML classifier with a generative GA model, the algorithm systematically optimized non-CPP sequences in CPP sequences. The optimization algorithm's performance was evaluated by examining the amino acid substitutions, focusing on their types and frequencies (Figure 17). As expected, positively charged amino acids like arginine (R) and lysine (K) were frequently inserted, increasing the net positive charge and thus, driving the peptides' penetration potential. In contrast, anionic amino acids like aspartic and glutamic acids (D and E) were rarely included in the optimized sequences and were often deleted from the original peptides. This aligns with the critical role of positive charge in CPP functionality[179–182] discussed previously.

In accordance with the feature importance analysis, hydrophobic amino acids such as alanine, cysteine, isoleucine, and leucine (A, C, I, and L, respectively) also played a pivotal role in optimizing the peptide sequences for membrane penetration. Interestingly, these residues were often inserted, but also removed from the original sequences, further remarking that their specific distribution within the peptide may significantly influence the penetration ability of CPPs. Cysteine emerged as another often inserted amino acid, but almost never deleted one, suggesting its potential in promoting penetration capability to CPPs (Figure 17), in line with recent studies showing cysteine rich penetrating peptides[190–192].

One particularly interesting observation is the substitution of valine (V) with isoleucine (I), which raised questions about the relationship between this specific amino acid substitution and the classification of peptides as CPPs or non-CPPs (Figure 17). The feature importance and impact analyses (Figure 16) revealed that a high valine content (PAAC_Xc1.V) correlates strongly with non-CPP predictions, suggesting that decreasing valine levels may shift the classification toward the CPP category. It should also be noted that amino acid substitutions are influenced by the similarity distance evaluation, which is based on the BLOSUM62 substitution matrix. Within this framework, the substitution of valine (V) with isoleucine (I) is identified as the most favorable after the homologous V-V substitution. However, the frequent occurrence of V-I substitutions raises questions about potential limitations of the current algorithm. Further research is needed to explore the biological significance and implications of this substitution pattern.

In conclusion, the design of CPPs using ML and GA represents a promising approach that enables the systematic optimization of peptide sequences for enhanced cellular penetration. This dual focus on generating effective CPPs and explaining the rationale behind their predicted performance represents a significant step forward in the design of interpretable tools for peptide research. By bridging the gap between predictive accuracy and explainability, LightCPPgen not only advances the study of CPPs but also lays the groundwork for innovative and reliable drug delivery solutions.

6. Exploring the TAT-RasGAP$_{317-326}$ Anti-Cancer Molecular Mechanisms

In the previous chapter, we developed an optimization algorithm designed to enhance the penetrability of non-CPPs sequences by transforming them into CPPs. This chapter focuses on exploring the anticancer properties of TAT-RasGAP$_{317-326}$ at the atomistic level through MD simulations. Building on this groundwork, this chapter delves into the mechanistic exploration of the anticancer properties of a multi-active peptide CPP called TAT-RasGAP$_{317-326}$. Through MD simulations, we aim to unravel the atomistic-level mechanisms by which TAT-RasGAP$_{317-326}$ induces cancer cell lysis.

Unlike conventional ACP, TAT-RasGAP$_{317-326}$ has demonstrated the ability to lyse cancer cells through a mechanism that is distinct from conventional programmed cell death pathways, such as apoptosis, necroptosis, parthanatos, pyroptosis, and autophagy[107]. This unique mode of action is attributed to its capacity to target and disrupt specific lipids within the plasma membrane, a process that bypasses the intracellular signaling pathways typically involved in regulated cell death. Because of this, the cytotoxic properties of TAT-RasGAP$_{317-326}$ may be particularly challenging for cancer cells to counteract through the development of resistance mechanisms often seen with therapies targeting traditional cell death pathways[193].

A single point mutation within the RasGAP domain, specifically the substitution of tryptophan at position 317 with alanine (W317A), completely abolishes the peptide's anticancer and antimicrobial activities[107,112,116]. In this connection, the W317A mutation serves as a key factor in unraveling the molecular mechanisms through which TAT-RasGAP$_{317-326}$ exerts its anticancer activity. MD simulations are well-suited for this task, offering a powerful and versatile approach to studying biomolecular systems at atomic resolution[194,195]. Through MD, the dynamic behavior of TAT-RasGAP$_{317-326}$ and its mutant form can be thoroughly investigated within biologically relevant environments, such as lipid bilayers mimicking cellular membranes. This computational technique allows for a detailed examination of how the mutation alters key structural and functional properties of the peptide, including its interaction with specific plasma membrane lipids, conformational stability, and contact probability.

The unconventional killing mechanism of TAT-RasGAP$_{317-326}$ opens new possibilities for its use in anticancer treatments. Moreover, the unique lipid-targeting nature of TAT-RasGAP$_{317-326}$ underscores its potential as part of a broader strategy to address heterogeneity in cancer cell death responses, ultimately paving the way for the development of more robust and versatile therapeutic regimens.

6.1. Materials and Methods

The interactions of TAT-RasGAP$_{317-326}$ (DTRLNTVWMWGGRRRQRRKKRG) and W317A mutant (DTRLNTVWMAGGRRRQRRKKRG) peptides in the retro-inverse configuration with different types of lipid bilayers was investigated through MD simulations (Figure 19). The TAT-RasGAP$_{317-326}$ and L-W317A tridimensional structures were modelled using PEP-FOLD3 server[196].

Subsequently, the retro-inverse configurations of TAT-RasGAP$_{317-326}$ and W317A peptides were generated following a procedure described recently[197].

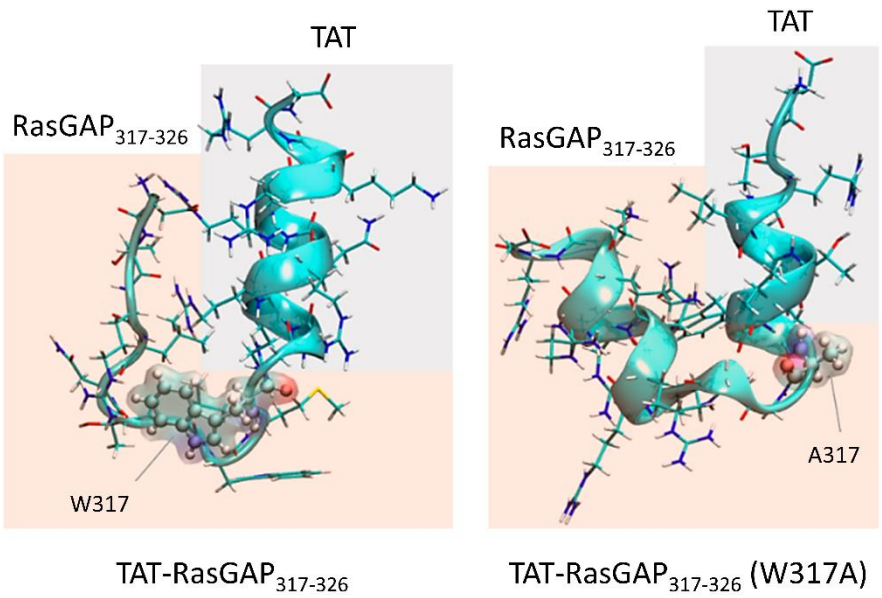


Figure 19: Qualitative visualization of the structures of TAT-RasGAP$_{317-326}$ and its W317A mutant taken at the start of the MD simulations, prior to any interaction between the peptides and the membranes[118].

Both peptides were inserted at a minimum distance of 2 nm from the different symmetric multi-component bilayers, constructed and solvated according to the TIP3P water model[198] using CHARMM-GUI[199–201]. The composition of each layer of the different types of membrane is described in Table 2. Both peptides were simulated in 4 replicas for all the different membrane composition, generating a total number of 48 systems. The systems were composed of 38'000 particles (48'000 for INNER system), after addition of sodium and chloride ions at a concentration of 0.15 M. The CHARMM36[202] force field was used to define phospholipid and protein topology through an all-atom approach.

Each system was minimized using the steepest descent method. We then performed equilibration procedure through one MD simulation of 50 ps under NVT ensemble and four MD simulations of 50 ps, 100 ps, 100 ps, and 200 ps under the NPT ensemble, with gradually removed position restraints. For the equilibration protocol, the v-rescaling[130] temperature coupling algorithm with a time constant of 1 ps was applied to keep the temperature at 310 K. Berendsen semi-isotropic pressure coupling algorithm[129] with a reference pressure of 1 bar and a time constant of 5 ps was employed. Then, all systems were simulated for the production run in the NPT ensemble with 2 fs time steps, using Nose-Hoover thermostat[131] and Parrinello-Rahman barostat[134].

The overall sampled time for all systems was 200 ns, except for the inner leaflet system, which was simulated for 500 ns. Electrostatic interactions were calculated by applying the particle-mesh Ewald (PME) method[127] and van der Waals interactions[126] were defined within a cut-off of 1.2 nm. Trajectories were collected every 2 ps and the Visual Molecular Dynamics (VMD) package[203] was employed to visually inspect the simulated systems. GROMACS 2018

package[204,205] was used for simulations and data analysis. The last 20 ns of the 200 ns production run of each simulation were considered for analysis, except when the inner leaflet system (last 50 ns of the 500 ns).

Table 2: Detailed description of the different membrane phospholipids composition per layer.

Phospholipids	*PC100*	*PS20*	*PA20*	*PI(P2)5*	*CAR10*	*INNER*
Phosphatidylcholine (POPC)	55	44	44	52	49	14
Phosphatidylethanolamine (SLPE)	0	0	0	0	0	20
Sphingomyelin (PSM)	0	11	0	0	0	8
Phosphatidylserine (SOPS)	0	0	0	0	0	9
Phosphatidylinositol (SAPI)	0	0	0	0	0	4
Phosphatidic acid (POPA)	0	0	11	0	0	1
Phosphatidylinositol (4,5) Biphosphate (POPI25)	0	0	0	2	0	1
Cholesterol (CHL)	0	0	0	0	0	23
Cardiolipin (TOCAR)	0	0	0	0	6	0
Total	**55**	**55**	**55**	**55**	**55**	**80**

6.2. Results

6.2.1. TAT-RasGAP$_{317-326}$ interaction behavior with negatively charged membranes

To characterize the interaction dynamics of TAT-RasGAP$_{317-326}$ and its mutant W317A with various phospholipid bilayers, a comprehensive set of analyses was performed. These include the evaluation of Buried Surface (BS), hydrogen bond (H-Bond) formation, and α-helix content (Figure 20). These analyses are instrumental in unraveling the structural and functional differences between the wild-type peptide and its mutated counterpart, shedding light on how the W317A mutation compromises the peptide's ability to engage with cellular membranes effectively.

The BS analysis provides insights into the extent of membrane contact achieved by each peptide. TAT-RasGAP$_{317-326}$ demonstrates a significantly higher BS compared to W317A across all membrane compositions analyzed (Figure 20A). This enhanced BS capability is primarily attributed to the RasGAP moiety, which exhibits more substantial interaction with the membranes compared to the TAT moiety (Figure 20B-C). Notably, both peptides exhibit minimal interaction with PC100 membranes, indicating a weak affinity for this particular lipid composition. These findings suggest that the RasGAP moiety plays a critical role in facilitating strong membrane interactions, which are significantly reduced in the W317A mutant.

The analysis of H-Bond formation further emphasizes the superior interaction capabilities of TAT-RasGAP$_{317-326}$. The wild-type peptide forms more H-Bond with all membrane compositions compared to W317A (Figure 20D). Similar to the BS analysis, this higher capacity for H-Bond formation is primarily driven by the RasGAP moiety rather than the TAT moiety (Figure 20E-F). The INNER system highlights the most pronounced difference, with TAT-RasGAP$_{317-326}$ forming an average of 22.2 ± 2.98 H-Bonds, significantly higher than the 16.3 ± 2.04 formed by W317A. These results underscore the importance of the RasGAP moiety in mediating stable and specific interactions with membrane phospholipids, a property that is markedly diminished in the

W317A mutant. The weak H-Bond formation observed for both peptides in the PC100 systems aligns with their reduced BS in this lipid composition, reinforcing the notion of a composition-specific interaction profile.

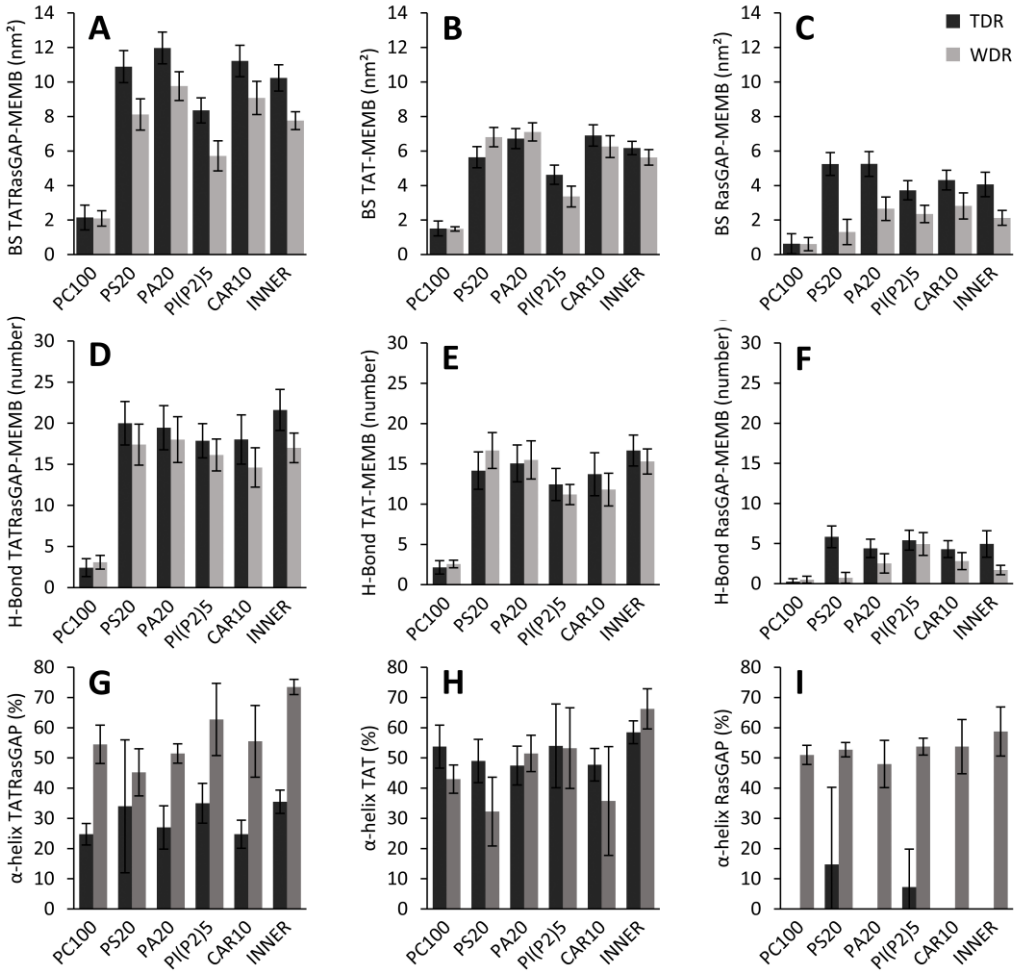


Figure 20: A) Total Buried Surface (BS) analysis of the two peptides with the different type of phospholipid bilayers. B) BS analysis between the TAT moiety of the two peptides with the different type of phospholipid bilayers. C) BS analysis between the $RasGAP_{317-326}$ moiety of the two peptides with the different type of phospholipid bilayers. D) H-Bond analysis of the two peptides with the different types of phospholipid bilayers. E) H-Bond analysis of the two peptides TAT moiety with the different type of phospholipid bilayers. F) H-Bond analysis of the two peptides $RasGAP_{317-326}$ moiety with the different type of phospholipid bilayers. G) α-helix analysis of the two peptides with the different types of phospholipid bilayers. H) α-helix analysis of the two peptides TAT moiety with the different type of phospholipid bilayers. I) α-helix analysis of the two peptides $RasGAP_{317-326}$ moiety with the different type of phospholipid bilayers. TAT-$RasGAP_{317-326}$ is shown in black and W317A is shown in grey.

To complete the characterization, secondary structure analysis was conducted to explore the impact of the W317A mutation on the peptide's conformational preferences. Interestingly, the W317A mutant displays a significantly higher propensity to adopt α-helix structures compared to TAT-$RasGAP_{317-326}$ (Figure 20G). This behavior is primarily driven by the RasGAP moiety of the mutant, which exhibits a strong tendency to arrange itself in α-helical conformations (Figure 20I). Remarkably, the W317A mutant's RasGAP moiety adopts α-helical structures in 50%-60% of its sequence, whereas TAT-$RasGAP_{317-326}$ exhibits an almost complete lack of α-helix formation

in its RasGAP region. This striking difference suggests that the mutation not only alters its intrinsic structural properties but also affects the peptide's interaction with membranes, potentially contributing to its reduced functional activity.
These findings collectively highlight the critical role of the RasGAP moiety in driving the interaction of TAT-RasGAP$_{317-326}$ with membranes. The W317A mutation allows a more α-helix formation of the peptide and disrupts its functional aspects, leading to diminished membrane contact, weaker hydrogen bonding. By providing a detailed atomistic view of these changes, this analysis lays the groundwork for understanding the molecular mechanisms underlying the peptide's anticancer and antimicrobial activities.

6.2.2. TAT-RasGAP$_{317-326}$ interaction behavior with inner-like plasmatic membranes

To understand the anticancer properties of TAT-RasGAP$_{317-326}$, it is crucial to investigate its interaction with biologically relevant models of plasma membranes. Inner-like plasma membranes, which mimic the composition and properties of cytoplasmic leaflets, provide an ideal environment for studying the peptide's binding, orientation, and potential disruptive effects on lipid bilayers. By comparing the wild-type and mutant peptides, it is possible to identify key differences in their interaction behavior with the inner-like membrane that could explain the loss of function.
To probe these differences, several analyses have been performed, focusing on metrics that capture critical aspects of peptide-membrane interactions. Specifically, Contact Probability (CP), minimum distance between peptides 10th aa and the phospholipids PO_4 groups, and angle between the vector linking GLY-ARG Cα's and the z axis analyses have been performed to try to understand how the mutation may change the interaction of the peptide with the membrane phospholipids (Figure 21). Such analyses have been performed only in the INNER systems in the last 50 ns of simulation time.
The CP analysis highlights the great difference between high CP demonstrated by TAT-RasGAP$_{317-326}$ and the low CP demonstrated by W317A mutant (Figure 21A). In greater detail, the W317A mutation triggers also a lower CP capability of the adjacent amino acids in comparison to the TAT-RasGAP$_{317-326}$ wild-type. Indeed, the entire sequence 'TVWMWGG' of W317A's RasGAP moiety have conspicuous lower CP. Such higher CP capability resulted in a higher propensity for the TAT-RasGAP$_{317-326}$ tryptophan to interact with the phospholipid PO_4 groups (Figure 21B). The distance between the latter is around 0.3 nm, which is a distance that allows the formation of H-Bonds. The higher CP exhibited by TAT-RasGAP$_{317-326}$ in combination with the higher capability to infiltrate 10th the tryptophan amino acid deeper into the membrane is allowed by a specific interaction behavior that TAT-RasGAP$_{317-326}$ have with the INNER membrane.
Building on these observations, the orientation of the TAT moiety relative to the z-axis of the membrane offers an additional layer of understanding. Differences in the angular disposition of the peptide highlight how the mutation alters its structural alignment during interaction, potentially affecting its ability to penetrate and destabilize the membrane. Indeed, the TAT-Z axis angle analysis shows a different inclination of the TAT moiety between TAT-RasGAP$_{317-326}$ and W317A when interacting with the membrane (Figure 21C). In detail, the TAT moiety of TAT-RasGAP$_{317-326}$ can stay more in a vertical conformation (parallel to the z axis), while the one of

W317A could stay in a more horizonal conformation (perpendicular to the z axis), when interacting with the INNER membrane. A qualitative representation of this behavior is shown in Figure 21D.

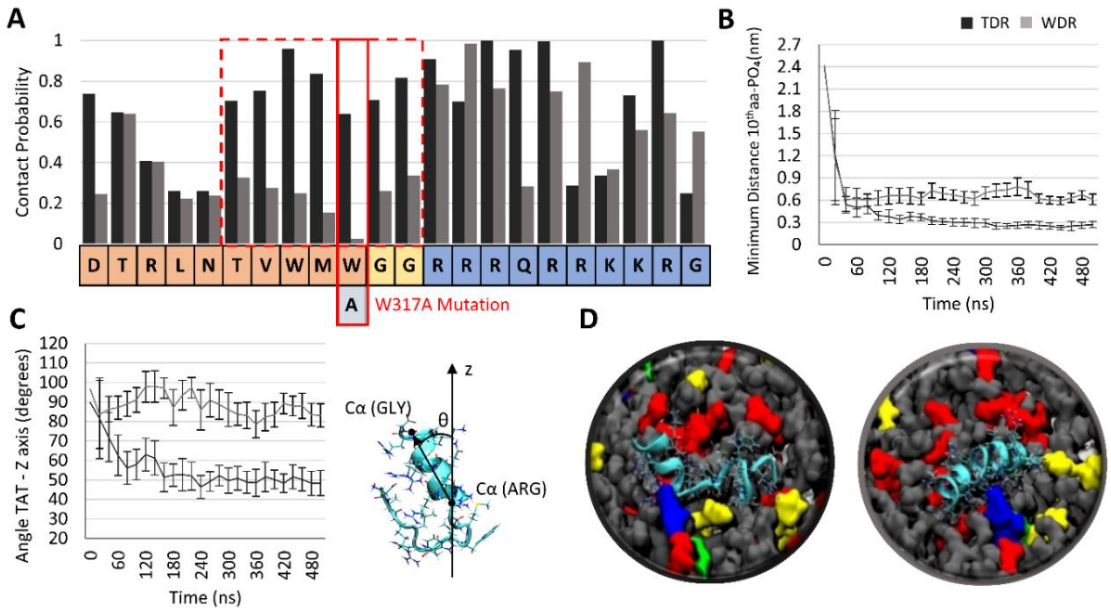


Figure 21. A) Contact probability analysis computed for each amino acid of the TAT-RasGAP$_{317-326}$ (black) and W317A (grey) peptides. The red box highlights the mutated amino acid, while the dashed red box highlights the influence of the single-point mutation on the neighboring amino acid. B) Distance between the PO_4 groups of the membrane and the 10th amino acid of TAT-RasGAP$_{317-326}$ (black) and W317A (grey) peptides. C) Angle analyses between the vector linking GLY-ARG Cα's and the z axis of TAT-RasGAP$_{317-326}$ (black) and W317A (grey) peptides. D) Visual inspection of TAT-RasGAP$_{317-326}$ (left) and W317 (right) peptides interacting with cellular inner leaflet-like membrane.

6.3. Discussion

MD was adopted to elucidate the molecular interactions between TAT-RasGAP$_{317-326}$, a CPP-based construct with anticancer and antimicrobial activities[112,118]. Several cationic peptides have demonstrated promising anticancer properties, including their ability to deliver drugs into cells and their intrinsic antimetastatic, antiangiogenic, and apoptosis-enhancing activities[122]. However, only a select group of these peptides exhibit significant membranolytic properties through interactions with phosphoinositides. Examples include certain plant defensins like NaD1 (Nicotiana alata Defensin 1) and TPP3, which disrupt the plasma membrane by interacting with phosphoinositides such as PI(4,5)P2[206–208].

This mechanism appears similar to the one observed for TAT-RasGAP$_{317-326}$. Indeed, the results shown in Figure 20 demonstrate the high ability of TAT-RasGAP$_{317-326}$ to complex with the highly anionic membranes in terms of BS and H-bonds. Contrariwise, the W317A mutant shows a lower ability to complex with the same membranes, but an higher α-helix structure content, which is in agreement with literature[118]. Such behavior is also in agreement with experimental data obtained through peptide binding experiments and permeabilization experiments (TAT-RasGAP$_{317-326}$ and W317A mutant) with multilamellar vesicles and giant unilamellar vesicle[118].

Moreover, the secondary structure analysis shows a greater difference in α-helix content between the two peptides (Figure 20). Noteworthy, W317A mutant exhibits conspicuously higher α-helix content, rather than TAT-RasGAP$_{317-326}$. Such behavior is also in agreement with experimental data obtained through Circular dichroism experiments[118]. In literature, it is widely accepted that tryptophan plays a crucial role in destabilizing membranes[209] and in CPP membrane anchoring[210]. Indeed, the W317A mutation of the RasGAP moiety is known to impact the peptide molecular mechanism of action[109,116].

We tried to understand how the W317A mutation may change the interaction of the peptide with the membrane phospholipids. In agreement with previous observations[210], it has been determined how the tryptophan residue of TAT-RasGAP$_{317-326}$ plays a pivotal role in stabilizing the entire CPP within the membrane, as demonstrated by the per-residue contact probability analysis (Figure 21A). The rupture of the WXW motif, which is crucial for the anticancer activity of the peptide, leads to a marked loss of contact probability in the spacer motif and in the RasGAP regions (Figure 21A). The different interaction behavior of the peptides with lipids also triggers a different propensity of the peptides to infiltrate the membrane (Figure 21B-C-D).

In line with experimental wet bench evidence, we have also demonstrated how the peptide displays binding preferences for some, but not all, anionic lipids. This cannot be explained by global differences in electrostatic interactions between the peptides and the membranes as both wild-type and mutant peptides carry the same net positive charges (+9). The two peptides do not share the same secondary structures, indeed, the W317A substitution strongly favors the α-helix formation in the RasGAP moiety (Figure 20I). This is likely an important determinant for the membranolytic activity of TAT-RasGAP$_{317–326}$. Indeed, the less structured wild-type peptide “attacks” membranes with a sharper angle allowing deeper insertion into the lipid core if compared to the flatter interaction of W317A mutant (Figure 21C-D).

The prospective clinical utility of TAT-RasGAP$_{317–326}$ appears to be constrained primarily to topical or on-target applications, given its suboptimal biodistribution and bioavailability for systemic use[112,115]. The present characterization of the mode-of-action allowing TAT-RasGAP$_{317–326}$ to destabilize the different membrane compositions may boost the exploration for small molecules mimicking the peptide’s structure while maintaining the peptide’s killing activity.

7. Unveiling the TAT-RasGAP$_{317-326}$ Anti-Bacterial Activity

In this conclusive chapter, the focus shifts to the intriguing antibacterial properties of TAT-RasGAP$_{317-326}$, expanding on the exploration of its multifaceted biological activities. Building on the previous investigation into its unique anticancer capabilities, in this section the attention will be shifted towards its potent antibacterial effects. This peptide is recognized for its broad-spectrum antibacterial activity, capable of targeting and eliminating various bacterial species, including E. coli[112,113,211]. Recent studies have revealed that its bactericidal action is mediated through the inhibition of the BamA sub-unit, a critical system in Gram-negative bacteria responsible for the folding and insertion of outer membrane proteins. By disrupting BamA's function, the peptide destabilizes the outer membrane's integrity, leaving E. coli susceptible to environmental stresses and enhancing its vulnerability to antimicrobial agents.
The exploration centers on a thorough investigation of three experimentally validated mutations that modulate the susceptibility of the E. Coli Bam protein to TAT-RasGAP$_{317-326}$[120]. Furthermore, our exploration extends beyond the boundaries of experimental evidence as we venture into predictive modeling. Through this approach, we have predicted two additional mutations within the Bam protein that are hypothesized to influence its interaction with TAT-RasGAP$_{317-326}$. These predictions are currently undergoing rigorous validation in wet laboratory experiments, bridging the gap between computational insights and empirical evidence. This forward-looking approach not only expands our understanding of TAT-RasGAP$_{317-326}$'s antibacterial capabilities but also contributes to the ongoing development of innovative strategies in the fight against bacterial infections.

7.1. Materials and Methods

In this study, we used the β-barrel subunit of Bam-A (BamA) protein in closed inward position from the PDB ID: 5D0O[212] and TAT-RasGAP$_{317-326}$ peptide in the retro-inverse configuration. The selected BAMA was studied using MD considering the wild-type (WT), A499V (1MUT-P), K798D (1MUT-P-MD), D498N (1MUT-N), D497N-D498K-D500N (3MUT-N), and Q664A-E800K (2MUT-N-MD) mutations configurations surrounded by a lipid bilayer with TAT-RasGAP$_{317-326}$. The docking of TAT-RasGAP$_{317-326}$ with BamA was performed with HADDOCK 2.4[213], using as receptor target the BamA loop 3 (from Q495 to T505) and as ligand the whole TAT-RasGAP$_{317-326}$ peptide allowing a full flexibility of the RasGAP$_{317-326}$ moiety. The first 10 poses of the WT, 1MUT-P, 1MUT-P-MD, 1MUT-N, and 3MUT-N systems were selected as starting points for the simulations.
Furthermore, the membrane was composed of 200 phospholipids in total, including 114 PYPE, 30 POPE, 16 OYPE and 40 POPG, constructed and solvated according to the TIP3P water model[198] using CHARMM-GUI[199–201]. All the systems were composed of around 95’000 particles, after addition of sodium and chloride ions at a concentration of 0.15 M. The CHARMM36[202] force field was used to define phospholipids and proteins topology through an all-atom approach. Each system was minimized using the steepest descent method. We then performed equilibration procedure through one MD simulation of 250 ps under NVT ensemble and four MD simulations of 250 ps, 500 ps, 1 ns, and 5 ns under the NPT ensemble, with gradually removed position

restraints. For the equilibration protocol, the v-rescale[130] temperature coupling algorithm with a time constant of 1.0 ps was applied to keep the temperature at 310 K. c-rescale[135] semi-isotropic pressure coupling algorithm with reference pressure of 1 bar and a time constant of 5.0 ps was employed.

Then, all systems were simulated for the production run in the NPT ensemble with 2 fs time steps, using v-rescale[130] thermostat and c-rescale[135] barostat, with a time constant of 1.0 ps and 5.0 ps, respectively. Each starting docking pose was simulated for 200 ns in 3 replicas. Electrostatic interactions were calculated by applying the particle-mesh Ewald (PME) method[127] and van der Waals interactions[126] were defined within a cut-off of 1.2 nm. Trajectories were collected every 2 ps and GROMACS 2023[204,205] package was used for simulations and data analysis. Visual Molecular Dynamics (VMD) package was employed to visually inspect the simulated systems[203]. The last 20 ns of the 200 ns production run of each simulation were considered for the analyses. The simulation summary is shown below:

System	*Docking Poses*	*Simulation Time*	*Replicas*	*Total sampled time*
β-BAMA-WT+TRG	10	200 ns	3	6 µs
β-BAMA-1MUT-P +TRG	10	200 ns	3	6 µs
β-BAMA-1MUT-P-MD +TRG	10	200 ns	3	6 µs
β-BAMA-1MUT-N +TRG	10	200 ns	3	6 µs
β-BAMA-3MUT-N +TRG	10	200 ns	3	6 µs
β-BAMA-2MUT-N-MD +TRG	10	200 ns	3	6 µs

7.2. Rationale

The data gathered from in vitro and in silico studies underscore the critical role of electrostatic interactions in modulating the inhibitory efficacy of TAT-RasGAP$_{317-326}$ on BamA[120]. Specifically, a reduction in the negative charge within the Q495-T505 range (loop3) of BamA leads to a marked decrease in TAT-RasGAP$_{317-326}$'s ability to inhibit BamA functionality. This phenomenon can be attributed to the destabilization of TAT-RasGAP$_{317-326}$ within loop3, which compromises its capacity to effectively coordinate with two critical functional regions of BamA: Q664-S665 (loop6) and D795-N805 (β-sheet16). These regions are vital for BamA's proper operation, with loop6 playing a role in the formation of the exit pore and β-sheet16 contributing to the lateral opening and sealing mechanism[214,215] (Figure 22).

TAT-RasGAP$_{317-326}$ acts as a molecular bridge by linking loop3 with loop6, impairing the formation of the exit pore, which is critical for BamA's function in folding and inserting outer membrane proteins. Similarly, the peptide also establishes interactions between loop3 and β-sheet16, disrupting the lateral sealing necessary for the β-barrel assembly process. These dual interactions highlight the strategic mechanism by which TAT-RasGAP$_{317-326}$ inhibits BamA, targeting its most essential functional elements.

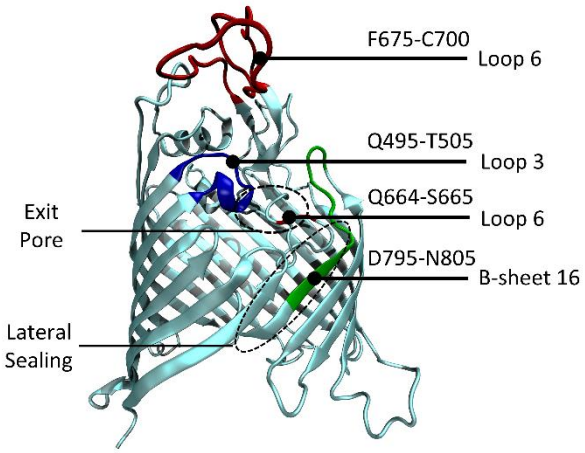


Figure 22: Detail of the BamA's Q495-T505 (loop3), Q664-S665 (loop6), F675-C700 (loop6), and D795-N805 (β-sheet16) ranges. In addition, the exit pore and the lateral sealing regions are shown on the left.

Given the importance of these interactions, a central hypothesis emerges: mutations in BamA that alter the electrostatic profile of loop3, loop6, or β-sheet16 could directly modulate TAT-RasGAP$_{317-326}$'s ability to inhibit BamA. To investigate this, hypothetical mutations were designed to either increase or decrease the negative charge within the D795-N805 range, aiming to respectively enhance or weaken the electrostatic attraction between TAT-RasGAP$_{317-326}$ and BamA in this region. These mutations could also indirectly influence the interactions between TAT-RasGAP$_{317-326}$ and the Q664-S665 range, owing to the proximity of loop6 to the D795-N805 region.

For example, a potential positive mutation, K798D, could amplify the negative charge within the D795-N805 range, thereby increasing the susceptibility of BamA to TAT-RasGAP$_{317-326}$ by strengthening electrostatic attraction. Conversely, a proposed negative mutation, Q664A-E800K, could alter the charge balance within loop6 and β-sheet16, leading to increased resistance of BamA to TAT-RasGAP$_{317-326}$ by reducing electrostatic attraction.

This rationale provides a framework for understanding how TAT-RasGAP$_{317-326}$'s interactions with BamA are governed by electrostatics and highlights the critical importance of loop3, loop6, and β-sheet16 in regulating BamA's function and its inhibition by peptides.

7.3. Results

7.3.1. TAT-RasGAP$_{317-326}$ interaction behavior with BamA

The interaction between TAT-RasGAP$_{317-326}$ and BamA represents a crucial aspect of understanding the peptide's functional mechanism and its potential to disrupt bacterial outer membrane assembly. BamA plays a vital role in inserting and folding outer membrane proteins, making it a compelling target for therapeutic interventions. By investigating how TAT-RasGAP$_{317-326}$ interacts with BamA, we aim to uncover insights into the molecular determinants governing their binding, hydrogen bonding, and overall buried surface area. Such analyses provide a deeper understanding of how the peptide engages with this essential bacterial protein and how specific mutations may alter these interactions. depiction of the optimal interaction poses

between TAT-RasGAP$_{317-326}$ and the specified BamA variants following molecular dynamics simulations is presented in Figure 23.

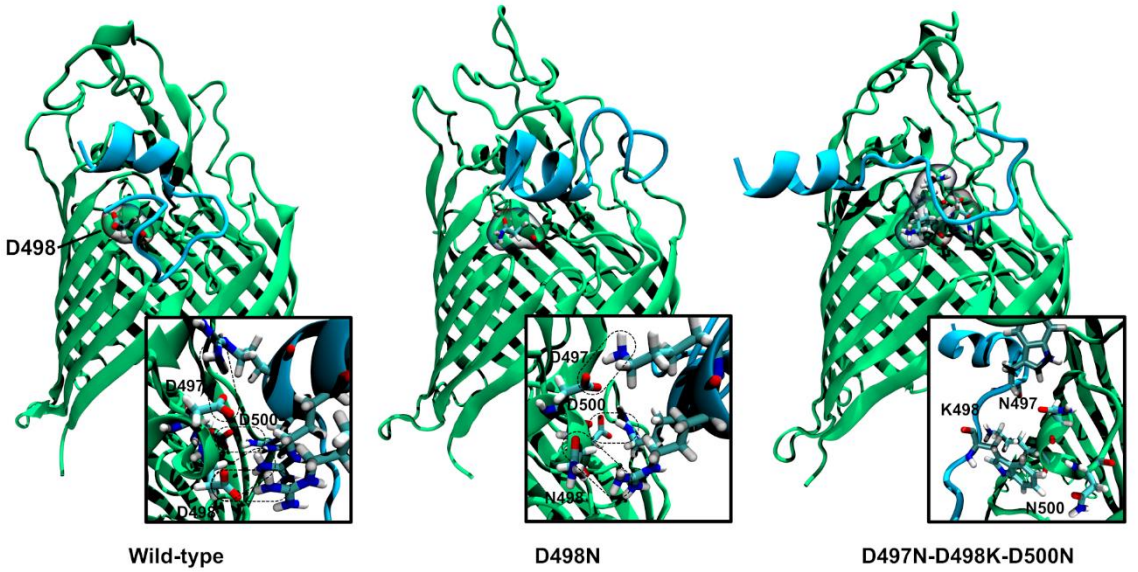


Figure 23: A visual depiction of the optimal interaction poses between TAT-RasGAP317-326 and the specified BamA variants following molecular dynamics simulations is presented[120]. Key residues, including D498 in the wild-type BamA and the corresponding mutations (D498N and D497N-D498K-D500N), are highlighted on the respective BamA structures. To enhance clarity, zoomed-in views of each interaction scenario are provided, offering a closer examination of the interactions between the peptide and BamA. Additionally, potential hydrogen bond interactions involving BamA's D497, D498, and D500 residues, or their respective mutations, with TAT-RasGAP317-326 are indicated using dashed boxes for a more detailed representation.

To characterize the interaction between BamA and TAT-RasGAP$_{317-326}$, we performed the H-Bond and Total Buried Surface (BS) analyses of the simulated WT, 1MUT-P, 1MUT-P-MD, 1MUT-N, 3MUT-N, 2MUT-N-MD systems. Each system was evaluated based on three replica simulations for each docking pose to ensure the robustness of our findings. These analyses not only assess the stability and strength of interactions between TAT-RasGAP$_{317-326}$ and BamA but also help evaluate the impact of mutations on the peptide's ability to bind and engage with BamA effectively.

Figure 24-A shows the H-Bond analysis between the whole structure of BamA and TAT-RasGAP$_{317-326}$. In detail, the diamond marks in Figure 24 represent the average of the 3 replicas for each docking pose of each system. It is worth noting that the 1MUT-P-MD has the highest H-Bonds (9.77±1.53) between BAMA and TAT-RasGAP$_{317-326}$ in comparison to the WT (9.67±1.78), 1MUT-P (9.58±1.40), 1MUT-N (8.20±1.20), 2MUT-N-MD (7.02±1.13), and 3MUT-N (5.54±1.01) systems. Interestingly, the 1MUT-P-MD, WT, and 1MUT-P systems have almost the same H-Bonds average values. In addition, the 3MUT-N system has the lowest H-Bonds between BamA and TAT-RasGAP$_{317-326}$, which is nearly half in comparison to the WT system. Instead, the proposed 2MUT-N-MD system exhibits an intermediate hydrogen bonding capacity compared to the 1MUT-N and 3MUT-N systems.

Figure 24-B shows the BS analysis between the whole structure of BamA and TAT-RasGAP$_{317-326}$. It is worth mentioning that the 1MUT-P system has the highest BS (6.83±1.00 nm^2) between

BamA and TAT-RasGAP$_{317-326}$ in comparison to the WT (6.44±0.89 nm²), 1MUT-P-MD (6.43±1.47 nm²), 1MUT-N (6.36±0.62 nm²), 2MUT-N-MD (5.50±1.08 nm²), and 3MUT-N (5.26±0.38 nm²) systems. In line with the H-bonds analysis, the proposed 2MUT-N-MD system exhibits an intermediate BS compared to the 1MUT-N and 3MUT-N systems.

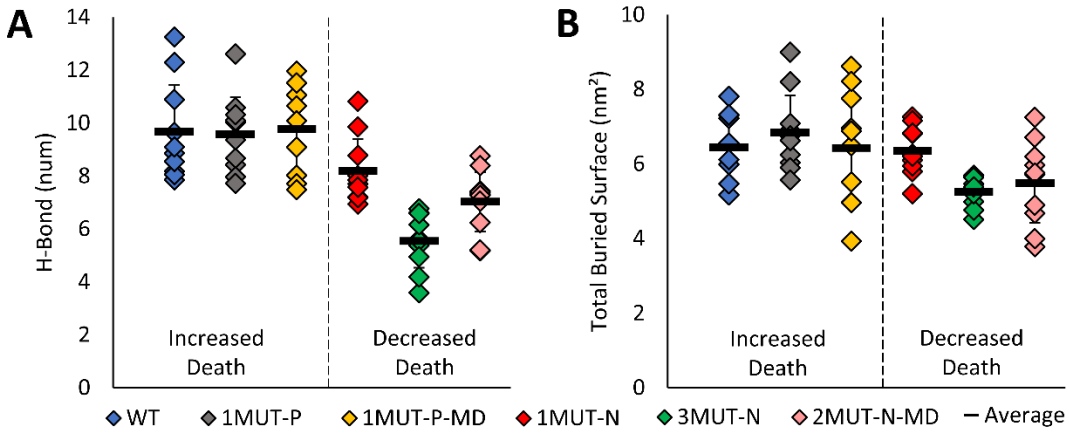


Figure 24: A) H-Bond analysis between the β-barrel subunit of Bam-A (BamA) and TAT-RasGAP$_{317-326}$ (TRG). B) Total buried surface analysis between the BamA and TAT-RasGAP$_{317-326}$ (TRG). In both analyses, the average of the 3 replicas is reported for each docking pose with the colored diamonds blue, grey, orange, red, and green (WT, 1MUT-P, 1MUT-P-MD, 1MUT-N, 3MUT-N, and 2MUT-N-MD). In addition, the average and standard deviation of all the simulated systems (WT, 1MUT-P, 1MUT-P-MD, 1MUT-N, 3MUT-N, and 2MUT-N-MD) are reported in figures A and B.

7.3.2. TAT-RasGAP$_{317-326}$ inhibition of BamA functionality

The ability of TAT-RasGAP$_{317-326}$ to inhibit BamA functionality stems from its interactions with key structural elements of BamA that are critical for its activity. BamA relies on specific functional regions, including loop3, loop6, and β-sheet16, to perform its role in folding and inserting outer membrane proteins[214–216].

To explore how TAT-RasGAP$_{317-326}$ interacts with BamA and how these interactions vary across different BamA mutants, we conducted a Contact Probability (CP) analysis (Figure 25-A). This analysis provides insights into the frequency and extent of interactions between the peptide and critical BamA regions under different conditions. By focusing on the Q495-T505 (loop3), Q664-S665 (loop6), F675-C700 (loop6), and D795-N805 (β-sheet16) ranges (Figure 25-B), we aimed to assess how these interactions are modulated by mutations and how they influence BamA's functionality. These regions were selected based on their established importance in BamA's activity, as highlighted in previous literature[214,215].

The CP analysis was performed using a cut-off distance of 0.3 nm and included the top five docking poses of each system, chosen based on their mean H-bond values. This selection is grounded in the hypothesis that electrostatic attraction, facilitating hydrogen bond formation, is the primary driving force in the interaction between TAT-RasGAP$_{317-326}$ and BamA. The findings from this analysis provide a detailed understanding of how mutations in BamA affect its interaction landscape with TAT-RasGAP$_{317-326}$.

The marks in Figure 25-A represent the ratios between the WT/WT, 1MUT-P/WT, 1MUT-P-MD/WT, 1MUT-N/WT, 3MUT-N/WT, and 2MUT-N-MD/WT averages of the selected amino acids ranges. Values above 1 mean higher CP, while values below 1 mean lower CP in comparison to the WT system. Noteworthy, the 1MUT-N, 2MUT-N-MD, and 3MUT-N systems have lower CP in comparison to WT for the Q495-T505, Q664-S665, and D795-N805 ranges. The major difference in CP is in the Q664-S665, and D795-N805 ranges. However, the 1MUT-N and 3MUT-N systems have higher CP in comparison to WT for the F675-C700 range.

It is reasonable to hypothesize that the decreasing of the negative charges in loop3 due to mutations decreases the electrostatic attraction between TAT-RasGAP$_{317-326}$ and loop3 and thus decreases the interacting capability between BamA's loop3 and TAT-RasGAP$_{317-326}$. Contrariwise, the same mutations increase the electrostatic attraction between TAT-RasGAP$_{317-326}$ and the F675-C700 range of loop6, which has several negative charges exposed. It is important to mention that the proposed 2MUT-N-MD system has slightly lower CP for the F675-C700 range and lower CP for the Q495-T505 range in comparison to WT. Additionally, it exhibits a CP within the pivotal Q664-S665 and D795-N805 intervals that lies intermediate to the values observed in the 1MUT-N and 3MUT-N systems. The unvaried negative charge within the Q495-T505 loop does not impede TAT-RasGAP$_{317-326}$'s capacity to establish complexes within this region, which is in agreement with the earlier hypothesis.

Furthermore, there is no evident translocation of TAT-RasGAP$_{317-326}$ towards the F675-C700 loop, as observed in the 1MUT-N and 3MUT-N systems, owing to the diminished negative charge within the Q495-T505 loop. Differently from the 1MUT-N and 3MUT-N systems, the 2MUT-N-MD proposed mutation seems to act predominantly in the key Q664-S665 and D795-N805 ranges, drastically decreasing the ability of TAT-RasGAP$_{317-326}$ to bind these protein regions.

Moreover, the 1MUT-P system has similar CP in comparison to WT for the Q495-T505 range, while it has higher CP in comparison to WT for the Q664-S665, F675-C700 ranges and D795-N805 ranges. It is important to mention that the proposed 1MUT-P-MD system has similar CP for the F675-C700 range and slightly lower CP for the Q495-T505 range in comparison to WT. However, it has an increased CP for the key Q664-S665 and D795-N805 ranges in comparison to the both WT and 1MUT-P systems.

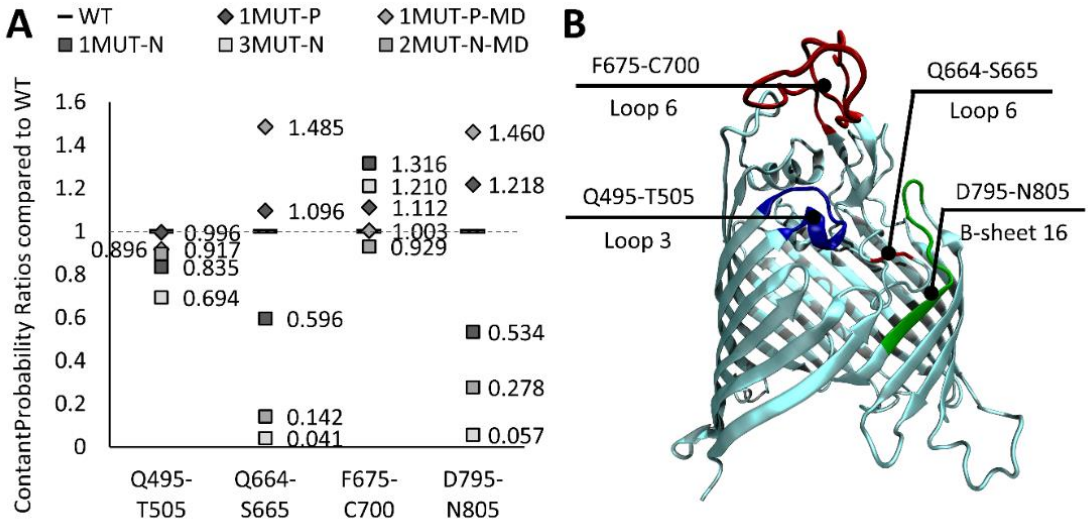


Figure 25: A) Contact Probability (CP) analysis using a cut-off distance of 0.3 nm between the BamA's loop3, loop6 and β-sheet16 and TAT-RasGAP$_{317-326}$ (TRG). In detail, the CP analysis is performed considering the BamA's Q495-T505 (loop3), Q664-S665 (loop6), F675-C700 (loop6), and D795-N805 (β-sheet16) ranges. The marks represent the ratios between the WT/WT, 1MUT-P/WT, 1MUT-P-MD/WT, 1MUT-N/WT, 3MUT-N/WT, and 2MUT-N-MD/WT averages of the selected amino acids ranges. Values above 1 mean higher CP, while values below 1 mean lower CP in comparison to WT system.

To comprehensively elucidate the interplay between BamA and TAT-RasGAP$_{317-326}$, a CP analysis (Table 3) was performed using a cut-off distance of 0.3 nm in the TAT-RasGAP$_{317-326}$ peptide. This analysis takes into considerations for the TAT, RasGAP, and the complete TAT-RasGAP entities. Noteworthy, the TAT moiety exhibits a similar CP value across all the systems except for the 3MUT-N, 2MUT-N-MD, and 1MUT-P-MD configurations, wherein was observed a diminished in the first two cases and an increased CP value in the last case. The principal divergence becomes evident within the RasGAP segment (Table 3), with the 1-MUT-P configuration exhibiting the highest CP value, followed by a diminishing order comprising 1-MUT-P-MD, WT, 1MUT-N, 2MUT-N-MD, and 3MUT-N, respectively.

These findings are in agreement with the previous analyses, thereby underscoring that the TAT moiety of the WT, 1MUT-P, and 1MUT-N systems exhibit comparable interaction propensities with BAMA. Instead, the 1MUT-P-MD, 2MUT-N-MD, and 3MUT-N systems display an increased interaction propensity in the first case and a diminished interaction propensity in the last two cases.

Table 3: Contact Probability (CP) analysis using a cut-off distance of 0.3 nm between the TAT-RasGAP$_{317-326}$ and BamA. The average CP values are presented for the TAT, RasGAP, and TAT-RasGAP moieties.

	WT	*1MUT-P*	*1MUT-P-MD*	*1MUT-N*	*3MUT-N*	*2MUT-N-MD*
TAT	0.4643	0.4639	0.5222	0.4652	0.4269	0.4365
RasGAP	0.3163	0.4269	0.3254	0.2885	0.2544	0.3001
TAT-RasGAP	0.3903	0.4454	0.4238	0.3769	0.3407	0.3683

7.4. Discussion

The findings of this study reveal key differences in the interaction dynamics between BamA and TAT-RasGAP$_{317-326}$ across various systems, underscoring the critical role of electrostatic interactions and hydrogen bonds in this complexation. Notably, the systems 1MUT-P-MD-TAT-RasGAP$_{317-326}$, WT-TAT-RasGAP$_{317-326}$, and 1MUT-P- TAT-RasGAP$_{317-326}$ exhibit the highest number of hydrogen bonds and the largest total BS area, indicating robust interactions between BamA and TAT-RasGAP$_{317-326}$ (Figure 23-Figure 24). In contrast, the 1MUT-N, 3MUT-N, and 2MUT-N-MD systems display reduced hydrogen bonding and lower BS area, suggesting significantly weaker interactions (Figure 23-Figure 24). The observed reduction in interaction strength in the 1MUT-N and 3MUT-N systems is attributed to the decreased negative charge within BamA's loop3 (Q495-T505 range) as a result of the mutations.

CP analysis further elucidates the interaction dynamics between TAT-RasGAP$_{317-326}$ and specific BamA regions: loop3 (Q495-T505), loop6 (Q664-S665 and F675-C700), and β-sheet16 (D795-N805). The results reveal that the negative mutant systems (1MUT-N, 3MUT-N, and 2MUT-N-MD) exhibit reduced CP in the Q495-T505, Q664-S665, and D795-N805 ranges compared to the WT system (Figure 25). This decrease suggests that these mutations alter the electrostatic landscape of BamA, leading to weaker interactions with TAT-RasGAP$_{317-326}$. Interestingly, the 1MUT-N and 3MUT-N systems show an increased CP within the F675-C700 region of loop6 compared to the WT system, likely due to enhanced electrostatic attraction resulting from the decreased negative charge in loop3. Conversely, the 2MUT-N-MD system exhibits only a slight reduction in CP for the F675-C700 range but significantly lower CP in the Q495-T505, Q664-S665, and D795-N805 ranges, reflecting its distinct mutational impact on BamA's electrostatic properties (Figure 25).

On the other hand, the positive mutant systems (1MUT-P and 1MUT-P-MD) demonstrate enhanced interactions with BamA. The 1MUT-P system exhibits comparable CP values to the WT in the Q495-T505 and F675-C700 ranges but shows increased CP in the Q664-S665 and D795-N805 ranges (Figure 25). This enhanced interaction is likely driven by the A499V mutation, which appears to stabilize the complex through improved hydrophobic interactions. The 1MUT-P-MD system, in contrast, exhibits similar BS area and higher hydrogen bond formation compared to the WT, indicating stronger electrostatic interactions (Figure 24). This is likely a result of the K798D mutation, which enhances the negative charge in the β-sheet16 region and contributes to increased CP in the Q664-S665 and D795-N805 ranges compared to both WT and the 1MUT-P system.

The results underscore the importance of electrostatic interactions, particularly hydrogen bonding, in the stability and functionality of the BamA-TAT-RasGAP$_{317-326}$ complex. In the positive systems (1MUT-P and 1MUT-P-MD), the stronger interactions observed are likely responsible for disrupting BamA's critical functions. By simultaneously interacting with key regions, such as loop3 and loop6, TAT-RasGAP$_{317-326}$ impairs the formation of the exit pore, which is essential for BamA's ability to fold and insert outer membrane proteins[214,215]. Additionally, interactions with loop3 and β-sheet16 prevent the lateral opening and sealing mechanism, further compromising BamA's functionality[214,215]. These synergistic interactions effectively block the essential processes required for BamA's proper operation.

Recent studies have demonstrated that certain designed molecules can inhibit the BamA complex by binding to specific regions and preventing the lateral gate from opening[121,217]. For instance, Similarly, Darobactin B stabilizes a closed conformation of the BamA lateral gate, effectively blocking its activity[217]. Peptide Targeting BamA-1 (PTB1) binds to an extracellular divalent cation-dependent site on BamA, locking it into a closed lateral gate conformation and thereby inhibiting its function[121]. Interestingly, TAT-RasGAP$_{317-326}$ also appears to disrupt BamA's functionality by stabilizing it in a closed conformation.

These findings align with recent literature, which has demonstrated that blocking the exit pore and lateral sealing mechanisms is sufficient to inhibit BamA's functionality[121,214,215,217]. The results provide a mechanistic explanation for the differential impact of TAT-RasGAP$_{317-326}$ on wild-type and mutant BamA systems and shed light on the role of specific BamA regions in maintaining its proper function. This study offers valuable insights into the molecular basis of BamA inhibition and underscores the potential of TAT-RasGAP$_{317-326}$ and its derivatives as tools for targeting BamA in therapeutic applications.

8. Conclusion and Future Perspective

In conclusion, this thesis represents a comprehensive integration of ML and GA to drive advancements in CPP design, while also delving into the molecular dynamics of TAT-RasGAP$_{317-326}$, a chimeric CPP with significant anticancer and antibacterial potential. Central to this work is the recognition of the intrinsic cell-penetrating capabilities of CPPs, which emerge as a cornerstone for developing innovative therapies targeting both cancer cells and bacterial infections. The implementation of a genetic algorithm significantly enhanced our ability to design and optimize CPPs, enabling the transformation of non-CPP sequences into functional CPPs. This powerful approach not only underscores the utility of de novo peptide design but also opens up new avenues for creating tailored peptides with multifunctional bioactivities.

Building upon the foundation of CPP design, the thesis focused in parallel on TAT-RasGAP$_{317-326}$. Through MD simulations, this work unraveled the intricate interactions of TAT-RasGAP$_{317-326}$ within plasma membrane-like environments, shedding light on its antitumoral mechanisms of action. Moreover, an in-depth investigation of the BamA-TAT-RasGAP$_{317-326}$ complex, supported by molecular dynamics and mutational analyses, provided critical insights into how this peptide induces the loss of functionality in E. coli BamA. These findings highlight the potential of TAT-RasGAP$_{317-326}$ as a dual-action therapeutic agent capable of targeting both cancer and bacterial infections.

The multifaceted approaches employed in this thesis bridge computational and experimental paradigms, significantly enriching our understanding of the multi-bioactive functionalities of peptides. This interdisciplinary perspective underscores the importance of leveraging computational tools to complement experimental research, setting the stage for future therapeutic breakthroughs. The exploration of the interactions between TAT-RasGAP$_{317-326}$ and both plasma membrane-like systems and BamA provides a strong foundation for designing targeted interventions, offering innovative strategies to combat antibiotic resistance and advance anticancer therapies.

Looking ahead, several promising avenues for future research emerge. First, the CPP design algorithm can be further refined by incorporating the ability to modulate sequence length and total charge. Such enhancements would provide experimental researchers with a more versatile tool for fine-tuning peptide characteristics to meet specific therapeutic requirements. Additionally, advancing the study of TAT-RasGAP$_{317-326}$'s anticancer effects through enhanced-sampling techniques, such as metadynamics, umbrella sampling, or replica exchange molecular dynamics will enable a more detailed exploration of its interactions with membranes of varying phospholipid compositions. These techniques will yield a richer understanding of the peptide's binding free energy landscape and dynamic interaction mechanisms.

Expanding the scope of this work, the optimization algorithm developed for CPP design in chapter 5 was directly applied to enhance the penetrability of TAT-RasGAP$_{317-326}$. The original peptide was assessed by the ML surrogate model, yielding a penetrability score of 0.71. Through iterative optimization, the algorithm proposed two improved variants: one with a single mutation (V320I) and another with two mutations (V320I-L323I), achieving higher predicted penetrability scores of 0.84. Structural predictions of these variants using PEP-FOLD3 (as detailed in Chapter 6 and visualized in Figure 26) highlighted distinct conformational differences.

The single-mutation variant retained an alpha-helical arrangement in the RasGAP segment, reminiscent of the W317A mutant characterized earlier in this thesis, which exhibited suboptimal membrane interaction dynamics. In contrast, the double-mutation variant displayed a random coil conformation in the RasGAP region, a structural feature previously correlated with enhanced membrane interaction and bioactivity in our MD simulations. Drawing on these insights, the two-mutation solution was prioritized for further validation. This decision aligns with the mechanistic principles elucidated in Chapter 6, where RasGAP disordered conformations were shown to facilitate dynamic membrane interactions. The optimized variant thus represents a promising candidate for subsequent in-silico refinement and experimental testing, offering a strategic balance between penetrability and preserved bioactivity.

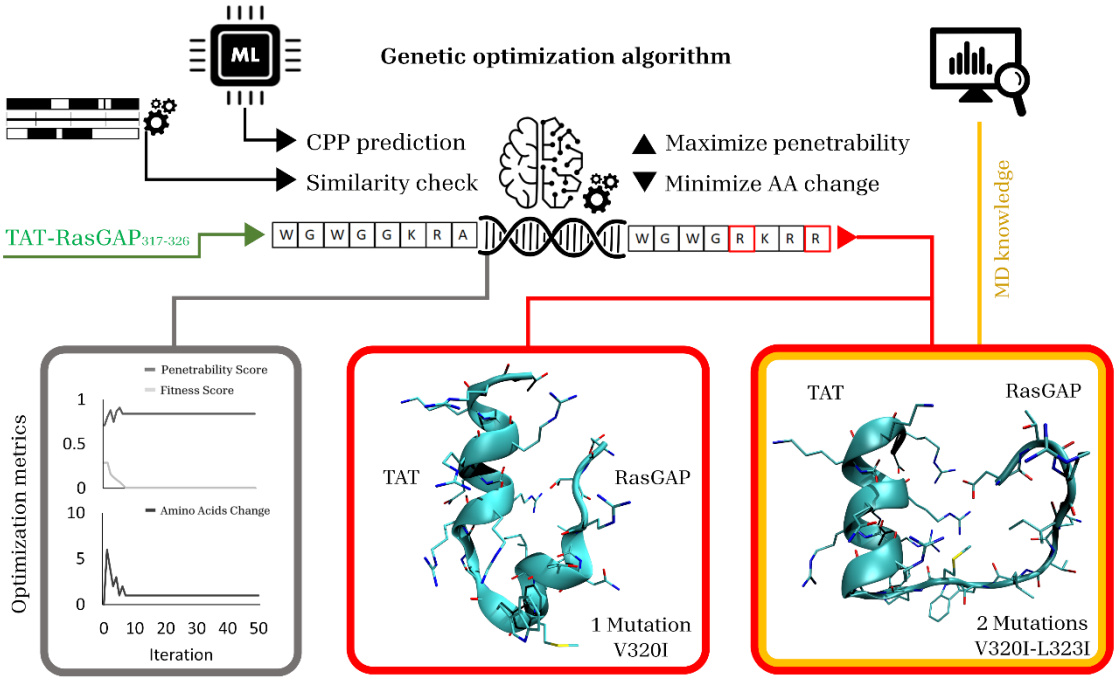


Figure 26: Qualitative illustration of TAT-RasGAP$_{317-326}$ optimization process using LightCPPgen (developed in Chapter 5). Two solutions were proposed (red squares): one with a single mutation (V320I) and another with two mutations (V320I-L323I), achieving higher predicted penetrability scores of 0.84. Structural predictions of these variants were done using PEP-FOLD3.

Experimental validation remains a critical step in translating computational findings into real-world applications. A key focus should be placed on validating the predicted BamA mutations that either enhance or reduce susceptibility to TAT-RasGAP$_{317-326}$. Such experimental efforts will not only confirm the computational predictions but also provide invaluable insights into therapeutic strategies targeting bacterial infections. By bridging computation with experimental fields, this research paves the way for transformative developments in the areas of peptide design, antibacterial resistance, and cancer therapy.

In summary, this thesis lays a robust groundwork for future advancements in CPP design and the exploration of multi-functional peptides like TAT-RasGAP$_{317-326}$. The integration of a data-driven approach with mechanistic explainability has proven to be an optimal synergistic framework, enabling a deeper exploration of the complex interplay governing multi-BPs such as TAT-

$RasGAP_{317-326}$. By combining predictive computational models with atomistic insights into molecular interactions, this dual strategy not only enhances design precision but also uncovers the structural and dynamic principles underlying therapeutic efficacy. The proposed future directions aim to refine and expand the current findings, driving innovation in peptide-based therapies. These efforts promise to unlock new strategies for addressing some of the most pressing medical challenges, including antibiotic resistance and effective cancer treatment, positioning this research at the forefront of biomedicine and therapeutic development.

9. Supporting Information

Section 1

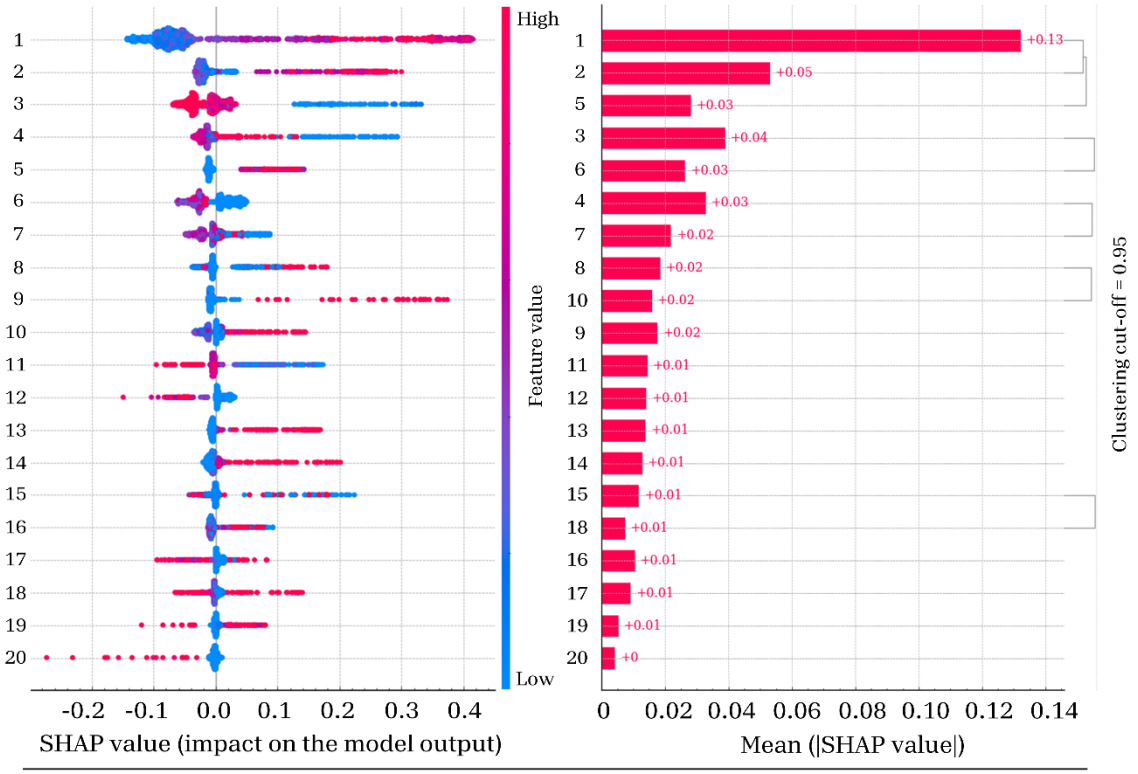


Figure S1: Feature importance and impact computed with SHAP on training set of MLCPP 2.0. The right bar plot ranks the variables by their average impact on model prediction. The left dot plot shows each data point with the signed contribution of each feature. Blue dots indicate low variable values, while red dots indicate high values.

Section 2

The variation of fitness score, penetration score, and changed amino acids across generations of all 40 peptides from Table S1 are shown in Figure S2 and Figure S3. The main objective was to determine if 100 generations were sufficient for the algorithm to converge. Remarkably, convergence was consistently achieved before 50 generations across all peptide length categories (Figure S2 and Figure S3). Even considering the longest peptide, with a length of between 30 and 40 amino acids, it reached convergence slightly before 50 generations. Therefore, it is evident that 50 generations provide sufficient time for the optimization algorithm to converge and reach a definitive CPP optimized sequence.

Table S1: 40 non-CPP randomly selected from MLCPP2.0 test set to tune the optimal generation number of the optimization algorithm.

Peptides	*Length*
AVGKIMKF	8
DAWRMHMQEFVAQLETR	17
DGLVCLLKKPFNRPQGVQPKTGP	23
DLVEAGVVDPTKVTRTALQNAASIAGLILTTDATVA	36
DQQHHGLISKTIQNKLQ	17
EAARALKAALAEEEPA	16
EIKGFTGVDD	10
EIVAEKKKEEVDAEMNPVTKQFQFGQSTVTLE	32
ESFSQMT	7
FGNPLNTAALFIILLHLIIESKSLTWIHMLVIT	33
FYIIGAITLNNLLLR	15
GDTLAAIDLFGIKDK	15
GGHLTHGSPVNVSGKWFKVVHYGVE	25
GPPMMNAAVIKMLKDLG	17
GYIAINISSPNTPG	14
HMLEEAEKRDHRKLGKQLDLFHIQE	25
IADYDADLWQAM	12
IFSNTALVNCMRQTLQDTGHNP	22
IIPRLREPKDLYGSSSQDGC	20
ILCAEFPDEPKWMGGAELSDDGRYVLL	27
ITLQPVRRHGVDAGIFFSDIVVPLKLA	27
IVKGGDQSKMEEGEVY	16
IVNGGAHADTG	11
KFYASVRIDI	10
LEENSVDAVKVRSVVSCD	18
LFNLMTDGDG	10
LIQADQNIENIVKESVTKT	19
MARIAIIGGGSIGEAL	16
NKVESLQSR	9
NLKRLGMKATVKQGDGR	17
NTVGLERSGF	10
QLSKKEASRHAIMRSPQMVSAIVRTM	26
QRINREKHLVLTAAHPSPLAANRGGFFGCKVFSK	34
REGRVTKRFAVLNARFH	17
TPATRQEWVKAA	12
VDETLRVLKAFQT	13
VDTVLFMVPADEARGKGDDMIIE	23
VGKDVTVANATI	12
VIKNFVLFWVTLLPYVKE	18
VLEMVPAALSAALTQALTHCP	21

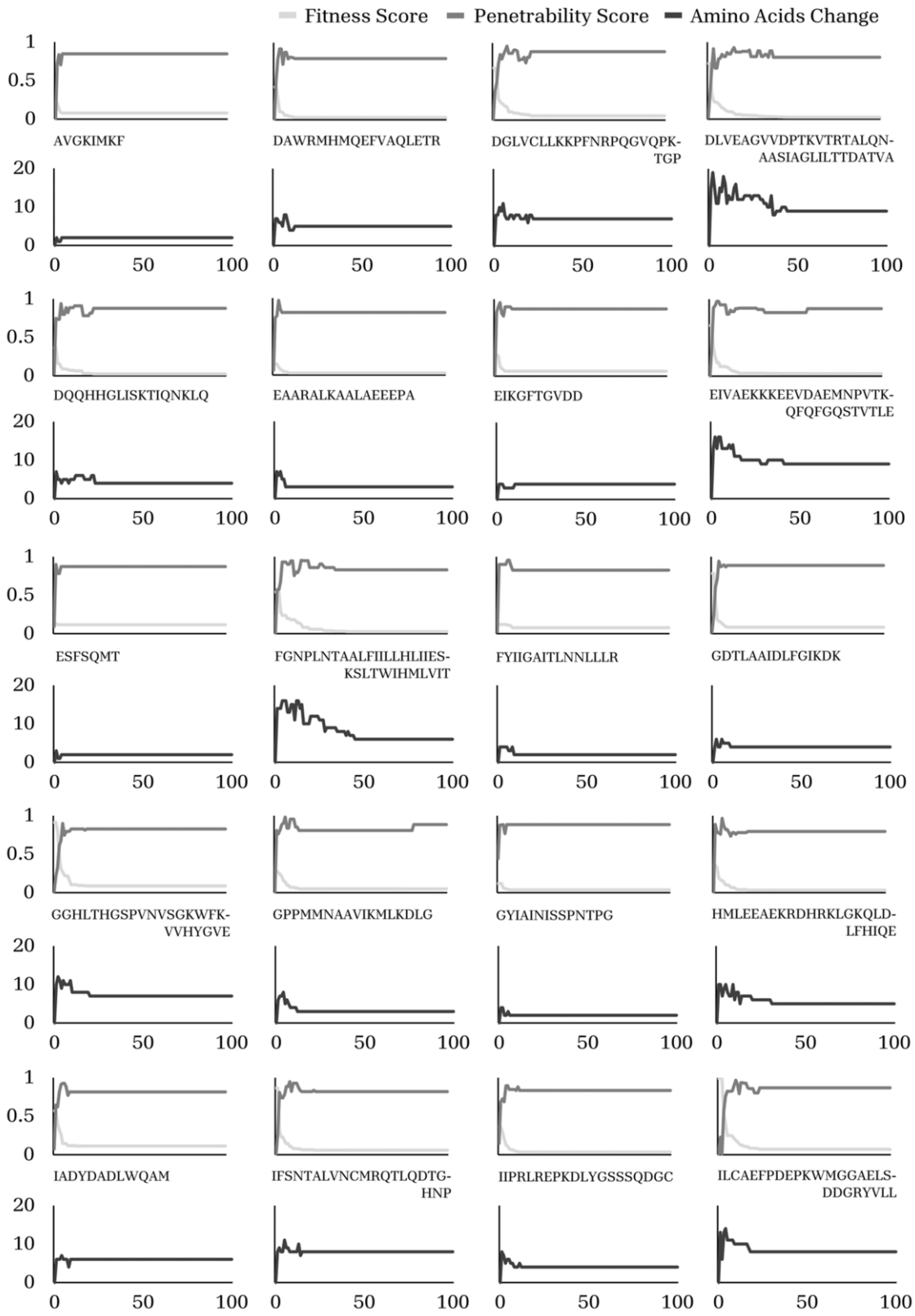


Figure S2: Fitness score, penetrability score, and amino acids change values as generations increase of the first 20 non-CPP from table S1. The data shown are performed using a population size of 500, generations count of 100 and BLOSUM62 substitution matrix.

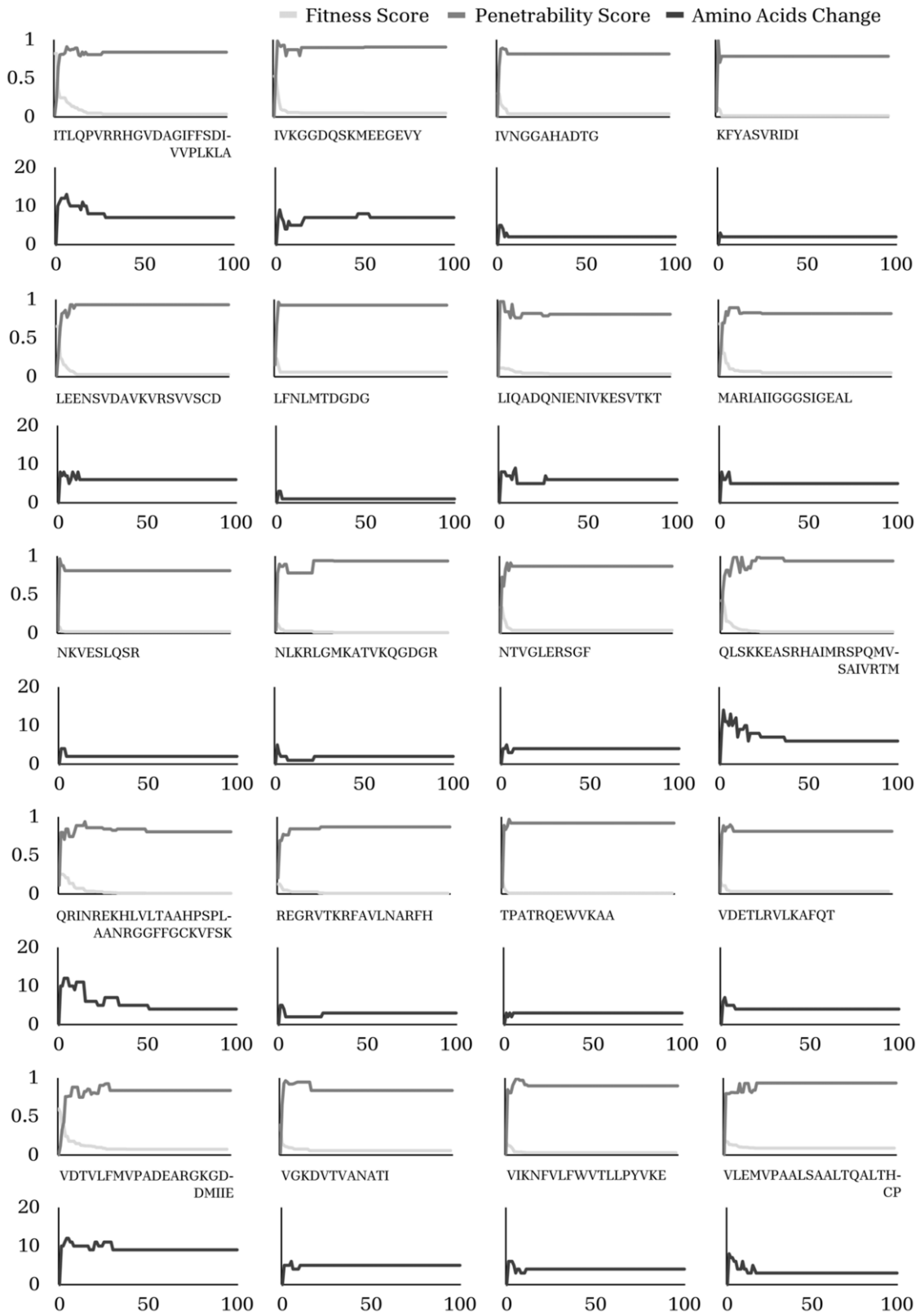
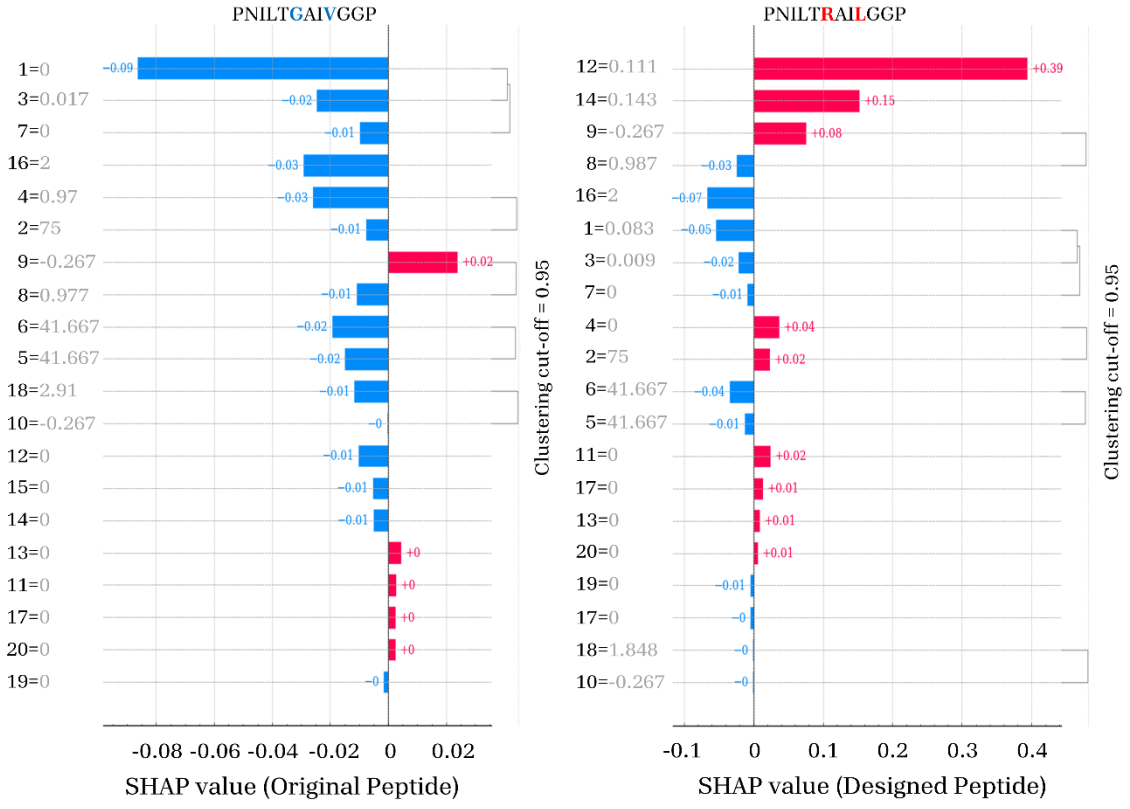


Figure S3: Fitness score, penetrability score, and amino acids change values as generations increase of the first 20 non-CPP from table S1. The data shown are performed using a population size of 500, generations count of 100 and BLOSUM62 substitution matrix.

Section 3



1. CTDC_charge.G1
2. CTDD_normwaalsvolume.2.residue100
3. APAAC_Pc2.Hydrophobicity.1
4. PAAC_Xc1.V
5. CTDD_hydrophobicity_ARGP820101.1.residue50
6. CTDD_hydrophobicity_FASG890101.2.residue50
7. KSCTriad_g5.g5.g5.gap1
8. APAAC_Pc1.A
9. DDE_LA
10. DDE_LP
11. ASDC_DM
12. CKSAAP_IR.gap2
13. QSOrder_Grantham.Xr.T
14. CKSAAP_LL.gap4
15. CKSAAGP_positivecharger.uncharger.gap1
16. Zscale_p1.z5
17. CKSAAGP_uncharger.aromatic.gap9
18. PAAC_Xc1.G
19. ASDC_AS
20. CKSAAP_TR.gap4

Figure S4: SHAP values for a randomly selected peptide (PNILTGAIVGGP) are displayed before optimization (left) and after optimization (right). The original deleted amino acid is highlighted in bold blue (top-left), while the newly inserted amino acids are marked in bold red (top-right). The values for each feature, both pre- and post-optimization, are shown in gray.

The amino acid substitutions in this peptide following the optimization process led to significant shifts in several features, transitioning them from non-CPP to CPP predictions, including CKSAAP_IR.gap2, CKSAAP_LL.gap4, and PAAC_Xc1.V, among others. Remarkably, the algorithm achieved an increase in penetrability with only two mutations. Specifically, replacing glycine (G) with arginine (R) and valine (V) with leucine (L) enhanced CKSAAP_IR.gap2 and CKSAAP_LL.gap4 while simultaneously decreasing PAAC_Xc1.V. Notably, this optimization did not result in a substantial increase in the peptide's net charge, demonstrating that the algorithm can steer sequence evolution toward CPP prediction through diverse mechanisms and patterns, beyond simply amplifying the overall charge.

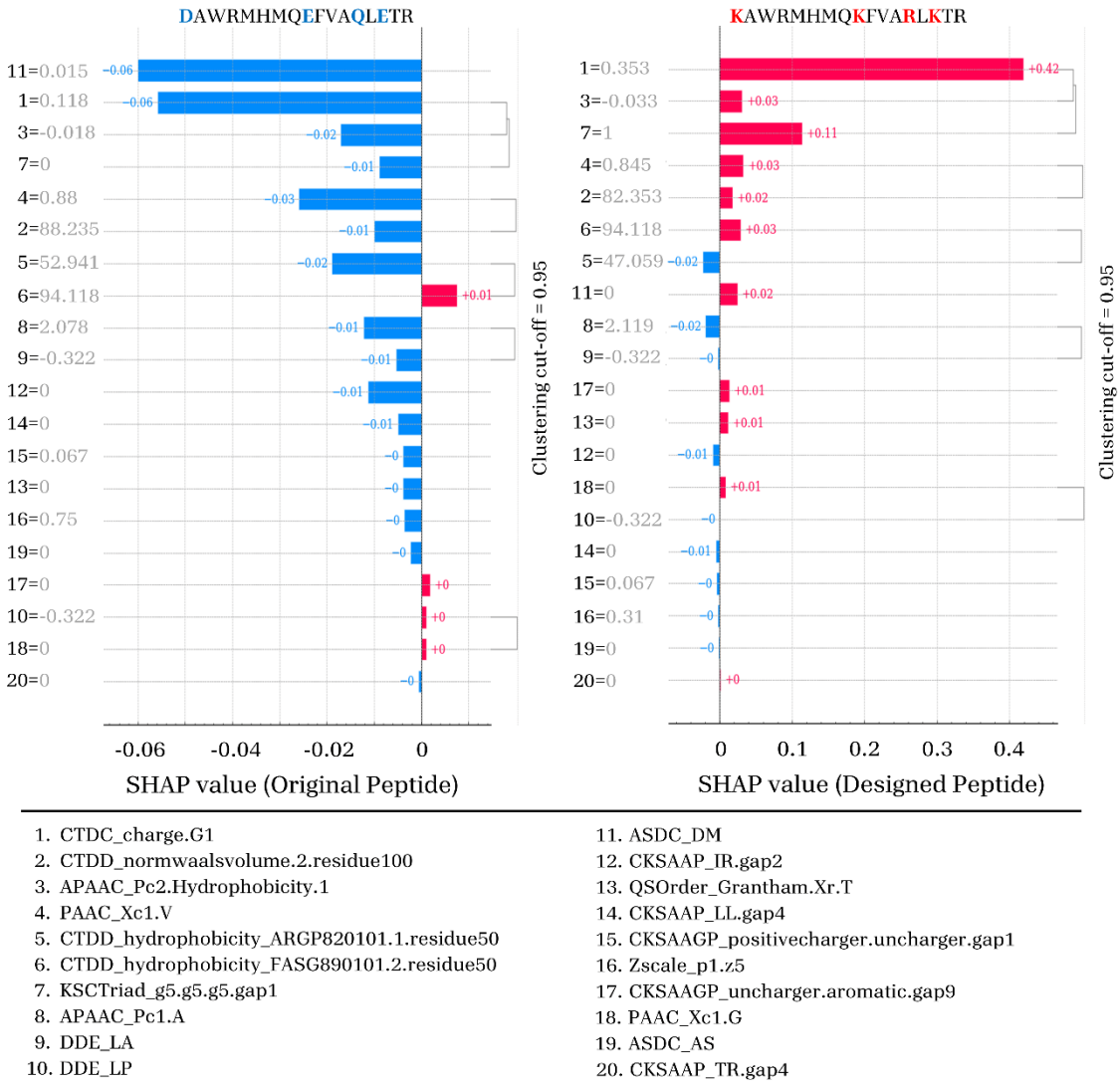


Figure S5: SHAP values for a randomly selected peptide (DAWRMHMQEFVAQLETR) are displayed before optimization (left) and after optimization (right). The original deleted amino acid is highlighted in bold blue (top-left), while the newly inserted amino acids are marked in bold red (top-right). The values for each feature, both pre- and post-optimization, are shown in gray.

Following the optimization process, the amino acid substitutions in this peptide resulted in the transition of several features from non-CPP to CPP predictions, including CTDC_charge.G1, KSCTriad_g5.g5.g5.gap1, PAAC_Xc1.V, and ASDC_DM, among others. In this instance, the algorithm replaced three negatively charged residues (D and E) and one glutamine (Q) with positively charged amino acids (R and K). These substitutions led to an increase in the CTDC_charge.G1 and KSCTriad_g5.g5.g5.gap1 features, while simultaneously reducing the values of PAAC_Xc1.V and ASDC_DM. The optimized sequence experienced a significant boost in total charge, with KSCTriad_g5.g5.g5.gap1 capturing the specific charge distribution pattern with 1 gap between 3 positively charged amino acids (R or K). Additionally, the algorithm not only focused on substituting negatively charged residues with positively charged ones but also optimized the sequence to exhibit a reduced ASDC_DM value, further enhancing its classification as a CPP.

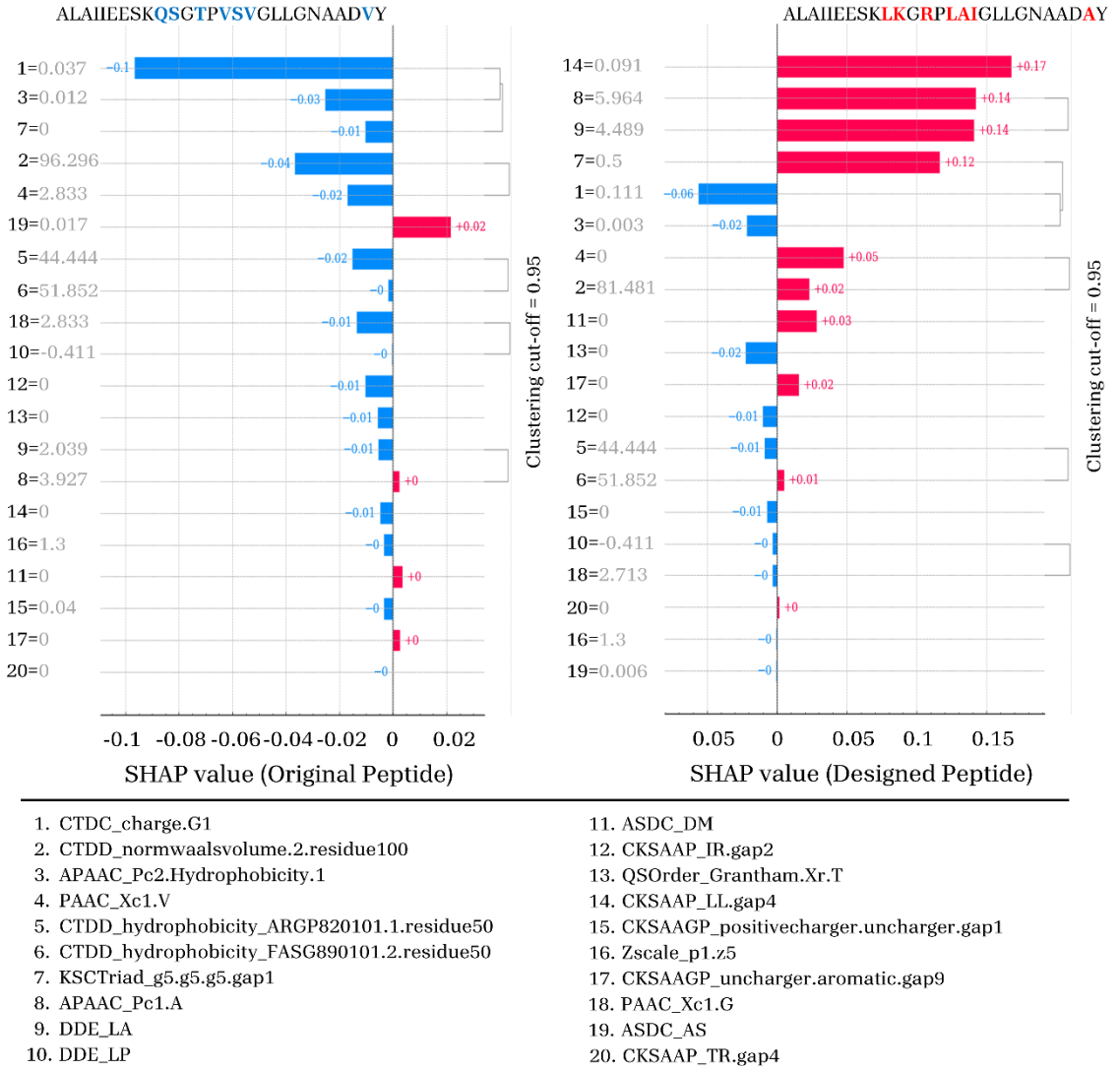


Figure S6: SHAP values for a randomly selected peptide (ALAIIEESKQSGTPVSVGLLGNAADVY) are displayed before optimization (left) and after optimization (right). The original deleted amino acid is highlighted in bold blue (top-left), while the newly inserted amino acids are marked in bold red (top-right). The values for each feature, both pre- and post-optimization, are shown in gray.

The amino acid substitutions in this peptide, following the optimization process, shifted several features from non-CPP to CPP predictions, including CKSAAP_LL.gap4, APAAC_Pc1.A, DDE_LA, KSCTriad_g5.g5.g5.gap1, PAAC_Xc1.V, and CTDD_normwaalsvolume.2.residue100, among others. In this case, the algorithm replaced three valines (V) and four polar amino acids (Q, S, and T) with two positively charged residues (R and K) and five hydrophobic residues (A, L, and I). These substitutions increased the values of CKSAAP_LL.gap4, APAAC_Pc1.A, DDE_LA, and KSCTriad_g5.g5.g5.gap1 features, while reducing the values of PAAC_Xc1.V and CTDD_normwaalsvolume.2.residue100. Notably, the optimized sequence did not experience a significant increase in overall charge. This highlights the algorithm's ability to guide sequence evolution toward CPP classification by strategically altering the hydrophobic pattern of the sequence rather than relying solely on boosting the net charge.

10. Bibliography

(1) Bray, F.; Laversanne, M.; Weiderpass, E.; Soerjomataram, I. The Ever-increasing Importance of Cancer as a Leading Cause of Premature Death Worldwide. *Cancer* **2021**, *127* (16), 3029–3030. https://doi.org/10.1002/cncr.33587.

(2) Bray, F.; Laversanne, M.; Sung, H.; Ferlay, J.; Siegel, R. L.; Soerjomataram, I.; Jemal, A. Global Cancer Statistics 2022: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA A Cancer J Clinicians* **2024**, *74* (3), 229–263. https://doi.org/10.3322/caac.21834.

(3) Mestrovic, T.; Robles Aguilar, G.; Swetschinski, L. R.; Ikuta, K. S.; Gray, A. P.; Davis Weaver, N.; Han, C.; Wool, E. E.; Gershberg Hayoon, A.; Hay, S. I.; Dolecek, C.; Sartorius, B.; Murray, C. J. L.; Addo, I. Y.; Ahinkorah, B. O.; Ahmed, A.; Aldeyab, M. A.; Allel, K.; Ancuceanu, R.; Anyasodor, A. E.; Ausloos, M.; Barra, F.; Bhagavathula, A. S.; Bhandari, D.; Bhaskar, S.; Cruz-Martins, N.; Dastiridou, A.; Dokova, K.; Dubljanin, E.; Durojaiye, O. C.; Fagbamigbe, A. F.; Ferrero, S.; Gaal, P. A.; Gupta, V. B.; Gupta, V. K.; Gupta, V. K.; Herteliu, C.; Hussain, S.; Ilic, I. M.; Ilic, M. D.; Jamshidi, E.; Joo, T.; Karch, A.; Kisa, A.; Kisa, S.; Kostyanev, T.; Kyu, H. H.; Lám, J.; Lopes, G.; Mathioudakis, A. G.; Mentis, A.-F. A.; Michalek, I. M.; Moni, M. A.; Moore, C. E.; Mulita, F.; Negoi, I.; Negoi, R. I.; Palicz, T.; Pana, A.; Perdigão, J.; Petcu, I.-R.; Rabiee, N.; Rawaf, D. L.; Rawaf, S.; Shakhmardanov, M. Z.; Sheikh, A.; Silva, L. M. L. R.; Skryabin, V. Y.; Skryabina, A. A.; Socea, B.; Stergachis, A.; Stoeva, T. Z.; Sumi, C. D.; Thiyagarajan, A.; Tovani-Palone, M. R.; Yesiltepe, M.; Zaman, S. B.; Naghavi, M. The Burden of Bacterial Antimicrobial Resistance in the WHO European Region in 2019: A Cross-Country Systematic Analysis. *The Lancet Public Health* **2022**, *7* (11), e897–e913. https://doi.org/10.1016/S2468-2667(22)00225-0.

(4) Murray, C. J. L.; Ikuta, K. S.; Sharara, F.; Swetschinski, L.; Robles Aguilar, G.; Gray, A.; Han, C.; Bisignano, C.; Rao, P.; Wool, E.; Johnson, S. C.; Browne, A. J.; Chipeta, M. G.; Fell, F.; Hackett, S.; Haines-Woodhouse, G.; Kashef Hamadani, B. H.; Kumaran, E. A. P.; McManigal, B.; Achalapong, S.; Agarwal, R.; Akech, S.; Albertson, S.; Amuasi, J.; Andrews, J.; Aravkin, A.; Ashley, E.; Babin, F.-X.; Bailey, F.; Baker, S.; Basnyat, B.; Bekker, A.; Bender, R.; Berkley, J. A.; Bethou, A.; Bielicki, J.; Boonkasidecha, S.; Bukosia, J.; Carvalheiro, C.; Castañeda-Orjuela, C.; Chansamouth, V.; Chaurasia, S.; Chiurchiù, S.; Chowdhury, F.; Clotaire Donatien, R.; Cook, A. J.; Cooper, B.; Cressey, T. R.; Criollo-Mora, E.; Cunningham, M.; Darboe, S.; Day, N. P. J.; De Luca, M.; Dokova, K.; Dramowski, A.; Dunachie, S. J.; Duong Bich, T.; Eckmanns, T.; Eibach, D.; Emami, A.; Feasey, N.; Fisher-Pearson, N.; Forrest, K.; Garcia, C.; Garrett, D.; Gastmeier, P.; Giref, A. Z.; Greer, R. C.; Gupta, V.; Haller, S.; Haselbeck, A.; Hay, S. I.; Holm, M.; Hopkins, S.; Hsia, Y.; Iregbu, K. C.; Jacobs, J.; Jarovsky, D.; Javanmardi, F.; Jenney, A. W. J.; Khorana, M.; Khusuwan, S.; Kissoon, N.; Kobeissi, E.; Kostyanev, T.; Krapp, F.; Krumkamp, R.; Kumar, A.; Kyu, H. H.; Lim, C.; Lim, K.; Limmathurotsakul, D.; Loftus, M. J.; Lunn, M.; Ma, J.; Manoharan, A.; Marks, F.; May, J.; Mayxay, M.; Mturi, N.; Munera-Huertas, T.; Musicha, P.; Musila, L. A.; Mussi-Pinhata, M. M.; Naidu, R. N.; Nakamura, T.; Nanavati, R.; Nangia, S.; Newton, P.; Ngoun, C.; Novotney, A.; Nwakanma, D.; Obiero, C. W.; Ochoa, T. J.; Olivas-Martinez, A.; Olliaro, P.; Ooko, E.; Ortiz-Brizuela, E.; Ounchanum, P.; Pak, G. D.; Paredes, J. L.; Peleg, A. Y.; Perrone, C.; Phe, T.; Phommasone, K.; Plakkal, N.; Ponce-de-Leon, A.; Raad, M.; Ramdin, T.; Rattanavong, S.; Riddell, A.; Roberts, T.; Robotham, J. V.; Roca, A.; Rosenthal, V. D.; Rudd, K. E.; Russell, N.; Sader, H. S.; Saengchan, W.; Schnall, J.; Scott, J. A. G.; Seekaew, S.; Sharland, M.; Shivamallappa, M.; Sifuentes-Osornio, J.; Simpson, A. J.; Steenkeste, N.; Stewardson, A. J.; Stoeva, T.; Tasak, N.; Thaiprakong, A.; Thwaites, G.; Tigoi, C.; Turner, C.; Turner, P.; Van Doorn, H. R.; Velaphi, S.; Vongpradith, A.; Vongsouvath, M.; Vu, H.; Walsh, T.; Walson, J. L.; Waner, S.; Wangrangsimakul, T.; Wannapinij, P.; Wozniak, T.; Young

Sharma, T. E. M. W.; Yu, K. C.; Zheng, P.; Sartorius, B.; Lopez, A. D.; Stergachis, A.; Moore, C.; Dolecek, C.; Naghavi, M. Global Burden of Bacterial Antimicrobial Resistance in 2019: A Systematic Analysis. *The Lancet* **2022**, *399* (10325), 629–655. https://doi.org/10.1016/S0140-6736(21)02724-0.

(5) Bizuayehu, H. M.; Ahmed, K. Y.; Kibret, G. D.; Dadi, A. F.; Belachew, S. A.; Bagade, T.; Tegegne, T. K.; Venchiarutti, R. L.; Kibret, K. T.; Hailegebireal, A. H.; Assefa, Y.; Khan, M. N.; Abajobir, A.; Alene, K. A.; Mengesha, Z.; Erku, D.; Enquobahrie, D. A.; Minas, T. Z.; Misgan, E.; Ross, A. G. Global Disparities of Cancer and Its Projected Burden in 2050. *JAMA Netw Open* **2024**, *7* (11), e2443198. https://doi.org/10.1001/jamanetworkopen.2024.43198.

(6) Van Den Boogaard, W. M. C.; Komninos, D. S. J.; Vermeij, W. P. Chemotherapy Side-Effects: Not All DNA Damage Is Equal. *Cancers* **2022**, *14* (3), 627. https://doi.org/10.3390/cancers14030627.

(7) Monsuez, J.-J.; Charniot, J.-C.; Vignat, N.; Artigou, J.-Y. Cardiac Side-Effects of Cancer Chemotherapy. *International Journal of Cardiology* **2010**, *144* (1), 3–15. https://doi.org/10.1016/j.ijcard.2010.03.003.

(8) Wagland, R.; Richardson, A.; Armes, J.; Hankins, M.; Lennan, E.; Griffiths, P. Treatment-Related Problems Experienced by Cancer Patients Undergoing Chemotherapy: A Scoping Review: Treatment-Related Problems in Cancer Chemotherapy. *Eur J Cancer Care (Engl)* **2015**, *24* (5), 605–617. https://doi.org/10.1111/ecc.12246.

(9) Petersen, C.; Würschmidt, F. Late Toxicity of Radiotherapy: A Problem or a Challenge for the Radiation Oncologist? *Breast Care* **2011**, *6* (5), 369–374. https://doi.org/10.1159/000334220.

(10) Nussinov, R.; Tsai, C.-J.; Jang, H. Anticancer Drug Resistance: An Update and Perspective. *Drug Resistance Updates* **2021**, *59*, 100796. https://doi.org/10.1016/j.drup.2021.100796.

(11) Taefehshokr, S.; Parhizkar, A.; Hayati, S.; Mousapour, M.; Mahmoudpour, A.; Eleid, L.; Rahmanpour, D.; Fattahi, S.; Shabani, H.; Taefehshokr, N. Cancer Immunotherapy: Challenges and Limitations. *Pathology - Research and Practice* **2022**, *229*, 153723. https://doi.org/10.1016/j.prp.2021.153723.

(12) Wang, L.; Dong, C.; Li, X.; Han, W.; Su, X. Anticancer Potential of Bioactive Peptides from Animal Sources. *Oncology Reports* **2017**, *38* (2), 637–651. https://doi.org/10.3892/or.2017.5778.

(13) Skjånes, K.; Aesoy, R.; Herfindal, L.; Skomedal, H. Bioactive Peptides from Microalgae: Focus on Anti-cancer and Immunomodulating Activity. *Physiologia Plantarum* **2021**, *173* (2), 612–623. https://doi.org/10.1111/ppl.13472.

(14) Soon, T. N.; Chia, A. Y. Y.; Yap, W. H.; Tang, Y.-Q. Anticancer Mechanisms of Bioactive Peptides. *PPL* **2020**, *27* (9), 823–830. https://doi.org/10.2174/0929866527666200409102747.

(15) Naghavi, M.; Vollset, S. E.; Ikuta, K. S.; Swetschinski, L. R.; Gray, A. P.; Wool, E. E.; Robles Aguilar, G.; Mestrovic, T.; Smith, G.; Han, C.; Hsu, R. L.; Chalek, J.; Araki, D. T.; Chung, E.; Raggi, C.; Gershberg Hayoon, A.; Davis Weaver, N.; Lindstedt, P. A.; Smith, A. E.; Altay, U.; Bhattacharjee, N. V.; Giannakis, K.; Fell, F.; McManigal, B.; Ekapirat, N.; Mendes, J. A.; Runghien, T.; Srimokla, O.; Abdelkader, A.; Abd-Elsalam, S.; Aboagye, R. G.; Abolhassani, H.; Abualruz, H.; Abubakar, U.; Abukhadijah, H. J.; Aburuz, S.; Abu-Zaid, A.; Achalapong, S.; Addo, I. Y.; Adekanmbi, V.; Adeyeoluwa, T. E.; Adnani, Q. E. S.; Adzigbli, L. A.; Afzal, M. S.; Afzal, S.; Agodi, A.; Ahlstrom, A. J.; Ahmad, A.; Ahmad, S.; Ahmad, T.; Ahmadi, A.; Ahmed, A.; Ahmed, H.; Ahmed, I.; Ahmed, M.; Ahmed, S.; Ahmed, S. A.; Akkaif, M. A.; Al Awaidy, S.; Al Thaher, Y.; Alalalmeh, S. O.; AlBataineh, M. T.; Aldhaleei, W. A.; Al-Gheethi, A. A. S.; Alhaji, N. B.; Ali, A.; Ali, L.; Ali, S. S.; Ali, W.; Allel, K.; Al-Marwani, S.; Alrawashdeh, A.; Altaf, A.; Al-Tammemi, A. B.; Al-Tawfiq, J. A.; Alzoubi, K. H.; Al-Zyoud, W. A.; Amos, B.; Amuasi, J. H.; Ancuceanu, R.; Andrews, J. R.; Anil, A.; Anuoluwa, I. A.; Anvari, S.; Anyasodor, A. E.; Apostol,

G. L. C.; Arabloo, J.; Arafat, M.; Aravkin, A. Y.; Areda, D.; Aremu, A.; Artamonov, A. A.; Ashley, E. A.; Asika, M. O.; Athari, S. S.; Atout, M. M. W.; Awoke, T.; Azadnajafabad, S.; Azam, J. M.; Aziz, S.; Azzam, A. Y.; Babaei, M.; Babin, F.-X.; Badar, M.; Baig, A. A.; Bajcetic, M.; Baker, S.; Bardhan, M.; Barqawi, H. J.; Basharat, Z.; Basiru, A.; Bastard, M.; Basu, S.; Bayleyegn, N. S.; Belete, M. A.; Bello, O. O.; Beloukas, A.; Berkley, J. A.; Bhagavathula, A. S.; Bhaskar, S.; Bhuyan, S. S.; Bielicki, J. A.; Briko, N. I.; Brown, C. S.; Browne, A. J.; Buonsenso, D.; Bustanji, Y.; Carvalheiro, C. G.; Castañeda-Orjuela, C. A.; Cenderadewi, M.; Chadwick, J.; Chakraborty, S.; Chandika, R. M.; Chandy, S.; Chansamouth, V.; Chattu, V. K.; Chaudhary, A. A.; Ching, P. R.; Chopra, H.; Chowdhury, F. R.; Chu, D.-T.; Chutiyami, M.; Cruz-Martins, N.; Da Silva, A. G.; Dadras, O.; Dai, X.; Darcho, S. D.; Das, S.; De La Hoz, F. P.; Dekker, D. M.; Dhama, K.; Diaz, D.; Dickson, B. F. R.; Djorie, S. G.; Dodangeh, M.; Dohare, S.; Dokova, K. G.; Doshi, O. P.; Dowou, R. K.; Dsouza, H. L.; Dunachie, S. J.; Dziedzic, A. M.; Eckmanns, T.; Ed-Dra, A.; Eftekharimehrabad, A.; Ekundayo, T. C.; El Sayed, I.; Elhadi, M.; El-Huneidi, W.; Elias, C.; Ellis, S. J.; Elsheikh, R.; Elsohaby, I.; Eltaha, C.; Eshrati, B.; Eslami, M.; Eyre, D. W.; Fadaka, A. O.; Fagbamigbe, A. F.; Fahim, A.; Fakhri-Demeshghieh, A.; Fasina, F. O.; Fasina, M. M.; Fatehizadeh, A.; Feasey, N. A.; Feizkhah, A.; Fekadu, G.; Fischer, F.; Fitriana, I.; Forrest, K. M.; Fortuna Rodrigues, C.; Fuller, J. E.; Gadanya, M. A.; Gajdács, M.; Gandhi, A. P.; Garcia-Gallo, E. E.; Garrett, D. O.; Gautam, R. K.; Gebregergis, M. W.; Gebrehiwot, M.; Gebremeskel, T. G.; Geffers, C.; Georgalis, L.; Ghazy, R. M.; Golechha, M.; Golinelli, D.; Gordon, M.; Gulati, S.; Gupta, R. D.; Gupta, S.; Gupta, V. K.; Habteyohannes, A. D.; Haller, S.; Harapan, H.; Harrison, M. L.; Hasaballah, A. I.; Hasan, I.; Hasan, R. S.; Hasani, H.; Haselbeck, A. H.; Hasnain, M. S.; Hassan, I. I.; Hassan, S.; Hassan Zadeh Tabatabaei, M. S.; Hayat, K.; He, J.; Hegazi, O. E.; Heidari, M.; Hezam, K.; Holla, R.; Holm, M.; Hopkins, H.; Hossain, M. M.; Hosseinzadeh, M.; Hostiuc, S.; Hussein, N. R.; Huy, L. D.; Ibáñez-Prada, E. D.; Ikiroma, A.; Ilic, I. M.; Islam, S. M. S.; Ismail, F.; Ismail, N. E.; Iwu, C. D.; Iwu-Jaja, C. J.; Jafarzadeh, A.; Jaiteh, F.; Jalilzadeh Yengejeh, R.; Jamora, R. D. G.; Javidnia, J.; Jawaid, T.; Jenney, A. W. J.; Jeon, H. J.; Jokar, M.; Jomehzadeh, N.; Joo, T.; Joseph, N.; Kamal, Z.; Kanmodi, K. K.; Kantar, R. S.; Kapisi, J. A.; Karaye, I. M.; Khader, Y. S.; Khajuria, H.; Khalid, N.; Khamesipour, F.; Khan, A.; Khan, M. J.; Khan, M. T.; Khanal, V.; Khidri, F. F.; Khubchandani, J.; Khusuwan, S.; Kim, M. S.; Kisa, A.; Korshunov, V. A.; Krapp, F.; Krumkamp, R.; Kuddus, M.; Kulimbet, M.; Kumar, D.; Kumaran, E. A. P.; Kuttikkattu, A.; Kyu, H. H.; Landires, I.; Lawal, B. K.; Le, T. T. T.; Lederer, I. M.; Lee, M.; Lee, S. W.; Lepape, A.; Lerango, T. L.; Ligade, V. S.; Lim, C.; Lim, S. S.; Limenh, L. W.; Liu, C.; Liu, X.; Liu, X.; Loftus, M. J.; M Amin, H. I.; Maass, K. L.; Maharaj, S. B.; Mahmoud, M. A.; Maikanti-Charalampous, P.; Makram, O. M.; Malhotra, K.; Malik, A. A.; Mandilara, G. D.; Marks, F.; Martinez-Guerra, B. A.; Martorell, M.; Masoumi-Asl, H.; Mathioudakis, A. G.; May, J.; McHugh, T. A.; Meiring, J.; Meles, H. N.; Melese, A.; Melese, E. B.; Minervini, G.; Mohamed, N. S.; Mohammed, S.; Mohan, S.; Mokdad, A. H.; Monasta, L.; Moodi Ghalibaf, A.; Moore, C. E.; Moradi, Y.; Mossialos, E.; Mougin, V.; Mukoro, G. D.; Mulita, F.; Muller-Pebody, B.; Murillo-Zamora, E.; Musa, S.; Musicha, P.; Musila, L. A.; Muthupandian, S.; Nagarajan, A. J.; Naghavi, P.; Nainu, F.; Nair, T. S.; Najmuldeen, H. H. R.; Natto, Z. S.; Nauman, J.; Nayak, B. P.; Nchanji, G. T.; Ndishimye, P.; Negoi, I.; Negoi, R. I.; Nejadghaderi, S. A.; Nguyen, Q. P.; Noman, E. A.; Nwakanma, D. C.; O'Brien, S.; Ochoa, T. J.; Odetokun, I. A.; Ogundijo, O. A.; Ojo-Akosile, T. R.; Okeke, S. R.; Okonji, O. C.; Olagunju, A. T.; Olivas-Martinez, A.; Olorukooba, A. A.; Olwoch, P.; Onyedibe, K. I.; Ortiz-Brizuela, E.; Osuolale, O.; Ounchanum, P.; Oyeyemi, O. T.; P A, M. P.; Paredes, J. L.; Parikh, R. R.; Patel, J.; Patil, S.; Pawar, S.; Peleg, A. Y.; Peprah, P.; Perdigão, J.; Perrone, C.; Petcu, I.-R.; Phommasone, K.; Piracha, Z. Z.; Poddighe, D.; Pollard, A. J.; Poluru, R.; Ponce-De-Leon, A.; Puvvula, J.; Qamar, F. N.; Qasim, N. H.; Rafai, C. D.; Raghav, P.; Rahbarnia, L.; Rahim, F.; Rahimi-Movaghar, V.; Rahman, M.; Rahman, M. A.; Ramadan, H.; Ramasamy, S. K.; Ramesh, P. S.; Ramteke, P. W.; Rana, R. K.; Rani, U.; Rashidi, M.-M.; Rathish, D.; Rattanavong, S.; Rawaf, S.; Redwan, E. M.

M.; Reyes, L. F.; Roberts, T.; Robotham, J. V.; Rosenthal, V. D.; Ross, A. G.; Roy, N.; Rudd, K. E.; Sabet, C. J.; Saddik, B. A.; Saeb, M. R.; Saeed, U.; Saeedi Moghaddam, S.; Saengchan, W.; Safaei, M.; Saghazadeh, A.; Saheb Sharif-Askari, N.; Sahebkar, A.; Sahoo, S. S.; Sahu, M.; Saki, M.; Salam, N.; Saleem, Z.; Saleh, M. A.; Samodra, Y. L.; Samy, A. M.; Saravanan, A.; Satpathy, M.; Schumacher, A. E.; Sedighi, M.; Seekaew, S.; Shafie, M.; Shah, P. A.; Shahid, S.; Shahwan, M. J.; Shakoor, S.; Shalev, N.; Shamim, M. A.; Shamshirgaran, M. A.; Shamsi, A.; Sharifan, A.; Shastry, R. P.; Shetty, M.; Shittu, A.; Shrestha, S.; Siddig, E. E.; Sideroglou, T.; Sifuentes-Osornio, J.; Silva, L. M. L. R.; Simões, E. A. F.; Simpson, A. J. H.; Singh, A.; Singh, S.; Sinto, R.; Soliman, S. S. M.; Soraneh, S.; Stoesser, N.; Stoeva, T. Z.; Swain, C. K.; Szarpak, L.; T Y, S. S.; Tabatabai, S.; Tabche, C.; Taha, Z. M.-A.; Tan, K.-K.; Tasak, N.; Tat, N. Y.; Thaiprakong, A.; Thangaraju, P.; Tigoi, C. C.; Tiwari, K.; Tovani-Palone, M. R.; Tran, T. H.; Tumurkhuu, M.; Turner, P.; Udoakang, A. J.; Udoh, A.; Ullah, N.; Ullah, S.; Vaithinathan, A. G.; Valenti, M.; Vos, T.; Vu, H. T. L.; Waheed, Y.; Walker, A. S.; Walson, J. L.; Wangrangsimakul, T.; Weerakoon, K. G.; Wertheim, H. F. L.; Williams, P. C. M.; Wolde, A. A.; Wozniak, T. M.; Wu, F.; Wu, Z.; Yadav, M. K. K.; Yaghoubi, S.; Yahaya, Z. S.; Yarahmadi, A.; Yezli, S.; Yismaw, Y. E.; Yon, D. K.; Yuan, C.-W.; Yusuf, H.; Zakham, F.; Zamagni, G.; Zhang, H.; Zhang, Z.-J.; Zielińska, M.; Zumla, A.; Zyoud, S. H. H.; Zyoud, S. H.; Hay, S. I.; Stergachis, A.; Sartorius, B.; Cooper, B. S.; Dolecek, C.; Murray, C. J. L. Global Burden of Bacterial Antimicrobial Resistance 1990–2021: A Systematic Analysis with Forecasts to 2050. *The Lancet* **2024**, *404* (10459), 1199–1226. https://doi.org/10.1016/S0140-6736(24)01867-1.

(16) Irfan, M.; Almotiri, A.; AlZeyadi, Z. A. Antimicrobial Resistance and Its Drivers—A Review. *Antibiotics* **2022**, *11* (10), 1362. https://doi.org/10.3390/antibiotics11101362.

(17) Lau, C. H.-F.; Van Engelen, K.; Gordon, S.; Renaud, J.; Topp, E. Novel Antibiotic Resistance Determinants from Agricultural Soil Exposed to Antibiotics Widely Used in Human Medicine and Animal Farming. *Appl Environ Microbiol* **2017**, *83* (16), e00989-17. https://doi.org/10.1128/AEM.00989-17.

(18) Karwowska, E. Antibiotic Resistance in the Farming Environment. *Applied Sciences* **2024**, *14* (13), 5776. https://doi.org/10.3390/app14135776.

(19) Singh, B.; Bhat, A.; Ravi, K. Antibiotics Misuse and Antimicrobial Resistance Development in Agriculture: A Global Challenge. *Environ. Health* **2024**, *2* (9), 618–622. https://doi.org/10.1021/envhealth.4c00094.

(20) Moosazadeh Moghaddam, M.; Eftekhary, M.; Erfanimanesh, S.; Hashemi, A.; Fallah Omrani, V.; Farhadihosseinabadi, B.; Lasjerdi, Z.; Mossahebi-Mohammadi, M.; Pal Singh Chauhan, N.; Seifalian, A. M.; Gholipourmalekabadi, M. Comparison of the Antibacterial Effects of a Short Cationic Peptide and 1% Silver Bioactive Glass against Extensively Drug-Resistant Bacteria, Pseudomonas Aeruginosa and Acinetobacter Baumannii, Isolated from Burn Patients. *Amino Acids* **2018**, *50* (11), 1617–1628. https://doi.org/10.1007/s00726-018-2638-z.

(21) Mansour, S. C.; De La Fuente-Núñez, C.; Hancock, R. E. W. Peptide IDR-1018: Modulating the Immune System and Targeting Bacterial Biofilms to Treat Antibiotic-resistant Bacterial Infections. *Journal of Peptide Science* **2015**, *21* (5), 323–329. https://doi.org/10.1002/psc.2708.

(22) Shahidi, F.; Zhong, Y. Bioactive Peptides. *Journal of AOAC INTERNATIONAL* **2008**, *91* (4), 914–931. https://doi.org/10.1093/jaoac/91.4.914.

(23) Sánchez, A.; Vázquez, A. Bioactive Peptides: A Review. *Food Quality and Safety* **2017**, *1* (1), 29–46. https://doi.org/10.1093/fqs/fyx006.

(24) Daliri, E.; Oh, D.; Lee, B. Bioactive Peptides. *Foods* **2017**, *6* (5), 32. https://doi.org/10.3390/foods6050032.

(25) Ahmed, T.; Sun, X.; Udenigwe, C. C. Role of Structural Properties of Bioactive Peptides in Their Stability during Simulated Gastrointestinal Digestion: A Systematic Review. *Trends in Food Science & Technology* **2022**, *120*, 265–273. https://doi.org/10.1016/j.tifs.2022.01.008.

(26) Contreras, M. D. M.; Sanchez, D.; Sevilla, M. Á.; Recio, I.; Amigo, L. Resistance of Casein-Derived Bioactive Peptides to Simulated Gastrointestinal Digestion. *International Dairy Journal* **2013**, *32* (2), 71–78. https://doi.org/10.1016/j.idairyj.2013.05.008.
(27) Tugyi, R.; Uray, K.; Iván, D.; Fellinger, E.; Perkins, A.; Hudecz, F. Partial D -Amino Acid Substitution: Improved Enzymatic Stability and Preserved Ab Recognition of a MUC2 Epitope Peptide. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102* (2), 413–418. https://doi.org/10.1073/pnas.0407677102.
(28) Khafagy, E.-S.; Morishita, M. Oral Biodrug Delivery Using Cell-Penetrating Peptide. *Advanced Drug Delivery Reviews* **2012**, *64* (6), 531–539. https://doi.org/10.1016/j.addr.2011.12.014.
(29) Kamei, N.; Morishita, M.; Eda, Y.; Ida, N.; Nishio, R.; Takayama, K. Usefulness of Cell-Penetrating Peptides to Improve Intestinal Insulin Absorption. *Journal of Controlled Release* **2008**, *132* (1), 21–25. https://doi.org/10.1016/j.jconrel.2008.08.001.
(30) Nielsen, E. J. B.; Yoshida, S.; Kamei, N.; Iwamae, R.; Khafagy, E.-S.; Olsen, J.; Rahbek, U. L.; Pedersen, B. L.; Takayama, K.; Takeda-Morishita, M. In Vivo Proof of Concept of Oral Insulin Delivery Based on a Co-Administration Strategy with the Cell-Penetrating Peptide Penetratin. *Journal of Controlled Release* **2014**, *189*, 19–24. https://doi.org/10.1016/j.jconrel.2014.06.022.
(31) Guidotti, G.; Brambilla, L.; Rossi, D. Cell-Penetrating Peptides: From Basic Research to Clinics. *Trends in Pharmacological Sciences* **2017**, *38* (4), 406–424. https://doi.org/10.1016/j.tips.2017.01.003.
(32) Ruseska, I.; Zimmer, A. Internalization Mechanisms of Cell-Penetrating Peptides. *Beilstein Journal of Nanotechnology* **2020**, *11*, 101–123. https://doi.org/10.3762/bjnano.11.10.
(33) Frankel, A. D.; Pabo, C. O. Cellular Uptake of the Tat Protein from Human Immunodeficiency Virus. *Cell* **1988**, *55* (6), 1189–1193. https://doi.org/10.1016/0092-8674(88)90263-2.
(34) Green, M.; Loewenstein, P. M. Autonomous Functional Domains of Chemically Synthesized Human Immunodeficiency Virus Tat Trans-Activator Protein. *Cell* **1988**, *55* (6), 1179–1188. https://doi.org/10.1016/0092-8674(88)90262-0.
(35) Elliott, G.; O'Hare, P. Intercellular Trafficking and Protein Delivery by a Herpesvirus Structural Protein. *Cell* **1997**, *88* (2), 223–233. https://doi.org/10.1016/S0092-8674(00)81843-7.
(36) Pooga, M.; Soomets, U.; Hällbrink, M.; Valkna, A.; Saar, K.; Rezaei, K.; Kahl, U.; Hao, J. X.; Xu, X. J.; Wiesenfeld-Hallin, Z.; Hökfelt, T.; Bartfai, T.; Langel, Ü. Cell Penetrating PNA Constructs Regulate Galanin Receptor Levels and Modify Pain Transmission in Vivo. *Nature Biotechnology* **1998**, *16* (9), 857–861. https://doi.org/10.1038/nbt0998-857.
(37) Futaki, S.; Suzuki, T.; Ohashi, W.; Yagami, T.; Tanaka, S.; Ueda, K.; Sugiura, Y. Arginine-Rich Peptides. An Abundant Source of Membrane-Permeable Peptides Having Potential as Carriers for Intracellular Protein Delivery. *Journal of Biological Chemistry* **2001**, *276* (8), 5836–5840. https://doi.org/10.1074/jbc.M007540200.
(38) Klauschenz, E.; Scheller, A.; Wiesner, B.; Bienert, M.; Krause, E.; Oehlke, J.; Melzig, M.; Beyermann, M. Cellular Uptake of an α-Helical Amphipathic Model Peptide with the Potential to Deliver Polar Compounds into the Cell Interior Non-Endocytically. *Biochimica et Biophysica Acta (BBA) - Biomembranes* **2002**, *1414* (1–2), 127–139. https://doi.org/10.1016/s0005-2736(98)00161-8.
(39) Bechara, C.; Sagan, S. Cell-Penetrating Peptides: 20 Years Later, Where Do We Stand? *FEBS Letters*. No longer published by Elsevier June 2013, pp 1693–1702. https://doi.org/10.1016/j.febslet.2013.04.031.
(40) Gori, A.; Lodigiani, G.; Colombarolli, S. G.; Bergamaschi, G.; Vitali, A. Cell Penetrating Peptides: Classification, Mechanisms, Methods of Study, and Applications. *ChemMedChem* **2023**, *18* (17), e202300236. https://doi.org/10.1002/cmdc.202300236.

(41) Hayashi, T.; Shinagawa, M.; Kawano, T.; Iwasaki, T. Drug Delivery Using Polyhistidine Peptide-Modified Liposomes That Target Endogenous Lysosome. *Biochemical and Biophysical Research Communications* **2018**, *501* (3), 648–653. https://doi.org/10.1016/j.bbrc.2018.05.037.

(42) Mansur, A. A. P.; Carvalho, S. M.; Lobato, Z. I. P.; Leite, M. de F.; Cunha, A. da S.; Mansur, H. S. Design and Development of Polysaccharide-Doxorubicin-Peptide Bioconjugates for Dual Synergistic Effects of Integrin-Targeted and Cell-Penetrating Peptides for Cancer Chemotherapy. *Bioconjugate Chemistry* **2018**, *29* (6), 1973–2000. https://doi.org/10.1021/acs.bioconjchem.8b00208.

(43) Liu, A.; Xu, H.; Gao, Y.; Luo, D.; Li, Z.; Voss, C.; Li, S. S. C.; Cao, X. (Arg)9-SH2 Superbinder: A Novel Promising Anticancer Therapy to Melanoma by Blocking Phosphotyrosine Signaling. *Journal of Experimental & Clinical Cancer Research* **2018**, *37* (1), 138. https://doi.org/10.1186/s13046-018-0812-5.

(44) Miao, J.; Guo, H.; Chen, F.; Zhao, L.; He, L.; Ou, Y.; Huang, M.; Zhang, Y.; Guo, B.; Cao, Y.; Huang, Q. Antibacterial Effects of a Cell-Penetrating Peptide Isolated from Kefir. *J. Agric. Food Chem.* **2016**, *64* (16), 3234–3242. https://doi.org/10.1021/acs.jafc.6b00730.

(45) Cruz, G. S.; Santos, A. T. D.; Brito, E. H. S. D.; Rádis-Baptista, G. Cell-Penetrating Antimicrobial Peptides with Anti-Infective Activity against Intracellular Pathogens. *Antibiotics* **2022**, *11* (12), 1772. https://doi.org/10.3390/antibiotics11121772.

(46) Oh, D.; Sun, J.; Nasrolahi Shirazi, A.; LaPlante, K. L.; Rowley, D. C.; Parang, K. Antibacterial Activities of Amphiphilic Cyclic Cell-Penetrating Peptides against Multidrug-Resistant Pathogens. *Mol. Pharmaceutics* **2014**, *11* (10), 3528–3536. https://doi.org/10.1021/mp5003027.

(47) Niu, Z.; Samaridou, E.; Jaumain, E.; Coëne, J.; Ullio, G.; Shrestha, N.; Garcia, J.; Durán-Lobato, M.; Tovar, S.; Santander-Ortega, M. J.; Lozano, M. V.; Arroyo-Jimenez, M. M.; Ramos-Membrive, R.; Peñuelas, I.; Mabondzo, A.; Préat, V.; Teixidó, M.; Giralt, E.; Alonso, M. J. PEG-PGA Enveloped Octaarginine-Peptide Nanocomplexes: An Oral Peptide Delivery Strategy. *Journal of Controlled Release* **2018**, *276*, 125–139. https://doi.org/10.1016/j.jconrel.2018.03.004.

(48) Chen, C.; Liu, K.; Xu, Y.; Zhang, P.; Suo, Y.; Lu, Y.; Zhang, W.; Su, L.; Gu, Q.; Wang, H.; Gu, J.; Li, Z.; Xu, X. Anti-Angiogenesis through Noninvasive to Minimally Invasive Intraocular Delivery of the Peptide CC12 Identified by in Vivo-Directed Evolution. *Biomaterials* **2017**, *112*, 218–233. https://doi.org/10.1016/j.biomaterials.2016.09.022.

(49) Kamei, N.; Yamaoka, A.; Fukuyama, Y.; Itokazu, R.; Takeda-Morishita, M. Noncovalent Strategy with Cell-Penetrating Peptides to Facilitate the Brain Delivery of Insulin through the Blood–Brain Barrier. *Biological and Pharmaceutical Bulletin* **2018**, *41* (4), 546–554. https://doi.org/10.1248/bpb.b17-00848.

(50) Suda, K.; Murakami, T.; Gotoh, N.; Fukuda, R.; Hashida, Y.; Hashida, M.; Tsujikawa, A.; Yoshimura, N. High-Density Lipoprotein Mutant Eye Drops for the Treatment of Posterior Eye Diseases. *Journal of Controlled Release* **2017**, *266*, 301–309. https://doi.org/10.1016/j.jconrel.2017.09.036.

(51) Du, Y.; Wang, L.; Wang, W.; Guo, T.; Zhang, M.; Zhang, P.; Zhang, Y.; Wu, K.; Li, A.; Wang, X.; He, J.; Fan, J. Novel Application of Cell Penetrating R11 Peptide for Diagnosis of Bladder Cancer. *Journal of Biomedical Nanotechnology* **2018**, *14* (1), 161–167. https://doi.org/10.1166/jbn.2018.2499.

(52) Simón-Gracia, L.; Scodeller, P.; Fuentes, S. S.; Vallejo, V. G.; Ríos, X.; San Sebastián, E.; Sidorenko, V.; Di Silvio, D.; Suck, M.; De Lorenzi, F.; Rizzo, L. Y.; von Stillfried, S.; Kilk, K.; Lammers, T.; Moya, S. E.; Teesalu, T. Application of Polymersomes Engineered to Target P32 Protein for Detection of Small Breast Tumors in Mice. *Oncotarget* **2018**, *9* (27), 18682–18697. https://doi.org/10.18632/oncotarget.24588.

(53) Ji, X.; Lv, H.; Guo, J.; Ding, C.; Luo, X. A DNA Nanotube-Peptide Biocomplex for mRNA Detection and Its Application in Cancer Diagnosis and Targeted Therapy. *Chemistry - A European Journal* **2018**, *24* (40), 10171–10177. https://doi.org/10.1002/chem.201801347.
(54) Collard, R.; Majtan, T.; Park, I.; Kraus, J. P. Import of TAT-Conjugated Propionyl Coenzyme A Carboxylase Using Models of Propionic Acidemia. *Molecular and Cellular Biology* **2018**, *38* (6). https://doi.org/10.1128/MCB.00491-17.
(55) Tsai, C.-W.; Lin, Z.-W.; Chang, W.-F.; Chen, Y.-F.; Hu, W.-W. Development of an Indolicidin-Derived Peptide by Reducing Membrane Perturbation to Decrease Cytotoxicity and Maintain Gene Delivery Ability. *Colloids and Surfaces B: Biointerfaces* **2018**, *165*, 18–27. https://doi.org/10.1016/j.colsurfb.2018.02.007.
(56) Vaissière, A.; Aldrian, G.; Konate, K.; Lindberg, M. F.; Jourdan, C.; Telmar, A.; Seisel, Q.; Fernandez, F.; Viguier, V.; Genevois, C.; Couillaud, F.; Boisguerin, P.; Deshayes, S. A Retro-Inverso Cell-Penetrating Peptide for siRNA Delivery. *Journal of Nanobiotechnology* **2017**, *15* (1), 34. https://doi.org/10.1186/s12951-017-0269-2.
(57) Lee, Y. W.; Hwang, Y. E.; Lee, J. Y.; Sohn, J.-H.; Sung, B. H.; Kim, S. C. VEGF siRNA Delivery by a Cancer-Specific Cell-Penetrating Peptide. *Journal of Microbiology and Biotechnology* **2018**, *28* (3), 367–374. https://doi.org/10.4014/jmb.1711.11025.
(58) Meng, Z.; Kang, Z.; Sun, C.; Yang, S.; Zhao, B.; Feng, S.; Meng, Q.; Liu, K. Enhanced Gene Transfection Efficiency by Use of Peptide Vectors Containing Laminin Receptor-Targeting Sequence YIGSR. *Nanoscale* **2018**, *10* (3), 1215–1227. https://doi.org/10.1039/C7NR05843H.
(59) Soudah, T.; Mogilevsky, M.; Karni, R.; Yavin, E. CLIP6-PNA-Peptide Conjugates: Non-Endosomal Delivery of Splice Switching Oligonucleotides. *Bioconjugate Chemistry* **2017**, *28* (12), 3036–3042. https://doi.org/10.1021/acs.bioconjchem.7b00638.
(60) Liu, F.; Lou, J.; Hristov, D. X-Ray Responsive Nanoparticles with Triggered Release of Nitrite, a Precursor of Reactive Nitrogen Species, for Enhanced Cancer Radiosensitization. *Nanoscale* **2017**, *9* (38), 14627–14634. https://doi.org/10.1039/C7NR04684G.
(61) Ma, N.; Liu, P.; He, N.; Gu, N.; Wu, F.-G.; Chen, Z. Action of Gold Nanospikes-Based Nanoradiosensitizers: Cellular Internalization, Radiotherapy, and Autophagy. *ACS Applied Materials & Interfaces* **2017**, *9* (37), 31526–31542. https://doi.org/10.1021/acsami.7b09599.
(62) Fan, Y.-X.; Liang, Z.-X.; Liu, Q.-Z.; Xiao, H.; Li, K.-B.; Wu, J.-Z. Cell Penetrating Peptide of Sodium-Iodide Symporter Effect on the I-131 Radiotherapy on Thyroid Cancer. *Experimental and Therapeutic Medicine* **2017**, *13* (3), 989–994. https://doi.org/10.3892/etm.2017.4079.
(63) Holm, T.; Johansson, H.; Lundberg, P.; Pooga, M.; Lindgren, M.; Langel, Ü. Studying the Uptake of Cell-Penetrating Peptides. *Nature Protocols* **2006**, *1* (2), 1001–1005. https://doi.org/10.1038/nprot.2006.174.
(64) Derakhshankhah, H.; Jafari, S. Cell Penetrating Peptides: A Concise Review with Emphasis on Biomedical Applications. *Biomedicine & Pharmacotherapy* **2018**, *108*, 1090–1096. https://doi.org/10.1016/j.biopha.2018.09.097.
(65) Illien, F.; Rodriguez, N.; Amoura, M.; Joliot, A.; Pallerla, M.; Cribier, S.; Burlina, F.; Sagan, S. Quantitative Fluorescence Spectroscopy and Flow Cytometry Analyses of Cell-Penetrating Peptides Internalization Pathways: Optimization, Pitfalls, Comparison with Mass Spectrometry Quantification. *Scientific Reports* **2016**, *6* (1), 36938. https://doi.org/10.1038/srep36938.
(66) Mueller, J.; Kretzschmar, I.; Volkmer, R.; Boisguerin, P. Comparison of Cellular Uptake Using 22 CPPs in 4 Different Cell Lines. *Bioconjugate Chemistry* **2008**, *19* (12), 2363–2374. https://doi.org/10.1021/bc800194e.
(67) Vasconcelos, L.; Pärn, K.; Langel, Ü. Therapeutic Potential of Cell-Penetrating Peptides. *Therapeutic Delivery* **2013**, *4* (5), 573–591. https://doi.org/10.4155/tde.13.22.

(68) Trabulo, S.; Cardoso, A. L.; Mano, M.; De Lima, M. C. P. Cell-Penetrating Peptides—Mechanisms of Cellular Uptake and Generation of Delivery Systems. *Pharmaceuticals* **2010**, *3* (4), 961–993. https://doi.org/10.3390/ph3040961.
(69) Jones, A. T.; Sayers, E. J. Cell Entry of Cell Penetrating Peptides: Tales of Tails Wagging Dogs. *Journal of Controlled Release* **2012**, *161* (2), 582–591. https://doi.org/10.1016/j.jconrel.2012.04.003.
(70) Madani, F.; Lindberg, S.; Langel, Ü.; Futaki, S.; Gräslund, A. Mechanisms of Cellular Uptake of Cell-Penetrating Peptides. *Journal of Biophysics* **2011**, *2011*, 1–10. https://doi.org/10.1155/2011/414729.
(71) Verdurmen, W. P. R.; Thanos, M.; Ruttekolk, I. R.; Gulbins, E.; Brock, R. Cationic Cell-Penetrating Peptides Induce Ceramide Formation via Acid Sphingomyelinase: Implications for Uptake. *Journal of Controlled Release* **2010**, *147* (2), 171–179. https://doi.org/10.1016/j.jconrel.2010.06.030.
(72) Ziegler, A. Thermodynamic Studies and Binding Mechanisms of Cell-Penetrating Peptides with Lipids and Glycosaminoglycans. *Advanced Drug Delivery Reviews* **2008**, *60* (4–5), 580–597. https://doi.org/10.1016/j.addr.2007.10.005.
(73) Bechara, C.; Pallerla, M.; Zaltsman, Y.; Burlina, F.; Alves, I. D.; Lequin, O.; Sagan, S. Tryptophan within Basic Peptide Sequences Triggers Glycosaminoglycan- Dependent Endocytosis. *FASEB Journal* **2013**, *27* (2), 738–749. https://doi.org/10.1096/fj.12-216176.
(74) Hirose, H.; Takeuchi, T.; Osakada, H.; Pujals, S.; Katayama, S.; Nakase, I.; Kobayashi, S.; Haraguchi, T.; Futaki, S. Transient Focal Membrane Deformation Induced by Arginine-Rich Peptides Leads to Their Direct Penetration into Cells. *Molecular Therapy* **2012**, *20* (5), 984–993. https://doi.org/10.1038/mt.2011.313.
(75) Melikov, K.; Hara, A.; Yamoah, K.; Zaitseva, E.; Zaitsev, E.; Chernomordik, L. V. Efficient Entry of Cell-Penetrating Peptide Nona-Arginine into Adherent Cells Involves a Transient Increase in Intracellular Calcium. *Biochemical Journal* **2015**, *471* (2), 221–230. https://doi.org/10.1042/BJ20150272.
(76) Hemmati, S.; Rasekhi Kazerooni, H. Polypharmacological Cell-Penetrating Peptides from Venomous Marine Animals Based on Immunomodulating, Antimicrobial, and Anticancer Properties. *Marine Drugs* **2022**, *20* (12), 763. https://doi.org/10.3390/md20120763.
(77) Xuan, J.; Feng, W.; Wang, J.; Wang, R.; Zhang, B.; Bo, L.; Chen, Z.-S.; Yang, H.; Sun, L. Antimicrobial Peptides for Combating Drug-Resistant Bacterial Infections. *Drug Resistance Updates* **2023**, *68*, 100954. https://doi.org/10.1016/j.drup.2023.100954.
(78) Macyszyn, J.; Chyży, P.; Burmistrz, M.; Lobka, M.; Miszkiewicz, J.; Wojciechowska, M.; Trylska, J. Structural Dynamics Influences the Antibacterial Activity of a Cell-Penetrating Peptide (KFF)3K. *Sci Rep* **2023**, *13* (1), 14826. https://doi.org/10.1038/s41598-023-38745-y.
(79) Rowland, L.; Marjault, H.-B.; Karmi, O.; Grant, D.; Webb, L. J.; Friedler, A.; Nechushtai, R.; Elber, R.; Mittler, R. A Combination of a Cell Penetrating Peptide and a Protein Translation Inhibitor Kills Metastatic Breast Cancer Cells. *Cell Death Discov.* **2023**, *9* (1), 325. https://doi.org/10.1038/s41420-023-01627-3.
(80) Hadjicharalambous, A.; Bournakas, N.; Newman, H.; Skynner, M. J.; Beswick, P. Antimicrobial and Cell-Penetrating Peptides: Understanding Penetration for the Design of Novel Conjugate Antibiotics. *Antibiotics* **2022**, *11* (11), 1636. https://doi.org/10.3390/antibiotics11111636.
(81) Ruoslahti, E. Tumor Penetrating Peptides for Improved Drug Delivery. *Advanced Drug Delivery Reviews* **2017**, *110–111*, 3–12. https://doi.org/10.1016/j.addr.2016.03.008.
(82) Wang, L.; Qu, L.; Lin, S.; Yang, Q.; Zhang, X.; Jin, L.; Dong, H.; Sun, D. Biological Functions and Applications of Antimicrobial Peptides. *CPPS* **2022**, *23* (4), 226–247. https://doi.org/10.2174/1389203723666220519155942.

(83) Kordi, M.; Borzouyi, Z.; Chitsaz, S.; Asmaei, M. H.; Salami, R.; Tabarzad, M. Antimicrobial Peptides with Anticancer Activity: Today Status, Trends and Their Computational Design. *Archives of Biochemistry and Biophysics* **2023**, *733*, 109484. https://doi.org/10.1016/j.abb.2022.109484.
(84) Lath, A.; Santal, A. R.; Kaur, N.; Kumari, P.; Singh, N. P. Anti-Cancer Peptides: Their Current Trends in the Development of Peptide-Based Therapy and Anti-Tumor Drugs. *Biotechnology and Genetic Engineering Reviews* **2023**, *39* (1), 45–84. https://doi.org/10.1080/02648725.2022.2082157.
(85) Vakili, B.; Jahanian-Najafabadi, A. Application of Antimicrobial Peptides in the Design and Production of Anticancer Agents. *Int J Pept Res Ther* **2023**, *29* (2), 28. https://doi.org/10.1007/s10989-023-10501-w.
(86) Mohan, K. V. K.; Rao, S. S.; Atreya, C. D. Antiviral Activity of Selected Antimicrobial Peptides against Vaccinia Virus. *Antiviral Research* **2010**, *86* (3), 306–311. https://doi.org/10.1016/j.antiviral.2010.03.012.
(87) Vanzolini, T.; Bruschi, M.; Rinaldi, A. C.; Magnani, M.; Fraternale, A. Multitalented Synthetic Antimicrobial Peptides and Their Antibacterial, Antifungal and Antiviral Mechanisms. *IJMS* **2022**, *23* (1), 545. https://doi.org/10.3390/ijms23010545.
(88) Huan, Y.; Kong, Q.; Mou, H.; Yi, H. Antimicrobial Peptides: Classification, Design, Application and Research Progress in Multiple Fields. *Front. Microbiol.* **2020**, *11*, 582779. https://doi.org/10.3389/fmicb.2020.582779.
(89) Lee, P.; Yen, C.; Lin, C.; Lung, F. T. Designing the Antimicrobial Peptide with Centrosymmetric and Amphipathic Characterizations for Improving Antimicrobial Activity. *Journal of Peptide Science* **2023**, *29* (11), e3510. https://doi.org/10.1002/psc.3510.
(90) Garvey, M. Antimicrobial Peptides Demonstrate Activity against Resistant Bacterial Pathogens. *Infectious Disease Reports* **2023**, *15* (4), 454–469. https://doi.org/10.3390/idr15040046.
(91) Van Eijk, M.; Boerefijn, S.; Cen, L.; Rosa, M.; Morren, M. J. H.; Van Der Ent, C. K.; Kraak, B.; Dijksterhuis, J.; Valdes, I. D.; Haagsman, H. P.; De Cock, H. Cathelicidin-Inspired Antimicrobial Peptides as Novel Antifungal Compounds. *Medical Mycology* **2020**, *58* (8), 1073–1084. https://doi.org/10.1093/mmy/myaa014.
(92) Elnagdy, S.; AlKhazindar, M. The Potential of Antimicrobial Peptides as an Antiviral Therapy against COVID-19. *ACS Pharmacol. Transl. Sci.* **2020**, *3* (4), 780–782. https://doi.org/10.1021/acsptsci.0c00059.
(93) Lee, H.; Lim, S. I.; Shin, S.-H.; Lim, Y.; Koh, J. W.; Yang, S. Conjugation of Cell-Penetrating Peptides to Antimicrobial Peptides Enhances Antibacterial Activity. *ACS Omega* **2019**, *4* (13), 15694–15701. https://doi.org/10.1021/acsomega.9b02278.
(94) Buccini, D. F.; Cardoso, M. H.; Franco, O. L. Antimicrobial Peptides and Cell-Penetrating Peptides for Treating Intracellular Bacterial Infections. *Front. Cell. Infect. Microbiol.* **2021**, *10*, 612931. https://doi.org/10.3389/fcimb.2020.612931.
(95) Zeiders, S. M.; Chmielewski, J. Antibiotic–Cell-penetrating Peptide Conjugates Targeting Challenging Drug-resistant and Intracellular Pathogenic Bacteria. *Chem Biol Drug Des* **2021**, *98* (5), 762–778. https://doi.org/10.1111/cbdd.13930.
(96) Tornesello, A. L.; Borrelli, A.; Buonaguro, L.; Buonaguro, F. M.; Tornesello, M. L. Antimicrobial Peptides as Anticancer Agents: Functional Properties and Biological Activities. *Molecules* **2020**, *25* (12), 2850. https://doi.org/10.3390/molecules25122850.
(97) Li, F.-M.; Wang, X.-Q. Identifying Anticancer Peptides by Using Improved Hybrid Compositions. *Sci Rep* **2016**, *6* (1), 33910. https://doi.org/10.1038/srep33910.
(98) Chinnadurai, R. K.; Khan, N.; Meghwanshi, G. K.; Ponne, S.; Althobiti, M.; Kumar, R. Current Research Status of Anti-Cancer Peptides: Mechanism of Action, Production, and Clinical

Applications. *Biomedicine & Pharmacotherapy* **2023**, *164*, 114996. https://doi.org/10.1016/j.biopha.2023.114996.

(99) Fu, Y.; Yang, S.; Liu, Y.; Liu, J.; Wang, Q.; Li, F.; Shang, X.; Teng, Y.; Guo, N.; Yu, P. Peptide Modified Albumin–Paclitaxel Nanoparticles for Improving Chemotherapy and Preventing Metastasis. *Macromolecular Bioscience* **2022**, *22* (3), 2100404. https://doi.org/10.1002/mabi.202100404.

(100) Szlasa, W.; Zendran, I.; Zalesińska, A.; Tarek, M.; Kulbacka, J. Lipid Composition of the Cancer Cell Membrane. *J Bioenerg Biomembr* **2020**, *52* (5), 321–342. https://doi.org/10.1007/s10863-020-09846-4.

(101) Zhao, J.; Hao, X.; Liu, D.; Huang, Y.; Chen, Y. In Vitro Characterization of the Rapid Cytotoxicity of Anticancer Peptide HPRP-A2 through Membrane Destruction and Intracellular Mechanism against Gastric Cancer Cell Lines. *PLoS ONE* **2015**, *10* (9), e0139578. https://doi.org/10.1371/journal.pone.0139578.

(102) Norouzi, P.; Mirmohammadi, M.; Houshdar Tehrani, M. H. Anticancer Peptides Mechanisms, Simple and Complex. *Chemico-Biological Interactions* **2022**, *368*, 110194. https://doi.org/10.1016/j.cbi.2022.110194.

(103) Gabernet, G.; Müller, A. T.; Hiss, J. A.; Schneider, G. Membranolytic Anticancer Peptides. *Med. Chem. Commun.* **2016**, *7* (12), 2232–2245. https://doi.org/10.1039/C6MD00376A.

(104) Chiangjong, W.; Chutipongtanate, S.; Hongeng, S. Anticancer Peptide: Physicochemical Property, Functional Aspect and Trend in Clinical Application (Review). *Int J Oncol* **2020**, *57* (3), 678–696. https://doi.org/10.3892/ijo.2020.5099.

(105) Karami Fath, M.; Babakhaniyan, K.; Zokaei, M.; Yaghoubian, A.; Akbari, S.; Khorsandi, M.; Soofi, A.; Nabi-Afjadi, M.; Zalpoor, H.; Jalalifar, F.; Azargoonjahromi, A.; Payandeh, Z.; Alagheband Bahrami, A. Anti-Cancer Peptide-Based Therapeutic Strategies in Solid Tumors. *Cell Mol Biol Lett* **2022**, *27* (1), 33. https://doi.org/10.1186/s11658-022-00332-w.

(106) Liu, S.; Yang, H.; Wan, L.; Cheng, J.; Lu, X. Penetratin-Mediated Delivery Enhances the Antitumor Activity of the Cationic Antimicrobial Peptide Magainin II. *Cancer Biotherapy and Radiopharmaceuticals* **2013**, *28* (4), 289–297. https://doi.org/10.1089/cbr.2012.1328.

(107) Heulot, M.; Chevalier, N.; Puyal, J.; Margue, C.; Michel, S.; Kreis, S.; Kulms, D.; Barras, D.; Nahimana, A.; Widmann, C. The TAT-RasGAP317-326 Anti-Cancer Peptide Can Kill in a Caspase-, Apoptosis-, and Necroptosis-Independent Manner. *Oncotarget* **2016**, *7* (39), 64342–64359. https://doi.org/10.18632/oncotarget.11841.

(108) Michod, D.; Yang, J.-Y.; Chen, J.; Bonny, C.; Widmann, C. A RasGAP-Derived Cell Permeable Peptide Potently Enhances Genotoxin-Induced Cytotoxicity in Tumor Cells. *Oncogene* **2004**, *23* (55), 8971–8978. https://doi.org/10.1038/sj.onc.1207999.

(109) Barras, D.; Lorusso, G.; Rüegg, C.; Widmann, C. Inhibition of Cell Migration and Invasion Mediated by the TAT-RasGAP317–326 Peptide Requires the DLC1 Tumor Suppressor. *Oncogene* **2014**, *33* (44), 5163–5172. https://doi.org/10.1038/onc.2013.465.

(110) Chevalier, N.; Gross, N.; Widmann, C. Assessment of the Chemosensitizing Activity of TAT-RasGAP317-326in Childhood Cancerse0120487. *PLoS ONE* **2015**, *10* (3), e0120487. https://doi.org/10.1371/journal.pone.0120487.

(111) Pittet, O.; Petermann, D.; Michod, D.; Krueger, T.; Cheng, C.; Ris, H.-B.; Widmann, C. Effect of the TAT-RasGAP317–326 Peptide on Apoptosis of Human Malignant Mesothelioma Cells and Fibroblasts Exposed to Meso-Tetra-Hydroxyphenyl-Chlorin and Light. *Journal of Photochemistry and Photobiology B: Biology* **2007**, *88* (1), 29–35. https://doi.org/10.1016/j.jphotobiol.2007.04.009.

(112) Heulot, M.; Jacquier, N.; Aeby, S.; Le Roy, D.; Roger, T.; Trofimenko, E.; Barras, D.; Greub, G.; Widmann, C. The Anticancer Peptide TAT-RasGAP317-326 Exerts Broad Antimicrobial Activity. *Frontiers in Microbiology* **2017**, *8* (JUN). https://doi.org/10.3389/fmicb.2017.00994.

(113) Heinonen, T.; Hargraves, S.; Georgieva, M.; Widmann, C.; Jacquier, N. The Antimicrobial Peptide TAT-RasGAP317-326 Inhibits the Formation and Expansion of Bacterial Biofilms in Vitro. *Journal of Global Antimicrobial Resistance* **2021**, *25*, 227–231. https://doi.org/10.1016/j.jgar.2021.03.022.

(114) Tsoutsou, P.; Annibaldi, A.; Viertl, D.; Ollivier, J.; Buchegger, F.; Vozenin, M. C.; Bourhis, J.; Widmann, C.; Matzinger, O. TAT-RasGAP317-326 Enhances Radiosensitivity of Human Carcinoma Cell Lines in Vitro and in Vivo through Promotion of Delayed Mitotic Cell Death. *Radiation Research* **2017**, *187* (5), 562–569. https://doi.org/10.1667/RR14509.1.

(115) Michod, D.; Annibaldi, A.; Schaefer, S.; Dapples, C.; Rochat, B.; Widmann, C. Effect of RasGAP N2 Fragment-Derived Peptide on Tumor Growth in Mice. *Journal of the National Cancer Institute* **2009**, *101* (11), 828–832. https://doi.org/10.1093/jnci/djp100.

(116) Barras, D.; Chevalier, N.; Zoete, V.; Dempsey, R.; Lapouge, K.; Olayioye, M. A.; Michielin, O.; Widmann, C. AWXWMotif Is Required for the Anticancer Activity of the TAT-RasGAP 317-326 Peptide. *Journal of Biological Chemistry* **2014**, *289* (34), 23701–23711. https://doi.org/10.1074/jbc.M114.576272.

(117) Trofimenko, E.; Grasso, G.; Heulot, M.; Chevalier, N.; Deriu, M. A.; Dubuis, G.; Arribat, Y.; Serulla, M.; Michel, S.; Vantomme, G.; Ory, F.; Dam, L. C.; Puyal, J.; Amati, F.; Lüthi, A.; Danani, A.; Widmann, C. Genetic, Cellular, and Structural Characterization of the Membrane Potential-Dependent Cell-Penetrating Peptide Translocation Pore. *eLife* **2021**, *10*, e69832. https://doi.org/10.7554/eLife.69832.

(118) Serulla, M.; Ichim, G.; Stojceski, F.; Grasso, G.; Afonin, S.; Heulot, M.; Schober, T.; Roth, R.; Godefroy, C.; Milhiet, P. E.; Das, K.; García-Sáez, A. J.; Danani, A.; Widmann, C. TAT-RasGAP317-326 Kills Cells by Targeting Inner-Leaflet-Enriched Phospholipids. *Proceedings of the National Academy of Sciences of the United States of America* **2020**, *117* (50), 31871–31881. https://doi.org/10.1073/pnas.2014108117.

(119) Georgieva, M.; Heinonen, T.; Vitale, A.; Hargraves, S.; Causevic, S.; Pillonel, T.; Eberl, L.; Widmann, C.; Jacquier, N. Bacterial Surface Properties Influence the Activity of the TAT-RasGAP317-326 Antimicrobial Peptide. *iScience* **2021**, *24* (8), 102923. https://doi.org/10.1016/j.isci.2021.102923.

(120) Georgieva, M.; Stojceski, F.; Wüthrich, F.; Sosthène, C.; Blanco Pérez, L.; Grasso, G.; Jacquier, N. Mutations in the Essential Outer Membrane Protein BamA Contribute to Escherichia Coli Resistance to the Antimicrobial Peptide TAT-RasGAP317-326. *Journal of Biological Chemistry* **2025**, *301* (1), 108018. https://doi.org/10.1016/j.jbc.2024.108018.

(121) Sun, D.; Storek, K. M.; Tegunov, D.; Yang, Y.; Arthur, C. P.; Johnson, M.; Quinn, J. G.; Liu, W.; Han, G.; Girgis, H. S.; Alexander, M. K.; Murchison, A. K.; Shriver, S.; Tam, C.; Ijiri, H.; Inaba, H.; Sano, T.; Yanagida, H.; Nishikawa, J.; Heise, C. E.; Fairbrother, W. J.; Tan, M.-W.; Skelton, N.; Sandoval, W.; Sellers, B. D.; Ciferri, C.; Smith, P. A.; Reid, P. C.; Cunningham, C. N.; Rutherford, S. T.; Payandeh, J. The Discovery and Structural Basis of Two Distinct State-Dependent Inhibitors of BamA. *Nat Commun* **2024**, *15* (1), 8718. https://doi.org/10.1038/s41467-024-52512-1.

(122) Raucher, D.; Ryu, J. S. Cell-Penetrating Peptides: Strategies for Anticancer Treatment. *Trends in Molecular Medicine* **2015**, *21* (9), 560–570. https://doi.org/10.1016/j.molmed.2015.06.005.

(123) Lindahl, E. R. Molecular Dynamics Simulations. In *Molecular Modeling of Proteins*; Kukol, A., Ed.; Walker, J. M., Series Ed.; Methods in Molecular Biology; Humana Press: Totowa, NJ, 2008; Vol. 443, pp 3–23. https://doi.org/10.1007/978-1-59745-177-2_1.

(124) Brooks, C. L.; Case, D. A.; Plimpton, S.; Roux, B.; Van Der Spoel, D.; Tajkhorshid, E. Classical Molecular Dynamics. *The Journal of Chemical Physics* **2021**, *154* (10), 100401. https://doi.org/10.1063/5.0045455.

(125) Car, R.; Parrinello, M. Unified Approach for Molecular Dynamics and Density-Functional Theory. *Phys. Rev. Lett.* **1985**, *55* (22), 2471–2474. https://doi.org/10.1103/PhysRevLett.55.2471.
(126) Johnson, J. K.; Zollweg, J. A.; Gubbins, K. E. The Lennard-Jones Equation of State Revisited. *Molecular Physics* **1993**, *78* (3), 591–618. https://doi.org/10.1080/00268979300100411.
(127) Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G. A Smooth Particle Mesh Ewald Method. *The Journal of Chemical Physics* **1995**, *103* (19), 8577–8593. https://doi.org/10.1063/1.470117.
(128) Gibbs, J. W. *Elementary Principles in Statistical Mechanics: Developed with Especial Reference to the Rational Foundation of Thermodynamics*, 1st ed.; Cambridge University Press, 2010. https://doi.org/10.1017/CBO9780511686948.
(129) Berendsen, H. J. C.; Postma, J. P. M.; Van Gunsteren, W. F.; DiNola, A.; Haak, J. R. Molecular Dynamics with Coupling to an External Bath. *The Journal of Chemical Physics* **1984**, *81* (8), 3684–3690. https://doi.org/10.1063/1.448118.
(130) Bussi, G.; Donadio, D.; Parrinello, M. Canonical Sampling through Velocity Rescaling. *The Journal of Chemical Physics* **2007**, *126* (1), 014101. https://doi.org/10.1063/1.2408420.
(131) Evans, D. J.; Holian, B. L. The Nose–Hoover Thermostat. *The Journal of Chemical Physics* **1985**, *83* (8), 4069–4074. https://doi.org/10.1063/1.449071.
(132) Andersen, H. C. Molecular Dynamics Simulations at Constant Pressure and/or Temperature. *The Journal of Chemical Physics* **1980**, *72* (4), 2384–2393. https://doi.org/10.1063/1.439486.
(133) Ke, Q.; Gong, X.; Liao, S.; Duan, C.; Li, L. Effects of Thermostats/Barostats on Physical Properties of Liquids by Molecular Dynamics Simulations. *Journal of Molecular Liquids* **2022**, *365*, 120116. https://doi.org/10.1016/j.molliq.2022.120116.
(134) Parrinello, M.; Rahman, A. Polymorphic Transitions in Single Crystals: A New Molecular Dynamics Method. *Journal of Applied Physics* **1981**, *52* (12), 7182–7190. https://doi.org/10.1063/1.328693.
(135) Bernetti, M.; Bussi, G. Pressure Control Using Stochastic Cell Rescaling. *The Journal of Chemical Physics* **2020**, *153* (11), 114107. https://doi.org/10.1063/5.0020514.
(136) Kr[illegible]tler, V.; Van Gunsteren, W. F.; H[illegible]enberger, P. H. A Fast SHAKE Algorithm to Solve Distance Constraint Equations for Small Molecules in Molecular Dynamics Simulations. *J. Comput. Chem.* **2001**, *22* (5), 501–508. https://doi.org/10.1002/1096-987X(20010415)22:5<501::AID-JCC1021>3.0.CO;2-V.
(137) Hess, B.; Bekker, H.; Berendsen, H. J. C.; Fraaije, J. G. E. M. LINCS: A Linear Constraint Solver for Molecular Simulations. *J. Comput. Chem.* **1997**, *18* (12), 1463–1472. https://doi.org/10.1002/(SICI)1096-987X(199709)18:12<1463::AID-JCC4>3.0.CO;2-H.
(138) Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.-Y. Lightgbm: A Highly Efficient Gradient Boosting Decision Tree. *Advances in neural information processing systems* **2017**, *30*.
(139) Breiman, L. Random Forests. *Machine Learning* **2001**, *45* (1), 5–32. https://doi.org/10.1023/A:1010933404324.
(140) Molnar, C. *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*; Leanpub: Victoria, British Columbia, 2020.
(141) Breunig, M. M.; Kriegel, H.-P.; Ng, R. T.; Sander, J. LOF: Identifying Density-Based Local Outliers. *SIGMOD Rec.* **2000**, *29* (2), 93–104. https://doi.org/10.1145/335191.335388.
(142) Kramer, O. Genetic Algorithms. In *Genetic Algorithm Essentials*; Studies in Computational Intelligence; Springer International Publishing: Cham, 2017; Vol. 679, pp 11–19. https://doi.org/10.1007/978-3-319-52156-5_2.

(143) Katoch, S.; Chauhan, S. S.; Kumar, V. A Review on Genetic Algorithm: Past, Present, and Future. *Multimed Tools Appl* **2021**, *80* (5), 8091–8126. https://doi.org/10.1007/s11042-020-10139-6.

(144) Manavalan, B.; Subramaniyam, S.; Shin, T. H.; Kim, M. O.; Lee, G. Machine-Learning-Based Prediction of Cell-Penetrating Peptides and Their Uptake Efficiency with Improved Accuracy. *J. Proteome Res.* **2018**, *17* (8), 2715–2726. https://doi.org/10.1021/acs.jproteome.8b00148.

(145) Pandey, P.; Patel, V.; George, N. V.; Mallajosyula, S. S. KELM-CPPpred: Kernel Extreme Learning Machine Based Prediction Model for Cell-Penetrating Peptides. *J. Proteome Res.* **2018**, *17* (9), 3214–3222. https://doi.org/10.1021/acs.jproteome.8b00322.

(146) Damiati, S. A.; Alaofi, A. L.; Dhar, P.; Alhakamy, N. A. Novel Machine Learning Application for Prediction of Membrane Insertion Potential of Cell-Penetrating Peptides. *International Journal of Pharmaceutics* **2019**, *567*, 118453. https://doi.org/10.1016/j.ijpharm.2019.118453.

(147) Wei, L.; Xing, P.; Su, R.; Shi, G.; Ma, Z. S.; Zou, Q. CPPred-RF: A Sequence-Based Predictor for Identifying Cell-Penetrating Peptides and Their Uptake Efficiency. *J. Proteome Res.* **2017**, *16* (5), 2044–2053. https://doi.org/10.1021/acs.jproteome.7b00019.

(148) Tang, H.; Su, Z.-D.; Wei, H.-H.; Chen, W.; Lin, H. Prediction of Cell-Penetrating Peptides with Feature Selection Techniques. *Biochemical and Biophysical Research Communications* **2016**, *477* (1), 150–154. https://doi.org/10.1016/j.bbrc.2016.06.035.

(149) Fu, X.; Cai, L.; Zeng, X.; Zou, Q. StackCPPred: A Stacking and Pairwise Energy Content-Based Prediction of Cell-Penetrating Peptides and Their Uptake Efficiency. *Bioinformatics* **2020**, *36* (10), 3028–3034. https://doi.org/10.1093/bioinformatics/btaa131.

(150) Kumar, V.; Raghava, G. P. S. In Silico Design of Chemically Modified Cell-Penetrating Peptides. In *Cell Penetrating Peptides*; Langel, Ü., Ed.; Methods in Molecular Biology; Springer US: New York, NY, 2022; Vol. 2383, pp 63–71. https://doi.org/10.1007/978-1-0716-1752-6_4.

(151) Shi, K.; Xiong, Y.; Wang, Y.; Deng, Y.; Wang, W.; Jing, B.; Gao, X. PractiCPP: A Deep Learning Approach Tailored for Extremely Imbalanced Datasets in Cell-Penetrating Peptide Prediction. *Bioinformatics* **2024**, *40* (2), btae058. https://doi.org/10.1093/bioinformatics/btae058.

(152) Qiang, X.; Zhou, C.; Ye, X.; Du, P.; Su, R.; Wei, L. CPPred-FL: A Sequence-Based Predictor for Large-Scale Identification of Cell-Penetrating Peptides by Feature Representation Learning. *Briefings in Bioinformatics* **2018**. https://doi.org/10.1093/bib/bby091.

(153) Fu, X.; Ke, L.; Cai, L.; Chen, X.; Ren, X.; Gao, M. Improved Prediction of Cell-Penetrating Peptides via Effective Orchestrating Amino Acid Composition Feature Representation. *IEEE Access* **2019**, *7*, 163547–163555. https://doi.org/10.1109/ACCESS.2019.2952738.

(154) Kumar, V.; Agrawal, P.; Kumar, R.; Bhalla, S.; Usmani, S. S.; Varshney, G. C.; Raghava, G. P. S. Prediction of Cell-Penetrating Potential of Modified Peptides Containing Natural and Chemically Modified Residues. *Front. Microbiol.* **2018**, *9*, 725. https://doi.org/10.3389/fmicb.2018.00725.

(155) de Oliveira, E. C. L.; Santana, K.; Josino, L.; Lima e Lima, A. H.; de Souza de Sales Júnior, C. Predicting Cell-Penetrating Peptides Using Machine Learning Algorithms and Navigating in Their Chemical Space. *Sci Rep* **2021**, *11* (1), 7628. https://doi.org/10.1038/s41598-021-87134-w.

(156) Manavalan, B.; Patra, M. C. MLCPP 2.0: An Updated Cell-Penetrating Peptides and Their Uptake Efficiency Predictor. *Journal of Molecular Biology* **2022**, *434* (11), 167604. https://doi.org/10.1016/j.jmb.2022.167604.

(157) Zhang, X.; Wei, L.; Ye, X.; Zhang, K.; Teng, S.; Li, Z.; Jin, J.; Kim, M. J.; Sakurai, T.; Cui, L.; Manavalan, B.; Wei, L. SiameseCPP: A Sequence-Based Siamese Network to Predict Cell-

Penetrating Peptides by Contrastive Learning. *Briefings in Bioinformatics* **2023**, *24* (1), bbac545. https://doi.org/10.1093/bib/bbac545.

(158) Gautam, A.; Chaudhary, K.; Kumar, R.; Raghava, G. P. S. Computer-Aided Virtual Screening and Designing of Cell-Penetrating Peptides. In *Cell-Penetrating Peptides*; Langel, Ü., Ed.; Methods in Molecular Biology; Springer New York: New York, NY, 2015; Vol. 1324, pp 59–69. https://doi.org/10.1007/978-1-4939-2806-4_4.

(159) Fu, L.; Niu, B.; Zhu, Z.; Wu, S.; Li, W. CD-HIT: Accelerated for Clustering the next-Generation Sequencing Data. *Bioinformatics* **2012**, *28* (23), 3150–3152. https://doi.org/10.1093/bioinformatics/bts565.

(160) Agrawal, P.; Bhalla, S.; Usmani, S. S.; Singh, S.; Chaudhary, K.; Raghava, G. P. S.; Gautam, A. CPPsite 2.0: A Repository of Experimentally Validated Cell-Penetrating Peptides. *Nucleic Acids Res* **2016**, *44* (D1), D1098–D1103. https://doi.org/10.1093/nar/gkv1266.

(161) Basith, S.; Manavalan, B.; Hwan Shin, T.; Lee, G. Machine Intelligence in Peptide Therapeutics: A Next-generation Tool for Rapid Disease Screening. *Medicinal Research Reviews* **2020**, *40* (4), 1276–1314. https://doi.org/10.1002/med.21658.

(162) Mienye, I. D.; Sun, Y. Performance Analysis of Cost-Sensitive Learning Methods with Application to Imbalanced Medical Data. *Informatics in Medicine Unlocked* **2021**, *25*, 100690. https://doi.org/10.1016/j.imu.2021.100690.

(163) Gautam, A.; Chaudhary, K.; Kumar, R.; Sharma, A.; Kapoor, P.; Tyagi, A.; Open source drug discovery consortium; Raghava, G. P. S. In Silico Approaches for Designing Highly Effective Cell Penetrating Peptides. *J Transl Med* **2013**, *11* (1), 74. https://doi.org/10.1186/1479-5876-11-74.

(164) Greg Landrum; Paolo Tosco; Brian Kelley; Ric; David Cosgrove; sriniker; gedeck; Riccardo Vianello; NadineSchneider; Eisuke Kawashima; Gareth Jones; Dan N; Andrew Dalke; Brian Cole; Matt Swain; Samo Turk; AlexanderSavelyev; Alain Vaucher; Maciej Wójcikowski; Ichiru Take; Vincent F. Scalfani; Daniel Probst; Kazuya Ujihara; guillaume godin; Axel Pahl; Rachel Walker; Juuso Lehtivarjo; Francois Berenger; jasondbiggs; strets123. Rdkit/Rdkit: 2023_09_4 (Q3 2023) Release, 2024. https://doi.org/10.5281/ZENODO.591637.

(165) Cock, P. J. A.; Antao, T.; Chang, J. T.; Chapman, B. A.; Cox, C. J.; Dalke, A.; Friedberg, I.; Hamelryck, T.; Kauff, F.; Wilczynski, B.; De Hoon, M. J. L. Biopython: Freely Available Python Tools for Computational Molecular Biology and Bioinformatics. *Bioinformatics* **2009**, *25* (11), 1422–1423. https://doi.org/10.1093/bioinformatics/btp163.

(166) Chen, Z.; Liu, X.; Zhao, P.; Li, C.; Wang, Y.; Li, F.; Akutsu, T.; Bain, C.; Gasser, R. B.; Li, J.; Yang, Z.; Gao, X.; Kurgan, L.; Song, J. *iFeatureOmega:* An Integrative Platform for Engineering, Visualization and Analysis of Features from Molecular Sequences, Structural and Ligand Data Sets. *Nucleic Acids Research* **2022**, *50* (W1), W434–W447. https://doi.org/10.1093/nar/gkac351.

(167) Chou, K.-C. Using Amphiphilic Pseudo Amino Acid Composition to Predict Enzyme Subfamily Classes. *Bioinformatics* **2005**, *21* (1), 10–19. https://doi.org/10.1093/bioinformatics/bth466.

(168) Argos, P.; Rao, J. K. M.; Hargrave, P. A. Structural Prediction of Membrane-Bound Proteins. *European Journal of Biochemistry* **1982**, *128* (2–3), 565–575. https://doi.org/10.1111/j.1432-1033.1982.tb07002.x.

(169) *Prediction of Protein Structure and the Principles of Protein Conformation*; Fasman, G. D., Ed.; Plenum Press: New York, 1989.

(170) Saravanan, V.; Gautham, N. Harnessing Computational Biology for Exact Linear B-Cell Epitope Prediction: A Novel Amino Acid Composition-Based Feature Descriptor. *OMICS: A Journal of Integrative Biology* **2015**, *19* (10), 648–658. https://doi.org/10.1089/omi.2015.0095.

(171) Wei, L.; Zhou, C.; Chen, H.; Song, J.; Su, R. ACPred-FL: A Sequence-Based Predictor Using Effective Feature Representation to Improve the Prediction of Anti-Cancer Peptides. *Bioinformatics* **2018**, *34* (23), 4007–4016. https://doi.org/10.1093/bioinformatics/bty451.
(172) Grantham, R. Amino Acid Difference Formula to Help Explain Protein Evolution. *Science* **1974**, *185* (4154), 862–864. https://doi.org/10.1126/science.185.4154.862.
(173) Sandberg, M.; Eriksson, L.; Jonsson, J.; Sjöström, M.; Wold, S. New Chemical Descriptors Relevant for the Design of Biologically Active Peptides. A Multivariate Characterization of 87 Amino Acids. *J. Med. Chem.* **1998**, *41* (14), 2481–2491. https://doi.org/10.1021/jm9700575.
(174) Henikoff, S.; Henikoff, J. G. Amino Acid Substitution Matrices from Protein Blocks. *Proc. Natl. Acad. Sci. U.S.A.* **1992**, *89* (22), 10915–10919. https://doi.org/10.1073/pnas.89.22.10915.
(175) Tran, D. P.; Tada, S.; Yumoto, A.; Kitao, A.; Ito, Y.; Uzawa, T.; Tsuda, K. Using Molecular Dynamics Simulations to Prioritize and Understand AI-Generated Cell Penetrating Peptides. *Sci Rep* **2021**, *11* (1), 10630. https://doi.org/10.1038/s41598-021-90245-z.
(176) Ma, H.; Zhou, X.; Zhang, Z.; Weng, Z.; Li, G.; Zhou, Y.; Yao, Y. AI-Driven Design of Cell-Penetrating Peptides for Therapeutic Biotechnology. *Int J Pept Res Ther* **2024**, *30* (6), 69. https://doi.org/10.1007/s10989-024-10654-2.
(177) Yoshida, M.; Hinkley, T.; Tsuda, S.; Abul-Haija, Y. M.; McBurney, R. T.; Kulikov, V.; Mathieson, J. S.; Galiñanes Reyes, S.; Castro, M. D.; Cronin, L. Using Evolutionary Algorithms and Machine Learning to Explore Sequence Space for the Discovery of Antimicrobial Peptides. *Chem* **2018**, *4* (3), 533–543. https://doi.org/10.1016/j.chempr.2018.01.005.
(178) Boone, K.; Wisdom, C.; Camarda, K.; Spencer, P.; Tamerler, C. Combining Genetic Algorithm with Machine Learning Strategies for Designing Potent Antimicrobial Peptides. *BMC Bioinformatics* **2021**, *22* (1), 239. https://doi.org/10.1186/s12859-021-04156-x.
(179) Wender, P. A.; Mitchell, D. J.; Pattabiraman, K.; Pelkey, E. T.; Steinman, L.; Rothbard, J. B. The Design, Synthesis, and Evaluation of Molecules That Enable or Enhance Cellular Uptake: Peptoid Molecular Transporters. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97* (24), 13003–13008. https://doi.org/10.1073/pnas.97.24.13003.
(180) Koren, E.; Torchilin, V. P. Cell-Penetrating Peptides: Breaking through to the Other Side. *Trends in Molecular Medicine* **2012**, *18* (7), 385–393. https://doi.org/10.1016/j.molmed.2012.04.012.
(181) Nagel, Y. A.; Raschle, P. S.; Wennemers, H. Effect of Preorganized Charge-Display on the Cell-Penetrating Properties of Cationic Peptides. *Angewandte Chemie* **2017**, *129* (1), 128–132. https://doi.org/10.1002/ange.201607649.
(182) Schmidt, N.; Mishra, A.; Lai, G. H.; Wong, G. C. L. Arginine-rich Cell-penetrating Peptides. *FEBS Letters* **2010**, *584* (9), 1806–1813. https://doi.org/10.1016/j.febslet.2009.11.046.
(183) Di Pisa, M.; Chassaing, G.; Swiecicki, J.-M. Translocation Mechanism(s) of Cell-Penetrating Peptides: Biophysical Studies Using Artificial Membrane Bilayers. *Biochemistry* **2015**, *54* (2), 194–207. https://doi.org/10.1021/bi501392n.
(184) Gao, X.; Hong, S.; Liu, Z.; Yue, T.; Dobnikar, J.; Zhang, X. Membrane Potential Drives Direct Translocation of Cell-Penetrating Peptides. *Nanoscale* **2019**, *11* (4), 1949–1958. https://doi.org/10.1039/C8NR10447F.
(185) Khalil, I. A.; Kogure, K.; Futaki, S.; Harashima, H. High Density of Octaarginine Stimulates Macropinocytosis Leading to Efficient Intracellular Trafficking for Gene Expression. *Journal of Biological Chemistry* **2006**, *281* (6), 3544–3551. https://doi.org/10.1074/jbc.M503202200.
(186) Gestin, M.; Dowaidar, M.; Langel, Ü. Uptake Mechanism of Cell-Penetrating Peptides. In *Peptides and Peptide-based Biomaterials and their Biomedical Applications*; Sunna, A., Care, A., Bergquist, P. L., Eds.; Advances in Experimental Medicine and Biology; Springer

International Publishing: Cham, 2017; Vol. 1030, pp 255–264. https://doi.org/10.1007/978-3-319-66095-0_11.

(187) Chen, L.; Zhang, Q.; Yuan, X.; Cao, Y.; Yuan, Y.; Yin, H.; Ding, X.; Zhu, Z.; Luo, S.-Z. How Charge Distribution Influences the Function of Membrane-Active Peptides: Lytic or Cell-Penetrating? *The International Journal of Biochemistry & Cell Biology* **2017**, *83*, 71–75. https://doi.org/10.1016/j.biocel.2016.12.011.

(188) Allen, J.; Pellois, J.-P. Hydrophobicity Is a Key Determinant in the Activity of Arginine-Rich Cell Penetrating Peptides. *Sci Rep* **2022**, *12* (1), 15981. https://doi.org/10.1038/s41598-022-20425-y.

(189) Oba, M.; Nakajima, S.; Misao, K.; Yokoo, H.; Tanaka, M. Effect of Helicity and Hydrophobicity on Cell-Penetrating Ability of Arginine-Rich Peptides. *Bioorganic & Medicinal Chemistry* **2023**, *91*, 117409. https://doi.org/10.1016/j.bmc.2023.117409.

(190) Jha, D.; Mishra, R.; Gottschalk, S.; Wiesmüller, K.-H.; Ugurbil, K.; Maier, M. E.; Engelmann, J. CyLoP-1: A Novel Cysteine-Rich Cell-Penetrating Peptide for Cytosolic Delivery of Cargoes. *Bioconjugate Chem.* **2011**, *22* (3), 319–328. https://doi.org/10.1021/bc100045s.

(191) Ponnappan, N.; Budagavi, D. P.; Chugh, A. CyLoP-1: Membrane-Active Peptide with Cell-Penetrating and Antimicrobial Properties. *Biochimica et Biophysica Acta (BBA) - Biomembranes* **2017**, *1859* (2), 167–176. https://doi.org/10.1016/j.bbamem.2016.11.002.

(192) Kim, W.-J.; Kim, G.-R.; Cho, H.-J.; Choi, J.-M. The Cysteine-Containing Cell-Penetrating Peptide AP Enables Efficient Macromolecule Delivery to T Cells and Controls Autoimmune Encephalomyelitis. *Pharmaceutics* **2021**, *13* (8), 1134. https://doi.org/10.3390/pharmaceutics13081134.

(193) Peng, F.; Liao, M.; Qin, R.; Zhu, S.; Peng, C.; Fu, L.; Chen, Y.; Han, B. Regulated Cell Death (RCD) in Cancer: Key Pathways and Targeted Therapies. *Sig Transduct Target Ther* **2022**, *7* (1), 286. https://doi.org/10.1038/s41392-022-01110-y.

(194) Ulmschneider, J. P.; Ulmschneider, M. B. Molecular Dynamics Simulations Are Redefining Our View of Peptides Interacting with Biological Membranes. *Acc. Chem. Res.* **2018**, *51* (5), 1106–1116. https://doi.org/10.1021/acs.accounts.7b00613.

(195) Kabelka, I.; Vácha, R. Advances in Molecular Understanding of α-Helical Membrane-Active Peptides. *Acc. Chem. Res.* **2021**, *54* (9), 2196–2204. https://doi.org/10.1021/acs.accounts.1c00047.

(196) Lamiable, A.; Thévenet, P.; Rey, J.; Vavrusa, M.; Derreumaux, P.; Tufféry, P. PEP-FOLD3: Faster *de Novo* Structure Prediction for Linear Peptides in Solution and in Complex. *Nucleic Acids Res* **2016**, *44* (W1), W449–W454. https://doi.org/10.1093/nar/gkw329.

(197) Garton, M.; Nim, S.; Stone, T. A.; Wang, K. E.; Deber, C. M.; Kim, P. M. Method to Generate Highly Stable D-Amino Acid Analogs of Bioactive Helical Peptides Using a Mirror Image of the Entire PDB. *Proc. Natl. Acad. Sci. U.S.A.* **2018**, *115* (7), 1505–1510. https://doi.org/10.1073/pnas.1711837115.

(198) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of Simple Potential Functions for Simulating Liquid Water. *The Journal of Chemical Physics* **1983**, *79* (2), 926–935. https://doi.org/10.1063/1.445869.

(199) Jo, S.; Kim, T.; Iyer, V. G.; Im, W. CHARMM-GUI: A Web-Based Graphical User Interface for CHARMM. *J. Comput. Chem.* **2008**, *29* (11), 1859–1865. https://doi.org/10.1002/jcc.20945.

(200) Jo, S.; Kim, T.; Im, W. Automated Builder and Database of Protein/Membrane Complexes for Molecular Dynamics Simulations. *PLoS ONE* **2007**, *2* (9), e880. https://doi.org/10.1371/journal.pone.0000880.

(201) Lee, J.; Cheng, X.; Swails, J. M.; Yeom, M. S.; Eastman, P. K.; Lemkul, J. A.; Wei, S.; Buckner, J.; Jeong, J. C.; Qi, Y.; Jo, S.; Pande, V. S.; Case, D. A.; Brooks, C. L.; MacKerell, A. D.; Klauda, J. B.; Im, W. CHARMM-GUI Input Generator for NAMD, GROMACS, AMBER,

OpenMM, and CHARMM/OpenMM Simulations Using the CHARMM36 Additive Force Field. *J. Chem. Theory Comput.* **2016**, *12* (1), 405–413. https://doi.org/10.1021/acs.jctc.5b00935.

(202) Huang, J.; Rauscher, S.; Nawrocki, G.; Ran, T.; Feig, M.; De Groot, B. L.; Grubmüller, H.; MacKerell, A. D. CHARMM36m: An Improved Force Field for Folded and Intrinsically Disordered Proteins. *Nat Methods* **2017**, *14* (1), 71–73. https://doi.org/10.1038/nmeth.4067.

(203) Humphrey, W.; Dalke, A.; Schulten, K. VMD: Visual Molecular Dynamics. *Journal of Molecular Graphics* **1996**, *14* (1), 33–38. https://doi.org/10.1016/0263-7855(96)00018-5.

(204) Abraham, M. J.; Murtola, T.; Schulz, R.; Páll, S.; Smith, J. C.; Hess, B.; Lindahl, E. GROMACS: High Performance Molecular Simulations through Multi-Level Parallelism from Laptops to Supercomputers. *SoftwareX* **2015**, *1–2*, 19–25. https://doi.org/10.1016/j.softx.2015.06.001.

(205) Kutzner, C.; Páll, S.; Fechner, M.; Esztermann, A.; Groot, B. L.; Grubmüller, H. More Bang for Your Buck: Improved Use of GPU Nodes for GROMACS 2018. *J Comput Chem* **2019**, *40* (27), 2418–2431. https://doi.org/10.1002/jcc.26011.

(206) Baxter, A. A.; Poon, I. K.; Hulett, M. D. The Plant Defensin NaD1 Induces Tumor Cell Death via a Non-Apoptotic, Membranolytic Process. *Cell Death Discov.* **2017**, *3* (1), 16102. https://doi.org/10.1038/cddiscovery.2016.102.

(207) Baxter, A. A.; Richter, V.; Lay, F. T.; Poon, I. K. H.; Adda, C. G.; Veneer, P. K.; Phan, T. K.; Bleackley, M. R.; Anderson, M. A.; Kvansakul, M.; Hulett, M. D. The Tomato Defensin TPP3 Binds Phosphatidylinositol (4,5)-Bisphosphate via a Conserved Dimeric Cationic Grip Conformation To Mediate Cell Lysis. *Molecular and Cellular Biology* **2015**, *35* (11), 1964–1978. https://doi.org/10.1128/MCB.00282-15.

(208) Poon, I. K.; Baxter, A. A.; Lay, F. T.; Mills, G. D.; Adda, C. G.; Payne, J. A.; Phan, T. K.; Ryan, G. F.; White, J. A.; Veneer, P. K.; Van Der Weerden, N. L.; Anderson, M. A.; Kvansakul, M.; Hulett, M. D. Phosphoinositide-Mediated Oligomerization of a Defensin Induces Cell Lysis. *eLife* **2014**, *3*, e01808. https://doi.org/10.7554/eLife.01808.

(209) Chen, L.; Gan, L.; Liu, M.; Fan, R.; Xu, Z.; Hao, Z.; Chen, L. Destabilization of Artificial Biomembrane Induced by the Penetration of Tryptophan. *Applied Surface Science* **2011**, *257* (11), 5070–5076. https://doi.org/10.1016/j.apsusc.2011.01.023.

(210) Situ, A. J.; Kang, S.-M.; Frey, B. B.; An, W.; Kim, C.; Ulmer, T. S. Membrane Anchoring of α-Helical Proteins: Role of Tryptophan. *The Journal of Physical Chemistry B* **2018**, *122* (3), 1185–1194. https://doi.org/10.1021/acs.jpcb.7b11227.

(211) Vizzarro, G.; Jacquier, N. In Vitro Synergistic Action of TAT-RasGAP317-326 Peptide with Antibiotics against Gram-Negative Pathogens. *Journal of Global Antimicrobial Resistance* **2022**, *31*, 295–303. https://doi.org/10.1016/j.jgar.2022.10.003.

(212) Gu, Y.; Li, H.; Dong, H.; Zeng, Y.; Zhang, Z.; Paterson, N. G.; Stansfeld, P. J.; Wang, Z.; Zhang, Y.; Wang, W.; Dong, C. Structural Basis of Outer Membrane Protein Insertion by the BAM Complex. *Nature* **2016**, *531* (7592), 64–69. https://doi.org/10.1038/nature17199.

(213) De Vries, S. J.; Van Dijk, M.; Bonvin, A. M. J. J. The HADDOCK Web Server for Data-Driven Biomolecular Docking. *Nat Protoc* **2010**, *5* (5), 883–897. https://doi.org/10.1038/nprot.2010.32.

(214) Noinaj, N.; Kuszak, A. J.; Balusek, C.; Gumbart, J. C.; Buchanan, S. K. Lateral Opening and Exit Pore Formation Are Required for BamA Function. *Structure* **2014**, *22* (7), 1055–1062. https://doi.org/10.1016/j.str.2014.05.008.

(215) Wu, R.; Bakelar, J. W.; Lundquist, K.; Zhang, Z.; Kuo, K. M.; Ryoo, D.; Pang, Y. T.; Sun, C.; White, T.; Klose, T.; Jiang, W.; Gumbart, J. C.; Noinaj, N. Plasticity within the Barrel Domain of BamA Mediates a Hybrid-Barrel Mechanism by BAM. *Nat Commun* **2021**, *12* (1), 7131. https://doi.org/10.1038/s41467-021-27449-4.

(216) Tomasek, D.; Kahne, D. The Assembly of β-Barrel Outer Membrane Proteins. *Current Opinion in Microbiology* **2021**, *60*, 16–23. https://doi.org/10.1016/j.mib.2021.01.009.
(217) Haysom, S. F.; Machin, J.; Whitehouse, J. M.; Horne, J. E.; Fenn, K.; Ma, Y.; El Mkami, H.; Böhringer, N.; Schäberle, T. F.; Ranson, N. A.; Radford, S. E.; Pliotas, C. Darobactin B Stabilises a Lateral-Closed Conformation of the BAM Complex in *E. Coli* Cells. *Angew Chem Int Ed* **2023**, *62* (34), e202218783. https://doi.org/10.1002/anie.202218783.