.

Predicting the affinity of compstatin peptide with non-natural amino acids to human C3c protein by scoring molecular dynamics simulations

by

Kelly Yohana Muñoz Gomez

A dissertation submitted in partial fulfillment
of the requirements for the degree of

Master of Physics

Advisor: Pilar Cossio Tejada, Ph.D.
co-advisors: Johans Restrepo Cárdenas, Ph.D,
Rodrigo Ochoa Deossa, Ph.D

Biophysics of Tropical Diseases - Max Planck Tandem Group
Faculty of Exact and Natural Sciences
October, 2022

Abstract

Peptides are chemical entities composed of natural and non-natural amino acids that have been used successfully as drugs, vaccines, biomarkers, among others. However, these can be easily cleaved and degraded by proteases, where their breaking of a chemical bond in peptides gives smaller molecules or radicals, causing instability in some biological environments when we use peptides therapeutically or as medicines. One possible solution is the use of peptides with non-natural amino acids (NNAA). In the present study, we assessed the prediction of affinities in complexes between human Complement component 3 (C3c) protein bound to multiple compstatin peptide analogs with NNAAs. We used molecular dynamics simulations and six scoring functions to correlate the average score with the experimental binding data obtained from previous studies. Several correlation coefficients above 0.7 and one above 0.85 were detected, indicating an excellent correlation between these two variables. We found the highest Spearman correlation for the Nnscore and Cyscore scoring function, suggesting that these are the most adequate for ranking the binding of modified peptides to a protein target.

Contents

List of figures

List of Tables

List of Abbreviations

3D Three-dimensional

C3c Complement component 3 protein

IC_{50} Experimental half maximum inhibitory concentration

MD Molecular dynamics

NNAA non-natural amino acids

PDB Protein Data Bank

SF Scoring function

1 Introduction

Peptides are biomolecules with chemical and physical properties associated with their amino acid composition. They have dominant roles in multiple molecular recognition and signalling events, especially in living systems [1]. They are also used as drugs or diagnostics tools for several biomedical applications [2]. However, working with peptides has a limitation: they can be cleaved and easily degraded. A potential solution to this problem is the design of peptides using non-natural amino acids (NNAA). These modified-peptides can mimic the mechanism of a peptide while being resistant to enzymatic degradation and displaying a significant activity, for example against several pathogens [3, 4, 5]. Moreover, incorporating NNAA can enhance the affinity [6, 7], selectivity [8, 9, 10], and stability of drug leads [11], as well as expand their applications in different fields in biochemistry [12, 13, 14, 15] and protein engineering [16, 17, 18, 19, 20, 21].

Databases with information of NNAAs are available, in particular, the SwissSideChain database which provides biophysical, structural, and molecular data for hundreds of commercially available non-natural amino acid side chains, both in L- and D-configurations [22]. Moreover, there is a force field called Forcefield NCAA, designed for the purpose of discovery and therapeutic design of proteins and peptides with non-natural amino acids. This force field has information of multiple non-natural amino acids that can be used in applications such as protein structure prediction and *de novo* protein design [23]. On the other hand, a tool called PEPstrMOD is available for the prediction of the tertiary structure of peptides including NNAAs. This tool is useful for performing mutations of natural amino acids to non-natural ones, and it enables obtaining the topology of these mutated peptides to be included in molecular dynam-

ics simulations [24].

For the purpose of drug screening, one would like to rank different modified peptides according to their binding affinity to a protein target. Experiments are costly and time consuming and computational methods offer a solution to estimate an initial rank. To evaluate the potential binding of the modified peptides, receptor-ligand complexes have to be built, and their binding has to be evaluated. For this purpose, molecular dynamics (MD) simulations enable sampling of the bound conformations. The obtained conformations are evaluated with scoring functions, which are mathematical functions used to roughly predict the binding affinity between two molecules. This method was used to classify natural peptide binders to the major histocompatibility class II receptor [25], which was applied for the design of multi-allele binding peptides [26]. However, few studies are available on complexes containing NNAAs since most MD simulation force fields are not parameterized for novel chemical entities [27, 28, 29, 30, 31, 32].

This is why, in the present work, we studied the human complement component 3 (C3c) protein in complex with the compstatin peptide and with different analogs of the peptide containing NNAAs. This is possible due to the availability of C3c structures in complex with different compstatin peptide analogs. The main goal of this work is to predict the rank of the experimental binding affinity of the modified-peptides to the C3c protein. We considered a set of nine peptides with known binding affinity to the target. We started from a C3c-compstatin crystal structure, and the peptide analogs were modelled by modifying single-positions by the reported NNAAs. We generated parameters for each NNAA to subsequently submit all the protein-peptide complexes to MD simulations. The obtained conformations were then scored using 6 different protein-ligand scoring functions. Finally, with the affinity scores for the complexes obtained from the scoring functions, we correlated with the experimental data using Spearman's correlation. A schematic of the workflow used is shown in the Figure 1.1.

1.1 Hypothesis and objectives

1.1.1 Hypothesis

NNAA can be parameterized using knowledge of amino acidic units, and the peptide bonds. These can then be used in protein-peptide complexes with NNAA to predict the affinity of the complexes with molecular dynamics sampling and binding affinity calculations.

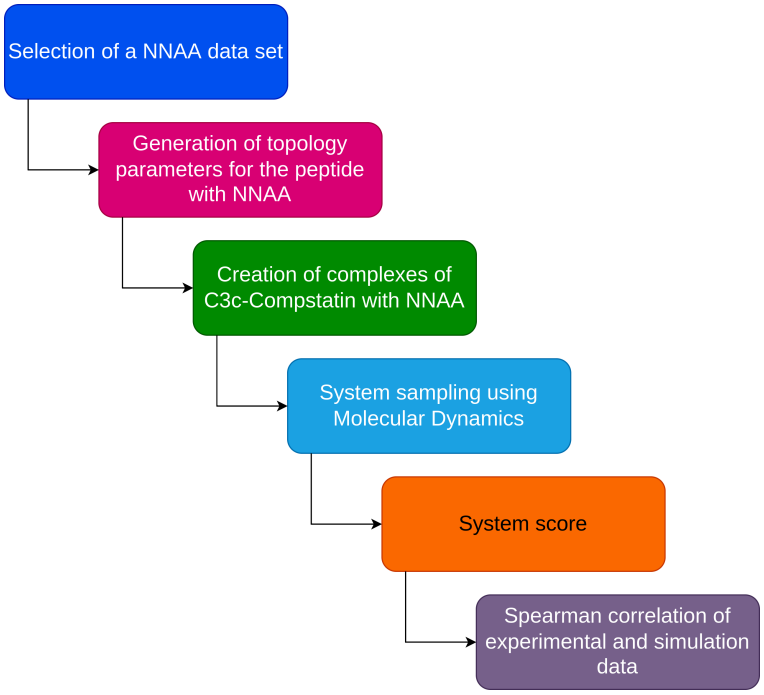

Figure 1.1: Schematic representation of the workflow used. First, we select a NNAA dataset. Second, generation of topology parameters for different peptides with NNAA from the selected data set. Third, the creation of complexes of C3c with the compstatin peptide with the selected NNAA. Fourth, the sampling of the system using molecular dynamics. Fifth, the punctuation of the trajectories obtained with the scoring functions. Finally, the Spearman correlation of experimental and simulation data.

1.1.2 Objectives

General objective

Predict the affinity of modified peptides with single NNAAs to the human complement component 3 (C3c) protein by scoring conformations from molecular dynamics simulations.

Specific objectives

- Parameterize non-natural amino acids found in the compstatin peptide analogs.
- Simulate the bound complexes using molecular dynamics.
- Score the affinities of the protein-peptide complexes to correlate the available experimental affinities.

1.2 Document outline

This document has the following organization. First, in chapter 1, we have the introduction and the objectives of the research. Secondly, in chapter 2, we introduce the theoretical concepts. Then, in chapter 3, we present the methods and results of the molecular dynamics and scoring of compstatin peptide with NNAAs bound to human C3c protein. Finally, in chapters 4 and 5, we present the conclusions and perspectives, respectively.

2
Theoretical Background

Some key concepts for understanding this research are introduced below. We present basic concepts of amino acids and peptides, followed by the introduction to molecular dynamics simulations, force fields, scoring functions, and Spearman's rank correlation.

2.1 Natural and non-natural amino acids (NNAA)

Amino acids are simple organic chemical entities having a common set of atoms that form the amino acid backbone, including a carboxyl and amino groups [33]. Attached to the central carbon atom (the alpha carbon), there are additional atoms that varies among the amino acids, making them different in terms of their properties. This is called the R group or amino acid side chain [34], figure 2.1.

Amino acids are the basic building blocks to form peptides and proteins [35]. In nature, we find a finite number of side chains (20) that conform to the natural amino acids. However, if non-natural side chains are used, it is possible to have more chemical diversity and broaden applications in both peptides and proteins in the field of biochemistry [15], protein engineering, and drug design [21]. For this purpose, non-natural side chains are increasingly used in experimental studies [22].

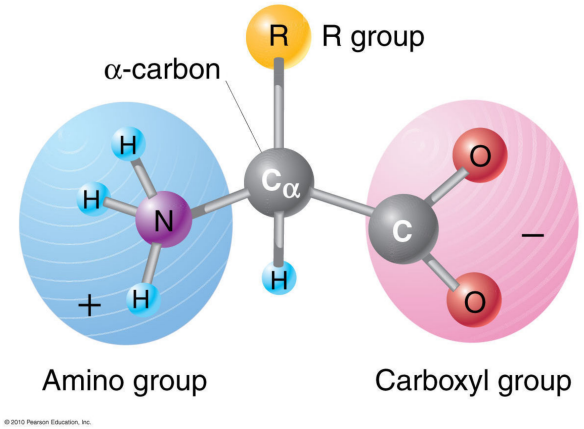

Figure 2.1: Amino acid structure [36].

2.2 Peptides

A peptide is a short chain of natural amino acids, figure 2.2. The amino acids are connected with each other through peptide bonds. Typically, peptides are distinguished from proteins by their shorter length, although the cut-off number of amino acids to define a protein can vary (normally around 100 amino acids). Peptides are generally considered to be short strings of 2 to 50 amino acids [37, 38].

Peptides have been used as alkaloids [39], antimicrobial agents [40] or hormones [41]. They can act as growth factors [42], anti-oxidant agents [43], or can be used for clinical diagnosis [44]. Peptides can undergo modifications in some parts of their sequences, in which a natural amino acid can be replaced by a NNAA. Benefits such as improved affinity, selectivity, and stability of peptide drugs can be obtained through the use of these NNAA.

These modified peptides are designed to overcome some limitations associated with traditional peptides: stability against proteolysis and low bioavailability. Additional properties, such as selectivity or receptor potency, could also be improved. Therefore, peptides with NNAAs have great potential in drug discovery because they offer an opportunity to improve the drug design process.

2.3 Protein structure

Knowledge of the three-dimensional (3D) structures of proteins is relevant to understand how the protein works, predict which molecules bind to that protein, and understand various biological interactions [45]. Generally, 3D structures are obtained using X-ray crystallography, nuclear magnetic resonance, or

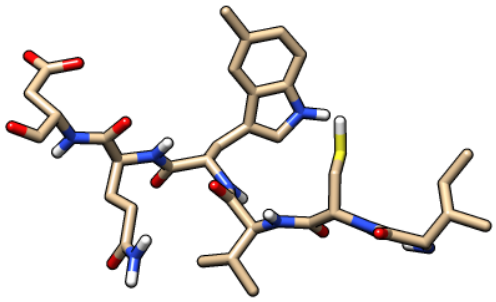

Figure 2.2: Peptide.

cryo-electron microscopy techniques. The Protein Data Bank (PDB) is the main repository containing experimentally determined 3D molecular structures [46, 47, 48], and its main purpose is to maintain a single archive of macromolecular structural data, which is free to use and available to the global community.

This structural information is also relevant in the field of structure-based drug design [49]. Knowledge of 3D structures is essential to understand how the molecular targets can interact with drug-like molecules in order to rationally optimize their chemical structures. This is useful in many stages of drug development, such as target validation, drug delivery, side-off effects, among others [50, 51]. Recently, this 3D information have been used in the design of new proteins with desired functions [50, 52].

Protein–ligand complex

A protein-ligand complex is a complex of a protein bound to a ligand through non-covalent interactions such as hydrogen bonding, metal coordination, hydrophobic forces, van der Waals forces, π–π interactions, or electrostatic effects [53]. The ligand can be any molecule that binds to the protein with high affinity and specificity. Therefore, a detailed understanding of protein-ligand interactions is critical to understand biology at the molecular level [54]. In addition, knowledge of the mechanisms responsible for molecular recognition and protein-ligand binding also contributes to the discovery, design, and development of drugs [55].

Half maximal inhibitory concentration (IC_{50})

In order to measure the binding affinity between ligand-protein complexes, it is useful to use the IC_{50} measurement. This can be determined with functional assays or with competitive binding assays. The IC_{50} evaluates the ability of a substance or a drug to inhibit a biological process. It is the concentration of the molecule or drug required to reduce the activity by 50%. This measure gives us an account of the capacity of a drug, and the lower the value of this IC_{50} measurement, the more effective the substance will be [56].

2.4 Molecular dynamics

MD is a computational method used to simulate classical physical motions of atoms and molecules. This movement depends on the simulation time and the chosen force field under different conditions, such as temperature, and pressure, among others [57]. MD requires an initial conformation of the system of study (either obtained from the PDB or by modelling tools like AlphaFold [58, 59]).

At time zero (t_0) the initial conformation $\mathbf{x}(t_0)$ has the position of all the atoms in the system. Then each of the atoms is assigned an initial velocity $v(t_0)$, which is chosen from a thermostat distribution at the temperature of study (see details about the thermostat in section 2.4.3). The force over each atom i is calculated using a classical potential ($V(\vec{x})$), or force field,

$$\vec{F}_i = \frac{-dV(\mathbf{x})}{d\vec{x}_i}, \tag{2.1}$$

where $\vec{F}_i$ is the force and $\vec{x}_i$ denotes the coordinates of atom i. These forces are derived from potentials that describe bond energies, valence angles, torsion angles, and Lennard-Jones interactions [60], details in section 2.5.

The atoms then evolve by numerically solving the equations of motion with a given time step. Newton's laws of motion are used to predict the spatial position of each atom as a function of time. From Newton's second law, we can solve and obtain the acceleration $\vec{a}_i$ of atom i.

$$\vec{a}_i = \frac{\vec{F}_i}{m_i}, \tag{2.2}$$

where m_i is the atom's mass. Next, an integration method is used to numerically integrate the equations of motion. This integration is done numerically: every time step, one calculates the forces on each atom and then uses those forces to update the position and velocity of each atom. We then evolve the system with small-time steps (dt), where the new positions and velocities are obtained after numerically integrating

Newton's equations of motion over time. We can use the leapfrog algorithm, where Δt is the size of each time step. At time step $i + 1$, the velocities are

$$\vec{v}_{i+1} = \vec{v}_i + \frac{1}{2}(\vec{a}_i + \vec{a}_{i+1})\Delta t, \tag{2.3}$$

and the positions are

$$\vec{x}_{i+1} = \vec{x}_i + \vec{v}_i \Delta t + \frac{1}{2}\vec{a}_i \Delta t^2. \tag{2.4}$$

The configurations are stored, and the interatomic forces are recalculated at each time step. This algorithm allows calculating velocities and positions in time steps as small as the vibratory motion of the system, thus reducing integration errors [61]. Finally, the system's evolution is a 3D trajectory that describes at atomic-level the configurations of the system at each point during the simulated time interval.

The importance of MD simulations stems from several reasons. One is that with MD it is possible to know the position and movement of each atom at each time, which has a high degree of complexity with respect to any experimental technique. On the other hand, the conditions of the simulation can be precisely known and can be controlled, such as, the initial conformation of a protein, what ligands are bound to it, if it has mutations, what other molecules are present in its environment, temperature, pressure, and many others. Therefore, by comparing simulations performed under different conditions, the effects of a wide variety of molecular perturbations can be identified [62].

MD applications are extensive. It is valuable for evaluating the mobility or flexibility of various regions of a biomolecule. In particular, by examining an MD simulation it is possible to quantify how many regions of the molecule move and what types of structural fluctuations they experience, thus complementing the determination of the 3D structure of biological systems. Furthermore, with MD simulations, it is possible to determine the dynamic behavior of water molecules and salt ions, which are often critical for protein function and ligand binding, thus improving structural models of biological systems. Another utility of MD is to test the accuracy of a modeled structure, or even refine it. MD is often used to test modeled binding poses of ligands because a pose that is stable in the simulation is more likely to be more accurate than one that is unstable [63]. Additionally, MD simulations can yield information about ligand binding to proteins and other macromolecules. This application is very useful in this study because it is the objective of this investigation.

Widely used programs for MD are GROMACS [64] (GROningen MAchine for Chemical Simulations) - package mainly designed for simulations of proteins, lipids, and nucleic acids, CHARMM [65] (Chemistry at HARvard Molecular Mechanics) – originally developed at Harvard, widely used for both small molecules and macromolecules, AMBER [66] (Assisted Model Building and Energy Refinement) – widely used for proteins and DNA, among others programs.

In the following, there are some relevant concepts for MD simulations:

2.4.1 Periodic boundary conditions

Periodic boundary conditions are necessary during MD simulations to preserve thermodynamic properties such as temperature, pressure, and density. In other words, with periodic boundary conditions, it is possible to approximate an infinite system using a unit cell or a periodic box. To implement them, the unit cell is surrounded by copies of the unit cell in all directions to approximate an infinitely large system. When a molecule diffuses across the simulation box boundary, it reappears on the opposite side. Therefore, each molecule always interacts with its neighbors, even though they may be on opposite sides of the simulation box [67].

2.4.2 Canonical ensemble (NVT) and (NPT)

A canonical ensemble represents the possible states of a mechanical system in equilibrium with a heat bath at a fixed temperature. To simulate the canonical ensemble in the field of MD simulations, we couple the system to a thermostat (an infinite heat bath), and it has no particle exchange with this bath [68].

For the canonical ensemble, the amount of substance (N), the volume (V), and the temperature (T) are conserved. In NVT, the energy of endothermic and exothermic processes is exchanged with a thermostat. Besides, the amount of substance (N), volume (V), and temperature (T) are kept constant. In the isothermal–isobaric ensemble, the amount of substance (N), pressure (P), and temperature (T) are conserved.

2.4.3 Temperature and pressure

At the molecular level, the temperature of a system is defined by the average kinetic energy of all the particles (atoms, molecules) that make up the system. In MD simulations to maintain a constant temperature, thermostat algorithms are used to allow energy to enter and leave the simulated system. In practice, thermostats do this by modifying the velocities of subsets of particles. The objective of controlling the temperature is to control the speed of the particles[69]. Temperature coupling is usually done with a Berendsen thermostat [70, 71, 72], which is an algorithm for rescaling particle velocities in molecular dynamics simulations to control the simulation temperature. Nose-Hoover thermostat [73], uses a Nose-Hoover extended ensemble. And, Andersen thermostat [74] is based on the reallocation of the speed of a chosen atom or molecule. The new speed is given by the Maxwell-Boltzmann statistics for the given

temperature.

In barostat algorithms, the goal is to keep the pressure in the simulation system constant or to apply external stress to the simulated system. The pressure is kept at a constant value by adjusting the volume of a periodic simulation system [75]. Pressure coupling is usually done with a Berendsen barostat [70], this uses exponential relaxation pressure coupling with a time constant. Parrinello-Rahman barostat [76], uses an extended-ensemble pressure coupling, where the box vectors are subject to an equation of motion. And Martyna-Tuckerman-Tobias-Klein barostat [77, 78], is similar to Parrinello-Rahman. In the present study, we use the Berendsen thermostat and Parrinello-Rahman barostat.

2.5 Force Fields

A molecule can be described as a series of charged points (atoms), which are held together by springs (bonds), in the field of classical MD. Specifically, a force field is used to describe bond lengths, bond angles, and twists, as well as non-bonding van der Waals interactions and electrostatic interactions between atoms. They are a collection of equations and associated constants designed to reproduce the molecular geometry and selected properties of tested structures [62]. Since each structure refers to a configuration in the conformational space of proteins, the force field gives us a complete description of the potential energy surface of proteins. Force fields help us to describe the energy of the protein in terms of its atomic coordinates [79].

Force fields consist of two components, Figure 2.3, the bond terms, and the non-bond terms. Bonding terms are those that describe interactions of bonds (stretching or compression of a pair of bonded atoms), angles (increase or decrease in bond angle), and dihedral (rotation of the dihedral angle), while non-bonding terms represent distant interactions, and describe electrostatics (modeled by Coulomb's Law) and Van der Waals interactions (modeled by the Lennard-Jones potential).

There are multiple force fields with different approaches. In the present study, we use the AMBER03 force field [81]. The Assisted Model Building with Energy Refinement (AMBER) is a force field for molecular dynamics. It is used to describe organic molecules and biological molecules, and it was developed primarily for the investigation of protein and nucleic acid systems [82].

In our particular case, for the C3c protein where all the amino acids are natural, the AMBER03 force field was used directly from the MD Gromacs program to generate the topology file. Additionally, in the case of the compstatin peptide ligand containing NNAA, it was not possible to execute the same procedure, in this case, we obtained the topology file through the PEPstrMOD server, but using the same force field.

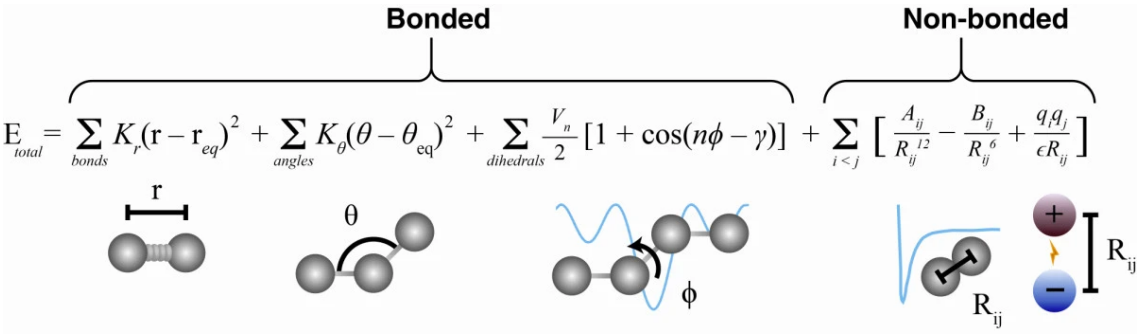

Figure 2.3: Typical force field model [80]. The atomic forces that govern molecular motion can be divided into those caused by interactions between atoms that are chemically bonded to each other and those that are not bonded. Bond terms are those that describe bond interactions: angles, dihedral. While the non-bonding terms represent distant interactions, they describe electrostatic interactions and Van der Waals interactions.

The force field parameters implemented in the PEPstrMOD server are adopted from Forcefield NCAA (Force Field for Noncanonical Amino Acids) [23], Forcefield PTM (Force Field for Post-Translational Modifications) [83], and SwissSideChain [22]. Importantly, the parameters derived from the PEPstrMOD are compatible with the ff03 force field in AMBER software package, which is the force field we used.

Topology file

To begin an MD simulation of the biological systems, two things are necessary. The first thing is to know the atomic connectivity information (to tell the MD program that one atom is linked to another). The second is to know the stiffness and equilibrium length of the bonds, angles, etc.These terms are described by a topology file [84].

Sometimes, one finds the need to simulate molecules for which topology and parameter information do not exist. In these cases, it is necessary to generate the topology file separately. For example, working with NNAA we find this limitation. This is why we use an alternative computational tool to generate the topology file, called PEPstrMOD server.

The PEPstrMOD server [24, 85] predicts the tertiary structure of small peptides, having a sequence length between 7 and 25 residues. In addition, it handles peptides having various modifications such as NNAAs, terminal modifications (acetylation/amidation), cyclization (N-C, disulfide bridges), conversion of L- to D-amino acids, and post-translational modifications. In the end, the resulting structure is further refined with energy minimization and molecular dynamics simulations. With this tool it is possible to make mutations of a natural amino acid by a NNAA, extracting the topology files of the new system

obtained with the AMBER force field.

2.6 Scoring functions

The scoring functions (SF) are an additive function that includes representations of various interactions between a ligand and a target receptor. These representations generally describe the electrostatic, hydrophobic, solvation, and hydrogen-bonding interactions between receptor and ligand [86]. Scoring functions assess their performance and effectiveness with various metrics, "scoring power", "ranking power", "docking power" and "screening power". The "scoring power" measures the ability of a SF to produce binding scores in a linear correlation with experimental binding data. The "ranking power" is the ability of a SF to correctly rank known ligands of a given target protein by their binding affinities, where the precise binding poses of those ligands are given. The "docking power" refers to the ability of a SF to identify the position of native ligand binders between computer-generated decoys, and, the "screening power" refers to the facility of a SF to identify the true binders to a given target protein among a pool of random molecules [87].

Typically, scoring functions have been created to evaluate the interaction of protein-protein complexes [88] and protein-ligand complexes [89] [90]. They can be used to determine the binding modeof a ligand, predict binding affinities, and identify potential drugs for a given target protein [91]. Scoring functions are essential for modern *in silico* drug discovery, but accurate prediction of binding affinity using these programs remains a difficult task. Besides, the performance of the scoring functions vary depending on the different target classes [92].

There are different types of scoring functions with diverse approaches, the most common are:

- Force-Field Based SF

  Force-field scoring functions generally quantify the interaction energy between the receptor and the ligand, and the internal energy of the ligand [93]. Similarly to the MD force field of a molecular system, atomic interactions are typically broken down into bond stretching energies, bond angle bending energies, bond torsion energies, hydrogen bond energies, van der Waals energies and electrostatic Coulomb energies.

- Empirical SF

  Empirical scoring functions are developed to replicate experimental affinity data on the assumption that binding free energy can be correlated with a set of unrelated variables [94]. By regression analysis using known binding affinity data of experimentally determined structures, it is possible to obtain the coefficients associated with the functional terms. Empirical scoring functions evaluate each system for specific terms that explain intermolecular interactions, such as van der Waals and electrostatic potentials [95].

- Knowledge-based SF

  Knowledge-based scoring function is also known as statistical potentials. It employs energy potentials that are derived from structural information embedded in experimentally determined atomic structures [96]. In other words, they are based on the statistical analysis of interacting atom pairs of protein-ligand complexes with available three-dimensional structures [95].

- Machine Learning based SF

  Scoring functions that use machine learning to predict receptor-ligand binding affinity do so by implementing multiple algorithms to predict affinities, rather than using just one. Furthermore, from complex structures with known binding affinities, these SFs can predict binding affinities for unknown molecules [97].
  Due to a large amount of data in biology, including protein structures, gene sequences, and binding data, the use of scoring functions with this approach is possible [98], as well as being useful for extracting features and learning patterns from complex data [99].

2.6.1 Details of the scoring functions

The scoring functions used in this work are from various categories such as empirical, knowledge-based, and machine learning-based. Below are some details of the scoring functions and the specific versions used.

- DLIGAND2 (VERSION 2) SF

  It is an improved energy function from the first version (Dligand) based on knowledge of protein-ligand interactions using the distance-scaled finite ideal-gas reference state. This SF has improvements in implementing a recently updated dataset containing 12,450 monomeric protein chains for training. Furthermore, it has a consistent improvement over the first version of Dligand in predicting binding affinities for native complex structures or docking-generated poses [100].

- CYSCORE SF

  Cyscore is an empirical scoring function. It is composed of hydrophobic free energy, van der Waals interaction energy, hydrogen bond interaction energy and ligand's conformational entropy. Those terms follow the classic approaches, except the hydrophobic free energy. This SF has significant improvements in prediction accuracy through the use of a new curvature-dependent surface area model, which is able to distinguish convex, planar, and concave surfaces in the calculation of hydrophobic free energy [101].

- DLSCORE SF

  The Dlscore scoring function takes a deep learning approach. Its main goal is to accurately predict the binding affinities between a receptor and a ligand. It consists of a set of neural networks, trained with the PDBBind (v2016) [102] database. Dligand uses a wide data set to improve the accuracy of its predictions. This SF is better than others in the consistency of its results, since it has less variability and fewer differences in terms of predicted affinity and experimental data [98].

- NNSCORE (VERSION 2.0) SF

  It is the second version of the NNScore scoring function. It is used to characterize the potency of receptor-ligand complexes. Furthermore, it is based on neural networks and computational models that simulate the microscopic organization of the brain [103, 104]. With this SF, better results could be obtained than with others, due to the use of different neural networks, and the implementation of various algorithms for affinity calculations, instead of having only one.

- VINA (VERSION 1.1.2) SF

  The scoring function is based on a combination of knowledge-based potentials and empirical scoring functions. It extracts empirical information both from the conformational preferences of the

receptor-ligand complexes and from the experimental affinity measurements. Its main goal is to predict bound conformations and binding affinity [105]. This SF has a significant improvement in speed and accuracy of binding mode prediction over its previous versions.

- SMINA SF

  Smina is based on AutoDock Vina 1.1.2. This empirical scoring function was specially designed to improve scoring and minimization. Additionally, it is optimized to support high-performance, user-specified custom scoring functions. [106].

2.7 SPEARMAN'S RANK CORRELATION COEFFICIENT

In order to discover the strength of a link between two data sets, it is useful to use Spearman's rank correlation coefficient [107]. This measure is ideal when we use variables that are not normally distributed. [108].

Spearman's correlation coefficient is useful for measuring the strength and direction of the correlation between two ranked variables [109]. To compute the coefficient, we assume that there are n pairs of observations from continuous distributions. The observations are classified into two separate samples, and ranked from smallest to largest. Let u_i be the rank of the i^{th} observation in the first sample, and let v_i be the rank of the i^{th} observation in the second sample. Spearman's rank correlation coefficient, r_S, is a measure of the correlation between ranks, calculated using the ranks instead of the actual observations [110, 111]:

$$r_S = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}, \tag{2.5}$$

where $d_i = u_i - v_i$.

Spearman's rank correlation coefficient, equation 2.5, can take values from +1 to -1. When r_s takes the value of +1 it means a perfect association between the ranks, in other words, the two variables have the same behavior, when one quantity increases the other also increases, or when one quantity decreases the other quantity also decreases. When $r_s = 0$ it means that there is no association between the ranks. Likewise, when r_s takes the value of -1, it means a perfect negative association of the ranks, that is, when one quantity increases, the other decreases, or when one quantity decreases the other increases. That is, when the values of r_s tend to 1 or -1, the correlation between variables is better.

The interpretation of the Spearman's correlation coefficient values is described as a "weak", "moderate" or "strong" relationship [112, 113], based on the following values reflected in the table 2.1.

Table 2.1: Interpretation of Spearman's Correlation Coefficient

Spearman's correlation coefficient	Interpretation
0.00 - 0.09	Negligible correlation
0.10 - 0.39	Weak correlation
0.40 - 0.69	Moderate correlation
0.70 - 0.89	Strong correlation
0.90 - 1.00	Very strong correlation

3

Scoring and molecular dynamics to rank compstatin peptide with NNAA bound to the Human C3c protein

This chapter includes the description of the system under study and the dataset of peptides with NNAA. It also includes, the results of the MD simulations and the scoring. Finally, the correlation between the experimental data and those from the MD simulations are presented.

3.1 Background

Complement component 3 (C3c) is a protein of the immune system. It plays a central role in the activation of the complement system, which is a part of the immune system and contributes to the innate immune system. In humans, it is encoded on chromosome 19 by a gene called C3. Compstatin, is a polypeptide with 13-residues, which has the following sequence ICVWQDWGAHRCT.

The compstatin peptide bound to protein C3c has been extensively studied, as well as compstatin analogs including NNAAs [114, 115, 116, 117]. Various analogs of compstatin with NNAAs have been analyzed both experimentally and computationally. The experiments have measured the experimental half-maximum inhibitory concentration (IC_{50}) of the modified peptides with the target.

This complex was taken as a reference for the elaboration of a force field called the NCAA Force Field. In this force field, more than 147 NNAAs are taken into account, and among them there are the ones used in this study. The importance of this new force field lies in the inclusion of these non-canonical amino acids because most force fields are designed to work with only natural amino acids [66, 118, 119, 120].

On the other hand, the family of compstatin peptides has been studied, which is made up of peptides that bind to the C3 protein and inhibit the activation of the complement system. In situations where there is inappropriate activation of the complement system, controlled inhibition with compstatin peptide is useful. This is desirable in cases of various autoimmune, inflammatory and pathological diseases[121]. In addition, the C3c-compstatin complex has been characterized and used for the study of other diseases [122], as well as, as a biomarker [123].

The purpose of this chapter is to run MD simulations of modified peptides bound to C3c and then to assess the MD conformations with scoring functions, with the specific goal of predicting the binding affinity between the complexes, following ideas from ref. [25]. To do this, we start from different analogs of the compstatin peptide having single-point mutations of NNAAs, and with affinity values reported. We ran the MD simulations with the complexes formed by the modified peptides bound to the C3c protein. After that, we score the conformations using six small-molecule scoring functions. Finally, we evaluated the performance in predicting the experimental rank of the peptides.

3.2 Methods

3.2.1 Biological system

The crystal structure of compstatin in complex with C3c (PDB id: 2QKI), is the system under study [124]. We choose this complex because it has reported information on bioactivity data on the binding of the C3c protein with analogs of the Compstatin peptide that contain non-natural amino acids [23]. It has been studied both in clinical trials and computationally, which allows a detailed view of the binding mode of the peptide compstatin with the protein, this being useful for the rational design of peptides and mimetics with improved activity [125]. We also note that this complex has been used in various studies because the peptide compstatin does not alter the conformation of C3c, while the peptide compstatin does undergo a large conformational change when bound to the protein. Such information is relevant to the development of compstatin for potential therapeutic use [126]. For these reasons, we were motivated to choose this crystal structure.

Due to the large size of the original protein and the unresolved and missing residues in the crystal structure, using the full structure of the C3c-compstatin complex could hinder calculations and add noise

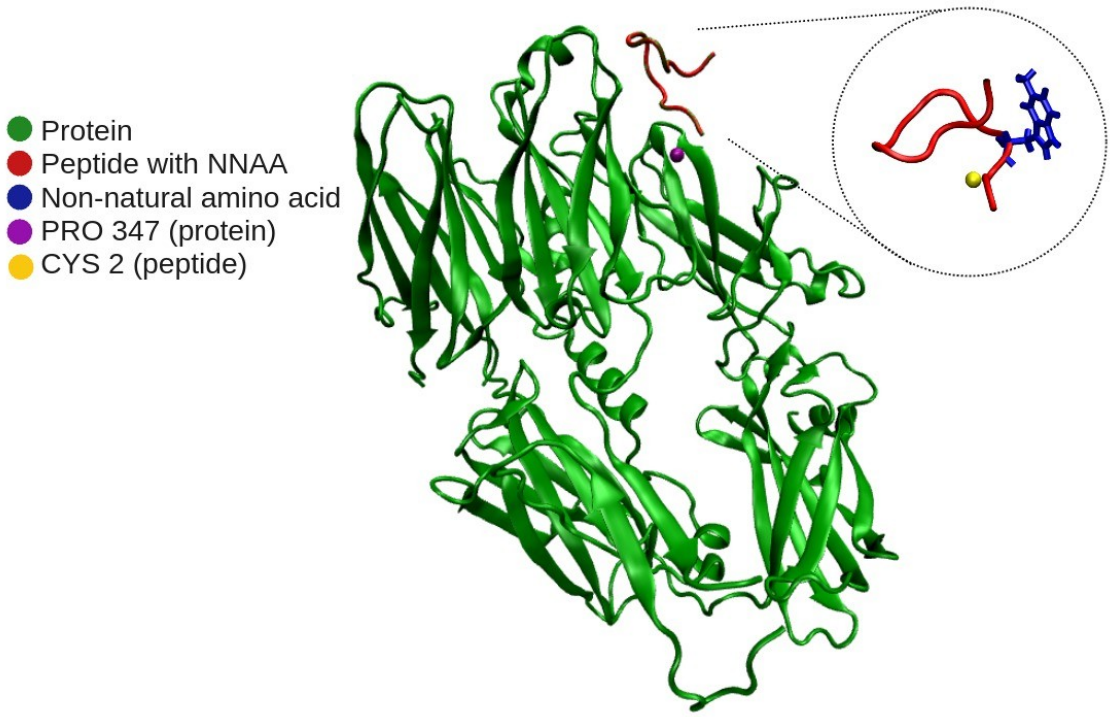

Figure 3.1: C3c protein structure bound to compstatin peptide. The PDB structure id is 2QKI. The interface is characterized by the chain A of the C3c protein (green color), the compstatin peptide (red color), and 5-methyltryptophan (MTR) which is the NNAA (blue color).

to the simulations. For this reason, to facilitate the simulation, we only used the chain A of the protein in complex with the chain G of the peptide compstatin. This is possible since the region where compstatin binds to C3c is confined and localized to one site, so it is not necessary to use the entire complex [23]. The chain linked to the compstatin peptide was taken as a template to model different peptides with NNAA, Figure 3.1.

3.2.2 NNAA selection and peptide formation

C3c protein is the selected structure in complex with the compstatin peptide. The compstatin peptide has the original sequence ICVWQDWGAHRCT and this sequence is modified to form different complexes. In this sequence, one natural amino acid is mutated by NNAA, thus forming the peptides with NNAA. The selected NNAA (Figure 3.2) are: 1-methyltryptophan (OMW), introduced at position 4, 5-methyltryptophan (MTR), introduced at position 4, O-methyltyrosine (OMY), introduced at position

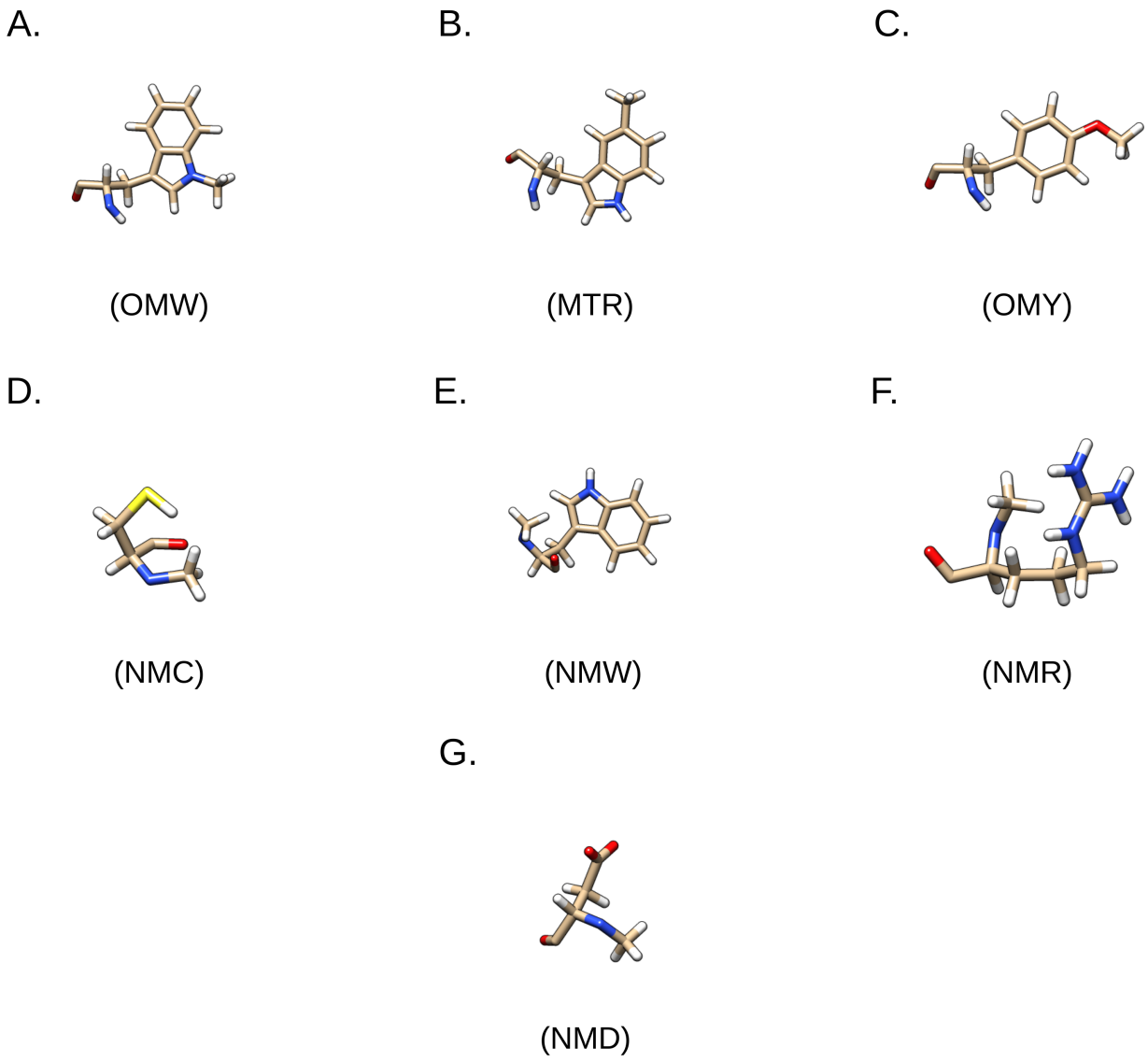

Figure 3.2: Non-natural amino acids: (A) 1-methyltryptophan (OMW), (B) 5-methyltryptophan (MTR), (C) O-methyltyrosine (OMY), (D) N-methylcysteine (NMC), (E) N-methyltryptophan (NMW), (F) N-methylarginine (NMR), (G) N-methylaspartic acid (NMD).

4, N-methylcysteine (NMC), introduced at position 2, and 12, N-methyltryptophan (NMW), introduced at position 7, N-methylarginine (NMR), introduced at position 11, and N-methylaspartic acid (NMD), introduced at position 6.

Experimental background information (IC_{50}) of compstatin peptide analogs with NNAA was obtained from Forcefield NCAA (Force Field for Noncanonical Amino Acids) [23], Table 3.1.

The list of complexes formed with the C3c protein and compstatin peptide with NNAA are:

- Complex 1: C3c protein-ICVWQDWGAHRCT
- Complex 2: C3c protein-ICV(OMW)QDWGAHRCT
- Complex 3: C3c protein-ICV(MTR)QDWGAHRCT
- Complex 4: C3c protein-ICV(OMY)QDWGAHRCT
- Complex 5: C3c protein-I(NMC)VYQDWGAHRCT
- Complex 6: C3c protein-ICVYQD(NMW)GAHRCT
- Complex 7: C3c protein-ICVYQDWGAH(NMR)CT
- Complex 8: C3c protein-ICVYQ(NMD)WGAHRCT
- Complex 9: C3c protein-ICVYQDWGAHR(NMC)T

3.2.3 Modeling of peptides with NNAA bound to Human C3c receptor

Starting from the selected C3c-compstatin with NNAAs complexes, the first step is to take the original peptide sequence of compstatin and mutate one of its natural amino acids for a NNAA, according to the sequences selected in Table 3.1. With the PEPstrMOD server, we performed these mutations. Then, the same program generates the topology files and initial conformation that will be used in the MD simulations. This process is similar to adding a ligand to a protein.

3.2.4 Molecular dynamics simulations

Each C3c-compstatin with NNAA complex was subjected to 100 ns of MD simulations with previous minimization and NVT/NPT equilibrating phases. GROMACS v2020 [64] was used to perform the simulations. The Amber03 protein forcefield [81] and TIP3P water model [127] were used for the protein and solvent, respectively. The parameters for the NNAA were found with the PEPstrMOD server (as described above). The protein was solvated in a cubic box of water with periodic boundaries at a distance of at least 8 Å from any atom of the protein. After solvation, counterions of Na+ and Cl– were included in the solvent to make the box neutral. The simulation was run using a modified Berendsen thermostat

Table 3.1: List of selected peptides with NNAA, template peptide (bold sequence).

Full sequence	IC_{50} [μ M]
ICVWQDWGAHRCT	1.20
ICV(OMW)QDWGAHRCT	0.21
ICV(MTR)QDWGAHRCT	0.87
ICV(OMY)QDWGAHRCT	1.30
I(NMC)VYQDWGAHRCT	7.50
ICVYQD(NMW)GAHRCT	25.00
ICVYQDWGAH(NMR)CT	32.00
ICVYQ(NMD)WGAHRCT	44.00
ICVYQDWGAHR(NMC)T	154.00

[70, 71, 72] at 310K temperature-coupling, and the Parrinello-Rahman barostat [76] at 1bar pressure-coupling. The electrostatic interactions were calculated using the Particle Mesh Ewald (PME) method with 1.0 nm short-range electrostatic and van der Waals cutoffs [128]. The equations of motion were solved with the leapfrog integrator [129] using a time step of 2 femtoseconds (fs).

3.2.5 Scoring functions

Six different scoring functions for protein-peptide with NNAA interactions were used to calculate scores over the conformations from all of each molecular dynamics trajectory. The scoring functions used were Cyscore, Dligand2, Dlscore, Nnscore, Smina, and Vina (see Section 2.6.1).

Scoring functions are used to assess the affinity between C3c-compstatin complexes with NNAA by scoring the trajectories. Conformations were recorded every 500 ps. For the scoring functions, Cyscore, Dligand2, Smina, and Vina give us an affinity result in molar units, in this case, the lower the result, the more affinity the complex has. On the other hand, for Nnscore and Dlscore SF, the affinity results are given in pK_d units, that is, the higher the result, the better the affinity. For these two scoring functions, we are going to multiply the result of the scores obtained by -1, to work all the SFs with the same metric.

3.2.6 Spearman correlation analysis

Spearman's correlation coefficient, r_s, measures the strength and direction of the association between two ranked variables. In our case, it will be used to relate the following variables. The first variable is the experimental IC_{50} values found in Table 3.1, and the second variable is the affinity scores obtained by the scoring functions. Specifically, MD conformations were evaluated with 6 scoring functions. The

conformations obtained were recorded every 500 ps, which would be a total of 200 structures for each C3c-compstatin complex with NNAA.

To calculate the Spearman rank correlation, it is first necessary to separate the two variables under study by ranks, from highest to lowest. Then, make sure that the variables under study follow a non-normal distribution, as is the case with our variables. Finally, obtain the correlation coefficients.

3.3 Results and Discussion

We generated topology parameters for the compstatin peptide analogs that include NNAAs. From these modified peptides bound to the C3c protein, we formed complexes to be sampled with MD simulations for a total of 100ns. Additionally, in order to corroborate the results, a replica of the system was made for 100 ns and the simulations were extended up to 200 ns. MD trajectories in the period 0 to 100 ns, 100 to 200 ns, and 0 to 200 ns were scored, different programs were used, and correlations against experimental activities were calculated.

3.3.1 Molecular dynamics

The convergence of the simulations MD was checked by calculating the all-atom root-mean-square deviation (RMSD) of the protein and peptide with NNAA. Additionally, the root-mean-square fluctuation (RMSF) of both the protein and peptide with NNAA. We found that the majority of the protein-peptide complexes with NNAA remained stable during the MD simulation. We monitored the RMSD, RMSF, and the distance between the receptor and ligand. As reference, we measured the distance between the alpha carbon atom of proline 347 from protein C3c and the alpha carbon atom of the cysteine (residue number 2) from the compstatin peptide, Figure 3.1. Further, figure 3.3, shows these observables for modified peptide (ICV(MTR)QDWGAHRCT), and C3c protein. The results for the additional complexes are found in Appendix A.1, in Figure A.1 to A.8.

From the RMSD plots for the protein and the peptide, (figure 3.3, and A.1), it is observed that of the protein the C3c protein-ICV(MTR)QDWGAHRCT, and C3c protein-I(NMC)VYQDWGAHRCT complexes are the most stable with variations of approximately 1Å. In the other complexes, the variations are up to 4Å. For the compstatin peptides with NNAA, in most plots, the complexes will remain stable with variations of 1Å, except for the C3c protein-ICVYQDWGAH(NMR)CT and C3c protein-ICVYQDWGAHR(NMC)T complexes which have variations of up to 4Å. In general, it is observed that the complexes remain stable over time.

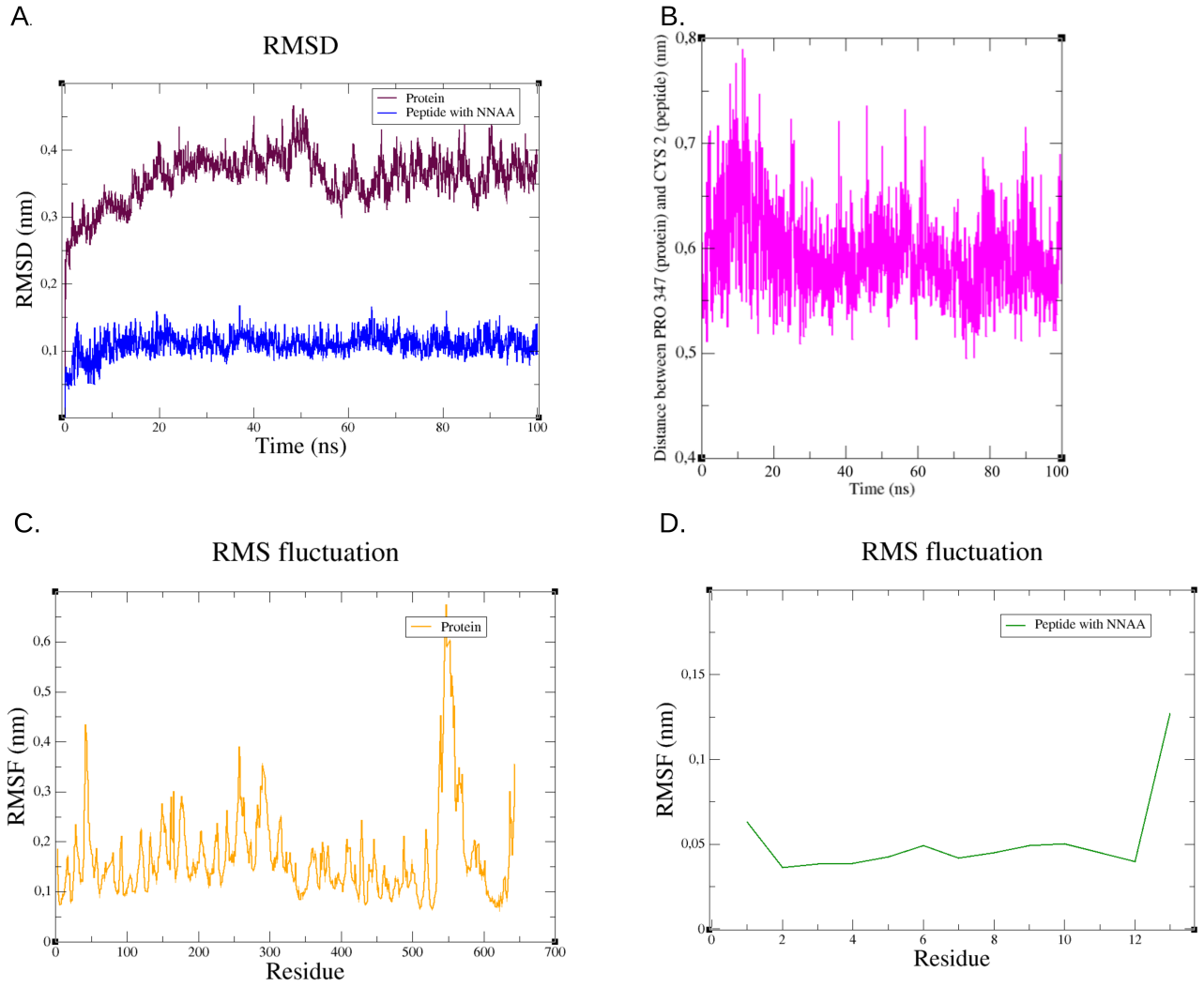

Figure 3.3: (A) RMSD of protein and peptide with NNAA. (B) Distance between the alpha carbon atom of proline 347 from protein C3c and the alpha carbon atom of the cysteine (residue number 2) from the compstatin peptide. (C) RMSF of the protein. (D) RMSF of the peptide with NNAA. All of the above for the complex C3c protein-ICV(MTR)QDWGAHRCT.

In order to validate the results, we ran one replica of each system for a period of 100 ns. The RMSD and RMSF graphs of the replica of these systems can be found in figure A.9 to A.17. Further, due to the low convergence of some complex systems and in order to validate the results, the simulations were extended by 100 ns, for a total of 200 ns. The RMSD and RMSF plots for these systems are found in Figure A.18 to A.26.

From the results obtained, instability is found in some of these complexes, such as: C3c protein-ICVWQDWGAHRCT and C3c protein-ICVYQD(NMW)GAHRCT. The potential reasons for these instabilities are that the force-field parameterization is poor. Additionally, the MD simulations were not performed with the original system of the C3c protein in complex with the peptide compstatin only chain A and G was used; so it is possible that the part of the protein that we excluded is important and stabilizes the system and/or that we are not using the exact environmental factors, for example, the temperature that was in the experiments was 300 K, instead, we used 310 K, which could also destabilize the complex.

3.3.2 Scoring the Molecular Dynamics conformations

Six different scoring functions for protein-peptide interactions were used to calculate the scores over the period 0 to 100 ns, 100 to 200 ns, and 0 to 200 ns. In the tables of Appendix A.2, the scores obtained with each of the scoring functions are found. On average, we found that the best binding affinity results are the complexes evaluated in the period from 0 to 100 ns, followed by 0 to 200 ns, and finally, the complexes in the period from 100 to 200 ns.

3.3.3 Spearman Correlation Statistical Analysis

Spearman's rank correlation was used to compare the scores obtained from the MD conformations and the experimental IC_{50} ranking. The ranks of the scoring functions versus the experimental ranks for all the scoring functions in the period from 0 to 100 ns, figure 3.4. The points in blue color represent the data by ranks and the orange line represents the best fit line, also called the trend line or linear regression. This is a straight line that helps us see if there is a relationship or correlation between the two factors being studied [130]. The orange lines show the best fit, following the general trend of the scoring functions and the experimental data. As you can see, some points are on the line, while others are above or below. The data that is closest to the line is the data that correlates best. From the graphs, we observe that the Nnscore, and Cyscore function have the best predictive power. In addition, the graphs show lines with a positive slope. Data with a positive slope means that when one of the variables increases, the other also increases, or if one decreases, the other also decreases.

Once we have the data of the two variables under study, it is possible to calculate the Spearman rank correlation, according to equation 2.5. Regarding the correlations obtained, the best correlations are the data of the coefficients that are closer to 1. For the system, Table 3.2, the scoring functions Nnscore and Cyscore that the data tend to be >0.75, it means that the affinity score prediction correlates with the experimental data of IC_{50}. i.e. a strong correlation is obtained [113, 112]. For the scoring functions Dlscore, Smina, Vina and Dligand2, there is a moderate correlation.

Table 3.2: Spearman's rank correlation for each scoring function at 310K.

Scoring Functions	Correlation 0 - 100 ns	Correlation 100 - 200 ns	Correlation 0 - 200 ns
Cyscore	0.783	0.733	0.733
DLIGAND2	0.833	0.333	0.667
dlscore	0.433	0.717	0.617
nnscore	0.866	0.783	0.799
smina	0.799	0.483	0.700
vina	0.799	0.483	0.799

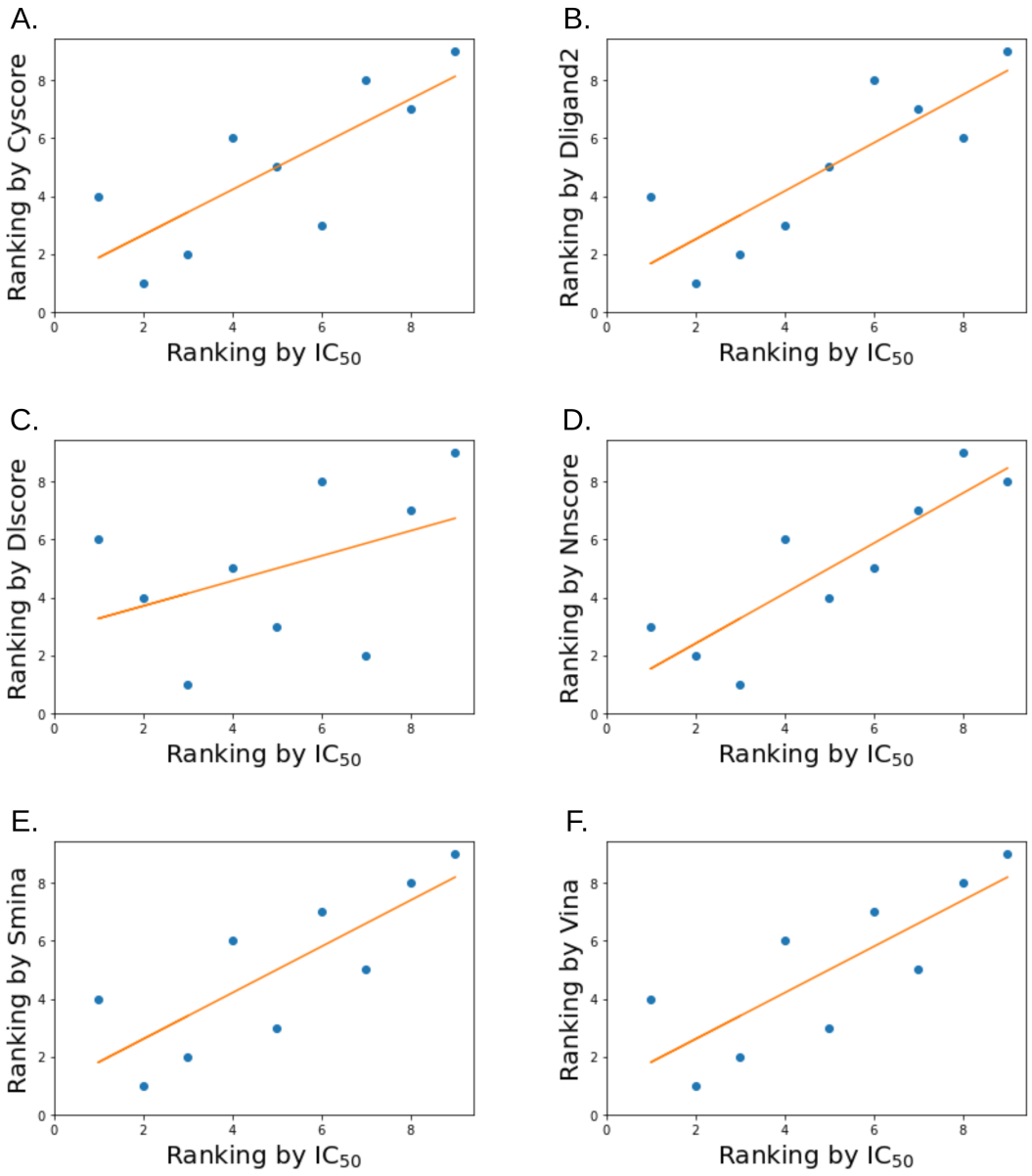

Figure 3.4: Experimental rank IC_{50} vs rank scoring function, in the period from 0 to 100 ns. (A) Cyscore SF. (B) Dligand2 SF. (C) Dlscore SF. (D) Nnscore SF. (E) Smina SF. (F) Vina SF.

4
Conclusions

In this study, we evaluated the binding affinity for various complexes formed from the C3c protein in binding with the compstatin peptide using a computational approach.

We selected a small set of modified peptides to rank. The challenge was to include them in the MD simulations, since most force field programs do not consider NNAAs. For this reason, it was necessary to look for program alternatives for the generation of topology parameters, which define which atoms are connected to each other through chemical bonds and how they interact in the MD simulations. With the help of an online server called PEPstrMOD, we generated the topology files.

Based on the set of previously defined protein-peptide topologies with NNAA, we performed the MD simulations. The convergence of the MD simulations was evaluated by calculating the RMSD and RMSF. In general, the complexes remained stable, that is, the complex binding it does not drastically change in time.

In order to evaluate the obtained MD conformations, we use the scoring functions to predict the binding affinity between the protein-peptide complex. The scoring functions we use are based on different principles, such as empirical, knowledge-based, deep learning-based, and neural network-based, with the intention of considering different approaches. In order to evaluate the entire space of conformations obtained in the MD simulations, we calculated the average of each scoring function in the period 0 to 100 ns, 100 to 200 ns, and 0 to 200 ns. Additionally, to validate the results, we ran one replica of each system

for a period of 100 ns, and the simulation time was extended to 200 ns. Finally, with the scores obtained from the MD simulations and with experimental data obtained from other studies [23], it is possible to correlate these two variables. We find the highest Spearman correlation for the Nnscore and Cyscore scoring function, suggesting that these are the most adequate for ranking the binding of modified peptides.

5
Perspectives

Here, we have implemented a computational methodology using the peptide-bound protein C3c compstatin to predict the binding affinity between protein-ligand complexes. Various NNAA containing analogs of the compstatin peptide were used to form the C3c-compstatin complexes. In future studies, it would be worthwhile to expand the number of compstatin peptide analogs containing various experimental binding values. This would be useful since it would broaden the spectrum of the experimental variables, providing a better sampling of the affinity data. Moreover, optimizing the starting NNAA side chain conformations, as done for natural amino acids in ref. [131].

On the other hand, regarding the scoring of the MD conformations, six scoring functions with different approaches were used in this work. However, it would be convenient in future research to include scoring functions with other approaches to more exhaustively sample the entire conformation space.

As a perspective, due to the use of NNAA in the study complex, it is possible to use this work as a comparative evaluation of receptor-ligand complexes that include NNAA. Specifically, it is possible to use this research as benchmarking in peptide design protocols, such as the PARCE protocol [132]. The purpose of PARCE is to design peptides and proteins with better affinities toward a particular target, including proteins. For this reason, the inclusion of this study as benchmarking in this protocol could expand the set of binding molecule proposals and include NNAAs.

A
Appendix

A.1 RMSD and RMSF graphs for all complexes of protein C3 and peptide Compstatin with NNAA

RMSD and RMSF graphs to observe the convergence of the MD simulations of the complex of C3c protein and compstatin peptide with NNAA, during 100 ns.

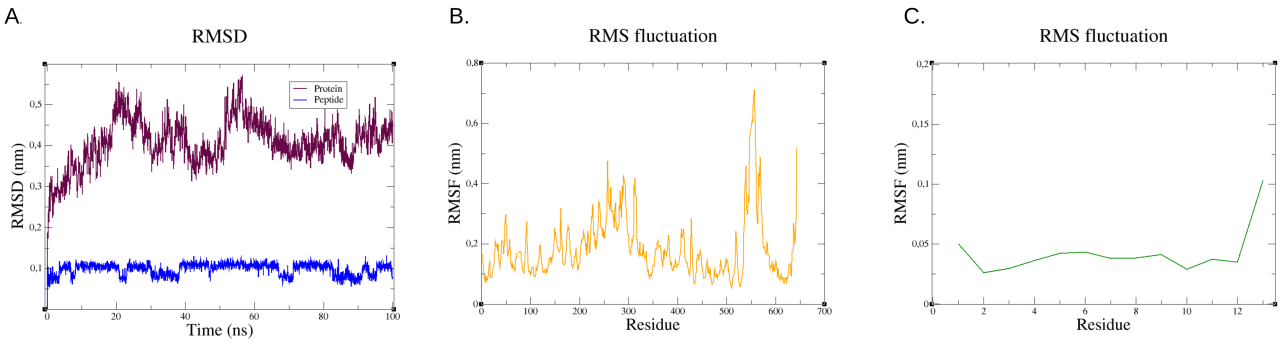

Figure A.1: (A) RMSD of protein and peptide with NNAA. (B) RMSF of the protein. (D) RMSF of the peptide with NNAA. All of the above for the complex C3c protein-ICVWQDWGAHRCT.

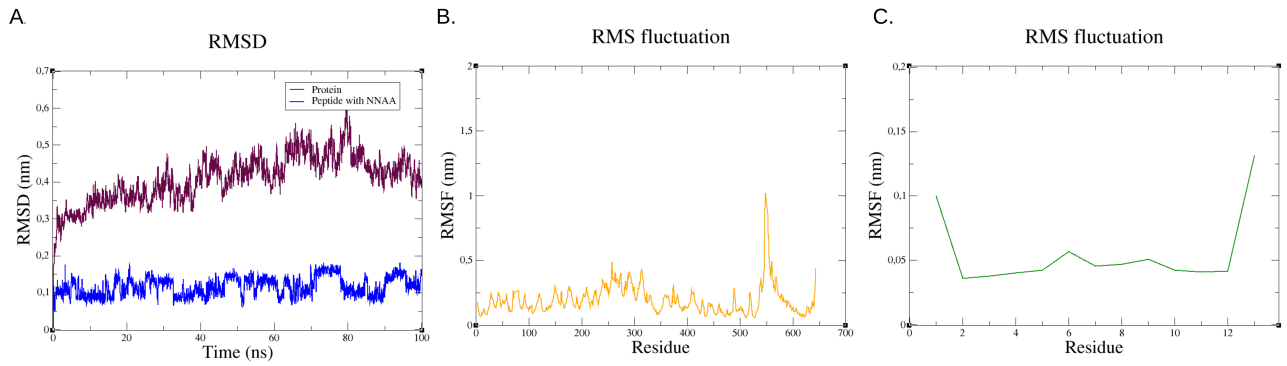

Figure A.2: (A) RMSD of protein and peptide with NNAA. (B) RMSF of the protein. (D) RMSF of the peptide with NNAA. All of the above for the complex C3c protein-ICV(OMW)QDWGAHRCT.

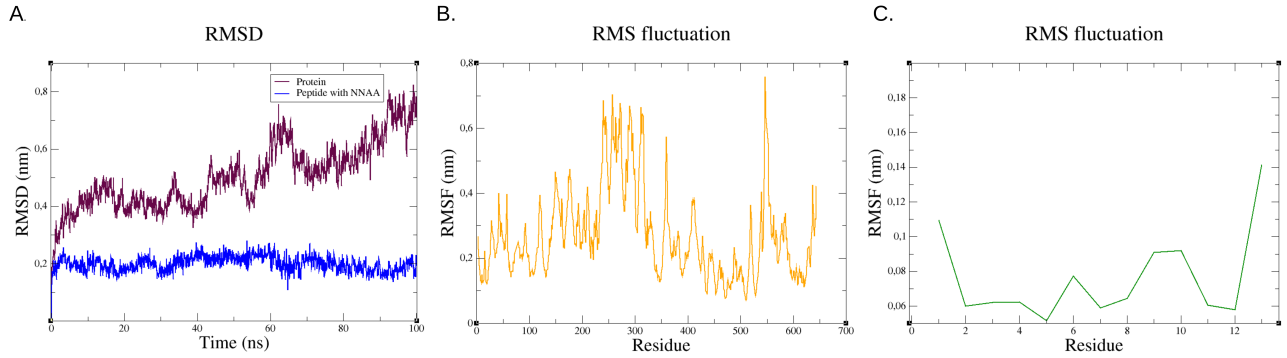

Figure A.3: (A) RMSD of protein and peptide with NNAA. (B) RMSF of the protein. (D) RMSF of the peptide with NNAA. All of the above for the complex C3c protein-ICV(OMY)QDWGAHRCT.

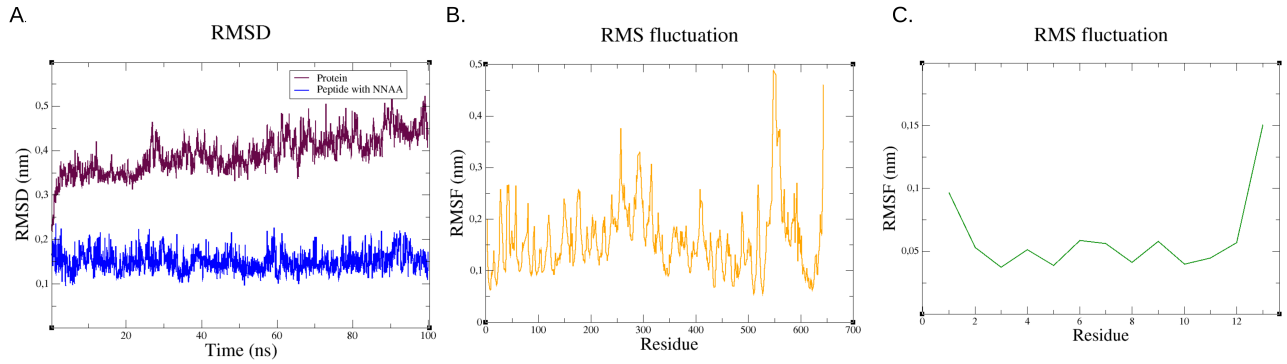

Figure A.4: (A) RMSD of protein and peptide with NNAA. (B) RMSF of the protein. (D) RMSF of the peptide with NNAA. All of the above for the complex C3c protein-I(NMC)VYQDWGAHRCT.

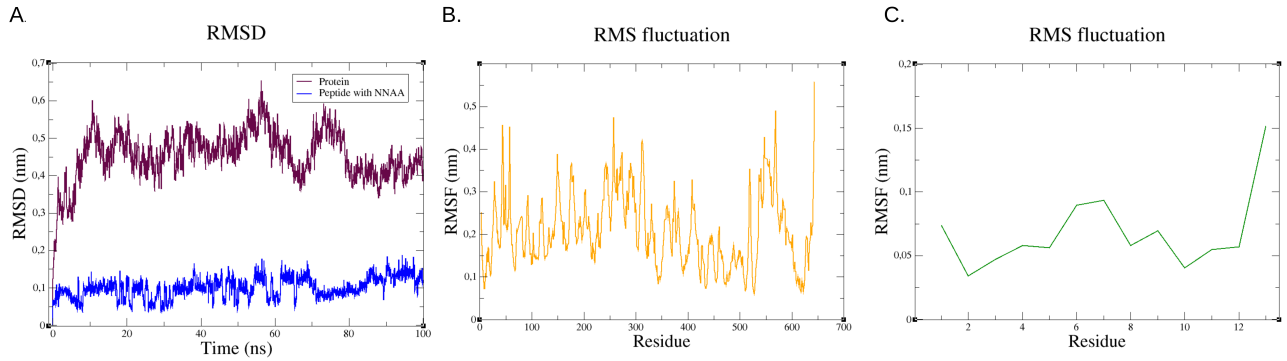

Figure A.5: (A) RMSD of protein and peptide with NNAA. (B) RMSF of the protein. (D) RMSF of the peptide with NNAA. All of the above for the complex C3c protein-ICVYQD(NMW)GAHRCT.

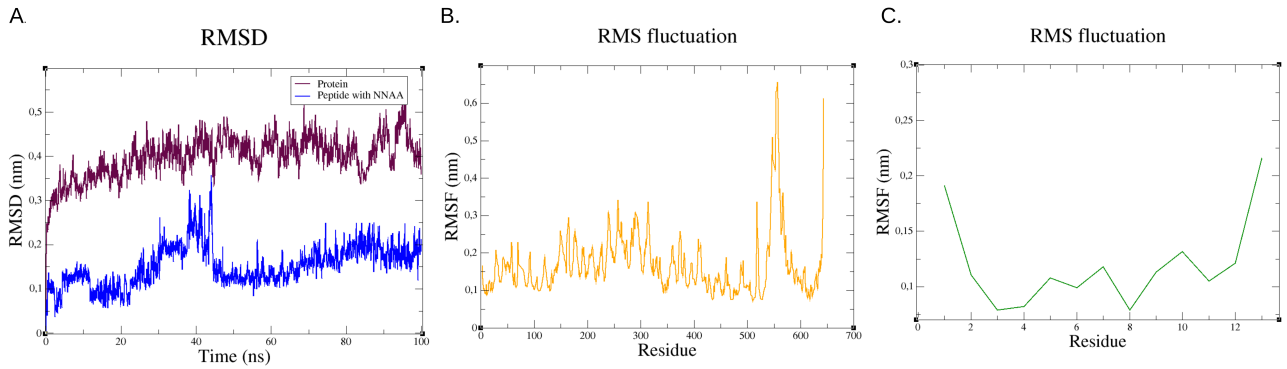

Figure A.6: (A) RMSD of protein and peptide with NNAA. (B) RMSF of the protein. (D) RMSF of the peptide with NNAA. All of the above for the complex C3c protein-ICVYQDWGAH(NMR)CT.

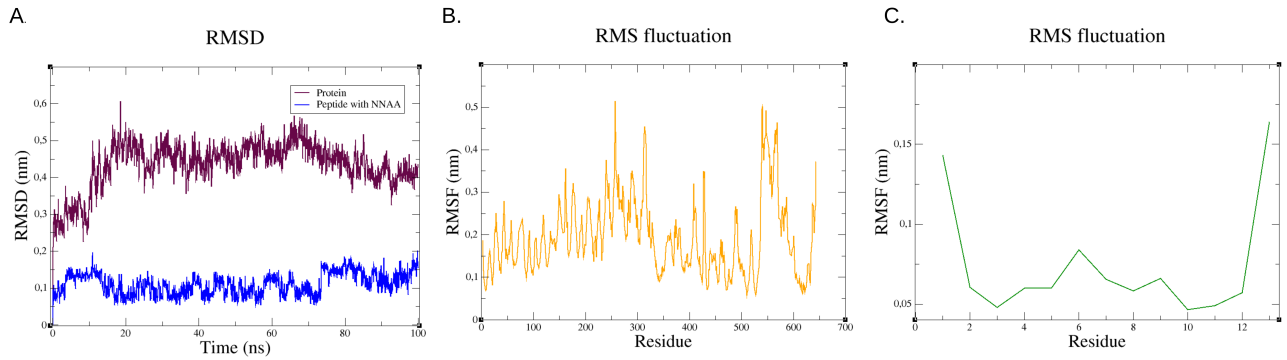

Figure A.7: (A) RMSD of protein and peptide with NNAA. (B) RMSF of the protein. (D) RMSF of the peptide with NNAA. All of the above for the complex C3c protein-ICVYQ(NMD)WGAHRCT.

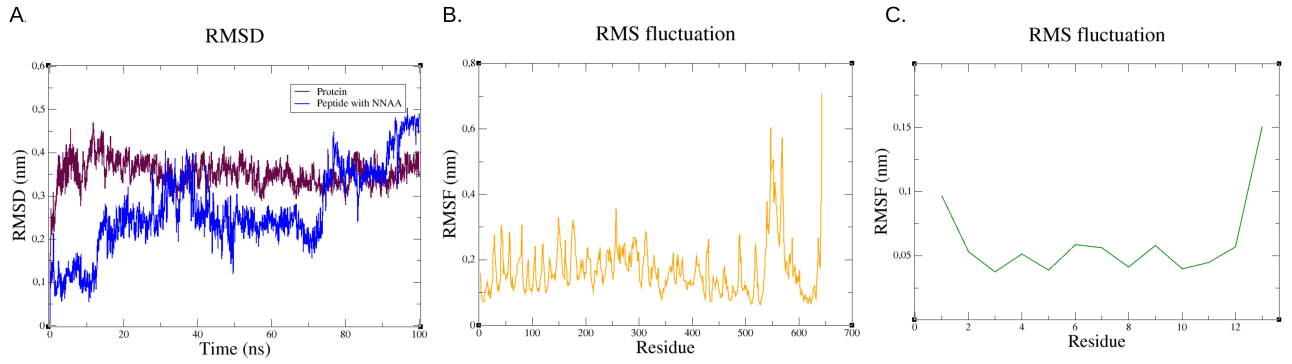

Figure A.8: (A) RMSD of protein and peptide with NNAA. (B) RMSF of the protein. (D) RMSF of the peptide with NNAA. All of the above for the complex C3c protein-ICVYQDWGAHR(NMC)T.

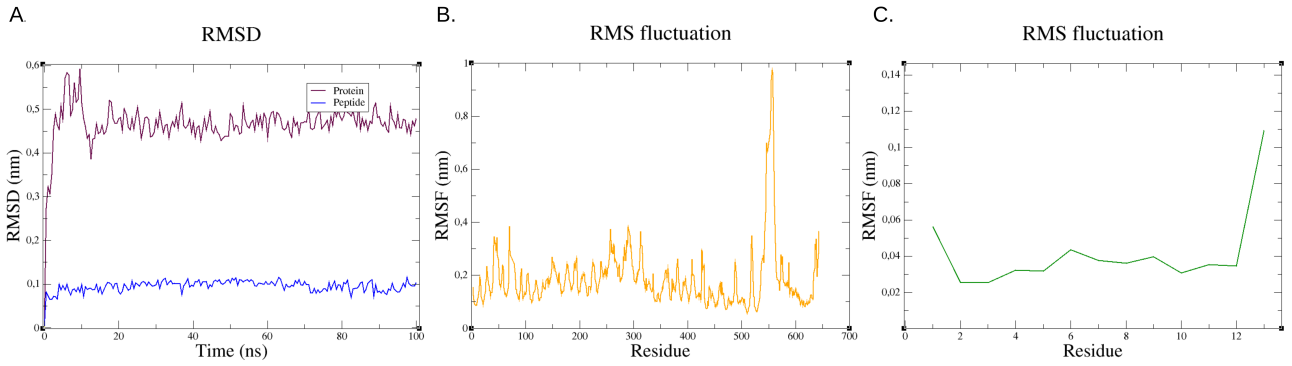

Figure A.9: (A) RMSD of protein and peptide with NNAA. (B) RMSF of the protein. (D) RMSF of the peptide with NNAA. All of the above for the complex C3c protein-ICVWQDWGAHRCT replica.

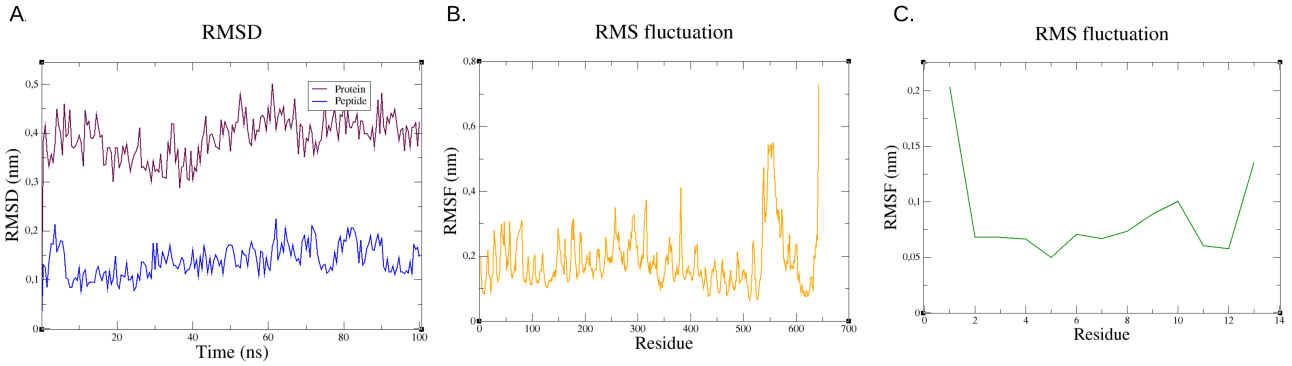

Figure A.10: (A) RMSD of protein and peptide with NNAA. (B) RMSF of the protein. (D) RMSF of the peptide with NNAA. All of the above for the complex C3c protein-ICV(OMW)QDWGAHRCT replica.

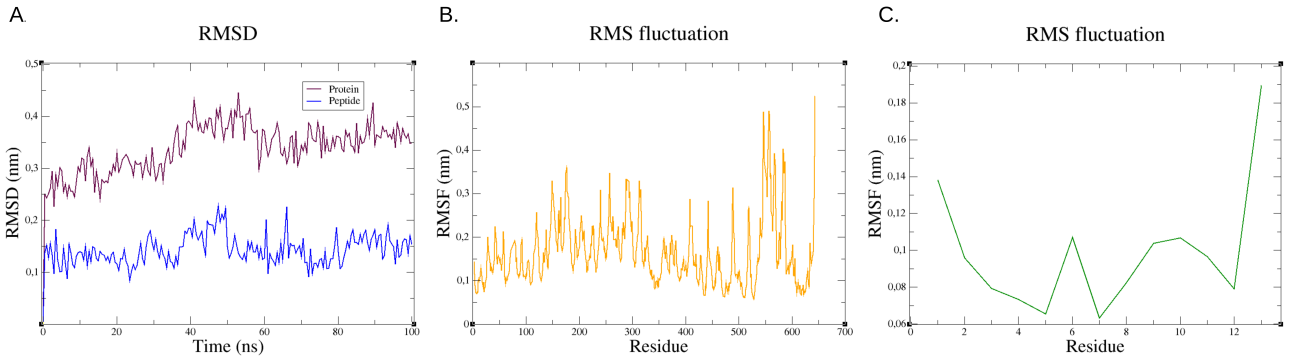

Figure A.11: (A) RMSD of protein and peptide with NNAA. (B) RMSF of the protein. (D) RMSF of the peptide with NNAA. All of the above for the complex C3c protein-ICV(MTR)QDWGAHRCT replica.

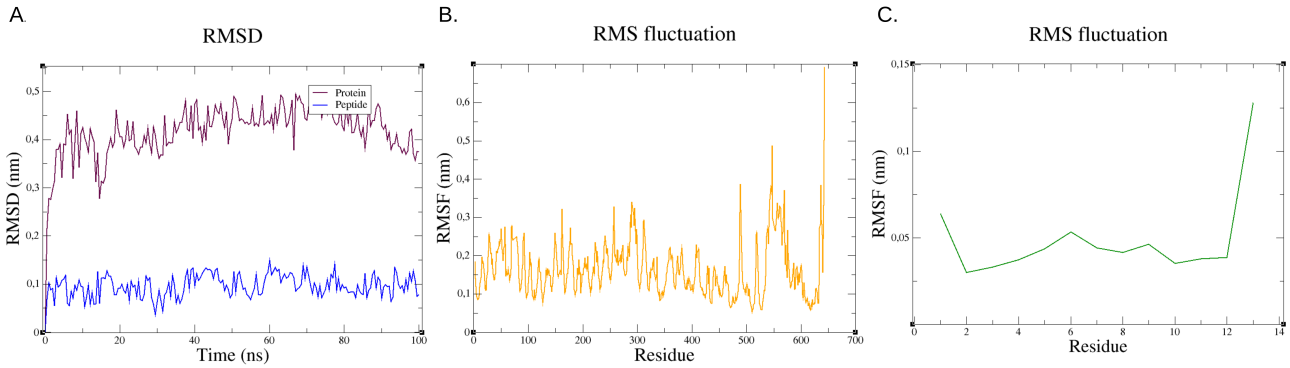

Figure A.12: (A) RMSD of protein and peptide with NNAA. (B) RMSF of the protein. (D) RMSF of the peptide with NNAA. All of the above for the complex C3c protein-ICV(OMY)QDWGAHRCT replica.

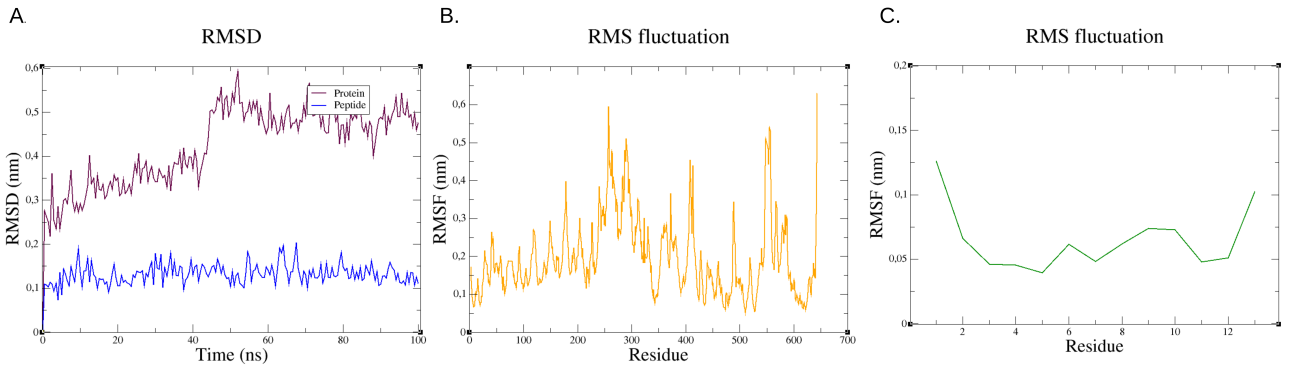

Figure A.13: (A) RMSD of protein and peptide with NNAA. (B) RMSF of the protein. (D) RMSF of the peptide with NNAA. All of the above for the complex C3c protein-I(NMC)VYQDWGAHRCT replica.

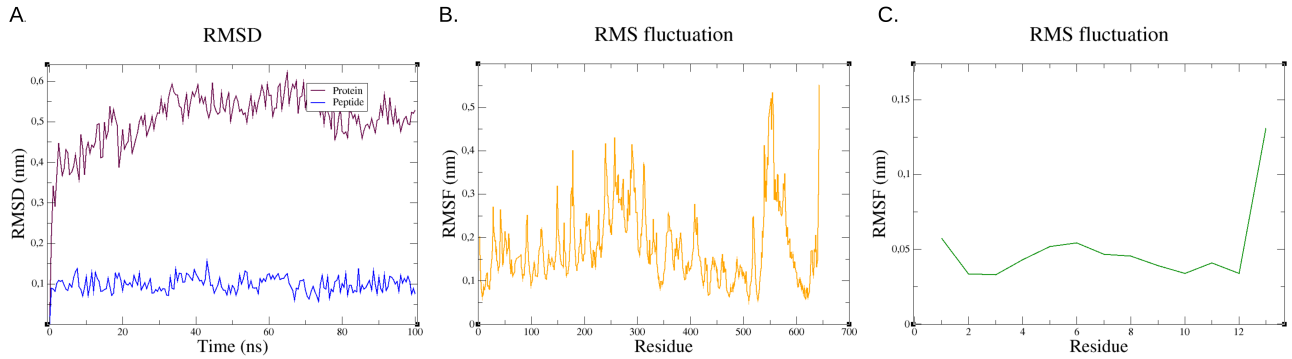

Figure A.14: (A) RMSD of protein and peptide with NNAA. (B) RMSF of the protein. (D) RMSF of the peptide with NNAA. All of the above for the complex C3c protein-ICVYQD(NMW)GAHRCT replica.

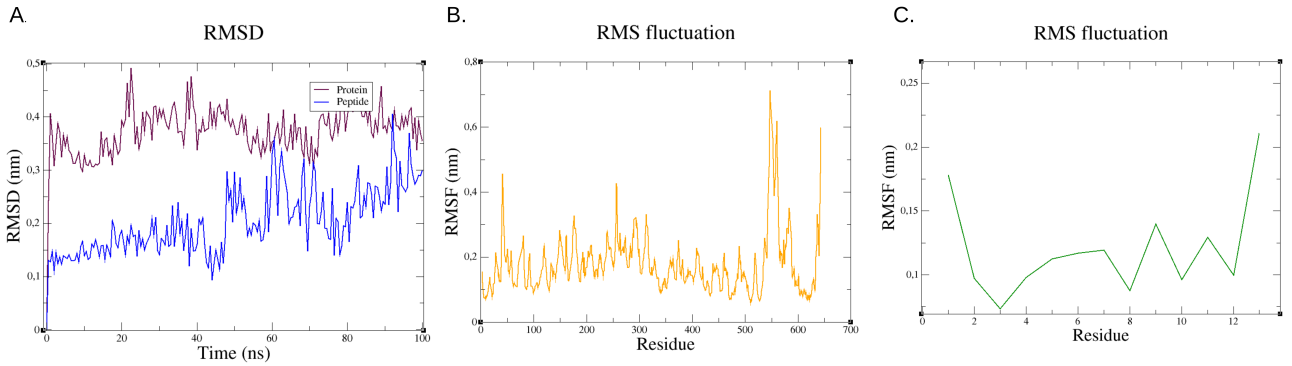

Figure A.15: (A) RMSD of protein and peptide with NNAA. (B) RMSF of the protein. (D) RMSF of the peptide with NNAA. All of the above for the complex C3c protein-ICVYQDWGAH(NMR)CT replica.

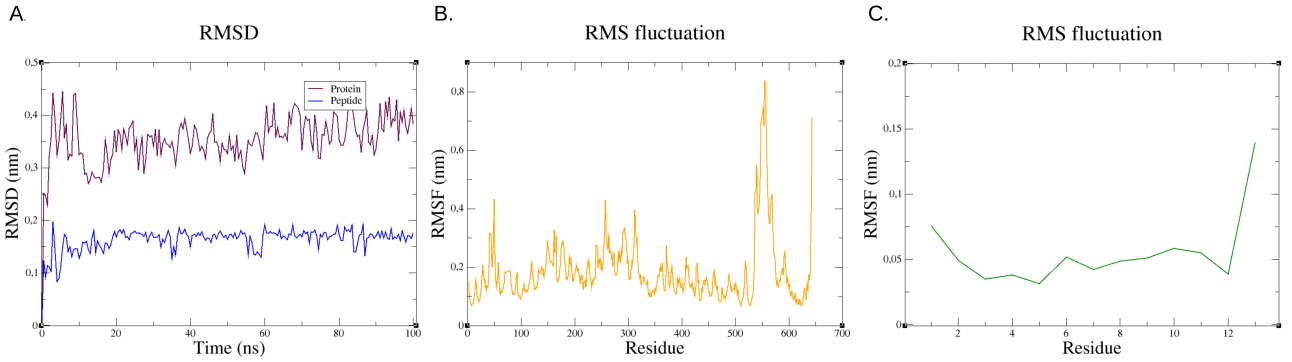

Figure A.16: (A) RMSD of protein and peptide with NNAA. (B) RMSF of the protein. (D) RMSF of the peptide with NNAA. All of the above for the complex C3c protein-ICVYQ(NMD)WGAHRCT replica.

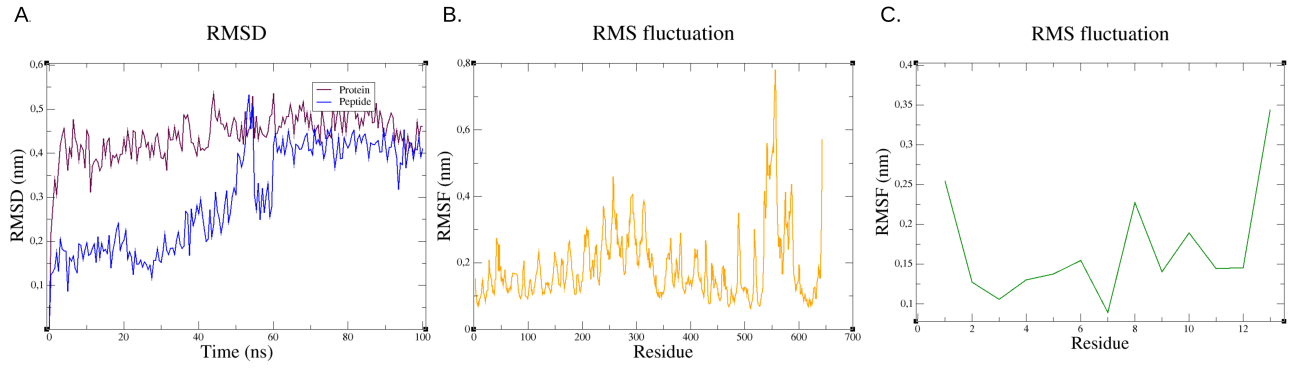

Figure A.17: (A) RMSD of protein and peptide with NNAA. (B) RMSF of the protein. (D) RMSF of the peptide with NNAA. All of the above for the complex C3c protein-ICVYQDWGAHR(NMC)T replica.

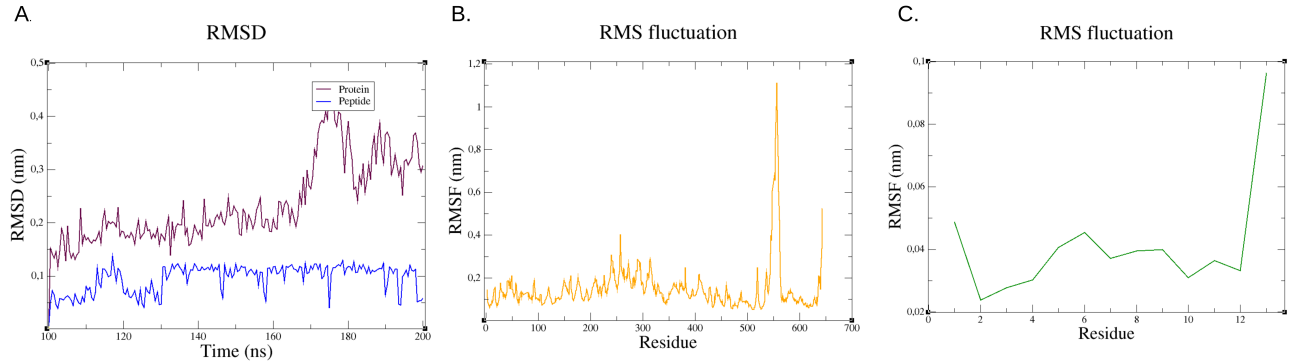

Figure A.18: (A) RMSD of protein and peptide with NNAA. (B) RMSF of the protein. (D) RMSF of the peptide with NNAA. All of the above for the complex C3c protein-ICVWQDWGAHRCT, in the period of 100 to 200 ns.

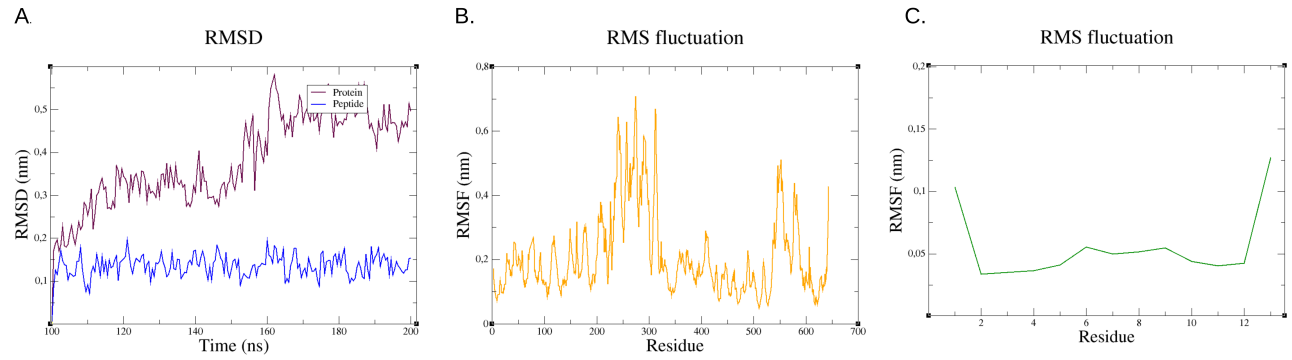

Figure A.19: (A) RMSD of protein and peptide with NNAA. (B) RMSF of the protein. (D) RMSF of the peptide with NNAA. All of the above for the complex C3c protein-ICV(OMW)QDWGAHRCT, in the period of 100 to 200 ns.

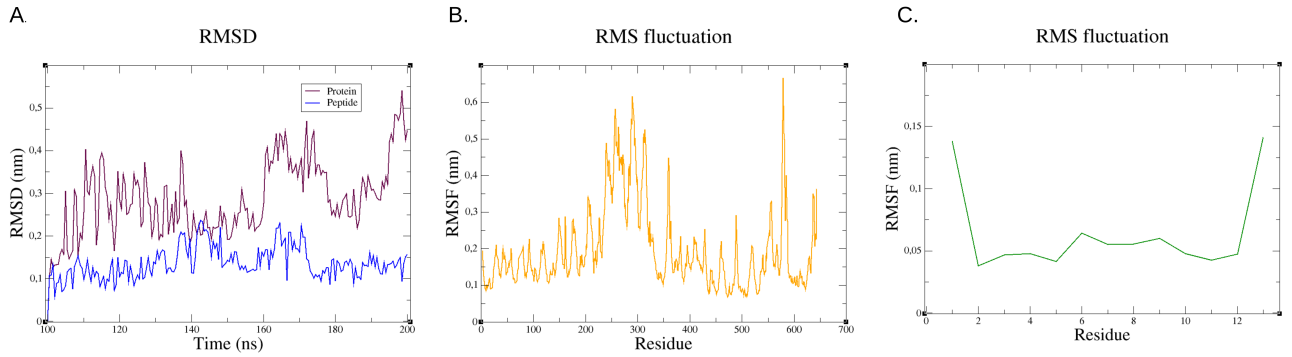

Figure A.20: (A) RMSD of protein and peptide with NNAA. (B) RMSF of the protein. (D) RMSF of the peptide with NNAA. All of the above for the complex C3c protein-ICV(OMY)QDWGAHRCT, in the period of 100 to 200 ns.

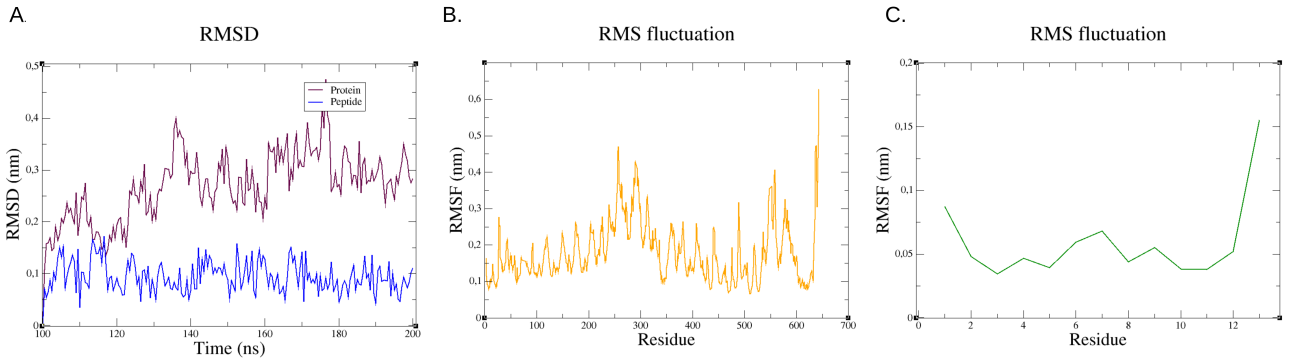

Figure A.21: (A) RMSD of protein and peptide with NNAA. (B) RMSF of the protein. (D) RMSF of the peptide with NNAA. All of the above for the complex C3c protein-I(NMC)VYQDWGAHRCT, in the period of 100 to 200 ns.

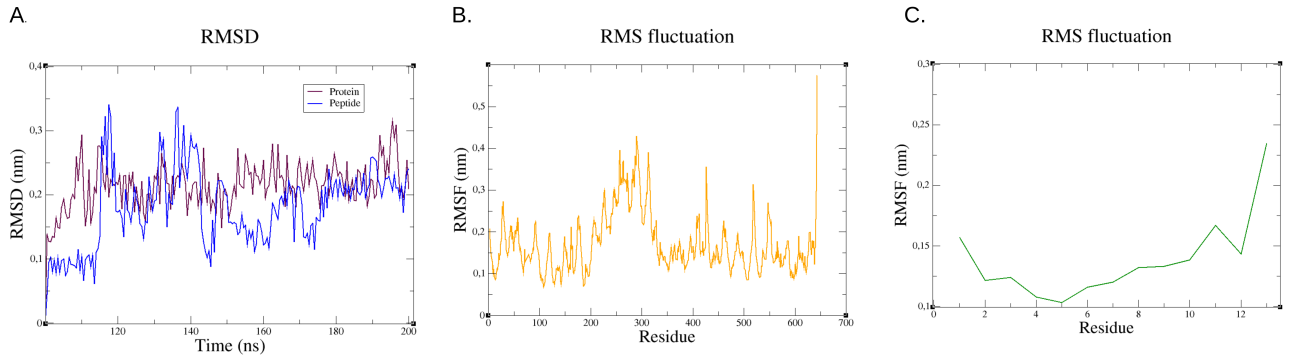

Figure A.22: (A) RMSD of protein and peptide with NNAA. (B) RMSF of the protein. (D) RMSF of the peptide with NNAA. All of the above for the complex C3c protein-ICVYQD(NMW)GAHRCT, in the period of 100 to 200 ns.

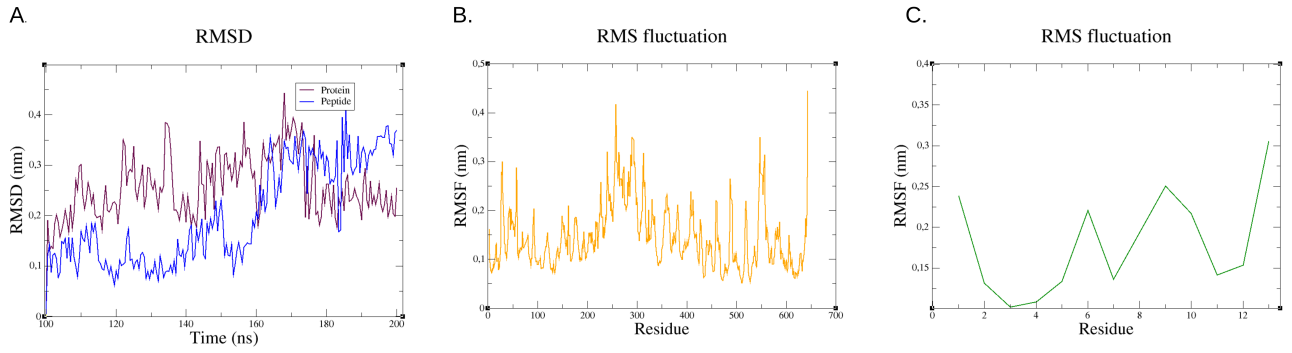

Figure A.23: (A) RMSD of protein and peptide with NNAA. (B) RMSF of the protein. (D) RMSF of the peptide with NNAA. All of the above for the complex C3c protein-ICVYQDWGAH(NMR)CT, in the period of 100 to 200 ns.

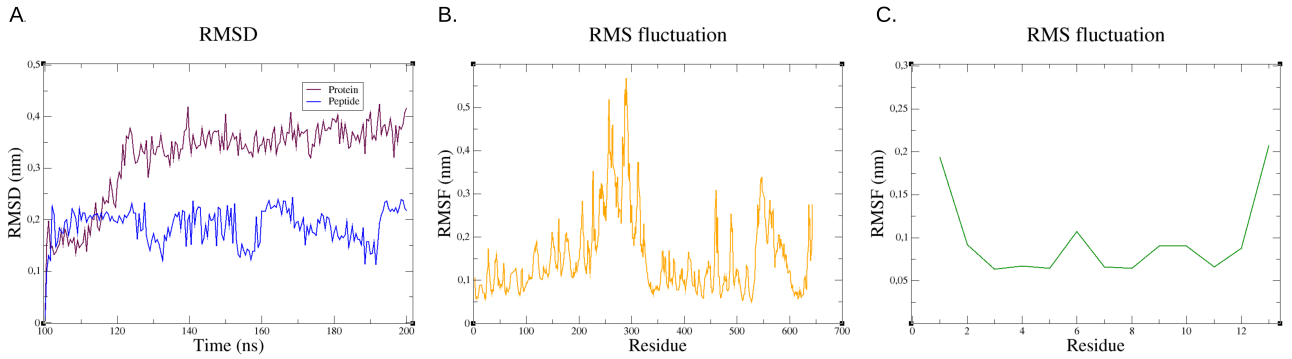

Figure A.24: (A) RMSD of protein and peptide with NNAA. (B) RMSF of the protein. (D) RMSF of the peptide with NNAA. All of the above for the complex C3c protein-ICVYQ(NMD)WGAHRCT, in the period of 100 to 200 ns.

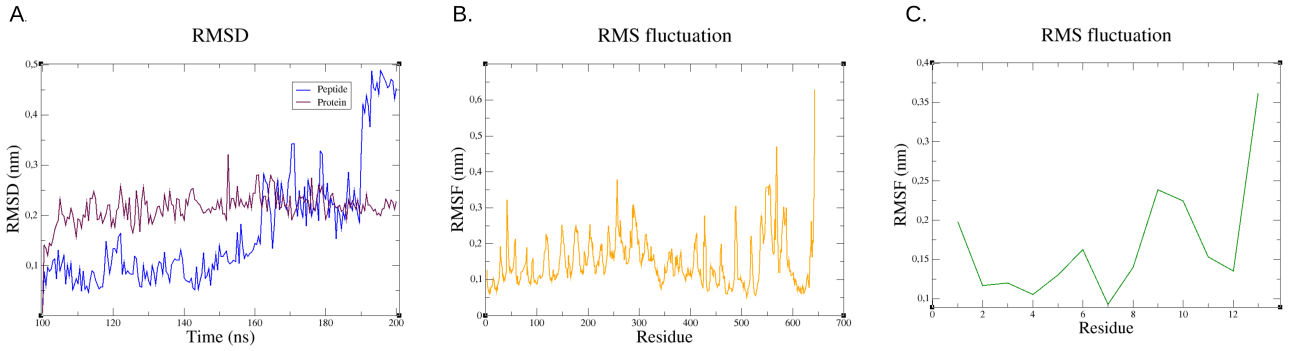

Figure A.25: (A) RMSD of protein and peptide with NNAA. (B) RMSF of the protein. (D) RMSF of the peptide with NNAA. All of the above for the complex C3c protein-ICVYQDWGAHR(NMC)T, in the period of 100 to 200 ns.

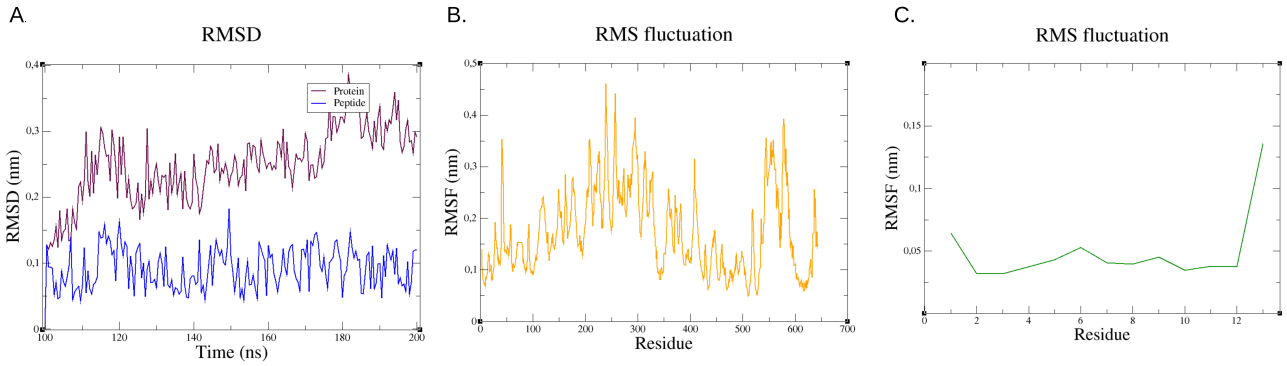

Figure A.26: (A) RMSD of protein and peptide with NNAA. (B) RMSF of the protein. (D) RMSF of the peptide with NNAA. All of the above for the complex C3c protein-ICV(MTR)QDWGAHRCT, in the period of 100 to 200 ns.

A.2 Score of C3c protein and compstatin peptide complexes with different NNAA of all scoring functions

Table A.1: Cyscore score in the period 0 to 100 ns, 100 to 200 ns, and 0 to 200 ns.

Complex	Score 0 - 100 ns	Score 100 - 200 ns	Score 0 - 200 ns
1	-1,847	-2,053	-1,950
2	-1,672	-1,180	-1,426
3	-1,988	-2,153	-2,070
4	-1,355	-1,711	-1,533
5	-1,641	-1,410	-1,525
6	-1,797	-1,191	-1,494
7	-1,308	-1,162	-1,235
8	-1,337	-0,114	-0,726
9	-1,242	-1,077	-1,159

Table A.2: Dligand2 score in the period 0 to 100 ns, 100 to 200 ns, and 0 to 200 ns.

Complex	Score 0 - 100 ns	Score 100 - 200 ns	Score 0 - 200 ns
1	-18,756	-18,692	-18,728
2	-17,944	-17,024	-17,466
3	-18,880	-19,045	-18,958
4	-18,249	-18,258	-18,254
5	-17,863	-17,524	-17,696
6	-17,319	-14,705	-15,700
7	-17,350	-13,589	-15,018
8	-17,554	-19,296	-17,695
9	-15,327	-13,634	-14,566

Table A.3: Dlscore score in the period 0 to 100 ns, 100 to 200 ns, and 0 to 200 ns.

Complex	Score 0 - 100 ns	Score 100 - 200 ns	Score 0 - 200 ns
1	7.400	7.307	7.354
2	6.837	6.587	6.712
3	7.130	7.185	7.158
4	6.950	6.854	6.902
5	7.287	6.881	7.084
6	6.628	5.997	6.312
7	7.328	6.525	6.926
8	6.686	5.688	6.187
9	6.460	6.063	6.261

Table A.4: Nnscore score in the period 0 to 100 ns, 100 to 200 ns, and 0 to 200 ns.

Complex	Score 0 - 100 ns	Score 100 - 200 ns	Score 0 - 200 ns
1	7.616	7.884	7.750
2	6.768	5.928	6.348
3	7.056	7.317	7.186
4	6.307	6.613	6.460
5	6.716	6.004	6.360
6	6.617	5.822	6.220
7	6.304	5.893	6.099
8	5.886	4.255	5.071
9	6.105	5.519	5.812

Table A.5: Smina score in the period 0 to 100 ns, 100 to 200 ns, and 0 to 200 ns.

Complex	Score 0 - 100 ns	Score 100 - 200 ns	Score 0 - 200 ns
1	-6.925	-7,067	-6,996
2	-6.302	-4,736	-5,519
3	-6.949	-7,247	-7,098
4	-6.117	-6,160	-6,139
5	-6.372	-5,741	-6,056
6	-6.058	-4,537	-5,298
7	-6.288	-4,777	-5,533
8	-5.581	-0,402	-2,991
9	-5.404	-4,807	-5,105

Table A.6: Vina score in the period 0 to 100 ns, 100 to 200 ns, and 0 to 200 ns.

Complex	Score 0 - 100 ns	Score 100 - 200 ns	Score 0 - 200 ns
1	-6.696	-6.833	-6.765
2	-6.095	-4.585	-5.340
3	-6.732	-7.018	-6.875
4	-5.936	-5.975	-5.956
5	-6.108	-5.504	-5.806
6	-5.800	-4.341	-5.071
7	-6.050	-4.597	-5.323
8	-5.339	-0.386	-2.862
9	-5.168	-4.597	-4.882

References

[1] Adam Urbach and Vijay Ramalingam. Molecular recognition of amino acids, peptides, and proteins by cucurbit[n]uril receptors. *Israel Journal of Chemistry*, 51:664 – 678, 05 2011.

[2] D Tesauro, A Accardo, and Diaferia C. Peptide-based drug-delivery systems in biotechnological applications: Recent advances and perspectives. *Molecules*, 2019.

[3] Patricia Méndez-Samperio. Peptidomimetics as a new generation of antimicrobial agents: Current progress. *Infection and drug resistance*, 7:229–37, 08 2014.

[4] Jane Hua, R.W. Scott, and Gill Diamond. Activity of antimicrobial peptide mimetics in the oral cavity: Ii. activity against periopathogenic biofilms and anti-inflammatory activity. *Molecular oral microbiology*, 25:426–32, 12 2010.

[5] Johan Isaksson, Bjørn Brandsdal, Magnus Engqvist, Gøril Flaten, John Svendsen, and Wenche Stensen. A synthetic antimicrobial peptidomimetic (ltx 109): Stereochemical impact on membrane disruption. *Journal of medicinal chemistry*, 54:5786–95, 08 2011.

[6] Daniel Meister, Nazanin Taimoory, and John Trant. Unnatural amino acids improve affinity and modulate immunogenicity: Developing peptides to treat mhc type ii autoimmune disorders. *Peptide Science*, 111:e24058, 03 2018.

[7] Krzysztof Bzymek, Kendra Avery, Yuelong Ma, David Horne, and John Williams. Natural and non-natural amino-acid side-chain substitutions: Affinity and diffraction studies of meditope-fab complexes. *Acta Crystallographica. Section F, Structural Biology Communications*, 72:820–830, 10 2016.

[8] Merel van de Plassche, Thomas J. O'Neill, Thomas Seeholzer, Boris Turk, Daniel Krappmann, and Steven Verhelst. Use of non-natural amino acids for the design and synthesis of a selective, cell-permeable malt1 activity-based probe. *Journal of Medicinal Chemistry*, 63:3996–4004, 03 2020.

[9] M Poreba, Paulina Kasperkiewicz, Scott Snipas, Domenico Fasci, G Salvesen, and Marcin Drag. Unnatural amino acids increase sensitivity and provide for the design of highly selective caspase substrates. *Cell death and differentiation*, 21, 05 2014.

[10] Minwoo Yang and Woon Song. Diverse protein assembly driven by metal and chelating amino acids with selectivity and tunability. *Nature communications*, 10:5545, 12 2019.

[11] J.S. Ma. Unnatural amino acids in drug discovery. *Chimica Oggi*, 21:65–68, 06 2003.

[12] Tamara Hendrickson, Valérie Crécy-Lagard, and Paul Schimmel. Incorporation of nonnatural amino acids into proteins. *Annu Rev Biochem*, 73:147 – 176, 11 2003.

[13] Ellen Minnihan, Kenichi Yokoyama, and Joanne Stubbe. Unnatural amino acids: better than the real things? *F1000 biology reports*, 1:88, 11 2009.

[14] Pamela England. Unnatural amino acid mutagenesis: A precise tool for probing protein structure and function †. *Biochemistry*, 43:11623–9, 10 2004.

[15] Qian Wang, Angela R Parrish, and Lei Wang. Expanding the genetic code for biological studies. *Chem. Biol.*, 16(3):323–336, March 2009.

[16] Shaheena Parween, Ashraf Ali, and Virander Chauhan. Non-natural amino acids containing peptide-capped gold nanoparticles for drug delivery application. *ACS applied materials interfaces*, 5, 06 2013.

[17] Michael Goldflam and Christopher G. Ullman. Recent advances toward the discovery of drug-like peptides de novo. *Frontiers in Chemistry*, 3, 2015.

[18] Aijun Wang, Natalie Nairn, and Marcello Marelli. *Protein Engineering with Non-Natural Amino Acids*. 02 2012.

[19] William H Zhang, Gottfried Otting, and Colin J Jackson. Protein engineering with unnatural amino acids. *Current Opinion in Structural Biology*, 23(4):581–587, 2013.

[20] Inchan Kwon and Sung Lim. Non-natural amino acids for protein engineering and new protein chemistries. *Macromolecular Chemistry and Physics*, 214:1295–1301, 06 2013.

[21] Anup Adhikari, Bibek Bhattarai, Ashika Aryal, Niru Thapa, Puja Kc, Ashma Adhikari, Sushila Maharjan, Prem Chanda, Bishnu Regmi, and Niranjan Parajuli. Reprogramming natural proteins using unnatural amino acids. *RSC Advances*, 11:38126–38145, 11 2021.

[22] David Gfeller, Olivier Michielin, and Vincent Zoete. Swisssidechain: A molecular and structural database of non-natural sidechains. *Nucleic acids research*, 41, 10 2012.

[23] George Khoury, James Smadbeck, Phanourios Tamamis, Andrew Vandris, Chris Kieslich, and Christodoulos Floudas. Forcefield ncaa: Ab initio charge parameters to aid in the discovery and design of therapeutic proteins and peptides with unnatural amino acids and their application to complement inhibitors of the compstatin family. *ACS Synthetic Biology*, 2014.

[24] Sandeep Singh, Harinder Singh, Abhishek Tuknait, Kumardeep Chaudhary, Balvinder Singh, S. Kumaran, and Gajendra Raghava. Pepstrmod: Structure prediction of peptides containing natural, non-natural and modified residues. *Biology Direct*, 10:73, 12 2015.

[25] Rodrigo Ochoa, Alessandro Laio, and Pilar Cossio. Predicting the affinity of peptides to mhc class ii by scoring molecular dynamics simulations. *Journal of Chemical Information and Modeling*, 59, 07 2019.

[26] Rodrigo Ochoa, V Lunardelli, D Rosa, Alessandro Laio, and Pilar Cossio. Multiple-allele mhc class ii epitope engineering by a molecular dynamics-based evolution protocol. *Frontiers in immunology*, 13, 2022.

[27] Olujide O. Olubiyi and Birgit Strodel. Topology and parameter data of thirteen non-natural amino acids for molecular simulations with charmm22. *Data in Brief*, 9:642–647, 2016.

[28] Xiaowen Wang and Wenjin Li. Development and testing of force field parameters for phenylalanine and tyrosine derivatives. *Frontiers in Molecular Biosciences*, 7, 2020.

[29] Joan Gimenez-Dejoz and Keiji Numata. Molecular dynamics study of the internalization of cell-penetrating peptides containing unnatural amino acids across membranes. *Nanoscale Adv.*, 4:397–407, 2022.

[30] Sam Giannakoulias, Sumant Shringari, John Ferrie, and E. Petersson. Biomolecular simulation based machine learning models accurately predict sites of tolerability to the unnatural amino acid acridonylalanine. *Scientific Reports*, 11:18406, 09 2021.

[31] Haibo Yu, Xavier Daura, and Wilfred F. van Gunsteren. Molecular dynamics simulations of peptides containing an unnatural amino acid: Dimerization, folding, and protein binding. *Proteins: Structure*, 54, 2004.

[32] Tiffany Clark, Libero (Lee) Bartolotti, and Rickey Hicks. The application of dosy nmr and molecular dynamics simulations to explore the mechanism(s) of micelle binding of antimicrobial peptides containing unnatural amino acids. *Biopolymers*, 99:548–561, 08 2013.

[33] David L. Nelson and Michael M. Cox. *Lehninger Principles of Biochemistry, Fourth Edition*. W. H. Freeman, 4 edition, April 2004.

[34] Center for BioMolecular Modeling. Primary Structure amino acids are the building blocks of proteins, 2002.

[35] MedlinePlus. Amino acids, 2022.

[36] Russell Peter. *Igenetics : A Molecular Approach*. San Francisco; Benjamin Cummings, 2010.

[37] Ian W. Hamley. *Introduction to Peptide Science*. Wiley, 2020.

[38] J Forbes and K Krishnamurthy. *Biochemistry, Peptide. In StatPearls*. StatPearls Publishing, 2021.

[39] W. BromerH. EggeK. EiterR. EyjólfssonD. GrossH. HikinoY. HikinoB. G. JacksonR. B. MorinJ. E. PikeE. W. WarnhoffH. WiegandtE. Wong. *Fortschritte der Chemie Organischer Naturstoffe / Progress in the Chemistry of Organic Natural Products*. Spring, 2013.

[40] Jun Lei, Lichun Sun, Siyu Huang, Chenhong Zhu, Ping Li, Jun He, Vienna Mackey, David Coy, and Quanyong He. The antimicrobial peptides and their potential clinical applications. *American journal of translational research*, 11:3919–3931, 07 2019.

[41] Olivier Mirabeau, Emerald Perlas, Cinzia Severini, Enrica Audero, Olivier Gascuel, Roberta Possenti, Ewan Birney, Nadia Rosenthal, and Cornelius Gross. Identification of novel peptide hormones in the human proteome by hidden markov model screening. *Genome research*, 17:320–7, 04 2007.

[42] Yoshikatsu Matsubayashi. Exploring peptide hormones in plants: Identification of four peptide hormone-receptor pairs and two post-translational modification enzymes. *Proceedings of the Japan Academy, Series B*, 94:59–74, 02 2018.

[43] Ramak Esfandi, Mallory Walters, and Apollinaire Tsopmo. Antioxidant properties and potential mechanisms of hydrolyzed proteins and peptides from cereals. *Heliyon*, 5:e01538, 04 2019.

[44] Jerry Palmer. C-peptide in the natural history of type 1 diabetes. *Diabetes/metabolism research and reviews*, 25:325–8, 05 2009.

[45] Corvalan Claudia. Why does predicting a protein's 3d structure matter? the biological impact of ai system alphafold. `https://piip.co.kr/`. Accessed: 2020-12-10.

[46] Helen Berman, Kim Henrick, and Haruki Nakamura. Berman, h, henrick, k and nakamura, h. announcing the worldwide protein data bank. nat struct biol 10: 980. *Nature structural biology*, 10:980, 01 2004.

[47] Helen Berman, Kim Henrick, Haruki Nakamura, and John Markley. The worldwide protein data bank (wwpdb): ensuring a single, uniform archive of pdb data. nucleic acids res 35:d301-d303. *Nucleic acids research*, 35:D301–3, 02 2007.

[48] wwPDB consortium. Protein Data Bank: the single global archive for 3D macromolecular structure data. *Nucleic Acids Research*, 47(D1):D520–D528, 10 2018.

[49] Roman Laskowski. Protein structure databases. *Molecular biotechnology*, 48:183–98, 06 2011.

[50] Helen Berman and Lila Gierasch. How the protein data bank changed biology: An introduction to the jbc reviews thematic series, part 1. *Journal of Biological Chemistry*, 296:100608, 03 2021.

[51] Stephen Burley. Impact of structural biologists and the protein data bank on small-molecule drug discovery and development. *Journal of Biological Chemistry*, 296:100559, 03 2021.

[52] Andrew Senior, Richard Evans, John Jumper, James Kirkpatrick, Laurent Sifre, Tim Green, Chongli Qin, Augustin Žídek, Alexander Nelson, Alex Bridgland, Hugo Penedones, Stig Petersen, Karen Simonyan, Steve Crossan, Pushmeet Kohli, David Jones, David Silver, Koray Kavukcuoglu, and Demis Hassabis. Improved protein structure prediction using potentials from deep learning. *Nature*, 577:1–5, 01 2020.

[53] Walther Caminati and Jens-Uwe Grabow. Chapter 15 - microwave spectroscopy: Molecular systems. In Jaan Laane, editor, *Frontiers of Molecular Spectroscopy*, pages 455–552. Elsevier, Amsterdam, 2009.

[54] Xing Du, Yi Li, Yuan-Ling Xia, Shi-Meng Ai, Jing Liang, Xing-Lai Ji, and Shu-Qun Liu. Insights into protein–ligand interactions: Mechanisms, models, and methods. *International Journal of Molecular Sciences*, 17:144, 01 2016.

[55] Xianjin Xu and Xiaoqin Zou. Predicting protein–peptide complex structures by accounting for peptide flexibility and the physicochemical environment. *Journal of Chemical Information and Modeling*, 62, 12 2021.

[56] Saravanan Rajan and Sachdev S. Sidhu. Chapter 1 - simplified synthetic antibody libraries. In K. Dane Wittrup and Gregory L. Verdine, editors, *Protein Engineering for Therapeutics, Part A*, volume 502 of *Methods in Enzymology*, pages 3–23. Academic Press, 2012.

[57] Pijush Samui. *Handbook of Research on Advanced Computational Techniques for Simulation-Based Engineering*. 12 2019.

[58] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon Kohl, Andrew Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, and Demis Hassabis. Highly accurate protein structure prediction with alphafold. *Nature*, 596:1–11, 08 2021.

[59] Mihaly Varadi, Stephen Anyango, Mandar Deshpande, Sreenath Nair, Cindy Natassia, Galabina Yordanova, David Yuan, Oana Stroe, Gemma Wood, Agata Laydon, Augustin Žídek, Tim Green, Kathryn Tunyasuvunakool, Stig Petersen, John Jumper, Ellen Clancy, Richard Green, Ankur Vora, Mira Lutfi, and Sameer Velankar. Alphafold protein structure database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic acids research*, 50, 11 2021.

[60] Bernard Monasse and Frédéric Boussinot. Determination of forces from a potential in molecular dynamics, 2014.

[61] Benedict Leimkuhler, Sebastian Reich, Konrad-zuse Zentrum, Heilbronner Str, and Robert Skeel. Integration methods for molecular dynamics. 82, 02 1995.

[62] Scott Hollingsworth and Ron Dror. Molecular dynamics simulation for all. *Neuron*, 99:1129–1143, 09 2018.

[63] Anthony Clark, Pratyush Tiwary, Ken Borrelli, Shulu Feng, Edward Miller, Robert Abel, Richard Friesner, and Bruce Berne. Prediction of protein-ligand binding poses via a combination of induced fit docking and metadynamics simulations. *Journal of chemical theory and computation*, 12, 05 2016.

[64] Mark Abraham, Teemu Murtola, Roland Schulz, Szilárd Páll, Jeremy Smith, Berk Hess, and Erik Lindahl. Gromacs: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX*, 1, 07 2015.

[65] B. Brooks, C. Brooks, Alexander MacKerell, L Nilsson, Robert Petrella, B Roux, Y Won, Georgios Archontis, Christian Bartels, Boresch S, Amedeo Caflisch, Leo Caves, Qiang Cui, Dinner AR, Feig M, Stefan Fischer, Gao J, Hodoscek M, W Im, and M. Karplus. Charmm: The biomolecular simulation program. *Journal of computational chemistry*, 30:1545–614, 06 2009.

[66] David Case, Thomas Cheatham, Tom Darden, Holger Gohlke, Ray Luo, Kenneth Merz, Alexey Onufriev, Carlos Simmerling, Bing Wang, and Robert Woods. The amber biomolecular simulation programs. *Journal of computational chemistry*, 26:1668–88, 12 2005.

[67] P.H. Hünenberger and J.A. McCammon. Effect of artificial periodicity in simulations of biomolecules under ewald boundary conditions: A continuum electrostatics study. *Biophysical chemistry*, 78:69–88, 05 1999.

[68] Josiah Willard Gibbs. *Elementary Principles in Statistical Mechanics: Developed with Especial Reference to the Rational Foundation of Thermodynamics*. Cambridge Library Collection - Mathematics. Cambridge University Press, 2010.

[69] Philippe H. Hünenberger. *Thermostat Algorithms for Molecular Dynamics Simulations*, pages 105–149. Springer Berlin Heidelberg, Berlin, Heidelberg, 2005.

[70] Herman Berendsen, J.P.M. Postma, Wilfred van Gunsteren, AD DiNola, and J.R. Haak. Molecular-dynamics with coupling to an external bath. *The Journal of Chemical Physics*, 81:3684, 10 1984.

[71] A. Lemak and Nikolay Balabaev. On the berendsen thermostat. *Molecular Simulation*, 13:177–187, 09 1994.

[72] Giovanni Bussi, Davide Donadio, and Michele Parrinello. Canonical sampling through velocity rescaling. *The Journal of chemical physics*, 126:014101, 02 2007.

[73] Denis Evans and Brad Holian. The nose–hoover thermostat. *The Journal of Chemical Physics*, 83:4069–4074, 10 1985.

[74] Hideki Tanaka, Koichiro Nakanishi, and Nobuatsu Watanabe. Constant temperature molecular dynamics calculation on lennard-jones fluid and its application to water. *Chemical Physics - CHEM PHYS*, 78:2626–2634, 03 1983.

[75] Marco Rozgic. Integration methods, thermostats and barostats in molecular dynamics. 09 2015.

[76] M. Parrinello and A. Rahman. Crystal structure and pair potentials: A molecular-dynamics study. *Physical Review Letters - PHYS REV LETT*, 45:1196–1199, 10 1980.

[77] Glenn Martyna, Douglas Tobias, and Michael Klein. Constant-pressure molecular-dynamics algorithms. *The Journal of Chemical Physics*, 101, 09 1994.

[78] Glenn Martyna, Mark Tuckerman, D. Tobias, and M. Klein. Explicit reversible integrators for extended systems dynamics. *Molecular Physics*, 87:1117–1157, 04 1996.

[79] Olgun Guvench and Alexander MacKerell. Comparison of protein force fields for molecular dynamics simulations. *Methods in molecular biology (Clifton, N.J.)*, 443:63–88, 02 2008.

[80] Jacob Durrant and J McCammon. Molecular dynamics simulations and drug discovery. *BMC biology*, 9:71, 10 2011.

[81] Yong Duan, Chun Wu, Shibasish Chowdhury, Mathew Lee, Guoming Xiong, Wei Zhang, Rong Yang, Piotr Cieplak, Ray Luo, Tai-Sung Lee, James Caldwell, Junmei Wang, and Peter Kollman. A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. *Journal of computational chemistry*, 24:1999–2012, 12 2003.

[82] Yunshuo Guo. Introduction of the amber force field. 11 2019.

[83] George Khoury, Jeff Thompson, James Smadbeck, Chris Kieslich, and Christodoulos Floudas. Forcefield ptm: Ab initio charge and amber forcefield parameters for frequently occurring post-translational modifications. *Journal of chemical theory and computation*, 9:5653–5674, 12 2013.

[84] Aaron Oakley, Timothy Isgro, and Yi Wang. Topology file tutorial, 2012.

[85] Harpreet Saini, Aarti Garg, and Gajendra Raghava. Pepstr: A de novo method for tertiary structure prediction of small bioactive peptides. *Protein and peptide letters*, 14:626–31, 02 2007.

[86] Khaled Barakat, Jonathan Mane, and Jack Tuszynski. *Virtual Screening: An Overview on Methods and Applications*, pages 28–60. 01 2011.

[87] Minyi Su, Qifan Yang, Yu Du, Feng Guoqin, Zhihai Liu, Yan Li, and Renxiao Wang. Comparative assessment of scoring functions: The casf-2016 update. *Journal of Chemical Information and Modeling*, 59, 11 2018.

[88] Iain Moal, Rocco Moretti, David Baker, and Juan Fernandez-Recio. Scoring functions for protein-protein interactions. *Current opinion in structural biology*, 23, 07 2013.

[89] Philippe Ferrara, Holger Gohlke, Daniel Price, Gerhard Klebe, and Charles Brooks. Assessing scoring functions for protein–ligand interactions. *Journal of medicinal chemistry*, 47:3032–47, 07 2004.

[90] Zhiqiang Yan and Jin Wang. *Scoring Functions of Protein-Ligand Interactions*. 01 2017.

[91] Jin Li, Ailing Fu, and Le Zhang. An overview of scoring functions used for protein–ligand interactions in molecular docking. *Interdisciplinary Sciences: Computational Life Sciences*, 11, 03 2019.

[92] Isabella Guedes, Andre Barreto, Diogo Marinho, Eduardo Krempser, Melaine Kuenemann, Olivier Sperandio, Laurent Dardenne, and Maria Miteva. New machine learning and physics-based scoring functions for drug discovery. *Scientific Reports*, 11, 02 2021.

[93] Jui-Chih Wang and Jung-Hsin Lin. Scoring functions for prediction of protein-ligand interactions. *Current pharmaceutical design*, 19, 09 2012.

[94] Lukas Pason and Christoph Sotriffer. Empirical scoring functions for affinity prediction of protein-ligand complexes. *Molecular Informatics*, 35, 07 2016.

[95] Isabella Guedes, Felipe Pereira, and Laurent Dardenne. Empirical scoring functions for structure-based virtual screening: Applications, critical aspects, and challenges. *Frontiers in Pharmacology*, 9, 09 2018.

[96] Zhiqiang Yan and Jin Wang. *Scoring Functions of Protein-ligand Interactions*. 01 2016.

[97] Hongjian Li, Kam-Heung Sze, Gang Lu, and Pedro Ballester. Machine-learning scoring functions for structure-based virtual screening. *WIREs Computational Molecular Science*, 11, 04 2020.

[98] Mahmudulla Hassan, Daniel Mogollon, Olac Fuentes, and suman sirimulla. Dlscore: A deep learning model for predicting protein-ligand binding affinities. 01 2018.

[99] Chensi Cao, Feng Liu, Hai Tan, Deshou Song, Wenjie Shu, Weizhong Li, Yiming Zhou, Xiaochen Bo, and Zhi Xie. Deep learning and its applications in biomedicine. *Genomics, Proteomics Bioinformatics*, 16, 03 2018.

[100] Pin Chen, Yaobin Ke, Yutong Lu, Yunfei Du, Jiahui Li, Hui Yan, Huiying Zhao, Yaoqi Zhou, and Yuedong Yang. Dligand2: an improved knowledge-based energy function for protein–ligand interactions using the distance-scaled, finite, ideal-gas reference state. *Journal of Cheminformatics*, 11, 08 2019.

[101] Yang Cao and Li Lei. Improved protein-ligand binding affinity prediction by using a curvature-dependent surface-area model. *Bioinformatics (Oxford, England)*, 30, 02 2014.

[102] Renxiao Wang, Xueliang Fang, Yipin Lu, Chao-Yie Yang, and Shaomeng Wang. The pdbbind database: Methodologies and updates. *Journal of medicinal chemistry*, 48:4111–9, 07 2005.

[103] Jacob Durrant and J McCammon. Nnscore 2.0: A neural-network receptor-ligand scoring function. *Journal of chemical information and modeling*, 51:2897–903, 11 2011.

[104] Jacob Durrant and J McCammon. Nnscore: A neural-network-based scoring function for the characterization of protein-ligand complexes. *Journal of chemical information and modeling*, 50:1865–71, 10 2010.

[105] Oleg Trott and Arthur J. Olson. Autodock vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of Computational Chemistry*, 31, 2010.

[106] David Koes, Matthew Baumgartner, and Carlos Camacho. Lessons learned in empirical scoring with smina from the csar 2011 benchmarking exercise. *Journal of chemical information and modeling*, 53, 02 2013.

[107] Barcelona Field Studies Centre. Spearman's Rank Correlation Coefficient, 2000.

[108] Ronald N. Forthofer, Eun Sul Lee, and Mike Hernandez. 3 - descriptive methods. In Ronald N. Forthofer, Eun Sul Lee, and Mike Hernandez, editors, *Biostatistics (Second Edition)*, pages 21–69. Academic Press, San Diego, second edition edition, 2007.

[109] Laerd Statistics. Spearman's Rank-Order Correlation, 2020.

[110] Stephen Kokoska and Daniel Zwillinger. Crc standard probability and statistics tables and formulae, student edition. 1999.

[111] Andrew P. King and Robert J. Eckersley. Chapter 2 - descriptive statistics ii: Bivariate and multivariate statistics. In Andrew P. King and Robert J. Eckersley, editors, *Statistics for Biomedical Engineers and Scientists*, pages 23–56. Academic Press, 2019.

[112] Patrick Schober, Christa Boer, and Lothar Schwarte. Correlation coefficients: Appropriate use and interpretation. *Anesthesia Analgesia*, 126:1, 02 2018.

[113] Haldun Akoglu. User's guide to correlation coefficients. *Turkish Journal of Emergency Medicine*, 18(3):91–93, 2018.

[114] Didier Devaurs, Dinler Antunes, and Lydia Kavraki. Computational analysis of complement inhibitor compstatin using molecular dynamics. *Journal of Molecular Modeling*, 26:231, 08 2020.

[115] Arvind Sahu, Dimitrios Morikis, and John Lambris. Compstatin, a peptide inhibitor of complement, exhibits species-specific binding to complement component c3. *Molecular immunology*, 39:557–66, 02 2003.

[116] Madan Katragadda, Dimitrios Morikis, and John D. Lambris. Thermodynamic studies on the interaction of the third complement component and its inhibitor, compstatin*. *Journal of Biological Chemistry*, 279(53):54987–54995, 2004.

[117] Daniel Ricklin, Edimara Reis, Dimitrios Mastellos, Piet Gros, and John Lambris. Complement component c3 – the "swiss army knife" of innate immunity and host defense. *Immunological Reviews*, 274:33–58, 11 2016.

[118] Kjeld Rasmussen, Søren Engelsen, Jesper Fabricius, and Birgit Rasmussen. *The Consistent Force Field: Development of Potential Energy Functions for Conformational Analysis*, pages 381–419. 01 1993.

[119] Titus Beu, Andrada Ailenei, and Alexandra Farcaş. Charmm force field for protonated polyethyleneimine. *Journal of Computational Chemistry*, 39, 10 2018.

[120] Shirley Siu, Kristyna Pluhackova, and Rainer Böckmann. Optimization of the opls-aa force field for long hydrocarbons. *Journal of Chemical Theory and Computation*, 8:1459–1470, 03 2012.

[121] Dimitrios Mastellos, Despina Yancopoulou, Petros Kokkinos, Markus Huber-Lang, George Hajishengallis, Ali-Reza Biglarnia, Florea Lupu, Bo Nilsson, Antonio Risitano, Daniel Ricklin, and John Lambris. Compstatin: A c3-targeted complement inhibitor reaching its prime for bedside intervention. *European journal of clinical investigation*, 45, 02 2015.

[122] Nermin El-Halawany, Abd-El-Monsif A. Shawky, Ahmed F. M. AL-TOHAMY, Lamees Hegazy, Hamdy Abdel-Shafy, Magdy A. Abdel-Latif, Yasser A. Ghazi, Christiane Neuhoff, Dessie Salilew-Wondim, and Karl Schellander. Complement component 3: characterization and association with mastitis resistance in egyptian water buffalo and cattle. *Journal of Genetics*, 96:65–73, 2017.

[123] Iván Arias de la Rosa, Pilar Font, Alejandro Escudero-Contreras, María López-Montilla, Carlos Pérez-Sánchez, Mari Carmen Abalos Aguilera, Lourdes Ladehesa-Pineda, Alejandro Ibáñez-Costa, Carmen Torres-Granados, Yolanda Jimenez-Gomez, Alejandra Patiño-Trives, María Luque, María Castro-Villegas, Jerusalem Calvo-Gutiérrez, Rafaela Ortega-Castro, Chary Lopez-Pedrera, Eduardo Collantes-Estevez, and Nuria Barbarroja. Complement component 3 as biomarker of disease activity and cardiometabolic risk factor in rheumatoid arthritis and spondyloarthritis. *Therapeutic Advances in Chronic Disease*, 11:204062232096506, 10 2020.

[124] Bert Janssen, Els Halff, John Lambris, and Piet Gros. Structure of compstatin in complex with complement component c3c reveals a new mechanism of complement inhibition. *The Journal of biological chemistry*, 282:29241–7, 11 2007.

[125] Daniel Ricklin and John Lambris. Compstatin: A complement inhibitor on its way to clinical application. *Advances in experimental medicine and biology*, 632:273–92, 02 2008.

[126] Bert J.C. Janssen, Els F. Halff, John D. Lambris, and Piet Gros. Structure of compstatin in complex with complement component c3c reveals a new mechanism of complement inhibition*. *Journal of Biological Chemistry*, 282(40):29241–29247, 2007.

[127] William Jorgensen, Jayaraman Chandrasekhar, Jeffry Madura, Roger Impey, and Michael Klein. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.*, 79:926–935, 07 1983.

[128] Michele Pierro, Ron Elber, and Ben Leimkuhler. A stochastic algorithm for the isobaric-isothermal ensemble with ewald summations for all long range forces. *Journal of Chemical Theory and Computation*, 11, 11 2015.

[129] Dusanka Janezic and Franci Merzel. An efficient symplectic integration algorithm for molecular dynamics simulations. *Journal of Chemical Information and Computer Sciences*, 35:321–326, 03 1995.

[130] Line of best fit: Definition, equation examples. https://study.com/academy/lesson/line-of-best-fit-definition-equation-examples.html. Accessed: 2022-6-6.

[131] Rodrigo Ochoa, Miguel Soler, Alessandro Laio, and Pilar Cossio. Assessing the capability of in silico mutation protocols for predicting the finite temperature conformation of amino acids. *Physical Chemistry Chemical Physics*, 20, 10 2018.

[132] Rodrigo Ochoa, Miguel Soler, Alessandro Laio, and Pilar Cossio. Parce: Protocol for amino acid refinement through computational evolution. *Computer Physics Communications*, 260:107716, 03 2021.